12

Correlation and Regression

In the last chapter, we looked at the idea of the association between two categorical variables. In doing so, we explored the idea of two variables being tied to one another by something other than chance. In this chapter, we extend our understanding of the idea of association as we take up two procedures appropriate for situations involving two interval/ratio level variables. First, we'll examine Pearson's *r*, or simple correlation analysis. Following that, we'll explore a related procedure known as regression analysis. As a prelude to both, we'll explore the use of scatter plots as a means to visually represent the association between two variables.

Before We Begin

As you encounter this twelfth and final chapter, I'm going to ask you to explore some additional dimensions related to relationships. First, you're going to be introduced to the notions of the strength and direction of relationships. In doing so, you'll deal with the question of how closely tied one variable is to the other, as well as how the variables vary together, so to speak. You'll also be dealing with the matter of prediction—the idea that if you know something about the way two variables are related, you're then in a position to make predictions. For example, if you have some knowledge as to the strength and direction of the relationship between two variables, X and Y, it is possible to make a prediction about the likely value of Y, given a certain value of X.

All of that may strike you as a little bit abstract as we begin this chapter, but I can assure you that you're already familiar with a lot of concepts that you're going to encounter. Means, standard deviations, and Z scores are about to reenter the picture. If you think you're a little rusty on some of those concepts, particularly standard deviations or Z scores, take a little time to reread previous material on the topics. The review will serve you well.

Scatter Plots

A **scatter plot** is an extremely useful tool when it comes to looking at the association between two variables. In short, a scatter plot allows us to simultaneously view the values of two variables on a case-by-case basis. A typical example used to illustrate the utility of a scatter plot is one involving the association between height and weight. Table 12-1 shows a hypothetical distribution of values of those variables (height and weight) for 20 cases.

A visual representation of the same data in the form of a scatter plot is shown in Figure 12-1. Height measurement values are shown along the horizontal or *X*-axis of the graph; weight measurement values are shown along the vertical or *Y*-axis of the graph. Focusing on case number 1, shown in the lower left corner of the scatter plot, we can interpret the point as reflecting a person (case) with a height of 59 inches (or 4' 11") and a weight of 92 pounds. Each of the 20 points can be interpreted in the same fashion—a reflection of the values of two variables (height and weight) for a given case.

Note that the scales along the *X*- and *Y*-axes are different. The variable of height is expressed in inches, but the variable of weight is expressed in pounds. As you learned earlier when you encountered *Z* scores, though, the fact that the scales are based on different units of measurement is not a deterrent to statistical analysis. Indeed, correlation analysis is a technique that is perfectly suited for such situations. All of that in good time, though. For the moment, let's take a closer look at scatter plots and what they can tell us.

Table 12-1 Height/Weight Data for 20 Cases (Young Adult Females): Raw Scores

Case	Height (Inches)	Weight (Pounds)
1	59	92
2	61	105
3	61	100
4	62	107
5	62	114
6	63	112
7	63	120
8	63	130
9	64	132
10	64	137
11	65	132
12	65	138
13	65	120
14	66	136
15	66	132
16	67	140
17	67	143
18	68	139
19	68	134
20	69	153

