Simple and Multiple Linear Regression

LEARNING OBJECTIVES

- Calculate and apply a linear regression equation.
- Measure uncertainty in regression predictions.
- Describe how multiple regression works.

14.1 Simple Linear Regression

14.2 Error in Regression

14.3 Multiple Regression

CHAPTER OVERVIEW

Psychology, the study of behavior and mental processes, has four goals. The first goal is to *describe* behavior and mental processes, and the second goal is to *understand* what causes them. If the causes are known, then the third goal is *predicting* how an organism will behave, think, or feel. Accurate predictions help with the fourth goal, *influencing* how an organism behaves, thinks, or feels.

The difference between the second goal (understanding) and third goal (predicting) is the same as the difference between correlation (Chapter 13) and regression (this chapter). Correlation is about finding associations between variables, about understanding how one variable relates to another variable. This chapter, on regression, looks at the procedure for predicting one variable from the other variable. The procedure is called linear regression, and it is used to make statements like, "A person with 18 years of education is predicted to earn an annual salary of $72,000."

14.1 Simple Linear Regression

In **linear regression**, one or more predictor variables are used to predict cases' scores on an outcome variable. For example:

- If a person has X level of depression, what will be his or her level of depression after 12 sessions of cognitive-behavioral therapy?
- If we reduce truancy by X amount, how much will the high school graduation rate improve?
- If a child is bullied at age X, how will that affect her self-esteem?
- What effect does the height of the mother, the height of the father, and the annual family income have on the height of a child?

In **simple linear regression,** one predictor variable, X, is used to predict Y, the outcome variable.

Simple linear regression uses the Pearson r to develop an equation, called a regression equation, to predict Y from X. All the predictions made by a regression equation won't be perfectly accurate, but using a regression equation helps one to arrive at better decisions overall. Of course, making predictions only makes sense if there is evidence that a relationship exists between X and Y. That means simple linear regression should only be used with a statistically significant Pearson r.

Using a Regression Line for Prediction

To see how linear regression works, let's start with a straightforward example. Figure 14.1 shows a perfect correlation ($r = 1.00$) between temperature measured in Fahrenheit and in Celsius. All six data points in Figure 14.1 fall on a straight line.

For these six data points, their values on X (Fahrenheit) and Y (Celsius) are known. For example, the point on the bottom left of the scatterplot has a Fahrenheit value of 32° and a Celsius value of 0°. These six are known, but what about all the other possible Fahrenheit values? If an object's temperature is measured and found to be 86° Fahrenheit, what would it be in Celsius?

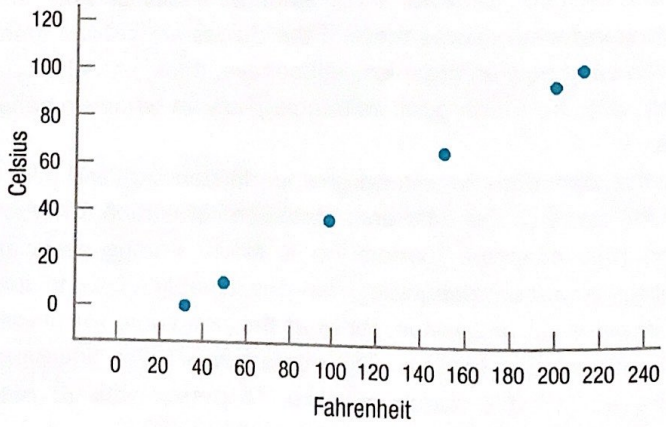


Figure 14.1 Temperature Measured in Fahrenheit and Celsius In a perfect relationship, all points in a scatterplot fall on a straight line.

Figure 14.2 shows how to estimate Y for a given value of X, like 86°. In Figure 14.2, the six points have been connected with a line. This line, called the **regression line,** allows one to find a Y value for any X value. Here's how to do it:

- Draw a vertical line from 86° on the X-axis up to the diagonal line.
- Draw a horizontal line over to the Y-axis from the point on the diagonal line.
- Estimate the value of Y where the horizontal line intersects the Y-axis, say, 30°.
- Thus, the predicted value of Y is approximately 30°.

Before moving on to the next example, let's add some terminology. The six data points in Figure 14.1 have X scores and Y scores. In the case above, we had an X score (86°F), but no Y score. The Y score that was found, 30°C, is a predicted or estimated value. A predicted value of Y has a special name, ***Y* prime**, abbreviated Y'.

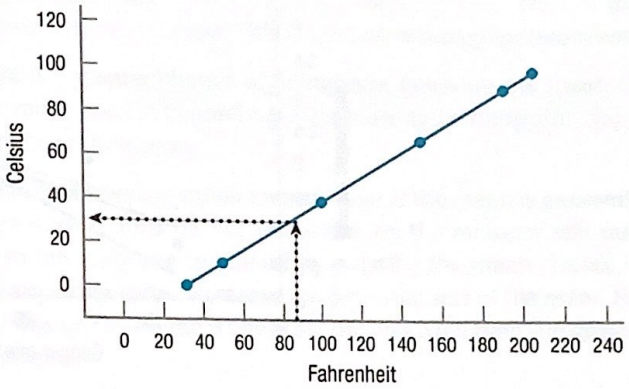


Figure 14.2 Regression Line for Predicting Celsius from Fahrenheit The regression line inserted in this scatterplot makes it easy to predict an object's temperature in Celsius if we know its temperature in Fahrenheit. An object that is 86°F would be about 30°C.

($\hat{Y}$, called "Y hat," is also commonly used as an abbreviation for the predicted value of Y.)

In Figure 14.1, all the data points fall on a line, so it is clear where to place the regression line. It is less clear what to do in a situation like that found in **Figure 14.3**, which displays Dr. Paik's marital satisfaction data from Chapter 13. In that study, a marital therapist randomly selected eight couples and found a statistically significant, positive relationship between the husband's gender role flexibility and the wife's marital satisfaction [$r(6) = .76$, $p < .05$].

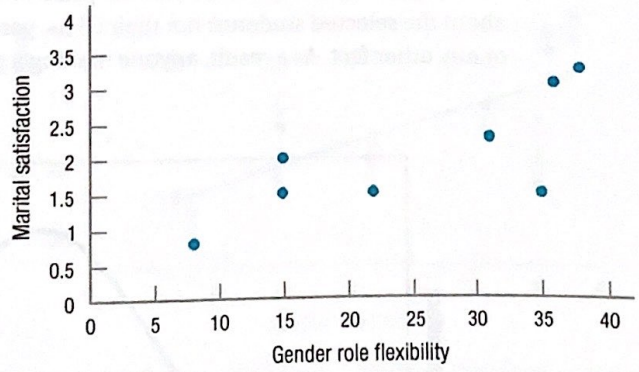


Figure 14.3 Relationship Between Gender Role Flexibility and Marital Satisfaction Though there is a strong ($r = .76$) relationship between the two variables in this scatterplot, it is not clear where the best place is to draw a regression line for predicting Y from X.

The relationship between gender role flexibility and marital satisfaction is a strong one. And a look at Figure 14.3 shows that it is a linear relationship. But where the best place would be to draw the regression line is not clear. **Figure 14.4** illustrates the marital satisfaction data with three different potential lines (labeled I, II, and III). Which one is the best regression line? Which one is the worst?

