TABLE 11.7 SPSS output for repeated-measures ANOVA for work problems

	Grade	Mean	Std. Deviation	N
Mastery Emphasis, Fall, Year 1	9	3.67	.65	301
	11	3.52	.80	253
	Total	3.60	.73	554
Mastery Emphasis, Spring, Year 1	9	3.41	.80	301
	11	3.49	.79	253
	Total	3.45	.80	554
Mastery Emphasis, Fall, Year 2	10	3.58	.73	301
	12	3.55	.75	253
	Total	3.57	.74	554
Mastery Emphasis, Spring, Year 2	10	3.38	.81	301
	12	3.29	.86	253
	Total	3.34	.83	554

Tests for Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	26751.461	1	26751.461	19257.280	.000	.972
Grade Level	1.164	1	1.164	.838	.360	.002
Error	766.817	552	1.389			

Tests Involving Within-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Mastery Emphasis	23.454	3	7.818	23.254	.000	.040
Mastery Emphasis * Grade Level	3.828	3	1.276	3.796	.010	.007
Error (Mastery Emphasis)	556.738	1656	.336			

6. Using the information from the graphs that you created and the statistics about the main effects and the interaction effects, write a summary statement about what you now know about how students' perceptions of their teachers' emphasis on Mastery (i.e., learning and understanding) changed over a two-year period during high school, and how those changes are similar or different for early and later grade levels in high school.

For answers to these work problems, and for additional work problems, please refer to the website that accompanies this book.

CHAPTER 12
Correlation

In several of the previous chapters, we examined statistics and parameters that describe a single variable at a time, such as the mean, standard deviation, *z* scores, and standard errors. Although such single-variable statistics are important, researchers are often interested in examining the relations among two or more variables. One of the most basic measures of the association among variables, and a foundational statistic for several more complex statistics, is the **correlation coefficient**. Although there are a number of different types of correlation coefficients, the most commonly used in social science research is the **Pearson product-moment correlation coefficient**. Most of this chapter is devoted to understanding this statistic, with a brief description of three other types of correlations: the **point-biserial coefficient**, the **Spearman rho coefficient**, and the **phi coefficient**.

When to Use Correlation and What it Tells Us

Researchers compute correlation coefficients when they want to know how two variables are related to each other. For a Pearson product-moment correlation, both of the variables must be measured on an interval or ratio scale and are known as **continuous variables**. For example, suppose I want to know whether there is a relationship between the amount of time students spend studying for an exam and their scores on the exam. I suspect that the more hours students spend studying, the higher their scores will be on the exam. But I also suspect that there is not a perfect correspondence between time spent studying and test scores. Some students will probably get low scores on the exam even if they study for a long time, simply because they may have a hard time understanding the material. Indeed, there will probably be a number of students who spend an inordinately long period of time studying for the test precisely *because* they are having trouble understanding the material. On the other hand, there will probably be some students who do very well on the test without spending very much time studying. Despite these "exceptions" to my rule, I still hypothesize that, in general, as the amount of time spent studying increases, so do students' scores on the exam.

There are two fundamental characteristics of correlation coefficients researchers care about. The first of these is the **direction** of the correlation coefficient. Correlation coefficients can be either positive or negative. A **positive correlation** indicates that the values on the two variables being analyzed move in the same direction, and they are associated with each other in a predictable manner. That is, as scores on one variable go up, scores on the other variable go up as well, in general. Similarly, as scores on one variable go down, scores on the other variable go down. Returning to my earlier example, if there is a positive correlation between the amount of time students spend studying and their test scores, I can tell that, in general, the more time students spend studying, the higher their scores are on the test. This is *equivalent* to saying that the *less* time they spend studying, the *lower* their scores are on the test. Both of these represent a *positive* correlation between time spent studying and test scores. (*Note*: I keep saying "in general" because

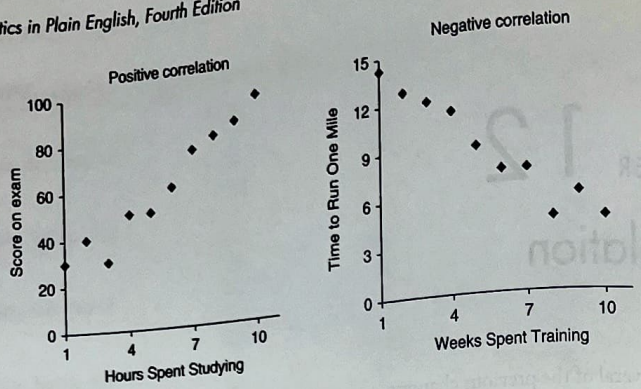

FIGURE 12.1 Examples of positive and negative correlations.

it is important to note that the presence of a correlation between two variables does *not* mean that this relationship holds true for each member of the sample or population. Rather, it means that, in general, there is a relationship of a given direction and strength between two variables in the sample or population.)

A **negative correlation** indicates that the values on the two variables being analyzed move in *opposite* directions. That is, as scores on one variable go up, scores on the other variable go down, and vice versa. For example, the more time runners spend training, the less time it takes them to run a mile. As training time increases, mile time decreases, in general. Similarly, with a negative correlation I would also conclude that the *less* time runners spend training, the *longer* it takes them to run a mile. These positive and negative correlations are represented by the **scattergrams** (also known as **scatterplots** or **scatter graphs**) in Figure 12.1. Scattergrams are simply graphs that indicate the scores of each case in a sample simultaneously on two variables. For example, in the positive correlation scattergram in Figure 12.1, the first case in the sample studied for 1 hour and got a score of 30 on the exam. The second case studied for 2 hours and scored 40 on the exam. Notice that each dot on the scattergram indicates an individual's score on two variables simultaneously.

The second fundamental characteristic of correlation coefficients is the **strength** or **magnitude** of the relationship. The magnitude of the relationship between two variables is indicated by the numerical value of ***r***, regardless of whether the value is positive or negative. Correlation coefficients range in strength from −1.00 to +1.00. A correlation coefficient of .00 indicates that there is no relationship between the two variables being examined. That is, scores on one of the variables are not related in any meaningful way to scores on the second variable. The closer the correlation coefficient is to either −1.00 or +1.00, the stronger the relationship is between the two variables. A **perfect negative correlation** of −1.00 indicates that for every member of the sample or population, a higher score on one variable is related to a lower score on the other variable. A **perfect positive correlation** of +1.00 reveals that for every member of the sample or population, a higher score on one variable is related to a higher score on the other variable.

Perfect correlations are never found in actual social science research. Generally, correlation coefficients stay between −.70 and +.70. Some textbook authors suggest that correlation coefficients between −.20 and +.20 indicate a weak association between two variables; those between

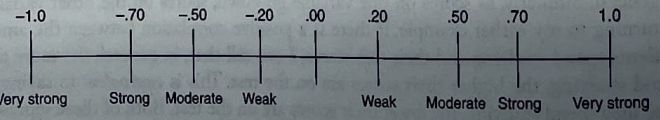

FIGURE 12.2 General guide to the strength of correlation coefficients.

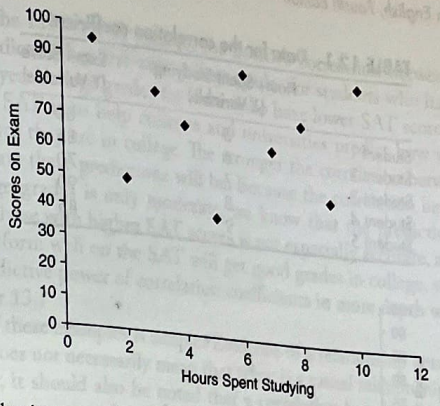

FIGURE 12.3 Scatterplot showing no correlation between hours spent studying and exam scores.

.20 and .50 (either positive or negative) represent a moderate association, and those larger than .50 (either positive or negative) represent a strong association. (See Figure 12.2 for a general guide to interpreting the strength of correlation coefficients.) These general rules of thumb for judging the relevance of correlation coefficients must be taken with a grain of salt. For example, even a "small" correlation between alcohol consumption and liver disease (e.g., +.15) is important, whereas a strong correlation between how much children like vanilla and chocolate ice cream (e.g., +.70) may not be so important.

The scattergrams presented in Figure 12.1 represent very strong positive and negative correlations (r = .97 and r = −.97 for the positive and negative correlations, respectively; r is the symbol for the sample Pearson correlation coefficient). In Figure 12.3, a scattergram representing virtually no correlation between the number of hours spent studying and scores on the exam is presented. Notice that there is no discernible pattern between the scores on the two variables. In other words, the data presented in Figure 12.3 reveal that it would be virtually impossible to predict an individual's test score simply by knowing how many hours the person studied for the exam.

Pearson Correlation Coefficients in Depth

The first step in understanding what the Pearson correlation coefficient is, and what it tells us, is to notice that we are concerned with a sample's scores on *two* variables at the same time. Returning to our previous example of study time and test scores, suppose that we randomly select a sample of five students and measure the time they spent studying for the exam and their exam scores. The data are presented in Table 12.1 (with a scattergram in Figure 12.4).

For these data to be used in a correlation analysis, it is critical that the scores on the two variables are *paired*. That is, for each student in my sample, the score on the *X* variable (hours spent studying) is paired with his or her own score on the *Y* variable (exam score). If I want to determine the relation between hours spent studying and exam scores, I cannot pair Student 1's hours spent studying with Student 4's test score. I must match each student's score on the *X* variable with his or her own score on the *Y* variable. Once I have done this, I can determine whether hours spent studying is related to exam scores.

What the Correlation Coefficient Does, and Does Not, Tell Us

Correlation coefficients such as the Pearson coefficient are very powerful statistics. They allow us to determine whether the values on one variable are *associated with* the values on a second variable. This can be very useful information, but people, including social scientists, are often

TABLE 12.1 Data for the correlation coefficient

	Hours Spent Studying (X Variable)	Exam Score (Y Variable)
Student 1	5	80
Student 2	6	85
Student 3	7	70
Student 4	8	90
Student 5	9	85

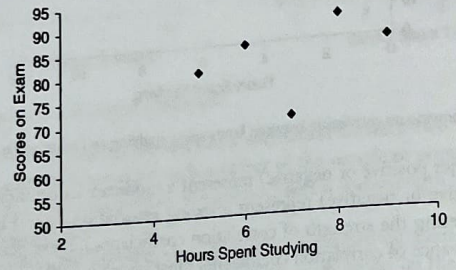

FIGURE 12.4 Scatterplot of data from Table 12.1.

tempted to ascribe more meaning to correlation coefficients than they deserve. Namely, people often confuse the concepts of *correlation* and **causation**. Correlation (co-relation) simply means that variation in the scores on one variable *corresponds* with variation in the scores on a second variable. Causation means that variation in the scores on one variable *causes* or *creates* variation in the scores on a second variable.

When we make the leap from correlation to causation, we may be wrong. As an example, I offer this story, which I heard in my introductory psychology class. As the story goes, one winter shortly after World War II, there was an explosion in the number of storks nesting in some northern European country (I cannot remember which). Approximately nine months later, there was a large jump in the number of babies that were born. Now, the link between storks and babies being what it is, many concluded that this correlation between the number of storks and the number of babies represented a causal relationship. Fortunately, science tells us that babies do not come from storks after all, at least not human babies. However, there is something that storks and babies have in common: both can be "summoned" by cold temperatures and warm fireplaces. It seems that storks like to nest in warm chimneys during cold winters. As it happens, cold winter nights also foster baby-making behavior. The apparent cause-and-effect relationship between storks and babies was in fact caused by a third variable: a cold winter.[1]

For a more serious example, we can look at the relationship between SAT scores and first-year college grade point averages. The correlation between these two variables is about .40. Although these two variables are moderately correlated, it would be difficult to argue that higher SAT scores *cause* higher achievement in the first year of college. Rather, there is probably some other variable, or set of variables, responsible for this relationship. For example, we know that taking a greater number of advanced math courses in high school is associated with higher SAT scores *and* with higher grades in first-year math courses in college.

Although correlations alone cannot tell us whether there is a causal association between two variables, they do allow us to use information from one variable to make **predictions** about scores on a second variable. Returning to our example of the correlation between SAT scores and grades in the first year of college, the positive correlation between these two variables allows us to predict higher first-year college grades for students who have higher SAT scores, and lower first-year college grades for those who have lower SAT scores. Indeed, the primary purpose of the SAT is to help colleges and universities predict how well college applicants will perform once they are in college. The stronger the correlation between the two variables, the more accurate these predictions will be. Because the correlation between SAT scores and first-year college grades is only moderate, we know that the prediction of higher first-year grades among those with higher SAT scores is not especially accurate, and that many students who do not perform well on the SAT will get good grades in college, and vice versa. (We will discuss the predictive power of correlation coefficients in more depth when we discuss regression in Chapter 13.)