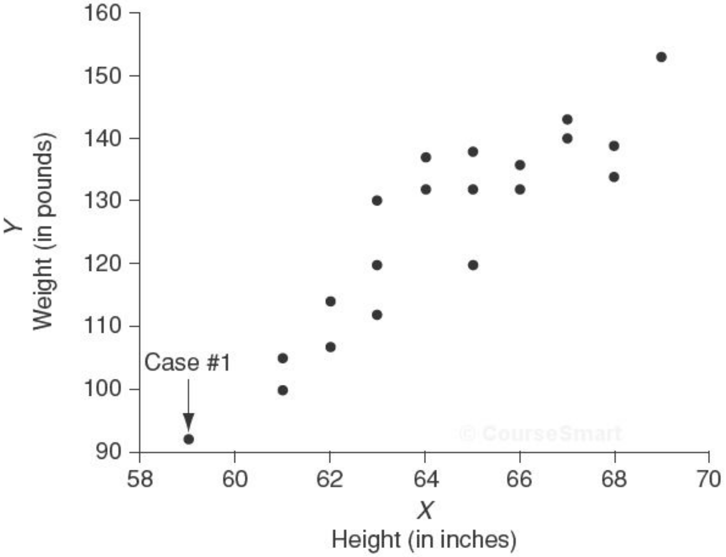

Figure 12-1 Height/Weight Data for 20 Cases (Young Adult Females): Scatter Plot

LEARNING CHECK

Question: What is a scatter plot?
Answer: It's a visual representation of the values of two variables on a case-by-case basis.

Linear Associations: Direction and Strength

When two variables (Variable X and Variable Y) are associated, they can be associated in several ways. A scatter plot can provide a graphic and concise statement as to the general relationship or association between two variables. In short, a scatter plot tells us something about the *direction* and *strength* of association. To better grasp the variety of relationships or associations that are possible, take a look at Figures 12-2 through 12-5. These illustrations reflect data on two variables—Variable X and Variable Y. As you consider the illustrations, don't worry about what the X and Y variables represent or how they might be measured. Just look at each axis as a scale that has low to high values. Treat the illustrations as abstract representations; focus on the general trends or associations that may or may not be reflected in the scatter plots.

One of the first observations we might make about the illustrations in Figure 12-2 is that they reflect *linear* patterns of association. To suggest that an association between two variables is linear is to suggest that the pattern could be described as approximating a straight line. Looking at both illustrations in Figure 12-2, however, we note that **linear associations** can take different forms.

LEARNING CHECK

Question: What is a linear association between two variables?
Answer: It's an association that can be described in general terms as approximating a straight line.

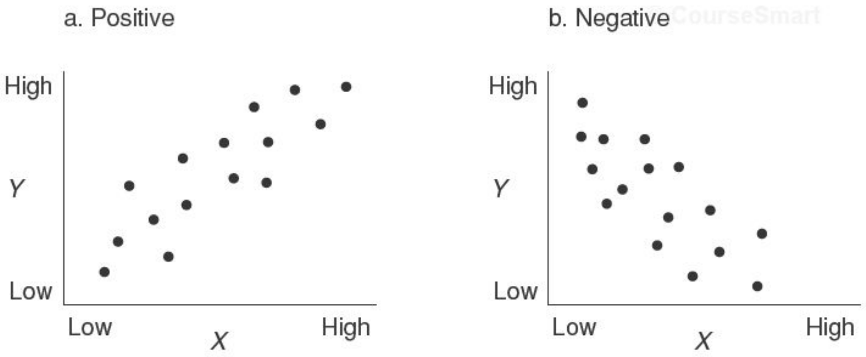

Figure 12-2 Moderate Positive (Direct) and Negative (Inverse) Associations

Direction of Association. Figure 12-2a depicts what we refer to as a **positive** or **direct association**. To say that two variables are related or associated in a positive or direct fashion is to say that they track together; it means that a high value on Variable X is generally associated with a high value on Variable Y, and a low value on Variable X is generally associated with a low value of Variable Y. If, however, the variables are related in a negative or inverse fashion, an opposite pattern appears (see Figure 12-2b). In a **negative** or **inverse association**, high values on Variable X are associated with low values on Variable Y, and low values on Variable X are associated with high values on Variable Y. In short, the variables track in opposite directions.

LEARNING CHECK

Question: What is a positive or direct association?
Answer: It's an association in which the variables track together. As one variable increases in value, the other variable increases in value. As one variable decreases in value, the other variable decreases in value.

Question: What is a negative or inverse association?
Answer: It's an association in which the variables track in opposite directions. As one variable increases in value, the other variable decreases in value. As one variable decreases in value, the other variable increases in value.

Strength of Association. Figure 12-3 presents similar patterns, but with one important difference. There is less dispersion of the points in the plots (when compared to the patterns shown in Figure 12-2), and the general trend (either positive or negative) is more easily detected. In that sense, the illustrations in Figure 12-3 reflect associations that are stronger than those represented in Figure 12-2.