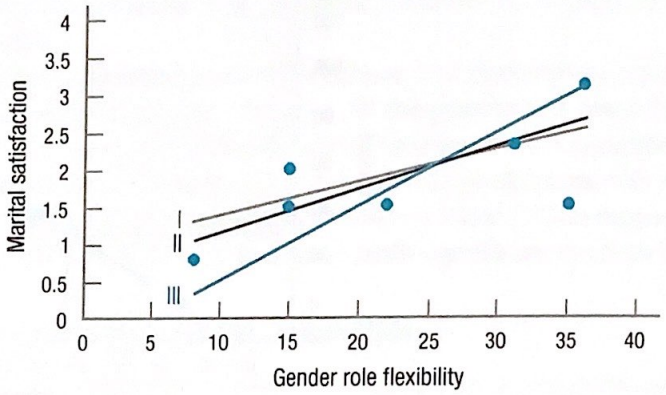


Figure 14.4 Three Potential Regression Lines for Predicting Marital Satisfaction Which of these three lines best "fits" the data in this scatterplot? By what criterion should one decide? Statisticians use the "least squares" criterion, which means the best-fitting line is the one that, overall, minimizes the discrepancies between actual Y scores and predicted Y scores.

How to Judge Whether Prediction Is Good

To learn how statisticians decide which line is the best, imagine that a memory test is given to all the students at a college. Scores can range from 1.00 to 19.00, and the mean score is calculated to be 10.00. Further, the scores are normally distributed with a standard deviation of 3.00. A frequency distribution of the memory scores is shown in Figure 14.5.

Now, imagine 12 students are randomly selected from this college and a contest is held to guess what their scores on the memory test are. Any prediction from 1.00 to 19.00 is fair and there is a substantial prize for guessing correctly. Nothing is known about the selected students: not their GPAs, years in school, histories of head trauma, or any other fact. As a result, anyone making a prediction is guessing blindly. What to

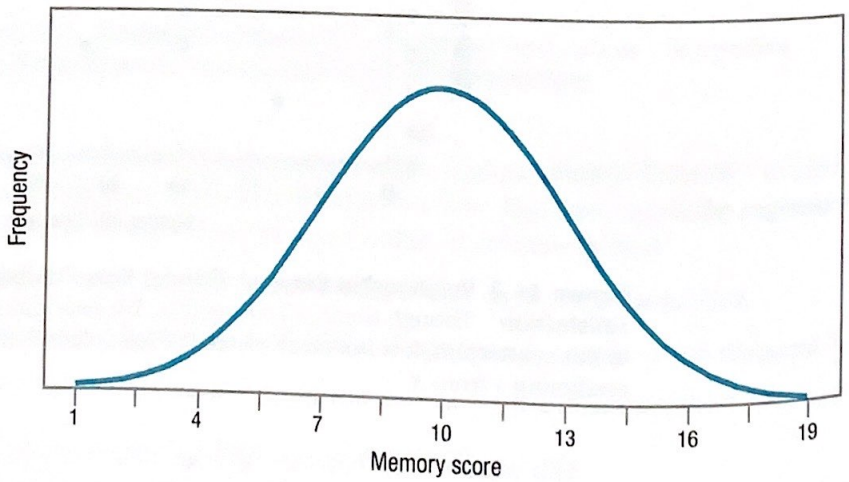


Figure 14.5 Frequency Distribution of Memory Scores If memory scores are normally distributed with a mean of 10, then the most commonly occurring score is 10. If asked to guess what a randomly selected person's memory score is, one is more likely to be right by guessing the mean than any other value.

do? A statistician would say, "Guess the mean for each one." That is, make the same guess, 10.00, twelve times in a row. This is the best strategy for two reasons:

1. First, there is a greater chance of being right guessing the mean than guessing any other value. Look at Figure 14.5—the score at the midpoint, the mean, occurs with the greatest frequency.
2. Second, the errors will be smaller, on average, if the mean is guessed for each person. Here's how to think of this. The scores on the memory test range from 1.00 to 19.00, so the most one can be off by guessing the mean (10.00) is 9.00 points. Guessing any other value increases the potential size of the error. For example, if the guess was 14.50 and the student's score was 2.00, then the guess would be off by 12.50 points.

The best prediction is the one that yields the smallest errors between predicted outcomes and actual outcomes.

The second point, about minimizing errors, is important because it is how statisticians judge prediction. The best prediction is the one that yields the smallest errors between predicted outcomes and actual outcomes. In fact, minimizing errors is how the regression line is defined—it is the best-fitting straight line by the least squares criterion. The **least squares criterion** means that the prediction errors are squared and the best-fitting line is the one that has the smallest sum of squared errors. Why are we concerned with squared values? Let's return to Dr. Paik's study.

Figure 14.6 shows the scatterplot for the marital satisfaction data with line II from Figure 14.4. In Figure 14.6, double-headed arrows are used to mark the distance for the eight cases from their actual values (the dots) to the line. These distances represent errors in prediction: the distance from Y (the wives' real satisfaction scores) to Y' (their predicted satisfaction scores) is the error in prediction. Sometimes the errors are small, as for points A, B, and E. Sometimes the errors are large, as for point F.

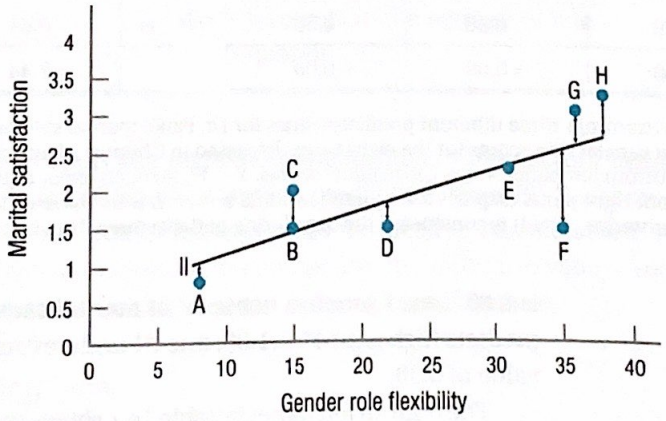


Figure 14.6 Errors in Prediction This figure compares the cases' actual marital satisfaction scores to their predicted scores for line II from Figure 14.4. The double-headed arrows from the actual cases to the line show the sizes of the errors. Notice that for some cases (e.g., A, B, and E), the errors are small and the predicted values are close to the actual values. For other cases, like F, the error is large and the predicted value is far from the actual value.

Look at the top panel in Table 14.1. The top panel shows the Y scores for the eight marital satisfaction cases. The first column gives the actual Y value for each case and the next three columns give the predicted Y scores, one for each of the three lines in Figure 14.4. The first row is for case A, where the wife's marital satisfaction score

TABLE 14.1 Three Prediction Lines for Dr. Paik's Marital Satisfaction Data

	Y: Actual Marital Satisfaction Score	Y': Marital Satisfaction Score Predicted by Line I	Y': Marital Satisfaction Score Predicted by Line II	Y': Marital Satisfaction Score Predicted by Line III
A	0.80	1.32	1.03	0.30
B	1.50	1.60	1.43	1.00
C	2.00	1.60	1.43	1.00
D	1.50	1.88	1.83	1.70
E	2.30	2.24	2.34	2.60
F	1.50	2.40	2.57	3.00
G	3.10	2.44	2.62	3.10
H	3.30	2.52	2.74	3.30

	$Y - Y'$ for Line I	$Y - Y'$ for Line II	$Y - Y'$ for Line III
A	−0.52	−0.23	0.50
B	−0.10	0.07	0.50
C	0.40	0.57	1.00
D	−0.38	−0.33	−0.20
E	0.06	−0.04	−0.30
F	−0.90	−1.07	−1.50
G	0.66	0.48	0.00
H	0.78	0.56	0.00
	$\Sigma = 0.00$	$\Sigma = 0.00$	$\Sigma = 0.00$

	$(Y - Y')^2$ for Line I	$(Y - Y')^2$ for Line II	$(Y - Y')^2$ for Line III
A	0.27	0.05	0.25
B	0.01	0.00	0.25
C	0.16	0.32	1.00
D	0.14	0.11	0.04
E	0.00	0.00	0.09
F	0.81	1.14	2.25
G	0.44	0.23	0.00
H	0.61	0.32	0.00
	$\Sigma = 2.44$	$\Sigma = 2.17$	$\Sigma = 3.88$

This set of tables examines three different prediction lines for Dr. Paik's marital satisfaction data. Each row in the top panel shows the actual marital satisfaction scores for the eight cases discussed in Chapter 13 and the scores predicted by each of the linear equations. The bottom left panel shows the residual scores, $Y - Y'$, for each linear equation. For each line, the residual scores sum to zero. The bottom right panel displays the squared residual scores. It is by the least squares criterion that the best-fitting line is selected. By this criterion, line II is considered the best-fitting of these three lines because it minimizes the sum of the squared residual scores.

is 0.80. Line I predicts her level of marital satisfaction to be higher, 1.32; line II also predicts high with $Y' = 1.03$; line III underestimates her satisfaction with a predicted value of 0.30.

