11

EXPLORING THE RELATIONSHIP BETWEEN TWO VARIABLES

The Linear Regression Model

In the last chapter, we reviewed two models for testing hypotheses about the relationship between two variables. By making one simple assumption (i.e., when there is no relationship between the two variables, any *X*-value in our data could occur with any *Y*-value in our data), we can perform a randomization (permutation) test. We can also use the classical statistical model to conduct hypothesis tests about the parameter of a bivariate normal distribution called ρ, which describes the degree of relationship between the two variables for that population. The latter tests are not restricted to whether or not there is a relationship between the two variables (does $\rho = 0$?); we can also test hypotheses about the specific value of ρ (for example, does $\rho = .5$?) and about whether the population correlations are the same or different in different bivariate normal distributions (does $\rho_1 = \rho_2$?).

In this chapter, we will explore another classical statistical model for the relationship between two variables called the **linear regression model**. This model makes different assumptions about the population from which we sample. Although we can only test the hypothesis that $\rho = 0$ with this model, it is more useful because it provides the foundation for analysis of variance, multivariate analysis, and hypothesis tests when the relationship is not linear.

To illustrate the use of the linear regression model, we will return our focus to the research study conducted by McManus, Feyes, and Saucier (2011) that we introduced in Chapter 10. Recall that the researchers were interested in examining the factors associated with prejudicial attitudes toward individuals with intellectual disabilities. As we previously described, they used a sample of 125 undergraduate students and assessed their quantity of contact with individuals with intellectual disabilities, their quality of contact with individuals with intellectual disabilities, and the amount of knowledge they reported that they possessed about intellectual disabilities. The participants reported their levels of these variables on several items on scales from 1 (*strongly disagree*) to 9 (*strongly agree*), and their responses were averaged to produce scores for each of

these three factors, with higher scores reflecting higher levels of quantity of contact, quality of contact, and perceived knowledge, respectively. The researchers assessed the participants' levels of prejudicial attitudes toward individuals with intellectual disabilities using the Mental Retardation Attitude Inventory–Revised (MRAI-R; Antonak & Harth, 1994). This measure also employed several items using scales from 1 (*strongly disagree*) to 9 (*strongly agree*), and the participants' responses were averaged, with higher scores reflecting more positive (i.e., less prejudicial) attitudes toward individuals with intellectual disabilities.

As we previously noted, consistent with their predictions, McManus et al. (2011) found that participants' levels of quantity and quality of contact with individuals with intellectual disabilities and their perceived knowledge about intellectual disabilities were positively correlated with their scores on the MRAI-R. Thus, as levels of quantity and quality of contact increased and as perceived knowledge increased, so did positive attitudes toward individuals with intellectual disabilities.

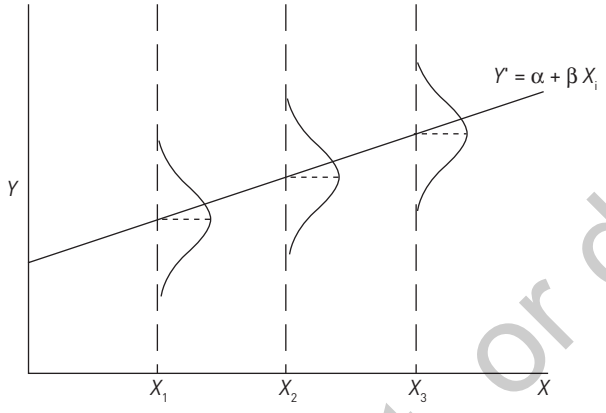
In Chapter 10, we focused on the relationship between the participants' quality of contact with individuals with intellectual disabilities and their attitudes toward individuals with intellectual disabilities. In this chapter, we will again focus on this relationship by using the participants' quality of contact with individuals with intellectual disabilities as a predictor variable and the participants' scores on the MRAI-R as a criterion variable.

ASSUMPTIONS FOR THE LINEAR REGRESSION MODEL

Unlike with the bivariate normal distribution model, we make no assumptions about the distribution of the values on the X*-axis. However, we assume that for every value of* X*, the* Y *scores are normally distributed around a straight line* $Y' = \alpha + \beta X$ *and the variances of those normal distributions are equal.*

In the linear regression model, we make *no assumptions about the distribution of the variable we call* X *or how we obtain the values of* X: *X* can have a normal distribution as in the bivariate normal distribution model, and we can randomly sample the values of *X* (as in that model). On the other hand, *X* could be a nominal variable for which the values are preselected. However, no matter what the distribution of *X* is or how we obtain the values of *X*, in this model, we do assume the following for each *X*: the values of *Y* are normally distributed, all of these normal distributions of *Y* values have the same variance, and the means of these normal distributions lie on a straight line. We can estimate the parameters and test hypotheses by taking a random sample from this population. These assumptions can be summarized as follows:

Figure 11.1 ■ This population consists of three values of X (X_1, X_2, X_3). For each X_i, the Y-values have a normal distribution. The variances of these normal distributions are the same, and the means of these normal distributions are on the line $Y' = \alpha + \beta X$.



1. The variable we call X can have any distribution, and it does not matter whether the values of X were obtained by random sampling or were preselected.
2. For every value of X, the associated values of Y have a normal distribution.
3. The means of these normal distributions of Y values lie on a straight line. The equation for that line is $Y' = \alpha + \beta X$, where α is the Y-intercept of the line and β is the slope. This line is called **population regression line**.
4. These normal distributions of Y scores all have the same variance around that straight line. (This assumption is called **homoscedasticity**.)
5. We can estimate the values of the two parameters α and β by taking a random sample from this population.

These assumptions are represented graphically in Figure 11.1.

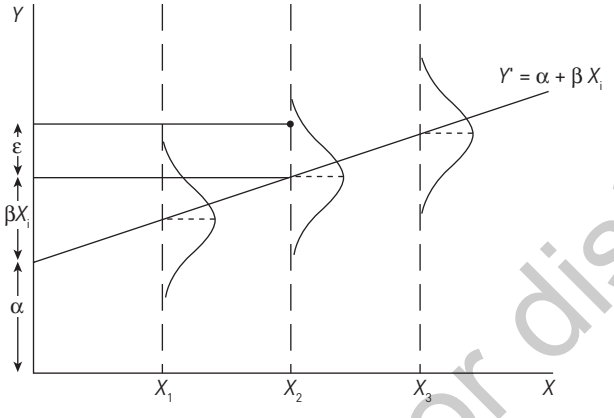
In Figure 11.1, we can locate every Y-value in any of the normal distributions from the following equation:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{11.1}$$

where Y_i is the Y-value we are seeking, α is the Y-intercept of the line on which the means of the normal distributions of the Y-values lie, β is the slope of that line, X_i is the X-value for the normal distribution in which we are seeking the Y-value, and ε_i is the distance the particular Y-value is above or below the line. These components are shown on the graph in Figure 11.2.

