CHAPTER

16

Introduction to Regression

Tools You Will Need

The following items are considered essential background material for this chapter. If you doubt your knowledge of any of these items, you should review the appropriate chapter or section before proceeding.

- Sum of squares (*SS*) (Chapter 4)
  - Computational formula
  - Definitional formula
- *z*-scores (Chapter 5)
- Analysis of variance (Chapter 12)
  - *MS* values and *F*-ratios
- Pearson correlation (Chapter 15)
  - Sum of products (*SP*)

Preview

In Chapter 15, we noted that one common application of correlations is for purposes of prediction. Whenever there is a consistent relationship between two variables, it is possible to use the value of one variable to predict the value of another. Managers at the electric company, for example, can use the weather forecast to predict power demands for upcoming days. If exceptionally hot summer weather is forecast, they can anticipate an exceptionally high demand for electricity. In the field of psychology, a known relationship between certain personality characteristics and eating disorders can allow clinicians to predict that individuals who show specific characteristics are more likely to develop disorders. A common prediction that is especially relevant for college students (and potential college students) is based on the relationship between scores on aptitude tests (such as the SAT) and future grade point averages in college. Each year, SAT scores from thousands of high school students are used to help college admissions officers decide who should be admitted and who should not.

The Problem: The correlations introduced in Chapter 15 allow researchers to measure and describe relationships, and the hypothesis tests allow researchers to evaluate the significance of correlations. However, we now want to go one step further and actually use a correlation to make predictions.

The Solution: In this chapter we introduce some of the statistical techniques that are used to make predictions based on correlations. Whenever there is a linear relationship (Pearson correlation) between two variables, it is possible to compute an equation that provides a precise, mathematical description of the relationship. With the equation, it is possible to plug in the known value for one variable (for example, your SAT score), and then calculate a predicted value for the second variable (for example, your college grade point average). The general statistical process of finding and using a prediction equation is known as *regression*.

Beyond finding a prediction equation, however, it is reasonable to ask how good its predictions are. For example, I can make predictions about the outcome of a coin toss by simply guessing. However, my predictions are correct only about 50% of the time. In statistical terms, my predictions are not significantly better than chance. In the same way, it is appropriate to challenge the significance of any prediction equation. In this chapter we introduce the techniques that are used to find prediction equations, as well as the techniques that are used to determine whether their predictions are statistically significant. Incidentally, although there is some controversy about the practice of using SAT scores to predict college performance, there is a great deal of research showing that SAT scores really are valid and significant predictors (Camera & Echternacht, 2000; Geiser & Studley, 2002).

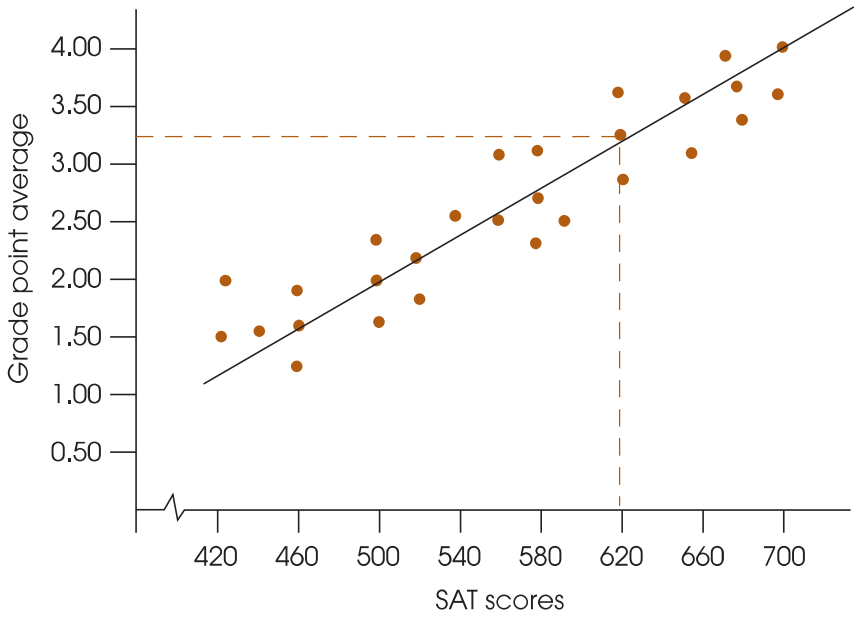
16.1 INTRODUCTION TO LINEAR EQUATIONS AND REGRESSION

In the previous chapter, we introduced the Pearson correlation as a technique for describing and measuring the linear relationship between two variables. Figure 16.1 presents hypothetical data showing the relationship between SAT scores and college grade point average (GPA). Note that the figure shows a good, but not perfect, positive relationship. Also note that we have drawn a line through the middle of the data points. This line serves several purposes:

1. The line makes the relationship between SAT scores and GPA easier to see.
2. The line identifies the center, or *central tendency,* of the relationship, just as the mean describes central tendency for a set of scores. Thus, the line provides a simplified description of the relationship. For example, if the data points were removed, the straight line would still give a general picture of the relationship between SAT scores and GPA.
3. Finally, the line can be used for prediction. The line establishes a precise, one-to-one relationship between each X value (SAT score) and a corresponding Y value (GPA). For example, an SAT score of 620 corresponds to a GPA of 3.25 (see Figure 16.1). Thus, the college admissions officers could use the straight-line

FIGURE 16.1

Hypothetical data showing the relationship between SAT scores and GPA with a regression line drawn through the data points. The regression line defines a precise, one-to-one relationship between each *X* value (SAT score) and its corresponding *Y* value (GPA).



relationship to predict that a student entering college with an SAT score of 620 should achieve a college GPA of approximately 3.25.

Our goal in this section is to develop a procedure that identifies and defines the straight line that provides the best fit for any specific set of data. This straight line does not have to be drawn on a graph; it can be presented in a simple equation. Thus, our goal is to find the equation for the line that best describes the relationship for a set of *X* and *Y* data.

LINEAR EQUATIONS

In general, a *linear relationship* between two variables *X* and *Y* can be expressed by the equation

$$Y = bX + a \tag{16.1}$$

where *a* and *b* are fixed constants.

For example, a local video store charges a membership fee of $5 per month, which allows you to rent videos and games for $2 each. With this information, the total cost for 1 month can be computed using a *linear equation* that describes the relationship between the total cost (*Y*) and the number of videos and games rented (*X*).

$$Y = 2X + 5$$

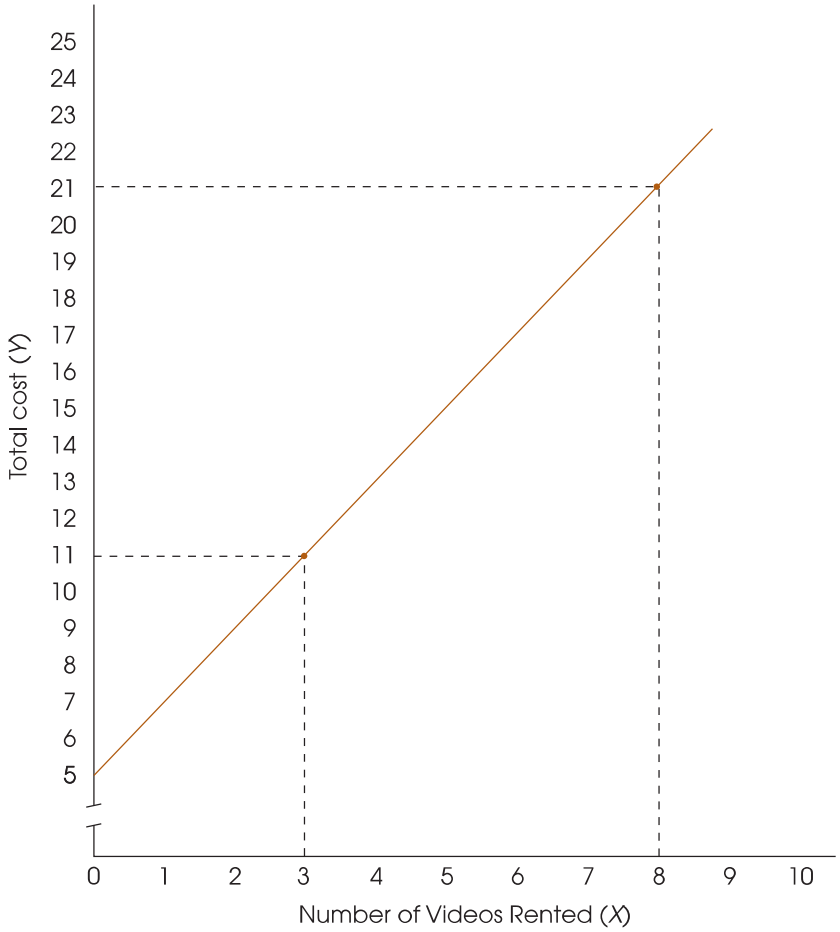
Note that a positive slope means that *Y* increases when *X* is increased, and a negative slope indicates that *Y* decreases when *X* is increased.

In the general linear equation, the value of *b* is called the *slope*. The slope determines how much the *Y* variable changes when *X* is increased by 1 point. For the video store example, the slope is $b = 2$ and indicates that your total cost increases by $2 for each video you rent. The value of *a* in the general equation is called the Y-*intercept* because it determines the value of *Y* when $X = 0$. (On a graph, the *a* value identifies the point where the line intercepts the *Y*-axis.) For the video store example, $a = 5$; there is a $5 membership charge even if you never rent a video.

Figure 16.2 shows the general relationship between the monthly cost and number of videos for the video store example. Notice that the relationship results in a straight

FIGURE 16.2

The relationship between total cost and number of videos rented each month. The video store charges a $5 monthly membership fee and $2 for each video or game rented. The relationship is described by a linear equation $Y = 2X + 5$ where Y is the total cost and X is the number of videos.



line. To obtain this graph, we picked any two values of X and then used the equation to compute the corresponding values for Y. For example,

when $X = 3$:	when $X = 8$:
$Y = bX + a$	$Y = bX + a$
$= \$2(3) + \$5$	$= \$2(8) + \$5$
$= \$6 + \$5$	$= \$16 + \$5$
$= \$11$	$= \$21$

When drawing a graph of a linear equation, it is wise to compute and plot at least three points to be certain that you have not made a mistake.

Next, these two points are plotted on the graph: one point at $X = 3$ and $Y = 11$, the other point at $X = 8$ and $Y = 21$. Because two points completely determine a straight line, we simply drew the line so that it passed through these two points.