The bottom left panel in Table 14.1 shows the differences between the actual scores and the predicted scores. These values are sizes of the errors. They are what is left over after Y' is subtracted out, so they are called **residuals**. For example, case A in the first row has a Y' for line I that is off by −0.52 points, for line II off by −0.23 points, and for line III off by 0.50 points. Notice that each column is a mixture of positive and negative residuals, of overestimates and underestimates. For these three lines, the residuals for each column sum to zero, meaning that the positive and negative errors balance each other out. Thus, comparing the sums of the error scores does not make one of these lines stand out over the others. So, how can one differentiate these three lines?

The answer is to square the residual scores. As a result, the squared error scores are all positive (see the bottom panel of Table 14.1) and sum to a positive number

when added together. The squared error scores sum to 2.44 for line I, 2.17 for line II, and 3.88 for line III. Linear regression uses the least squares criterion, which minimizes the sum of the squared errors, so we can now conclude that line II is the best-fitting line of these three and line III is the worst-fitting line.

Line II is the best-fitting of these three lines, but is it the best-fitting line out of all other possible lines? The regression formula we are about to learn determines the equation for the best-fitting line.

The Linear Regression Equation

Most students remember the formula for a straight line from algebra. The abbreviations may have been different, but it looked something like this:

$$Y = bX + a$$

In this equation, Y is the value being calculated; b is the slope of the line; X is the value for which Y is being calculated; and a is the point where the line intersects the Y-axis, the Y-intercept.

The regression line equation, Equation 14.1, is similar, but it calculates Y', the predicted value of Y, not Y.

Equation 14.1 Formula for Calculating a Regression Line

$$Y' = bX + a$$

where Y' = predicted value of Y
b = slope of the regression line (Equation 14.2)
X = value of X for which one wants to find Y'
a = Y-intercept of the regression line (Equation 14.3)

In order to apply the regression line formula, three factors need to be known: (1) the X value for which one wants to predict a Y value; (2) the slope, b; and (3) the Y-intercept, a. The first of these, X, does not need to be calculated. It will either be given to or determined by the researcher. But the other two values, the slope and the Y-intercept, need to be calculated in order to apply Equation 14.1.

Understanding Slope

Slope represents the tilt of the line. It tells how much up or down change in Y is predicted for each 1-unit change in X. This is often called "rise over run."

- If the slope is positive, then the line is moving up and to the right. (The slope is positive for direct relationships where increases on one variable are associated with increases on the other variable.)
- If the slope is negative, then the line is moving down and to the right. (The slope is negative for inverse relationships. In an inverse relationship, increases in X are associated with decreases in Y.)
- If the slope is zero, then the line is horizontal.

Here's the formula for calculating the slope.

Equation 14.2 Formula for the Slope, b, of the Regression Line

$$b = r\left(\frac{s_Y}{s_X}\right)$$

where b = slope of the regression line
r = observed correlation between X and Y
s_Y = standard deviation of the Y scores
s_X = standard deviation of the X scores

For Dr. Paik's marital satisfaction study, $r = .76$, $s_Y = 0.86$, and $s_X = 11.49$. He would calculate the slope as follows:

$$\begin{aligned} b &= r\left(\frac{s_Y}{s_X}\right) \\ &= .76\left(\frac{0.86}{11.49}\right) \\ &= .76 \times 0.0748 \\ &= 0.0568 \\ &= 0.06 \end{aligned}$$

The slope, 0.06, is positive. This was expected because the correlation coefficient, .76, was positive. The value of the slope, 0.06, means that, on average, for every 1-point increase in a husband's level of gender role flexibility, there is a predicted increase of 0.06 points in the wife's level of marital satisfaction. It also means that a 1-point *decrease* in gender role flexibility is associated with a 0.06-point *decrease* in marital satisfaction.

It is important to note the careful use of language here. Correlational designs give information about association, not cause and effect. Dr. Paik is careful *not* to say that a 1-point increase in gender role flexibility causes a 0.06-point increase in marital satisfaction.

Understanding the Y-Intercept

The slope was calculated first because it is needed to calculate a, the Y-intercept. The ***Y*-intercept** indicates the spot where the regression line would pass through the Y-axis. It gives information about the "altitude" of the line, how high or low it is:

- If the Y-intercept is positive, the line passes through the Y-axis above zero.
- If the Y-intercept is negative, the line passes through the Y-axis below zero.
- If the Y-intercept is zero, the line passes through the Y-axis at zero.
- The bigger the absolute value of the Y-intercept, the further away from zero the intercept passes through the Y-axis.

Here is the formula for calculating the *Y*-intercept.

Equation 14.3 Formula for the Y-Intercept, *a*, for the Regression Line

$$a = M_Y - bM_X$$

where a = *Y*-intercept for the regression line
M_Y = mean of the *Y* scores
b = slope of the regression line (Equation 14.2)
M_X = mean of the *X* scores

Dr. Paik has already calculated the slope and found it to be 0.0568, which he rounded to $b = 0.06$. Consulting his data, he finds $M_Y = 2.00$ and $M_X = 25.00$. Using these values, the *Y*-intercept is calculated as follows:

$$\begin{aligned} a &= M_Y - bM_X \\ &= 2.00 - (0.0568 \times 25.00) \\ &= 2.00 - (1.4200) \\ &= 0.5800 \\ &= 0.58 \end{aligned}$$

(*Note:* Because very precise numbers are needed for an example to work later in the chapter, this equation uses a value of the slope to four decimal places, $b = 0.0568$.)

The *Y*-intercept, the spot where the regression line would intersect the *Y*-axis, is 0.58. Now that the slope, $b = 0.06$, and the *Y*-intercept, $a = 0.58$, are known, Dr. Paik can complete the regression formula, Equation 14.1:

$$\begin{aligned} Y' &= bX + a \\ &= 0.06X + 0.58 \end{aligned}$$

Predicting *Y*

Here is how a researcher could apply the formula and use it to draw the regression line. Dr. Paik needs to select an *X* value for which to predict a *Y* score. He must select a value that is within the range used to develop the regression formula. So, he selects a gender role flexibility score of 30 and substitutes that for *X* in Equation 14.1:

$$\begin{aligned} Y' &= bX + a \\ &= (0.06 \times 30) + 0.58 \\ &= 1.8000 + 0.58 \\ &= 2.3800 \\ &= 2.38 \end{aligned}$$

Dr. Paik has just predicted that a man with a gender role flexibility score of 30 will have a partner who rates her level of marital satisfaction as 2.38. Given that marital satisfaction is rated on a 4-point scale like GPA, this means she's predicted to rate her marriage at the C+ level.

Drawing the Regression Line

Putting a regression line into a scatterplot helps to highlight the relationship between the two variables. Any two points can be connected with a straight line, so the regression line can be drawn once two points are known. All Dr. Paik needs is the two points.

