

# Correlation and Linear Regression



Exercise 61 lists the movies with the largest world box office sales and their world box office budget. Is there a correlation between the world box office sales for a movie and the total amount spent making the movie? Comment on the association between the two variables. (See Exercise 61 and L02.)

## Learning Objectives

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

- L01** Define the terms *independent variable* and *dependent variable*.
- L02** Calculate, test, and interpret the relationship between two variables using the correlation coefficient.
- L03** Apply regression analysis to estimate the linear relationship between two variables.
- L04** Interpret the regression analysis.
- L05** Evaluate the significance of the slope of the regression equation.
- L06** Evaluate a regression equation to predict the dependent variable.
- L07** Calculate and interpret the coefficient of determination.
- L08** Calculate and interpret confidence and prediction intervals.

## 13.1 Introduction



Chapters 2 through 4 presented *descriptive statistics*. We organized raw data into a frequency distribution and computed several measures of location and measures of dispersion to describe the major characteristics of the distribution. In chapters 5 through 7, we described probability, and from probability statements, we created probability distributions. In Chapter 8, we began the study of *statistical inference*, where we collected a sample to estimate a population parameter such as the population mean or population proportion. In addition, we used the sample data to test an inference or hypothesis about a population mean or a population proportion, the difference between two population means, or the equality of several population means. Each of these tests involved just *one* interval- or ratio-level variable, such as the profit made on a car sale, the income of bank presidents, or the number of patients admitted each month to a particular hospital.

In this chapter, we shift the emphasis to the study of relationships between two interval- or ratio-level variables. In all business fields, identifying and studying relationships between variables can provide information on ways to increase profits, methods to decrease costs, or variables to predict demand. In marketing products, many firms use price reductions through coupons and discount pricing to increase sales. In this example, we are interested in the relationship between two variables: price reductions and sales. To collect the data, a company can test-market a variety of price reduction methods and observe sales. We hope to confirm a relationship that decreasing price leads to increased sales. In economics, you will find many relationships between two variables that are the basis of economics, such as supply and demand and demand and price.

As another familiar example, recall in Section 4.6 in Chapter 4 we used the Applewood Auto Group data to show the relationship between two variables with a scatter diagram. We plotted the profit for each vehicle sold on the vertical axis and the age of the buyer on the horizontal axis. See the statistical software output on page 125. In that diagram, we observed that as the age of the buyer increased, the profit for each vehicle also increased.

Other examples of relationships between two variables are:

- Does the amount Healthtex spends per month on training its sales force affect its monthly sales?
- Is the number of square feet in a home related to the cost to heat the home in January?
- In a study of fuel efficiency, is there a relationship between miles per gallon and the weight of a car?
- Does the number of hours that students study for an exam influence the exam score?

In this chapter, we carry this idea further. That is, we develop numerical measures to express the relationship between two variables. Is the relationship strong or weak? Is it direct or inverse? In addition, we develop an equation to express the relationship between variables. This will allow us to estimate one variable on the basis of another.

To begin our study of relationships between two variables, we examine the meaning and purpose of **correlation analysis**. We continue by developing an equation that will allow us to estimate the value of one variable based on the value of another. This is called **regression analysis**. We will also evaluate the ability of the equation to accurately make estimations.



### Statistics in Action

The space shuttle *Challenger* exploded on January 28, 1986. An investigation of the cause examined four contractors: Rockwell International for the shuttle and engines, Lockheed Martin for ground support, Martin Marietta for the external fuel tanks, and Morton Thiokol for the solid fuel booster rockets. After several months, the investigation blamed the explosion on defective O-rings produced by Morton Thiokol. A study of the contractor's stock

(continued)

prices showed an interesting happenstance. On the day of the crash, Morton Thiokol stock was down 11.86% and the stock of the other three lost only 2 to 3%. Can we conclude that financial markets predicted the outcome of the investigation?

## 13.2 What Is Correlation Analysis?

When we study the relationship between two interval- or ratio-scale variables, we often start with a scatter diagram. This procedure provides a visual representation of the relationship between the variables. The next step is usually to calculate the correlation coefficient. It provides a quantitative measure of the strength of the relationship between two variables. As an example, the sales manager of Copier Sales of America, which has a large sales force throughout the United States and Canada, wants to determine whether there is a relationship between the number of sales calls made in a month and the number of copiers sold that month. The manager selects a random sample of 10 representatives and determines the number of sales calls each representative made. This information is reported in Table 13–1.

**TABLE 13–1** Number of Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

By reviewing the data, we observe that there does seem to be some relationship between the number of sales calls and the number of units sold. That is, the salespeople who made the most sales calls sold the most units. However, the relationship is not “perfect” or exact. For example, Soni Jones made fewer sales calls than Jeff Hall, but she sold more units.

In addition to the graphical techniques in Chapter 4, we will develop numerical measures to precisely describe the relationship between the two variables, sales calls and copiers sold. This group of statistical techniques is called **correlation analysis**.

**CORRELATION ANALYSIS** A group of techniques to measure the relationship between two variables.

The basic idea of correlation analysis is to report the relationship between two variables. The usual first step is to plot the data in a **scatter diagram**. An example will show how a scatter diagram is used.

### Example

Copier Sales of America sells copiers to businesses of all sizes throughout the United States and Canada. Ms. Marcy Bancerc was recently promoted to the position of national sales manager. At the upcoming sales meeting, the sales representatives from all over the country will be in attendance. She would like to impress upon them the importance of making that extra sales call each day. She decides to gather some information on the relationship between the number of sales calls and the number of copiers sold. She selects a random sample of 10 sales representatives and determines the number of sales calls they made last month and the number of copiers they sold. The sample information is reported in Table 13–1. What observations can you make about the relationship between the number of sales calls and the number of copiers sold? Develop a scatter diagram to display the information.

## Solution

**L01** Define the terms *independent variable* and *dependent variable*.

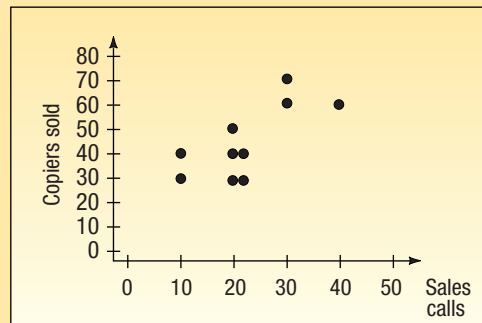
Based on the information in Table 13–1, Ms. Bancerc suspects there is a relationship between the number of sales calls made in a month and the number of copiers sold. Soni Jones sold the most copiers last month, and she was one of three representatives making 30 or more sales calls. On the other hand, Susan Welch and Carlos Ramirez made only 10 sales calls last month. Ms. Welch, along with two others, had the lowest number of copiers sold among the sampled representatives.

The implication is that the number of copiers sold is related to the number of sales calls made. As the number of sales calls increases, it appears the number of copiers sold also increases. We refer to number of sales calls as the **independent variable** and number of copiers sold as the **dependent variable**.

The independent variable provides the basis for estimation. It is the predictor variable. For example, we would like to predict the expected number of copiers sold if a salesperson makes 20 sales calls. Notice that we choose this value. The independent variable is not a random number.

The dependent variable is the variable that is being predicted or estimated. It can also be described as the result or outcome for a known value of the independent variable. The dependent variable is random. That is, for a given value of the independent variable, there are many possible outcomes for the dependent variable. In this example, notice that five different sales representatives made 20 sales calls. The result or outcome of making 20 sales calls is three different values of the dependent variable.

It is common practice to scale the dependent variable (copiers sold) on the vertical or *Y*-axis and the independent variable (number of sales calls) on the horizontal or *X*-axis. To develop the scatter diagram of the Copier Sales of America sales information, we begin with the first sales representative, Tom Keller. Tom made 20 sales calls last month and sold 30 copiers, so  $X = 20$  and  $Y = 30$ . To plot this point, move along the horizontal axis to  $X = 20$ , then go vertically to  $Y = 30$  and place a dot at the intersection. This process is continued until all the paired data are plotted, as shown in Chart 13–1.



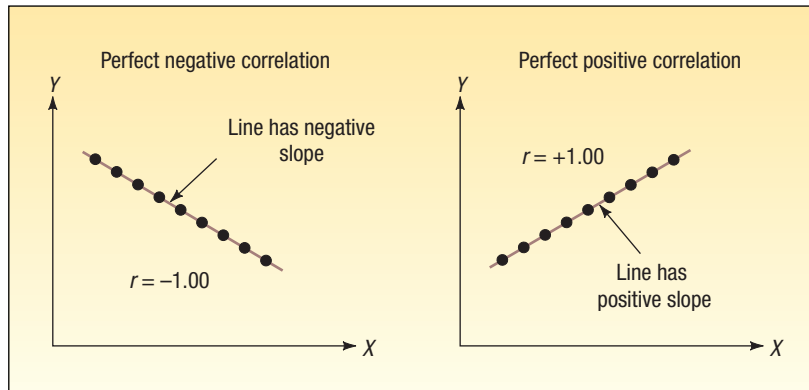
**CHART 13–1** Scatter Diagram Showing Sales Calls and Copiers Sold

The scatter diagram shows graphically that the sales representatives who make more calls tend to sell more copiers. It is reasonable for Ms. Bancerc, the national sales manager at Copier Sales of America, to tell her salespeople that, the more sales calls they make, the more copiers they can expect to sell. Note that, while there appears to be a positive relationship between the two variables, all the points do not fall on a line. In the following section, you will measure the strength and direction of this relationship between two variables by determining the correlation coefficient.

