CHAPTER 16

Bivariate, Multiple, and Logistic Regression

In Chapter 3, we considered how bivariate correlation can be used to describe the relationship between two variables. Then, in Chapter 9, we looked at how bivariate correlations are dealt with in an inferential manner. In this chapter, our focus is on a topic closely related to correlation. This topic is called *regression*.

Three different kinds of regression are considered here: **bivariate regression, multiple regression,** and **logistic regression.** Bivariate regression is similar to bivariate correlation, because both are designed for situations in which there are just two variables. Multiple and logistic regression, however, were created for cases in which there are three or more variables. Although many other kinds of regression procedures have been developed, the three considered here are by far the ones used most frequently by applied researchers.

The three regression procedures considered in this chapter are like correlation in that they are concerned with relationships among variables. Because of this, you may be tempted to think that regression is simply another way of talking about, or measuring, correlation. Resist that temptation, because these two statistical procedures differ in three important respects: their purpose, the way variables are labeled, and the kinds of inferential tests applied to the data.

The first difference between correlation and regression concerns the purpose of each technique. As indicated in Chapter 3, bivariate correlation is designed to illuminate the relationship, or connection, between two variables. The computed correlation coefficient may suggest that the relationship being focused on is direct and strong, or indirect and moderate, or so weak that it would be unfair to think of the relationship as being either direct or indirect. Regardless of how things turn out, each of the two variables is equally responsible for the nature and strength of the link between the two variables.

Whereas correlation concentrates on the relationship, or link, that exists *between* variables, regression focuses on the variable(s) that exist on one or the other *ends* of the link. Depending on which end is focused on, regression tries to accomplish one or the other of two goals: **prediction** or **explanation.**

In some studies, regression is utilized to predict scores on one variable based on information regarding the other variable(s). For example, a college might use regression in an effort to predict how well applicants will handle its academic curriculum. Each applicant's college grade point average (GPA) might be the main focus of the regression, with predictions made on the basis of available data on other variables (e.g., an entrance exam, the applicant's essay, and the recommendations written by high school teachers). If used in this manner, regression's focus would be on the one variable toward which predictions are made: college GPA.

In other investigations, regression is used in an effort to explain why the study's people, animals, or things score differently on a particular variable of interest. For example, a researcher might be interested in why people differ in the degree to which they seem satisfied with life. If such a study were to be conducted, a questionnaire might be administered to a large group of individuals for the purpose of measuring life satisfaction. Those same individuals would also be measured on several other variables that might explain why some people are quite content with what life has thrown at them whereas others seem to grumble incessantly because they think life has been cruel and unfair to them. Such variables might include health status, relationships with others, and job enjoyment. If used in this manner, regression's focus would be on the variables that potentially explain why people differ in their levels of life satisfaction.

Excerpts 16.1 and 16.2 illustrate the two different purposes of regression. In the first of these excerpts, the clear objective was to use regression analyses to help predict people's adjustment to living in a nursing home. In Excerpt 16.2, the goal

EXCERPTS 16.1–16.2 • *The Two Purposes of Regression: Prediction and Explanation*

Relocation to a nursing home is regarded as one of the most stressful events a person can experience. . . . The purpose of this study was to identify predictors of nursing home adjustment for elderly residents using a direct measure of nursing home adjustment. . . . Descriptive analysis was used and multiple linear regression was performed to identify the predictors of adjustment for nursing home residents.

Source: Lee, G. E. (2010). Predictors of adjustment to nursing home life of elderly residents: A cross-sectional survey. *International Journal of Nursing Studies, 47*(6), 1–8.

Need for recovery (NFR) after work is an indicator for work-related fatigue. . . . This study aims to establish the prevalence of high work-related fatigue [and] explain group differences categorized by gender, age, and education. The study

(*continued*)

EXCERPTS 16.1–16.2 • (*continued*)

particularly aims to clarify prevalence and explanatory factors in highly educated women. [Our regression] analyses give an indication of the factors that may explain the difference in the prevalence of high NFR between the compared groups, and of the degree to which the combination of [several] demographic, health, and work-related factors can explain the difference in the prevalence of high NFR.

Source: Verdonk, P., Hooftman, W. E., van Veldhoven, M. J. P. M., Boelens, L. R. M., & Koppes, L. L. J. (2010). Work-related fatigue: The specific case of highly educated women in the Netherlands. *International Archives of Occupational and Environmental Health, 83*(3), 309–321.

was explanation, not prediction. Here, the researchers wanted to know which factors explain why some people are afflicted by work-related fatigue whereas others are not. The focus was primarily on women workers, a trait called *need for recovery* after a day's work, and demographic variables.

The second difference between regression and correlation concerns the labels attached to the variables. This difference can be seen most easily in the case in which data on just two variables have been collected. Let's call these variables A and B. In a correlation analysis, variables A and B have no special names; they are simply the study's two variables. With no distinction made between them, their location in verbal descriptions or in pictorial representations can be switched without changing what is being focused on. For example, once r becomes available, it can be described as the correlation between A and B *or* it can be referred to as the correlation between B and A. Likewise, if a scatter diagram is used to show the relationship between the two variables, it does not matter which variable is positioned on the abscissa.

In a regression analysis involving A and B, an important distinction between the two variables must be made. In regression, one of the two variables needs to be identified as the **dependent variable** and the other variable must be seen as the **independent variable.**[1] This distinction is important because (1) the scatter diagram in bivariate regression always is set up such that the vertical axis corresponds with the dependent variable whereas the horizontal axis represents the independent variable, and (2) the names of the two variables cannot be interchanged in verbal descriptions of the regression. For example, the regression of A on B is not the same as the regression of B on A.[2]

Excerpts 16.3 and 16.4 come from two studies that were quite different. In the first study, only two variables were involved in the single regression that was

[1]The terms **criterion variable, outcome variable,** and **response variable** are synonymous with the term **dependent variable,** whereas the terms **predictor variable** or **explanatory variable** mean the same thing as **independent variable.**

[2]When the phrase "regression of _____ on _____" is used, the variable appearing in the first blank is the

Stress reactions to uncertainty were measured [via] the Physicians' Reactions to Uncertainty Scale (PRUS). . . . Epistemology was measured using the Physicians' Belief Scale (PBS). . . . Our primary hypothesis was tested [by means of] a simple bivariate regression with PRUS scores as the dependent variable and PBS scores as the independent variable.

Source: Evans, L., & Trotter, D. R. M. (2009). Epistemology and uncertainty in primary care: An exploratory study. *Family Medicine, 41*(5), 319–326.

To assess the extent to which students' perceptions of parenting style predicted evaluations of their parents' preferred music we performed a multiple regression analysis using the mean rating score of parent music (ratings given to only those pieces indicated as a favorite by a student's parent) as the dependent variable and scores on the caring and the autonomy dimensions of the PBI, as well as students' age and gender, as independent variables.

Source: Serbun, S. J., & DeBono, K. G. (2010). On appreciating the music of our parents: The role of the parent–child bond. *North American Journal of Psychology, 12*(1), 93–102.

conducted. In the second excerpt, there was one dependent variable and four independent variables. Despite these differences, notice how the researchers associated with each excerpt clearly designate the status of each variable as being a dependent variable or an independent variable.

The third difference between correlation and regression concerns the focus of inferential tests and confidence intervals. With correlation, there is just one thing that can be focused on: the sample correlation coefficient. With regression, however, inferences focus on the correlation coefficient, the regression coefficient(s), the intercept, the change in the regression coefficient, and something called the *odds ratio*. We consider these different inferential procedures as we look at bivariate regression, multiple regression, and logistic regression.

Although correlation and regression are not the same, correlational concepts serve as some (but not all) of regression's building blocks. With that being the case, you may wonder why this chapter is positioned here rather than immediately after Chapter 9. If this question has popped into your head, there is a simple answer. This chapter is located here because certain concepts from the analysis of variance and the analysis of covariance also serve as building blocks in some regression analyses. For example, researchers sometimes base their regression predictions (or explanations) on the interactions between independent variables. Also, regressions are sometimes conducted with one or more covariate variables controlled or held constant. Without knowing about interactions and covariates, you would be unable to understand these particular components of regression analyses.

We now turn our attention to the simplest kind of regression used by applied researchers. Take good mental notes as you study this material, for the concepts you now encounter provide a foundation for the other two kinds of regression to be considered later in the chapter.

Bivariate Regression

The simplest kind of regression analysis is called **bivariate regression.** First, we must clarify the purpose of and the data needed for this kind of regression. Then, we consider scatter diagrams, lines of best fit, and prediction equations. Finally, we discuss inferential procedures associated with bivariate regression.

Purpose and Data

As you would suspect based on its name, bivariate regression involves just two variables. One of the variables serves as the dependent variable whereas the other functions as the independent variable. The purpose of this kind of regression can be either prediction or explanation; however, bivariate regression is used most frequently to see how well scores on the dependent variable can be predicted from data on the independent variable.

To illustrate how bivariate regression can be used in a predictive manner, imagine that Sam, a 41-year-old tennis player, has been plagued by a knee injury that for months has failed to respond to nonsurgical treatment. Consequently, arthroscopic surgery is scheduled to repair Sam's bad knee. Even though he knows that arthroscopic procedures usually permit a rapid return to usual activity, Sam would like to know how long he will be out of commission following surgery. His presurgery question to the doctor is short and sweet: "When will I be able to play again?" Clearly, Sam wants his doctor to make a prediction.

Although Sam's doctor might be inclined to answer this question concerning down-time by telling Sam about the *average* length of convalescence for tennis players following arthroscopic knee surgery, that is really not what Sam wants to know. Obviously, some people bounce back from surgery more quickly than do others. Therefore, Sam wants the doctor to consider his (i.e., Sam's) individual case and make a prediction about how long *he* will have to interrupt his on-court activity. If Sam's doctor is aware of what has happened with other tennis players who have had arthroscopic knee surgery, and if the doctor has a computer program that can perform a bivariate regression, he could provide Sam with a better-than-average answer to the question about postsurgical down time.