The regression equation is meant to make predictions for the range of values it was based on. So, Dr. Paik will find Y' for the lowest X value (8), and Y' for the largest (38). (Again, because precise numbers are needed for an example later in the chapter, a four decimal place version of slope $b = 0.0568$ instead of $b = .06$ will be used.)

$$\begin{aligned} Y' &= bX + a \\ &= (0.0568 \times 8) + 0.58 \\ &= 0.4544 + 0.58 \\ &= 1.0344 \\ &= 1.03 \end{aligned}$$

$$\begin{aligned} Y' &= bX + a \\ &= (0.0568 \times 38) + 0.58 \\ &= 2.1584 + 0.58 \\ &= 2.7384 \\ &= 2.74 \end{aligned}$$

Dr. Paik now knows two points that anchor the line: (8, 1.03) and (38, 2.74). Figure 14.7 shows the scatterplot with the two points marked and a line drawn through them. Yes, this is the same as line II in Figure 14.4.

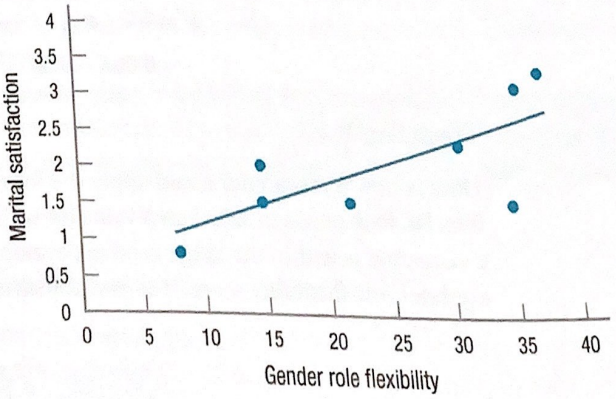


Figure 14.7 Regression Line for Predicting Marital Satisfaction Finding Y' for the two points at two ends of the range of X scores allows a researcher to draw a regression line. Remember, the regression line should only be used to predict Y' for the range of X scores used to derive the regression equation.

Worked Example 14.1 For another example of developing a regression equation from start to finish, imagine the following: A large and representative sample of cigarette smokers ($N = 2{,}500$) was obtained in order to see if there were a relationship between how much a person smoked and his or her physical health. To measure amount

Drawing the Regression Line

Putting a regression line into a scatterplot helps to highlight the relationship between the two variables. Any two points can be connected with a straight line, so the regression line can be drawn once two points are known. All Dr. Paik needs is the two points.

The regression equation is meant to make predictions for the range of values it was based on. So, Dr. Paik will find Y' for the lowest X value (8), and Y' for the largest (38). (Again, because precise numbers are needed for an example later in the chapter, a four decimal place version of slope $b = 0.0568$ instead of $b = .06$ will be used.)

$$\begin{aligned} Y' &= bX + a \\ &= (0.0568 \times 8) + 0.58 \\ &= 0.4544 + 0.58 \\ &= 1.0344 \\ &= 1.03 \end{aligned}$$

$$\begin{aligned} Y' &= bX + a \\ &= (0.0568 \times 38) + 0.58 \\ &= 2.1584 + 0.58 \\ &= 2.7384 \\ &= 2.74 \end{aligned}$$

Dr. Paik now knows two points that anchor the line: (8, 1.03) and (38, 2.74). Figure 14.7 shows the scatterplot with the two points marked and a line drawn through them. Yes, this is the same as line II in Figure 14.4.

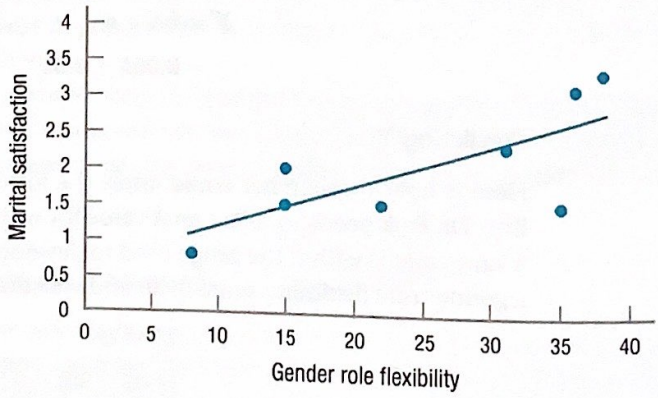


Figure 14.7 Regression Line for Predicting Marital Satisfaction Finding Y' for the two points at two ends of the range of X scores allows a researcher to draw a regression line. Remember, the regression line should only be used to predict Y' for the range of X scores used to derive the regression equation.

Worked Example 14.1 For another example of developing a regression equation from start to finish, imagine the following: A large and representative sample of cigarette smokers ($N = 2{,}500$) was obtained in order to see if there were a relationship between how much a person smoked and his or her physical health. To measure amount

of smoking, each person reported how many years he or she had been smoking cigarettes. The mean, M_X, was 22 years, with a standard deviation of 9. As a measure of physical health, each person's lung function was measured. This was reported as a percent of the predicted normal level, so lower scores mean worse functioning. A lung function score of 100 would mean that the person's lung capacity was normal for his or her age and sex. A score of 50 would mean that the smoker's level of lung function was only 50% of what was expected for a person of the same age and sex. The mean level of function was 76%, with a standard deviation of 13. The average person had been smoking for 22 years and had lungs functioning at 76% of what was expected.

Not surprisingly, the relationship between years of smoking and degree of lung function was negative and strong: [$r(2,498) = -.68$, $p < .05$]. As years of smoking went up, the percent of normal lung function went down (see the scatterplot in **Figure 14.8**).

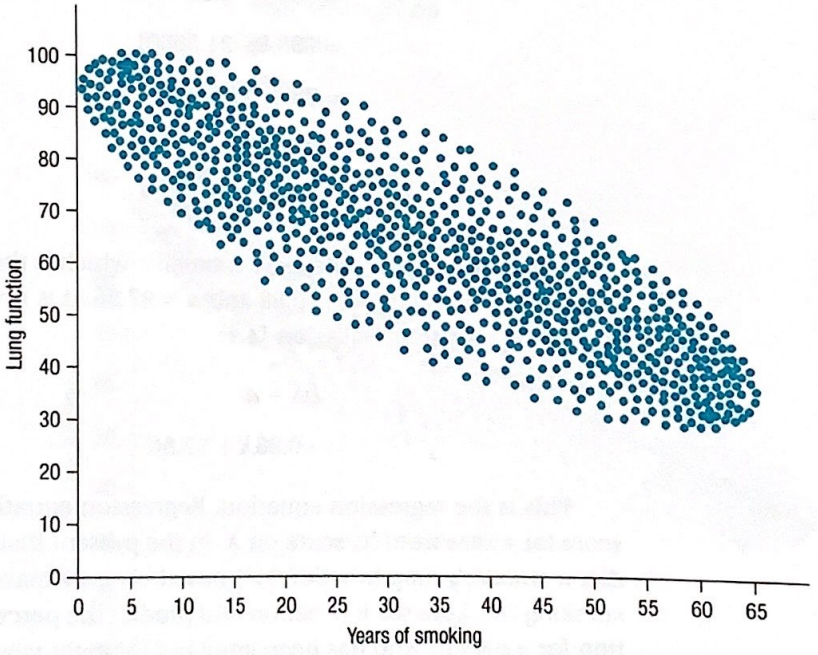


Figure 14.8 Relationship Between Years of Cigarette Smoking and Lung Function In this scatterplot, the relationship between years of smoking and loss of lung function is strong, statistically significant, and inverse.