The point of these examples is simple: Evidence of a relationship between two variables (i.e., a correlation) does not necessarily mean that there is a causal relationship between the two variables. However, it should also be noted that a correlation between two variables is a *necessary ingredient* of any argument that the two variables are causally related. In other words, I cannot claim that one variable causes another (e.g., that smoking causes cancer) if there is no correlation between smoking and cancer. If I do find a correlation between smoking and cancer, I must rule out other factors before I can conclude that it is smoking that causes cancer.

In addition to the correlation-causation issue, there are a few other important features of correlations worth noting. First, simple Pearson correlations are designed to examine *linear* relations among variables. In other words, they describe *straight* associations among variables. For example, if you find a positive correlation between two variables, you can predict how much the scores in one variable will increase with each corresponding increase in the second variable. But not all relations between variables are linear. For example, there is a **curvilinear** relationship between anxiety and performance on a number of academic and nonacademic behaviors. When taking a math test, for example, a little bit of anxiety may actually help performance. However, once a student becomes too nervous, this anxiety can interfere with performance. We call this a curvilinear relationship because what began as a positive relationship between performance and anxiety at lower levels of anxiety becomes a negative relationship at higher levels of anxiety. This curvilinear relationship is presented graphically in Figure 12.5. Because correlation coefficients show the relation between two variables *in general*, when the relationship between two variables is curvilinear, the correlation coefficient can be quite small, suggesting a weaker relationship than may actually exist.

Another common problem that arises when examining correlation coefficients is the problem of **truncated range**. This problem is encountered when the scores on one or both of the variables in the analysis do not have much variety in the distribution of scores, possibly due to a ceiling or

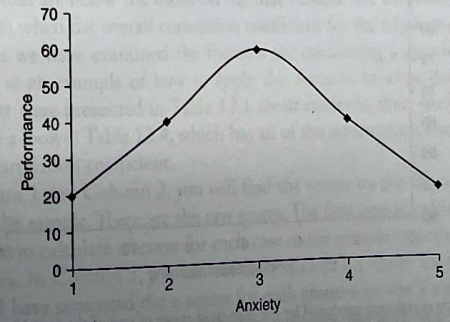

FIGURE 12.5 A curvilinear relationship between anxiety and performance.

TABLE 12.2 Data for the demonstration of truncated range and ceiling effects for association between hours spent studying and exam score

	Hours Spent Studying (*X* Variable)	Exam Score (*Y* Variable)
Student 1	5	95
Student 2	7	97
Student 3	1	98
Student 4	3	100
Student 5	12	96
Student 6	10	99
Student 7	6	98
Student 8	8	98
Student 9	4	97
Student 10	2	95

floor effect. For example, suppose that I gave a sample of students a very easy test with a possible high score of 100. Then suppose I wanted to see if there was a correlation between scores on my test and how much time students spent studying for the test. Suppose I got the data presented in Table 12.2.

In this example, all of my students did well on the test, whether they spent many hours studying for it or not. Because the test was too easy, a ceiling effect may have occurred, thereby truncating the range of scores on the exam. (See Figure 12.6 for an illustration of the ceiling effect and the weak correlation produced by the truncated range of test scores.) Although there may be a relationship between how much time students spent studying and their knowledge of the material, my test was not sensitive enough to reveal this relationship. The weak correlation that will be produced by the data in Table 12.2 may not reflect the true relationship between how much students study and how much they learn.

Calculating the Pearson correlation coefficient by hand is a complicated process. There are different formulas for the calculation, some involving *z* scores and others involving raw scores. Because these formulas are a bit complicated, and because it is easy to calculate correlation coefficients with computer programs and calculators, I have put the details about how to perform these calculations in a separate box in this chapter. Please take a look at the box below if you are interested in the specifics of how to calculate a correlation coefficient.

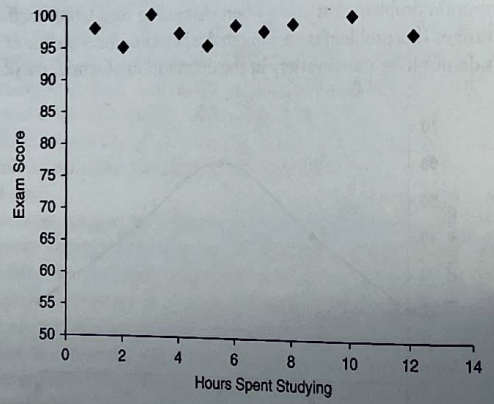

FIGURE 12.6 A small or zero correlation produced by the truncated range of one of the variables, in this case the *Y* variable, test scores.

UNDERSTANDING THE FORMULA FOR CALCULATING THE CORRELATION COEFFICIENT

There are several different formulas that can be used to calculate Pearson product-moment correlation coefficients. These formulas produce the same results and differ only in their ease of use. In fact, none of them is particularly easy to use, and you will probably never need to calculate a correlation coefficient by hand. But it is important to take a look at the formula for calculating the Pearson correlation coefficient (r) so you can gain insights into how this statistic works and what it tells you.

TABLE 12.3 Definitional formula for the Pearson correlation coefficient

$$r = \frac{\Sigma(z_x z_y)}{N}$$

where r is the Pearson product-moment correlation coefficient,
z_x is a z score for variable X,
z_y is a paired z score for variable Y,
N is the number of pairs of X and Y scores.

If you look at the numerator in the formula, you will notice that it is the sum of the z scores (i.e., standardized scores) multiplied together. Each case in the sample will have a score on each of the two variables in the correlation. After standardizing these two variables, each case will have a z score on each of the two variables. So for each case in the sample, multiply the two z scores together, and add up these **cross products**. Recall from Chapter 5 that z scores will be positive for those scores in the sample that are greater than the mean, and negative for those scores that are lower than the mean. If scores above the mean on the first variable (X) are generally associated with scores above the mean on the second variable (Y), the sum of the cross products will be a positive value, and this will produce a positive correlation coefficient. But if scores that are above the mean on the X variable are generally associated with scores that are below the mean on the Y variable, the sum of the cross products will be negative, producing a negative correlation coefficient.

The denominator of the formula in Table 12.3 is N, which refers to the total number of cases in the sample. Whenever we divide by the number of cases, we are getting an average. So in this formula, what we get is the average cross product of standardized scores on the two variables in the correlation. So the correlation coefficient that is produced by this formula tells us the strength and direction of the two standardized variables *in general*. For some cases in our sample, this general pattern will not hold true. For example, some individuals may have scores that are below the mean on the first variable but above the mean on the second variable, even when the overall correlation coefficient for the whole sample is positive.

Now that we have examined the formula for calculating a correlation coefficient, let's take a look at an example of how to apply this formula to some data. I am going to use the data that were presented in Table 12.1 about students, their study time, and their test scores. Take a look at Table 12.4, which has all of the information that we need to calculate a Pearson correlation coefficient.

In Column 1 and Column 2, you will find the scores on the X and Y variables for each student in the sample. These are the raw scores. The first step in calculating the correlation coefficient is to calculate z scores for each case in the sample or population on each of the two variables. In Column 3, you can see each student's z score^2 for the X variable, and in Column 4 I have presented the z scores for each student on the Y variable. Once you have the z scores for each case on each of the two variables, the next step is to multiply the two

TABLE 12.4 Data for calculating the correlation coefficient (*r*)

	Column 1 Hours Spent Studying (X)	Column 2 Exam Score (Y)	Column 3 *z* Score for X	Column 4 *z* Score for Y	Column 5 $z_X z_Y$
Student 1	5	80	-1.41	-.29	.42
Student 2	6	85	-.71	.44	-.31
Student 3	7	70	.00	-1.77	.00
Student 4	8	90	.71	1.18	.83
Student 5	9	85	1.41	.44	.63
		82			Sum = 1.56
Mean	7	6.78			r = .31
Standard Deviation (Population Formula)	1.41				

z scores together for each case. These multiplied *z* scores are presented in Column 5. Next, you add these *z* score products together, creating the sum of 1.56 that is presented in the lower right portion of Table 12.4. The final step in the calculation is to divide that sum by the number of cases: 1.56/5 = .31. So our Pearson correlation coefficient in this example is .31, which is a positive, moderately sized correlation. We can take this a step further and calculate the coefficient of determination: $.31^2 = .10$. So we can say that 10 percent of the variance in exam scores is explained by the variance in hours spent studying.

The Coefficient of Determination

Although correlation coefficients give an idea of the strength of the relationship between two variables, they often seem a bit nebulous. If you get a correlation coefficient of .40, is that a strong relationship? Fortunately, correlation coefficients can be used to obtain a seemingly more concrete statistic: the **coefficient of determination**. Even better, it is easy to calculate.

When we want to know if two variables are related to each other, we are really asking a somewhat more complex question: Are the variations in the scores on one variable somehow associated with the variations in the scores on a second variable? Put another way, a correlation coefficient tells us whether we can know anything about the scores on one variable if we already know the scores on a second variable. In common statistical vernacular, what we want to be able to do with a measure of association, like the correlation coefficient, is to be able to *explain* some of the variance in the scores on one variable based on our knowledge of the scores on a second variable. The coefficient of determination tells us how much of the variance in the scores of one variable can be understood, or explained, by the scores on a second variable.

One way to conceptualize **explained variance** is to understand that when two variables are correlated with each other, they *share* a certain percentage of their variance. Consider an example. If we have a sample of 10 people, and we measure their height and their weight, we've got 10 scores on each of two variables. Assuming that my 10 people differ in how tall they are, there will be some total amount of variance in their scores on the height variable. There will also be some total amount of variance in their scores on the weight variable, assuming that they do not all weigh the same amount. These total variances are depicted in Figure 12.7 as two full squares, each representing 100 percent of the variance in their respective variables. Notice how they do not overlap.

When two variables are related, or correlated, with each other, there is a certain amount of **shared variance** between them. In Figure 12.7, the two squares are not touching each other, suggesting that all of the variance in each variable is independent of the other variable. There is no overlap. But when two variables are correlated, there is some *shared* variance. The stronger the correlation, the greater the amount of shared variance, and the more variance you can explain in one variable by knowing the scores on the second variable. The precise percentage of shared, or explained, variance can be determined by squaring the correlation coefficient. This squared correlation coefficient is known as the **coefficient of determination**. Some examples of different

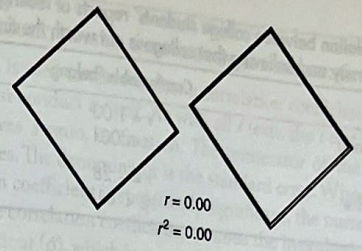

FIGURE 12.7 Uncorrelated variables.

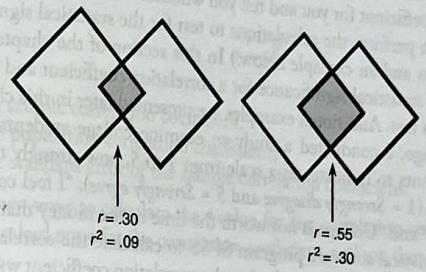

FIGURE 12.8 Examples of different coefficients of determination.

coefficients of determination are presented in Figure 12.8. As the strength of the association between the two variables increases (i.e., larger correlation coefficient), the greater the amount of shared variance, and the higher the coefficient of determination. Notice that on the left side of Figure 12.8, the overlap between the boxes is fairly small, reflecting a correlation coefficient of .30. When this is squared, it produces the coefficient of determination, indicating that 9 percent of the variance in the two variables is shared. On the right side of Figure 12.8, the boxes overlap more, reflecting a larger correlation and more shared variance between the two variables. It is still important to remember that even though the coefficient of determination is used to tell us how much of the variance in one variable can be explained by the variance in a second variable, coefficients of determination do not necessarily indicate a *causal* relationship between the two variables.

Statistically Significant Correlations

When researchers calculate correlation coefficients, they often want to know whether a correlation found in sample data represents the existence of a relationship between two variables in the population from which the sample was selected. In other words, they want to test whether the correlation coefficient is statistically significant (see Chapter 7 for a discussion of statistical significance). To test whether a correlation coefficient is statistically significant, the researcher begins with the null hypothesis that there is absolutely no relationship between the two variables in the population, or that the correlation coefficient in the population equals zero. The alternative hypothesis is that there is, in fact, a statistical relationship between the two variables in the population and that the population correlation coefficient is *not* equal to zero. These two competing hypotheses can be expressed with symbols:

$$H_0 : \rho = 0$$

$$H_0 : \rho \neq 0$$

where ρ is rho, the population correlation coefficient.