In the study conducted with people like Sam, imagine that there are 12 tennis players who had one of their knees repaired via arthroscopic surgery. Also imagine that data exist on each person regarding two variables: age and number of postsur-

EXCERPTS 16.3–16.4 • *Dependent and Independent Variables*

Stress reactions to uncertainty were measured [via] the Physicians' Reactions to Uncertainty Scale (PRUS). . . . Epistemology was measured using the Physicians' Belief Scale (PBS). . . . Our primary hypothesis was tested [by means of] a simple bivariate regression with PRUS scores as the dependent variable and PBS scores as the independent variable.

Source: Evans, L., & Trotter, D. R. M. (2009). Epistemology and uncertainty in primary care: An exploratory study. *Family Medicine, 41*(5), 319–326.

To assess the extent to which students' perceptions of parenting style predicted evaluations of their parents' preferred music we performed a multiple regression analysis using the mean rating score of parent music (ratings given to only those pieces indicated as a favorite by a student's parent) as the dependent variable and scores on the caring and the autonomy dimensions of the PBI, as well as students' age and gender, as independent variables.

Source: Serbun, S. J., & DeBono, K. G. (2010). On appreciating the music of our parents: The role of the parent–child bond. *North American Journal of Psychology, 12*(1), 93–102.

conducted. In the second excerpt, there was one dependent variable and four independent variables. Despite these differences, notice how the researchers associated with each excerpt clearly designate the status of each variable as being a dependent variable or an independent variable.

The third difference between correlation and regression concerns the focus of inferential tests and confidence intervals. With correlation, there is just one thing that can be focused on: the sample correlation coefficient. With regression, however, inferences focus on the correlation coefficient, the regression coefficient(s), the intercept, the change in the regression coefficient, and something called the *odds ratio*. We consider these different inferential procedures as we look at bivariate regression, multiple regression, and logistic regression.

Although correlation and regression are not the same, correlational concepts serve as some (but not all) of regression's building blocks. With that being the case, you may wonder why this chapter is positioned here rather than immediately after Chapter 9. If this question has popped into your head, there is a simple answer. This chapter is located here because certain concepts from the analysis of variance and the analysis of covariance also serve as building blocks in some regression analyses. For example, researchers sometimes base their regression predictions (or explanations) on the interactions between independent variables. Also, regressions are sometimes conducted with one or more covariate variables controlled or held constant. Without knowing about interactions and covariates, you would be unable to understand these particular components of regression analyses.

We now turn our attention to the simplest kind of regression used by applied researchers. Take good mental notes as you study this material, for the concepts you now encounter provide a foundation for the other two kinds of regression to be considered later in the chapter.

Bivariate Regression

The simplest kind of regression analysis is called **bivariate regression.** First, we must clarify the purpose of and the data needed for this kind of regression. Then, we consider scatter diagrams, lines of best fit, and prediction equations. Finally, we discuss inferential procedures associated with bivariate regression.

Purpose and Data

As you would suspect based on its name, bivariate regression involves just two variables. One of the variables serves as the dependent variable whereas the other functions as the independent variable. The purpose of this kind of regression can be either prediction or explanation; however, bivariate regression is used most frequently to see how well scores on the dependent variable can be predicted from data on the independent variable.

To illustrate how bivariate regression can be used in a predictive manner, imagine that Sam, a 41-year-old tennis player, has been plagued by a knee injury that for months has failed to respond to nonsurgical treatment. Consequently, arthroscopic surgery is scheduled to repair Sam's bad knee. Even though he knows that arthroscopic procedures usually permit a rapid return to usual activity, Sam would like to know how long he will be out of commission following surgery. His presurgery question to the doctor is short and sweet: "When will I be able to play again?" Clearly, Sam wants his doctor to make a prediction.

Although Sam's doctor might be inclined to answer this question concerning down-time by telling Sam about the *average* length of convalescence for tennis players following arthroscopic knee surgery, that is really not what Sam wants to know. Obviously, some people bounce back from surgery more quickly than do others. Therefore, Sam wants the doctor to consider his (i.e., Sam's) individual case and make a prediction about how long *he* will have to interrupt his on-court activity. If Sam's doctor is aware of what has happened with other tennis players who have had arthroscopic knee surgery, and if the doctor has a computer program that can perform a bivariate regression, he could provide Sam with a better-than-average answer to the question about postsurgical down time.

In the study conducted with people like Sam, imagine that there are 12 tennis players who had one of their knees repaired via arthroscopic surgery. Also imagine that data exist on each person regarding two variables: age and number of postsur-

TABLE 16.1 *Data for Bivariate Regression Example*

Post-Surgical Down-Time and Age for 12 Adult Tennis Players

Player	*Age*	*Down-Time (days)*
Kathy	41	7
Alex	47	6
Nancy	36	5
David	29	5
Pat	41	6
Andrew	22	3
Allison	21	4
Gary	38	6
Emily	19	5
Candace	31	4
Ted	32	6
Barbara	24	4

Scatter Diagrams, Regression Lines, and Regression Equations

The component parts and functioning of regression can best be understood by examining a scatter diagram. In Figure 16.1, such a picture has been generated using the data from Table 16.1. There are 12 dots in this "picture," each positioned so as to reveal the age and postsurgical convalescent time for one of the tennis players.

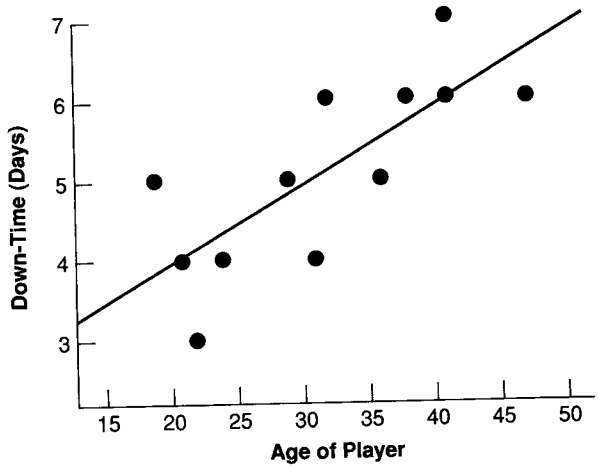

FIGURE 16.1 *Scatter Plot with Regression Line*

The scatter diagram in Figure 16.1 was set up with days of convalescence on the ordinate and age on the abscissa. These two axes of the scatter diagram were labeled like this because it makes sense to treat convalescence as the dependent variable. It is the variable toward which predictions will eventually be made for Sam and other tennis players who are similar to those who supplied the data we are currently considering. Age, however, is positioned on the abscissa because it is the independent variable. It is the variable that "supplies" data used to make the predictions.[3]

As you can see, a slanted line passes through the data points of the scatter diagram. This line is called the **regression line** or the **line of best fit,** and it functions as the tool our hypothetical doctor will use in order to predict how long Sam will have to refrain from playing tennis. As should be apparent, the regression line is positioned so as to be as close as possible to all of the dots. A special formula determines the precise location of this line; however, you do not need to know anything about that formula except that it is based on a statistical concept called *least squares.*[4]

Let's make a prediction for Sam, pretending now that we are his doctor. All we must do is turn to the scatter diagram and take a little trip with our index finger or our eyes. Our trip begins on the abscissa at a point equal to Sam's age. (Remember, Sam is 41 years old.) We move vertically from that point up into the scatter diagram until we reach the regression line. Finally, we travel horizontally (to the left) from that point on the regression line until reaching the ordinate. Wherever this little trip causes us to end up on the ordinate becomes our prediction for Sam's down time. According to our information, our prediction is that Sam will be out of commission for approximately six days.

Notice that our prediction of Sam's down time would have been shorter if he had been younger and longer if he had been older. For example, we would have predicted about four days if he had been 21 years old, or five days if he had been 31. These alternative predictions for a younger Sam are brought about by the tilt of the regression line. Because there is a positive correlation between the independent and dependent variables, the regression line tilts from lower left to upper right.

Although it is instructive to see how predictions are made by means of a regression line that passes through the data points of a scatter diagram, the exact same objective can be achieved more quickly and more scientifically by means of something called the **regression equation.** In bivariate, linear regression, this equation always has the following form:

$$Y' = a + b \cdot X$$

[3]Because we are dealing with regression (and not correlation), it would be improper to switch the two variables in the scatter diagram. The dependent variable always goes on the ordinate; the independent variable always goes on the abscissa.

[4]The *least squares principle* simply means that when the squared vertical distances of the data points from the regression line are added together, they yield a smaller sum than would be the case for any other straight

where Y' stands for the predicted score on the dependent variable, a is a constant, b is the **regression coefficient,** and X is the known score on the independent variable. This equation is simply the technical way of describing the regression line. For the data shown in Table 16.1 (and Figure 16.1), the regression equation turns out like this:

$$Y' = 1.978 + 0.098 \cdot X$$

To make a prediction for Sam by using the regression equation, we simply substitute Sam's age for X and then work out the simple math. Because the numbers in the regression equation are so close to being whole numbers, let's round things off a bit and rewrite the regression equation as $Y' = 2 + (0.10)X$. We now have a simple prediction model that says to add 2 to one-tenth of person's age, with the result being a guess as to that individual's down-time. When we do this for Sam, $Y' = 6.1$. If we don't round off, $Y' = 5.996$. The fact that these values are very similar to what we predicted from the scatter diagram should not be at all surprising. This is because the regression equation is nothing more than a precise mechanism for telling us where we should end up if, in a scatter diagram, we first move vertically from some point on the abscissa up to the regression line and then move horizontally from the regression line to the ordinate.

Whereas scatter diagrams with regression lines appear only rarely in research reports, regression equations show up more frequently. We will see an example shortly; first, however, let's consider the study that was conducted. In this investigation, the researchers collected data from 67 individuals with multiple sclerosis (MS) who had been on a home-based self-medication program. First, the researcher asked each patient to estimate the percentage of self-injections that had been missed during the previous two months. Then, self-medication adherence was electronically monitored during the next two-month period. These data were used to see if retrospective self-reports could predict prospective (i.e., future) adherence.

The data assessing self-medication adherence during the second half of the study did not come from patients' self-reports. Instead, the researchers provided patients with special containers into which they put their disposable needles after they were used. These containers—called MEMS (Medication Event Monitoring System)—had been designed to record electronically the precise date and time any needle was put into the container. At the end of the full four-month period of the study, the researchers performed a regression analysis in which retrospective self-report of adherence served as the independent (predictor) variable and the electronic MEMS-based measure of adherence was the dependent (criterion) variable.