To calculate the slope (b) for the regression line, plug $r = -.68$, $s_X = 9$, and $s_Y = 13$ into Equation 14.2:

$$b = r\left(\frac{s_Y}{s_X}\right)$$

$$= -.68\left(\frac{13}{9}\right)$$

$$= -.68 \times 1.4444$$

$$= -0.9822$$

$$= -0.98$$

For the regression line for the years of smoking/lung function data, the slope is $b = -0.98$. A slope of *negative* 0.98 means that for every 1-point *increase* in X, there's a predicted *decrease* of 0.98 points in Y. To put this in the context of this example, every year of smoking is associated with an additional decrease of 0.98 percentage points from normal lung function. In this way, a slope can be a meaningful tool for interpreting regression.

To calculate the Y-intercept, one needs the slope, which was just found to be -0.98, and the two means, M_X and M_Y. The predictor variable is years of smoking so $M_X = 22$; the predicted variable is percent of normal function and $M_Y = 76$. These values can be plugged into Equation 14.3 to find the Y-intercept:

$$\begin{aligned} a &= M_Y - bM_X \\ &= 76 - (-0.98 \times 22) \\ &= 76 - (-21.5600) \\ &= 76 + (21.5600) \\ &= 97.5600 \\ &= 97.56 \end{aligned}$$

The Y-intercept for the regression line, which is the predicted value of Y when $X = 0$, is 97.56. Given $b = -0.98$ and $a = 97.56$, it is now possible to complete the regression equation, Equation 14.1:

$$\begin{aligned} Y' &= bX + a \\ &= -0.98X + 97.56 \end{aligned}$$

This is the regression equation. Regression equations are used to predict a Y score for a case from its score on X. In the present instance, it can be used to predict a smoker's lung function (Y') based on how many years he or she has been smoking (X). Let's see it in action and predict the percentage of normal lung function for a person who has been smoking for eight years. Or, phrased mathematically, if $X = 8$, what is Y'? Applying Equation 14.1 to answer that question, it is predicted that a person who has been smoking for eight years will have lungs that function at 89.72% of normal capacity:

$$\begin{aligned} Y' &= bX + a \\ &= -0.98X + 97.56 \\ &= (-0.98 \times 8) + 97.56 \\ &= -7.8400 + 97.56 \\ &= 89.7200 \\ &= 89.72 \end{aligned}$$

Now let's draw the regression line. The regression line should only span the range of existing X values. Look at the scatterplot in Figure 14.8 and see that the X values range from 1 to 65. Below, Y' scores for these two X values are calculated

and they are used to draw the regression line seen in Figure 14.9 from (1, 96.58) to (65, 33.86):

$$\begin{aligned} Y' &= bX + a \\ &= -0.98X + 97.56 \\ &= (-0.98 \times 1) + 97.56 \\ &= -0.9800 + 97.56 \\ &= 96.5800 \\ &= 96.58 \end{aligned}$$

$$\begin{aligned} Y' &= bX + a \\ &= -0.98X + 97.56 \\ &= (-0.98 \times 65) + 97.56 \\ &= -63.7000 + 97.56 \\ &= 33.8600 \\ &= 33.86 \end{aligned}$$

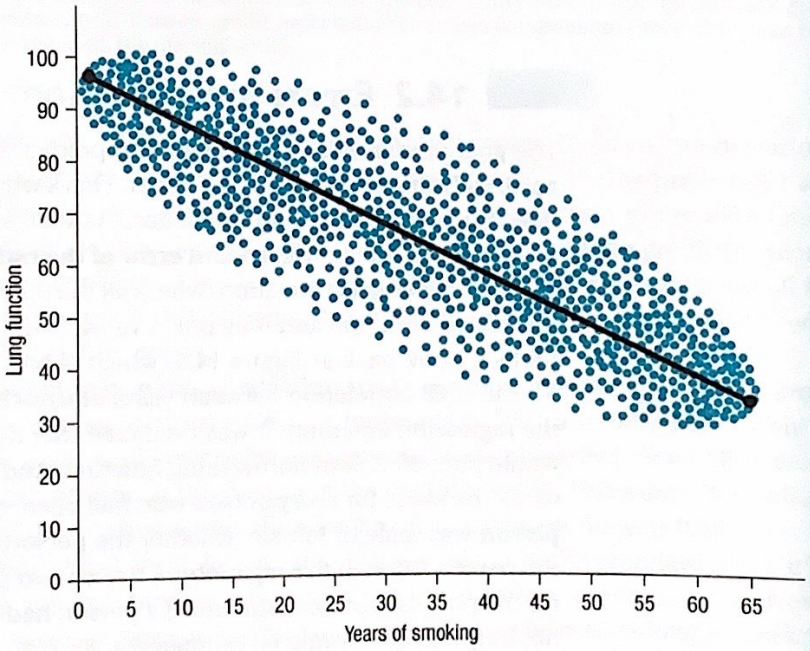


Figure 14.9 Regression Line for Years of Cigarette Smoking and Lung Function Study The regression line for this data set has a negative slope because the relationship is inverse—as the years of smoking go up, lung function goes down.

A Common Question

Q Predicting that a person who has smoked for eight years will have lungs that function at 89.72% of normal capacity sounds quite exact. Is such a precise prediction accurate?

A No. A prediction like 89.72% is a point estimate. An interval estimate, which gives a range within which Y' probably falls, is a better way to go. We will discuss such an interval, called a predication interval, in the next section, though calculating the prediction interval is beyond the scope of this text.

Practice Problems 14.1

Apply Your Knowledge

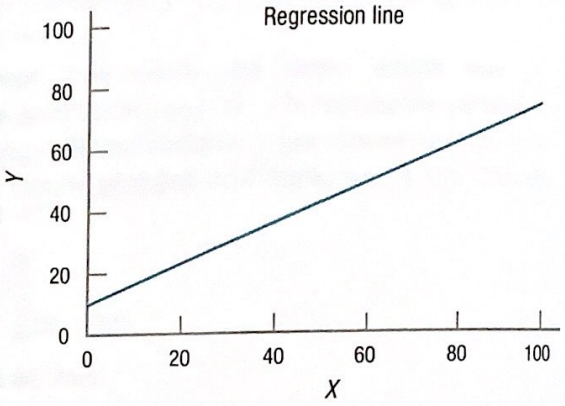
14.01 Given $r = -.37$, $s_X = 12.88$, and $s_Y = 9.33$, find the slope.

14.02 Given $M_X = 10.65$, $M_Y = 45.64$, and $b = 4.54$, find the Y-intercept.

14.03 Given a slope of 1.80 and a Y-intercept of -12.42, form a regression equation.

14.04 Given $Y' = 1.50X + 4.50$, (a) find the predicted values of Y for $X = 2$ and $X = 12$, and (b) draw the regression line.

14.05 Given the regression line to the right, estimate Y' if $X = 40$.



14.2 Errors in Regression

The goal of regression is to be able to predict Y values for X values. And the more accurately this can be done, the better. That's why researchers need a way to measure how much error occurs in prediction. As we'll soon learn, the measure for error in prediction is called the **standard error of the estimate**. Here's how it works.

With a perfect correlation, where all the data points fall on a line, each X value is associated with one and only one Y value. The story is more complex when r is not perfect. Look back at Figure 14.9, which shows the scatterplot and regression line for the $-.68$ correlation between years of smoking and loss of lung function. Using the regression equation, it was predicted that a person with eight years of smoking would have 89.72% of normal lung function. And that exact same prediction, 89.72%, would be made for every person who had been smoking for eight years. Whether the person was male or female, whether the person exercised regularly or not, whether the person smoked five cigarettes a day or two packs—all those things don't matter for the purposes of this estimate. If a person had been smoking for eight years, his or her lung capacity would be estimated at 89.72%.

But from the scatterplot in Figure 14.9, it is apparent that people who smoke for eight years have lung function scores that range from about 70% all the way up to 100%. There is variability in the scores. How much the actual scores deviate from the predicted scores is a measure of the degree of error in prediction. The more deviation there is—the larger the error is—the less sure a researcher is of the accuracy of the prediction. The statistic that summarizes how much error exists is called the standard error of the estimate.