TABLE 12.5 Correlation between college students' reports of feeling comfortable at their university and believing that college is not worth the time or money that it costs

	Comfortable/belong	College not worth it
Feel comfortable/belong at this university	$r = 1.00$ $p < .001$	
College is not worth the time or money	$r = -.28$ $p < .01$	$r = 1.00$ $p < .001$

Fortunately, there are many computer programs, and even some calculators, that will compute the correlation coefficient for you and tell you whether it is statistically significant. (If you would like to see how to perform the calculations to test for the statistical significance of r, I have provided instructions and an example below.) In this section of the chapter, I present information from the tests of statistical significance for a correlation coefficient and explain how to interpret the results of this test. Additional examples are presented later in this chapter.

A few years ago, I conducted a study to examine college students' perceptions of college. I asked the students to indicate, on a scale from 1 to 5, how strongly they agreed with each of these statements (1 = *Strongly disagree* and 5 = *Strongly agree*): "I feel comfortable, like I belong, at this university" and "College is not worth the time or the money that it costs."

I used the statistical software program SPSS to calculate the correlation coefficient between these two variables and to indicate whether the correlation coefficient was statistically significant. The results of this analysis are presented in Table 12.5.

There are three boxes in this table that have correlation coefficients and p values presented inside of them. The two boxes along the diagonal of the table are simply variables that are correlated with themselves. These correlations are always perfect and positive (i.e., $r = 1.00$), and are not particularly meaningful. But the correlation between the two variables is presented in the third box, and shows that the scores on these two variables are fairly weakly, negatively correlated ($r = -.28$). This means that, in this sample of 130 college students, those who feel more comfortable at the university are less likely to feel that college is not worth the time or money that it costs. The p value that is reported below the correlation coefficient, $p < .01$, indicates that the probability of obtaining a correlation coefficient this large, from a randomly selected sample of 130, is less than 1 in 100, or less than a 1 percent chance. In Chapter 7, we learned that in most social science research, a p value less than .05 leads us to the conclusion that the result did *not* occur by chance, and is therefore statistically significant. So our correlation coefficient in this example would be considered statistically significant, and we would conclude that in the *population* of college students that this sample represents, there is a negative relationship between feeling that one is comfortable and belongs at the university and the belief that college is not worth the time or money that it costs.

It is important to remember that just because a correlation is statistically significant, that does not mean it is either particularly important or that the association between the two variables is a causal one. The coefficient of determination for this correlation ($-.28^2 = .08$) indicates that only 8 percent of the variance in beliefs about college not being worth it is explained by feelings of comfort and belonging at the university. This is not a lot of variance explained, which suggests that our efforts to predict whether students believe college is worth the cost simply from their feelings of comfort and belonging would not be very accurate. In addition, we cannot tell from this correlation whether feelings of comfort and belonging *affect* whether students feel that college is worth the time and money. The correlation between these two variables may be explained by some third variable that was not included in the analysis, such as how well students are performing in their classes, how well the college football team is performing on the field, or any number of other variables.

If you are interested in the details about how to calculate a t value to determine whether a correlation coefficient is statistically significant, please take a look at the box below.

CALCULATIONS TO TEST FOR STATISTICAL SIGNIFICANCE OF r

The t distribution is used to test whether a correlation coefficient is statistically significant. Therefore, we must conduct a t test. As with all t tests, the t test that we use for correlation coefficients involves a ratio, or fraction. The numerator of the fraction is the difference between two values. The denominator is the standard error. When we want to see whether a sample correlation coefficient is statistically significant, the numerator of the t test formula will be the sample correlation coefficient, r, minus the hypothesized value of the population correlation coefficient (ρ), which in our null hypothesis is zero. The denominator will be the standard error of the sample correlation coefficient:

$$t = \frac{r - \rho}{s_r}$$

where r is the sample correlation coefficient,
ρ is the population correlation coefficient,
s_r is the standard error of the sample correlation coefficient.

There are two ways to calculate the t value for the correlation coefficient. First, we can calculate a standard error for the correlation coefficient, and use that as the denominator of our t formula:

$$s_r = \sqrt{\frac{(1 - r^2)}{(N - 2)}}$$

where r^2 is the correlation coefficient squared and
N is the number of cases in the sample.

With the help of a little algebra, we can also use the formula below to calculate a t value without going through the process of calculating a separate standard error:

$$t = (r)\sqrt{\frac{N - 2}{1 - r^2}}$$

where **degrees of freedom** is the number of cases in the sample minus two ($df = N - 2$).

To illustrate this formula in action, let's consider an example. Some research suggests that there is a relationship between the number of hours of sunlight people are exposed to during the day and their mood. People living at extreme northern latitudes, for example, are exposed to very little sunlight in the depths of winter and may go days or weeks without more than a few hours of sunlight per day. There is some evidence that such sunlight deprivation is related to feelings of depression and sadness. In fact, there is even a name for the condition: seasonal affective disorder, or SAD. To examine this relationship for myself, I randomly select 100 people from various regions of the world, measure the time from sunrise to sunset on a given day where each person lives, and get a measure of each person's mood on a scale from 1 to 10 (1 = *Very sad*; 10 = *Very happy*). Because the members of my sample live at various latitudes, the number of daylight hours will vary. If I conduct my study in January, those participants living in the north will have relatively short days, whereas those living in the south will have long days.

Suppose that I compute a Pearson correlation coefficient with these data and find that the correlation between number of sunlight hours in the day and scores on the mood scale

is $r = .25$. Is this a statistically significant correlation? To answer that question, we must find a t value associated with this correlation coefficient and determine the probability of obtaining a t value of this size by chance (see Chapter 7). In this example,

$$t = (.25)\sqrt{\frac{100-2}{1-.25^2}}$$

$$t = (.25)\sqrt{\frac{98}{1-.25^2}}$$

$$t = (.25)\sqrt{\frac{98}{1-.0625}}$$

$$t = (.25)\sqrt{\frac{98}{1-.9375}}$$

$$t = (.25)\sqrt{104.53}$$

$$t = (.25)10.22$$

$$t = 2.56,\ df = 98$$

To see whether this t value is statistically significant, we must look at the table of t values in Appendix B. There we can see that, because our degrees of freedom = 98, we must look at t values in both the $df = 60$ row and the $df = 120$ row. Looking at the $df = 60$ row, we can see that a t value of 2.56 has a probability of between .01 and .02 (for a two-tailed test). We get the same results when looking in the $df = 120$ row. Therefore, we conclude that our p value is between .01 and .02. If our alpha level is the traditional .05, we would conclude that our correlation coefficient is statistically significant. In other words, we would conclude that, on the basis of our sample statistic, in the larger population of adults, the longer the daylight hours, the better their mood, in general. We could convey all of that information to the informed reader of statistics by writing, "We found a significant relationship between number of daylight hours and mood ($r = .25$, $t_{(98)} = 2.56$, $p < .05$)."

This example also provides a good opportunity to once again remind you of the dangers of assuming that a correlation represents a causal relationship between two variables. Although it may well be that longer days cause the average adult to feel better, these data do not prove it. An alternative causal explanation for our results is that shorter days are also associated with *colder* days, whereas longer days are generally associated with *warmer* days. It may be the case that *warmth* causes better moods and the lack of warmth causes depression and sadness. If people had warm, short days, they might be just as happy as if they had warm, long days. So remember: Just because two variables are correlated, it does not mean that one causes the other.

To watch a video demonstrating how to calculate the t value and determine whether a correlation coefficient is statistically significant, please refer to the website that accompanies this book.

A Brief Word on Other Types of Correlation Coefficients

Although Pearson correlation coefficients are probably the most commonly used and reported in the social sciences, they are limited by the requirement that both variables be measured on interval or ratio scales. Fortunately, there are methods available for calculating the strength of the relationship between two variables even if one or both variables are not measured using interval or ratio scales. In this section, I briefly describe three of these "other" correlation coefficients. It is important to note that all of these statistics are very similar to the Pearson correlation coefficient and each produces a correlation coefficient that is similar to the Pearson r. They are simply specialized versions of the Pearson correlation coefficient that can be used when one or both of the variables are not measured using interval or ratio scales.

Point-Biserial Correlation

When one of our variables is a continuous variable (i.e., measured on an interval or ratio scale) and the other is a two-level categorical (a.k.a. nominal) variable (also known as a **dichotomous variable**), we need to calculate a point-biserial correlation coefficient. This coefficient is a specialized version of the Pearson correlation coefficient discussed earlier in this chapter. For example, suppose I want to know whether there is a relationship between whether a person owns a car (yes or no) and their score on a written test of traffic rule knowledge, such as the tests one must pass to get a driver's license. In this example, we are examining the relation between one categorical variable with two categories ("owns a car" or "does not own a car") and one continuous variable (one's score on the driver's test). Therefore, the point-biserial correlation is the appropriate statistic in this instance.

Phi

Sometimes researchers want to know whether two dichotomous variables are correlated. In this case, we would calculate a phi coefficient (Φ), which is another specialized version of the Pearson r. For example, suppose I want to know whether gender (male or female) is associated with whether one smokes cigarettes or not (smoker or nonsmoker). In this case, with two dichotomous variables, I would calculate a phi coefficient. (*Note*: Those readers familiar with chi-square analysis will notice that two dichotomous variables can also be analyzed using chi squared.)

Spearman Rho

Sometimes data are recorded as ranks. Because ranks are a form of ordinal data, and the other correlation coefficients discussed so far involve either continuous (interval, ratio) or dichotomous variables, we need a different type of statistic to calculate the correlation between two variables that use ranked data. In this case, the **Spearman rho**, a specialized form of the Pearson r, is appropriate. For example, many schools use students' grade point averages (a continuous scale) to rank students (an ordinal scale). In addition, students' scores on standardized achievement tests can be ranked. To see whether a student's rank in their school is related to their rank on the standardized test, a Spearman rho coefficient can be calculated.

Example: The Correlation Between Grades and Test Scores

Student achievement can be measured in a variety of ways. Two of these methods are the grade point average (GPA), usually measured on a scale from 0 to 4 (with 4 = "A"), and standardized test scores. Grades are given to students by their teachers, and standardized test scores indicate how well students have performed on tests produced by test makers who are not affiliated with the students' schools, which are usually scored by computers.

TABLE 12.6 SPSS printout of correlation analysis

	GPA	Naglieri Test Scores
GPA	1.0000 (*n* = 314)	
Naglieri test scores	.4291 (*n* = 314) *p* = .000	1.0000 (*n* = 314)

My colleague, Carol Giancarlo, and I collected data from a sample of 314 eleventh-grade students at a high school in California. Among the data we collected were their cumulative GPAs (i.e., their GPAs accumulated from the time they began high school up to the time the data were collected). In addition, we gave students the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1996), a nonverbal test of general mental reasoning and critical thinking skills. To see if there was a statistically significant correlation between these two measures of ability, I used the SPSS statistical software program to calculate a correlation coefficient and a *p* value. The SPSS printout from this analysis is presented in Table 12.6.

The results presented in Table 12.6 provide several pieces of information. First, there are three correlation coefficients presented. The correlations on the diagonal show the correlation between a single variable and itself. Therefore, the first correlation coefficient presented reveals that GPA is correlated with itself perfectly ($r = 1.0000$). Because we always get a correlation of 1.00 when we correlate a variable with itself, these correlations presented on the diagonal are meaningless. That is why there is not a *p* value reported for them. The numbers in the parentheses, just below the correlation coefficients, report the sample size. There were 314 eleventh-grade students in this sample. The correlation coefficient that is *off* the diagonal is the one we're interested in. Here, we can see that students' GPAs were moderately correlated with their scores on the Naglieri test ($r = .4291$). This correlation is statistically significant, with a *p* value of less than .0001 ($p < .0001$).

To gain a clearer understanding of the relationship between GPA and Naglieri test scores, we can calculate a coefficient of determination. We do this by squaring the correlation coefficient. When we square this correlation coefficient ($.4291 \times .4291 = .1841$), we see that GPA explains a little over 18 percent of the variance in the Naglieri test scores. Although this is a substantial percentage, it still leaves more than 80 percent of the ability test scores unexplained. Because of this large percentage of unexplained variance, we must conclude that teacher-assigned grades reflect something substantially different from general mental reasoning abilities and critical thinking skills, as measured by the Naglieri test.