In Excerpt 16.5, we see the regression equation that appeared in the research report that summarized the self-medication study. As indicated in the excerpt, the data used to create this regression equation came from all research participants except four who reported very poor adherence over the first two-month period of the study.

EXCERPT 16.5 • *The Regression Equation in Bivariate Regression*

[We assessed] the relationship between self-reported retrospective adherence and prospective electronic monitoring among [63] patients who did not report poor adherence at the outset of the study. The data fit the regression line $y = 2.00x + 0.42$ (x = retrospective self-report and y = MEMS % days missed), suggesting that the prediction estimates of poor objective prospective adherence may be achieved by doubling patient reports of retrospective missed doses.

Source: Bruce, J. M., Hancock, L. M., & Lynch, S. G. (2010). Objective adherence monitoring in multiple sclerosis: Initial validation and association with self-report. *Multiple Sclerosis, 16*(1), 112–120.

It should be noted that there are two kinds of regression equations that can be created in any bivariate regression analysis. One of these is called an **unstandardized regression equation.** This is the kind we have considered thus far, and it has the form $Y' = a + b \cdot X$. The other kind of regression equation (that can be generated using the same data) is called a **standardized regression equation.** A standardized regression equation has the form $z'_Y = \beta \cdot z_X$. These two kinds of regression equations differ in three respects. First, a standardized regression equation involves z-scores on both the independent and the dependent variables, not raw scores. Second, the standardized regression equation does not have a constant (i.e., a term for a). Finally, the symbol β (called a **beta weight**) is used in place of the regression coefficient, b.

Interpreting a, b, r, and r^2 in Bivariate Regression

When used for predictive purposes, the regression equation has the form $Y' = a + b \cdot X$. Now that you understand how this equation works, let's take a closer look at its two main ingredients, a and b. In addition, let's now pin down the regression meaning of r and r^2.

Earlier, I referred to a as the *constant*. Alternatively, this component of the regression equation is called the **intercept.** Simply stated, a indicates where the regression line in the scatter diagram would, if extended to the left, intersect the ordinate. It indicates, therefore, the value of Y' for the case where $X = 0$. In many studies, it may be quite unrealistic (or downright impossible) for there to be a case where $X = 0$; nonetheless, $Y' = a$ when $X = 0$.

Earlier, we considered data concerning the post-surgical down-time of 12 adult tennis players. In the regression equation based on those data, the constant was equal to 1.978. That is not a very realistic number, for it indicates the predicted number of down-time days for a tennis player whose age is 0! Clearly, a may be totally devoid of meaning within the context of a study's independent and dependent

within a scatter diagram, because a indicates the point where the regression line intercepts the ordinate.

The other main component of the regression is b, the regression coefficient. When the regression line has been positioned within the data points of a scatter diagram, b simply indicates the **slope** of that line. As you probably recall from your high school math courses, *slope* means "rise over run." In other words, the value of b signifies how many predicted units of change (either up or down) in the dependent variable there are for any one unit increase in the independent variable. In Figure 16.1, the regression equation has a slope equal to .098. This means that the predicted down time for our hypothetical patient Sam would be about one-tenth of a day longer if the surgery is put off a year (assuming Sam's knee problem, health status, and fitness level do not change).

When researchers use bivariate regression, they sometimes will focus on either b or β more than anything else. Consider, for example, Excerpt 16.6. In the study associated with this excerpt, several college-age men and women were measured on several traits, one of which was aggression. In addition, the research participants had their pain tolerance measured via a device that sent increasing levels of electrical current into their non-dominant hands. After collecting the data, the researchers did a bivariate regression within each gender group to investigate the connection between trait aggression and pain tolerance. Notice how the researchers focused their attention on the beta weights when comparing the men versus the women.

EXCERPT 16.6 • *Focusing on the Regression Coefficient*

A sample of 195 collegiate men and women completed trait measures and a laboratory assessment of pain tolerance. . . . To determine whether the relationship between pain tolerance and aggression differed by sex of participant, we . . . computed simple regression coefficients of pain tolerance and trait aggression for men and women. Analyses indicated that while there was no relationship between pain tolerance and trait aggression for women [$\beta = -.02$], there was a significant positive relationship for men [$\beta = .31$]. . . . Pain tolerance was significantly and positively related to trait aggression in men, but in women the relation between pain tolerance and trait aggression was nil and nonsignificant.

Source: Reidy, D. E., Dimmick, K., MacDonald, K., & Zeichner, A. (2009). The relationship between pain tolerance and trait aggression: Effects of sex and gender role. *Aggressive Behavior, 35*(5), 422–429.

When summarizing the results of a regression analysis, researchers will normally indicate the value of r (the correlation between the independent and dependent variables) or r^2. You already know, of course, that such values for r and r^2 measure the strength of the relationship between the independent and dependent variables. However, each has a special meaning, within the regression context, that is worth learning.

As you might expect, the value of r is high to the extent that the scatter diagram's data points are located close to the regression line. Although that is undeniably true, there is a more precise way to conceptualize the regression meaning of r. Once the regression equation has been generated, that equation could be used to predict Y for each person who provided the scores used to develop the equation. In one sense, that would be a very silly thing to do, because predicted scores are unnecessary in light of the fact that *actual* scores on the dependent variable are available for these people. However, by comparing the predicted scores for these people against their actual scores (both on the dependent variable), we would be able to see how well the regression equation works. The value of r does exactly this. It quantifies the degree to which the predicted scores correlate, or match up with, the actual scores.

Just as r has an interpretation in regression that focuses on the dependent variable, so it is with r^2. Simply stated, the coefficient of determination indicates the proportion of variability in the dependent variable that is explained by the independent variable. As illustrated in Excerpt 16.7, r^2 is usually turned into a percentage when it is reported in research reports.

EXCERPT 16.7 • *Variability in the Dependent Variable Explained by Variability in the Independent Variable*

> Bivariate regression analyses [revealed] MLSS [running speed at maximal lactate steady state] as the strongest individual predictor [$r = 0.93, r^2 = 0.87$] for 2-mile running performance. [This predictor] explained . . . 87% of the variance in running performance.
>
> *Source:* Tolfrey, K., Hansen, S. A., Dutton, K., McKee, T., & Jones, A. M. (2009). Physiological correlates of 2-mile run performance as determined using a novel on-demand treadmill. *Applied Physiology, Nutrition & Metabolism, 34*(4), 763–772.

Inferential Tests in Bivariate Regression

The data used to generate the regression line or the regression equation are typically considered to have come from a sample, not a population. Thus the component parts of a regression analysis—a, b, and r—are usually viewed as sample statistics, not population parameters. Accordingly, it should not come as a surprise that researchers conduct one or more inferential tests whenever they perform a regression analysis.

In bivariate regression, a test on r is mathematically equivalent to a test on b or β. Therefore, you are unlikely to see a case where both r and b (or r and β) are tested, because such tests would be fully redundant with each other. Researchers have the freedom to have their test focus on r or b or β. If r is tested, try to remember the things we considered in Chapter 9. In particular, keep in mind that

in the population is equal to 0.00. Also keep in mind that a test of this null hypothesis operates properly only if certain assumptions (e.g., linearity) are met.

When bivariate regression is involved in a study, you are likely to see a test of significance on either b or β rather than r. The null hypothesis in this alternative (but equivalent) kind of test says that the population value of the regression or beta weight is 0. Stated differently, the null hypothesis in such tests is that the regression line has no tilt, thus meaning that the independent variable provides no assistance in predicting scores on the dependent variable. In Excerpt 16.8, we see a case where such a test was applied. This excerpt is worth considering for two reasons. First, can you tell which of the two variables was the dependent variable? Second, do you agree the researchers deserve a pat on the back for presenting what they did immediately after they cite the test's p-level?

EXCERPT 16.8 • *Testing a Regression Coefficient*

> The aim of this study, part of a cross-sectional blood pressure survey, was to study the influence of ambient temperature on blood pressure in a rural West African adult population. . . . Blood pressure, anthropometric, time of blood pressure and room temperature measurements were taken in 574 adult males and females. . . . Linear regression analysis showed that SBP [systolic blood pressure] was significantly and inversely related to ambient temperature ($b = -0.98, p = 0.02$, 95% confidence interval: -1.19 to -0.11).
>
> *Source:* Kunutsor, S. K., & Powles, J. W. (2010). The effect of ambient temperature on blood pressure in a rural West African adult population: A cross-sectional study. *Cardiovascular Journal of Africa, 21*(1), 17–20.

Multiple Regression

We now turn our attention to the most popular regression procedure of all, **multiple regression.** This form of regression involves, like bivariate regression, a single dependent variable. In multiple regression, however, there are two or more independent variables. Stated differently, multiple regression involves just one Y variable but two, three, or more X variables.[5]

In three important respects, multiple regression is identical to bivariate regression. First, a researcher's reason for using multiple regression is the same as the reason for using bivariate regression, either prediction (with a focus on the dependent variable) or explanation (with a focus on the independent variables). Second, a regression equation is involved in both of these regression procedures. Third, both bivariate and multiple regression almost always involve inferential tests and a

[5]Recall that the dependent variable (Y) is sometimes referred to as the *criterion, outcome,* or *response variable,* whereas the independent variable (X) is sometimes referred to as the *predictor* or *explanatory variable.*

measure of the extent to which variability among the scores on the dependent variable has been explained or accounted for.

Although multiple regression and bivariate regression are identical in some respects, they also differ in three extremely important ways. As you will see, multiple regression can be done in *different ways* that lead to different results, it can be set up to accommodate *covariates* that the researcher wishes to control, and it can involve (as predictor variables) one or more *interactions* between independent variables. Bivariate regression has none of these characteristics.

In upcoming sections, these three unique features of multiple regression are discussed. We begin, however, with a consideration of the regression equation that comes from the analysis of data on one dependent variable and multiple independent variables. This equation functions as the most important stepping stone between the raw scores collected in a study and the findings extracted from the investigation.