Let's use Dr. Paik's marital satisfaction data set, with only eight cases, to calculate the standard error of the estimate. The first two columns in Table 14.2 contain his data set and the third column the predicted Y values for each of the X values. Case A, for example, has a gender role flexibility score of 8 and a marital satisfaction score of 0.8. Using the linear regression equation, its predicted marital satisfaction

TABLE 14.2 Actual, Predicted, and Residual Scores

	X: Gender Role Flexibility Score	Y: Marital Satisfaction Score	Y′: Predicted Marital Satisfaction Score	Y − Y′: Residual Score
A	8	0.8	1.03	−0.23
B	15	1.5	1.43	0.07
C	15	2	1.43	0.57
D	22	1.5	1.83	−0.33
E	31	2.3	2.34	−0.04
F	35	1.5	2.57	−1.07
G	36	3.1	2.62	0.48
H	38	3.3	2.74	0.56
	$M = 25.00$ $s = 11.49$	$M = 2.00$ $s = 0.86$		$s = 0.56$

This table shows two things. First, the standard deviation of the residual scores is the definitional formula for the standard error of the estimate. Second, it shows how the variability in Y scores can be broken down into two components—that which can be explained by X, Y', and that which is left unexplained, the residual score.

score is 1.03. The final column is labeled residual scores. It shows the deviation of the predicted score from the actual score, calculated as $Y - Y'$. The predicted Y score for case A was off by 0.23 points and, because of the direction of the difference, is reported as −0.23. This difference is a measure of how wrong the predicted score is, so it is a measure of error. Case B, for example, where the predicted score was off from the actual score by 0.07 points, had less error in its predicted score than case C, where $Y - Y'$ was off by 0.57 points.

The last column contains deviation scores (error scores), which sum to zero. So, how can the average amount of error be represented? With a standard deviation! And that is what a standard error of the estimate is, the standard deviation of the residual scores. As can be seen at the bottom of the column, for Dr. Paik's data, the standard deviation of the residual scores, the standard error of the estimate, is 0.56.

The definitional formula requires that we calculate the standard error of the estimate as the standard deviation of the residual scores. Equation 14.4 gives the easier-to-use computational formula, a formula that can be used as long as one knows r and s_Y.

Equation 14.4 Formula for the Standard Error of the Estimate

$$s_{Y-Y'} = s_Y\sqrt{1 - r^2}$$

where $s_{Y-Y'}$ = standard error of the estimate
s_Y = standard deviation of the Y scores
r = the Pearson r value

This equation says that the standard error of the estimate may be calculated by (1) squaring the correlation coefficient, (2) subtracting the square from 1, (3) taking

the square root of the difference, and (4) multiplying this square root by the standard deviation of the Y scores. Here are the calculations for Dr. Paik's data. The value about to be calculated, $s_{Y-Y'} = 0.56$, is exactly the same value that was found as the standard deviation of the difference scores in Table 14.2:

$$\begin{aligned} s_{Y-Y'} &= s_Y\sqrt{1 - r^2} \\ &= 0.86\sqrt{1 - .76^2} \\ &= 0.86\sqrt{1 - .5776} \\ &= 0.86\sqrt{.4224} \\ &= 0.86 \times .6499 \\ &= 0.5589 \\ &= 0.56 \end{aligned}$$

What does a standard error of the estimate of 0.56 mean? Loosely, one can think of standard error of the estimate as the average residual score, the average difference between the actual Y scores and the predicted Y scores. Is 0.56 a lot of error? It depends on the possible range of scores. Here the variable being predicted is marital satisfaction, which is measured on a scale ranging from 0 to 4. Being off by 0.56 points, on average, on a 4-point scale means being off, on average, by 14%. That's not good.

Want a concrete example? Suppose Neil goes to a county fair and stops at an "I'll Guess Your Weight" booth. The carny guesses Neil's weight as 150 pounds. But, if he's off by 14%, Neil could weigh 171 pounds and the carny underestimated his weight. The error could go the other way as well. The carny could have overestimated Neil's weight. Maybe Neil only weighs 129 pounds, which is off from 150 by 14% in the opposite direction.

This range, from 129 pounds to 171 pounds, gives the general idea of what a prediction interval is. A **prediction interval** gives a range within which there is some certainty that a case's real Y score falls. The calculation of the interval is based on the estimated Y score and the standard error of the estimate. The smaller the standard error of the estimate, the narrower the prediction interval and the better the prediction.

Worked Example 14.2 For another example of calculating the standard error of the estimate, a return to the cigarette smoking and lung function study is in order. In that study, 2,500 smokers reported how many years they had been smoking ($M = 22$, $s = 9$) and had their lung function measured as a percentage of normal ($M = 76$, $s = 13$). There was a strong and statistically significant inverse relationship, $r = -.68$: the longer people smoked, the lower their lung function. After a regression equation was developed, it was used to predict that a person who had been smoking for eight years would have lungs functioning at 89.72% of normal capacity. How much confidence should we have that this estimate is accurate?

The way to answer this is by calculating $s_{Y-Y'}$ using Equation 14.4:

$$\begin{aligned} s_{Y-Y'} &= s_Y\sqrt{1-r^2} \\ &= 13\sqrt{1-(-.68^2)} \\ &= 13\sqrt{1-.4624} \\ &= 13\sqrt{.5376} \\ &= 13 \times .7332 \\ &= 9.5316 \\ &= 9.53 \end{aligned}$$

This standard error of the estimate of 9.53 means that the actual Y scores and Y' scores for the 2,500 people in the sample differed by almost 10 points, on average, on a 100-point scale. That seems like a fair amount of error. This suggests that predictions based on this regression equation aren't very accurate.

A Common Question

Q So far, both examples have had standard errors of the estimate that are large, suggesting prediction is not very good. What does it take to have a small standard error of the estimate?

A As r grows larger and s_Y becomes smaller, $s_{Y-Y'}$ gets smaller.

Practice Problems 14.2

Apply Your Knowledge

14.06 Given $r = .42$ and $s_Y = 5.64$, find $s_{Y-Y'}$.

14.07 Dr. Binet developed a regression equation to predict adult IQ from childhood language abilities. IQ can range from 55 to 145. The standard error of the estimate for the regression equation is 14. Is that a large error of the estimate?

14.3 Multiple Regression

Here's a thought experiment. In which scenario could one more accurately predict a student's GPA?

A. Knowing how many hours the student spends on schoolwork each week

B. Knowing how many hours the student spends on schoolwork each week *plus* his or her high school GPA, his or her IQ, and how much alcohol the student consumes each week

Most people believe the additional information in Scenario B is relevant to predicting academic performance and they are correct. In Scenario B, the prediction should be more accurate because more factors are taken into account.

The difference between Scenario A and Scenario B is the difference between simple regression and multiple regression. The previous section focused on simple linear regression. **Simple regression** uses just one predictor variable to calculate Y'. **Multiple regression** uses several predictor variables to calculate Y'. If the different X variables have different influences on the outcome variable being predicted, when they are combined, they will do a better job of prediction than any one variable by itself.

r^2, the percentage of variability in the outcome variable that is accounted for by the predictor variable(s), is called R^2 in multiple regression. Better prediction means a higher percentage of variability is accounted for with multiple regression than with simple regression. In this way, multiple regression is a more powerful technique than simple regression.

Deriving a multiple regression equation is beyond the scope of this text. But, here is an example to show how it works. Every year, colleges have many more applicants than they can admit. Part of the admissions process involves deciding which applicants can do college-level work. Multiple regression plays a role in predicting which applicants will fare well in college.