Worked Example: Statistical Significance and Confidence Interval

Suppose that I want to examine whether there is a correlation between how much time high school students spend working on homework and their scores on a standardized achievement test. I select a random sample of 64 high school students and find the correlation between the hours they spend working on homework, per week, and their scores on a standardized achievement test. I find that the correlation between these two variables is $r = .40$. From this statistic I know that, *in this sample*, there is a positive, moderate association between hours spent on homework and scores on the achievement test. But does this correlation in the sample indicate a positive correlation between these two variables in the larger population of high school students? To answer that question we need to calculate a *t* value (see above). Because we are also going to calculate a confidence interval, and because we have already seen a worked example using the simplified *t* test formula earlier in the chapter, this time let's calculate the standard error of the correlation coefficient:

$$s_r = \sqrt{(1 - .4^2)/(64 - 2)}$$
$$s_r = \sqrt{(.84)/(62)}$$
$$s_r = .12$$

When we plug this standard error into the standard *t* test formula for correlation coefficients, we find that our observed *t* value is

$$t = \frac{.40}{.12} = 3.33$$

Using our degrees of freedom of $N - 2$, we find a critical *t* value of 2.000. Notice that we used an alpha level of .05, a two-tailed test, and 60 degrees of freedom in Appendix B because that was the closest value we could find to our actual degrees of freedom. Because our observed *t* value of 3.33 is larger than our critical *t* value of 2.000, we conclude that our sample correlation coefficient is statistically significant. In other words, it is significantly different from zero, which was the value of the correlation coefficient proposed in the null hypothesis. Now we can conclude that in the population of high school students, there is a positive, moderately strong correlation between amount of hours spent doing homework per week and scores on this achievement test. Note that we *cannot* conclude that spending more time on homework *causes* higher scores on the achievement test. There could well be other explanations for this positive correlation.

Now let's calculate a 99 percent confidence interval for the sample correlation coefficient. We already know the standard error and the sample *r*, so we just need to look up the *t* value for 60 *df* and a two-tailed alpha level of .01. In Appendix B we see that this gives us a *t* value of 2.617. (*Note*: Many students make the mistake of plugging the *observed t* value that they calculated for the *t* test into the confidence interval formula. Be sure you use the correct *t* value from Appendix B in your confidence interval formula.)

$$CI_{99} = .4 \pm (.12)(2.617) \rightarrow .4 \pm (.31) \rightarrow .09, .71$$

Now we can wrap words around our results: We are 99 percent confident that the population correlation coefficient is contained within the interval ranging from .09 to .71.

To watch a video demonstrating how to calculate and interpret a confidence interval for *r*, please refer to the website that accompanies this book.

Writing it Up

When researchers write about bivariate correlations in scientific reports and journal articles, they often present a table that includes the correlation coefficients between the variables, some indication of whether the correlation coefficients are statistically significant, and descriptive statistics (i.e., means and standard deviations) for those variables. In addition, they provide a brief description of the correlations in the text.

In a study that I conducted some time ago, I wanted to compare two different survey measures of achievement goals. Achievement goals represent what students may be trying to accomplish in school, and are divided into mastery goals (the desire to develop knowledge and skills) and performance goals (the desire to do better than others). So I gave a sample of college students two surveys with items assessing mastery and performance goals. One survey was called the Achievement Goals Questionnaire (AGQ; Elliot & McGregor, 2001) and the other was the

TABLE 12.7 SPSS printout of correlation analysis

	Mastery AGQ[1]	Performance AGQ[1]	Mastery PALS[2]	Performance PALS[2]	Mean (s.d.)[3]
Mastery AGQ	1.00				5.56 (.96)
Performance AGQ	.24**	1.00			4.16 (1.46)
Mastery PALS	.75**	.18*	1.00		5.61 (.84)
Performance PALS	.03	.77**	−.02	1.00	4.17 (1.31)

Notes:
1 AGQ: Achievement Goals Questionnaire
2 PALS: Patterns of Adaptive Learning Survey
3 s.d. = standard deviation
* indicates $p < .05$
** indicates $p < .01$

Patterns of Adaptive Learning Survey (PALS; Midgley et al., 1998). I was curious whether these two different surveys were measuring the same thing, so I needed to see how strongly students' scores on the goal measures from one survey were correlated with their scores on the same goal measures from the other survey. The results of the correlation analyses, along with the means and standard deviations for each variable, are presented in Table 12.7.

If I were writing these results up for publication I would write something like the following:

> "As the correlation coefficients presented in Table 12.7 reveal, the correlations between the comparable goal variables in the AGQ and the PALS were quite strong. The correlation between the two measures of mastery goals was $r = .75$. The correlation between the performance goal scales of the AGQ and the PALS were similarly strong ($r = .77$). Although these correlations are quite strong, the coefficients of determination ($r^2 = .56$ and .59 for the mastery and performance goals, respectively) indicate that there is still quite a bit of variance in the goal measures from the PALS that is not explained by the corresponding goal measures of the AGQ. This suggests that there are important differences in students' interpretations of the goal items on these two survey measures."

Wrapping Up and Looking Forward

Correlation coefficients, in particular Pearson correlation coefficients, provide a way to determine both the direction and the strength of the relationship between two variables measured on a continuous scale. This index can provide evidence that two variables are related to each other, or that they are not, but does not, in and of itself, demonstrate a *causal* association between two variables. In this chapter, I also introduced the concepts of explained or shared variance and the coefficient of determination. Determining how much variance in one variable is shared with, or explained by, another variable is at the core of all of the statistics that are discussed in the remaining chapters of this book. In particular, correlation coefficients are the precursors to the more sophisticated statistics involved in multiple regression (Chapter 13). In the next chapter, we examine *t* tests, which allow us to look at the association between a two-category independent variable and a continuous dependent variable.

Work Problems

Suppose you want to know whether there is a correlation between how much time healthy young adults spend exercising and how much sleep they get. You select a random sample of 25 healthy young adults and ask them how many hours they spend exercising per week and how many hours they spend sleeping per week. You find there is a correlation in this sample of .45. Please answer the following questions based on this information.

1. What is the direction of this correlation coefficient? How do you know?
2. What is the strength of this correlation coefficient? How do you know?
3. Interpret this correlation coefficient. What does it tell you?
4. Calculate the coefficient of determination and interpret it. What does it tell you?
5. Using Appendix B and an alpha level of .05, find the critical *t* value.
6. Calculate the observed *t* value and decide whether this correlation coefficient is statistically significant using an alpha level of .05. Interpret your results. What do they tell you?
7. Calculate a 95 percent confidence interval for this correlation coefficient and interpret it. What does it tell you?

For answers to these work problems, and for additional work problems, please refer to the website that accompanies this book.

Notes

1 There is a website that has several excellent examples of **spurious correlations**—correlations between variables that are coincidental or caused by some third variable. Some examples from the website include the number of people who drowned by falling into a pool and how many films Nicolas Cage appeared in across several years ($r = .66$), and a strong correlation ($r = .99$) between the divorce rate in Maine and per capita consumption of margarine over time. For more examples, take a look at the website: http://tylervigen.com/spurious-correlations.

2 When calculating *z* scores for use in the correlation coefficient, the standard deviation that is used is the standard deviation for the population, not the standard deviation for the sample.

CHAPTER 13

Regression

In Chapter 12, the concept of correlation was introduced. Correlation involves a measure of the degree to which two variables are related to each other. A closely related concept, coefficient of determination, was also introduced in that chapter. This statistic provides a measure of the strength of the association between two variables in terms of percentage of variance explained. Both of these concepts are present in regression. In this chapter, the concepts of **simple linear regression** and **multiple regression** are introduced.

Regression is a very common statistical method in the social sciences. One of the reasons it is such a popular technique is because it is so versatile. Regression, particularly multiple regression, allows researchers to examine the nature and strength of the relations between variables, the relative predictive power of several independent variables on a dependent variable, and the unique contribution of one or more independent variables when controlling for one or more covariates. It is also possible to test for interactions in multiple regression. With all of the possible applications of multiple regression, it is clear that it is impossible to describe all of the functions of regression in this brief chapter. Therefore, the focus of this chapter is to provide an introduction to the concept and uses of regression, and to refer the reader to resources providing additional information.

Simple vs. Multiple Regression

The difference between simple and multiple regression is similar to the difference between one-way and factorial ANOVA. Like one-way ANOVA, simple regression analysis involves a single **independent**, or **predictor variable** and a single **dependent**, or **outcome variable**. This is the same number of variables as used in a simple correlation analysis. The difference between a Pearson correlation coefficient and a simple regression analysis is that while the correlation does not distinguish between independent and dependent variables, in a regression analysis there is always a designated predictor variable and a designated dependent variable (although, just as with correlations, regression analyses do not allow researchers to claim that there is a *causal* association between variables). That is because the purpose of regression analysis is to make *predictions* about the value of the dependent variable given certain values of the predictor variable. This is a simple extension of a correlation analysis. If I am interested in the relationship between height and weight, for example, I could use a simple regression analysis to answer this question: If I know a man's height, what would I *predict* his weight to be? Notice that the variable I am interested in predicting is designated as my dependent variable, and the variable or variables that I am using to predict my dependent variable become my predictor, or independent, variable or variables. Of course, the accuracy of my prediction will only be as good as my correlation will allow, with stronger correlations leading to more accurate predictions. Therefore, simple linear regression is not really a more powerful tool than simple correlation analysis. But it does give me another way of conceptualizing the relation between two variables, a point I elaborate on shortly.

The real power of regression analysis can be found in multiple regression. Like factorial ANOVA, multiple regression involves models that have two or more predictor variables and a single dependent variable. For example, suppose that, again, I am interested in predicting how much a person weighs (i.e., weight is the **dependent variable**). Now suppose that in addition to height, I know how many minutes of exercise that person gets per day, and how many calories a day he consumes. Now I've got three predictor variables (height, exercise, and calories consumed) to help me make an educated guess about the person's weight. Multiple regression analysis allows me to see, among other things, (a) how much these three **predictor variables**, as a group, are related to weight, (b) the strength of the relationship between each predictor variable and the dependent variable *while controlling for the other predictor variables in the model*, (c) the *relative* strength of each predictor variable, and (d) whether there are interaction effects between the predictor variables. As you can see, multiple regression is a particularly versatile and powerful statistical technique.

Variables Used in Regression

As with correlation analysis, in regression the dependent and **independent variables** (a.k.a. the predictor variables) need to be measured on an interval or ratio scale. **Dichotomous** (i.e., categorical variables with two categories) predictor variables can also be used.[1] There is a special form of regression analysis, logit regression, that allows us to examine dichotomous dependent variables, but this type of regression is beyond the scope of this book. In this chapter, we limit our consideration of regression to those types that involve a continuous dependent variable and either continuous or dichotomous predictor variables.

Regression in Depth

Regression, particularly simple linear regression, is a statistical technique that is very closely related to correlations (discussed in Chapter 12). In fact, when examining the relationship between two continuous (i.e., measured on an interval or ratio scale) variables, either a correlation coefficient or a regression equation can be used. Indeed, the Pearson correlation coefficient is nothing more than a simple linear regression coefficient that has been standardized. The benefits of conducting a regression analysis rather than a correlation analysis are (a) regression analysis yields more information, particularly when conducted with one of the common statistical software packages, and (b) the regression equation allows us to think about the association between the two variables of interest in a more intuitive way. Whereas the correlation coefficient provides us with a single number (e.g., $r = .40$), which we can then try to interpret, the regression analysis yields a formula for calculating the **predicted value** of one variable when we know the actual value of the second variable. Here's how it works. There are a few assumptions of regression that must be met or the results of the regression may not be valid. First, as mentioned, the dependent variable should be measured on an interval/ratio scale and the predictor variables should either be interval/ratio or dichotomous variables. Second, it is assumed that there is a linear association between the predictor variable and the dependent variable. (Recall from Chapter 12 that non-linear associations, such as curvilinear relationships, can create difficulties in identifying the true association among the variables.) Third, all of the variables in the regression analysis should have normal distributions. A fourth assumption (of multiple regression in particular) is that the predictor variables are not too strongly correlated with each other. It is also assumed that the errors in prediction are independent of each other. Finally, there is the assumption of **heteroscedasticity**. This means that errors in the prediction of Y are about the same, in both size and direction, at all levels of X.

The key to understanding regression is to understand the formula for the regression equation. So I begin by presenting the most simple form of the regression equation, describing how it works, and then moving on to more complicated forms of the equation. In Table 13.1, the

TABLE 13.1 The regression equation

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

where $\hat{Y}$ is the predicted value of the Y variable,
b is the unstandardized regression coefficient, or the slope,
X is a given value of the independent variable,
a is the intercept (i.e., the point where the regression line intercepts the Y axis).

regression equation used to find the predicted value of Y is presented along with definitions of the components.

In simple linear regression, we begin with the assumption that the two variables are **linearly** related. In other words, if the two variables are actually related to each other, we assume that every time there is an increase of a given size in value on the X variable (called the **predictor** or **independent** variable), there is a corresponding increase (if there is a positive correlation) or decrease (if there is a negative correlation) *of a given size* in the Y variable (called the **dependent**, **outcome**, or **criterion** variable). In other words, if the value of X increases from a value of 1 to a value of 2, and Y increases by 2 points, then when X increases from 2 to 3, we would predict that the value of Y would increase another 2 points.