LEARNING CHECK

1. A local gym charges a $25 monthly membership fee plus $2 per hour for aerobics classes. What is the linear equation that describes the relationship between the total monthly cost (Y) and the number of class hours each month (X)?
2. For the following linear equation, what happens to the value of Y each time X is increased by 1 point?

$$Y = -3X + 7$$

3. Use the linear equation $Y = 2X - 7$ to determine the value of Y for each of the following values of X: 1, 3, 5, 10.

4. If the slope constant (b) in a linear equation is positive, then a graph of the equation is a line tilted from lower left to upper right. (True or false?)

ANSWERS

1. $Y = 2X + 25$

2. The slope is -3, so Y decreases by 3 points each time X increases by 1 point.

3.

X	Y
1	-5
3	-1
5	3
10	13

4. True. A positive slope indicates that Y increases (goes up in the graph) when X increases (goes to the right in the graph).

REGRESSION

Because a straight line can be extremely useful for describing a relationship between two variables, a statistical technique has been developed that provides a standardized method for determining the best-fitting straight line for any set of data. The statistical procedure is *regression,* and the resulting straight line is called the *regression line.*

DEFINITION

The statistical technique for finding the best-fitting straight line for a set of data is called **regression,** and the resulting straight line is called the **regression line.**

The goal for regression is to find the best-fitting straight line for a set of data. To accomplish this goal, however, it is first necessary to define precisely what is meant by "best fit." For any particular set of data, it is possible to draw lots of different straight lines that all appear to pass through the center of the data points. Each of these lines can be defined by a linear equation of the form $Y = bX + a$ where b and a are constants that determine the slope and Y-intercept of the line, respectively. Each individual line has its own unique values for b and a. The problem is to find the specific line that provides the best fit to the actual data points.

THE LEAST-SQUARES SOLUTION

To determine how well a line fits the data points, the first step is to define mathematically the distance between the line and each data point. For every X value in the data, the linear equation determines a Y value on the line. This value is the predicted Y and is called $\hat{Y}$ ("Y hat"). The distance between this predicted value and the actual Y value in the data is determined by

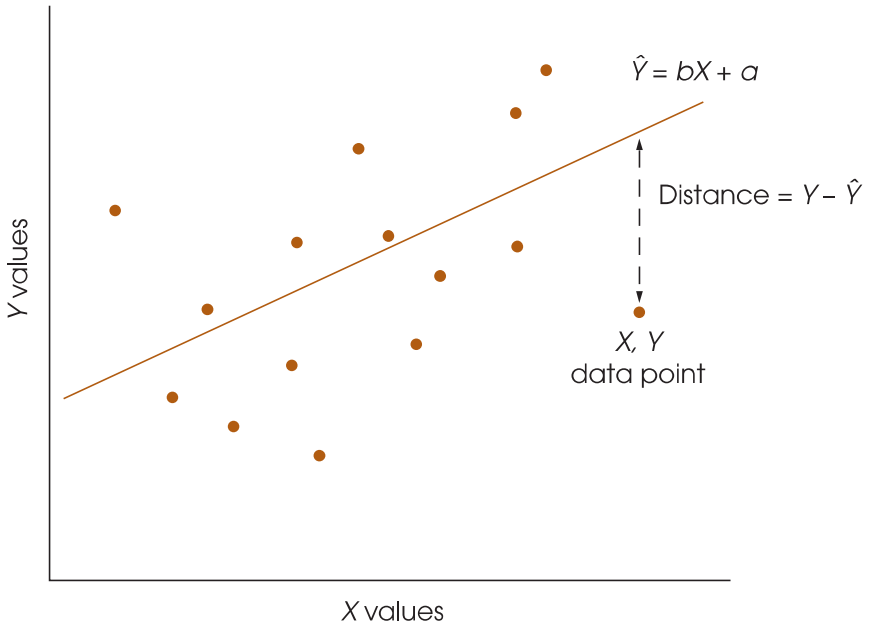
$$\text{distance} = Y - \hat{Y}$$

Note that we simply are measuring the vertical distance between the actual data point (Y) and the predicted point on the line. This distance measures the error between the line and the actual data (Figure 16.3).

Because some of these distances are positive and some are negative, the next step is to square each distance to obtain a uniformly positive measure of error. Finally, to

FIGURE 16.3

The distance between the actual data point (Y) and the predicted point on the line ($\hat{Y}$) is defined as $Y - \hat{Y}$. The goal of regression is to find the equation for the line that minimizes these distances.



determine the total error between the line and the data, we add the squared errors for all of the data points. The result is a measure of overall squared error between the line and the data:

$$\text{total squared error} = \Sigma(Y - \hat{Y})^2$$

Now we can define the *best-fitting* line as the one that has the smallest total squared error. For obvious reasons, the resulting line is commonly called the *least-squared-error solution*. In symbols, we are looking for a linear equation of the form

$$\hat{Y} = bX + a$$

For each value of X in the data, this equation determines the point on the line ($\hat{Y}$) *that gives the best prediction of* Y. The problem is to find the specific values for a and b that make this the best-fitting line.

The calculations that are needed to find this equation require calculus and some sophisticated algebra, so we do not present the details of the solution. The results, however, are relatively straightforward, and the solutions for b and a are as follows:

$$b = \frac{SP}{SS_X} \tag{16.2}$$

where SP is the sum of products and SS_X is the sum of squares for the X scores.

A commonly used alternative formula for the slope is based on the standard deviations for X and Y. The alternative formula is

$$b = r\frac{s_Y}{s_X} \tag{16.3}$$

where s_Y is the standard deviation for the Y scores, s_X is the standard deviation for the X scores, and r is the Pearson correlation for X and Y. The value of the constant a in the equation is determined by

$$a = M_Y - bM_X \tag{16.4}$$

Note that these formulas determine the linear equation that provides the best prediction of Y values. This equation is called the *regression equation for* Y.

DEFINITION

The **regression equation for *Y*** is the linear equation

$$\hat{Y} = bX + a \tag{16.5}$$

where the constant b is determined by Equation 16.2, or 16.3 and the constant a is determined by Equation 16.4. This equation results in the least squared error between the data points and the line.

EXAMPLE 16.1

The scores in the following table are used to demonstrate the calculation and use of the regression equation for predicting Y.

X	Y	$X - M_X$	$Y - M_Y$	$(X - M_X)^2$	$(Y - M_X)^2$	$(X - M_x)(Y - M_Y)$
2	3	−2	−5	4	25	10
6	11	2	3	4	9	6
0	6	−4	−2	16	4	8
4	6	0	−2	0	4	0
7	12	3	4	9	16	12
5	7	1	−1	1	1	−1
5	10	1	2	1	4	2
3	9	−1	1	1	1	−1
				$SS_X = 36$	$SS_Y = 64$	$SP = 36$

For these data, $\Sigma X = 32$, so $M_X = 4$. Also, $\Sigma Y = 64$, so $M_Y = 8$. These values have been used to compute the deviation scores for each X and Y value. The final three columns show the squared deviations for X and for Y, and the products of the deviation scores.

Our goal is to find the values for b and a in the regression equation. Using Equations 16.2 and 16.4, the solutions for b and a are

$$b = \frac{SP}{SS_X} = \frac{36}{36} = 1.00$$

$$a = M_Y - bM_X = 8 - 1(4) = 4.00$$

The resulting equation is

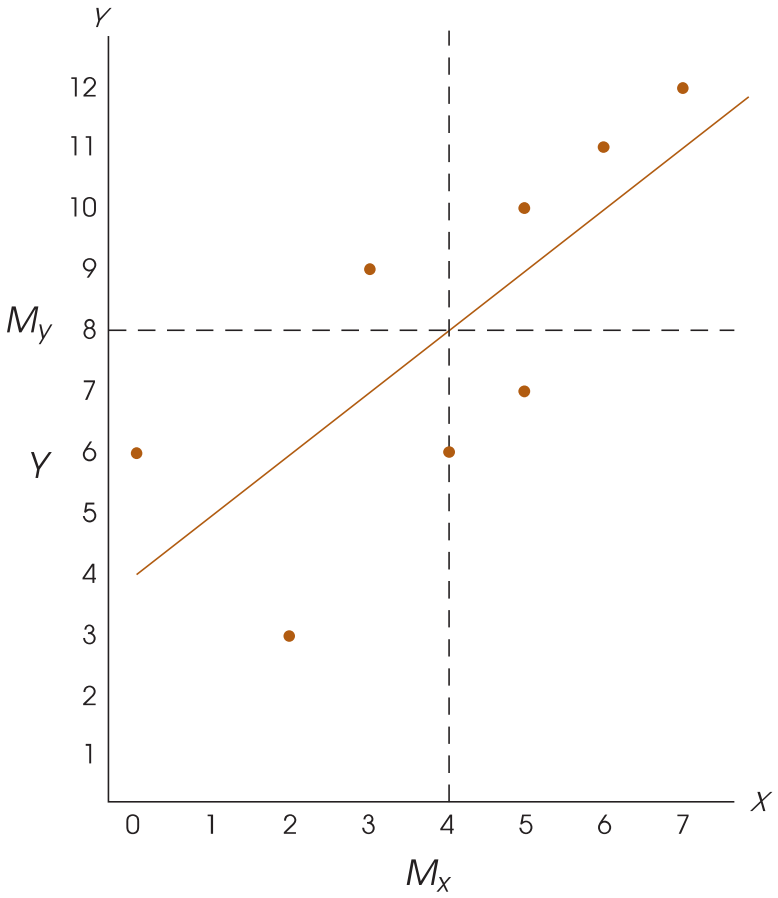
$$\hat{Y} = X + 4$$

The original data and the regression line are shown in Figure 16.4.

The regression line shown in Figure 16.4 demonstrates some simple and very predictable facts about regression. First, the calculation of the Y-intercept (Equation 16.4) ensures that the regression line passes through the point defined by the mean for X and the mean for Y. That is, the point identified by the coordinates M_X, M_Y will always be on the line. We have included the two means in Figure 16.4 to show that

FIGURE 16.4

The X and Y data points and the regression line for the $n = 8$ pairs of scores in Example 16.1.



the point they define is on the regression line. Second, the sign of the correlation (+ or −) is the same as the sign of the slope of the regression line. Specifically, if the correlation is positive, then the slope is also positive and the regression line slopes up to the right. On the other hand, if the correlation is negative, then the slope is negative and the line slopes down to the right. A correlation of zero means that the slope is also zero and the regression equation produces a horizontal line that passes through the data at a level equal to the mean for the Y values. Note that the regression line in Figure 16.4 has a positive slope. One consequence of this fact is that all of the points on the line that are above the mean for X are also above the mean for Y. Similarly, all of the points below the mean for X are also below the mean for Y. Thus, every individual with a positive deviation for X is predicted to have a positive deviation for Y, and everyone with a negative deviation for X is predicted to have a negative deviation for Y.