### 13.3 The Correlation Coefficient

**L02** Calculate, test, and interpret the relationship between two variables using the correlation coefficient.

Originated by Karl Pearson about 1900, the **correlation coefficient** describes the strength of the relationship between two sets of interval-scaled or ratio-scaled variables. Designated  $r$ , it is often referred to as *Pearson's  $r$*  and as the *Pearson product-moment correlation coefficient*. It can assume any value from  $-1.00$  to  $+1.00$  inclusive. A correlation coefficient of  $-1.00$  or  $+1.00$  indicates *perfect correlation*. For example, a correlation coefficient for the preceding example computed to be  $+1.00$  would indicate that the number of sales calls and the number of copiers sold are perfectly related in a positive linear sense. A computed value of  $-1.00$  reveals that sales calls and the number of copiers sold are perfectly related in an inverse linear sense. How the scatter diagram would appear if the relationship between the two sets of data were linear and perfect is shown in Chart 13-2.

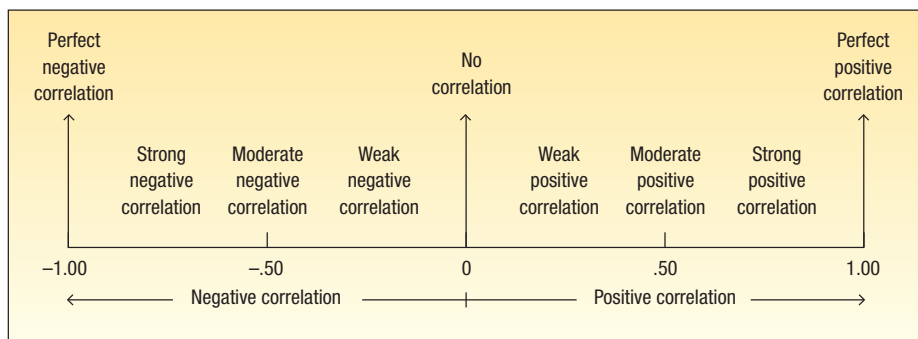


**CHART 13-2** Scatter Diagrams Showing Perfect Negative Correlation and Perfect Positive Correlation

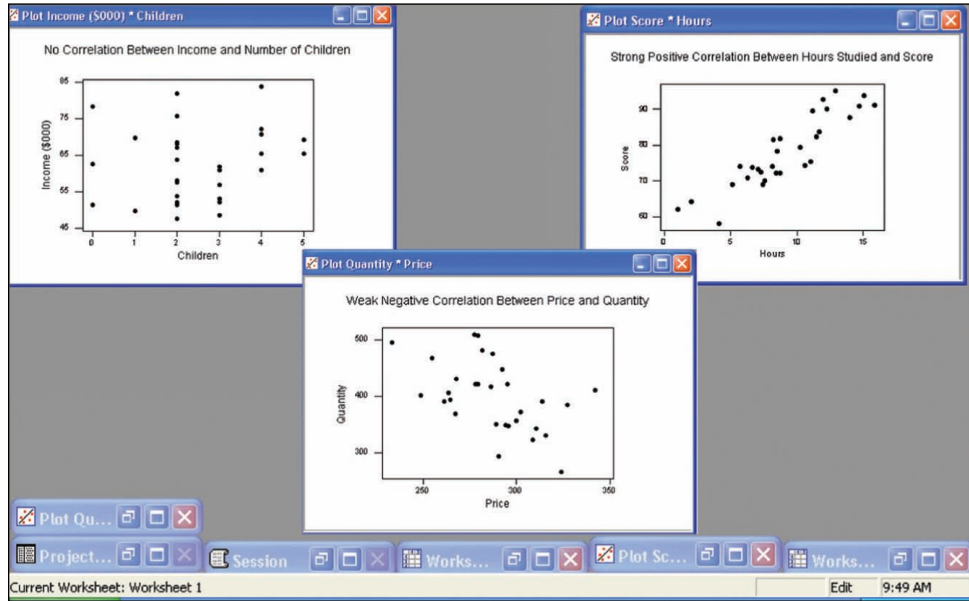
If there is absolutely no relationship between the two sets of variables, Pearson's  $r$  is zero. A correlation coefficient  $r$  close to 0 (say,  $.08$ ) shows that the linear relationship is quite weak. The same conclusion is drawn if  $r = -.08$ . Coefficients of  $-.91$  and  $+.91$  have equal strength; both indicate very strong correlation between the two variables. Thus, *the strength of the correlation does not depend on the direction (either - or +)*.

Scatter diagrams for  $r = 0$ , a weak  $r$  (say,  $-.23$ ), and a strong  $r$  (say,  $+.87$ ) are shown in Chart 13-3. Note that, if the correlation is weak, there is considerable scatter about a line drawn through the center of the data. For the scatter diagram representing a strong relationship, there is very little scatter about the line. This indicates, in the example shown on the chart, that hours studied is a good predictor of exam score.

The following drawing summarizes the strength and direction of the correlation coefficient.



Examples of degrees of correlation



**CHART 13–3** Scatter Diagrams Depicting Zero, Weak, and Strong Correlation

**CORRELATION COEFFICIENT** A measure of the strength of the linear relationship between two variables.

The characteristics of the correlation coefficient are summarized below.

**CHARACTERISTICS OF THE CORRELATION COEFFICIENT**

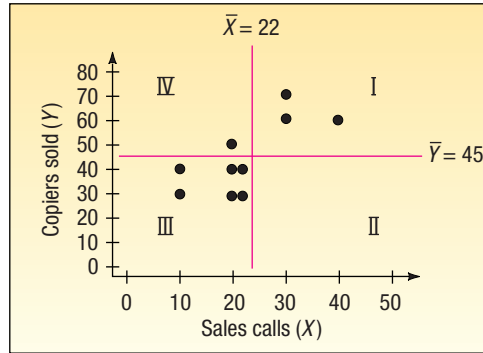
1. The sample correlation coefficient is identified by the lowercase letter *r*.
2. It shows the direction and strength of the linear relationship between two interval- or ratio-scale variables.
3. It ranges from  $-1$  up to and including  $+1$ .
4. A value near  $0$  indicates there is little relationship between the variables.
5. A value near  $1$  indicates a direct or positive relationship between the variables.
6. A value near  $-1$  indicates inverse or negative relationship between the variables.

How is the value of the correlation coefficient determined? We will use the Copier Sales of America data, which are reported in Table 13–2, as an example. We begin

**TABLE 13–2** Sales Calls and Copiers Sold for 10 Salespeople

Sales Representative	Sales Calls, (X)	Copiers Sold, (Y)
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70
Total	220	450

with a scatter diagram, similar to Chart 13-2. Draw a vertical line through the data values at the mean of the  $X$ -values and a horizontal line at the mean of the  $Y$ -values. In Chart 13-4, we've added a vertical line at 22.0 calls ( $\bar{X} = \Sigma X/n = 220/10 = 22$ ) and a horizontal line at 45.0 copiers ( $\bar{Y} = \Sigma Y/n = 450/10 = 45.0$ ). These lines pass through the "center" of the data and divide the scatter diagram into four quadrants. Think of moving the origin from (0, 0) to (22, 45).



**CHART 13-4** Computation of the Correlation Coefficient

Two variables are positively related when the number of copiers sold is above the mean and the number of sales calls is also above the mean. These points appear in the upper-right quadrant (labeled Quadrant I) of Chart 13-4. Similarly, when the number of copiers sold is less than the mean, so is the number of sales calls. These points fall in the lower-left quadrant of Chart 13-4 (labeled Quadrant III). For example, the last person on the list in Table 13-2, Soni Jones, made 30 sales calls and sold 70 copiers. These values are above their respective means, so this point is located in Quadrant I which is in the upper-right quadrant. She made 8 ( $X - \bar{X} = 30 - 22$ ) more sales calls than the mean and sold 25 ( $Y - \bar{Y} = 70 - 45$ ) more copiers than the mean. Tom Keller, the first name on the list in Table 13-2, made 20 sales calls and sold 30 copiers. Both of these values are less than their respective mean; hence this point is in the lower-left quadrant. Tom made 2 less sales calls and sold 15 less copiers than the respective means. The deviations from the mean number of sales calls and for the mean number of copiers sold are summarized in Table 13-3 for the 10 sales representatives. The sum of the products of the deviations from the respective means is 900. That is, the term  $\Sigma(X - \bar{X})(Y - \bar{Y}) = 900$ .

In both the upper-right and the lower-left quadrants, the product of  $(X - \bar{X})(Y - \bar{Y})$  is positive because both of the factors have the same sign. In our example, this

**TABLE 13-3** Deviations from the Mean and Their Products

Sales Representative	Calls, $X$	Sales, $Y$	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	-2	-15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	-2	-5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	-12	-15	180
Carlos Ramirez	10	40	-12	-5	60
Rich Niles	20	40	-2	-5	10
Mike Kiel	20	50	-2	5	-10
Mark Reynolds	20	30	-2	-15	30
Soni Jones	30	70	8	25	200
					<u>900</u>

happens for all sales representatives except Mike Kiel. We can therefore expect the correlation coefficient to have a positive value.

If the two variables are inversely related, one variable will be above the mean and the other below the mean. Most of the points in this case occur in the upper-left and lower-right quadrants, that is, Quadrant II and IV. Now  $(X - \bar{X})$  and  $(Y - \bar{Y})$  will have opposite signs, so their product is negative. The resulting correlation coefficient is negative.

What happens if there is no linear relationship between the two variables? The points in the scatter diagram will appear in all four quadrants. The negative products of  $(X - \bar{X})(Y - \bar{Y})$  offset the positive products, so the sum is near zero. This leads to a correlation coefficient near zero. So, the term  $\sum(X - \bar{X})(Y - \bar{Y})$  drives the strength as well as the sign of the relationship between the two variables.