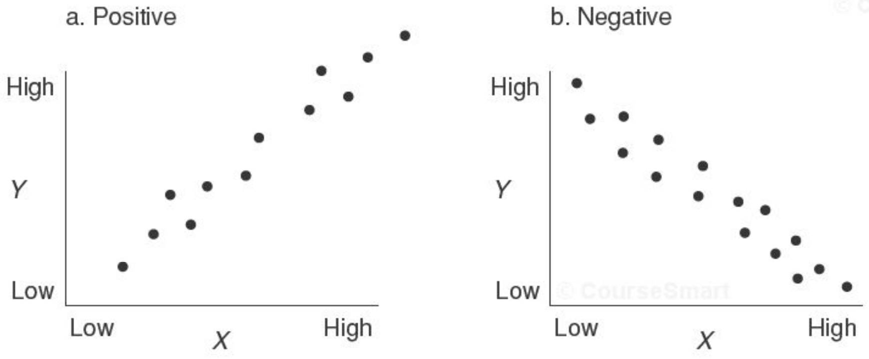

Figure 12-3 Stronger Positive and Negative Associations

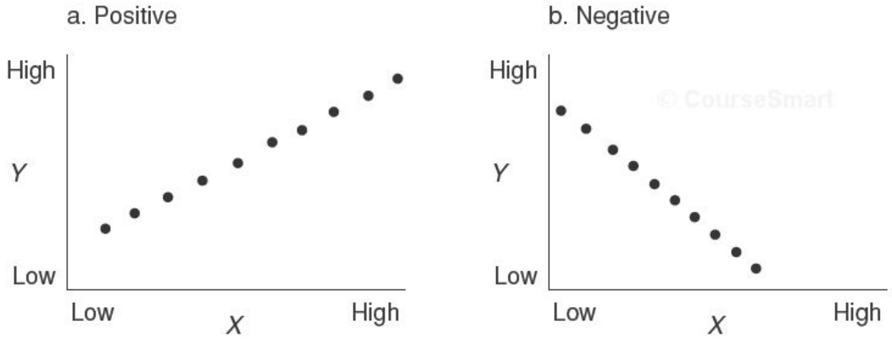

Figure 12-4 Perfect Positive and Negative Associations

Now take a look at Figure 12-4. Note that the association between the variables is shown as being even stronger. Indeed, the points in the scatter plots appear to be aligned in a straight line. Although such associations are rare in the real world, we could characterize the relationships shown in Figure 12-4 as **perfect associations**. They're perfect in the sense that the value on one variable could serve as a perfect or precise predictor of the value on the other variable.

As noted above, we rarely encounter perfect associations in the world of social science research. Even so, the idea of a perfect relationship is useful, because it helps us understand what is meant by **strength of association**. In many respects, the strength of an association is just an expression of how close we might be to being able to predict the value on one variable from knowledge of the value on another variable. Some associations are stronger than others in the sense that they come closer than others to the notion of perfect predictability.

LEARNING CHECK

Question: What is meant by the term *strength of association?*

Answer: It's an expression of the extent to which the value of one variable can be predicted on the basis of the value of another variable.

Other Types of Association

Finally, take a look at Figure 12-5. In Figure 12-5a, the points in the scatter plot are widely dispersed, and there isn't any discernable pattern. For all practical purposes, the relationship or association is *non-existent*. In the case of Figure 12-5b, a clear pattern is evident, but it's a **curvilinear association** (as opposed to the more linear relationships depicted in the previous illustrations). As the values of Variable *X* increase, the values of Variable *Y* also increase—up to a point. Eventually, however, the pattern begins to flatten and then reverses; as the values on Variable *X* increase, the values on Variable *Y* decrease. In short, a curvilinear association is one that is best described by a curved line.