USING THE REGRESSION EQUATION FOR PREDICTION

As we noted at the beginning of this section, one common use of regression equations is for prediction. For any given value of X, we can use the equation to compute a predicted value for Y. For the equation from Example 16.1, an individual with a score of $X = 1$ would be predicted to have a Y score of

$$\hat{Y} = X + 4 = 1 + 4 = 5$$

Although regression equations can be used for prediction, a few cautions should be considered whenever you are interpreting the predicted values:

1. The predicted value is not perfect (unless $r = +1.00$ or -1.00). If you examine Figure 16.4, it should be clear that the data points do not fit perfectly on the line. In general, there is some error between the predicted Y values (on the line) and the actual data. Although the amount of error varies from point to point, on average the errors are directly related to the magnitude of the correlation. With a correlation near 1.00 (or -1.00), the data points generally are clustered close to the line and the error is small. As the correlation gets nearer to zero, the points move away from the line and the magnitude of the error increases.
2. The regression equation should not be used to make predictions for X values that fall outside of the range of values covered by the original data. For Example 16.1, the X values ranged from $X = 0$ to $X = 7$, and the regression equation was calculated as the best-fitting line within this range. Because you have no information about the X-Y relationship outside this range, the equation should not be used to predict Y for any X value lower than 0 or greater than 7.

STANDARDIZED FORM OF THE REGRESSION EQUATIONS

So far we have presented the regression equation in terms of the original values, or raw scores, for X and Y. Occasionally, however, researchers standardize the scores by transforming the X and Y values into z-scores before finding the regression equation. The resulting equation is often called the standardized form of the regression equation and is greatly simplified compared to the raw-score version. The simplification comes from the fact that z-scores have standardized characteristics. Specifically, the mean for a set of z-scores is always zero and the standard deviation is always 1. As a result, the standardized form of the regression equation becomes

$$\hat{z}_Y = (\text{beta})z_X \tag{16.6}$$

First notice that we are now using the z-score for each X value (z_X) to predict the z-score for the corresponding Y value (z_Y). Also, note that the slope constant that was identified as b in the raw-score formula is now identified as beta. Because both sets of z-scores have a mean of zero, the constant a disappears from the regression equation. Finally, when one variable, X, is being used to predict a second variable, Y, the value of beta is equal to the Pearson correlation for X and Y. Thus, the standardized form of the regression equation can also be written as

$$\hat{z}_Y = rz_X \tag{16.7}$$

Because the process of transforming all of the original scores into z-scores can be tedious, researchers usually compute the raw-score version of the regression equation (Equation 16.5) instead of the standardized form. However, most computer programs report the value of beta as part of the output from linear regression, and you should understand what this value represents.

LEARNING CHECK

1. Sketch a scatter plot for the following data—that is, a graph showing the X, Y data points:

X	Y
1	4
3	9
5	8

a. Find the regression equation for predicting Y from X. Draw this line on your graph. Does it look like the best-fitting line?

b. Use the regression equation to find the predicted Y value corresponding to each X in the data.

ANSWERS

1. a. $SS_X = 8$, $SP = 8$, $b = 1$, $a = 4$. The equation is $\hat{Y} = X + 4$.

b. The predicted Y values are 5, 7, and 9.

THE STANDARD ERROR OF ESTIMATE

It is possible to determine a regression equation for any set of data by simply using the formulas already presented. The linear equation you obtain is then used to generate predicted Y values for any known value of X. However, it should be clear that the accuracy of this prediction depends on how well the points on the line correspond to the actual data points—that is, the amount of error between the predicted values, $\hat{Y}$, and the actual scores, Y values. Figure 16.5 shows two different sets of data that have exactly the same regression equation. In one case, there is a perfect correlation ($r = +1$) between X and Y, so the linear equation fits the data perfectly. For the second set of data, the predicted Y values on the line only approximate the real data points.

A regression equation, by itself, allows you to make predictions, but it does not provide any information about the accuracy of the predictions. To measure the precision of the regression, it is customary to compute a *standard error of estimate.*

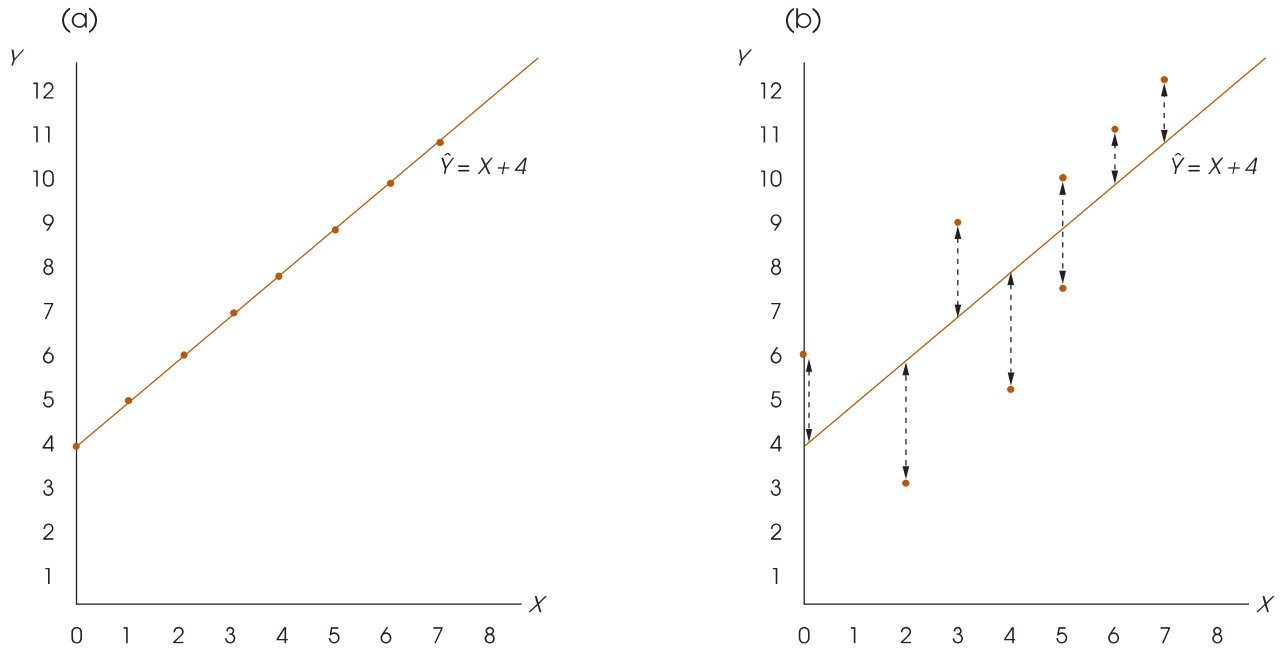


FIGURE 16.5

(a) A scatter plot showing data points that perfectly fit the regression line defined by the equation $\hat{Y} = X + 4$. Note that the correlation is $r = +1.00$. (b) A scatter plot for the data from Example 16.1. Notice that there is error between the actual data points and the predicted Y values of the regression line.

DEFINITION

The **standard error of estimate** gives a measure of the standard distance between the predicted Y values on the regression line and the actual Y values in the data.

Conceptually, the standard error of estimate is very much like a standard deviation: Both provide a measure of standard distance. Also, the calculation of the standard error of estimate is very similar to the calculation of standard deviation.

To calculate the standard error of estimate, we first find the sum of squared deviations (SS). Each deviation measures the distance between the actual Y value (from the data) and the predicted Y value (from the regression line). This sum of squares is commonly called SS_{residual} because it is based on the remaining distance between the actual Y scores and the predicted values.

$$SS_{\text{residual}} = \Sigma(Y - \hat{Y})^2 \tag{16.8}$$

The obtained SS value is then divided by its degrees of freedom to obtain a measure of variance. This procedure should be very familiar:

$$\text{Variance} = \frac{SS}{df}$$

The degrees of freedom for the standard error of estimate are $df = n - 2$. The reason for having $n - 2$ degrees of freedom, rather than the customary $n - 1$, is that we now are measuring deviations from a line rather than deviations from a mean. To find the equation for the regression line, you must know the means for both the X and the Y scores. Specifying these two means places two restrictions on the variability of the data, with the result that the scores have only $n - 2$ degrees of freedom. (*Note*: the $df = n - 2$ for SS_{residual} is the same $df = n - 2$ that we encountered when testing the significance of the Pearson correlation on page 529.)

Recall that variance measures the average squared distance.

The final step in the calculation of the standard error of estimate is to take the square root of the variance to obtain a measure of standard distance. The final equation is

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{df}} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} \tag{16.9}$$

The following example demonstrates the calculation of this standard error.

EXAMPLE 16.2

The same data that were used in Example 16.1 are used here to demonstrate the calculation of the standard error of estimate. These data have the regression equation

$$\hat{Y} = X + 4$$

Using this regression equation, we have computed the predicted Y value, the residual, and the squared residual for each individual, using the data from Example 16.1.

Data		Predicted Y Values	Residual	Squared Residual
X	Y	$\hat{Y} = X + 4$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
2	3	6	−3	9
6	11	10	1	1
0	6	4	2	4
4	6	8	−2	4
5	7	9	−2	4

(continued)

Data		Predicted Y Values	Residual	Squared Residual
X	Y	$\hat{Y} = X + 4$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
7	12	11	1	1
5	10	9	1	1
3	9	7	2	4
			0	$SS_{\text{residual}} = 28$

First note that the sum of the residuals is equal to zero. In other words, the sum of the distances above the line is equal to the sum of the distances below the line. This is true for any set of data and provides a way to check the accuracy of your calculations. The squared residuals are listed in the final column. For these data, the sum of the squared residuals is $SS_{\text{residual}} = 28$. With $n = 8$, the data have $df = n - 2 = 6$, so the standard error of estimate is

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{df}} = \sqrt{\frac{28}{6}} = 2.16$$

Remember: The standard error of estimate provides a measure of how accurately the regression equation predicts the Y values. In this case, the standard distance between the actual data points and the regression line is measured by standard error of estimate = 2.16.