The correlation coefficient also needs to be unaffected by the units of the two variables. For example, if we had used hundreds of copiers sold instead of the number sold, the correlation coefficient would be the same. The correlation coefficient is independent of the scale used if we divide the term  $\sum(X - \bar{X})(Y - \bar{Y})$  by the sample standard deviations. It is also made independent of the sample size and bounded by the values +1.00 and -1.00 if we divide by  $(n - 1)$ .

This reasoning leads to the following formula:

#### CORRELATION COEFFICIENT

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y}$$

[13-1]

To compute the correlation coefficient, we use the standard deviations of the sample of 10 sales calls and 10 copiers sold. We could use formula (3-12) to calculate the sample standard deviations or we could use a software package. For the specific Excel and Minitab commands, see the **Software Commands** section at the end of Chapter 3. The following is the Excel output. The standard deviation of the number of sales calls is 9.189 and of the number of copiers sold 14.337.

num 1 dstats [Compatibility Mode]									
	A	B	C	D	E	F	G	H	I
1	Calls	Sales		Calls			Sales		
2	20	30		Mean	22.000		Mean	45.000	
3	40	60		Standard Error	2.906		Standard Error	4.534	
4	20	40		Median	20.000		Median	40.000	
5	30	60		Mode	20.000		Mode	30.000	
6	10	30		Standard Deviation	9.189		Standard Deviation	14.337	
7	10	40		Sample Variance	84.444		Sample Variance	205.556	
8	20	40		Kurtosis	0.396		Kurtosis	-1.001	
9	20	50		Skewness	0.601		Skewness	0.566	
10	20	30		Range	30.000		Range	40.000	
11	30	70		Minimum	10.000		Minimum	30.000	
12				Maximum	40.000		Maximum	70.000	
13				Sum	220.000		Sum	450.000	
14				Count	10.000		Count	10.000	
15									

We now insert these values into formula (13-1) to determine the correlation coefficient:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n - 1)s_x s_y} = \frac{900}{(10 - 1)(9.189)(14.337)} = 0.759$$

How do we interpret a correlation of 0.759? First, it is positive, so we conclude there is a direct relationship between the number of sales calls and the number of copiers sold. This confirms our reasoning based on the scatter diagram,



Chart 13–4. The value of 0.759 is fairly close to 1.00, so we conclude that the association is strong.

We must be careful with the interpretation. The correlation of 0.759 indicates a strong positive association between the variables. Ms. Bancroft would be correct to encourage the sales personnel to make that extra sales call, because the number of sales calls made is related to the number of copiers sold. However, does this mean that more sales calls *cause* more sales? No, we have not demonstrated cause and effect here, only that the two variables—sales calls and copiers sold—are related.

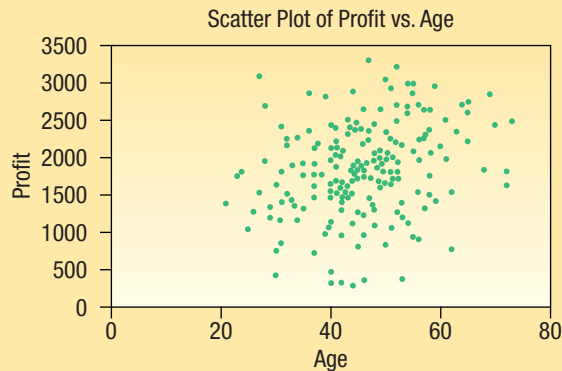
If there is a strong relationship (say, .91) between two variables, we are tempted to assume that an increase or decrease in one variable *causes* a change in the other variable. For example, it can be shown that the consumption of Georgia peanuts and the consumption of aspirin have a strong correlation. However, this does not indicate that an increase in the consumption of peanuts *caused* the consumption of aspirin to increase. Likewise, the incomes of professors and the number of inmates in mental institutions have increased proportionately. Further, as the population of donkeys has decreased, there has been an increase in the number of doctoral degrees granted. Relationships such as these are called **spurious correlations**. What we can conclude when we find two variables with a strong correlation is that there is a relationship or association between the two variables, not that a change in one causes a change in the other.

### Example

The Applewood Auto Group's marketing department believes younger buyers purchase vehicles on which lower profits are earned and the older buyers purchase vehicles on which higher profits are earned. They would like to use this information as part of an upcoming advertising campaign to try to attract older buyers on which the profits tend to be higher. Develop a scatter diagram depicting the relationship between vehicle profits and age of the buyer. Use statistical software to determine the correlation coefficient. Would this be a useful advertising feature?

### Solution

Using the Applewood Auto Group example, the first step is to graph the data using a scatter plot. It is shown in Chart 13-5.



**CHART 13–5** Scatter Diagram of Applewood Auto Group Data

The scatter diagram suggests that a positive relationship does exist between age and profit; however, that relationship does not appear strong.

The next step is to calculate the correlation coefficient to evaluate the relative strength of the relationship. Statistical software provides an easy way to calculate the value of the correlation coefficient. The Excel output follows.

H	I	J	K	L	M
Age	Profit		Applewood Auto Group Correlation Coefficient Between Profit and Age		
21	\$1387				
23	\$1754				
24	\$1817				
25	\$1040				
26	\$1273				$r = 0.262$
27	\$1529				

For this data,  $r = 0.262$ . To evaluate the relationship between a buyer's age and the profit on a car sale:

1. The relationship is positive or direct. Why? Because the sign of the correlation coefficient is positive. This confirms that as the age of the buyer increases, the profit on a car sale also increases.
2. The relationship between the two variables is weak. For a positive relationship, values of the correlation coefficient close to one indicate stronger relationships. In this case,  $r = 0.262$ . It is closer to zero, and we would observe that the relationship is not very strong.

It is not recommended that Applewood use this information as part of an advertising campaign to attract older more profitable buyers.

### Self-Review 13-1



Haverty's Furniture is a family business that has been selling to retail customers in the Chicago area for many years. The company advertises extensively on radio, TV, and the Internet, emphasizing low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising. Below is information on sales and advertising expense for the last four months.

Month	Advertising Expense (\$ million)	Sales Revenue (\$ million)
July	2	7
August	1	3
September	3	8
October	4	10

- (a) The owner wants to forecast sales on the basis of advertising expense. Which variable is the dependent variable? Which variable is the independent variable?
- (b) Draw a scatter diagram.
- (c) Determine the correlation coefficient.
- (d) Interpret the strength of the correlation coefficient.

## Exercises

connect™

1. The following sample observations were randomly selected.




$X$	4	5	3	6	10
$y$	4	6	5	7	7

Determine the correlation coefficient and interpret the relationship between  $X$  and  $Y$ .


2. The following sample observations were randomly selected. 

<b>X</b>	5	3	6	3	4	4	6	8
<b>Y</b>	13	15	7	12	13	11	9	5

Determine the correlation coefficient and interpret the relationship between  $X$  and  $Y$ .

3. Bi-lo Appliance Super-Store has outlets in several large metropolitan areas in New England. The general sales manager aired a commercial for a digital camera on selected local TV stations prior to a sale starting on Saturday and ending Sunday. She obtained the information for Saturday–Sunday digital camera sales at the various outlets and paired it with the number of times the advertisement was shown on the local TV stations. The purpose is to find whether there is any relationship between the number of times the advertisement was aired and digital camera sales. The pairings are: 

Location of TV Station	Number of Airings	Saturday–Sunday Sales (\$ thousands)
Providence	4	15
Springfield	2	8
New Haven	5	21
Boston	6	24
Hartford	3	17

- What is the dependent variable?
  - Draw a scatter diagram.
  - Determine the correlation coefficient.
  - Interpret these statistical measures.
4. The production department of Celltronics International wants to explore the relationship between the number of employees who assemble a subassembly and the number produced. As an experiment, two employees were assigned to assemble the subassemblies. They produced 15 during a one-hour period. Then four employees assembled them. They produced 25 during a one-hour period. The complete set of paired observations follows. 

Number of Assemblers	One-Hour Production (units)
2	15
4	25
1	10
5	40
3	30

The dependent variable is production; that is, it is assumed that different levels of production result from a different number of employees.

- Draw a scatter diagram.
  - Based on the scatter diagram, does there appear to be any relationship between the number of assemblers and production? Explain.
  - Compute the correlation coefficient.
5. The city council of Pine Bluffs is considering increasing the number of police in an effort to reduce crime. Before making a final decision, the council asked the chief of police to survey other cities of similar size to determine the relationship between the number

of police and the number of crimes reported. The chief gathered the following sample information.



City	Police	Number of Crimes	City	Police	Number of Crimes
Oxford	15	17	Holgate	17	7
Starksville	17	13	Carey	12	21
Danville	25	5	Whistler	11	19
Athens	27	7	Woodville	22	6

- Which variable is the dependent variable and which is the independent variable? Hint: If you were the Chief of Police, which variable would you decide? Which variable is the random variable?
  - Draw a scatter diagram.
  - Determine the correlation coefficient.
  - Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?
6. The owner of Maumee Ford-Mercury-Volvo wants to study the relationship between the age of a car and its selling price. Listed below is a random sample of 12 used cars sold at the dealership during the last year.



Car	Age (years)	Selling Price (\$000)	Car	Age (years)	Selling Price (\$000)
1	9	8.1	7	8	7.6
2	7	6.0	8	11	8.0
3	11	3.6	9	10	8.0
4	12	4.0	10	12	6.0
5	8	5.0	11	6	8.6
6	7	10.0	12	6	8.0

- Draw a scatter diagram.
- Determine the correlation coefficient.
- Interpret the correlation coefficient. Does it surprise you that the correlation coefficient is negative?