The College Board, the folks who created the SAT, provide a service to colleges that it calls ACES, the Admitted Class Evaluation Service. ACES uses admissions information from a first-year class to develop a multiple regression equation to predict first-year GPA. Once this equation is developed, the college can apply it in subsequent years to applicants to predict what their GPAs will be. The college can decide whom to admit, objectively, on the basis of predicted GPA.

The College Board offers a sample ACES report on its website (collegeboard.com). Using hypothetical data, the Board examines how well four variables—SAT reading subtest scores, SAT writing subtest scores, SAT math subtest scores, and high school class rank—predict first-year GPA for about a thousand students at one college.* [High school class rank is transformed to range from 100 (the best student) to 0 (the worst student).] Here are the Pearson r correlation coefficients for each of these variables predicting GPA by itself:

- SAT reading test, $r = .42$
- SAT writing test, $r = .42$
- SAT math test, $r = .39$
- High school class rank, $r = .52$

These r's are all fairly close to each other in terms of size. The r with the strongest correlation with GPA, meaning the one that is the strongest predictor, is high school class rank. There is certainly some overlap in what these four variables measure and how well they predict GPA. For example, general intelligence level plays a role in all four scores and intelligence plays a role in determining GPA. But, each of the four variables also measures something unique. For example, part of how well one does on the math test is not a result of one's general level of intelligence or the reading and writing skills that help on any test. But, to some degree, performance on a math test is determined by specialized math skills. And, to some degree, these same specialized math skills play a role in some of the courses that determine the GPA. Multiple regression adds together the unique predictive power of each variable. As a result, multiple regression usually accounts for a bigger percentage of the variability in the outcome variable than is accounted for by any single variable.

* Note: This example is based on the three-section SAT in use prior to 2016. Beginning March 2016, the SAT includes only two sections, Reading/Writing and Math.

When the four College Board variables are combined together to predict GPA in a multiple regression, the correlation climbs to $R = .57$. (The abbreviation for the correlation coefficient for multiple regression is R, not r.) This doesn't sound like much of an increase from the .52 correlation between class rank and GPA. But, it is. The percentage of variance explained changes from 27.04% to 32.49%. Predicting an extra 5 percentage points of variability is very worthwhile.

The multiple regression equation the College Board develops for a college can be used to predict GPA from SAT scores and high school rank for a potential student. Their equation is a more complex version of the linear regression equation from earlier in this chapter. The equation has "weights" for each of the predictor variables. The weights are like the slope in the linear regression equation. And there is a constant that is like the Y-intercept. When all of this information is put together, it makes for a long equation. Here is how estimated GPA, GPA', would be calculated:

$$GPA' = (SAT_{\text{ReadingScore}} \times Weight_{\text{ReadingScore}}) + (SAT_{\text{WritingScore}} \times Weight_{\text{WritingScore}}) + (SAT_{\text{MathScore}} \times Weight_{\text{MathScore}}) + (HSRank \times Weight_{\text{HSRank}}) + Constant$$

Here are the four weights and the constant for the College Board example:

- Reading weight = 0.0012
- Writing weight = 0.0013
- Math weight = 0.0006
- HS rank weight = 0.0029
- Constant = 0.7821

If an applicant were good at reading (SAT score = 600), not so good at writing (SAT score = 450), very good at math (SAT score = 760), and had a very good class rank (90), then her predicted GPA would be

$$\begin{aligned} GPA' &= (SAT_{\text{ReadingScore}} \times Weight_{\text{ReadingScore}}) + (SAT_{\text{WritingScore}} \times Weight_{\text{WritingScore}}) + \\ &\quad (SAT_{\text{MathScore}} \times Weight_{\text{MathScore}}) + (HSRank \times Weight_{\text{HSRank}}) + Constant \\ &= (600 \times 0.0012) + (450 \times 0.0013) + (760 \times 0.0006) + (90 \times 0.0029) + 0.7821 \\ &= 0.7200 + 0.5850 + 0.4560 + 0.2610 + 0.7821 = 2.8041 = 2.80 \end{aligned}$$

A person with those SAT scores and class rank would be predicted to end up with a GPA of 2.80 at the end of her first year.

The multiple regression equation is built from cases where the first-year GPA is known. The equation can be used to predict first-year GPA for these students. As a result, the students have both actual and predicted GPAs, and it is possible to see how well the predicted GPA predicts the actual GPA. The correlation between the two is .46. Multiple regression makes objective predictions that minimize errors in prediction *overall*. In that sense, it makes better decisions. But, unless $R = 1.00$, it doesn't make perfect predictions.

Worked Example 14.3 Many Americans are trying to lose weight, either by counting calories or by using Weight Watchers®. One of the two Weight Watchers plans counts points, not calories. In its system, foods are assigned a point value based on a mysterious combination of how much protein, carbohydrates, fat, and fiber the food contains. The number of calories in the food is not part of the equation. With the point system,

a cup of lettuce is worth 0.3 points and a McDonald's Quarter Pounder is worth 13.4 points.

Imagine that a nutritionist, Dr. Feldman, wanted to crack the secret equation that Weight Watchers uses and figure out how these four variables—protein, carbs, fat, and fiber—are combined to generate a point score. This calls for multiple regression.

First, Dr. Feldman draws a random sample of foods and for each one he finds out how much protein, carbs, fat, and fiber the food contains. He also consults the Weight Watchers Web site and locates the point value for each food. Armed with these four predictor variables (protein, carbs, fat, and fiber) and the one outcome variable (points), he uses SPSS to find the multiple regression equation. The equation that calculates *Points'*, the estimated number of points, is

$$\begin{aligned} Points' &= (Grams_{Protein} \times Weight_{Protein}) + (Grams_{Carbs} \times Weight_{Carbs}) + (Grams_{Fat} \times \\ &\quad Weight_{Fat}) + (Grams_{Fiber} \times Weight_{Fiber}) + Constant \\ &= (Grams_{Protein} \times 0.074) + (Grams_{Carbs} \times 0.096) + (Grams_{Fat} \times 0.279) + \\ &\quad (Grams_{Fiber} \times -0.101) + 0.112 \end{aligned}$$

Note that three of the weights are positive, but the weight for fiber is negative. This reveals that fiber plays a different role in determining points than do the other three variables. As the levels of proteins, carbohydrates, and fats in a food go up, so does the point value for the food. However, as the amount of fiber in a food goes up, the point value goes down.

An important use of a regression equation is to predict a value for a new case. Suppose Dr. Feldman is about to eat a BLT and wants to know how many points it is worth. From the menu, he learns that the sandwich has 15 grams of protein, 28 grams of carbohydrates, 17 grams of fat, and 3 grams of fiber. Here's how he would calculate its points:

$$\begin{aligned} Points\ BLT' &= (Grams_{Protein} \times 0.074) + (Grams_{Carbs} \times 0.096) + (Grams_{Fat} \times 0.279) \\ &\quad + (Grams_{Fiber} \times -0.101) + 0.112 = (15 \times 0.074) + (28 \times 0.096) \\ &\quad + (17 \times 0.279) + (3 \times -0.101) + 0.112 \\ &= 1.1100 + 2.6880 + 4.7430 - 0.3030 + 0.112 = 8.3500 = 8.35 \end{aligned}$$

Dr. Feldman has now estimated (predicted) that eating a BLT at lunch will use up 8.35 of a person's daily point allowance.

Practice Problems 14.3

Review Your Knowledge

14.08 Explain why multiple regression explains a larger percentage of variability in the predicted variable than does simple regression.

Apply Your Knowledge

14.09 A multiple regression equation has a constant of 55.12, a weight of 13.17 for variable 1, and a weight of 4.55 for variable 2. If a case has a score of 12 on variable 1 and a score of 33 on variable 2, what is Y'?

Application Demonstration

Let's see multiple regression in action. In this cost-conscious era, hospitals try to save money by reducing the length of stay of their patients. It would be beneficial to a hospital if it could predict a patient's length of stay at the time of admission. If so, therapeutic resources could be directed to the patients predicted to be in the hospital for a long time, in order to help them get better more quickly.