According to Equation 11.1, we locate the point in the normal distribution for any X_i by going up a distance α, then go up another distance βX_i. That takes you to the mean

Figure 11.2 ■ How Equation 11.1 locates any *Y*-value for any *X*-value using the linear regression model.



of the *Y*-values for X_i. Then go up or down the distance ε_i to get to Y_i. Every point (X_i, Y_i) can be defined that way.

The symbol ε_i in Equation 11.1 represents the distance each *Y*-value is from the mean of the normal distribution in which it resides. Because those means are on the line $Y' = \alpha + \beta X$, ε represents the distance each *Y* value is from the line, that is $\varepsilon = Y - Y'$. Clearly, if *Y* has a normal distribution for each *X*, then ε also has a normal distribution with $\mu_\varepsilon = 0$ and $\sigma_\varepsilon^2 = \sigma_Y^2$ for each *X* (homoscedasticity).

ESTIMATING PARAMETERS WITH THE LINEAR REGRESSION MODEL

We use the method of least squares to estimate the intercept and slope of the population regression line $Y = \alpha + \beta X$. *We start by taking a random sample from the population and finding the values of the estimators of* α *and* β *(*A *and* B *respectively) that minimize the function* $\Sigma[Y - (A + BX)]^2$.

The relationship between the two variables *X* and *Y* is described by the population line $Y' = \alpha + \beta X$. There are two parameters in this equation (α and β). To estimate these parameters, we take a random sample from the population and attempt to find the line that best fits the sample data. The equation for that line is $Y' = A + BX$, where *A* is the estimate of α and *B* is the estimate of β. Clearly, there are an infinite number of lines that can be drawn on top of the sample data. Some of these lines fit the sample data better

than others. What criterion should we use to choose among these lines? The criterion that has been adopted is the **least squares criterion**, which instructs us to *find the values for* A *and* B *that minimize the sum of the squared deviations of the data points from the line*. Put simply, we are attempting to identify the line that comes the closest to all of the data points collectively. Applying this criterion is the basis for the **method of least squares**.

The method of least squares works as follows:

1. Draw a random sample from the population.
2. Try to fit the equation $Y' = A + BX$ to the sample data by finding values for A and B that minimize the value of $\Sigma[Y - (A + BX)]^2$. Because $Y' = A + BX$, this expression can be rewritten as finding the values for A and B that minimize $\Sigma(Y - Y')^2$ (that is, finding the values that minimize the sum of the squared deviations between the sample values, Y, and the corresponding value on the sample regression line, Y').

The reader who is familiar with differential calculus can use that technique to find A and B: Find the first derivative of $\Sigma[Y - (A + BX)]^2$ with respect to A and again with respect to B. Set these terms equal to zero, thereby yielding two simultaneous equations with two unknowns. You can use algebra to find the values of A and B that satisfy those two simultaneous equations. The reader who is not familiar with differential calculus will have to take on faith that the following formulas give us the values of A and B that make $\Sigma[Y - (A + BX)]^2$ a minimum:

$$A = \overline{Y} - r\left(\frac{S_Y}{S_X}\right)\overline{X} \tag{11.2}$$

$$B = r\left(\frac{S_Y}{S_X}\right) \tag{11.3}$$

Substituting these expressions into the equation for the sample regression line ($Y' = A + BX$), we get the following formula:

$$Y' = \left[\overline{Y} - r\left(\frac{S_Y}{S_X}\right)\overline{X}\right] + \left[r\left(\frac{S_Y}{S_X}\right)X\right] \tag{11.4a}$$

Equation 11.4 is the **sample regression line**. It is based on five easily obtainable sample statistics: $\overline{Y}$, $\overline{X}$, S_Y, S_X, and r. When there is a non-zero correlation between X and Y, we can use Equation 11.4 to help us predict Y-scores from knowledge of X-scores.

REGRESSION AND PREDICTION

We can use the sample regression line to predict values on the Y*-axis for any value on the* X*-axis. We scale the accuracy of our prediction in terms of the variance in the observed scores around the predicted scores. The higher the value of the Pearson r, the lower the variance, and the closer our predicted scores are to the observed scores. We can also use the method of least squares to find a regression line for predicting the* X*-scores from the* Y*-scores. The phenomenon of regression toward the mean is a consequence of using the method of least squares to estimate the population regression line with the resulting sample regression line.*

Equation 11.4a is the sample regression line. When two variables are related, knowledge of an individual's score on one variable should help us to predict his or her score on the other variable; that is, after we find the sample regression line for one sample from a population, we can use that regression line to predict the *Y*s from the *X*s for other individuals sampled from that population. The predicted value will be the value of Y' from Equation 11.4a for a given value of *X*.

For the data from McManus et al. (2011), $\bar{X}_{\text{Quality}} = 6.13$, $\bar{Y}_{\text{MRAI-R}} = 6.52$, $\sigma_{\text{Quality}} = 1.78$, $\sigma_{\text{MRAI-R}} = 1.04$, and $r = .615$. Inserting these values into Equation 11.4a yields the following sample regression line:

$$Y' = 4.35 + 0.35X \tag{11.4b}$$

where Y' is the predicted score on the MRAI-R scale and X is the individual's score on the quality of contact scale. This sample regression line has 4.35 as the *Y*-intercept and 0.35 as the slope. These values are the least squares estimates for the parameters α and β of the population regression line. Therefore, if someone scores 7 on the quality of contact scale, we predict the score on the MRAI-R scale would be 6.79. The value of Y' is our best prediction for someone's score on the MRAI-R scale (indicating the positivity of their attitudes toward individuals with intellectual disabilities) given their score on the quality of contact (indicating the quality of their past experiences with individuals with intellectual disabilities). This prediction demonstrates that, consistent with the positive relationship between these variables, individuals' relatively high scores on quality of contact are associated with their having relatively high levels of positive attitudes. More importantly, this result is consistent with the researchers' hypothesis that higher levels of quality of contact with individuals with intellectual disabilities would be associated with more positive attitudes toward individuals with intellectual disabilities. This relationship suggests

that positive interactions with members of a stigmatized group may contribute to lower levels of prejudice toward them.