## 13.4 Testing the Significance of the Correlation Coefficient

Recall that the sales manager of Copier Sales of America found the correlation between the number of sales calls and the number of copiers sold was 0.759. This indicated a strong positive association between the two variables. However, only 10 salespeople were sampled. Could it be that the correlation in the population is actually 0? This would mean the correlation of 0.759 was due to chance. The population in this example is all the salespeople employed by the firm.

Resolving this dilemma requires a test to answer the obvious question: Could there be zero correlation in the population from which the sample was selected? To put it another way, did the computed  $r$  come from a population of paired observations with zero correlation? To continue our convention of allowing Greek letters to represent a population parameter, we will let  $\rho$  represent the correlation in the population. It is pronounced “rho.”

Could the correlation in the population be zero?

We will continue with the illustration involving sales calls and copiers sold. We employ the same hypothesis testing steps described in Chapter 10. The null hypothesis and the alternate hypothesis are:

$$\begin{aligned} H_0: \rho &= 0 && \text{(The correlation in the population is zero.)} \\ H_1: \rho &\neq 0 && \text{(The correlation in the population is different from zero.)} \end{aligned}$$

From the way  $H_1$  is stated, we know that the test is two-tailed. The formula for  $t$  is:

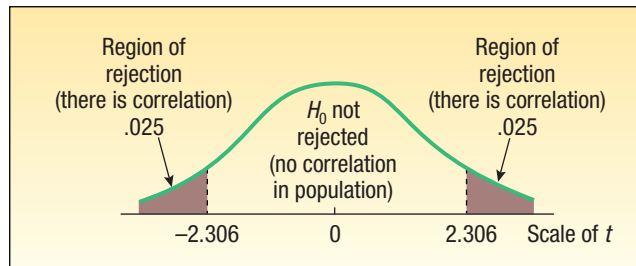
**$t$  TEST FOR THE CORRELATION COEFFICIENT**

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

with  $n - 2$  degrees of freedom

**[13-2]**

Using the .05 level of significance, the decision rule in this instance states that if the computed  $t$  falls in the area between plus 2.306 and minus 2.306, the null hypothesis is not rejected. To locate the critical value of 2.306, refer to Appendix B.2 for  $df = n - 2 = 10 - 2 = 8$ . See Chart 13-6.



**CHART 13-6** Decision Rule for Test of Hypothesis at .05 Significance Level and 8  $df$

Applying formula (13-2) to the example regarding the number of sales calls and units sold:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.759\sqrt{10-2}}{\sqrt{1-.759^2}} = 3.297$$

The computed  $t$  is in the rejection region. Thus,  $H_0$  is rejected at the .05 significance level. Hence we conclude the correlation in the population is not zero. From a practical standpoint, it indicates to the sales manager that there is correlation with respect to the number of sales calls made and the number of copiers sold in the population of salespeople.

We can also interpret the test of hypothesis in terms of  $p$ -values. A  $p$ -value is the likelihood of finding a value of the test statistic more extreme than the one computed, when  $H_0$  is true. To determine the  $p$ -value, go to the  $t$  distribution in Appendix B.2 and find the row for 8 degrees of freedom. The value of the test statistic is 3.297, so in the row for 8 degrees of freedom and a two-tailed test, find the value closest to 3.297. For a two-tailed test at the .02 significance level, the critical value is 2.896, and the critical value at the .01 significance level is 3.355. Because 3.297 is between 2.896 and 3.355, we conclude that the  $p$ -value is between .01 and .02.

Both Minitab and Excel will report the correlation between two variables. In addition to the correlation, Minitab reports the  $p$ -value for the test of hypothesis that the correlation in the population between the two variables is 0. The Minitab output is at the top of the following page.

	C1-T	C2	C3	C4	C5	C6	C7	C8	C9	C10
	Sales Representative	Calls	Sales							
1	Tom Keller	20	30							
2	Jeff Hall	40	60							
3	Brian Virost	20	40							
4	Greg Fish	30	60							
5	Susan Welch	10	30							
6	Carlos Ramirez	10	40							
7	Rich Niles	20	40							
8	Mike Kiel	20	50							
9	Mark Reynolds	20	30							
10	Soni Jones	30	70							

Session	
Welcome to Minitab, press F1 for help.	
<b>Correlations: Calls, Units Sold</b>	
Pearson correlation of Calls and Units Sold = 0.759	
P-Value = 0.011	

## Example

In the Example on page 470, we found that the correlation coefficient between the profit on the sale of a vehicle by the Applewood Auto Group and the age of the person that purchased the vehicle was 0.262. Because the sign of the correlation coefficient was positive, we concluded there was a direct relationship between the two variables. However, because the amount of correlation was low—that is, near zero—we concluded that an advertising campaign directed toward the older buyers, where there is a large profit, was not warranted. Does this mean we should conclude that there is no relationship between the two variables? Use the .05 significance level.

## Solution

To begin to answer the question in the last sentence above, we need to clarify the sample and population issues. Let's assume that the data collected on the 180 vehicles sold by the Applewood Group is a sample from the population of *all* vehicles sold over many years by the Applewood Auto Group. The Greek letter  $\rho$  is the correlation coefficient in the population and  $r$  the correlation coefficient in the sample.

Our next step is to set up the null hypothesis and the alternate hypothesis. We test the null hypothesis that the correlation coefficient is equal to zero. The alternate hypothesis is that there is positive correlation between the two variables.

$$H_0: \rho \leq 0 \quad (\text{The correlation in the population is zero.})$$

$$H_1: \rho > 0 \quad (\text{The correlation in the population is positive})$$

This is a one-tailed test because we are interested in confirming a positive association between the variables. The test statistic follows the  $t$  distribution with  $n - 2$  degrees of freedom, so the degrees of freedom is  $180 - 2 = 178$ . However, 178 degrees of freedom is not in Appendix B.2. The closest value is 180, so we will use that value. Our decision rule is to reject the null hypothesis if the computed value of the test statistic is greater than 1.653.

We use formula 13-2 to find the value of the test statistic.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.262\sqrt{180-2}}{\sqrt{1-0.262^2}} = 3.622$$

Comparing the value of our test statistic of 3.622 to the critical value of 1.653, we reject the null hypothesis. We conclude that the sample correlation coefficient of 0.262 is too large to have come from a population with no correlation. To put our results another way, there is a positive correlation between profits and age in the population.

This result is confusing and seems contradictory. On one hand, we observed that the correlation coefficient did not indicate a very strong relationship and that the Applewood Auto Group marketing department should not use this information for its promotion and advertising decisions. On the other hand, the hypothesis test indicated that the correlation coefficient is not equal to zero and that a positive relationship between age and profit exists. How can this be? We must be very careful about the interpretation of the hypothesis test results. The conclusion is that the correlation coefficient is not equal to zero and that there is a positive relationship between the amount of profit earned and the age of the buyer. The result of the hypothesis test only shows that a relationship exists. The hypothesis test makes no claims regarding the *strength* of the relationship.

### Self-Review 13-2



A sample of 25 mayoral campaigns in medium-sized cities with populations between 50,000 and 250,000 showed that the correlation between the percent of the vote received and the amount spent on the campaign by the candidate was .43. At the .05 significance level, is there a positive association between the variables?

## Exercises

connect™

7. The following hypotheses are given.

$$H_0: \rho \leq 0$$

$$H_1: \rho > 0$$

A random sample of 12 paired observations indicated a correlation of .32. Can we conclude that the correlation in the population is greater than zero? Use the .05 significance level.

8. The following hypotheses are given.

$$H_0: \rho \geq 0$$

$$H_1: \rho < 0$$

A random sample of 15 paired observations have a correlation of  $-.46$ . Can we conclude that the correlation in the population is less than zero? Use the .05 significance level.

9. Pennsylvania Refining Company is studying the relationship between the pump price of gasoline and the number of gallons sold. For a sample of 20 stations last Tuesday, the correlation was .78. At the .01 significance level, is the correlation in the population greater than zero?
10. A study of 20 worldwide financial institutions showed the correlation between their assets and pretax profit to be .86. At the .05 significance level, can we conclude that there is positive correlation in the population?
11. The Airline Passenger Association studied the relationship between the number of passengers on a particular flight and the cost of the flight. It seems logical that more passengers on the flight will result in more weight and more luggage, which in turn will result in higher fuel costs. For a sample of 15 flights, the correlation between the number of passengers and total fuel cost was .667. Is it reasonable to conclude that there is positive association in the population between the two variables? Use the .01 significance level.
12. The Student Government Association at Middle Carolina University wanted to demonstrate the relationship between the number of beers a student drinks and their blood alcohol content (BAC). A random sample of 18 students participated in a study in which each participating student was randomly assigned a number of 12-ounce cans of beer to drink. Thirty minutes after consuming their assigned number of beers a member of the

local sheriff's office measured their blood alcohol content. The sample information is reported below.



Student	Beers	BAC	Student	Beers	BAC
1	6	0.10	10	3	0.07
2	7	0.09	11	3	0.05
3	7	0.09	12	7	0.08
4	4	0.10	13	1	0.04
5	5	0.10	14	4	0.07
6	3	0.07	15	2	0.06
7	3	0.10	16	7	0.12
8	6	0.12	17	2	0.05
9	6	0.09	18	1	0.02

Use a statistical software package to answer the following questions.

- Develop a scatter diagram for the number of beers consumed and BAC. Comment on the relationship. Does it appear to be strong or weak? Does it appear to be positive or inverse?
- Determine the correlation coefficient.
- At the .01 significance level, is it reasonable to conclude that there is a positive relationship in the population between the number of beers consumed and the BAC? What is the  $p$ -value?

## 13.5 Regression Analysis

**L03** Apply regression analysis to estimate the linear relationship between two variables.

In the previous sections of this chapter, we evaluated the direction and the significance of the linear relationship between two variables by finding the correlation coefficient. If the correlation coefficient is significantly different from zero, then the next step is to develop an equation to express the *linear* relationship between the two variables. Using this equation, we will be able to estimate the value of the dependent variable  $Y$  based on a selected value of the independent variable  $X$ . The technique used to develop the equation and provide the estimates is called **regression analysis**.