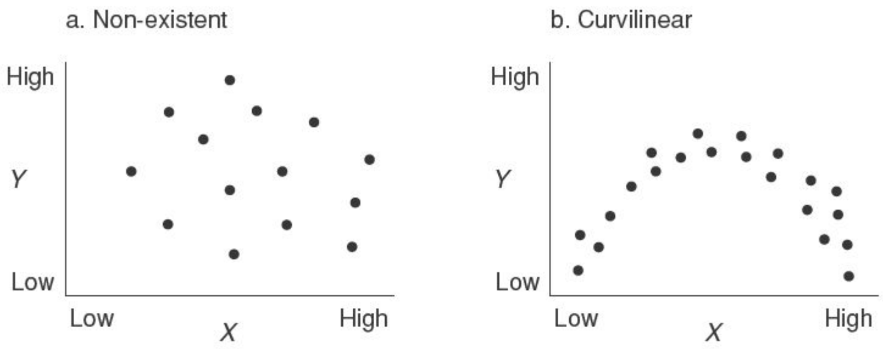

Figure 12-5 Non-existent and Curvilinear Associations

LEARNING CHECK

Question: In terms of a scatter plot, what is a curvilinear relationship?
Answer: It's a scatter plot in which the general pattern of the plot conforms to a curved line.

Question: In terms of a scatter plot, what does it mean to say that an association is non-existent?
Answer: It's a scatter plot in which there is no apparent association or pattern.

By now, you should be getting the picture. The association between two variables can take many forms. It can be positive, negative, weak, or strong. It can also be linear, curvilinear, or non-existent.

It's clear that a scatter plot can be a helpful tool in statistical analysis. It provides a visual statement about the general nature of an association and does so in a concise format. Statisticians, however, generally want something more exact—some sort of quantitative expression about the nature of associations—and for that they turn to correlation analysis.

Correlation Analysis

Correlation analysis is a technique developed by Karl Pearson; thus, it's often referred to as **Pearson's *r***. The popularity of Pearson's *r* stems from what it tells us about the direction and strength of association between two variables. When Pearson's *r* is calculated, the result will be a value that ranges from −1.0 to +1.0, depending on the direction and strength of association. Without going into the mathematics of the calculation just yet, let's take a closer look at what it means to say that the value of *r* has a known range between −1.0 and +1.0.

LEARNING CHECK

Question: What is Pearson's *r*?
Answer: It's a measure of the strength and direction of association between two variables.

Question: What is the range of *r*?
Answer: The value of *r* can range from –1.0 to +1.0.

Assuming that the two variables under investigation are related in a linear fashion (as opposed to a curvilinear fashion), the sign and value of *r* will tell us quite a lot about the relationship. A positive sign signals a positive or direct association; a negative sign signals a negative or inverse association. The closer the value is to +1.0 or –1.0, the stronger the association is between the two variables. An *r* value of +1.0 would indicate a perfect positive or direct association between the variables; an *r* value of –1.0 would indicate a perfect negative or inverse association between the variables. As I mentioned before, a perfect association, whether positive or negative, is one in which there would be perfect predictability. Knowledge of the value of one variable would allow us to make an exact prediction of the value of the other variable.

LEARNING CHECK

Question: If the value of *r* has a positive sign, what does that mean?
Answer: It means that the variables are associated in a positive or direct fashion.

Question: If the value of *r* has a negative sign, what does that mean?
Answer: It means that the variables are associated in a negative or inverse fashion.

Keep in mind that none of this necessarily indicates causation. Just because two variables appear to be closely associated with one another, it doesn't necessarily mean that one variable *causes* the other variable. Remember some of the points we covered in the last chapter. Recall that outside of highly controlled experimental research, it's virtually impossible to legitimately infer causality. Also keep in mind that what may look like a case of causality may be nothing more than the fact that two variables are expressions of a common concept. That said, we move on to our discussion of Pearson's *r*, or simple correlation analysis.

Two Variables: X and Y

Our discussion of scatter plots and Pearson's *r* has thus far revolved around the notion of two variables, usually referred to as Variable *X* and Variable *Y*.

We could have used the symbols Variable 1 and Variable 2, or almost any other designation, but as a matter of convention, we typically speak in terms of Variable *X* and Variable *Y*. Those notations appear time and time again in our discussions, so some additional commentary is warranted.