To illustrate this point, let's consider the following set of data. Suppose I want to know whether there is a relationship between the level of education people have and their monthly income. Education level is measured in years, beginning with kindergarten and extending through graduate school. Income is measured in thousands of dollars. Suppose that I randomly select a sample of 10 adults and measure their level of education and their monthly income, getting the data provided in Table 13.2.

When we look at these data, we can see that, in general, monthly income increases as the level of education increases. This is a general, rather than an absolute, trend because in some cases a person with more years of education makes less money per month than someone with fewer years of education (e.g., Case 10 and Case 9, Case 6 and Case 5). So although not every person with more years of education makes more money, *on average* more years of education are associated with higher monthly incomes. The correlation coefficient that describes the relation of these two variables is $r = .83$, which is a very strong, positive correlation (see Chapter 12 for a more detailed discussion of correlation coefficients).

If we were to plot these data on a simple graph, we would produce a **scatterplot**, such as the one provided in Figure 13.1. In this scatterplot, there are 10 data points, one for each case in the study.

TABLE 13.2 Income and education level data

	Education Level (X) (in Years)	Monthly Income (Y) (in Thousands of Dollars)
Case 1	6	1
Case 2	8	1.5
Case 3	11	1
Case 4	12	2
Case 5	12	4
Case 6	13	2.5
Case 7	14	5
Case 8	16	6
Case 9	16	10
Case 10	21	8
Mean	12.90	4.10
Standard Deviation	4.25	3.12
Correlation Coefficient	.83	

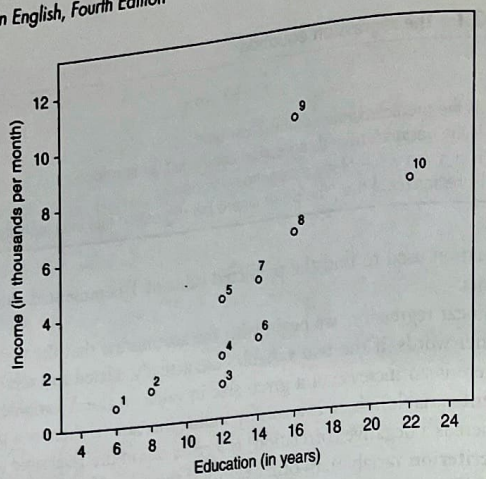

FIGURE 13.1 Scatterplot for education and income.

Note that each data point marks the spot where education level (the *X* variable) and monthly income (the *Y* variable) meet for each case. For example, the point that has a value of 10 on the *Y* axis (income) and 16 on the *X* axis (education level) is the data point for the 10th case in our sample. These 10 data points in our scatterplot reveal a fairly distinct trend. Notice that the points rise somewhat uniformly from the lower-left corner of the graph to the upper-right corner. This shape is a clear indicator of the positive relationship (i.e., correlation) between education level and income. If there had been a perfect correlation between these two variables (i.e., $r = 1.0$), the data points would be aligned in a perfectly straight line, rising from the lower left to the upper right on the graph. If the relationship between these two variables were weaker (e.g., $r = .30$), the data points would be more widely scattered, making the lower-left to upper-right trend much less clear.

With the data provided in Table 13.2, we can calculate all of the pieces of the regression equation. The regression equation allows us to do two things. First, it lets us find predicted values for the *Y* variable for any given value of the *X* variable. In other words, we can predict a person's monthly income if we know how many years of education he or she has. Second, the regression equation allows us to produce the **regression line**. The regression line is the basis for linear regression and can help us understand how regression works.

There are a number of different types of regression formulas, but the most commonly used is called **ordinary least squares regression**, or **OLS**. OLS is based on an idea that we have seen before: the *sum of squares* (see Chapters 3 and 9). If you wanted to, you could draw a number of straight lines that go through the cluster of data points presented in the scatterplot in Figure 13.1. For example, you could draw a horizontal line that extends out from the number 5 on the *Y* axis. Similarly, you could draw a straight line that extends down from the number 10 on the *Y* axis to the number 22 on the *X* axis. No matter how you decide to draw your straight line, notice that at least some of the data points in the scatterplot will not fall exactly *on* the line. Some or all will fall above the line, some may fall directly on the line, and some or all will fall below the line. Any data point that does not fall directly on the line will have a certain amount of distance between the point and the line. Now, if you were to calculate the distance between the data point and the line you have drawn, and then square that distance, you would have a *squared deviation* for that individual data point. If you calculated the squared deviation for each data point that did not fall on the line, and added all of these squared deviations together, you would end up with the *sum of squared deviations*, or *sum of squares*.

Now here is the key: The sum of the squared deviations, or sum of squares, will differ depending on where you draw your line. In any scatterplot, there is only *one* line that produces the *smallest* sum of squares. This line is known as the line of *least squares*, and this is the regression line. So, the reason this type of regression is called ordinary *least squares* regression is because in this form of regression, the regression line represents the straight line that produces the *smallest* sum of squared deviations from the line. This regression line represents the *predicted* values of *Y* at any given value of *X*. Of course, when we predict a value of *Y* for a given value of *X*, our prediction may be off. This error in prediction is represented by the distance between the regression line and the actual data point(s) in the scatterplot. To illustrate how this works, we first need to calculate the properties of the regression line (i.e., its slope and intercept). Then we draw this regression line into the scatterplot, and you can see how well it "fits" the data (i.e., how close the data points fall to the regression line).

If you take a look at the formula for the regression equation in Table 13.1, you will notice that there are four components: (a) $\hat{Y}$ is the predicted value of the *Y* variable, (b) ***b*** is the unstandardized **regression coefficient**, and also the **slope** of the regression line, (c) ***X*** is the value of the *X* variable, and (d) ***a*** is the value of the **intercept** (i.e., where the regression line crosses the *Y* axis). Because $\hat{Y}$ is the value produced by the **regression equation**, let's save that one for last. And because *X* is just a given value on the *X* variable, there is not really anything to work out with that one. So let's take a closer look at *a* and *b*.

We cannot calculate the intercept before we know the slope of the regression line, so let's begin there. The formula for calculating the regression coefficient is

$$b = r \times \frac{s_y}{s_x}$$

where *b* is the regression coefficient,
r is the correlation between the *X* and *Y* variables,
s_y is the standard deviation of the *Y* variable,
s_x is the standard deviation of the *X* variable.

Looking at the data in Table 13.2, we can see that $r = .83$, $s_y = 3.12$, and $s_x = 4.25$. When we plug these numbers into the formula, we get the following:

$$b = .83 \times \frac{3.12}{4.25}$$

$$b = (.83) \times (.73)$$

$$b = .61$$

Notice that the regression coefficient is simply the correlation coefficient times the ratio of the standard deviations for the two variables involved. When we multiply the correlation coefficient by this ratio of standard deviations, we are roughly transforming the correlation coefficient into the scales of measurement used for the two variables. Notice that there is a smaller range, or less variety, of scores on our *Y* variable than there is on our *X* variable in this example. This is reflected in the ratio of standard deviations used to calculate *b*.

Now that we've got our *b*, we can calculate our intercept, *a*. The formula for *a* is as follows:

$$a = \bar{Y} - b\bar{X}$$

where $\bar{Y}$ is the average value of *Y*,
$\bar{X}$ is the average value of *X*,
b is the regression coefficient.

When we plug in the values from Table 13.2, we find that

$$a = 4.1 - (.61)(12.9)$$

$$a = 4.1 - 7.87$$

$$a = -3.77$$

This value of a indicates that the intercept for the regression line is –3.77. In other words, the regression line crosses the Y axis at a value of –3.77. In still other words, this intercept tells us that when $X = 0$, we would predict the value of Y to be –3.77. Of course, in the real world, it is not possible to have a monthly income of negative 3.77 thousand dollars. Such unrealistic values remind us that we are dealing with *predicted* values of Y. Given our data, if a person has absolutely no formal education, we would *predict* that person to make a negative amount of money.

Now we can start to fill out our regression equation. The original formula

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

now reads

$$\hat{Y} = -3.77 + .61X$$

It is important to remember that when we use the regression equation to find predicted values of Y for different values of X, we are not calculating the *actual* value of Y. We are only making *predictions* about the value of Y. Whenever we make predictions, we will sometimes be incorrect. Therefore, there is bound to be some **error** (e) in our predictions about the values of Y at given values of X. The stronger the relationship (i.e., correlation) between my X and Y variables, the less error there will be in my predictions. The error is the difference between the actual, or observed, value of Y and the predicted value of Y. Because the predicted value of Y is simply $a + bX$, we can express the formula for the error in two ways:

$$e = Y - \hat{Y}$$

$$e = Y - a + bX$$

So, rather than a single regression equation, there are actually two. One of them, the one presented in Table 13.1, is for the *predicted* value of Y ($\hat{Y}$). The other one is for the actual, or *observed* value of Y. This equation takes into account the errors in our predictions, and is written as $Y = bX + a + e$.

Now that we've got our regression equation, we can put it to use. First, let's wrap words around it, so that we can make sure we understand what it is telling us. Our regression coefficient tells us that "For every unit of increase in X, there is a corresponding predicted increase of .61 units in Y." Applying this to our variables, we can say that "For every additional year of education, we would *predict* an increase of .61 times $1,000, or $610, in monthly income." We know that the predicted value of Y will *increase* when X increases, and vice versa, because the regression coefficient is *positive*. Had it been negative, we would predict a decrease in Y when X increases.

Next, let's use our regression equation to find predicted values of Y at given values of X. For example, what would we predict the monthly income to be for a person with 9 years of formal education? To answer this question, we plug in the value of 9 for the X variable and solve the equation:

$$\hat{Y} = -3.77 + .61(9)$$

$$\hat{Y} = -3.77 + 5.59$$

$$\hat{Y} = 1.82$$

So we would predict that a person with 9 years of education would make $1,820 per month, plus or minus our error in prediction (e).

Finally, we can use our regression equation to compute our regression line. We already know, from the value of the intercept, that our regression line will cross the Y axis at a value of –3.77. To draw a straight line, all we need to do is calculate one additional point. To make sure we include all of the points in our scatterplot, let's just calculate a predicted value of Y for a person with 25 years of education:

$$\hat{Y} = -3.77 + .61(25)$$

$$\hat{Y} = -3.77 + 15.25$$

$$\hat{Y} = 11.48$$

If we were to draw a regression line through our scatterplot, using the two points we found from our intercept and the predicted Y value, we would get something like the line presented in Figure 13.2.

With the regression line added to the scatterplot, some of the concepts mentioned earlier in this chapter may be easier to understand. First, notice that our regression line does not accurately predict the actual Y values for any of our cases except for Case 7. That data point is precisely on the regression line. For each of the other nine cases, there is some amount of error present in the prediction. In some cases, the amount of error is very small (e.g., Case 8), whereas in others the amount of error is quite large (e.g., Case 9). These errors in prediction are known as **residuals**. (See Figure 13.3 for an illustration of the residual for Case 10.) In some cases, our predicted value was less than our **observed value** (e.g., Cases 1, 2, 5, 8, and 9). For these cases, we have **underpredicted** their income based on their level of education. Such underpredictions produce *positive* residuals (because the residual = observed scores – predicted score). For other cases (Cases 3, 4, 6, and 10) we **overpredicted** the Y value, creating *negative* residuals. Second, notice the distance

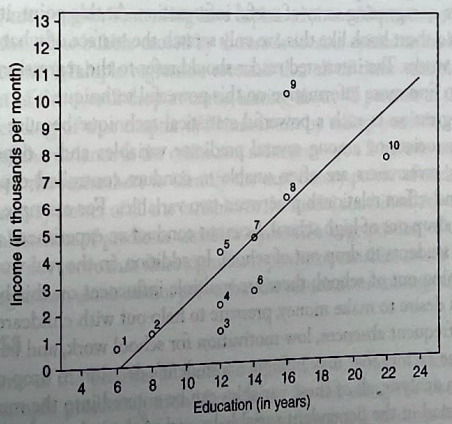

FIGURE 13.2 Scatterplot for education and income with regression line.

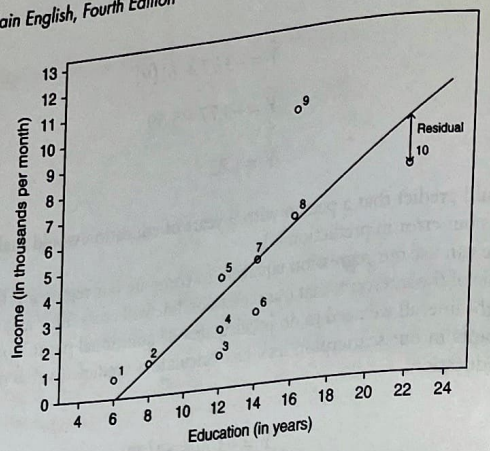

FIGURE 13.3 Illustration of the residual for Case 10.

between each case and the line. When we square each of these distances and then add them all together, we get the sum of squares. Third, notice that the regression line marks the line where the sum of the squared distances is smallest. To test this, try drawing some other lines and note the way this increases the overall amount of error in prediction. Finally, notice where the regression line crosses the *Y* axis (the intercept) and how much higher up the *Y* axis the regression line goes for each increase of one unit value in *X* (the slope). The slope and the intercept will correspond with the values that we found for *b* and *a*, respectively.