The Regression Equation

When a regression analysis involves one dependent variable and two independent variables, the regression equation takes the form

$$Y' = a + b_1X_1 + b_2X_2$$

where Y' stands for the predicted score on the dependent variable, a stands for the constant, b_1 and b_2 are regression coefficients, and X_1 and X_2 represent the two independent variables. As indicated previously, multiple regression can accommodate more than two independent variables. In such cases, the regression equation is simply extended to the right, with an extra term (made up of a new b multiplied by the new X) added for each additional independent variable. The presence of these extra terms, of course, does not alter the fact that the regression equation contains only one Y' term (located on the left side of the equation) and only one a term (located on the right side of the equation).

In Excerpts 16.9 and 16.10, we see regression equations that were created for the situations where there were two or three independent variables, respectively. In

EXCERPTS 16.9–16.10 • *Regression Equations with Different Numbers of Independent Variables*

Multiple regression analysis was conducted to evaluate how well overall contact conditions predicted post-test scores on the MGUDS-S. . . . The raw coefficients for the predictive equation were as follows:

MGUDS-S post-test score = 24.09 + .67(MGUDS-S pretest) + 1.08(SICS total).

Source: Seaman, J., Beightol, J., Shirilla, P., & Crawford, B. (2010). Contact theory as a framework for experiential activities as diversity education: An exploratory study. *Journal of Experiential Education, 32*(3), 207–225.

EXCERPTS 16.9–16.10 • (*continued*)

Results showed that social systems did have an impact on nurses' spiritual intelligence. . . . In general, the study yielded a regression equation associated with independent variables as follows: spiritual intelligence = 4.10 + 3.32 (childhood spirituality) + 1.01 (social system) + .03 (age).

Source: Yang, K-P., & Wu, X. J. (2009). Spiritual intelligence of nurses in two Chinese social systems: A cross-sectional comparison study. *Journal of Nursing Research, 17*(3), 189–198.

these regression equations, note that the numerical values of 24.09 (in Excerpt 16.9) and 4.10 (in Excerpt 16.10) represent a (i.e., the constants). The other numerical values in each equation are the bs (i.e., the regression coefficients), each of which is paired with a particular independent variable.

In each of the regression equations shown in Excerpts 16.9 and 16.10, the algebraic sign between any two adjacent terms on the right side of the equation is positive, meaning that the sign of every regression coefficient was positive. In some multiple regression equations, one of more of the bs ends up being negative. The sign of a regression coefficient simply indicates the nature of the relationship between that particular X variable and the dependent variable. Thus, if the nurses in the study that gave us Excerpt 16.10 had also been measured on how extensively they feel independently in control of their own lives, this predictor variable's regression coefficient would likely have a negative sign in front of it, thereby implying an inverse relationship between feeling independently powerful and level of spiritual intelligence.

Regardless of whether the multiple regression is being conducted for predictive or explanatory purposes, the researcher is usually interested in comparing the independent variables to see the extent to which each one helps the regression analysis achieve its objective. In other words, there is usually interest in finding out the degree to which each independent variable contributes to successful predictions or valid explanations. Although you (as well as a fair number of researchers) may be tempted to look at the bs in order to find out how well each independent variable works, this should not be done because each regression coefficient is presented in the units of measurement used to measure its corresponding X. Thus, if one of the independent variables in a multiple regression is height, its b will differ in size depending on whether height measurements are made in centimeters, inches, feet, or miles.

To determine the relative importance of the different independent variables, the researcher must look at something other than an unstandardized regression equation like those we have seen thus far. Instead, a standardized regression equation can be examined. This kind of regression equation, for the case of three independent variables, takes the form

$$z'_Y = \beta_1 z_{X_1} + \beta_2 z_{X_2} + \beta_3 z_{X_3}$$

Note that this equation presents the dependent and independent variables in terms of z, it has no constant term, and it uses the symbol β instead of b. These βs are like standardized regression coefficients, and they are called **beta weights.**

Although standardized regression equations are rarely included in research reports, researchers often extract the beta weights from such equations and present the numerical values of these βs. In Excerpt 16.11, we see an instance in which this was done. Notice that the beta weights in this excerpt were compared, with the researchers pointing out that the first beta weight was more than twice as large as the second beta, and more than three times as large as the third. Unstandardized regression coefficients cannot be compared like this.

EXCERPT 16.11 • *Beta Weights*

The betas from regression model were used to determine the relative weights of each factor. [Results indicated that] attitude toward the behavior had the most substantial impact ($\beta = 0.569$) on teachers' intentions to use computers to create and deliver lessons, producing a change of 0.569 units in behavioral intention for each unit change in attitude. This influence on intention is more than twice that of subjective norm ($\beta = 0.229$) and more than three times that of perceived behavioral control ($\beta = 0.144$).

Source: Lee, J., Cerreto, F. A., & Lee, J. (2010). Theory of planned behavior and teachers' decisions regarding use of educational technology. *Journal of Educational Technology & Society, 13*(1), 152–164.

Before concluding our discussion of regression equations, three important points must be made. First, one or more of the independent variables in a regression analysis can be categorical in nature. For example, gender is often used in multiple regression to help accomplish the researcher's predictive or explanatory objectives. As you see the technique of multiple regression used in different studies, you are likely to see a wide variety of categorical independent variables included, such as marital status (single, married, divorced), highest educational degree (high school diploma, bachelor's degree, Master's degree, Ph.D.), and race (Black, White, Hispanic). Such variables are sometimes referred to as **dummy variables.**

Second, researchers often include a term in the regression equation that represents the interaction between two independent variables. Just as two independent variables in a two-way ANOVA can be examined to see if they interact, so too can the interaction of independent variables be assessed in regression contexts. We consider this feature of multiple regression later in the chapter; for now, all I want to do is alert you to the fact that interactions are often used as independent variables

My third and final comment about regression equations is an important warning. Simply stated, be aware that the regression coefficients (or beta weights) associated with the independent variables can change dramatically if the analysis is repeated with one of the independent variables discarded or another independent variable added. Thus, regression coefficients (or beta weights) do not provide a pure and absolute assessment of any independent variable's worth. Instead, they are *context dependent.*

Three Kinds of Multiple Regression

Different kinds of multiple regression exist because there are different orders in which data on the independent variables can be entered into the analysis. In this section, we consider the three most popular versions of multiple regression: simultaneous multiple regression, stepwise multiple regression, and hierarchical multiple regression.

In **simultaneous multiple regression,** the data associated with all independent variables are considered at the same time. This kind of multiple regression is analogous to the process used in preparing vegetable soup where all ingredients are thrown into the pot at the same time, stirred, and then cooked together. In Excerpt 16.12, we see an example of simultaneous multiple regression.

EXCERPT 16.12 • *Simultaneous Multiple Regression*

> A series of [bivariate] linear regression analyses was conducted and analyses indicated that all of the predictor variables independently predicted rebuilding the marriage relationship [i.e., mid-life marital satisfaction]. . . . A simultaneous multiple regression analysis was then conducted with adaptive appraisal, social support, and compensating experiences as predictor variables and rebuilding the marriage relationship as the criterion variable ($n = 476$).
>
> *Source:* Huber, C. H., Navarro, R. L., Womble, M. W., & Mumme, F. L. (2010). Family resilience and midlife marital satisfaction. *The Family Journal: Counseling and Therapy for Couples and Families, 18*(2), 136–145.

Stepwise multiple regression analysis is analogous to the process of preparing a soup in which the ingredients are tossed into the pot based on the amount of each ingredient. Here the stock goes in first (because there is more of that than anything else), followed by the vegetables, the meat, and finally the seasoning. Each of these different ingredients is meant to represent an independent variable, with "amount of ingredient" equated, somewhat, to the size of the bivariate correlation between a given independent variable and the dependent variable. Here, in **stepwise multiple regression,** the computer determines the order in which the independent variables become a part of the regression equation. In Excerpt 16.13, we see an example of this kind of multiple regression.

EXCERPT 16.13 • *Stepwise Multiple Regression*

Patients completed a symptom-limited exercise treadmill test. . . . Exercise test time (ETT) was recorded in seconds and taken as a measure of exercise capacity. . . . Stepwise multiple regression analysis was performed to examine predictors of exercise test time in our cohort. Variables entered into the model included traditional cardiovascular risk factors (age, sex, presence/absence of hypertension, diabetes, hyperlipidemia, family history of cardiovascular disease (CVD), and BMI), BAD, FMD and NMD.

Source: Heffernan, H. S., Karas, R. H., Patvardhan, E. A., & Kuvin, J. T. (2010). Endothelium-dependent vasodilation is associated with exercise capacity in smokers and non-smokers. *Vascular Medicine, 15*(2), 119–125.

Instead of preparing our vegetable soup by simply tossing everything into the pot at once or by letting the amount of an ingredient dictate its order of entry, we could put things into the pot on the basis of concerns regarding flavor and tenderness. If we wanted garlic to flavor everything else, we would put it in first, even though there is only a small amount of it required by the recipe. Similarly, we would hold back some of the vegetables (and not put them in with the others) if they are tender to begin with and we want to avoid overcooking them. **Hierarchical multiple regression** is like cooking the soup in this manner, for in this form of regression the independent variables are entered into the analysis in stages. Often, as illustrated in Excerpt 16.14, the independent variables that are entered first correspond with things the researcher wishes to control. After these **control variables** are allowed to explain as much variability in the dependent variable as they can, then the other variables are entered to see if they can contribute above and beyond the independent variables that went in first.

EXCERPT 16.14 • *Hierarchical Multiple Regression*

Hierarchical multiple regression was used to analyze the relative importance of personal and peer attitudes supporting sexual aggression in predicting men's willingness to intervene against sexual aggression. We included several demographic variables that were potentially or theoretically related to our variables of interest [year in school, race, fraternity membership, sports team membership, and sexual orientation]. These demographic control variables were entered on the first step of the multiple regression. MCSDS scores were entered on the second step to control for social desirability. Personal and peer attitudes supporting sexual aggression were entered simultaneously on the third step.

Source: Brown, A. L., & Messman-Moore, T. L. (2009). Personal and perceived peer attitudes supporting sexual aggression as predictors of male college students' willingness to intervene against sexual aggression. *Journal of Interpersonal Violence, 25*(3), 503–517.