RELATIONSHIP BETWEEN THE STANDARD ERROR AND THE CORRELATION

It should be clear from Example 16.2 that the standard error of estimate is directly related to the magnitude of the correlation between X and Y. If the correlation is near 1.00 (or −1.00), then the data points are clustered close to the line, and the standard error of estimate is small. As the correlation gets nearer to zero, the data points become more widely scattered, the line provides less accurate predictions, and the standard error of estimate grows larger.

Earlier (p. 524), we observed that squaring the correlation provides a measure of the accuracy of prediction. The squared correlation, r^2, is called the coefficient of determination because it determines what proportion of the variability in Y is predicted by the relationship with X. Because r^2 measures the predicted portion of the variability in the Y scores, we can use the expression $(1 - r^2)$ to measure the unpredicted portion. Thus,

$$\text{predicted variability} = SS_{\text{regression}} = r^2 SS_Y \tag{16.10}$$

$$\text{unpredicted variability} = SS_{\text{residual}} = (1 - r^2)SS_Y \tag{16.11}$$

For example, if $r = 0.80$, then the *predicted variability* is $r^2 = 0.64$ (or 64%) of the total variability for the Y scores and the remaining 36% $(1 - r^2)$ is the *unpredicted variability*. Note that when $r = 1.00$, the prediction is perfect and there are no residuals. As the correlation approaches zero, the data points move farther off the line and the residuals grow larger. Using Equation 16.11 to compute SS_{residual}, the standard error of estimate can be computed as

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{df}} = \sqrt{\frac{(1 - r^2)SS_Y}{n - 2}} \tag{16.12}$$

Because it is usually much easier to compute the Pearson correlation than to compute the individual $(Y - \hat{Y})^2$ values, Equation 16.11 is usually the easiest way to compute $SS_{residual}$, and Equation 16.12 is usually the easiest way to compute the standard error of estimate for a regression equation. The following example demonstrates this new formula.

EXAMPLE 16.3

We use the same data used in Examples 16.1 and 16.2, which produced $SS_X = 36$, $SS_Y = 64$, and $SP = 36$. For these data, the Pearson correlation is

$$r = \frac{36}{\sqrt{36(64)}} = \frac{36}{48} = 0.75$$

With $SS_Y = 64$ and a correlation of $r = 0.75$, the predicted variability from the regression equation is

$$SS_{regression} = r^2 SS_Y = (0.75^2)(64) = 0.5625(64) = 36.00$$

Similarly, the unpredicted variability is

$$SS_{residual} = (1 - r^2)SS_Y = (1 - 0.75^2)(64) = 0.4375(64) = 28.00$$

Notice that the new formula for $SS_{residual}$ produces exactly the same value that we obtained by adding the squared residuals in Example 16.2. Also note that this new formula is generally much easier to use because it requires only the correlation value (r) and the SS for Y. The primary point of this example, however, is that $SS_{residual}$ and the standard error of estimate are closely related to the value of the correlation. With a large correlation (near +1.00 or −1.00), the data points are close to the regression line, and the standard error of estimate is small. As a correlation gets smaller (near zero), the data points move away from the regression line, and the standard error of estimate gets larger.

Because it is possible to have the same regression equation for several different sets of data, it is also important to consider r^2 and the standard error of estimate. The regression equation simply describes the best-fitting line and is used for making predictions. However, r^2 and the standard error of estimate indicate how accurate these predictions are.

LEARNING CHECK

1. Describe what is measured by the standard error of estimate for a regression equation.
2. As the numerical value of a correlation increases, what happens to the standard error of estimate?
3. A sample of $n = 6$ pairs of X and Y scores produces a correlation of $r = 0.80$ and $SS_Y = 100$. What is the standard error of estimate for the regression equation?

ANSWERS

1. The standard error of estimate measures the average, or standard, distance between the predicted Y values on the regression line and the actual Y values in the data.
2. A larger correlation means that the data points are clustered closer to the line, which means the standard error of estimate is smaller.
3. The standard error of estimate $= \sqrt{36/4} = 3$.

16.2 ANALYSIS OF REGRESSION: TESTING THE SIGNIFICANCE OF THE REGRESSION EQUATION

As we noted in Chapter 15, a sample correlation is expected to be representative of its population correlation. For example, if the population correlation is zero, then the sample correlation is expected to be near zero. Note that we do not expect the sample correlation to be exactly equal to zero. This is the general concept of *sampling error* that was introduced in Chapter 1 (p. 8). The principle of sampling error is that there is typically some discrepancy or error between the value obtained for a sample statistic and the corresponding population parameter. Thus, when there is no relationship whatsoever in the population, a correlation of $\rho = 0$, you are still likely to obtain a nonzero value for the sample correlation. In this situation, however, the sample correlation is caused by chance and a hypothesis test usually demonstrates that the correlation is not significant.

Whenever you obtain a nonzero value for a sample correlation, you also obtain real, numerical values for the regression equation. However, if there is no real relationship in the population, both the sample correlation and the regression equation are meaningless—they are simply the result of sampling error and should not be viewed as an indication of any relationship between *X* and *Y*. In the same way that we tested the significance of a Pearson correlation, we can test the significance of the regression equation. In fact, when a single variable, *X*, is being used to predict a single variable, *Y*, the two tests are equivalent. In each case, the purpose for the test is to determine whether the sample correlation represents a real relationship or is simply the result of sampling error. For both tests, the null hypothesis states that there is no relationship between the two variables in the population. A more specific null hypothesis for testing the significance of a regression equation is that the equation does not account for a significant proportion of the variance in the *Y* scores. An alternative version of H_0 states that the values of *b* or beta that are computed for the regression equation do not represent any real relationship between *X* and *Y* but rather are simply the result of chance or sampling error. In other words, the true population value of *b* or beta is zero.

The process of testing the significance of a regression equation is called *analysis of regression* and is very similar to the analysis of variance (ANOVA) presented in Chapter 12. As with ANOVA, the regression analysis uses an *F*-ratio to determine whether the variance predicted by the regression equation is significantly greater than would be expected if there were no relationship between *X* and *Y*. The *F*-ratio is a ratio of two variances, or mean square (*MS*) values, and each variance is obtained by dividing an *SS* value by its corresponding degrees of freedom. The numerator of the *F*-ratio is $MS_{\text{regression}}$, which is the variance in the *Y* scores that is predicted by the regression equation. This variance measures the systematic changes in *Y* that occur when the value of *X* increases or decreases. The denominator is MS_{residual}, which is the unpredicted variance in the *Y* scores. This variance measures the changes in *Y* that are independent of changes in *X*. The two *MS* value are defined as

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{regression}}} \text{ with } df = 1 \quad \text{and} \quad MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} \text{ with } df = n - 2$$

The *F*-ratio is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} \quad \text{with } df = 1, n - 2 \tag{16.13}$$

The complete analysis of *SS* and degrees of freedom is diagrammed in Figure 16.6. The analysis of regression procedure is demonstrated in the following example, using the same data that we used in Examples 16.1, 16.2, and 16.3.

EXAMPLE 16.4

The data consist of $n = 8$ pairs of scores with a correlation of $r = 0.75$ and $SS_Y = 64$. The null hypothesis either states that there is no relationship between *X* and *Y* in the population, or that the regression equation does not account for a significant portion of the variance for the *Y* scores.

The *F*-ratio for the analysis of regression has $df = 2, n - 2$. For these data, $df = 1, 6$. With $\alpha = .05$, the critical value is 5.99.

As noted in the previous section, the *SS* for the *Y* scores can be separated into two components: the predicted portion corresponding to r^2 and the unpredicted, or residual, portion corresponding to $(1-r^2)$. With $r = 0.75$, we obtain $r^2 = 0.5625$ and

$$\text{predicted variability} = SS_{\text{regression}} = 0.5625(64) = 36$$

$$\text{unpredicted variability} = SS_{\text{residual}} = (1 - 0.5625)(64) = 0.4375(64) = 28$$

Using these *SS* values and the corresponding *df* values, we calculate a variance, or *MS*, for each component. For these data the *MS* values are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{residual}}} = \frac{36}{1} = 36$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} = \frac{28}{6} = 4.67$$

Finally, the *F*-ratio for evaluating the significance of the regression equation is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{36.00}{4.67} = 7.71$$

The *F*-ratio is in the critical region, so we reject the null hypothesis and conclude that the regression equation does account for a significant portion of the variance for the *Y* scores. The complete analysis of regression is summarized in Table 16.1, which is a common format for computer printouts of regression analysis.

FIGURE 16.6

The partitioning of *SS* and *df* for analysis of regression. The variability in the original *Y* scores (both SS_Y and df_Y) is partitioned into two components: (1) the variability that is explained by the regression equation, and (2) the residual variability.

TABLE 16.1

A summary table showing the results of the analysis of regression in Example 16.4.

Source	*SS*	*df*	*MS*	*F*
Regression	36	1	36.60	7.71
Residual	28	6	4.67	
Total	64	7		

SIGNIFICANCE OF REGRESSION AND SIGNIFICANCE OF THE CORRELATION

As noted earlier, in situation with a single *X* variable and a single *Y* variable, testing the significance of the regression equation is equivalent to testing the significance of the Pearson correlation. Therefore, whenever the correlation between two variables is significant, you can conclude that the regression equation is also significant. Similarly, if a correlation is not significant, then the regression equation is also not significant. For the data in Example 16.3, we concluded that the regression equation is significant. This conclusion is perfectly consistent with the corresponding test for the significance of the Pearson correlation. For these data, the Pearson correlation is $r = 0.75$ with $n = 8$. Checking Table B.6 in Appendix B, you should find a critical value of 0.707. Our correlation exceeds this criterion, so we conclude that the correlation is also significant. In fact, the critical values listed in Table B.6 were developed using the *F*-ratio (Equation 16.13) from analysis of regression.

LEARNING CHECK

1. A set of $n = 18$ pairs of scores produces a Pearson correlation of $r = 0.60$ with $SS_Y = 100$. Find $SS_{\text{regression}}$ and SS_{residual} and compute the *F*-ratio to evaluate the significance of the regression equation of predicting *Y*.

ANSWER

1. $SS_{\text{regression}} = 36$ with df = 1. $SS_{\text{residual}} = 64$ with $df = 16$. $F = 9.00$. With $df = 1, 16$, the *F*-ratio is significant with either $\alpha = .05$ or $\alpha = .01$.