Multiple Regression

Now that we've discussed the elements of simple linear regression, let's move on to a consideration of **multiple regression**. Despite the impressive qualities of simple linear regression, the plain truth is that when we only have two variables, simple linear regression does not provide much more information than would a simple correlation coefficient. Because of this, you rarely see a simple linear regression with two variables reported in a published study. But multiple regression is a whole different story. Multiple regression is a very powerful statistic that can be used to provide a staggering array of useful information. At this point, it may be worth reminding you that in a short book like this, we only scratch the surface of what multiple regression can do and how it works. The interested reader should refer to the regression related readings in the bibliography to find more information on this powerful technique.

Multiple regression is such a powerful statistical technique because it allows researchers to examine the associations among several predictor variables and a dependent variable. In the social sciences, researchers are often unable to conduct controlled experiments that isolate a single cause-and-effect relationship between two variables. For example, if I want to know why some students drop out of high school, I cannot conduct an experiment as it would be unethical to cause some students to drop out of school. In addition, in the real world there is not a single cause of dropping out of school; there are multiple influences on this behavior. Low academic achievement, a desire to make money, pressure to help out with childcare and other responsibilities at home, frequent absences, low motivation for school work, and being socially isolated are just some of the factors that may influence a student's decision to drop out of school. In a multiple regression analysis, all of these variables can be entered into the model, and the amount of variance explained in the dependent variable by *each* of the independent variables can be examined, while controlling for the effects of the other variables. This is powerful stuff.

Two examples of real-world uses of multiple regression can demonstrate the power of the technique. First, much of the current data-driven analytics that are being used by companies to sell you products are taking advantage of multiple regression. For example, have you ever noticed which advertisements pop up on your screen when you are on a social media website? Well, it costs a lot of money to advertise on these websites, so advertisers want to make sure they show you ads that are likely to appeal to you. How do they know which ads will catch your attention? Your purchasing behavior is predicted by several factors including your gender, your age, where you live, whether you are single or in a relationship, what your friends like to do, what you like that your friends do, how much money you spend online, and what products you search for online. All of these variables can be plugged into a multiple regression to help companies determine the *strongest* predictor of your shopping behavior and show you advertisements that are targeted toward that predictor. If age is the strongest predictor of shopping behavior, you will see advertisements targeting your age group (e.g., age-related drugs for senior citizens, sugary energy drinks for adolescents and young adults). Some of the earliest adopters of multiple regression analysis techniques were insurance companies. To decide how much to charge each driver for auto insurance, companies look at predictors of safe driving behavior, like age (older drivers tend to be safer), where you live (people who live in more densely populated areas tend to get in more accidents), how well you do in school (higher achievers tend to drive more carefully), and so on.

In one particularly interesting example of an application of multiple regression, described in the book *Freakonomics* (Levitt & Dubner, 2009), researchers examined crime rates in the United States. After many years of rising crime rates that reached a peak in the early 1990s, crime rates began to fall dramatically from the mid-1990s through 2010. What caused this dramatic drop in crime? Several possible explanations have been offered, including more and better policing, tougher sentencing laws for criminals, a strong economy in the 1990s, increased use of capital punishment, tougher gun laws, and an aging population, among other explanations. Using multiple regression, researchers were able to determine how much of the variance in the drop in crime rates could be explained by *each* of these variables, controlling for all of the other variables. They found that some predictor variables (e.g., a strong economy, increased use of capital punishment, tougher gun laws, an aging population) had no effect on the crime rate. It is worth noting that some of these predictor variables—an improving economy in the 1990s, an aging population—were significantly *correlated* with the decrease in crime, but when the effects of other predictor variables were controlled in the multiple regression, these predictor variables were no longer significantly associated with the dependent variable. In contrast, several predictor variables remained significantly associated with the drop in crime, even after controlling for the effects of other variables. More and better policing explained about 10 percent of the drop in crime, tougher sentencing laws explained about 33 percent of the drop in crime, and the bursting of the crime bubble caused by the crack cocaine epidemic explained about 15 percent of the drop in crime. Multiple regression revealed that all of these variables combined explained about 58 percent of the overall drop in crime, leaving about 42 percent of the drop unexplained. (I won't spoil the ending of that chapter of *Freakonomics* by telling you what the authors think may explain the rest of the variance in the crime rate reduction—you'll have to read it yourself.) Multiple regression allowed the researchers to identify the explanatory power of each variable, as well as the combined explanatory power of all the variables in the model combined. Always remember that even though this combination of predictor variables *explained* a large portion of the variance in the drop in crime, this does not prove that these significant predictor variables *caused* the drop in crime. They may have, but regression alone cannot prove that the associations among variables are causal.

An Example Using SPSS

To illustrate some of the benefits of multiple regression, let's add a second *predictor* variable to our previous example predicting monthly income. So far, using the data from Table 13.2, we have examined the relationship between education level and income. In this example, education level has been used as our *predictor* or *independent* variable and income has been used as our

dependent or *outcome* variable. We found that, on average, in our sample, one's monthly salary is predicted to increase by $610 for every additional year of schooling the individual has received. But there was some error in our predictions, indicating that there are other variables that predict how much money one makes. One such predictor may be the length of time one has been out of school. Because people tend to make more money the longer they have been in the workforce, it stands to reason that those adults in our sample who finished school a long time ago may be making more than those who finished school more recently. Although Case 4 and Case 5 each had 12 years of schooling, Case 5 makes more money than Case 4. Perhaps this is due to Case 5 being in the workforce longer than Case 4.

When we add this second predictor variable to the model, we get the following regression equation:

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

where $\hat{Y}$ is the predicted value of the dependent variable,
X_1 is the value of the first predictor variable,
X_2 is the value of the second predictor variable.
b_1 is the regression coefficient for the first independent variable,
b_2 is the regression coefficient for the second independent variable.

This regression equation with two predictor variables will allow me to examine a number of different questions. First, I can see whether my two predictor variables combined are significantly related to, or predictive of, my dependent variable, and how much of the variance my predictor variables explain in my dependent variable. Second, I can test whether each of my predictor variables is significantly related to my dependent variable *when controlling for the other predictor variable.* When I say "controlling for the other predictor variable," I mean that I can examine whether a predictor variable is related to the dependent variable after I partial out, or take away, the portion of the variance in my dependent variable that has already been accounted for by my other independent variable. Third, I can see which of my two predictor variables is the stronger predictor of my dependent variable. Fourth, I can test whether one predictor variable is related to my dependent variable after controlling for the other predictor variable, thus conducting a sort of ANCOVA (see Chapter 10 for a discussion of ANCOVA). There are many other things I can do with multiple regression, but I will limit my discussion to these four.

Suppose that for the 10 cases in my sample, I also measure the number of years that they have been in the workforce, and I get the data presented in Table 13.3. These data reveal that both years of education and years in the workforce are positively correlated with monthly income. But how much of the variance in income can these two predictor variables explain *together*? Will years of education still predict income when we control for the effects of years in the workforce? In other words, after I partial out the portion of the variance in income that is accounted for by years in the workforce, will years of education still be able to help us predict income? Which of these two independent variables will be the stronger predictor of income? And will each make a unique contribution in explaining variance in income?

To answer these questions, I use the SPSS statistical software package to analyze my data. (*Note*: With only 10 cases in my sample, it is not wise to run a multiple regression. I am doing so for illustration purposes only. When conducting multiple regression analyses, you should have at least 30 cases plus 10 cases for each predictor variable in the model.) I begin by computing the Pearson correlation coefficients for all three of the variables in the model. The results are presented in Table 13.4.

These data reveal that both level of education and years in the workforce are correlated with monthly income (r = .826 and r = .695 for education and workforce with income, respectively). In Table 13.4, we can also see that there is a small-to-moderate correlation between our two predictors, years of education and years in the workforce (r = .310). Because this correlation is fairly weak, we can infer that both of these independent variables may predict education level.

TABLE 13.3 Income, years of employment, and education level data

	Education Level (X_1) (in Years)	Years in Workforce (X_2)	Monthly Income (Y) (in Thousands of Dollars)
Case 1	6	10	1
Case 2	8	14	1.5
Case 3	11	8	1
Case 4	12	7	2
Case 5	12	20	4
Case 6	13	15	2.5
Case 7	14	17	5
Case 8	16	22	6
Case 9	16	30	10
Case 10	21	10	8
Mean	12.90	15.00	4.10
Standard Deviation	4.25	7.20	3.12
Correlation With Income	r = .83	r = .70	

Remember that in a multiple regression, we've got multiple predictor variables trying to explain variance in the dependent variable. For a predictor variable to explain variance in a dependent variable, it must be *related* to the dependent variable (see Chapter 12 and the discussion on the coefficient of determination). In our current example, both of our predictor variables are strongly correlated with our dependent variable, so this condition is met. In addition, for each of our predictor variables to explain a *unique*, or *independent* portion of the variance in the dependent variable, our two predictor variables cannot be too strongly related to *each other.* If our two predictor variables are strongly correlated with each other, they will overlap each other and will not be able to explain unique portions of variance in the dependent variable.

For an illustration of how uncorrelated predictor variables explain unique portions of variance in a dependent variable, see Figure 13.4. In this illustration, 100 percent of the variance in the dependent variable is contained within the hexagon. The two shaded ovals represent two independent variables. Notice that each of the ovals overlaps with the hexagon, but the ovals do not overlap with each other. The portion of each independent variable (i.e., oval) that is overlapping with the dependent variable (i.e., hexagon) is the portion of variance in the dependent variable that is *explained* by the independent variables. In other words, each independent variable is explaining a *unique* portion of the variance in the dependent variable.

Contrast that with Figure 13.5, in which two correlated predictor variables each explain smaller unique portions of variance in the dependent variable. In this illustration, the two ovals representing the predictor variables overlap with each other, signifying that the two predictor variables are correlated with each other. Notice that although each of the ovals overlaps with the hexagon (i.e., each predictor variable is correlated with the dependent variable), there is quite a bit of overlap between the two ovals within the hexagon. The space where the two ovals overlap inside of the hexagon indicates the **shared variance** between the two predictor variables in their explanation of shared variance in the dependent variable. The part of each oval that is inside the hexagon but not overlapping with the other oval is the **unique variance** in the dependent variable that is being explained by the independent variable. You may notice that the portion

TABLE 13.4 Correlations among variables in the regression model

	Years of Education	Years in Workforce	Monthly Income
Years of Education	1.000		
Years in Workforce	.310	1.000	
Monthly Income	.826	.695	1.000

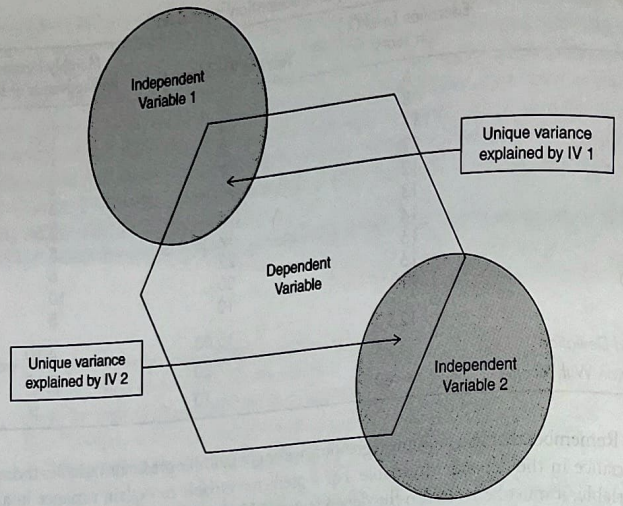

FIGURE 13.4 Portions of unique variance in the dependent variable explained by two uncorrelated independent variables.

of unique variance that is explained by the second independent variable is quite large, but the unique variance explained by the first independent variable is very small.

