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.**

As you will see, three different kinds of regression will be 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, on the other hand, 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! That's 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 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 will be trying to accomplish one or the other of two goals. These two goals involve prediction on the one hand and explanation on the other.

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 GPA would 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 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 while 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 the reading and spelling performance of young children, two markers of early literacy competence. In Excerpt 16.2, the goal was explanation, not prediction. Here, the researchers wanted to know which factors explain why some urban-dwelling African-American women with osteoarthritis (OA) or rheumatoid arthritis (RA) exercise more than others.

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's 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 doesn't 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

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

It is important to gather data on the effectiveness of the ELS [Early Literacy Support] in relation to other forms of intervention to enable schools to identify efficient and cost effective ways of preventing or overcoming difficulties in early literacy. In addition, an understanding of what predicts individual differences in children's responsiveness to the strategy is required so that modifications can be made to the curriculum in the light of children's special needs. . . . The correlation matrix [not shown here] suggests that reading and spelling at T2 are significantly predicted by all variables apart from receptive vocabulary at T1. However, given the association between predictor variables, the matrix does not allow us to identify the extent to which variables contribute unique variance to predicting reading and spelling. In order to do that, it is necessary to conduct regression analyses.

Source: Hatcher, P. J., Goetz, K., Snowling, M. J., Hulme, C., Gibbs, S., and Smith, G. (2006). Evidence for the effectiveness of the Early Literacy Support programme. *British Journal of Educational Psychology, 76,* pp. 353, 362.

The purpose of this study was to answer the following questions: (1) What factors explain physical activity and exercise behavior in urban adults with OA and RA who are predominantly African-American and female? and (2) Are the factors that explain physical activity and exercise behavior the same for people with OA and RA in this population? . . . Forward stepwise multiple linear regression was used to build a model of explanatory variables that best explained the variance in physical activity.

Source: Greene, B. L., Haldeman, G. F., Kaminski, A., Neal, K., Sam Lim, S., and Conn, D. L. (2006). Factors affecting physical activity behavior in urban adults with arthritis who are predominantly African-American and female. *Physical Therapy, 86*(4), pp. 512, 514.

---

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 while 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]

[1]The terms **criterion variable, outcome variable,** and **response variable** are synonymous with the term **dependent variable,** while 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 will be the dependent variable whereas the variable(s) appearing in the second blank will be the independent variable(s).

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 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.

EXCERPTS 16.3–16.4 • *Dependent and Independent Variables*

First, the subjects were divided into three groups, according to their resource levels (high, medium, or low), and then bivariate regressions were calculated for each resource level, with distress as the independent variable and suicidal ideation as the dependent variable.

Source: Lieberman, Z., Solomon, Z., and Ginzburg, K. (2005). Suicidal ideation among young adults: Effects of perceived social support, self-esteem, and adjustment. *Journal of Loss & Trauma, 10*(2), pp. 174–175.

To compare the predictive power of self-efficacy and SARS fear [as measured by the SARS Fear Scale subtests], we conducted a multiple regression analysis with CIES–R total as the dependent variable and perceived self-efficacy, SFS infection, SFS insecurity, and SFS instability as independent variables.

Source: Ho, S. M. Y., Kwong-Lo, R. S. Y., Mak, C. W. Y., and Wong, J. S. (2005). Fear of severe acute respiratory syndrome (SARS) among health care workers. *Journal of Consulting and Clinical Psychology, 73*(2), p. 347.

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, you will see that 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 will 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 will now encounter provide a foundation for the other two kinds of regression to be considered later in the chapter.

Bivariate Regression

Research Navigator.com
Bivariate regression

The simplest kind of regression analysis is called **bivariate regression.** First, we need to clarify the purpose of and the data needed for this kind of regression. Then, we will consider scatter diagrams, lines of best fit, and prediction equations. Finally, we will 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 will serve as the dependent variable while 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 30-year-old weight lifter, has been plagued by a shoulder injury that for months has failed to respond to nonsurgical treatment. Consequently, arthroscopic surgery is scheduled to repair Sam's bad shoulder. Even though he knows that arthroscopic procedures usually permit a rapid return to normal activity, he would like to know how long he'll be out of commission following surgery. His presurgery question to the doctor is short and sweet: "When will I be able to lift 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 weight lifters following arthroscopic shoulder surgery, that's 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'll have to interrupt his training. If Sam's doctor has seen the results of a recent study dealing with weight lifters who had arthroscopic shoulder 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 actual study conducted with people like Sam, there were 10 weight lifters who had shoulder injuries. Although data on several variables were collected in this real investigation, let's consider the data on just two: age and number of postsurgical days of down time. Excerpt 16.5 presents the data on these two variables.

EXCERPT 16.5 • *Data on Two Variables*

TABLE 1 ***Age and Postsurgical Time Away from Sport for 10 Weight Lifters***

Patient	*Age*	*Return to sport (days)*
1	33	6
2	31	4
3	32	4
4	28	1
5	33	3
6	26	3
7	34	4
8	32	2
9	28	3
10	27	2

Source: Auge, W. K., and Fischer, R. A. (1998). Arthroscopic distal clavicle resection for isolated atraumatic osteolysis in weight lifters. *American Journal of Sports Medicine, 26*(2), p. 191. (Note: Table 1 was modified slightly for presentation here.)