R, R^2, ΔR^2, Adjusted R^2, and sr^2 in Multiple Regression

In multiple regression studies, the extent to which the regression analysis achieves its objective is usually quantified by means of R, R^2, or adjusted R^2. Sometimes two of these will be presented, and occasionally you will see all three reported for the same regression analysis. These elements of a multiple regression analysis are not superficial and optional add-ons; instead, they are as central to a regression analysis as the regression equation itself.

In bivariate regression, r provides an indication of how well the regression equation works. It does that by quantifying the degree to which the predicted scores match up with the actual scores (on the dependent variable) for the group of individuals used to develop the regression equation. The R of multiple regression can be interpreted in precisely the same way. **Multiple *R*** is what we get if we compute Pearson's r between Y and Y' scores for the individuals who provided scores on the independent and dependent variables.

Although the value of R sometimes appears when the results of a multiple regression are reported, researchers are far more likely to report the value of R^2 or to report the percentage equivalent of R^2. By so doing, the success of the regression analysis is quantified by reporting the proportion or percentage of the variability in the dependent variable that has been accounted for or explained by the study's independent variables. Excerpt 16.15 illustrates the way researchers use R^2 in an explained-variance manner.

EXCERPT 16.15 • *R^2 as an Index of Explained Variance*

> The correlation coefficient resulting from the [multiple regression] analysis shows that there is a relatively strong correlation ($R = .79$) between the four relationship dimensions and giving history. The coefficient of determination is relatively strong ($R^2 = .62$) and shows moderate strength in predicting past giving. Thus, 62% of the variance in the number of years the participants have donated to the organization is explained by the four relationship dimensions.
>
> *Source:* Waters, R. D. (2010). Increasing fundraising efficiency through evaluation: Applying communication theory to the nonprofit organization–donor relationship. *Nonprofit and Voluntary Sector Quarterly,* in press.

When a multiple regression analysis is conducted with the data from all independent variables considered simultaneously, only one R^2 can be computed. In stepwise and hierarchical regression, however, several R^2 values can be computed, one for each stage of the analysis wherein individual independent variables or sets of independent variables are added. These R^2 values get larger at each stage, and the *increase* from stage to stage is referred to as R^2 change. Another label for the

increment in R^2 that's observed as more and more independent variables are used as predictors is ΔR^2 where the symbol Δ stands for the two-word phrase *change in*.

Excerpt 16.16 illustrates nicely the concept of ΔR^2. In the first step of the hierarchical multiple regression, the control variables of SES and gender were entered into the regression model, producing an R^2 of .06. In the second step of the regression analysis, two more independent variables (reading score and reading self-efficacy) entered the model, and they explained an additional 21 percent of variability in the students' English grades. In the last step of the regression analysis, the researcher entered the main variable he was concerned about, a measure of each child's "confidence to manage learning." As indicated in the excerpt, this final independent variable explained an additional eight percent of variability above and beyond what already had been explained by the first four variables.

EXCERPT 16.16 • ΔR^2 *in Stepwise or Hierarchical Multiple Regression*

> [We conducted] hierarchical multiple regression for the LD and NLD groups, with end-of-term English grade as the dependent variable. Control variables of SES (parent education level) and sex were entered at Step 1, followed by reading score and reading self-efficacy at Step 2, and finally SESRL at Step 3. For the LD group, the entry of the control variables at Step 1 did not significantly predict English grade [$R^2 = .06$]. The entry of reading score and reading self-efficacy at Step 2 significantly increased explained variance ($\Delta R^2 = .21$), as did the entry on the final step of SESRL ($\Delta R^2 = .08$), with a final R^2 of .35.
>
> *Source:* Klassen, R. M. (2010). Confidence to manage learning: The self-efficacy for self-regulated learning of early adolescents with learning disabilities. *Learning Disability Quarterly, 33*(1), 19–30.

Either in place of or in addition to R^2, something called **adjusted R^2** is often reported in conjunction with a multiple regression analysis. If reported, adjusted R^2 takes the form of a proportion or a percentage. It is interpreted just like R^2, because it indicates the degree to which variability in the dependent variable is explained by the set of independent variables included in the analysis. The conceptual difference between R^2 and adjusted R^2 is related to the fact that the former, being based on sample data, always yields an overestimate of the corresponding population value of R^2.

Adjusted R^2 removes the bias associated with R^2 by reducing its value. Thus, this adjustment anticipates the amount of so-called **shrinkage** that would be observed if the study were to be replicated with a much larger sample. As you might expect, the size of this adjustment is inversely related to study's sample size.[6]

[6]The size of the adjustment is also influenced by the number of independent variables. With more independent

When reporting the results of their multiple regression analyses, some researchers (who probably do not realize that R^2 provides an exaggerated index of predictive success) report just R^2. Of those who are aware of the positive bias associated with R^2, some include only adjusted R^2 in their reports whereas others include both R^2 and adjusted R^2.

In addition to assessing the effectiveness of the full regression model, many researchers evaluate the worth of each independent variable. Beta weights can help in this regard, but they do not indicate how much variability in the dependent variable is explained uniquely by each independent variable. To accomplish this goal, the square of the semi-partial correlation, symbolized sr^2, is computed for each variable used to help predict (or explain) variability in the dependent variable. In a very real sense, sr^2 is analogous to R^2, with the former index focused on a single predictor, whereas the latter is based on the full set of predictors. Excerpt 16.17 shows how sr^2 can help in the interpretation of results.

EXCERPT 6.17 • *Assessing the Worth of Individual Independent Variables with sr^2*

The value of the square of the coefficient of semi-partial correlation (sr^2) for each independent variable was also calculated, which allowed us to assess the unique contribution of this variable relative to R^2 in the set of variables included in the model. . . . The regression model found is significant, allowing explanation of 33.9% of the variance in the anxiety/depression symptoms. The only variable with a significant predictive value is the "negative reactivity" temperament dimension, which, once the variance explained by the remainder is controlled, is responsible for 19.5% of the variance.

Source: Lima, L., Guerra, M. P., & de Lemos, M. S. (2010). The psychological adjustments of children with asthma: Study of associated variables. *Spanish Journal of Psychology, 13*(1), 353–363.

Inferential Tests in Multiple Regression

Researchers can apply several different kinds of inferential tests when they perform a multiple regression. The three most frequently seen tests focus on β, R^2, and ΔR^2. Let's consider what each of these tests does and then look at an excerpt in which all three of these tests are used.

When the beta weight for a particular independent variable is tested, the null hypothesis says that the parameter value is equal to 0. If this were true, that particular independent variable would be contributing nothing to the predictive or explanatory objective of the multiple regression. Because of this, researchers frequently test each of the betas in an effort to decide (1) which independent variables should be included in the regression equation being built, or (2) which independent variables

included in an already-developed regression equation turned out to be helpful. Beta weights are usually tested with two-tailed *t*-tests.[7]

When R^2 is tested, the null hypothesis says that none of the variance in the dependent variable is explained by the collection of independent variables. (This H_0, of course, has reference to the study's population, not its sample.) This null hypothesis normally is evaluated via an *F*-test. In most studies, the researcher hopes that this H_0 will be rejected.[8]

When ΔR^2 is tested, the null hypothesis says that any new independent variables added to an existing regression equation are totally worthless in helping to explain variability in the dependent variable. As with the null hypotheses associated with tests on beta weights and R^2, this particular H_0 has reference to the study's population, not its sample. A special *F*-test is used to evaluate this null hypothesis. This kind of test, of course, logically fits into the procedures of stepwise and hierarchical multiple regression; however, it is never used within the context of a simultaneous multiple regression.[9]

Consider now Excerpt 16.18 which comes from a study involving a hierarchical multiple regression. Take the time to look at this excerpt closely, because it

EXCERPT 16.18 • *Inferential Tests in Multiple Regression*

Medication adherence was negatively associated with extreme violence ($r = -.21$, $p = .01$) as well as with substance coping ($r = -0.20, p = -.012$). Hierarchical multiple regression was used to assess the predictive ability of extreme violence and substance coping on medication adherence for the total sample, after controlling for both gender and time since diagnosis. Gender and time since diagnosis, which was calculated in months, were entered at Step 1, explaining 3% of the variance in medication adherence. After entry of extreme violence and substance use coping at Step 2, the total variance explained by the model as a whole was 18%, $F(4, 85) = 4.54$, $p = .002$. The two measures accounted for an additional 15% of the variance in adherence, after controlling for gender and time since diagnosis, R^2 change $= .15$, F change $(2, 85) = 7.56, p = .001$. In the final model, both extreme violence and substance use coping were statistically significant, with extreme violence recording a higher beta value (beta $= -0.29, p = .005$) than substance use coping (beta $= -0.20, p = 0.050$).

Source: Lopez, E. J., Jones, D. L., Villar-Loubet, O. M., Arheart, K. L., & Weis, S. M. (2010). Violence, coping, and consistent medication adherence in HIV-positive couples. *AIDS Education & Prevention, 22*(1), 61–68.

[7]The *df* for this kind of *t*-test is equal to the sample size minus one more than the number of independent variables.
[8]The first *df* for this kind of *F*-test is equal to the number of independent variables; the second *df* value is equal to the sample size minus one more than the number of independent variables.
[9]The *df* for this kind of *F*-test is equal to (1) the number of new independent variables and (2) the sample size

contains—in the second step of the analysis—tests of R^2, ΔR^2, and the beta weights for the two key independent variables.

Two additional features of Excerpt 16.18 are worth noting. First, the initial sentence contains the bivariate correlations between each of the primary independent variables and the dependent variable. Each of those *r*s has a negative sign, which is why the two betas, presented in the excerpt's final sentence, turned out to be negative rather than positive. The second thing to notice about Excerpt 16.18 is the fact that the value of R^2 increased dramatically from step 1 to step 2. Clearly, variation in medication adherence is associated with violence and substance abuse, even after gender and time since diagnosis are controlled.

Moderated and Mediated Multiple Regression

Researchers sometimes report that they have conducted a moderated multiple regression or a mediated multiple regression. Despite the similar-sounding names, these two kinds of regression are quite different. Let's briefly consider the goals and the procedure of these two special cases of multiple regression.