When we are talking about prediction in the case of regression, we are talking about it in the sense that when there is a correlation between *X* and *Y*, knowing something about a participant's score on *X*, while not necessarily allowing you to predict *Y* exactly, tells you something about the *Y*-score for a given individual: In the case of high positive correlation, higher *X*s tend to go with higher *Y*s, and lower *X*s tend to go with lower *Y*s. In the case of attitudes toward individuals with intellectual disabilities, a positive relationship such that higher levels of quality of contact are associated with more positive attitudes does not mean every participant with higher levels of quality of contact necessarily has more positive attitudes.

This situation raises a question. We know that our predicted scores for individuals are not going to be identical to their observed scores. We need some way to scale the accuracy of our predictions. We scale the accuracy of our predictions by looking at the *variation* of the observed scores from the predicted scores. Consider any given value of *X*. Equation 11.4a gives us a single predicted value for that *X*, but there are a large number of possible values associated with that *X*. The following formula gives us a measure of the accuracy of our predictions:

$$S^2_{Y|X} = \frac{\sum (Y - Y')^2}{n} \tag{11.5}$$[1]

When using Equation 11.5 as the measure of the accuracy of our predictions, we are *not* defining accuracy as the number or percentage of exact predictions we make. It is rare that we are going to predict the observed score exactly. Instead, Equation 11.5 measures accuracy of prediction in terms of the average of the sum of the squared deviations of the observed scores from the predicted scores. The closer the observed scores are to the predicted scores (and the regression line), the more accurate your predictions; the more spread out the observed scores are around the regression line, the poorer your predictions.

When $r = 1$, we would have perfect predictions because all of the observed scores would be on the regression line ($Y = Y'$); that is, all of the observed scores would equal the predicted scores. When this situation occurs, $S^2_{Y|X} = 0$. But when $r \neq 0$, then there

[1]The square root of Equation 11.5 is the *standard error of the estimate*.

will be some variation between the observed scores and predicted scores and $S^2_{Y|X} > 0$. Therefore, the smaller the value of $S^2_{Y|X}$, the better our predictions.

When $r = 0$, Equation 11.3 tells us that $B = 0$, Equation 11.2 tells us that $A = \overline{Y}$, and Equation 11.4 tells us that $Y' = \overline{Y}$. This means that when $r = 0$, our best prediction for the value of Y is $\overline{Y}$ because that value minimizes the $\Sigma\left(Y - Y'\right)^2$. Because our observed scores are scattered all over the place when $r = 0$, the sum of the squared deviations of the observed scores from the predicted scores is going to be the largest it can be for those data. Substituting $\overline{Y}$ for Y' in Equation 11.5 makes $S^2_{Y|X} = S^2_Y$.

The higher r is, the better our prediction. The better our prediction, the smaller the deviation of the observed scores from the predicted scores when you average them over all of the data. Therefore, there is an obvious relationship between $S^2_{Y|X}$ and r: As r increases, $S^2_{Y|X}$ decreases. Furthermore, when $r = 0, S^2_{Y|X} = S^2_Y$, and when $r = 1$, $S^2_{Y|X} = 0$. The exact form of the relationship between r and $S^2_{Y|X}$ is

$$S^2_{Y|X} = S^2_Y\left(1 - r^2\right) \tag{11.6}$$

Predicting in the Other Direction

When two variables are correlated, we should be able to make predictions in either direction (from X to Y and from Y to X). To make predictions from X to Y, we use the method of least squares to find the values of A and B in the equation $Y' = A + BX$ such that $\Sigma(Y - Y')$ is a minimum. If we want to make predictions from Y to X, the corresponding equation would be $X' = A^* + B^*Y$ (where A^* is the X-intercept and B^* is the slope), and we use the method of least squares to find the values of A^* and B^* such that $\Sigma(X - X')^2$ is a minimum. Here is the analytic solution:

$$A^* = \overline{X} - r\left(\frac{S_X}{S_Y}\right)\overline{Y} \tag{11.7}$$

and

$$B^* = r\left(\frac{S_X}{S_Y}\right) \tag{11.8}$$

Substituting these two terms into the equation for the sample regression line ($X' = A^* + B^*Y$), we get the following formula:

$$X' = \left[\overline{X} - r\left(\frac{S_X}{S_Y}\right)\overline{Y}\right] + \left[r\left(\frac{S_X}{S_Y}\right)Y\right] \tag{11.9}$$

BOX 11.1

REGRESSION TOWARD THE MEAN

An interesting thing happens when r is between 0 and ± 1: Suppose we take a value X_1 and use it to predict the value of Y_1'. Notice that the difference between X_1 and $\bar{X}$ and Y_1' and $\bar{Y}$ are not the same: $(Y_1' - \bar{Y})$ is smaller than $(X_1 - \bar{X})$. Now take Y_1' and go back in the other direction to predict X. When we do that, the predicted value for X (X_1') is closer to $\bar{X}$ than X_1 is; that is, $X_1' - \bar{X} < X_1 - \bar{X}$. Furthermore, $X_1' - \bar{X} < Y_1' - \bar{Y}$. In other words, *the predicted score is always closer to its mean than the score from which we started*, no matter which direction we go. This phenomenon is **regression toward the mean**.

Frances Galton discovered the phenomenon of regression toward the mean. He noticed that the average height of the children of tall parents was less than the average of the height of their parents and, for short parents, the average height of their children was greater than the average height of the parents. Furthermore, the average heights of the children of both tall and short parents were closer to the mean of all children than the average heights of the parents were to the average heights of all parents. There is a biological explanation for this phenomenon: Two things determine height. One is genetics (tall people tend to have tall children), and the other is nutrition. The nutritional factor accounts for the fact that in recent history, successive generations have tended to be taller than their parents.