In Table 13–1, we reported the number of sales calls and the number of units sold for a sample of 10 sales representatives employed by Copier Sales of America. Chart 13–1 portrayed this information in a scatter diagram. Recall that we tested the significance of the correlation coefficient ( $r = 0.759$ ) and concluded that a significant relationship exists between the two variables. Now we want to develop a linear equation that expresses the relationship between the number of sales calls, the independent variable, and the number of units sold, the dependent variable. The equation for the line used to estimate  $Y$  on the basis of  $X$  is referred to as the **regression equation**.



**REGRESSION EQUATION** An equation that expresses the linear relationship between two variables.

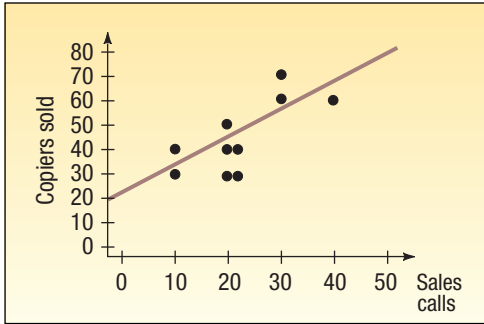
### Least Squares Principle

In regression analysis, our objective is to use the data to position a line that best represents the relationship between the two variables. Our first approach is to use a scatter diagram to visually position the line.

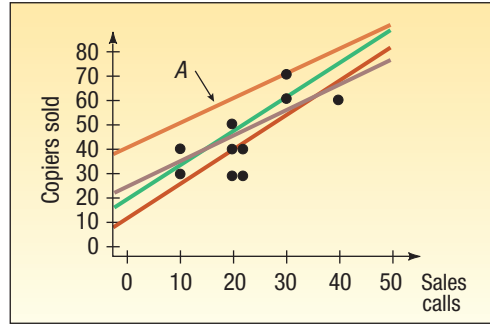
The scatter diagram in Chart 13–1 is reproduced in Chart 13–7, with a line drawn with a ruler through the dots to illustrate that a line would probably fit the data.



However, the line drawn using a straight edge has one disadvantage: Its position is based in part on the judgment of the person drawing the line. The hand-drawn lines in Chart 13–8 represent the judgments of four people. All the lines except line A seem to be reasonable. That is, each line is centered among the graphed data. However, each would result in a different estimate of units sold for a particular number of sales calls.



**CHART 13–7** Sales Calls and Copiers Sold for 10 Sales Representatives



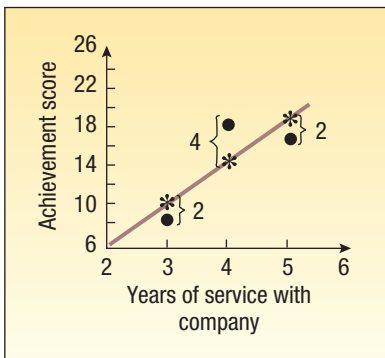
**CHART 13–8** Four Lines Superimposed on the Scatter Diagram

However, we would prefer a method that results in a single, best regression line. This method is called the least squares principle. It gives what is commonly referred to as the “best-fitting” line.

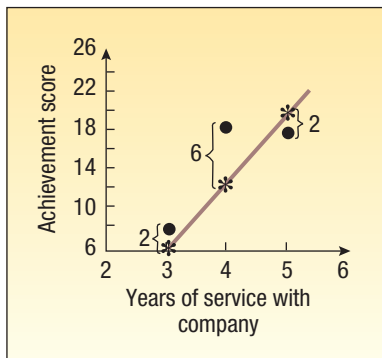
**LEAST SQUARES PRINCIPLE** A mathematical procedure that uses the data to position a line with the objective of minimizing the sum of the squares of the vertical distances between the actual  $Y$  values and the predicted values of  $Y$ .

To illustrate this concept, the same data are plotted in the three charts that follow. The dots are the actual values of  $Y$ , and the asterisks are the predicted values of  $Y$  for a given value of  $X$ . The regression line in Chart 13–9 was determined using the least squares method. It is the best-fitting line because the sum of the squares of the vertical deviations about it is at a minimum. The first plot ( $X = 3, Y = 8$ ) deviates by 2 from the line, found by  $10 - 8$ . The deviation squared is 4. The squared deviation for the plot  $X = 4, Y = 18$  is 16. The squared deviation for the plot  $X = 5, Y = 16$  is 4. The sum of the squared deviations is 24, found by  $4 + 16 + 4$ .

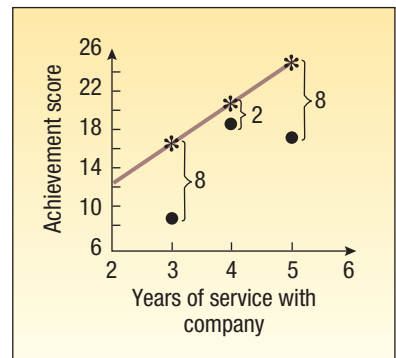
Assume that the lines in Charts 13–10 and 13–11 were drawn with a straight edge. The sum of the squared vertical deviations in Chart 13–10 is 44. For Chart 13–11,



**CHART 13–9** The Least Squares Line



**CHART 13–10** Line Drawn with a Straight Edge



**CHART 13–11** Different Line Drawn with a Straight Edge

it is 132. Both sums are greater than the sum for the line in Chart 13–9, found by using the least squares method.

The equation of a line has the form

$$\text{GENERAL FORM OF LINEAR REGRESSION EQUATION} \quad \hat{Y} = a + bX \quad [13-3]$$

where:

$\hat{Y}$ , read  $Y$  hat, is the estimated value of the  $Y$  variable for a selected  $X$  value.

$a$  is the  $Y$ -intercept. It is the estimated value of  $Y$  when  $X = 0$ . Another way to put it is:  $a$  is the estimated value of  $Y$  where the regression line crosses the  $Y$ -axis when  $X$  is zero.

$b$  is the slope of the line, or the average change in  $\hat{Y}$  for each change of one unit (either increase or decrease) in the independent variable  $X$ .

$X$  is any value of the independent variable that is selected.

The general form of the linear regression equation is exactly the same form as the equation of any line.  $a$  is the  $Y$  intercept and  $b$  is the slope. The purpose of regression analysis is to calculate the values of  $a$  and  $b$  to develop a linear equation that best fits the data.

The formulas for  $a$  and  $b$  are:

$$\text{SLOPE OF THE REGRESSION LINE} \quad b = r \frac{s_y}{s_x} \quad [13-4]$$

where:

$r$  is the correlation coefficient.

$s_y$  is the standard deviation of  $Y$  (the dependent variable).

$s_x$  is the standard deviation of  $X$  (the independent variable).

$$\text{Y-INTERCEPT} \quad a = \bar{Y} - b\bar{X} \quad [13-5]$$

where:

$\bar{Y}$  is the mean of  $Y$  (the dependent variable).

$\bar{X}$  is the mean of  $X$  (the independent variable).

### Example

Recall the example involving Copier Sales of America. The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. As a part of her presentation at the upcoming sales meeting, Ms. Bancor, the sales manager, would like to offer specific information about the relationship between the number of sales calls and the number of copiers sold. Use the least squares method to determine a linear equation to express the relationship between the two variables. What is the expected number of copiers sold by a representative who made 20 calls?

### Solution

The first step in determining the regression equation is to find the slope of the least squares regression line. That is, we need the value of  $b$ . On page 468, we determined the correlation coefficient  $r$  (.759). In the Excel output on the same page, we determined the standard deviation of the independent variable  $X$  (9.189) and the standard deviation of the dependent variable  $Y$  (14.337). The values are inserted in formula (13–4).

$$b = r \left( \frac{s_y}{s_x} \right) = .759 \left( \frac{14.337}{9.189} \right) = 1.1842$$

**LO4** Interpret the regression analysis.

Next we need to find the value of  $a$ . To do this, we use the value for  $b$  that we just calculated as well as the means for the number of sales calls and the number of copiers sold. These means are also available in the Excel printout on page 468. From formula (13-5):

$$a = \bar{Y} - b\bar{X} = 45 - 1.1842(22) = 18.9476$$

Thus, the regression equation is  $\hat{Y} = 18.9476 + 1.1842X$ . So if a salesperson makes 20 calls, he or she can expect to sell 42.6316 copiers, found by  $\hat{Y} = 18.9476 + 1.1842X = 18.9476 + 1.1842(20)$ . The  $b$  value of 1.1842 indicates that for each additional sales call made the sales representative can expect to increase the number of copiers sold by about 1.2. To put it another way, five additional sales calls in a month will result in about six more copiers being sold, found by  $1.1842(5) = 5.921$ .

The  $a$  value of 18.9476 is the point where the equation crosses the  $Y$ -axis. A literal translation is that if no sales calls are made, that is,  $X = 0$ , 18.9476 copiers will be sold. Note that  $X = 0$  is outside the range of values included in the sample and, therefore, should not be used to estimate the number of copiers sold. The sales calls ranged from 10 to 40, so estimates should be limited to that range.