16.3 INTRODUCTION TO MULTIPLE REGRESSION WITH TWO PREDICTOR VARIABLES

Thus far, we have looked at regression in situations in which one variable is being used to predict a second variable. For example, IQ scores can be used to predict academic performance for a group of college students. However, a variable such as academic performance is usually related to a variety of other factors. For example, college GPA is probably related to motivation, self-esteem, SAT score, rank in high school graduating class, parents' highest level of education, and many other variables. In this case, it is possible to combine several predictor variables to obtain a more accurate prediction. For example, IQ predicts some of academic performance, but you can probably get a better prediction if you use IQ and SAT scores together. The process of using several predictor variables to help obtain more accurate predictions is called *multiple regression.*

Although it is possible to combine a large number of predictor variables in a single multiple-regression equation, we limit our discussion to the two-predictor case. There are two practical reasons for this limitation.

1. Multiple regression, even limited to two predictors, can be relatively complex. Although we present equations for the two-predictor case, the calculations are

usually performed by a computer, so there is not much point in developing a set of complex equations when people are going to use a computer instead.

2. Usually, different predictor variables are related to each other, which means that they are often measuring and predicting the same thing. Because the variables may overlap with each other, adding another predictor variable to a regression equation does not always add to the accuracy of prediction. This situation is shown in Figure 16.7. In the figure, IQ overlaps with academic performance, which means that part of academic performance can be predicted by IQ. In this example, IQ overlaps (predicts) 40% of the variance in academic performance (combine sections a and b in the figure). The figure also shows that SAT scores overlap with academic performance, which means that part of academic performance can be predicted by knowing SAT scores. Specifically, SAT scores overlap, or predict, 30% of the variance (combine sections b and c). Thus, using both IQ and SAT scores to predict academic performance should produce better predictions than would be obtained from IQ alone. However, there is also a lot of overlap between SAT scores and IQ. In particular, much of the prediction from SAT scores overlaps with the prediction from IQ (section b). As a result, adding SAT scores as a second predictor only adds a small amount to the variance already predicted by IQ (section c). Because variables tend to overlap in this way, adding new variables beyond the first one or two predictors often does not add significantly to the quality of the prediction.

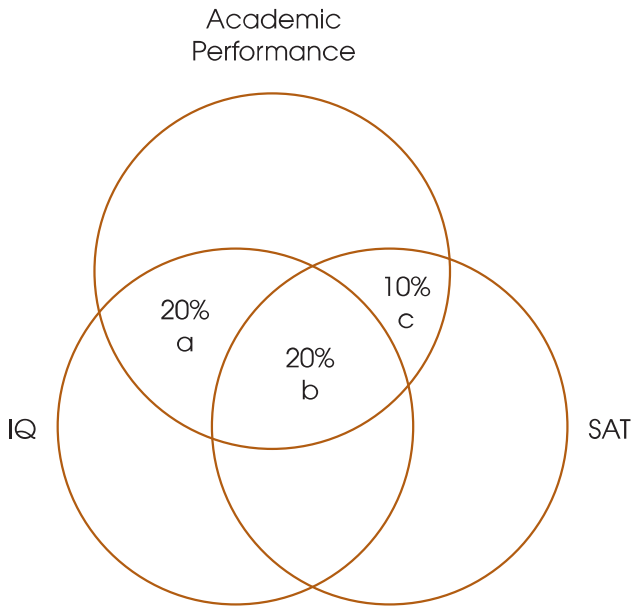
REGRESSION EQUATIONS WITH TWO PREDICTORS

We identify the two predictor variables as X_1 and X_2. The variable we are trying to predict is identified as Y. Using this notation, the general form of the multiple regression equation with two predictors is

$$\hat{Y} = b_1X_1 + b_2X_2 + a \tag{16.14}$$

FIGURE 16.7

Predicting the variance in academic performance from IQ and SAT scores. The overlap between IQ and academic performance indicates that 40% of the variance in academic performance can be predicted from IQ scores. Similarly, 30% of the variance in academic performance can be predicted from SAT scores. However, IQ and SAT also overlap, so that SAT scores contribute an additional predication of only 10% beyond what is already predicted by IQ.



If all three variables, X_1, X_2, and Y, have been standardized by transformation into z-scores, then the standardized form of the multiple regression equation predicts the z-score for each Y value. The standardized form is

$$\hat{z}_Y = (\text{beta}_1)z_{X1} + (\text{beta}_2)z_{X2} \tag{16.15}$$

Researchers rarely transform raw X and Y scores into z-scores before finding a regression equation, however, the beta values are meaningful and are usually reported by computer programs conducting multiple regression. We return to the discussion of beta values later in this section.

The goal of the multiple-regression equation is to produce the most accurate estimated values for Y. As with the single-predictor regression, this goal is accomplished with a least-squared solution. First, we define "error" as the difference between the predicted Y value from the regression equation and the actual Y value for each individual. Each error is then squared to produce uniformly positive values, and then we add the squared errors. Finally, we calculate values for b_1, b_2, and a that produce the smallest possible sum of squared errors. The derivation of the final values is beyond the scope of this text, but the final equations are as follows:

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} \tag{16.16}$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} \tag{16.17}$$

$$a = M_Y - b_1M_{X1} - b_2M_{X2} \tag{16.18}$$

In these equations, you should recognize the following SS and SP values:

SS_{X1} is the sum of squared deviations for X_1

SS_{X2} is the sum of squared deviations for X_2

SP_{X1Y} is the sum of products of deviations for X_1 and Y

SP_{X2Y} is the sum of products of deviations for X_2 and Y

SP_{X1X2} is the sum of products of deviations for X_1 and X_2

Note: More detailed information about the calculation of SS is presented in Chapter 4 (pp. 111–112) and information concerning SP is in Chapter 15 (pp. 515–516). The following example demonstrates multiple regression with two predictor variables.

EXAMPLE 16.5

We use the data in Table 16.2 to demonstrate multiple regression. Note that each individual has a Y score and two X scores that are used as predictor variables. Also note that we have already computed the SS values for Y and for both of the X scores, as well as all of the SP values. These values are used to compute the coefficients, b_1 and b_2, and the constant, a, for the regression equation.

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

TABLE 16.2

Hypothetical data consisting of three scores for each person. Two of the scores, X_1 and X_2, are used to predict the Y score for each individual.

Person	Y	X_1	X_2	
A	11	4	10	$SP_{X1Y} = 54$
B	5	5	6	$SP_{X2Y} = 47$
C	7	3	7	$SP_{X1X2} = 42$
D	3	2	4	
E	4	1	3	
F	12	7	5	
G	10	8	8	
H	4	2	4	
I	8	7	10	
J	6	1	3	
	$M_Y = 7$	$M_{X1} = 4$	$M_{X2} = 6$	
	$SS_Y = 90$	$SS_{X1} = 62$	$SS_{X2} = 64$	

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(54)(64) - (42)(47)}{(62)(64) - (42)^2} = 0.672$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(47)(62) - (42)(54)}{(62)(64) - (42)^2} = 0.293$$

$$a = M_Y - b_1M_{X1} - b_2M_{X2} = 7 - 0.672(4) - 0.293(6) = 7 - 2.688 - 1.758 = 2.554$$

Thus, the final regression equation is,

$$\hat{Y} = 0.672X_1 + 0.293X_2 + 2.554$$

Example 16.5 also demonstrates that multiple regression can be a tedious process. As a result, multiple regression is usually conducted on a computer. To demonstrate this process, we used the SPSS computer program to perform a multiple regression on the data in Table 16.2 and the output from the program is shown in Figure 16.8. At this time, focus on the Coefficients Table at the bottom of the printout. The values in the first column of Unstandardized Coefficients include the constant, b_1 and b_2 for the regression equation. We discuss other portions of the SPSS output later in this chapter.

LEARNING CHECK

1. A researcher computes a multiple-regression equation for predicting annual income for 40-year-old men based on their level of education (X_1 = number of years after high school) and their social skills (X_2 = score from a self-report questionnaire). The regression equation is $\hat{Y} = 8.3X_1 + 2.1X_2 + 3.5$ and predicts income in thousands of dollars. Two individuals are selected from the sample. One has $X_1 = 0$ and $X_2 = 16$; the other has $X_1 = 3$ and $X_2 = 12$. Compute the predicted income for each.

ANSWER

1. The first man has a predicted income of $\hat{Y} = 37.1$ thousand dollars and the second has $\hat{Y} = 53.6$ thousand dollars.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.746[a]	.557	.430	2.38788

a. Predictors: (Constant), VAR00003, VAR00002

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50.086	2	25.043	4.392	.058[a]
	Residual	39.914	7	5.702		
	Total	90.000	9			

a. Predictors: (Constant), VAR00003, VAR00002
b. Dependent Variable: VAR00001

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.552	1.944		1.313	.231
	VAR00002	.672	.407	.558	1.652	.142
	VAR00003	.293	.401	.247	.732	.488

a. Dependent Variable: VAR00001

FIGURE 16.8
The SPSS output for the multiple regression in Example 16.5.

R^2 AND RESIDUAL VARIANCE

In the same way that we computed an r^2 value to measure the percentage of variance accounted for with the single-predictor regression, it is possible to compute a corresponding percentage for multiple regression. For a multiple-regression equation, this percentage is identified by the symbol R^2. The value of R^2 describes the proportion of the total variability of the Y scores that is accounted for by the regression equation. In symbols,

$$R^2 = \frac{SS_{\text{regression}}}{SS_Y} \quad \text{or} \quad SS_{\text{regression}} = R^2 SS_Y$$

For a regression with two predictor variables, R^2 can be computed directly from the regression equation as follows:

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y} \tag{16.19}$$

For the data in Table 16.2, we obtain a value of

$$R^2 = \frac{0.672(54) + 0.293(47)}{90} = \frac{50.059}{90} = 0.5562 \text{ (or } 55.62\%)$$

In the computer printout in Figure 16.8, the value of R^2 is reported in the Model Summary table at the top.

Thus, 55.6% of the variance for the Y scores can be predicted by the regression equation. For the data in Table 16.2, SSY = 90, so the predicted portion of the variability is

$$SS_{\text{regression}} = R^2 SS_Y = 0.5562(90) = 50.06$$

The unpredicted, or residual, variance is determined by $1 - R^2$. For the data in Table 16.2, this is

$$SS_{\text{residual}} = (1 - R^2)SS_Y = 0.4438(90) = 39.94$$

COMPUTING R^2 AND $1 - R^2$ FROM THE RESIDUALS

The value of R^2 and $1 - R^2$ can also be obtained by computing the residual, or difference between the predicted Y and the actual Y for each individual, then computing the sum of the squared residuals. The resulting value is SS_{residual} and measures the unpredicted portion of the variability of Y, which is equal to $(1 - R^2)SS_Y$. For the data in Table 16.2, we first use the multiple-regression equation to compute the predicted value of Y for each individual. The process of finding and squaring each residual is shown in Table 16.3.