Galton was showing the effects of heredity and the environment in the following way. A person who is quite tall is so for two possible reasons. One is the genetic component, and the other is a favorable nutrition program. What were the chances in the next generation that such a combination would occur again? It is likely that the genetics will be passed on. But when the favorable nutritional environment is not repeated in the next generation, the height of the children of tall parents will be lower. Likewise, the children of short parents tend to be taller because, although the genes for shortness are passed on, better nutrition in the second generation would cause the children to be a little taller than their parents. Thus, in both cases, the fact that the nutritional environment was likely to vary from generation to generation leads to differences in height between parents and children. Of course, not all short parents had taller children and not all tall parents had shorter children, but the average height of the children was shorter or taller than the average height of their parents. The fact that the environmental factors can change from generation to generation allows researchers to separate the effects of heredity and the environment.

The same phenomenon is observed with IQ. Parents with high IQs will not necessarily have children whose IQs are higher than average. The children's IQ scores could be higher than average, of course, but not necessarily. The environment is an important factor also. In a constant environment, you would observe purely genetic factors at work.

Implicit in the formula for Y' is this idea of regression to the mean. This pattern becomes obvious when we convert both X and Y to standard scores. The resulting equations are $z_Y' = rz_X$ and $z_X' = rz_Y$. Clearly, the slopes of those lines are always going to be some value less than 1. Thus, since the slope of the regression line is less than 1, when you predict Y from X, the predicted value will be closer to the mean (zero in the case of z-scores). See Senn (2011) for more details about Galton's discovery of regression to the mean and its implications.

These two lines, X' and Y', are always different *except* when $r = +1$ or -1. On the other hand, when $r = 0$, $Y' = \overline{Y}$ and $X' = \overline{X}$. Therefore, when $r = 0$, the lines are perpendicular to each other and cross at the point $(\overline{X}, \overline{Y})$. When r is between 0 and ± 1, the two lines cross at the point $(\overline{X}, \overline{Y})$ with the angle between them less than 90°. The greater the absolute value of r, the smaller the angle between these two regression lines. When $r = \pm 1$, they are the same line. An interesting thing happens when r is between 0 and ± 1 (see Box 11.1).

VARIANCE AND CORRELATION

We can partition the variance of the Y*-scores into two parts: the variance of the predicted scores around the mean of* Y *(variance due to regression or explained variance) and the variance of the observed scores around the regression line (residual variance or unexplained variance). The proportions of the variance in* Y *that are due to regression and the residual variance are both functions of the Pearson* r *for a given set of data.*

In our sample data, the deviation of every Y-value from $\overline{Y}$ $(Y - \overline{Y})$ is the sum of the deviation of that Y-value from the predicted value of $Y (Y - Y')$ and the deviation of each predicted value Y' from $\overline{Y}$ $(Y' - \overline{Y})$. That is,

$$\left(Y - \overline{Y}\right) = \left(Y - Y'\right) + \left(Y' - \overline{Y}\right) \tag{11.10}$$

This concept is illustrated in Figure 11.3.

If we square both sides of Equation 11.10, add up all of the squared deviations on both sides, and divide by the sample size (n), we get the following:

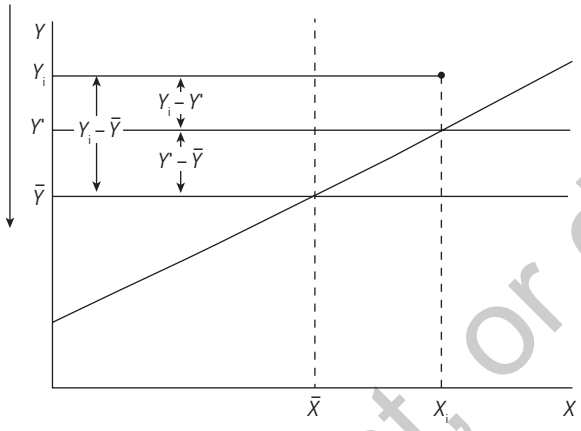
$$\sum \frac{(Y - \overline{Y})^2}{n} = \sum \frac{[(Y - Y') + (Y' - \overline{Y})]^2}{n} \tag{11.11}$$

If we complete the square on the right side of Equation 11.11 and distribute the summation sign and the division by n, we get the following equation:

$$\sum \frac{(Y - \overline{Y})^2}{n} = \sum \frac{(Y - Y')^2}{n} + \sum \frac{(Y' - \overline{Y})^2}{n} - \frac{}{n} + 2\sum \frac{(Y - Y')(Y' - \overline{Y})}{n} \tag{11.12}$$

The term to the left of the equal sign is the variance of the Y-scores. We call this the **total variance** of Y (S_Y^2). The first term after the equal sign is the variance of the Y-scores

FIGURE 11.3 ■ The deviation of every *Y*-score from $\bar{Y}$ ($Y - \bar{Y}$) is the sum of the deviation of that *Y*-value from the predicted value of *Y* ($Y - Y'$) and the deviation of each predicted value Y' from $\bar{Y}$ ($Y' - \bar{Y}$).



around the regression line. This term is referred to as either the **unexplained variance** or the **residual variance**. The symbol for this variance is $S^2_{Y|X}$. The second term after the equal sign is the variance of the scores on the line (predicted scores) around the mean of *Y*. This term is called the **explained variance**. The symbol for this variance is $S^2_{Y'}$. Substituting these symbols into Equation 11.12, we get the following:

$$S^2_Y = S^2_{Y|X} + S^2_{Y'} + 2\sum \frac{(Y - Y')(Y' - \bar{Y})}{n} \tag{11.13}$$

The last term in Equations 11.12 and 11.13 is the covariance between the deviations of the *Y* scores around the regression line ($Y - Y'$) and the deviations of the predicted scores around the mean of *Y* ($Y' - \bar{Y}$). As noted in the previous chapter, the covariance is related to the correlation between two variables, and the only time the correlation is zero is when there is no relationship between the two variables (correlation = 0). In this case, the covariance term in Equations 11.12 and 11.13 equals zero. In this case, there is no relationship between the deviations of the observed score from the line ($Y - Y'$) and the deviations of the points on the line from the mean of *Y* ($Y' - \bar{Y}$). If they were positively correlated, then where the deviation between Y' and $\bar{Y}$ is small, the deviation *Y* around the line (Y') would also be small, and where ($Y' - \bar{Y}$) is large, ($Y - Y'$) would also be large. In this case, we could not have equal variance around the line (homoscedasticity), because homoscedasticity means that the variance around the

regression line is the same everywhere along that line. This situation can only occur when the correlation (covariance) between $(Y - Y')$ and $(Y' - \overline{Y}) = 0$. Therefore,

$$S_Y^2 = S_{Y|X}^2 + S_{Y'}^2 \tag{11.14}$$

Equation 11.14 tells us that the *total variance* in Y (S_Y^2) is made up of two parts: One part is the *explained variance* ($S_{Y'}^2$), which is variation due to the correlation between X and Y; that is, this is the variance in Y that is due to or explained by variation in X (this component is also referred to as *variation due to regression*). The other part is the *unexplained or residual variance* ($S_{Y|X}^2$) which is the variation in Y that is not explained by variation in X.