When researchers conduct a **moderated multiple regression,** their goal is to see if the findings of the multiple regression are the same (or perhaps different) for different subgroups of people or different settings. For example, suppose data are collected from several young adults on a variety of independent variables that might explain variability in the study's dependent variable: level of satisfaction with a first date. In this situation, the researcher might choose to do a moderated multiple regression due to the thought that men and women could have different reasons for thinking that a first date was a terrific or terrible experience (or anywhere between these extremes).

In Excerpt 16.19, we see an example of a moderated multiple regression conducted in a business setting with data gathered from a large number of employees. The researchers wanted to see if employee commitment to an organization was related to the employee's perception of fit between him or her and the company he

EXCERPT 16.19 • *Moderated Multiple Regression*

> A moderated hierarchical multiple regression analysis was computed to predict organizational commitment (*H1*). In Step 1, strategy fit and job alternatives were entered, and in Step 2, the product of strategy fit and job alternatives was entered. The results are presented in Table 2 [not shown here]. Step 1, which includes both main effects (strategy fit and job alternatives), was statistically significant, $R^2 = .11, p < .01$, with strategy fit as the significant predictor, $\beta = .32, p < .01$. The change in R^2 when the product term was entered in Step 2 was also statistically significant ($\Delta R^2 = .02, p < .05$). Therefore, the results suggest that job alternatives moderate the relation between strategy fit and organizational commitment. . . . [I]f respondents
>
> (*continued*)

EXCERPT 16.19 • (*continued*)

perceive that there are numerous job alternatives, then the correlation between strategy fit and organizational commitment is positive. . . . However, when respondents perceive that there are few job alternatives, then there is no relationship between strategy fit and organizational commitment. When employees do not feel that they have other job alternatives, then their commitment to their organization is the same regardless of whether there is a fit or misfit in strategy. Therefore, as predicted in *H1*, job alternatives moderate the relation between strategy fit and organizational commitment.

Source: Da Silva, N., Hutcheson, J., & Wahl, G. D. (2010). Organizational strategy and employee outcomes: A person–organization fit perspective. *Journal of Psychology, 144*(2), 145–161.

or she worked for. The researchers also wanted to see if the strength of that relationship varies depending on whether other job opportunities exist for the employee. To answer these questions, a moderated multiple regression was conducted. The dependent variable was organizational commitment, with the analytic approach being hierarchical in nature. In step 1, the independent variables were strategy fit and job opportunities. Then in step 2, the researchers added a term to the regression model: the interaction between fit and other opportunities. It was this interaction term that caused the regression analysis to be of the moderated variety. Carefully read the material in Excerpt 16.19 and you get a good feel for the goal and procedures of a moderated multiple regression.

With a mediated multiple regression, a researcher's goal is to see whether the apparent causal influence of one variable on a second variable is attributable—totally, partially, or not at all—to the first variable having an influence through some other third variable. To illustrate, let's consider the plight of graduate students who are on teaching assistantships. These individuals probably feel varying levels of job stress because they are pressured to do three important things in the university setting: (1) earn high grades in the graduate courses they take, (2) perform well as instructors in the undergraduate courses they teach, and (3) get research papers published so they can be competitive in the job market once they complete their graduate degree programs.

We might hypothesize that the graduate students' levels of stress will impact their physical health, with those with more stress being more susceptible to colds, the flu, allergies, and other ailments. At this point in our example, we have an independent variable (job stress) that may have a causal impact on the dependent variable (illness). We could investigate this hypothesized relationship in a simple way by measuring a large group of graduate students on these two variables and then correlating the two sets of scores. By itself, a statistically significant positive correlation from our data would not prove that job stress causes illness; nevertheless,

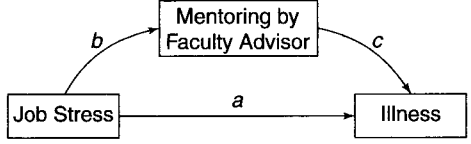

FIGURE 16.2 ***Diagram of Three-Variable Mediation Model***

Continuing our stress-and-illness example, suppose we now add a third variable to our causal model: faculty advisor mentoring. The quality of faculty mentoring varies, of course, with some graduate students having faculty advisors who are better than others at handling job stress, more willing to talk with their advisees about the difficulty of being a graduate teaching assistant, and more knowledgeable about the early-warning signals of impending illness. We might hypothesize that the level of job stress felt by graduate students is mediated by the quality of their mentors. In other words, we might conjecture that at least a portion of a graduate student's job stress flows through the relationship he or she has with his or her faculty advisor, with a good advisor functioning to lessen the stress (and ultimately lessen the likelihood of illness), whereas a not-so-good advisor would do little or nothing to mediate the causal impact of stress on illness.

A visual depiction of our example appears in Figure 16.2. The three lines are meant to represent the paths of the causal influence. Line *a* represents the *direct effect* of job stress on illness, whereas lines *b* and *c* represent the path of mediation. As Figure 16.2 illustrates, the effect of job stress may flow through, and may well be reduced by, the quality of mentoring provided by the graduate faculty advisor. Any reduction in the causal impact of the independent variable on the dependent variable is called the *indirect effect.*

To assess the worthiness of a mediated model, researchers do more than just draw diagrams with directional arrows between variable names. To test their hypotheses, they collect and analyze data. The usual data analytic strategy involves a four-step set of regression analyses. The sequence of tests and the needed results to establish mediation are shown in Excerpt 16.20. This excerpt comes from a study

EXCERPT 16.20 • *Mediated Multiple Regression*

To explore whether altruism had an indirect effect on intention we conducted mediation analyses. [M]ediation can be said to occur when four conditions are satisfied: (1) variation in the independent measure (e.g. altruism) accounts for significant variance in the dependent measure (e.g. intention); (2) variation in the independent measure accounts for significant variance in the mediator (e.g. moral norm); (3) variation in the mediator accounts for variance in the dependent measure while controlling for the influence of independent measure; and (4) the significant effect of the independent

(*continued*)

EXCERPT 16.20 • *(continued)*

measure on the dependent measure is significantly reduced after controlling for the effects of the mediator. . . . Regression analyses showed (1) an effect of altruism on intention, $B = 0.21, t(677) = 3.22, p < .001$, (2) an effect of altruism on moral norm, $B = 0.42, t(677) = 6.26, p < .001$, (3) an effect of moral norm on intention, $B = 0.44, t(677) = 13.13, p < .001$, and (4), the effect of altruism on intention was no longer significant (85% reduction), $B = 0.03, t(677) = 0.43, p = .67$, after including moral norm as additional predictor. This mediation effect was statistically significant, Sobel's $Z = 5.63, p < .001$.

Source: Lemmens, K. P. H., Abraham, C., Ruiter, R. A. C., Veldhuizen, I. J. T., Dehing, C. J. G., Bos, A. E. R., et al. (2009). Modelling antecedents of blood donation motivation among non-donors of varying age and education. *British Journal of Psychology, 100*(1), 71–90.

involving 687 Dutch residents who had never donated blood, even though they were eligible to do so. The researchers hypothesized that their participants' level of intention to donate was influenced by their sense of altruism. However, they also hypothesized that the variable of *norms* (i.e., the degree to which family members and friends donate blood and encourage others to do so) would functioned as a mediator. Examine this excerpt carefully to see the typical approach to a mediated multiple regression.

In the last sentence of Excerpt 16.20, reference is made to the **Sobel test,** one option for seeing if the reduction in the two involved regression coefficients is statistically significant. Another option for making this kind of test is **bootstrapping,** a procedure that can be used in many statistical situations. If bootstrapping is applied in this regression context, a computer is used to develop a sampling distribution (to evaluate the drop in the regression coefficient) by extracting many, many samples from the available data used in the study. This computer-based procedure is considered by some to be superior to the Sobel test.

Logistic Regression

The final kind of regression considered in this chapter is called **logistic regression.** Originally, only researchers from medical disciplines (especially epidemiology) used this form of regression. More recently, however, logistic regression has been discovered by those who conduct empirical investigations in a wide array of disciplines. Its popularity continues to grow at such a rate that it may soon overtake multiple regression and become the most frequently used regression tool of all.

Before considering how logistic regression differs from the forms of regression already considered, let's look at their similarities. First, logistic regression deals with

dependent (i.e., outcome or response) variable whereas the others are the independent (predictor or explanatory) variables. Second, the independent variables can be continuous or categorical in nature. Third, the purpose of logistic regression can be either prediction or explanation. Fourth, tests of significance can be and usually are conducted, with these tests targeted either at each individual independent variable or at the combined effectiveness of the full set of independent variables. Finally, logistic regression can be conducted in a simultaneous, stepwise, or hierarchical manner depending on the timing of and reasons for independent variables entering the equation.

There are, of course, important differences between logistic regression, on the one hand, and either bivariate or multiple regression, on the other hand. These differences are made clear in the next three sections. Logistic regression revolves around a core concept called the **odds ratio** that was not considered earlier in the chapter because it is not a feature of either bivariate or multiple regression. Before looking at this new concept, we must focus our attention on the kinds of data that go into a logistic regression and also the general reasons for using this kind of statistical tool.

Variables

As does any bivariate or multiple regression, logistic regression always involves two main kinds of variables. These are the study's *dependent* and *independent* variables. In the typical logistic regression (as in some applications of multiple regression), a subset of the independent variables is included for control purposes, with the label *control* (or *covariate*) designating any such variable. Data on these three variables constitute the only ingredients that go into the normal logistic regression, and the results of such analyses are inextricably tied, on a conceptual level, to these three kinds of variables. For these reasons, it is important for us to begin with a careful consideration of the logistic regression's constituent parts.