Note that the sum of the squared residuals, the unpredicted portion of SS_Y, is 39.960. This value corresponds to 44.4% of the variability for the Y scores:

$$\frac{SS_{\text{residual}}}{SS_Y} = \frac{39.96}{90} = 0.444 \text{ (or } 44.4\%)$$

Because the unpredicted portion of the variability is $1 - R^2 = 44.4\%$, we conclude that the predicted portion is $R^2 = 55.6\%$. Note that this answer is within rounding error of $R^2 = 55.62\%$ that we obtained from equation 16.19.

TABLE 16.3

The predicted Y values and the residuals for the data in Table 16.2. The predicted Y values were obtained using the values of X_1 and X_2 in the multiple-regression equation for each individual.

Actual Y	Predicted Y ($\hat{Y}$)	Residual ($Y - \hat{Y}$)	Squared Residual $(Y - \hat{Y})^2$
11	8.17	2.83	8.010
5	7.67	−2.67	7.129
7	6.62	0.38	0.144
3	5.07	−2.07	4.285
4	4.10	−0.10	0.010
12	8.72	3.28	10.758
10	10.27	−0.27	0.073
4	5.07	−1.07	1.145
8	10.19	−2.19	4.796
6	4.10	1.90	3.610
			39.960 = $SS_{\text{residuals}}$

THE STANDARD ERROR OF ESTIMATE

On page 567, we defined the standard error of estimate for a linear regression equation as the standard distance between the regression line and the actual data points. In more general terms, the standard error of estimate can be defined as the standard distance between the predicted Y values (from the regression equation) and the actual Y values (in the data). The more general definition applies equally well to both linear and multiple regression.

To find the standard error of estimate for either linear regression or multiple regression, we begin with SS_{residual}. For linear regression with one predictor, $SS_{\text{residual}} = (1 - r^2)SS_Y$ and has $df = n - 2$. For multiple regression with two predictors, $SS_{\text{residual}} = (1 - R^2)SS_Y$ and has $df = n - 3$. In each case, we can use the SS and df values to compute a variance or MS_{residual}.

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df}$$

The variance, or MS value, is a measure of the average squared distance between the actual Y values and the predicted Y values. By simply taking the square root, we obtain a measure of standard deviation or standard distance. This standard distance for the residuals is the standard error of estimate. Thus, for both linear regression and multiple regression

$$\text{the standard error of estimate} = \sqrt{MS_{\text{residual}}}$$

In the computer printout in Figure 16.8, the standard error of estimate is reported in the Model Summary table at the top.

For either linear or multiple regression, you do not expect the predictions from the regression equation to be perfect. In general, there is some discrepancy between the predicted values of Y and the actual values. The standard error of estimate provides a measure of how much discrepancy, on average, there is between the $\hat{Y}$ values and the actual Y values.

TESTING THE SIGNIFICANCE OF THE MULTIPLE REGRESSION EQUATION: ANALYSIS OF REGRESSION

Just as we did with the single-predictor equation, we can evaluate the significance of a multiple-regression equation by computing an F-ratio to determine whether the equation predicts a significant portion of the variance for the Y scores. The total variability of the Y scores is partitioned into two components, $SS_{\text{regression}}$ and SS_{residual}. With two predictor variables, $SS_{\text{regression}}$ has $df = 2$, and SS_{residual} has $df = n - 3$. Therefore, the two MS values for the F-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{2} \tag{16.20}$$

and

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{n - 3} \tag{16.21}$$

Because of rounding error, the value we obtain for SS_{residual} is slightly different from the value in Table 16.3.

The data for the $n = 10$ people in Table 16.2 have produced $R^2 = 0.5562$ (or 55.62%) and $SS_Y = 90$. Thus,

$$SS_{\text{regression}} = R^2SS_Y = 0.556(90) = 50.06$$

$$SS_{\text{residual}} = (1 - R^2)SS_Y = 0.4438(90) = 39.94$$

Therefore, $MS_{\text{regession}} = \dfrac{50.06}{2} = 25.03$ and $MS_{\text{residual}} = \dfrac{39.94}{7} = 5.71$

and $$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{25.03}{5.71} = 4.38$$

With $df = 2, 7$, this F-ratio is not significant with $\alpha = .05$, so we cannot conclude that the regression equation accounts for a significant portion of the variance for the Y scores.

The analysis of regression is summarized in the following table, which is a common component of the output from most computer versions of multiple regression. In the computer printout in Figure 16.8, this summary table is reported in the ANOVA table in the center.

Source	*SS*	*df*	*MS*	*F*
Regression	50.06	2	25.03	4.38
Residual	39.94	7	5.71	
Total	90.00	9		

LEARNING CHECK

1. Data from a sample of $n = 15$ individuals are used to compute a multiple-regression equation with two predictor variables. The equation has $R^2 = 0.20$ and $SS_Y = 150$.

a. Find $SS_{residual}$ and compute the standard error of estimate for the regression equation.

b. Find $SS_{regression}$ and compute the F-ratio to evaluate the significance of the regression equation.

ANSWER

1. a. $SS_{residual} = 120$. The standard error of estimate is $\sqrt{10} = 3.16$.

b. $SS_{regression} = 30$ with $df = 2$. $SS_{residual} = 120$ with $df = 12$. $F = 1.50$. With $df = 2, 12$, the F-ratio is not significant.

16.4 EVALUATING THE CONTRIBUTION OF EACH PREDICTOR VARIABLE

In addition to evaluating the overall significance of the multiple-regression equation, researchers are often interested in the relative contribution of each of the two predictor variables. Is one of the predictors responsible for more of the prediction than the other? Unfortunately, the b values in the regression equation are influenced by a variety of other factors and do not address this issue. If b_1 is larger than b_2, it does not necessarily mean that X_1 is a better predictor than X_2. However, in the standardized form of the regression equation, the relative size of the beta values is an indication of the relative contribution of the two variables. For the data in Table 16.3, the standardized regression equation is

$$\hat{z}_Y = (\text{beta}_1)z_{X1} + (\text{beta}_2)z_{X2}$$
$$= 0.558z_{X1} + 0.247z_{X2}$$

For the SPSS printout in Figure 16.8, the beta values are shown in the Coefficients table.

In this case, the larger beta value for the X_1 predictor indicates that X_1 predicts more of the variance than does X_2. The signs of the beta values are also meaningful. In this example, both betas are positive, indicating the both X_1 and X_2 are positively related to Y.

Beyond judging the relative contribution for each of the predictor variables, it also is possible to evaluate the significance of each contribution. For example, does variable X_2 make a significant contribution beyond what is already predicted by variable X_1? The null hypothesis states that the multiple-regression equation (using X_2 in addition to X_1) is not any better than the simple regression equation using X_1 as a single predictor variable. An alternative view of the null hypothesis is that the b_2 (or beta$_2$) value in the equation is not significantly different from zero. To test this hypothesis, we first determine how much more variance is predicted using X_1 and X_2 together than is predicted using X_1 alone.

Earlier we found that the multiple regression equation with both X_1 and X_2 predicted $R^2 = 55.62\%$ of the variance for the Y scores. To determine how much is predicted by X_1 alone, we begin with the correlation between X_1 and Y, which is

$$r = \frac{SP_{X1Y}}{\sqrt{(SS_{X1})(SS_Y)}} = \frac{54}{\sqrt{(62)(90)}} = \frac{54}{74.70} = 0.7229$$

Squaring the correlation produces $r^2 = (0.7229)^2 = 0.5226$ or 52.26%. This means that the relationship with X_1 predicts 52.26% of the variance for the Y scores. Therefore, the additional contribution made by adding X_2 to the regression equation can be computed as

$$(\%\text{ with both } X_1 \text{ and } X_2) - (\%\text{ with } X_1 \text{ alone})$$

$$= 55.62\% - 52.26\%$$

$$= 3.36\%$$

Because $SS_Y = 90$, the additional variability from adding X_2 as a predictor amounts to

$$SS_{\text{additional}} = 3.36\% \text{ of } 90 = 0.0336(90) = 3.024$$

This SS value has $df = 1$, and can be used to compute an F-ratio evaluating the significance of the contribution of X_2. First,

$$MS_{\text{additional}} = \frac{SS_{\text{additional}}}{1} = \frac{3.024}{1} = 3.024$$

This MS value is evaluated by computing an F-ratio with the MS_{residual} value from the multiple regression as the denominator. (*Note:* This is the same denominator that was used in the F-ratio to evaluate the significance of the multiple-regression equation.) For these data, we obtain

$$F = \frac{MS_{\text{additional}}}{MS_{\text{residual}}} = \frac{3.024}{5.71} = 0.5296$$

With $df = 1, 7$, this F-ratio is not significant. Therefore, we conclude that adding X_2 to the regression equation does not significantly improve the prediction compared to using X_1 as a single predictor. The computer printout shown in Figure 16.8 reports a t statistic instead of an F-ratio to evaluate the contribution for each predictor variable. Each t value is simply the square root of the F-ratio and is reported in the right-hand side of the Coefficients table. Variable X_2, for example, is reported as VAR00003 in the table and has $t = 0.732$, which is within rounding error of the F-ratio we obtained; $\sqrt{F} = \sqrt{0.5296} = 0.728$.

MULTIPLE REGRESSION AND PARTIAL CORRELATIONS

In Chapter 15 we introduced *partial correlation* as a technique for measuring the relationship between two variables while eliminating the influence of a third variable. At that time, we noted that partial correlations serve two general purposes:

1. A partial correlation can demonstrate that an apparent relationship between two variables is actually caused by a third variable. Thus, there is no direct relationship between the original two variables.
2. Partial correlation can demonstrate that there is a relationship between two variables even after a third variable is controlled. Thus, there really is a relationship between the original two variables that is not being caused by a third variable.

Multiple regression provides an alternative procedure for accomplishing both of these goals. Specifically, the regression analysis evaluates the contribution of each predictor variable after the influence of the other predictor has been considered. Thus, you can determine whether each predictor variable contributes to the relationship by itself or simply duplicates the contribution already made by another variable.