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 Excerpt 16.5. There are 10 dots in this "picture," each positioned so as to reveal the age and postsurgical convalescent time for one of the weight lifters.

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 weight lifters who are similar to those who supplied the data we are currently considering. Age, on the other hand, is positioned on the abscissa because it

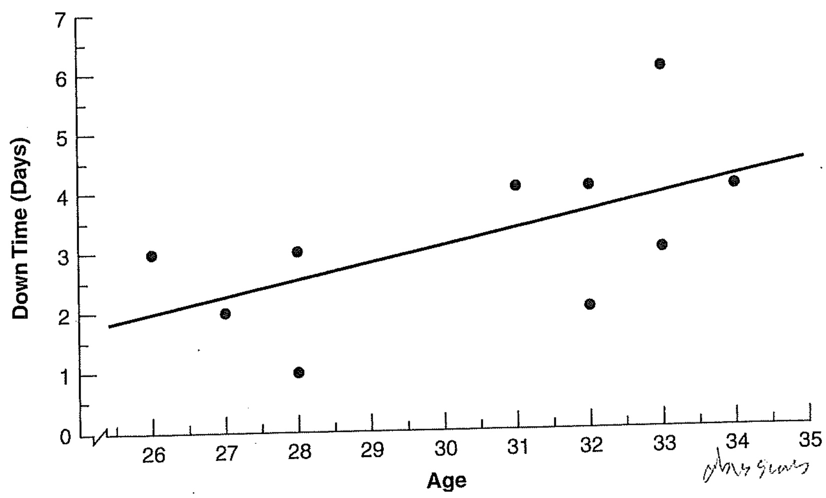

FIGURE 16.1 ***Regression Analysis Using Data in Excerpt 16.5***

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 will function as the tool our hypothetical doctor will use in order to predict how long Sam will have to refrain from lifting. 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 need to 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 30 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 three days.

[3]Since 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 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 line that could be drawn through the scatter diagram's data points.

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 two days if he had been 26 years old, or four days if he had been 34. These alternative predictions for a younger or older 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.

Research Navigator.com

Regression equation

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:

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

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 Excerpt 16.5 (and Figure 16.1), the regression equation turns out like this:

$$Y' = -5.05 + (.27)X.$$

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. When we do this, we find that $Y' = 3.05$. This is the predicted down time (in days) for Sam. The fact that this value is very similar to what we predicted earlier (when we took a trip through the scatter diagram) should not be at all surprising. That's because the regression equation is nothing more than a precise mechanism for telling us where we'll 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 and regression lines appear only rarely in research reports, regression equations show up quite frequently. In Excerpt 16.6, we see a case in point. As you can see from this passage, the regression equation was built so as to predict height from arm span. In this study, each person's height and arm span was measured in centimeters. Thus, the two numbers in the regression formula (0.87 and 20.54) must be interpreted as being in the scale of centimeters, not inches.

Let's use the regression equation presented in Excerpt 16.6 to make some height predictions. When I measure my own arm span, I find that it is equal to 175.1 cm. Using the regression equation, I find that my predicted height is equal to 20.54 + 0.87(175.1) = 172.88 cm. This predicted value turns out to be extremely close to my actual height of 172.70 cm. Now it's your turn. Get a metric measuring device, measure your arm span, and then use the regression formula in Excerpt 16.6 to calculate your predicted height. After doing this, check to see how closely your

EXCERPT 16.6 • *The Regression Equation in Bivariate Regression*

The purpose of this study is to determine the accuracy of arm span as a measure of height in young and middle-age adults. . . . A convenience sample of 83 subjects was studied. Subjects were between the ages of 20 and 61 years, with a mean age of 41.63 years ($SD = 11.10$). Fifty-seven (69%) were women, and 26 (31%) were men. . . . The first analysis was a simple regression of height on arm span to determine how well arm span, alone, predicted height. The prediction equation is as follows: Height = 0.87 (arm span) + 20.54.

Source: Brown, J. K., Whittemore, K. T., and Knapp, T. R. (2000). Is arm span an accurate measure of height in young and middle-age adults? *Clinical Nursing Research, 9*(1), pp. 90–91.

predicted height matches your actual height. You may be surprised (as I was) at how closely arm span predicts height!

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 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 is used in place of b (and is called a **beta weight** rather than a regression coefficient).

Research Navigator.com
Beta weight

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

When used for predictive purposes, the regression equation has the form $Y' = a + bX$. 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 $b = 0$.

In Excerpt 16.6, the constant in the regression equation for the 83 people was equal to 20.54. That is not a very realistic number, for it indicates the predicted height for a person with zero arm span. Likewise, the value of a for the regression line in Figure 16.1 is equal to -5.05. This number is nonsensical, of course, for it indicates the predicted down time following surgery for a weight lifter whose age is 0! Clearly, *a* may be totally devoid of meaning within the context of a study's independent and dependent variables. Nevertheless, it has an unambiguous and not-so-nonsensical meaning within a scatter diagram, for *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 .27. This means that the predicted down time for our hypothetical patient Sam would be about one-fourth of a day longer if the surgery is put off a year.