In any logistic regression, as in any bivariate or multiple regression, there is one and only one dependent variable. Here, however, the dependent variable is categorical. Although the dependent variable can have three or more categories, thus making the logistic regression multinomial in nature, we consider here only situations where the dependent variable is dichotomous. Examples of such variables used in recent studies include whether or not a person survives open heart surgery, whether an elderly and ill married person considers his or her spouse to be the primary caregiver, whether a young child chronically suffers from nightly episodes of coughing, and whether an adolescent drinks at least eight ounces of milk a day. As illustrated by these examples, dichotomous dependent variables in logistic regressions can represent either true or artificial dichotomies. Either way, our focus is on what sometimes is referred to as *binary logistic regression.*

In addition to the dependent variable, at least one independent variable is involved in any logistic regression. Almost always, two or more such variables are involved. As in multiple regression, these variables can be either quantitative or qualitative in nature. If of the former variety, scores on the independent variable are

construed to represent points along a numerical continuum. With qualitative independent variables, however, scores carry no numerical meaning and only serve the purpose of indicating group membership. In any given logistic regression, the independent variables can be all quantitative, all qualitative, or some of each. Moreover, independent variables can be used individually or jointly as an interaction.

When using logistic regression, applied researchers usually collect data on several independent variables, not just one. In the study alluded to earlier in which the dependent variable dealt with nighttime coughing among preschool children, the independent variables dealt with the child's sex and birth weight, the possible presence of pets and dampness problems in the home, whether the parents smoked or had asthma, and whether the child attended a day care center. It is not unusual to see this many independent variables utilized within logistic regression studies.

As indicated previously, some of the independent variables in a typical logistic regression are control variables. Such variables are included so the researcher can assess the "pure" relationship between the remaining independent variable(s) and the dependent variable. In a very real sense, control variables are included because of suspected confounding that would muddy the water if the connection between the independent and dependent variables were examined directly.

In any given logistic regression wherein control is being exercised by means of the inclusion of covariate variables, it may be that only one such variable is involved, or that two or three are used, or that all but one of the independent variables are covariates. It all depends, of course, on the study's purpose and the researcher's ability to identify and measure potentially confounding variables. In the study concerned with preschoolers and chronic coughing at night, all but one of the independent variables were included for control purposes; by so doing, the researchers considered themselves better able to examine the direct influence of day care versus home care on respiratory symptoms.

In Excerpt 16.21, we see a case in which the three kinds of variables of a typical logistic regression are clearly identified. It is worth the time to read this excerpt closely with an eye toward noting the nature and number of these three kinds of variables.

EXCERPT 16.21 • *Dependent, Independent, and Control Variables*

> Logistic regression was used to test the effects of the independent variables while controlling for relevant covariates. . . . The dependent variable [was] whether women completed the 30-day residential treatment program. . . . The independent variables were material and emotional support [from family and friends]. The following categorical demographic variables were included as control variables: marital status, education, drug treatment history, drug use in the past 30 days, ethnicity, and having children.
>
> *Source:* Lewandowski, C. A., & Hill, T. J. (2009). The impact of emotional and material social support on women's drug treatment completion. *Health & Social Work, 34*(3), 213–221.

Many logistic regression studies are like the one associated with Excerpt 16.21 in that they involve one dichotomous dependent variable, multiple independent variables, and multiple control variables. In some logistic regression studies, there are multiple independent variables and a single control variable. Or, there might be a single independent variable combined with several control variables. It all depends on the goals of the investigation and the researcher's ability to collect data on independent and control variables that are logically related to the dependent variable.

Objectives of a Logistic Regression

Earlier in this chapter, I pointed out that researchers use bivariate and multiple regression in order to achieve one of two main objectives: explanation or prediction. So it is with logistic regression. In many studies, the focus is on the noncontrol independent variables, with the goal being to identify the extent to which each one plays a role in explaining why people have the status they do on the dichotomous dependent variable. In other studies, the focus is primarily on the dependent variable and how to predict whether people end up in one or the other of the two categories of that outcome variable.

In Excerpt 16.22, we see a case in which logistic regression was used for predictive purposes. In the final sentence of this excerpt, the researchers point out which of their independent variables helped predict relapses among patients afflicted with schizophrenia.

EXCERPT 16.22 • *Logistic Regression and Prediction*

> Schizophrenia is a severe and chronic mental illness characterized by recurring relapses. . . . To determine predictors of relapse during the 1-year study period, a stepwise logistic regression analyses was conducted. . . . [T]his study identified a small set of variables that help predict subsequent relapse in the usual treatment of schizophrenia, demonstrating the predictive value of prior relapse as a robust marker, along with prior medication nonadherence, younger age at illness onset, having health insurance, and poorer level of functioning.
>
> *Source:* Ascher-Svanum, H., Baojin, Z., Faries, D. E., Salkever, D., Slade, E. P., Xiaomei, P., et al. (2010). The cost of relapse and the predictors of relapse in the treatment of schizophrenia. *BMC Psychiatry, 10,* 1–7.

Odds, Odds Ratios, and Adjusted Odds Ratios

Because the concept of *odds* is so important to logistic regression, let's consider a simple example that illustrates what this word does (and does not) mean. Suppose you have a pair of dice that are known to be fair and not loaded. If you were to roll these two little cubes and then look to see if you rolled a pair (two of the same

number), the answer is either yes or no. Altogether, there are 36 combinations of how the dice might end up, with six of these being pairs. On any roll, therefore, the probability of getting a pair is 6/36, or .167. (Naturally, the probability of not getting a pair is .833.) Clearly, it is more likely that you will fail than succeed in your effort to roll a pair. However, we can be even more precise than that. We could say that the odds are 5 to 1 against you, meaning that you are five times more likely to roll a nonpair than a pair.

Most researchers utilize logistic regression so they can discuss the explanatory or predictive power of each independent variable using the concept of odds. They want to be able to say, for example, that people are twice as likely to end up one way on the dependent variable if they have a particular standing on the independent variable being considered. For example, in a hypothetical study focused on the possible impact of car color on auto accidents, the researchers might summarize their findings by saying that "Red cars are three times as likely to be involved in an accident than white cars." Or, in a different study dealing with exercise and injuries, the research report might include a sentence saying that "Adults who stretched before exercising were found to be one-half as likely to incur a muscle cramp as compared with those who did not stretch."

After performing a logistic regression, researchers often cite the **odds ratio** for each independent variable, or at least for the independent variable(s) not being used for control purposes. The odds ratio is sometimes reported as OR, and it is analogous to r^2 in that it measures the strength of association between the independent variable and the study's dependent variable. However, the odds ratio is considered by many people to be a more user-friendly concept than the Pearson-based coefficient of determination. Because the odds ratio is so central to logistic regression, let's pause for a moment to consider what this index means.

Imagine that two very popular TV programs end up going head-to-head against each other in the same time slot on a particular evening. For the sake of our discussion, let's call these programs A and B. Also imagine that we conduct a survey of folks in the middle of this time slot in which we ask each person two questions: (1) What TV show are you now watching? and (2) Are you a male or a female? After eliminating people who either were not watching TV or were watching something other than program A or B, let's suppose we end up with data like that shown in Figure 16.3.

		TV Program Being Watched	
		Program A	Program B
Gender	Male	200	100
	Female	50	150

As you can see, both TV programs were equally popular among the 500 people involved in our study. Each was being watched by 250 of the people we called. Let's now look at how each gender group spread itself out between the two programs. To do this, we will arbitrarily select Program A and then calculate, first for males and then for females, the odds of watching Program A. For males, the odds of watching Program A are 200 ÷ 100 (or 2 to 1); for females, the odds of watching this same program are 50 ÷ 150 (or 1 to 3). If we now take these odds and divide the one for males by the one for females, we obtain the ratio of the odds for gender relative to Program A. This OR is equal to (2 ÷ 1) ÷ (1 ÷ 3), or 6. This result tells us that among our sampled individuals, males are six times more likely to be watching Program A than women. Stated differently, gender (our independent variable) appears to be highly related to which program is watched (our dependent variable).

In our example involving gender and the two TV programs, the odds ratio was easy to compute because there were only two variables involved. As we have seen, however, logistic regression is typically used in situations where there are more than two independent variables. When multiple independent variables are involved, the procedures for computing the odds ratio become quite complex; however, the basic idea of the odds ratio stays the same.

Consider Excerpts 16.23 and 16.24. Notice the phrase "about 30% lower" in the first of these excerpts, and the phrases "four times as likely" and "nearly three times as likely" in the second excerpt. Most people can understand conclusions such as these even though they are unfamiliar with the statistical formulas needed to generate an odds ratio type of conclusion. In addition, I suspect you can see, without difficulty, that whether an odds ratio ends up being greater than 1 or less than 1 is

EXCERPTS 16.23–16.24 • *Odds Ratio and Adjusted Odds Ratio*

The odds of graduation for Hispanics are about 30% lower compared to Whites [odds ratio = 0.66].

Source: Jones, M. T., Barlow, A. E. L, & Villarejo, M. (2010). Importance of undergraduate research for minority persistence and achievement in biology. *Journal of Higher Education, 81*(1), 82–115.

There were two important predictors of Emotional Cue Eating in this study: women were four times as likely [AOR = 4.0] to be emotional eaters, and those whose families offered food to comfort were nearly three times as likely (AOR = 2.6) to be emotional eaters.

Source: Brown, S. L., Schiraldi, G. R., & Wrobleski, P. P. (2009). Association of eating behaviors and obesity with psychosocial and familial influences. *American Journal of Health Education, 40*(2), 80–89.

quite arbitrary. It all depends on the way the sentence is structured. For example, the researchers who gave us Excerpt 16.23 would have presented an OR of 1.52 in the final sentence (and they would have said "about 50% higher") if the position of the words *Hispanics* and *Whites* had been reversed.

When the odds ratio is computed for a variable *without* considering the other independent variables involved in the study, it can be conceptualized as having come from a bivariate analysis. Such an OR is said to be a *crude* or *unadjusted odds ratio*. If, as is usually the case, the OR for a particular variable is computed in such a way that it takes into consideration the other independent variable(s), then it is referred to as an **adjusted odds ratio.** By considering all independent variables jointly so as to assess their connections to the dependent variable, researchers often say that they are performing a *multivariate analysis*.

To see an example of an adjusted odds ratio, consider once again Excerpt 16.24. Notice that the letters AOR, the abbreviation for this kind of odds ratio, appears twice in the excerpt. This study's other predictor variables (besides gender and whether a family offered food to comfort) do not appear in this excerpt, but there were many. These included ethnicity, a variety of personality variables, and several family characteristics.