We saw in Equation 11.6 that the value of $S_{Y|X}^2$ depends on the correlation between X and Y. When $r = 1$, all of the Y-scores are on the regression line, and there is no variation around that line; that is, there is no residual variance. In other words, when $r = 1$, all of the variation in Y is explained by the correlation with X, nothing is unexplained, and $S_{Y|X}^2 = 0$. On the other hand, when $r = 0$, $Y' = \overline{Y}$. In this case, no matter what the value of X, the predicted score is the same; in other words, there is no variation in the Y' values. Therefore, *all* of the variation in Y is unexplained variation or residual variance and $S_{Y|X}^2 = S_Y^2$.

Substituting Equation 11.6 into Equation 11.14 and rearranging the terms gives us the following equation:

$$S_{Y'}^2 = r^2 S_Y^2 \tag{11.15}$$

Equation 11.15 tells us that the explained variance is also a function of the correlation between X and Y (due to regression): When $r = 1$, $S_{Y'}^2 = S_Y^2$, and when $r = 0$, $S_{Y'}^2 = 0$.

We can rearrange Equation 11.15 to get the following important relationship:

$$r^2 = \frac{S_{Y'}^2}{S_Y^2} \tag{11.16}$$

Equation 11.16 tells us that the *square of the Pearson* r *equals the proportion of variance in* Y *explained by the correlation with* X. In other words, the higher the correlation, the greater the proportion of the total variance in Y that is explained by variation in X.

We can perform a similar arrangement of Equation 11.16 to get the following relationship:

$$r^2 = 1 - \frac{S_{Y|X}^2}{S_Y^2} \tag{11.17}$$

Equation 11.17 tells us that the square of the Pearson r equals 1 minus the proportion of variance that is not explained by the correlation with X (the residual variance).

TESTING HYPOTHESES WITH THE LINEAR REGRESSION MODEL

We use the ratio of two estimates of variance, one based on the variation among the predicted Y*-scores and the other based on the variation of the* Y*-scores around the regression line. When* $\rho = 0$*, both are estimating the variance of* Y*. When* $\rho \neq 0$*, they are estimating different things. The test statistic is an* F*-ratio. This concept is the basis of analysis of variance.*

We can test the hypothesis that $\rho = 0$ by comparing two estimates of the variance. One estimate is based on the deviations of the predicted scores of Y from $\overline{Y}$ in your sample, and the other is based on the deviations of the observed scores from the predicted scores in your sample. The test statistic is the ratio of those two estimates of variance. When the null hypothesis is true, both of these estimates are estimating the variance of Y for the population (σ_Y^2). When $\rho \neq 0$, one of those estimates of variance is estimating something greater than σ_Y^2, and the other is estimating something less than σ_Y^2.

In the linear regression model, all of the normal distributions of the Y values for each X value lie on a straight line $Y' = \alpha + \beta X$. Equations 11.2 and 11.3 give us the values of A and B (the least square estimates of α and β). It is also the case that

$$\alpha = \mu_Y - \rho\left(\frac{\sigma_Y}{\sigma_X}\right)\mu_X \text{ and } \beta = \rho\left(\frac{\sigma_Y}{\sigma_X}\right) \tag{11.18}$$

Therefore, $\beta = 0$ when $\rho = 0$. In this case, testing the hypothesis H_0: $\rho = 0$ is equivalent to testing the hypothesis H_0: $\beta = 0$.

We can construct a test for the hypothesis H_0: $\rho = 0$ from the following facts:

1. The total sum of the squared deviations of all of the Y-scores around the mean of Y can be divided into two additive parts: the sum of the squared deviations for each Y-score around the predicted value for that Y-value (Y') and the sum of the squared deviations of these predicted values (Y') from the mean of all of the Y-scores:

$$\sum\left(Y - \overline{Y}\right)^2 = \sum\left(Y - Y'\right)^2 + \sum\left(Y' - \overline{Y}\right)^2 \tag{11.19a}$$

   The sums of squared deviations to the left of the equal sign is called SS_{Total}, the first sum of squared deviations to the right of the equal sign is called the $SS_{\text{Unexplained}}$ or SS_{Residual}, and the second sum of squared deviations to the right of the equal sign is called the $SS_{\text{Explained}}$. Therefore,

$$SS_{\text{Total}} = SS_{\text{Unexplained}} + SS_{\text{Explained}} \tag{11.19b}$$

2. The total degrees of freedom equal sample size minus 1: $df_{\text{Total}} = n - 1$. We can use both of the definitions for degrees of freedom developed in an earlier chapter to arrive at this value.[2] The total degrees of freedom can be divided into two additive parts:
   a. *Degrees of freedom explained* ($\text{df}_{Explained}$) *or degrees of freedom regression* ($\text{df}_{Regression}$). The sums of squares explained involves the deviations' predicted values around the mean of *Y*. Those predicted values are all on a straight line. We know from Euclid's axioms in geometry that one and only one straight line can be drawn between two points. The regression line Y' goes through the point $(\bar{X}, \bar{Y})$. Therefore, using the first definition of degrees of freedom (how many of those predicted scores we have to know before we can know all the rest), if we know one other point on the line, we know the location of all of the points on the line. Therefore, $df_{\text{Explained}} = 1$.
   b. *Degrees of freedom unexplained* ($\text{df}_{Unexplained}$) *or degrees of freedom residual* ($\text{df}_{Residual}$). We use the second definition of degrees of freedom to find the unexplained degrees of freedom, which are related to the deviations of observed scores around the regression line: When we have only one point, we can draw an infinite number of regression lines through that point, and the sum of the squared deviations around the regression line is zero because the regression line has to go through that point based on the method of least squares. When we have two points, we minimize the sum of the squared deviations around the line by drawing the regression line through those two points. Here $r = 1$, and the sum of the squared deviations is zero. Therefore, we must have *at least three values* before the sum of squared deviations of the scores around the regression line can be a non-zero value. Therefore, $df_{\text{Unexplained}} = (n - 2)$.
   We can summarize this relationship with the following equation:

$$(n - 1) = (n - 2) + 1 \tag{11.20a}$$

$$df_{\text{Total}} = df_{\text{Unexplained}} + df_{\text{Explained}} \tag{11.20b}$$

3. Dividing sums of squares by degrees of freedom produces estimates of variance, and we can use the definition of unbiased estimates to determine what variances are being estimated:
   a. $\dfrac{\sum\left(Y - \bar{Y}\right)^2}{n-1}$ is an unbiased estimate of σ_Y^2. Therefore,

$$E\left[\frac{\sum\left(Y - \bar{Y}\right)^2}{n-1}\right] = \sigma_Y^2 \tag{11.21}$$

[2]The two definitions of *degrees of freedom* are (1) how many scores you need to know before all the rest are fixed and (2) how many scores you need to have before the sum of squares could be a non-zero value.