When researchers use bivariate regression, they sometimes will focus on either *b* or β more than anything else. Consider, for example, Excerpt 16.7. In the study associated with this excerpt, 50 individuals with multiple sclerosis were measured on three personality variables: existential (i.e., nonreligious) well-being (as measured by the EWB), perceived illness uncertainty (as measured by the MUIS), and psychosocial adjustment to illness (as measured by the PAIS-T). After dividing the full group of patients into high and low subgroups based on EWB scores, the researchers did a bivariate regression within each subgroup to investigate the connection between scores on the MUIS and PAIS-T. Notice how the researchers focused their attention on the beta weights when comparing the two subgroups of patients.

EXCERPT 16.7 • *Focusing on the Regression Coefficient*

Scores on the EWB were first dichotomized into two groups around the distribution mean (high EWB > 46, low EWB < 46). [Bivariate] regression analyses for MUIS and PAIS-T scores were then conducted for each group (high vs. low) separately. Inspecting slope . . . indicated that for the high-EWB group, increase in MUIS scores resulted in no change to PAIS-T scores ($\beta = .015$). However, for those in the low-EWB group, increase in MUIS scores resulted in decrease in PAIS-T scores ($\beta = -.30$). In other words, whereas no relationship was found between uncertainty and overall psychosocial adjustment in participants who scored high on EWB, increased uncertainty was associated with decreased psychosocial distress for those who scored low on EWB.

Source: McNulty, K., Livneh, H., and Wilson, L. M. (2004). Perceived uncertainty, spiritual well-being, and psychosocial adaptation in individuals with multiple sclerosis. *Rehabilitation Psychology, 49*(2), pp. 94, 96.

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 will be high to the extent that the scatter diagram's data points are located close to the regression line. Though 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, for 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 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 16.8, r^2 is usually turned into a percent when it is reported in research reports.

EXCERPT 16.8 • *Variability in the Dependent Variable Explained by Variability in the Independent Variable*

Bivariate regression analysis carried out to model how drug and regimen choice vary with SBP [systolic blood pressure] indicated a moderate linear relationship in the study population: about 15% (Pearson product-moment correlation coefficient $r^2 = 0.148$) of the variance of drug and regimen decision made by the participating physicians were associated with SBP.

Source: Erhun, W. O., Agbani, E. O., and Bolaji, E. E. (2003). Managing hypertension with combination diuretics and methyldopa in Nigerian Blacks at the primary care level. *Clinical Drug Investigation, 23*(9), p. 585.

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 typically considered to be 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.

The most frequently conducted test focuses on r. The null hypothesis in such a test will probably be set up to say that the correlation in the population is equal to 0 (i.e., H_0: $\rho = 0$). This kind of test was discussed in Chapter 9, and the considerations raised there apply equally to tests of r within the context of bivariate regression. In Excerpt 16.9, we see a case in which such a test was performed. Near the end of this excerpt, the researchers present the value of r^2 rather than r. If they had wanted to, they could have reported $r = .32$ in the place where they reported $r^2 = .10$. It doesn't matter, because a test of r is mathematically equivalent to a test of r^2.

EXCERPT 16.9 • *Testing r (or r^2) for Significance in Bivariate Correlation*

To test the first research question, whether perceived support for challenging racism, sexism, and social injustice from key social actors is associated with the reflection component of critical consciousness among urban adolescents, inverted SDO sum scores were regressed on the total perceived support variable (the sum of perceived support for challenging racism, sexism, and social injustice). This simple regression was statistically significant, $F(1, 91) = 9.70, p < .01, r^2 = .10$, indicating that 10% of the reflection component of critical consciousness variance was accounted for by total support. . . .

Source: Diemer, M. A., Kauffman, A., Koenig, N., Trahan, E., and Hsieh, C. (2006). Challenging racism, sexism, and social injustice: Support for urban adolescents' critical consciousness development. *Cultural Diversity & Ethnic Minority Psychology, 12*(3), pp. 448, 452.

In bivariate regression, a test on r is mathematically equivalent to a test on b or β. Therefore, you will never see a case where both r and b (or r and β) are tested, because these two tests would be fully redundant with each other. However, researchers have the freedom to have their test focus on r or b or β. In Excerpt 16.10, we see a case where a team of researchers chose to test β. The null hypothesis in

EXCERPT 16.10 • *Testing the Beta Weight*

A bivariate regression analysis found that age was a small, but significant, predictor of PTGI scores ($\beta = .17, p < .05$).

Source: Morris, B. A., Shakespeare-Finch, J., Rieck, M., and Newbery, J. (2005). Multidimensional nature of posttraumatic growth in an Australian population. *Journal of Traumatic Stress, 18*(5), p. 581.

this kind of test says that the population value of the 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.

Multiple Regression

Research Navigator.c⊕m
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 measure of the extent to which variability among the scores on the dependent variable has been explained or accounted for.

Though 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 will be 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

$$Y' = a + b_1 \cdot X_1 + b_2 \cdot X_2$$

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