For example, suppose that scores on a reading test are strongly correlated with scores on a writing test ($r = .70$). Now suppose that I want to use reading and writing test scores to predict students' grades in English class. Because reading and writing test scores are so highly correlated with each other, I will probably not explain any more of the variance in English class grades

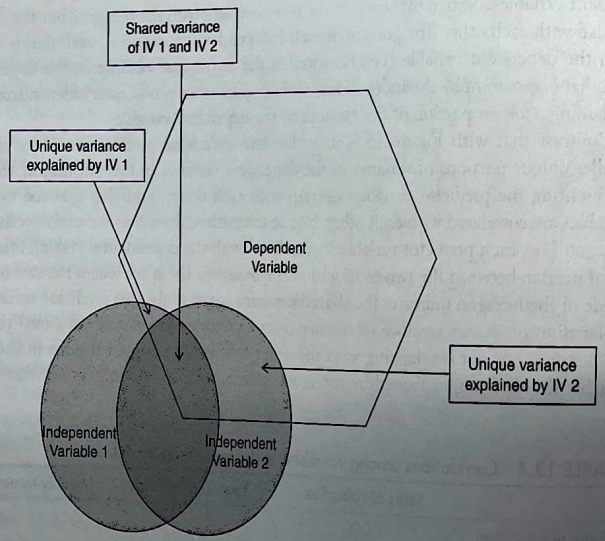

FIGURE 13.5 Portions of unique and shared variance in the dependent variable explained by two correlated independent variables.

using both predictor variables than if I use just one or the other. In other words, once I use reading test scores to predict English class grades, adding writing test scores to my regression model will probably not explain any more of the variance in my dependent variable, because reading and writing test scores are so closely related to each other. This is the concept that is represented graphically in Figure 13.5. Having strong correlations among predictor variables is called **multicollinearity** and it can cause problems in multiple regression analysis because it can make it difficult to identify the unique relation between each predictor variable and the dependent variable.

Returning to our example of using education level and years in the workforce to predict monthly income, when I conduct the regression analysis using SPSS, I get the results presented in Table 13.5. I have included several arrows to highlight particularly important statistics in the output and numbered each arrow to correspond with the discussion of these statistics. There are a variety of results produced with a multiple regression model. These results have been organized into three sections in Table 13.5. I have labeled the first section "Variance Explained." Here, we can see that we get an "*R*" value of .946 (Note 1 in Table 13.5). This is the **multiple correlation coefficient (*R*)**, and it provides a measure of the correlation between the two predictors *combined* and the dependent variable. It is also the correlation between the observed value of *Y* and the predicted value of $Y(\hat{Y})$. So together, years of education and years in the workforce have a very strong correlation with monthly income. Next, we get an "*R* Squared" value (symbolized as R^2) (Note 2). This is essentially the coefficient of determination (see Chapter 12) for my combined predictor variables and the dependent variables, and it provides us with a percentage of variance explained. So years of education and years in the workforce, combined, explain 89.6 percent of the variance in monthly income. When you consider that this leaves only about 10 percent of the variance in monthly income unexplained, you can see that this is a very large amount of variance explained. The R^2 statistic is the measure of effect size used in multiple regression. Because it is a measure of variance explained (like r^2 in correlation and eta squared in ANOVA), it provides a handy way of assessing the practical significance of the relation of the predictors to the

TABLE 13.5 Annotated SPSS output for regression analysis examining years of education and years in the workforce predicting monthly income

Variance Explained

R	*R* Squared	Adjusted *R* Squared	Std. Error of the Estimate
Note 1 → .946	Note 2 → .896	.866	1.1405

ANOVA Results

	Sum of Square	*df*	Mean Square	*F* Value	*p* Value
Regression	78.295	2	39.147	30.095	.000
Residual	9.105	7	1.301		
Total	87.400	9		Note 3	

Regression Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	*t* Value	*p* Value
Intercept	−5.504 ← Note 4	1.298		−4.241	.044 Note 9
Years Education	.495 ← Note 5	.094	Note 7 → .676	5.270	.001 ←
Years Work	.210 ← Note 6	.056	→ .485 Note 8	3.783	.007 ← Note 10

dependent variable. In this example, the effect size is large, suggesting practical significance as well as statistical significance. The "Adjusted *R* Squared" accounts for some of the error associated with multiple predictor variables by taking the number of predictor variables and the sample size into account, thereby adjusting the R^2 value down a little bit. Finally, there is a standard error for the *R* and R^2 values (see Chapter 6 for a discussion of standard errors).

Moving down the table to the "ANOVA Results" section, we get some statistics that help us determine whether our overall regression model is statistically significant. This section simply tells us whether our two predictor variables, combined, are able to explain a statistically significant portion of the variance in our dependent variable. The *F* value of 30.095, with a corresponding *p* value of .000, reveals that our regression model is statistically significant (Note 3 in Table 13.5). In other words, the relationship between years of education and years in the workforce combined (our predictor variables) and monthly income (our dependent variable) is statistically significant (i.e., greater than zero). Notice that these ANOVA statistics are quite similar to those presented in Chapter 10 on gender and GPA predicting valuing of English coursework among high school students. The sum of squares model in Table 10.2 corresponds to the sum of squares regression in Table 13.5. In both cases, we have sums of squares associated with the *combined predictors*, or the overall model.

Similarly, the sum of squares error in Table 10.2 is analogous to the sum of squares residual in Table 13.5. That is because residuals are simply another form of error. Just as the overall *F* value in Table 10.2 is produced by dividing the mean squares for the model by the mean square error, the overall *F* value produced in Table 13.5 is produced by dividing the mean square regression by the mean square residual. In both cases, we get an *F* value and a corresponding significance test, which indicates whether, overall, our predictors are significantly related to our dependent variable.

Finally, in the third section of Table 13.5, we get to the most interesting part of the table. Here we see our intercept (Note 4) and the regression coefficients for each predictor variable. These are the pieces of the regression equation. We can use these statistics to create the regression equation:

$$\hat{Y} = -5.504 + .495X_1 + .210X_2$$

where $\hat{Y}$ is the predicted value of *Y*,
X_1 is the value of the years of education variable,
X_2 is the value of the years in the workforce variable.

The unstandardized regression coefficients (Notes 5 and 6) can be found in the column labeled "*B*." Because years of education and years in the workforce are variables with different standard deviations, it is difficult to compare the size of the unstandardized regression coefficients. The variables are simply measured on different scales, making comparisons difficult. However, in the column labeled "Beta", the **standardized regression coefficients** are presented (Notes 7 and 8). These regression coefficients have been standardized, thereby converting the unstandardized coefficients into coefficients with the same scale of measurement (*z* scores; see Chapter 5 for a discussion of standardization). Here we can see that the two predictors are fairly close in their strength of relation to the dependent variable, but years of education is a bit stronger than years of work. In the next two columns, labeled "*t* Value" and "*p* Value," we get measures that allow us to determine whether each predictor variable is statistically significantly related to the dependent variable. Recall that earlier, in the ANOVA section of the table, we saw that the two predictor variables *combined* were significantly related to the dependent variable. Now we can use *t* tests to see if the slope for *each* predictor variable is significantly different from zero. The *p* values associated with each predictor variable are much smaller than .05, indicating that each of my independent variables is a significant predictor of my dependent variable. So both years of education (Note 9) and years in the workforce (Note 10) are statistically significant predictors of monthly income.

It is important to note that in this last section of Table 13.5, each regression coefficient shows the strength of the relationship between the predictor variable and the outcome variable *while controlling for the other predictor variable.* Recall that in the simple regression model with one predictor variable, I found that there was a relationship between years of education and monthly income. One of my questions in the multiple regression model was whether this education–income link would remain statistically significant when controlling for, or partialing out, the effects of years in the workforce. As the results presented in Table 13.5 indicate, even when controlling for the effects of years in the workforce, years of education is still a statistically significant predictor of monthly income. Similarly, when controlling for years of education, years in the workforce predicts monthly income as well.

For a brief video demonstrating how to read SPSS output for a multiple regression analysis, see the website that accompanies this book.

As you can see, multiple regression provides a wealth of information about the relations between predictor variables and dependent variables. Amazingly, in our previous example, we just scratched the surface of all that can be done with multiple regression analysis. Therefore, I strongly encourage you to read more about multiple regression using the references on regression that are provided in the bibliography. I also want to caution you about how to interpret regression analyses, whether they be simple or multiple regressions. Despite the uses of such terms as *predictor* and *dependent* variables, it is important to remember that regression analysis is based on good old correlations. Just as correlations should not be mistaken for proof of causal relationships between variables, regression analyses cannot prove that one variable, or set of variables, causes variation in another variable. Regression analyses can reveal how sets of variables are *related* to each other, but cannot prove causal relations among variables.

Example: Predicting the Use of Self-Handicapping Strategies

Sometimes students engage in behaviors that actually undermine their chances of succeeding academically. For example, they may procrastinate rather than study for an upcoming test, or they may spend time with their friends when they should be doing their homework. These behaviors are called "self-handicapping" because they actually inhibit students' chances of succeeding. One reason that students may engage in such behaviors is to provide an explanation for their poor academic performance, should it occur. If students fear that they may perform poorly on an academic task, they may not want others to think that the reason for this poor performance is that they lack ability, or intelligence. So some students strategically engage in self-handicapping to provide an alternative explanation for their poor performance. That is why these behaviors are called self-handicapping *strategies.*

Because self-handicapping strategies can undermine academic achievement and may be a sign of academic withdrawal on the part of students, it is important to understand the factors that are associated with the use of these strategies. Self-handicapping represents a concern with not appearing academically unable, even if that means perhaps sacrificing performance. Therefore, engaging in self-handicapping behaviors may be related to students' goals of avoiding appearing academically unable to others. In addition, because self-handicapping may be provoked by performance situations in which students expect to fail, it perhaps occurs more commonly among lower achieving students, who have a history of poor academic performance. Moreover, it is reasonable to suspect that when students lack confidence in their academic abilities, they will be more likely to use self-handicapping strategies. Finally, there may be gender differences in how concerned high school students are with appearing academically unable to others. Therefore, I conducted a multiple regression analysis to examine whether avoidance goals, self-efficacy, gender, and GPA, as a group and individually, predicted the use of self-handicapping strategies.

My colleague, Carol Giancarlo, and I collected data from 464 high school students. We used surveys to measure their self-reported use of self-handicapping strategies. In addition, the survey contained questions about their desire to avoid appearing academically unable (called "avoidance goals") and their confidence in their ability to perform academically (called "self-efficacy"). We also collected information about the students' gender (i.e., whether they were boys or girls) and their overall GPA in high school. Self-handicapping, avoidance goals, and self-efficacy were all measured using a 1–5 scale. Low scores indicated that students did not believe the items were true for them (i.e., they did not use self-handicapping strategies, were not confident in their abilities, were not concerned with trying to avoid appearing academically unable), whereas high scores indicated the opposite. Gender was a "dummy"-coded categorical variable (boys = 1, girls = 0), and GPA was measured using a scale from 0 to 4.0 (0 = F, 4.0 = A).

Once again, I used SPSS to analyze my data. The results of this multiple regression analysis are presented in Table 13.6. In the first section of the table, "Variance Explained," there is an R value of .347, and an R^2 value of .12. These statistics tell us that the four predictor variables combined have a moderate correlation with self-handicapping (multiple R = .347) and explain the variance in handicapping. This R^2 value is reduced to .113 when adjusted for the error associated with multiple predictor variables. In the second section of the table, "ANOVA Results," I see that I have an F value of 15.686 and a corresponding p value of .000. These results tell me that, as a group, my four predictor variables explain a statistically significant portion of the variance in self-handicapping. In other words, my overall regression model is statistically significant.

In the last section of the table, I find my unstandardized regression coefficients (column labeled "B") for each predictor variable in the model, as well as my intercept. These tell me that GPA and self-efficacy are negatively related to self-handicapping, whereas gender and avoidance goals are positively related to self-handicapping. Scanning toward the right side of the table, I find the standardized regression coefficients (column labeled "Beta"). These coefficients, which are all converted to the same standardized scale, reveal that GPA and self-efficacy appear to be more strongly related to self-handicapping than are avoidance goals and, in particular,

TABLE 13.6 Multiple regression results for predicting self-handicapping

Variance Explained				
	R	*R* Squared	Adjusted *R* Squared	Std. Error of the Estimate
	.347	.120	.113	.9005

ANOVA Results					
	Sum of Squares	*df*	Mean Square	*F* Value	*p* Value
Regression	50.877	4	12.719	15.686	.000
Residual	372.182	459	.811		
Total	423.059	463			

Regression Coefficients					
	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	*t* Value	*p* Value
Intercept	3.630	.264		13.775	.000
Avoidance Goals	.132	.045	.130	2.943	.003
Grades (GPA)	−.254	.054	−.209	−4.690	.000
Gender	.105	.085	.055	1.234	.218
Self-Efficacy	−.232	.052	−.198	−4.425	.000

gender. Notice that standardized and unstandardized regression coefficients can be either positive or negative. To make sense of them, it is important to remember how your variables are measured. In the current example, higher values of self-efficacy are actually a *bad* outcome, because self-handicapping is harmful behavior. With this in mind, the negative regression coefficient between GPA and self-handicapping tells us that higher academic achievement is associated with *less* self-handicapping, which is good. Conversely, a positive regression coefficient indicates that higher values on the predictor variable are associated with higher levels of self-handicapping. In this example, higher levels of avoidance goals predict higher levels of self-handicapping.