When researchers speak in terms of Variable *X* and Variable *Y*, they commonly propose some logical connection between the two. For example, the *X* variable is commonly regarded as the **independent variable**, and the *Y* variable is commonly regarded as the **dependent variable**. In the language of research, an independent variable is a variable that's presumed to influence another variable. The dependent variable, in turn, is the variable that's presumed to be influenced by another variable.

For example, it's common to assert that there's a connection between a person's level of education and level of income. Educational level would be treated as the independent variable, and level of income would be regarded as the dependent variable. In other words, education (independent) is thought to exert an influence on income (dependent).

☑ LEARNING CHECK

Question: What's the definition of an independent variable?
Answer: It's the variable that's presumed to influence another variable.

Question: What's the definition of a dependent variable?
Answer: It's the variable that's presumed to be influenced by another variable.

Once again, it's important to remember that the notion of one variable exerting an influence on another is not the same thing as pure causality. For example, it's easy to understand how one's level of education would have some connection to one's level of income, but it's difficult to imagine that education is the only variable that determines income. A person's level of education might be one influence on level of income, but that's hardly the same thing as saying level of education causes a person's level of income.

Although researchers rely on simple logic when it comes to identifying the independent and dependent variables, not all research situations are clear-cut. Consider the association between the level of unemployment in a community and the level of in-migration. The level of unemployment in a community may influence the amount of in-migration, but continued in-migration is likely to affect the level of unemployment. Job opportunities (expressed in low levels of unemployment) could attract significant numbers of job seekers, to the point that the level of unemployment is pushed upward. Much like the relationship between the temperature in a room and a thermostat, each variable has a way of affecting the other. Such relationships are said to be *reciprocal*—relationships or associations in which each variable is presumed to exert an influence on the other.

☑ LEARNING CHECK

Question: What is a reciprocal relationship?

Answer: A reciprocal relationship is one in which each variable is presumed to exert an influence on the other variable.

Given that, it's probably best to expand your thinking about Variable *X* and Variable *Y* as follows: When you can reasonably assign a variable's place in a logical sequence of events, it's reasonable to think in terms of Variable *X* as the independent variable and Variable *Y* as the dependent variable. When you can't make a reasonable assignment of place in a logical sequence of events, just think of Variable *X* and Variable *Y* as two variables—plain and simple, without regard for causality, logical sequencing, or anything else.

Later on, we'll deal with Variable *X* as the variable that you'd use to *predict* Variable *Y*, but all of that can wait until our discussion of regression. For the present, we'll continue with our discussion of simple correlation analysis as a measure of the association between Variable *X* and Variable *Y*. As before, we'll start with the logic.

The Logic of Correlation

In truth, correlation analysis takes many forms (such as multiple correlation or partial correlation); the one we're considering here is referred to as *simple correlation*. In short, simple correlation analysis allows us to measure the association between two interval/ratio level variables (assuming that the two variables, if associated, are associated in a linear fashion). The logic of correlation analysis traces back to the notion of *Z* scores (something you encountered in Chapter 2). You'll want to review the material in Chapter 2 if you think you're not quite up to speed on the topic. Assuming you feel comfortable with the concept, however, we can move to the topic of *Z* scores and what they allow us to do in the context of correlation analysis.

Earlier you learned how to transform a raw score into a *Z* score by finding the difference between a raw score and the mean of a distribution and dividing that difference by the standard deviation of the distribution. Just to refresh your memory on this point, consider the formula for a *Z* transformation and recall what it allows you to do. *Z* transformations allow you to convert the scores on different scales to a single scale based on *Z* scores (or points along the baseline of the normal curve).

$$Z = \frac{\text{Raw Score} - \text{Mean}}{\text{Standard Deviation}}$$

For example, look back at Figure 12-1. Note that the values along the horizontal and vertical axes are expressed in different scales or units of measurement: inches along the horizontal axis and pounds along the vertical axis. When we consider the raw scores of the points represented in the scatter plot, then, we are dealing with two different scales. The two sets of scores will be on the same scale, though, if they're transformed into Z scores. The same is true for any number of situations.