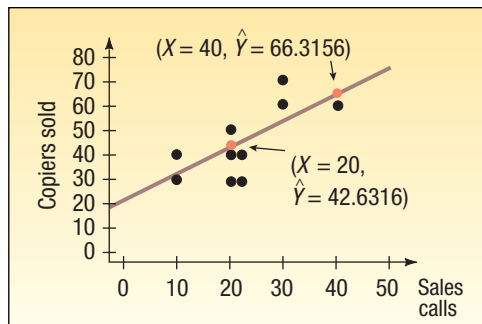
**Statistics in Action**

In finance, investors are interested in the trade-off between returns and risk. One technique to quantify risk is a regression analysis of a company's stock price (dependent variable) and an average measure of the stock market (independent variable). Often the Standard and Poor's (S&P) 500 Index is used to estimate the market. The regression coefficient, called beta in finance, shows the change in a company's stock price for a one-unit change in the S&P Index. For  
*(continued)*

**Drawing the Regression Line**

The least squares equation,  $\hat{Y} = 18.9476 + 1.1842X$ , can be drawn on the scatter diagram. The first sales representative in the sample is Tom Keller. He made 20 calls. His estimated number of copiers sold is  $\hat{Y} = 18.9476 + 1.1842(20) = 42.6316$ . The plot  $X = 20$  and  $\hat{Y} = 42.6316$  is located by moving to 20 on the  $X$ -axis and then going vertically to 42.6316. The other points on the regression equation can be determined by substituting the particular value of  $X$  into the regression equation. All the points are connected to give the line. See Chart 13-12.

Sales Representative	Sales Calls (X)	Estimated Sales (Ŷ)	Sales Representative	Sales Calls (X)	Estimated Sales (Ŷ)
Tom Keller	20	42.6316	Carlos Ramirez	10	30.7896
Jeff Hall	40	66.3156	Rich Niles	20	42.6316
Brian Virost	20	42.6316	Mike Kiel	20	42.6316
Greg Fish	30	54.4736	Mark Reynolds	20	42.6316
Susan Welch	10	30.7896	Soni Jones	30	54.4736



**CHART 13-12** The Line of Regression Drawn on the Scatter Diagram

example, if a stock has a beta of 1.5, then when the S&P index increases by 1%, the stock price will increase by 1.5%. The opposite is also true. If the S&P decreases by 1%, the stock price will decrease by 1.5%. If the beta is 1.0, then a 1% change in the index should show a 1% change in a stock price. If the beta is less than 1.0, then a 1% change in the index shows less than a 1% change in the stock price.

The least squares regression line has some interesting and unique features. First, it will always pass through the point  $(\bar{X}, \bar{Y})$ . To show this is true, we can use the mean number of sales calls to predict the number of copiers sold. In this example, the mean number of sales calls is 22.0, found by  $\bar{X} = 220/10$ . The mean number of copiers sold is 45.0, found by  $\bar{Y} = 450/10 = 45$ . If we let  $X = 22$  and then use the regression equation to find the estimated value for  $\hat{Y}$ , the result is:

$$\hat{Y} = 18.9476 + 1.1842(22) = 45$$

The estimated number of copiers sold is exactly equal to the mean number of copiers sold. This simple example shows the regression line will pass through the point represented by the two means. In this case, the regression equation will pass through the point  $X = 22$  and  $Y = 45$ .

Second, as we discussed earlier in this section, there is no other line through the data where the sum of the squared deviations is smaller. To put it another way, the term  $\sum(Y - \hat{Y})^2$  is smaller for the least squares regression equation than for any other equation. We use the Excel system to demonstrate this condition.

	A	B	C	D	E	F	G	H	I	J	K
1	Sales	Calls	Sales	Estimates Sales							
2	Representative	(X)	(Y)	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$	$Y^*$	$(Y - Y^*)^2$	$Y^{**}$	$(Y - Y^{**})^2$	
3	Tom Keller	20	30	42.6316	-12.6316	159.55731856	43	169	40	100	
4	Jeff Hall	40	60	66.3156	-6.3156	39.88680336	67	49	60	0	
5	Brian Virost	20	40	42.6316	-2.6316	6.92531856	43	9	40	0	
6	Greg Fish	30	60	54.4736	5.5264	30.54109696	55	25	50	100	
7	Susan Welch	10	30	30.7896	-0.7896	0.62346816	31	1	30	0	
8	Carlos Ramirez	10	40	30.7896	9.2104	84.83146816	31	81	30	100	
9	Rich Niles	20	40	42.6316	-2.6316	6.92531856	43	9	40	0	
10	Mike Kiel	20	50	42.6316	7.3684	54.29331856	43	49	40	100	
11	Mark Reynolds	20	30	42.6316	-12.6316	159.55731856	43	169	40	100	
12	Soni Jones	30	70	54.4736	15.5264	241.06909696	55	225	50	400	
13					0.00	784.21052640		786		900	
14											
15											
16											
17											

In Columns A, B, and C in the Excel spreadsheet above, we duplicated the sample information on sales and copiers sold from Table 13–1. In column D, we provide the estimated sales values, the  $\hat{Y}$  values, as calculated above.

In column E, we calculate the **residuals**, or the error values. This is the difference between the actual values and the predicted values. That is, column E is  $(Y - \hat{Y})$ . For Soni Jones,

$$\hat{Y} = 18.9476 + 1.1842(30) = 54.4736$$

Her actual value is 70. So the residual, or error of estimate, is

$$(Y - \hat{Y}) = (70 - 54.4736) = 15.5264$$

This value reflects the amount the predicted value of sales is “off” from the actual sales value.

Next, in Column F, we square the residuals for each of the sales representatives and total the result. The total is 784.2105.

$$\sum(Y - \hat{Y})^2 = 159.5573 + 39.8868 + \dots + 241.0691 = 784.2105$$

This is the sum of the squared differences or the least squares value. There is no other line through these 10 data points where the sum of the squared differences is smaller.

We can demonstrate the least squares criterion by choosing two arbitrary equations that are close to the least squares equation and determining the sum of the

squared differences for these equations. In column G, we use the equation  $Y^* = 19 + 1.2X$  to find the predicted value. Notice this equation is very similar to the least squares equation. In Column H, we determine the residuals and square these residuals. For the first sales representative, Tom Keller,

$$Y^* = 19 + 1.2(20) = 43$$

$$(Y - Y^*)^2 = (43 - 30)^2 = 169$$

This procedure is continued for the other nine sales representatives and the squared residuals totaled. The result is 786. This is a larger value (786 versus 784.2105) than the residuals for the least squares line.

In columns I and J on the output, we repeat the above process for yet another equation  $Y^{**} = 20 + X$ . Again, this equation is similar to the least squares equation. The details for Tom Keller are:

$$Y^{**} = 20 + X = 20 + 20 = 40$$

$$(Y - Y^{**})^2 = (30 - 40)^2 = 100$$

This procedure is continued for the other nine sales representatives and the residuals totaled. The result is 900, which is also larger than the least squares values.

What have we shown with the example? The sum of the squared residuals  $[\sum(Y - \hat{Y})^2]$  for the least squares equation is smaller than for other selected lines. The bottom line is you will not be able to find a line passing through these data points where the sum of the squared residuals is smaller.

**Self-Review 13–3**



Refer to Self-Review 13–1, where the owner of Haverty’s Furniture Company was studying the relationship between sales and the amount spent on advertising. The sales information for the last four months is repeated below.

Month	Advertising Expense (\$ million)	Sales Revenue (\$ million)
July	2	7
August	1	3
September	3	8
October	4	10

- (a) Determine the regression equation.
- (b) Interpret the values of  $a$  and  $b$ .
- (c) Estimate sales when \$3 million is spent on advertising.

**Exercises**



13. The following sample observations were randomly selected.




X:	4	5	3	6	10
Y:	4	6	5	7	7


- a. Determine the regression equation.
  - b. Determine the value of  $\hat{Y}$  when  $X$  is 7.
14. The following sample observations were randomly selected.




X:	5	3	6	3	4	4	6	8
Y:	13	15	7	12	13	11	9	5

- a. Determine the regression equation.  
 b. Determine the value of  $\hat{Y}$  when  $X$  is 7.
15. Bradford Electric Illuminating Company is studying the relationship between kilowatt-hours (thousands) used and the number of rooms in a private single-family residence. A random sample of 10 homes yielded the following. 

Number of Rooms	Kilowatt-Hours (thousands)	Number of Rooms	Kilowatt-Hours (thousands)
12	9	8	6
9	7	10	8
14	10	10	10
6	5	5	4
10	8	7	7

- a. Determine the regression equation.  
 b. Determine the number of kilowatt-hours, in thousands, for a six-room house.
16. Mr. James McWhinney, president of Daniel-James Financial Services, believes there is a relationship between the number of client contacts and the dollar amount of sales. To document this assertion, Mr. McWhinney gathered the following sample information. The  $X$  column indicates the number of client contacts last month, and the  $Y$  column shows the value of sales (\$ thousands) last month for each client sampled. 

Number of Contacts, $X$	Sales (\$ thousands), $Y$	Number of Contacts, $X$	Sales (\$ thousands), $Y$
14	24	23	30
12	14	48	90
20	28	50	85
16	30	55	120
46	80	50	110

- a. Determine the regression equation.  
 b. Determine the estimated sales if 40 contacts are made.
17. A recent article in *BusinessWeek* listed the “Best Small Companies.” We are interested in the current results of the companies’ sales and earnings. A random sample of 12 companies was selected and the sales and earnings, in millions of dollars, are reported below. 

Company	Sales (\$ millions)	Earnings (\$ millions)	Company	Sales (\$ millions)	Earnings (\$ millions)
Papa John’s International	\$89.2	\$4.9	Checkmate Electronics	\$17.5	\$ 2.6
Applied Innovation	18.6	4.4	Royal Grip	11.9	1.7
Integracare	18.2	1.3	M-Wave	19.6	3.5
Wall Data	71.7	8.0	Serving-N-Slide	51.2	8.2
Davidson & Associates	58.6	6.6	Daig	28.6	6.0
Chico’s FAS	46.8	4.1	Cobra Golf	69.2	12.8

Let sales be the independent variable and earnings be the dependent variable.