Continuing to scan toward the right side of the table, I find my t values and p values for each coefficient. These tell me which of my independent variables are statistically significant predictors of self-handicapping. The p values tell me that all of the independent variables, except for gender, are significant predictors of handicapping.

So what can we make of these results? First, my predictors explain a significant percentage of the variance in self-handicapping, although not a particularly large percentage (about 11 percent). Second, as we might expect, students with higher GPAs report engaging in less self-handicapping behavior than students with lower GPAs. Third, students with more confidence in their academic abilities engage in less self-handicapping than do students with less confidence in their abilities. Fourth, students who are concerned with not appearing academically unable in school are more likely to use self-handicapping strategies than are students without this concern. Finally, boys and girls do not differ significantly in their reported use of self-handicapping strategies. Although boys scored slightly higher than girls on the handicapping items (we know this because the regression coefficient was positive, and the gender variable was coded boys = 1, girls = 0), this difference was not statistically significant.

It is important to remember that the results for each independent variable are reported while controlling for the effects of the other independent variables. So the statistically significant relationship between self-efficacy and self-handicapping exists even when we control for the effects of GPA and avoidance goals. This is important, because one may be tempted to argue that the relationship between confidence and self-handicapping is merely a by-product of academic achievement. Those who perform better in school *should* be more confident in their abilities, and therefore *should* engage in less self-handicapping. What the results of this multiple regression reveal is that there is a statistically significant relationship between self-efficacy and self-handicapping *even after controlling for the effects of academic performance.* Confidence is associated with less self-handicapping *regardless* of one's level of academic achievement. Similarly, when students are concerned with not appearing incompetent in school (avoidance goals), *regardless* of their actual level of achievement (GPA), they are more likely to engage in self-handicapping behavior. The ability to examine both the combined and independent relations among predictor variables and a dependent variable is the true value of multiple regression analysis.

Writing it Up

To write up the results of the multiple regression summarized in Table 13.6 for a professional journal or conference, I would only need a few sentences: "A multiple regression analysis was conducted to examine the predictors of self-handicapping. Four predictors were simultaneously entered into the model: avoidance goals, GPA, gender, and self-efficacy. Together, these predictors accounted for 11 percent of the variance in self-handicapping. All of these variables except for gender (β = .06, p < .218) were significant predictors of self-handicapping. GPA (β = −.21, p < .001) and self-efficacy (β = −.20, p < .001) were the strongest predictors and were negatively associated with self-handicapping, whereas avoidance goals were positively associated with self-handicapping (β = .13, p = .003)."

Worked Examples

In this section I present two worked examples, one for a simple linear regression with a single predictor variable and one for multiple regression. For the first example, suppose that my independent variable, *X*, is the number of sugary drinks consumed per month by children between the ages of 10 and 15. Suppose that my dependent variable, *Y*, is weight. I collect data from 10 children, 5 boys and 5 girls, at each age between 10 and 15, giving me a total sample of 60 children (30 boys, 30 girls). I calculate the mean and standard deviation for the number of sugary drinks consumed in a month and weight for the total sample, and I get the data presented in Table 13.7.

With the information in Table 13.7, I can calculate the regression coefficient:

$$b = .59 \times \frac{21.29}{6.22}$$

$$b = .59 \times 3.42$$

$$b = 2.02$$

This regression coefficient tells me that for every increase of one sugary beverage consumed per month, there is a corresponding increase of 2.02 pounds, on average. Now that we have the regression coefficient, we can calculate the intercept:

$$a = 112.12 - (10.88)(2.02)$$

$$a = 112.12 - 21.98$$

$$a = 90.14$$

This intercept tells me that when the number of sugary drinks consumed per month is zero, we would predict the child's weight to be 90.14 pounds.

Now that we know our regression coefficient and our intercept, we can use these to predict *Y* values for various *X* values, using the formula presented in Table 13.1. For example, how much would we expect a child who consumes 10 sugary drinks per month to weigh?

$$\hat{Y} = 2.02(10) + 90.14$$

$$\hat{Y} = 20.20 + 90.14$$

$$\hat{Y} = 110.34$$

We would predict that a child who consumes 10 sugary drinks per month would weigh 110.34.

TABLE 13.7 Means and standard deviations for the number of sugary drinks consumed per month by children, and the correlation between sugary drinks consumed and weight

	Mean	Standard Deviation
Sugary Drinks per Month (*X*)	10.88	6.22
Weight (*Y*)	112.12	21.29
Correlation between *X* and *Y*	$r = .59$	

As we discussed earlier in this chapter, our prediction of 110.34 pounds for a child who has 10 sugary drinks per month is just an educated guess based on a correlation between these two variables. We also know that there are some kids who have 10 sugary drinks per month but weigh 75 pounds, and others who weight 150. Because the predictions based on regression coefficients often have a fair amount of error (i.e., residuals), and because we are smart people who have lived a little, we also know that there are several things that are associated with weight besides how many sugary drinks children consume. In our current example, we might expect the older kids (i.e., 14- and 15-year-olds) to weigh more than the younger kids (i.e., 10- and 11-year-olds), so age is probably a predictor of weight. Come to think of it, age might also be related to how many sugary drinks kids consume. As children get older, they often have more independence from their parents, and also get a little spending money from their parents. When left alone with money in their pockets, early adolescents often spend that money on sugary drinks.

For a brief video demonstration of how to calculate and interpret a regression coefficient, intercept, and predicted value for *Y*, refer to the website that accompanies this book.

So let's compute another regression, this time using two predictor variables (age and sugary drinks consumed per month) to predict weight. Because the calculations for a multiple regression are a little more complicated due to the shared variance between the predictor variables, I am going to let SPSS do the calculations for us. I'm going to run the regression twice, first with only the number of sugary drinks predicting weight, and then with both predictors (i.e., sugary drinks and age) in the regression model. This will show you how adding a second predictor variable affects the strength of the regression coefficient between the first independent (i.e., predictor) variable and the dependent variable in a regression. But first, let's take a look at Table 13.8 to see the bivariate correlations between the three variables that are going into our regression model.

As you can see from the correlations presented in Table 13.8, as age increased, so did weight, on average ($r = .55$). So our second predictor variable, age, is related to our dependent variable, weight. In addition, we can see that our two predictor variables (age and drinks consumed) are also positively correlated ($r = .44$). Because of this correlation between the two predictor variables, there is a good chance that when they are both put into a regression model, their *independent* association with the dependent variable (weight) will be smaller than the bivariate correlations between each predictor variable and the dependent variable.

Let's take a look. In Table 13.9 I present the results of two regression analyses using SPSS.[2] In the first regression I used sugary drinks consumed per month to predict weight. This is exactly the same analysis that we conducted earlier, and SPSS confirmed our prior results. The unstandardized regression coefficient (b) = 2.09 (*B* in the table), and the intercept (a) = 90.14. Notice that the "Beta" in this SPSS output, which is the *standardized* regression coefficient, is .59, and this is the exact same value as our bivariate correlation (r) between drinks consumed and weight. When we only have a single predictor variable in a regression, the correlation coefficient is the same as the standardized regression coefficient, also known as the beta. You can also see that the *R* Squared value, which is the percentage of variance in the dependent variable that is explained by the independent variable, is the same as the coefficient of determination: $(.59)^2 = .35$. So 35 percent of the variance in weight can be explained by the sugary drinks consumed variable.

Now if you look at the SPSS output for the second regression in Table 13.9, which is a multiple regression with two predictors of weight, there are a few interesting differences to note.

TABLE 13.8 Correlations between weight, sugary drinks consumed, and age

	Weight	Sugary Drinks	Age
Weight	–		
Sugary Drinks	.59	–	
Age	.55	.44	–

TABLE 13.9 SPSS output for two regression analyses predicting weight

Regression 1: Single Predictor

R	*R* Squared	Adjusted *R* Squared	Std. Error of Estimate
.59	.35	.34	17.30

ANOVA Results

	Sum of Squares	*df*	Mean Square	*F* Value	*p* Value
Regression	9392.41	1	9392.41	31.39	.000
Residual	17355.77	58	299.24		
Total	26748.18	59			

	Unstandardized Coefficients		Standardized Coefficient		
	B	Std. Error	Beta	*t*	*p* Value
Intercept	90.14	4.53		19.88	.000
Drinks per Month	2.09	.36	.59	5.60	.000

Regression 2: Multiple Predictors

R	*R* Squared	Adjusted *R* Squared	Std. Error of Estimate
.67	.45	.43	16.05

ANOVA Results

	Sum of Squares	*df*	Mean Square	*F* Value	*p* Value
Regression	12057.31	2	6028.66	23.39	.000
Residual	14690.87	57	257.74		
Total	26748.18	59			

	Unstandardized Coefficients		Standardized Coefficient		
	B	Std. Error	Beta	*t*	*p* Value
Intercept	41.45	15.68		2.64	.011
Drinks per Month	1.50	.38	.44	3.99	.001
Age	4.35	1.35	.35	3.22	.002

First, the *R* and *R* Squared values increased, compared to the output for the first regression in Table 13.9. Adding a second predictor to the regression model, age, increased the percentage of variance explained in weight from 35 percent to 45 percent. Second, notice that the regression coefficients for "Drinks per Month," both standardized and unstandardized, became smaller in the second regression. That is because when we added the second variable, age, to the model, it overlapped (i.e., was correlated with) the first predictor variable, and therefore cut into the percentage of variance that first predictor variable explained in the dependent variable, weight.

Wrapping Up and Looking Forward

The overlap between correlation (Chapter 12) and regression are plain. In fact, simple linear regression provides a statistic, the regression coefficient, that is simply the unstandardized version of the Pearson correlation coefficient. What may be less clear, but equally important, is that regression is also a close relative of ANOVA. As you saw in the discussion of Table 13.6, regression *is* a form of analysis of variance. Once again, we are interested in dividing up the variance of a dependent variable and explaining it with our independent variables. The major difference between ANOVA and regression generally involves the types of variables that are analyzed, with ANOVA using categorical independent variables and regression using continuous independent variables. As you learn more about regression on your own, you will learn that even this simple distinction is a false one, as categorical independent variables can be analyzed in regression.

In the next chapter, we turn our attention to nonparametric statistics, particularly the chi-square test of independence. Unlike the inferential statistics we have discussed in most of the previous chapters of the book, nonparametric tests are not based on the assumption that dependent variables are normally distributed or measured using an interval scale. In the last chapter of the book we will return to a consideration of statistical techniques—factor analysis and reliability analysis—that are based on correlation among variables.

Work Problems

Suppose I want to know something about the study habits of undergraduate college students. I collect a random sample of 200 students and find that they spend 12 hours per week studying, on average, with a standard deviation of 5 hours. I am curious how their social lives might be associated with their studying behavior, so I ask the students in my sample how many other students at their university they consider "close friends." The sample produces an average of 6 close friends with a standard deviation of 2. Please use this information to answer the following questions. The correlation between these two variables is –.40.

1. Assume that "hours spent studying" is the *Y* variable and "close friends" is the *X* variable. Calculate the regression coefficient (i.e., the slope) and wrap words around your results. What, exactly, does this regression coefficient tell you?
2. What would the value of the standardized regression coefficient be in this problem? How do you know?
3. Calculate the intercept and wrap words around your result.
4. If you know that somebody studied had 10 close friends, how many hours per week would you expect them to study?
5. What, exactly, is a residual (when talking about regression)?
6. Regression is essentially a matter of drawing a straight line through a set of data, and the line has a slope and an intercept. In regression, how is it decided where the line should be drawn? In other words, explain the concept of least squares.
7. Now suppose that I add a second predictor variable to the regression model: hours per week spent working for money. And suppose that the correlation between hours spent working and hours spent studying is –.50. The correlation between the two predictor variables (number of close friends and hours spent working for money) is –.30.
   a. What effect do you think the addition of this second predictor variable will have on the overall amount of variance explained (R^2) in the dependent variable? Why?
   b. What effect do you think the addition of this second predictor variable will have on the strength of the regression coefficient for the first predictor variable, compared to when only the first predictor variable was in the regression model? Why?

For answers to these work problems, and for additional work problems, please refer to the website that accompanies this book.

Notes

1 It is also possible to use categorical predictor variables with more than two categories in a multiple regression, but these variables must first be transformed into multiple dichotomous predictor variables and entered into the regression model as separate predictor variables.

2 I have modified the SPSS output to make it easier to read and rounded values to the nearest hundredth.