SUMMARY

1. When there is a general linear relationship between two variables, X and Y, it is possible to construct a linear equation that allows you to predict the Y value corresponding to any known value of X.

$$\text{predicted } Y \text{ value} = \hat{Y} = bX + a$$

The technique for determining this equation is called regression. By using a *least-squares* method to minimize the error between the predicted Y values and the actual Y values, the best-fitting line is achieved when the linear equation has

$$b = \frac{SP}{SS_X} = r\frac{s_Y}{s_X} \qquad \text{and} \qquad a = M_Y - bM_X$$

2. The linear equation generated by regression (called the regression equation) can be used to compute a predicted Y value for any value of X. However, the prediction is not perfect, so for each Y value, there is a predicted portion and an unpredicted, or residual, portion. Overall, the predicted portion of the Y score variability is measured by r^2, and the residual portion is measured by $1 - r^2$.

$$\text{predicted variability} = SS_{\text{regression}} = r^2 SS_Y$$

$$\text{unpredicted variability} = SS_{\text{residual}} = (1 - r^2)SS_Y$$

3. The residual variability can be used to compute the standard error of estimate, which provides a measure of the standard distance (or error) between the predicted Y values on the line and the actual data points. The standard error of estimate is computed by

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{n-2}} = \sqrt{MS_{\text{residual}}}$$

4. It is also possible to compute an F-ratio to evaluate the significance of the regression equation. The process is called analysis of regression and determines whether the equation predicts a significant portion of the variance for the Y scores. First a variance, or MS, value is computed for the predicted variability and the residual variability,

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df_{\text{regression}}} \qquad MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}}$$

where $df_{\text{regression}} = 1$ and $df_{\text{residual}} = n - 2$. Next, an F-ratio is computed to evaluate the significance of the regression equation.

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} \qquad \text{with } df = 1, n - 2$$

5. Multiple regression involves finding a regression equation with more than one predictor variable. With two predictors (X_1 and X_2), the equation becomes

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

with the values for b_1, b_2, and a computed using equations 16.16, 16.17, and 16.18.

6. For multiple regression, the value of R^2 describes the proportion of the total variability of the Y scores that is accounted for by the regression equation. With two predictor variables,

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y}$$

Predicted variability = $SS_{\text{regression}} = R^2 SS_Y$.
Unpredicted variability = $SS_{\text{residual}} = (1 - R^2) SS_Y$.

7. The residual variability for the multiple-regression equation can be used to compute a standard error of estimate, which provides a measure of the standard distance (or error) between the predicted Y values from the equation and the actual data points. For multiple regression with two predictors, the standard error of estimate is computed by

$$\text{standard error of estimate} = \sqrt{\frac{SS_{\text{residual}}}{n - 3}}$$

$$= \sqrt{MS_{\text{residual}}}$$

8. Evaluating the significance of the two-predictor multiple-regression equation involves computing an F-ratio that divides the $MS_{\text{regression}}$ (with $df = 2$) by the MS_{residual} (with $df = n - 3$). A significant F-ratio indicates that the regression equation accounts for a significant portion of the variance for the Y scores.

9. An F-ratio can also be used to determine whether a second predictor variable (X_2) significantly improves the prediction beyond what was already predicted by X_1. The numerator of the F-ratio measures the additional SS that is predicted by adding X_2 as a second predictor.

$$SS_{\text{additional}} = SS_{\text{regression with } X1 \text{ and } X2} - SS_{\text{regression with } X1 \text{ alone}}$$

This SS value has $df = 1$. The denominator of the F-ratio is the MS_{residual} from the two-predictor regression equation.

KEY TERMS

linear relationship (559)
linear equation (559)
slope (559)
Y-intercept (559)
regression (561)
regression line (561)
least-squared-error solution (562)
regression equation for *Y* (563)
standard error of estimate (567)
predicted variability ($SS_{\text{regression}}$) (568)
unpredicted variability (SS_{residual}) (568)
analysis of regression (570)
multiple regression (572)
partial correlation (581)

RESOURCES

Book Companion Website: www.cengage.com/psychology/gravetter

You can find a tutorial quiz and other learning exercises for Chapter 16 on the book companion website. The website also provides access to a workshop entitled *Correlation* that includes information on regression.

CENGAGEbrain.com

Log in to CengageBrain to access the resources your instructor requires. For this book, you can access:

Psychology CourseMate brings course concepts to life with interactive learning, study, and exam preparation tools that support the printed textbook. A textbook-specific website, Psychology CourseMate includes an integrated interactive eBook and other interactive learning tools including quizzes, flashcards, and more.

Visit www.cengagebrain.com to access your account and purchase materials.

SPSS

General instructions for using SPSS are presented in Appendix D. Following are detailed instructions for using SPSS to perform the **Linear Regression** and **Multiple Regression** presented in this chapter.

Data Entry

With one predictor variable (X), you enter the X values in one column and the Y values in a second column of the SPSS data editor. With two predictors (X_1 and X_2), enter the X_1 values in one column, X_2 in a second column, and Y in a third column.

Data Analysis

1. Click **Analyze** on the tool bar, select **Regression,** and click on **Linear.**
2. In the left-hand box, highlight the column label for the Y values, then click the arrow to move the column label into the **Dependent Variable** box.
3. For one predictor variable, highlight the column label for the X values and click the arrow to move it into the **Independent Variable(s)** box. For two predictor variables, highlight the X_1 and X_2 column labels, one at a time, and click the arrow to move them into the **Independent Variable(s)** box.
4. Click **OK.**

SPSS Output

We used SPSS to perform multiple regression for the data in Example 16.4 and the output is shown in Figure 16.8 (p. 576). The Model Summary table presents the values for R, R^2, and the standard error of estimate. (*Note:* For a single predictor, R is simply the Pearson correlation between X and Y.) The ANOVA table presents the analysis of regression evaluating the significance of the regression equation, including the F-ratio and the level of significance (the p value or alpha level for the test). The **Coefficients** table summarizes both the unstandardized and the standardized coefficients for the regression equation. For one predictor, the table shows the values for the constant (a) and the coefficient (b). For two predictors, the table shows the constant (a) and the two coefficients (b_1 and b_2). The standardized coefficients are the beta values. For one predictor, beta is simply the Pearson correlation between X and Y. Finally, the table uses a t statistic to evaluate the significance of each predictor variable. For one predictor variable, this is identical to the significance of the regression equation and you should find that t is equal to the square root of the F-ratio from the analysis of regression. For two predictor variables, the t values measure the significance of the contribution of each variable beyond what is already predicted by the other variable.

FOCUS ON PROBLEM SOLVING

1. A basic understanding of the Pearson correlation, including the calculation of *SP* and *SS* values, is critical for understanding and computing regression equations.
2. You can calculate $SS_{residual}$ directly by finding the residual (the difference between the actual *Y* and the predicted *Y* for each individual), squaring the residuals, and adding the squared values. However, it usually is much easier to compute r^2 (or R^2) and then find $SS_{residual} = (1 - r^2)SS_Y$.
3. The *F*-ratio for analysis of regression is usually calculated using the actual $SS_{regression}$ and $SS_{residual}$. However, you can simply use r^2 (or R^2) in place of $SS_{regression}$ and you can use $1 - r^2$ or $(1 - R^2)$ in place of $SS_{residual}$. *Note:* You must still use the correct *df* value for the numerator and the denominator.

DEMONSTRATION 16.1

LINEAR REGRESSION

The following data are used to demonstrate the process of linear regression. The scores and summary statistics are as follows:

Person	*X*	*Y*
A	0	4
B	2	1
C	8	10
D	6	9
E	4	6

$M_X = 4$ with $SS_X = 40$
$M_Y = 6$ with $SS_Y = 54$
$SP = 40$

These data produce a Pearson correlation of $r = 0.861$.

STEP 1 **Compute the values for the regression equation.** The general form of the regression equation is

$$\hat{Y} = bX + a \qquad \text{where } b = \frac{SP}{SS_X} \qquad \text{and} \qquad a = M_Y - bM_X$$

$$\text{For these data, } b = \frac{40}{40} = 1.00 \qquad \text{and} \qquad a = 6 - 1(4) = +2.00$$

Thus, the regression equation is $\hat{Y} = (1)X + 2.00$ or simply, $\hat{Y} = X + 2$.

STEP 2 **Evaluate the significance of the regression equation.** The null hypothesis states that the regression equation does not predict a significant portion of the variance for the *Y* scores. To conduct the test, the total variability for the *Y* scores, $SS_Y = 54$, is partitioned into the portion predicted by the regression equation and the residual portion.

$$SS_{regression} = r^2(SS_Y) = 0.741(54) = 40.01 \text{ with } df = 1$$

$$SS_{residual} = (1 - r^2)(SS_Y) = 0.259(54) = 13.99 \text{ with } df = n - 2 = 3$$

The two *MS* values (variances) for the *F*-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df} = \frac{40.01}{1} = 40.01$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df} = \frac{13.99}{3} = 4.66$$

And the *F*-ratio is

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{40.01}{4.66} = 8.59$$

With $df = 1, 3$ and $\alpha = .05$, the critical value for the *F*-ratio is 10.13. Therefore, we fail to reject the null hypothesis and conclude that the regression equation does not predict a significant portion of the variance for the *Y* scores.

DEMONSTRATION 16.2

MULTIPLE REGRESSION

The following data are used to demonstrate the process of multiple regression. Note that there are two predictor variables, X_1 and X_2, that are used to compute a predicted *Y* score for each individual.

Person	X_1	X_2	Y
A	0	5	2
B	3	1	4
C	5	2	7
D	6	0	9
E	8	4	5
F	2	6	3
	$M_{X1} = 4$	$M_{X2} = 3$	$M_Y = 5$
	$SS_{X1} = 42$	$SS_{X2} = 28$	$SS_Y = 34$
	$SP_{X1Y} = 27$	$SP_{X2Y} = -24$	$SP_{X1X2} = -15$

STEP 1 **Compute the values for the multiple regression equation.** The general form of the multiple-regression equation is

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

The values for the multiple regression equation are

$$b_1 = \frac{(SP_{X1Y})(SS_{X2}) - (SP_{X1X2})(SP_{X2Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(27)(28) - (-15)(-24)}{(42)(28) - (-15)^2} = 0.416$$

$$b_2 = \frac{(SP_{X2Y})(SS_{X1}) - (SP_{X1X2})(SP_{X1Y})}{(SS_{X1})(SS_{X2}) - (SP_{X1X2})^2} = \frac{(-24)(42) - (-15)(27)}{(42)(28) - (-15)^2} = -0.634$$

$$a = M_Y - b_1 M_{X1} - b_2 M_{X2} = 5 - 0.416(4) - -0.634(3) = 5 - 1.664 + 1.902 = 5.238$$

The multiple-regression equation is

$$\hat{Y} = 0.416X_1 - 0.634X_2 + 5.238$$

STEP 2 **Evaluate the significance of the regression equation.** The null hypothesis states that the regression equation does not predict a significant portion of the variance for the Y scores. To conduct the test, the total variability for the Y scores, $SS_Y = 34$, is partitioned into the portion predicted by the regression equation and the residual portion. To find each portion, we must first compute the value of R^2.

$$R^2 = \frac{b_1 SP_{X1Y} + b_2 SP_{X2Y}}{SS_Y}$$

$$\frac{(0.416)(27) + (-0.634)(-24)}{34} = 0.778 \text{ (or } 77.8\%)$$

Then, the two components for the F-ratio are

$$SS_{\text{regression}} = R^2(SS_Y) = 0.778(34) = 26.45 \text{ with } df = 2$$

$$SS_{\text{residual}} = (1 - R^2)(SS_Y) = 0.222(34) = 7.55 \text{ with } df = n - 3 = 3$$

The two MS values (variances) and the F-ratio are

$$MS_{\text{regression}} = \frac{SS_{\text{regression}}}{df} = \frac{26.45}{2} = 13.23$$

$$MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df} = \frac{7.55}{3} = 2.52$$

$$F = \frac{MS_{\text{regression}}}{MS_{\text{residual}}} = \frac{13.23}{2.52} = 5.25$$

with $df = 2, 3$, the F-ratio is not significant.