Some researchers turned their attention to predicting length of stay for patients admitted to a large, metropolitan psychiatric hospital (Huntley, Cho, Christman, & Csernansky, 1998). In a six-month period, almost 800 patients were admitted to the facility and they spent an average of 16.3 days in the hospital. The hospital database contained a lot of information about each patient, including each patient's sex, age, primary and secondary diagnoses, number of prior admissions, and legal status. The researchers combined these variables using multiple regression to see if length of stay could be predicted.

There turned out to be five variables that played a statistically significant role in predicting a patient's length of stay: (1) a primary diagnosis of schizophrenia, (2) the number of previous admissions, (3) a primary diagnosis of a mood disorder, (4) age, and (5) an alcohol or drug problem as a secondary diagnosis. These variables can be thought of as reflecting difficult cases. For example, someone with five previous psychiatric admissions probably has a more severe problem, one that may take longer to treat, than a patient for whom this admission is the first hospitalization.

Together, these five variables predicted 17% of the variance in length of stay. This may not sound like much, but Cohen (1988) would call it a medium effect. Is it enough to be useful?

So far, what these researchers did is not unusual. But, now their work took an interesting direction. They used their regression equation to calculate the predicted length of stay for each patient. As a result, there were two pieces of data for each patient—the actual length of stay and the predicted length of stay. The researchers then added a third variable for each patient—the psychiatrist in charge of the patient's care. There were 12 psychiatrists at this hospital and newly admitted patients were assigned to their care on a rotating basis. In essence, patients were randomly assigned to psychiatrists.

If patients are randomly assigned to psychiatrists and if all psychiatrists provide equivalent care, then the mean length of stay should be roughly the same for each psychiatrist. The blue bars in **Figure 14.10** show the mean length of stay of the patients for each of the psychiatrists—it ranges from less than 10 days (Psychiatrist 1) to more than 25 days (Psychiatrist 12). Either differences in the effectiveness of the psychiatrists exist or some psychiatrists had more or less than their fair share of hard-to-treat patients.

How could one tell if a psychiatrist were assigned difficult or easy patients? Difficult patients should have a longer predicted length of stay. So, calculating the mean predicted length of stay for each psychiatrist should answer that question. The grey bars in Figure 14.10 show the mean predicted length of stay corresponding to each psychiatrist.

Look at Psychiatrist 1. Earlier, when the focus was only on the blue bars showing actual length of stay, he appeared to be doing a good job because his patients had the shortest length of stay. Now, looking at the grey bar, it is apparent that

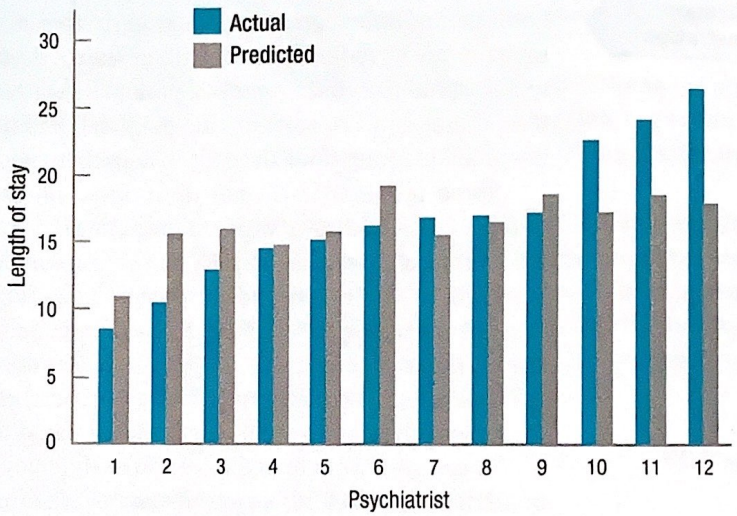


Figure 14.10 Actual and Predicted Length of Stay for Patients Treated by Different Psychiatrists Multiple regression was used to calculate the predicted length of stay for each patient. Comparing the two means allows one to speculate about a psychiatrist's skills. (Data from Huntley, Cho, Christman, & Csernansky, 1998.)

this psychiatrist was assigned the healthiest patients. No wonder he discharged them quickly.

The two bars, one the actual length of stay and the other predicted by multiple regression, allow us to think about this psychiatrist's performance in a more complex fashion. Psychiatrist 1's patients have a mean length of stay around 8 days but are predicted to need approximately 11 days. Why the difference? There are three likely explanations, two of which involve his skills and one that involves multiple regression. First, perhaps he is a phenomenal psychiatrist and cures people quickly. It could happen. Second, maybe he is a terrible psychiatrist who can't assess patients' progress and discharges them before they are ready. That could happen, too. Third, maybe there are errors in prediction. Maybe this is especially true for the healthier cases and their lengths of stay are overestimated.

Whatever the explanation turns out to be, this study shows how multiple regression is used in psychology and how predictions are made and utilized. Regression, either simple or multiple, is a useful tool that helps researchers understand their results in more detail.

SUMMARY

Calculate and apply a linear regression equation for a Pearson correlation coefficient.

- Linear regression predicts a value of Y, Y', for X when there is a statistically significant relationship between X and Y. The prediction equation uses the slope and Y-intercept to generate a regression line, the best-fitting line that minimizes the errors between Y and Y'. Slope indicates how much change in Y is predicted for each 1-unit change in X, and the Y-intercept tells where the line passes through the Y-axis.

- As r approaches zero, the regression line becomes horizontal and predicted Y values approach M_Y. When $r = 0$, then X doesn't predict Y and the best prediction that can be made for Y' is M_Y.

Measure uncertainty in regression predictions.

- Error in prediction is the difference between the actual score, Y, and the predicted score, Y'.
- The average amount of error is summarized in a statistic called the standard error of the estimate, which is the standard deviation of the residual scores.

Describe how multiple regression works.

- Simple regression uses a single predictor variable to predict Y'; multiple regression uses two or more predictor variables. By combining the unique predictive ability of multiple predictor variables, multiple regression accounts for more variability in the outcome variable.

KEY TERMS

least squares criterion – prediction errors are squared and the best-fitting regression line is the one that has the smallest sum of squared errors.

linear regression – a predictor variable is used to predict a case's score on another variable and the prediction equation takes the form of a straight line.

multiple linear regression – prediction in which multiple predictor variables are combined to predict an outcome variable.

prediction interval – a range around Y' within which there is some certainty that a case's real value of Y falls.

regression line – the best-fitting straight line for predicting Y from X.

residual – the difference between an actual score and a predicted score; the size of the error in prediction.

simple linear regression – prediction in which Y' is predicted from a single predictor variable.

slope – the tilt of the line; rise over run; how much up or down change in Y is predicted for each 1-unit change in X.

standard error of the estimate – the standard deviation of the residual scores, a measure of error in regression.

Y-intercept – the spot where the regression line would pass through the Y-axis.

Y prime – the value of Y predicted from X by a regression equation; Y'.

DIY

In the DIY of Chapter 13, you calculated the correlation between foot size and height. Now, take that same correlation coefficient and generate the regression equation to predict height from foot size. When you have arrived at the equation, use it to calculate Y' for the students on whom the equation was based. For each of the cases, calculate residual scores. Do they sum to zero? Now, find the standard deviation of the residual scores. Then, use Equation 14.4 to calculate the standard error of the estimate. Is that the same value you calculated for the standard deviation?

Want more fun? Select 10 new cases and use the regression equation to calculate Y' scores for them. Will the regression equation be as accurate for them as it was for the original group? Investigate this by calculating residual scores and finding their standard deviation. Is it larger or smaller than the first standard deviation? Why?