BOX 11.2

THE EXPECTED VALUE OF $\frac{\sum(Y-Y')^2}{n-2}$

$\frac{\sum(Y-Y')^2}{n-2}$ is an unbiased estimate of the unexplained (residual) variance in the population $\sigma^2_{Y|X}$. We saw earlier that the unexplained (residual) variance of the sample, $S^2_{Y|X}$, is a function of the value of the Pearson r: $S^2_{Y|X} = S^2_Y(1-r^2)$. The value of $\sigma^2_{Y|X}$ is a function of the population correlation ρ in a similar way; that is, $\sigma^2_{Y|X} = \sigma^2_Y(1-\rho^2)$. Combining these two pieces of information we arrive at the following equation:

$$E\,\frac{\sum(Y-Y')^2}{n-2} = \sigma^2_{Y|X} = \sigma^2_Y(1-\rho^2)$$

When $\rho = 0$, $E\,\frac{\sum(Y-Y')^2}{n-2} = \sigma^2_Y$; that is, $E\,\frac{\sum(Y-Y')^2}{n-2}$ is estimating σ^2_Y.

When $\rho \neq 0$, $E\,\frac{\sum(Y-Y')^2}{n-2}$ is estimating something *smaller* than σ^2_Y.

b. When $\rho = 0$,

$$E\left[\frac{\sum(Y-Y')^2}{n-2}\right] = \sigma^2_Y \tag{11.22}$$

Therefore, when $\rho = 0$, this formula is estimating σ^2_Y. When $\rho \neq 0$, this formula is estimating something *smaller* than σ^2_Y. (See Box 11.2.)

c. When $\rho = 0$,

$$E\left[\frac{\sum(Y'-\overline{Y})^2}{1}\right] = \sigma^2_Y \tag{11.23}$$

Therefore, when $\rho = 0$, this formula is estimating σ^2_Y. When $\rho \neq 0$, this formula is estimating something *larger* than σ^2_Y. (See Box 11.3.)

BOX 11.3

THE EXPECTED VALUE OF $\sum \frac{(Y'-\bar{Y})}{1}$

We can derive the expected value of $\frac{\sum(Y-\bar{Y})^2}{1}$ by starting with Equation 11.19a:

$$\sum(Y-\bar{Y})^2 = \sum(Y-Y')^2 + \sum(Y'-\bar{Y})^2$$

We can rearrange the terms to get

$$\sum(Y'-\bar{Y})^2 = \sum(Y-\bar{Y})^2 - \sum(Y-Y')^2$$

If we take the expected value of both sides and use the fact that the expected value of a difference is the difference in the expected values, we arrive at the following relationship

$$E\left[\sum(Y'-\bar{Y})^2\right] = \mathrm{E}\left[\sum(Y-\bar{Y})^2\right] - \mathrm{E}\left[\sum(Y-Y')^2\right] \quad \text{(A)}$$

We saw earlier that $E\left[\frac{\sum(Y-\bar{Y})^2}{n-1}\right] = \sigma_Y^2$. Applying the rules of algebra and expected values, we get $E\left[\sum(Y-\bar{Y})^2\right] = (n-1)\sigma_Y^2$.

It is also the case that because

$$E\left[\frac{\sum(Y-Y')^2}{n-2}\right] = \sigma_Y^2(1-\rho^2),$$

$$E\left[\sum(Y-Y')^2\right] = (n-2)\sigma_Y^2(1-\rho^2)$$

Substituting these into Equation A above, we get this:

$$E\left[\sum(Y-Y')^2\right] = (n-1)\sigma_Y^2 - (n-2)\sigma_Y^2(1-\rho^2) \quad \text{(B)}$$

By applying the rules of algebra, we obtain the following:

$$E\left[\sum(Y'-\bar{Y})^2\right] = \sigma_Y^2\left[1+(n-2)\rho^2\right] \quad \text{(C)}$$

Thus, when $\rho = 0$, $E\left[\frac{\sum(Y'-\bar{Y})^2}{1}\right] = \sigma_Y^2$; that is, $E\left[\frac{\sum(Y'-\bar{Y})^2}{1}\right]$ is estimating σ_Y^2, and when $\rho \neq 0$, $E\left[\frac{\sum(Y'-\bar{Y})^2}{1}\right]$ is estimating something *larger* than σ_Y^2.

We can use the results in Boxes 11.2 and 11.3 to construct a hypothesis test for $\rho = 0$ by comparing the two estimates of variance $\frac{\sum(Y-Y')^2}{n-2}$ and $\frac{\sum(Y'-\bar{Y})^2}{1}$. When $\rho = 0$, we expect