- a. Draw a scatter diagram.  
 b. Compute the correlation coefficient.  
 c. Determine the regression equation.  
 d. For a small company with \$50.0 million in sales, estimate the earnings.
18. We are studying mutual bond funds for the purpose of investing in several funds. For this particular study, we want to focus on the assets of a fund and its five-year performance. The question is: Can the five-year rate of return be estimated based on the assets of the

fund? Nine mutual funds were selected at random, and their assets and rates of return are shown below.



Fund	Assets (\$ millions)	Return (%)	Fund	Assets (\$ millions)	Return (%)
AARP High Quality Bond	\$622.2	10.8	MFS Bond A	\$494.5	11.6
Babson Bond L	160.4	11.3	Nichols Income	158.3	9.5
Compass Capital Fixed Income	275.7	11.4	T. Rowe Price Short-term	681.0	8.2
Galaxy Bond Retail	433.2	9.1	Thompson Income B	241.3	6.8
Keystone Custodian B-1	437.9	9.2			

- a. Draw a scatter diagram.
  - b. Compute the correlation coefficient.
  - c. Write a brief report of your findings for parts (b) and (c).
  - d. Determine the regression equation. Use assets as the independent variable.
  - e. For a fund with \$400.0 million in sales, determine the five-year rate of return (in percent).
19. Refer to Exercise 5.
- a. Determine the regression equation.
  - b. Estimate the number of crimes for a city with 20 police officers.
  - c. Interpret the regression equation.
20. Refer to Exercise 6.
- a. Determine the regression equation.
  - b. Estimate the selling price of a 10-year-old car.
  - c. Interpret the regression equation.

## 13.6 Testing the Significance of the Slope

**L05** Evaluate the significance of the slope of the regression equation.

In the prior section, we showed how to find the equation of the regression line that best fits the data. The method for finding the equation is based on the *least squares principle*. The purpose of the regression equation is to quantify a linear relationship between two variables.

The next step is to analyze the regression equation by conducting a test of hypothesis to see if the slope of the regression line is different from zero. Why is this important? If we can show that the slope of the line in the population is different from zero, then we can conclude that using the regression equation adds to our ability to predict or forecast the dependent variable based on the independent variable. If we cannot demonstrate that this slope is different from zero, then we conclude there is no merit to using the independent variable as a predictor. To put it another way, if we cannot show the slope of the line is different from zero, we might as well use the mean of the dependent variable as a predictor, rather than use the regression equation.

Following from the hypothesis-testing procedure in Chapter 10, the null and alternative hypotheses are:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

We use  $\beta$  (the Greek letter beta) to represent the population slope for the regression equation. This is consistent with our policy to identify population parameters by Greek letters. We assumed the information regarding Copier Sales of America, Table 13–2, and the Example for the Applewood Auto Group are samples. Be careful here. Remember, this is a single sample, but when we selected a particular salesperson we identified two pieces of information, how many customers they called on and how many copiers they sold. It is still a single sample, however.

We identified the slope value as  $b$ . So our computed slope “ $b$ ” is based on a sample and is an estimate of the population’s slope, identified as “ $\beta$ .” The null hypothesis is that the slope of the regression equation in the population is zero. If this is the case, the regression line is horizontal and there is no relationship between the independent variable,  $X$ , and the dependent variable,  $Y$ . In other words, the value of the dependent variable is the same for any value of the independent variable and does not offer us any help in estimating the value of the dependent variable.

What if the null hypothesis is rejected? If the null hypothesis is rejected and the alternate hypothesis accepted, this indicates that the slope of the regression line for the population is not equal to zero. That is, knowing the value of the independent variable allows us to make a better estimate of the dependent variable. To put it another way, a significant relationship exists between the two variables.

Before we test the hypothesis, we use statistical software to determine the needed regression statistics. We continue to use the Copier Sales of America data from Table 13–2 and use Excel to perform the necessary calculations. The following spreadsheet shows three tables to the right of the sample data.

complete reg analysis for 15e										
	A	B	C	D	E	F	G	H	I	J
1	Sales Representative	Calls	Sales		SUMMARY OUTPUT					
2	Tom Keller	20	30							
3	Jeff Hall	40	60		<i>Regression Statistics</i>					
4	Brian Virost	20	40		Multiple R	0.759				
5	Greg Fish	30	60		R Square	0.576				
6	Susan Welch	10	30		Adjusted R Square	0.523				
7	Carlos Ramirez	10	40		Standard Error	9.901				
8	Rich Niles	20	40		Observations	10				
9	Mike Kiel	20	50							
10	Mark Reynolds	20	30		ANOVA					
11	Soni Jones	30	70			<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12					Regression	1	1065.789	1065.789	10.872	0.011
13					Residual	8	784.211	98.026		
14					Total	9	1850.000			
15										
16						<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
17					Intercept	18.9474	8.4988	2.2294	0.05635	
18					Calls	1.18421	0.35914	3.29734	0.01090	
19										
20										

1. Starting on the top are the *Regression Statistics*. We will use this information later in the chapter, but notice that the “Multiple R” value is familiar. It is .759, which is the correlation coefficient we calculated in Section 13.2 using formula (13–1).
2. Next is an ANOVA table. This is a useful table for summarizing regression information. We will refer to it later in this chapter and use it extensively in the next chapter when we study multiple regression.
3. At the bottom, highlighted in blue, is the information needed to conduct our test of hypothesis regarding the slope of the line. It includes the value of the slope, which is 1.18421, and the intercept, which is 18.9474. (Note that these values for the slope and the intercept are slightly different from those computed on pages 478 and 479. These small differences are due to rounding.) In the column to the right of the regression coefficient is a column labeled “Standard Error.” This is a value similar to the standard error of the mean. Recall that the standard error of the mean reports the variation in the sample means. In a similar fashion, these standard errors report the possible variation in slope and intercept values. The standard error of the slope coefficient is 0.35914.

To test the null hypothesis, we use the  $t$ -distribution with  $(n - 2)$  and the following formula.



$$\text{TEST FOR THE SLOPE} \quad t = \frac{b - 0}{s_b} \quad \text{with } n - 2 \text{ degrees of freedom} \quad [13-6]$$

where:

$b$  is the estimate of the regression line's slope calculated from the sample information.

$s_b$  is the standard error of the slope estimate, also determined from sample information.

Our first step is to set the null and the alternative hypotheses. They are:

$$H_0: \beta \leq 0$$

$$H_1: \beta > 0$$

Notice that we have a one-tailed test. If we do not reject the null hypothesis, we conclude that the slope of the regression line in the population could be zero. This means the independent variable is of no value in improving our estimate of the dependent variable. In our case, this means that knowing the number of sales calls made by a representative does not help us predict the sales.

If we reject the null hypothesis and accept the alternative, we conclude the slope of the line is greater than zero. Hence, the independent variable is an aid in predicting the dependent variable. Thus, if we know the number of sales calls made by a representative, this will help us forecast that representative's sales. We also know, because we have demonstrated that the slope of the line is greater than zero—that is, positive—that more sales calls will result in the sale of more copiers.

The  $t$ -distribution is the test statistic; there are 8 degrees of freedom, found by  $n - 2 = 10 - 2$ . We use the .05 significance level. From Appendix B.2, the critical value is 1.860. Our decision rule is to reject the null hypothesis if the value computed from formula (13-6) is greater than 1.860. We apply formula (13-6) to find  $t$ .

$$t = \frac{b - 0}{s_b} = \frac{1.18421 - 0}{0.35814} = 3.297$$

The computed value of 3.297 exceeds our critical value of 1.860, so we reject the null hypothesis and accept the alternative hypothesis. We conclude that the slope of the line is greater than zero. The independent variable referring to the number of sales calls is useful for obtaining a better estimate of sales.

The table also provides us information on the  $p$ -value of this test. This cell is highlighted in purple. So we could select a significance level, say .05, and compare that value with the  $p$ -value. In this case, the calculated  $p$ -value in the table is .01090, so our decision is to reject the null hypothesis. An important caution is that the  $p$ -values reported in the statistical software are usually for a two-tailed test.

Before moving on, here is an interesting note. Observe that on page 473, when we conducted a test of hypothesis regarding the correlation coefficient for these same data using formula (13-2), we obtained the same value of the  $t$ -statistic,  $t = 3.297$ . Actually, the two-tests are equivalent and will always yield exactly the same values of  $t$  and the same  $p$ -values.

#### Self-Review 13-4



Refer to Self-Review 13-1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable, and advertising expense the independent variable. The regression equation in that study was  $\hat{Y} = 1.5 + 2.2X$  for a sample of five months. Conduct a test of hypothesis to show there is a positive relationship between advertising and sales. From statistical software, the standard error of the regression coefficient is 0.42. Use the .05 significance level.

## Exercises

connect™

21. Refer to Exercise 5. The regression equation is  $\hat{Y} = 29.29 - 0.96X$ , the sample size is 8, and the standard error of the slope is 0.22. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?
22. Refer to Exercise 6. The regression equation is  $\hat{Y} = 11.18 - 0.49X$ , the sample size is 12, and the standard error of the slope is 0.23. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?
23. Refer to Exercise 17. The regression equation is  $\hat{Y} = 1.85 + .08X$ , the sample size is 12, and the standard error of the slope is 0.03. Use the .05 significance level. Can we conclude that the slope of the regression line is *different from zero*?
24. Refer to Exercise 18. The regression equation is  $\hat{Y} = 9.9198 - 0.00039X$ , the sample size is 9, and the standard error of the slope is 0.0032. Use the .05 significance level. Can we conclude that the slope of the regression line is less than zero?