Tests of Significance

When using logistic regression, researchers usually conduct tests of significance. As in multiple regression, such tests can be focused on the odds ratios (which are like regression coefficients) associated with individual independent variables or on the full regression equation. Whereas tests on the full regression equation typically represent the most important test in multiple regression, tests on the odds ratios in logistic regression are considered to be the most critical tests the researcher can perform.

When the odds ratio or adjusted odds ratio associated with an independent variable is tested, the null hypothesis says that the population counterpart to the sample-based OR is equal to 1. If the null hypothesis were true (with OR = 1), it means that membership in the two different categories of the dependent variable is unrelated to the independent variable under consideration. For this null hypothesis to be rejected, the sample value of OR must deviate from 1 further than would be expected by chance.

Researchers typically use one of two approaches when they want to test an odds ratio or an adjusted odds ratio. One approach involves setting up a null hypothesis, selecting a level of significance, and then evaluating the H_0 either by comparing a test statistic against a critical value or by comparing the data-based p against α. The second way a researcher can test an odds ratio is through the use of a confidence interval (CI). The CI rule-of-thumb for deciding whether to reject or retain the null hypothesis is the same as when CIs are used to test means, rs, the

pinpoint number in the null hypothesis, the null hypothesis is retained; otherwise, H_0 is rejected. Excerpts 16.25 and 16.26 illustrate these two ways to test an odds ratio or an adjusted odds ratio.

EXCERPTS 16.25–16.26 • *Testing an OR or an AOR for significance*

Logistic regression was also used to explore what baseline variables could predict psychological distress at six months (predictive model). . . . Stroke severity (Wald's $\chi^2 = 7.95, P < 0.01$) was a significant predictor of psychological distress [OR = 1.24].

Source: Hilari, K., Northcott, S., Roy, P., Marshall, J., Wiggins, R. D., Chataway, J., et al. (2010). Psychological distress after stroke and aphasia: The first six months. *Clinical Rehabilitation, 24*(2), 181–190.

The goal of this study was to determine whether residential exposure to vehicular traffic was associated with SAB [spontaneous abortion]. . . . SAB was examined in relation to the traffic exposure measures using logistic regression adjusting for a number of demographic and lifestyle variables. ... Among women who were nonsmokers, significantly increased odds of SAB were observed in the highest traffic exposure group (AOR = 1.47; 95% CI, 1.07–2.04).

Source: Green, R. S., Malig, B., Windham, G. C., Fenster, L., Ostro, B., & Swan, S. (2009). Residential exposure to traffic and spontaneous abortion. *Environmental Health Perspectives, 117*(12), 1939–1944.

Notice in Excerpt 16.25 that a **Wald test** was used to see if the odds ratio was statistically significant. This test is highly analogous to the *t*-test in multiple regression that is used to see if a beta weight is statistically significant. These two tests are only analogous, however, for they differ not only in terms of the null hypothesis but also in the kinds of calculated and critical values used to test the H_0. As illustrated in Excerpt 16.25, the Wald test is tied to a theoretical distribution symbolized by χ^2 rather than *t*. (This is the *chi-square distribution.*)

Excerpt 16.26 illustrates how a CI can be used to test an odds ratio. Take the time to look closely at this excerpt's CI, note its ends, and then recall that the pinpoint number in the null hypothesis being tested is 1.0. I hope that you see why the researchers' AOR of 1.47 was reported to reflect "significantly increased odds of SAB" for one of the study's comparison groups.

As indicated previously, it is possible in logistic regression to assess whether the collection of independent and control variables do a better-than-chance job of accounting for the status of people on the dependent variable. Three popular procedures exist for doing this. One approach involves setting up and testing a single null hypothesis concerning the full equation, with a data-based *p*-level compared

against an alpha level to see if the independent variables, as a unified set, are linked to the dependent variable. A second approach involves computing **Nagelkerke's** R^2, an index that is highly analogous to the R^2 used in multiple regression.[10] A third approach involves determining the **hit rate** to see how successful the set of independent variables are at correctly classifying individuals into the categories of the dependent variable. Some researchers use one of these approaches, others use two, and a few, as illustrated in Excerpt 16.27, use all three.

EXCERPT 16.27 • *Evaluating the Full Logistic Regression Model*

The final logistic regression model was significant, $\chi^2(3, N = 31) = 25.48$, $p < .001$, indicating that combined performance on the three tasks distinguished adults with SLI from those with TL. The model explained 75% (Nagelkerke's R^2) of the variance in language status (i.e., affected or unaffected) and correctly classified 87% of cases, with a sensitivity of 85% and a specificity of 89% where cases with a .50 or greater predicted probability were classified as affected.

Source: Poll, G. H., Betz, S. K., & Miller, C. A. (2010). Identification of clinical markers of specific language impairment in adults. *Journal of Speech, Language, and Hearing Research, 53*(2), 414–429.

Excerpt 16.27 contains two new technical terms: sensitivity and specificity. **Sensitivity** is the hit rate for correctly classifying people as being in the first category—usually the category that has a disease or ailment—of the dependent variable, whereas **specificity** is the hit rate for correctly classifying people into the other category. Both sensitivity and specificity are based on the data used to build the logistic regression model, with each index computed as the percentage of people actually in a given category who are correctly classified as being members of that category.

Final Comments

As we conclude this chapter, we must consider four additional regression-related issues: assumptions, control, practical significance, and the inflated Type I error risk. If you keep these issues in mind as you encounter research reports based on bivariate, multiple, and logistic regression, you will be in a far better position to both decipher and critique such reports.

[10]Nagelkerke's R^2 sometimes is referred to as an *approximate* or *pseudo-measure of explained variability* because of the way it is computed. (It begins with the Cox and Snell's measure of R^2, which itself is only an

All three forms of regression considered in this chapter carry with them *underlying assumptions*. If these assumptions are violated, regression results can be misleading. Therefore, give credit to researchers who indicate that they attended to their regression's assumptions before analyzing their data to get answers to their research questions. Excerpt 16.28 illustrates this good practice.

EXCERPT 16.28 • *Concern for Assumptions*

> The assumptions of multiple regression were examined using a series of diagnostic graphs and tests for outliers, normality of residuals, homoscedasticity, multicollinearity, linearity, model specification, and independence. . . . The regression models provided an acceptable description of the data because no violations of the assumptions were observed.
>
> *Source:* Cowley, P. M., Ploutz-Snyder, L. L., Baynard, T., Heffernan, K., Jae, S. Y., Hsu, S., et al. (2010). Physical fitness predicts functional tasks in individuals with Down Syndrome. *Medicine and Science in Sports and Exercise, 42*(2), 388–393.

Two important terms in Excerpt 16.28 have not been discussed previously in this book: multicollinearity and model specification. **Multicollinearity** exists if two or more independent variables are too highly correlated with each other. This undesirable situation causes inferences about individual predictor variables to be untrustworthy. Accordingly, regression assumes that multicollinearity *does not* exist. **Model specification** is concerned with the researcher's decision regarding which variables to include in the regression model. If important variables are overlooked or if irrelevant variables are included, the regression model is said to be *misspecified.* Understandably, the assumption here is that the model has been specified properly, thereby avoiding the problem of misspecification.

In the discussions of both hierarchical multiple regression and logistic regression, we saw that researchers often incorporate control or covariate variables into their analyses. Try to remember that such *control* is very likely to be less than optimal for three reasons: First, one or more important confounding variables might be overlooked. Second, potential confounding variables that *are* measured are likely to be measured with instruments possessing less than perfect reliability. Finally, recall that the analysis of covariance undercorrects when used with nonrandom groups that come from populations that differ on the covariate variable(s). Regression suffers from this same undesirable characteristic.

My next concern relates to *the distinction between statistical significance and practical significance.* We considered this issue in connection with tests focused on means and *r*s, and it is just as relevant to the various inferential tests used by researchers within regression analyses. In Excerpts 16.29 and 16.30, we see two cases in which researchers attended to the important distinction between useful and

EXCERPTS 16.29–16.30 • *Practical versus Statistical Significance*

An a priori power analysis was conducted [so as] to determine the needed sample size to answer the research questions. For multiple linear regression analysis, with a significance level of 0.05, 80% power, a total of 15 predictor variables, and an estimated moderate effect size ($R^2 = 0.25$), 70 subjects were needed. A conservative R^2 was estimated from a study using the Quality of Life Index – Dialysis version with persons on hemodialysis, in which the R^2 was reported as 0.28.

Source: Kring, D. L., & Crane, P. B. (2009). Factors affecting quality of life in persons on hemodialysis. *Nephrology Nursing Journal, 36*(1), 15–55.

A doctoral degree in counselor education was the second [significant] contributor to the regression equation. Nevertheless, with an [increase in] R^2 of .04 when this variable is added, and a resulting effect size similar to that of the equation with just one predictor variable, it does not contribute to the prediction ability in a highly meaningful way. Thus, it is important to note that a doctoral degree contributes to cognitive complexity, but the relatively small change in the R^2 means that it does not have a high degree of practical significance.

Source: Granello, D. H. (2010). Cognitive complexity among practicing counselors: How thinking changes with experience. *Journal of Counseling & Development, 88*(1), 92–100.

trivial findings. The researchers associated with the first excerpt set a good example by conducting an a priori power analysis in the planning stage of their investigation. The researchers associated with the second of these excerpts deserve high praise for realizing (and for warning their readers) that inferential tests can yield results that are statistically significant without being important in a practical manner.

In many research reports, researchers make a big deal about a finding that seems small and of little importance. Perhaps such researchers are unaware of the important distinction between practical and statistical significance, or it may be that they know about this distinction but prefer not to mention it due to a realization that their statistically significant results do not matter very much. Either way, it is important that *you* keep this distinction in mind whenever you are on the receiving end of a research report. Remember, you have the right to evaluate a statistical finding as having little or no meaningfulness after you examine the research report's summary statistics, and you can draw such a conclusion even if your opinion is at odds with those of the researcher's.

As we have seen in the excerpts of previous chapters, competent researchers are sensitive to the inflated Type I error risk that occurs if a given level of significance, say .05, is used multiple times within the same study when different null hypotheses are tested. Give credit to researchers when they apply the Bonferroni