$$\frac{\sum(Y-Y')^2}{n-2} \approx \frac{\sum(Y'-\bar{Y})^2}{1} \tag{11.24}$$

And when $\rho \neq 0$, we expect

$$\frac{\sum(Y-Y')^2}{n-2} << \frac{\sum(Y'-\bar{Y})^2}{1} \tag{11.25}$$

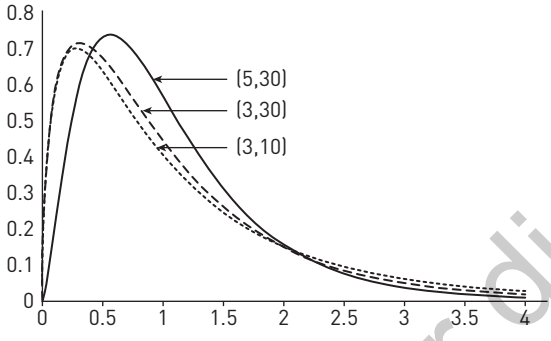
We can compare these two estimates of variance by forming their ratio. The symbol for this ratio is F (after R. A. Fisher, who derived its distribution). When the null hypothesis is true, the value of the F-ratio should be close to 1. If we put the value we expect to be larger in the numerator when the null hypothesis is false, the value of F will be a lot larger than 1 when the null hypothesis is false. When the null hypothesis is true, the F-ratio has an F-distribution. The exact shape of the central F-distribution depends on both the degrees of freedom for the estimate in the numerator and degrees of freedom for the estimate in the denominator. Because F is the ratio of two estimates of variance, which must be positive given that they are squared values, $F \geq 0$. When the null hypothesis is true, $\mu_F = \frac{df_{\text{denominator}}}{df_{\text{denominator}} - 2}$. The variance of F is a complicated formula that depends on both $df_{\text{numerator}}$ and $df_{\text{denominator}}$.[3] As either degrees of freedom increase, σ_F^2 decreases, and the distribution of F becomes narrower (see Figure 11.4).

We reject the null hypothesis when F is a large number. Therefore, the critical region for an F-test is always in the right tail of the null hypothesis true distribution (central F-distribution). When the null hypothesis is false, the F-ratio has a non-central F-distribution with

$$\mu_{F^*} = \left(\frac{df_{\text{den}}}{df_{\text{den}} - 2}\right)\left(\frac{df_{\text{num}} + \delta}{df_{\text{num}}}\right), \tag{11.26}$$

[3] $\sigma_F^2 = \frac{2\,df_{\text{den}}^2\,(df_{\text{num}} + df_{\text{den}} - 2)}{df_{\text{num}}\,(df_{\text{den}} - 2)^2\,(df_{\text{den}} - 4)}$ when $df_{\text{denominator}} > 4$.

FIGURE 11.4 ■ Probability density functions for *F* for various degrees of freedom. In the parentheses, the first numbers are $df_{numerator}$, and the second numbers are $df_{denominator}$.



where δ is the noncentrality parameter. For this situation,

$$\delta = n\left[\frac{\rho^2}{(1-\rho)^2}\right] \tag{11.27}$$

where ρ is the true population correlation.

When H_0 is true, $\rho = 0$ and $\delta = 0$, and μ_{F^*} reduces to μ_F.

Therefore, to test H_0: $\rho = 0$ when the relationship between the variables is linear, the *F*-ratio is as follows:

$$F = \frac{\sum\left(Y' - \bar{Y}\right)^2 / 1}{\sum\left(Y - Y'\right)^2 / (n-2)} \tag{11.28}$$

As noted in Equation 11.24, when the null hypothesis is true, the numerator and denominator are both estimating σ_Y^2, but when the null hypothesis is false (Equation 11.25), the numerator is estimating a value *larger* than σ_Y^2, and the denominator is estimating a value *smaller* than σ_Y^2. Therefore, large values of *F* lead to rejection of the null hypothesis.

Presenting the Results of an *F*-Test

When people perform *F*-tests, they may display the results in a standard format called an ***F*-table** (see Table 11.1). The *F*-table starts with the sums of squares and degrees of freedom for each component of the *F*-ratio, combines them into estimates of variance (called *mean squares*), and finally forms the *F*-ratio from the ratio of the mean squares.

The column labeled "Source of Variance" contains the names of the sources of variance that have been identified as adding up to the total variance in the situation under study.

TABLE 11.1 ■ *F*-Table

Source of Variance	Sums of Squares	Degrees of Freedom	Mean Square	*F*
Source 1	SS_1	df_1	$MS_1 = SS_1/df_1$	$F = MS_1/MS_2$
Source 2	SS_2	df_2	$MS_2 = SS_2/df_2$	
Total	SS_{Total}	df_{Total}		

How many sources of variance there are depends on the situation. The sums of squares and degrees of freedom in the second and third columns must add to the total sums of squares and degrees of freedom, respectively. Mean squares are estimates of variance for each source of variance in the table. They are the obtained by dividing the sums of squares by the corresponding degrees of freedom. Finally, the value of F is the ratio of two mean squares (estimates of variance).

In the test of the hypothesis that $\rho = 0$ for the linear regression situation, the *F*-table would look like Table 11.2.

The sums of squares for regression and residual in Table 11.2 are not easy to calculate. Equivalent formulas that are easier to use are presented in Box 11.4.

The results from Box 11.4 lead to Table 11.3.

By dividing the numerator and denominator of the *F*-ratio in the Table 11.3 by SS_{Total}, the test statistic can be simplified to the following:

$$F = \frac{r^2/1}{\left(1-r^2\right)/(n-2)} \tag{11.29}$$

TABLE 11.2 ■ *F*-Table for the Hypothesis That $\rho = 0$

Source of Variance	Sums of Squares	Degrees of Freedom	Mean Square	*F*
Regression	$\sum(Y'-\bar{Y})^2$	1	$\sum(Y'-\bar{Y})^2/1$	$\frac{\sum(Y'-\bar{Y})^2/1}{\sum(Y-Y')^2/(n-2)}$
Residual	$\sum(Y-Y')^2$	$n-2$	$\sum(Y-Y')^2/(n-2)$	
Total	$\sum(Y-\bar{Y})^2$	$n-1$		

BOX 11.4

COMPUTATIONAL FORMULAS FOR THE SUMS OF SQUARES FOR THE *F*-TEST OF H_0: $\rho = 0$

The sums of squares regression and residual in the table can be calculated from the Pearson *r* and the sums of squares total (SS_{Total}). The derivation of these formulas for the hypothesis test under consideration can be generalized to a wide variety of situations.

Equation 11.15 tells us that the explained variance (variance due to regression [$S^2_{Y'}$]) is equal to r^2 times the total variance in *Y*; that is,

$$S^2_{Y'} = r^2 S^2_Y$$

where S^2_Y is the total variance in *Y* (S^2_{Total}).

Because variance equals sums of squares divided by sample size (*n*), the above equation can be written as follows:

$$\frac{SS_{Regression}}{n} = r^2\left(\frac{SS_{Total}}{n}\right)$$

Multiplying both sides of this equation by *n* leaves

$$SS_{Regression} = r^2 \bullet SS_{Total}$$

where $SS_{Total} = \Sigma\left(Y - \overline{Y}\right)^2$.