## 13.7 Evaluating a Regression Equation's Ability to Predict

### The Standard Error of Estimate

**L06** Evaluate a regression equation to predict the dependent variable.

The results of the regression analysis for Copier Sales of America show a significant relationship between number of sales calls and the number of sales made. By substituting the names of the variables into the equation, it can be written as:

$$\text{Number of copiers sold} = 18.9476 + 1.1842 (\text{Number of sales calls})$$

The equation can be used to estimate the number of copiers sold for any given "number of sales calls" within the range of the data. For example, if the number of sales calls is 30, then we can predict the number of copiers sold. It is 54.4736, found by  $18.9476 + 1.1842(30)$ . However, the data show two sales representatives with sales of 60 and 70 copiers sold. Is the regression equation a good predictor of "Number of copiers sold"?

Perfect prediction, which is finding the *exact outcome*, in economics and business is practically impossible. For example, the revenue for the year from gasoline sales ( $Y$ ) based on the number of automobile registrations ( $X$ ) as of a certain date could no doubt be approximated fairly closely, but the prediction would not be exact to the nearest dollar, or probably even to the nearest thousand dollars. Even predictions of tensile strength of steel wires based on the outside diameters of the wires are not always exact, because of slight differences in the composition of the steel.

What is needed, then, is a measure that describes how precise the prediction of  $Y$  is based on  $X$  or, conversely, how inaccurate the estimate might be. This measure is called the **standard error of estimate**. The standard error of estimate is symbolized by  $s_{y \cdot x}$ . The subscript,  $y \cdot x$ , is interpreted as the standard error of  $y$  for a given value of  $x$ . It is the same concept as the standard deviation discussed in Chapter 3. The standard deviation measures the dispersion around the mean. The standard error of estimate measures the dispersion about the regression line for a given value of  $X$ .

**STANDARD ERROR OF ESTIMATE** A measure of the dispersion, or scatter, of the observed values around the line of regression for a given value of  $X$ .

The standard error of estimate is found using formula (13–7).

**STANDARD ERROR OF ESTIMATE**

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

**[13–7]**

The calculation of the standard error of estimate requires the sum of the squared differences between each observed value of  $Y$  and the predicted value of  $Y$ , which is identified as  $\hat{Y}$  in the numerator. This calculation is illustrated in the spreadsheet on page 484. See cell G13 in the spreadsheet. It is a very important value. It is the numerator in the calculation of the standard error of the estimate.

$$s_{y \cdot x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{784.211}{10 - 2}} = 9.901$$

This calculation can be eliminated by using statistical software such as Excel. The standard error of the estimate is included in Excel's regression analysis and highlighted in yellow on page 484. Its value is 9.901.

If the standard error of estimate is small, this indicates that the data are relatively close to the regression line and the regression equation can be used to predict  $Y$  with little error. If the standard error of estimate is large, this indicates that the data are widely scattered around the regression line, and the regression equation will not provide a precise estimate of  $Y$ .

## The Coefficient of Determination

Using the standard error of the estimate provides a relative measure of a regression equation's ability to predict. We will use it to provide more specific information about a prediction in the next section. In this section, another statistic is explained that will provide a more interpretable measure of a regression equation's ability to predict. It is called the coefficient of determination, or  $R$ -square.

**COEFFICIENT OF DETERMINATION** The proportion of the total variation in the dependent variable  $Y$  that is explained, or accounted for, by the variation in the independent variable  $X$ .

**L07** Calculate and interpret the coefficient of determination.

The coefficient of determination is easy to compute. It is the correlation coefficient squared. Therefore, the term  $R$ -square is also used. With the Copier Sales of America, the correlation coefficient for the relationship between the number of copiers sold and the number of sales calls is 0.759. If we compute  $(0.759)^2$ , the coefficient of determination is 0.576. See the blue (Multiple R) and green ( $R$ -square) highlighted cells in the spreadsheet on page 484. To better interpret the coefficient of determination, convert it to a percentage. Hence, we say that 57.6 percent of the variation in the number of copiers sold is explained, or accounted for, by the variation in the number of sales calls.

How well can the regression equation predict number of copiers sold with number of sales calls made? If it were possible to make perfect predictions, the coefficient of determination would be 100 percent. That would mean that the independent variable, number of sales calls, explains or accounts for all the variation in the number of copiers sold. A coefficient of determination of 100 percent is associated with a correlation coefficient of +1.0 or -1.0. Refer to Chart 13–2, which shows that a perfect prediction is associated with a perfect linear relationship where all the data points form a perfect line in a scatter diagram. Our analysis shows that only 57.6 percent of the variation in copiers sold is explained by the number of sales

calls. Clearly, this data does not form a perfect line. Instead, the data are scattered around the best-fitting, least squares regression line, and there will be error in the predictions. In the next section, the standard error of the estimate is used to provide more specific information regarding the error associated with using the regression equation to make predictions.

### Self-Review 13–5



Refer to Self-Review 13–1, where the owner of Haverty's Furniture Company studied the relationship between the amount spent on advertising in a month and sales revenue for that month. The amount of sales is the dependent variable, and advertising expense is the independent variable.

- Determine the standard error of estimate.
- Determine the coefficient of determination.
- Interpret the coefficient of determination.

## Exercises



(You may wish to use a software package such as Excel to assist in your calculations.)

- Refer to Exercise 5. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 6. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 15. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.
- Refer to Exercise 16. Determine the standard error of estimate and the coefficient of determination. Interpret the coefficient of determination.

## Relationships among the Correlation Coefficient, the Coefficient of Determination, and the Standard Error of Estimate

In Section 13.7, we described the standard error of estimate. Recall that it measures how close the actual values are to the regression line. When the standard error is small, it indicates that the two variables are closely related. In the calculation of the standard error, the key term is

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

If the value of this term is small, then the standard error will also be small.

The correlation coefficient measures the strength of the linear association between two variables. When the points on the scatter diagram appear close to the line, we note that the correlation coefficient tends to be large. Therefore, the correlation coefficient and the standard error of the estimate are inversely related. As the strength of a linear relationship between two variables increases, the correlation coefficient increases and the standard error of the estimate decreases.

We also noted that the square of the correlation coefficient is the coefficient of determination. The coefficient of determination measures the percentage of the variation in  $Y$  that is explained by the variation in  $X$ .

A convenient vehicle for showing the relationship among these three measures is an ANOVA table. See the yellow highlighted portion of the spreadsheet on page 489. This table is similar to the analysis of variance table developed in Chapter 12. In that chapter, the total variation was divided into two components: variation due to the *treatments* and that due to *random error*. The concept is similar in regression analysis. The total variation is divided into two components: (1) variation explained

by the *regression* (explained by the independent variable) and (2) the *error*, or *residual*. This is the unexplained variation. These three categories are identified in the first column of the spreadsheet ANOVA table. The column headed “*df*” refers to the degrees of freedom associated with each category. The total number of degrees of freedom is  $n - 1$ . The number of degrees of freedom in the regression is 1, because there is only one independent variable. The number of degrees of freedom associated with the error term is  $n - 2$ . The term “*SS*” located in the middle of the ANOVA table refers to the sum of squares. You should note that the total degrees of freedom is equal to the sum of the regression and residual (error) degrees of freedom, and the total sum of squares is equal to the sum of the regression and residual (error) sum of squares. This is true for any ANOVA table.

complete reg analysis for 15e										
	A	B	C	D	E	F	G	H	I	J
1	Sales Representative	Calls	Sales		SUMMARY OUTPUT					
2	Tom Keller	20	30							
3	Jeff Hall	40	60		Regression Statistics					
4	Brian Virost	20	40		Multiple R	0.759				
5	Greg Fish	30	60		R Square	0.576				
6	Susan Welch	10	30		Adjusted R Square	0.523				
7	Carlos Ramirez	10	40		Standard Error	9.901				
8	Rich Niles	20	40		Observations	10				
9	Mike Kiel	20	50							
10	Mark Reynolds	20	30		ANOVA					
11	Soni Jones	30	70		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12					Regression	1	1065.789	1065.789	10.872	0.011
13					Residual	8	784.211	98.026		
14					Total	9	1850.000			
15										
16					<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>		
17					Intercept	18.9474	8.4988	2.2294	0.05635	
18					Calls	1.18421	0.35914	3.29734	0.01090	

The ANOVA sum of squares are computed as follows:

$$\begin{aligned} \text{Regression Sum of Squares} &= \text{SSR} = \sum(\hat{Y} - \bar{Y})^2 = 1065.789 \\ \text{Residual or Error Sum of Squares} &= \text{SSE} = \sum(Y - \hat{Y})^2 = 784.211 \\ \text{Total Sum of Squares} &= \text{SS Total} = \sum(Y - \bar{Y})^2 = 1850.00 \end{aligned}$$

Recall that the coefficient of determination is defined as the percentage of the total variation (SS Total) explained by the regression equation (SSR). Using the ANOVA table, the reported value of *R*-square can be validated.

**COEFFICIENT OF DETERMINATION**

$$r^2 = \frac{\text{SSR}}{\text{SS Total}} = 1 - \frac{\text{SSE}}{\text{SS Total}}$$

**[13–8]**

Using the values from the ANOVA table, the coefficient of determination is  $1065.789/1850.00 = 0.576$ . Therefore, the more variation of the dependent variable (SS Total) explained by the independent variable (SSR), the higher the coefficient of determination.

We can also express the coefficient of determination in terms of the error or residual variation:

$$r^2 = 1 - \frac{\text{SSE}}{\text{SS Total}} = 1 - \frac{784.211}{1850.00} = 1 - 0.424 = 0.576$$

In this case, the coefficient of determination and the residual or error sum of squares are inversely related. The higher the unexplained or error variation as a percentage of the total variation, the lower is the coefficient of determination. In this case, 42.4 percent of the total variation in the dependent variable is error or residual variation.