PROBLEMS

1. Sketch a graph showing the line for the equation $Y = -2X + 4$. On the same graph, show the line for $Y = X - 4$.

2. The regression equation is intended to be the "best fitting" straight line for a set of data. What is the criterion for "best fitting"?

3. A set of $n = 20$ pairs of scores (X and Y values) has $SS_X = 16$, $SS_Y = 100$, and $SP = 32$. If the mean for the X values is $M_X = 6$ and the mean for the Y values is $M_Y = 20$.
   a. Calculate the Pearson correlation for the scores.
   b. Find the regression equation for predicting Y from the X values.

4. A set of $n = 25$ pairs of scores (X and Y values) produces a regression equation of $\hat{Y} = 3X - 2$. Find the predicted Y value for each of the following X scores: 0, 1, 3, -2.

5. Briefly explain what is measured by the standard error of estimate.

6. In general, how is the magnitude of the standard error of estimate related to the value of the correlation?

7. For the following set of data, find the linear regression equation for predicting Y from X:

X	Y
7	6
9	6
6	3
12	5
9	6
5	4

8. For the following data:
- **a.** Find the regression equation for predicting Y from X.
- **b.** Calculate the Pearson correlation for these data. Use r^2 and SS_Y to compute $SS_{residual}$ and the standard error of estimate for the equation.

X	Y
1	2
4	7
3	5
2	1
5	14
3	7

9. Does the regression equation from problem 8 account for a significant portion of the variance in the Y scores? Use $\alpha = .05$ to evaluate the F-ratio.

10. For the following scores,

X	Y
3	6
6	1
3	4
3	3
5	1

- **a.** Find the regression equation for predicting Y from X.
- **b.** Calculate the predicted Y value for each X.

11. Problem 12 in Chapter 15 examined the relationship between weight and income for a sample of $n = 10$ women. Weights were classified in five categories and had a mean of $M = 3$ with $SS = 20$. Income, measured in thousands, had a mean score of $M = 66$ with $SS = 7430$, and $SP = -359$.
- **a.** Find the regression equation for predicting income from weight. (Identify the income scores as X values and the weight scores as Y values.)
- **b.** What percentage of the variance in the income is accounted for by the regression equation? (Compute the correlation, r, then find r^2.)
- **c.** Does the regression equation account for a significant portion of the variance in income? Use $\alpha = .05$ to evaluate the F-ratio.

12. A professor obtains SAT scores and freshman grade point averages (GPAs) for a group of $n = 15$ college students. The SAT scores have a mean of $M = 580$ with $SS = 22{,}400$, and the GPAs have a mean of 3.10 with $SS = 1.26$, and $SP = 84$.
- **a.** Find the regression equation for predicting GPA from SAT scores.
- **b.** What percentage of the variance in GPAs is accounted for by the regression equation? (Compute the correlation, r, then find r^2.)
- **c.** Does the regression equation account for a significant portion of the variance in GPA? Use $\alpha = .05$ to evaluate the F-ratio.

13. Problem 14 in Chapter 15 described a study examining the effectiveness of a 7-Minute Screen test for Alzheimer's disease. The study evaluated the relationship between scores from the 7-Minute Screen and scores for the same patients from a set of cognitive exams that are typically used to test for Alzheimer's disease. For a sample of $n = 9$ patients, the scores for the 7-Minute Screen averaged $M = 7$ with $SS = 92$. The cognitive test scores averaged $M = 17$ with $SS = 236$. For these data, $SP = 127$.
- **a.** Find the regression equation for predicting the cognitive scores from the 7-Minute Screen score.
- **b.** What percentage of variance in the cognitive scores is accounted for by the regression equation?
- **c.** Does the regression equation account for a significant portion of the variance in the cognitive scores? Use $\alpha = .05$ to evaluate the F-ratio.

14. There appears to be some evidence suggesting that earlier retirement may lead to memory decline (Rohwedder & Willis, 2010). The researchers gave a memory test to men and women aged 60 to 64 years in several countries that have different retirement ages. For each country, the researchers recorded the average memory score and the percentage of individuals in the 60 to 64 age range who were retired. Note that a higher percentage retired indicates a younger retirement age for that country. The following data are similar to the results from the study. Use the data to find the regression equation for predicting memory scores from the percentage of people aged 60 to 64 who are retired.

Country	% Retired (X)	Memory Score (Y)
Sweden	39	9.3
U.S.A.	48	10.9
England	59	10.7
Germany	70	9.1
Spain	74	6.4
Netherlands	78	9.1
Italy	81	7.2
France	87	7.9
Belgium	88	8.5
Austria	91	9.0

15. The regression equation is computed for a set of $n = 18$ pairs of X and Y values with a correlation of $r = +80$ and $SS_Y = 100$.
- **a.** Find the standard error of estimate for the regression equation.
- **b.** How big would the standard error be if the sample size were $n = 38$?

16. **a.** One set of 20 pairs of scores, X and Y values, produces a correlation of $r = 0.70$. If $SS_Y = 150$, find the standard error of estimate for the regression line.
- **b.** A second set of 20 pairs of X and Y values produces of correlation of $r = 0.30$. If $SS_Y = 150$, find the standard error of estimate for the regression line.

17. **a.** A researcher computes the regression equation for a sample of $n = 25$ pairs of scores, X and Y values. If an analysis of regression is used to test the significance of the equation, what are the df values for the F-ratio?
- **b.** A researcher evaluating the significance of a regression equation obtains an F-ratio with $df = 1, 18$. How many pairs of scores, X and Y values, are in the sample?

18. For the following data:
- **a.** Find the regression equation for predicting Y from X.
- **b.** Use the regression equation to find a predicted Y for each X.
- **c.** Find the difference between the actual Y value and the predicted Y value for each individual, square the differences, and add the squared values to obtain $SS_{residual}$.
- **d.** Calculate the Pearson correlation for these data. Use r^2 and SS_Y to compute $SS_{residual}$ with Equation 16.11. You should obtain the same value as in part c.

X	Y
7	16
5	2
6	1
3	2
4	9

19. A multiple-regression equation with two predictor variables produces $R^2 = .22$.
- **a.** If $SS_Y = 20$ for a sample of $n = 18$ individuals, does the equation predict a significant portion of the variance for the Y scores? Test with $\alpha = .05$.
- **b.** If $SS_Y = 20$ for a sample of $n = 8$ individuals, does the equation predict a significant portion of the variance for the Y scores? Test with $\alpha = .05$.

20. A researcher obtained the following multiple-regression equation using two predictor variables: $\hat{Y} = 0.5X_1 + 4.5X_2 + 9.6$. Given that $SS_Y = 210$, the SP value for X_1 and Y is 40, and the SP value for X_2 and Y is 9, find R^2, the percentage of variance accounted for by the equation.

21. In Chapter 15 (p. 531), we presented an example showing the general relationship among the number of churches, the number of serious crimes, and the population for a set of cities. At that time, we used a partial correlation to evaluate the relationship between churches and crime while controlling population. It is possible to use multiple regression to accomplish essentially the same purpose. For the following data,

Number of Churches (X_1)	Population (X_2)	Number of Crimes (Y)
1	1	4
2	1	1
3	1	2
4	1	3
5	1	5
7	2	8
8	2	11
9	2	9
10	2	7
11	2	10
13	3	15
14	3	14
15	3	16
16	3	17
17	3	13

- **a.** Find the multiple regression equation for predicting the number of crimes using the number of churches and population as predictor variables.
- **b.** Find the value of R^2 for the regression equation.
- **c.** The correlation between the number of crimes and population is $r = 0.961$, which means that $r^2 = .924$ (92.4%) is the proportion of variance in the number of crimes that is predicted by population size. Does adding the number of churches as a second variable in the multiple

regression equation add a significant amount to the prediction? Test with $\alpha = .05$.

22. Problem 11 in Chapter 15 examined the TV-viewing habits of adopted children in relation to their biological parents and their adoptive parents. The data are reproduced as follows. If both the biological and adoptive parents are used to predict the viewing habits of the children in a multiple-regression equation, what percentage of the variance in the children's scores would be accounted for? That is, compute R^2.

Amount of Time Spent Watching TV		
Adopted Children Y	Birth Parents X_1	Adoptive Parents X_2
2	0	1
3	3	4
6	4	2
1	1	0
3	1	0
0	2	3
5	3	2
2	1	3
5	3	3
$SS_Y = 32$	$SS_{X1} = 14$	$SS_{X2} = 16$

$SP_{X1X2} = 8$
$SP_{X1Y} = 15$
$SP_{X2Y} = 3$

23. For the data in problem 22, the correlation between the children's scores and the biological parents' scores is $r = 0.709$. Does adding the adoptive parents' scores as a second predictor significantly improve the ability to predict the children's scores? Use $\alpha = .05$ to evaluate the F-ratio.

24. For the following data, find the multiple-regression equation for predicting Y from X_1 and X_2.

X_1	X_2	Y
1	3	1
2	4	2
3	5	6
6	9	8
4	8	3
2	7	4
$M = 3$	$M = 6$	$M = 4$
$SS_{X1} = 16$	$SS_{X2} = 28$	$SS_Y = 34$

$SP_{X1X2} = 18$
$SP_{X1Y} = 19$
$SP_{X2Y} = 21$

25. A researcher evaluates the significance of a multiple-regression equation and obtains an F-ratio with $df = 2, 36$. How many participants were in the sample?

This page intentionally left blank