We can derive a similar formula for $SS_{Residual}$. Equation 11.17 can be rewritten as

$$S^2_{Y|X} = (1 - r^2)S^2_Y$$

Applying the same logic as above,

$$\frac{SS_{Residual}}{n} = \left(1 - r^2\right)\left(\frac{SS_{Total}}{n}\right)$$

and

$$SS_{Residual} = \left(1 - r^2\right) \bullet SS_{Total}$$

TABLE 11.3 ■ *F*-Table for the Hypothesis That ρ = 0

Source of Variance	Sums of Squares	Degrees of Freedom	Mean Square	*F*
Regression	$r^2 SS_{Total}$	1	$r^2 SS_{Total} / 1$	$\frac{r^2 SS_{Total}/1}{\left(1-r^2\right)SS_{Total}/(n-2)}$
Residual	$(1 - r^2)SS_{Total}$	$(n - 2)$	$(1 - r^2)SS_{Total} / (n - 2)$	
Total	SS_{Total}	$(n - 1)$		

Therefore, we may now perform our hypothesis test using only r and n.

When the degrees of freedom numerator = 1 in an F-ratio, the square root of $F = t$, with degrees of freedom = $df_{\text{denominator}}$. Therefore, one can also use the following t-statistic to test H_0: $\rho = 0$:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{11.30}$$

When the null hypothesis is true, Equation 11.30 has a t-distribution with $(n - 2)$ degrees of freedom. When the null hypothesis is false, Equation 11.30 has a non-central t-distribution. We can use the t-test in Equation 11.30 to conduct one-tailed tests where the alternative hypothesis is $\rho > 0$ or $\rho < 0$. The F-test only allows for two-tailed tests because it uses r^2. That is, while the F-distribution only has one tail, it does not allow for the testing of directional tests that are typically referred to as "one-tailed."

Summary

The linear regression model starts with the population regression line $Y = \alpha + \beta X + \varepsilon$, which describes the relationship between the two variables X and Y (where X is the predictor and Y is the criterion) in the population. After drawing a random sample from that population, we can use the method of least squares to estimate the parameters α and β. The result is a line $Y' = \alpha + \beta X$, which minimizes the sum of the squared deviations between the individual data points in the sample (Y) and the corresponding points on the line (Y'). After the sample regression line (Y') is located, the total variance in the Y-values (S_Y^2) can be partitioned into two parts, the explained variance and the unexplained or residual variance. The explained variance is the variance of the points on the sample regression line (Y') around the mean of the Y-values in the sample ($\overline{Y}$). It is called *explained* because it is the variance in Y that is due to variance in X. The higher the value of the sample correlation (Pearson r), the greater the explained variance. The symbol for the explained variance is $S_{Y'}^2$. The unexplained, or residual, variance is the variance of the individual data points Y around Y'. It is that part of the variance in Y that is *not* due to the correlation, and the higher the value of the Pearson r, the less the residual variance. The symbol for the residual variance is $S_{Y|X}^2$.

The value of the square of the Pearson r can be calculated from the following two formulas:

$$r^2 = \frac{S_{Y'}^2}{S_Y^2}$$

and

$$r^2 = 1 - \frac{S_{Y|X}^2}{S_Y^2}$$

We can use r^2 to test the null hypothesis that the population correlation = 0 with the following test statistic:

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

The ideas summarized here can be generalized to situations in which there are more than one predictor (multiple regression) and in which the relationship between two variables is not linear. These ideas also provide a link between correlation and hypothesis tests on means, a connection that is the basis for measures of effect sizes in t-tests and analysis of variance. Finally, these ideas lay the foundation for the analysis of variance. We will show how these ideas apply to nonlinear situations and to situations in which there is more than one predictor (multiple regression) in the next chapters.

Conceptual Exercises

1. Describe how the method of least squares can be used to estimate parameters in the linear regression model. Indicate what parameters are being estimated. (You do *not* have to describe all the assumptions of this model to answer the question.)
2. Why does $df_{\text{residual}} = n - 2$ when testing the hypothesis that the slope of the regression line = 0?
3. In the *F*-test of the hypothesis H_0: $\rho = 0$ with the linear regression model, the numerator is:

$$\frac{\sum\left(Y' - \bar{Y}\right)^2}{1}$$

   What does this term represent (or estimate) when H_0 is true *and* when H_0 is false?
4. Why is the test statistic for the hypothesis H_0: $\rho = 0$ of the following form?

$$\frac{\frac{\sum\left(Y' - \bar{Y}\right)^2}{1} / 1}{\sum\left(Y - Y'\right)^2 / (n-2)}$$

(Hint: Think in terms of expected values.)

5. Why does $\frac{\sum\left(Y' - \bar{Y}\right)^2}{1}$ not necessarily equal 0 when $\rho = 0$?
6. We encounter two similar terms when discussing the linear regression model:

$$\frac{\sum\left(Y' - \bar{Y}\right)^2}{n} \text{ and } \frac{\sum\left(Y' - \bar{Y}\right)^2}{1}$$

   What do these terms represent?
7. To test the hypothesis H_0: $\rho = 0$ against the alternative hypothesis H_1: $\rho \neq 0$ using the linear regression model, we could use either of the following test statistics:

$$F = \frac{\sum\left(Y' - \bar{Y}\right) / 1}{\sum\left(Y' - \bar{Y}\right)^2 / (n-2)}$$

$$F = \frac{r^2 / 1}{\left(1 - r^2\right) / (n-2)}$$

   a. Describe the distributions of these test statistics when H_0 is true *and* when H_0 is false (their means, degrees of freedom, and forms). Also draw a diagram.
   b. Why does the power of this test increase as sample size increases?
   c. Describe how you could create a power function for this test statistic. Also, make a diagram and label the axes. (Your answer should indicate your understanding of the task here.)
   d. In the case of the *F*-statistic on the top, what can you say about the terms in the numerator and denominator when H_0 is true *and* when H_0 is false. That is, what are the terms estimating?
8. What do the sample statistics $s^2_{Y|X}$ and $s^2_{Y'}$ represent? (They are variances of what around what?) What are their values when $r = 0$ and $r = 1$? Why?

Student Study Site

Visit the Student Study Site at **https://study.sagepub.com/friemanstats** for a variety of useful tools including data sets, additional exercises, and web resources.