For example, student aptitude test scores (SAT scores) and grade point averages (GPAs), when expressed as raw scores, are based on very different underlying scales, but the raw scores can easily be transformed into Z scores to create a single scale of comparison. Data on education (expressed as the number of school years completed) and income (expressed in dollars) can share a common scale when transformed into Z scores. The list goes on. All we need is the mean and standard deviation of each distribution. It's a simple transformation: Subtract the mean from each raw score in the distribution, and divide each difference by the standard deviation.

Decades ago, Pearson discovered something very interesting about distributions of Z scores. In short, he discovered that it's possible to transform two distributions of raw scores (expressed as pairs of scores or values) into Z scores, perform some very minor calculations, and end up with a statistic that will always range between –1.0 and +1.0. What Pearson discovered became the basis for the computation of r.

In developing the correlation procedure, Pearson found that two distributions of closely associated Z scores that tracked together in a positive (direct) fashion would result in an r value approaching +1.0. He also discovered that two distributions of Z scores that were closely associated in a negative (inverse) direction would result in an r value approaching –1.0. All that's necessary is a couple of minor calculations, once the raw scores have been converted to Z scores. This brings us to the formula for r, so that's where we'll turn next.

The Formula for Pearson's r

Because the computational formula for r includes the steps necessary to convert raw scores to Z scores, it has a way of appearing extremely complex. Assuming you know the basis of Pearson's r (namely, the conversion of raw scores into Z scores), though, you're in a position to rely on a more *conceptual* formula—one I suspect you'll find very simple to follow. The heart of the more conceptual approach has to do with what we refer to as the *cross products of the Z scores*. That sounds like a mouthful until it's explained, so let's start with a look at the data presented in Table 12-2.

Table 12-2 shows pairs of values or scores associated with 10 cases. Columns 2 and 4 show the raw score distributions for the two variables, X and Y. The means and standard deviations of the raw score distributions are given at the bottom of the table. Columns 3 and 5 show the Z scores or transformations based on the associated raw scores. (Recall that these are calculated by subtracting the mean from each raw score and dividing by the standard deviation.) Case number 1, for example, has a raw score X value of 20 (shown in Column 2) and a Z score (Z_X) value of –1.49 (shown in Column 3). The raw score Y value for case number 1 is 105 (shown in Column 4), and the Z score (Z_Y) value of –1.57 (shown in Column 5).

The cross products are obtained by multiplying each Z_X value (the entry in Column 3) by the associated Z_Y value (the entry in Column 5). The results of the

Table 12-2 Cross Product Calculations for *X* and *Y* Variables (Positive Association)

(1) Case	(2) X	(3) Z_X	(4) Y	(5) Z_Y	(6) $Z_X \cdot Z_Y$
1	20	−1.49	105	−1.57	2.34
2	25	−1.16	126	−0.84	0.97
3	30	−0.83	122	−0.98	0.81
4	35	−0.50	130	−0.70	0.35
5	40	−0.17	155	0.17	−0.03
6	45	0.17	159	0.31	0.05
7	50	0.50	153	0.10	0.05
8	55	0.83	184	1.18	0.98
9	60	1.16	177	0.94	1.09
10	65	1.49	190	1.39	2.07
				Sum of Cross Products =	8.68

Mean of X = 42.50
Standard Deviation of X = 15.14

Mean of Y = 150.10
Standard Deviation of Y = 28.65

cross product multiplication are shown in Column 6 ($Z_X \cdot Z_Y$). Earlier I mentioned that the cross products of the *Z* scores lie at the heart of correlation analysis.

Now that you know how to derive the cross products, it's time to encounter the formula for the calculation of *r*. My guess is that you'll find it to be rather straightforward.

$$r = \frac{\sum(Z_X \bullet Z_Y)}{n - 1}$$

As in Table 12-2, the symbol Z_X denotes the *Z* scores for the *X* variable, and Z_Y denotes the *Z* scores for the *Y* variable. All you need to do is sum the cross products and divide the sum by the number of paired cases minus 1. The result is our calculated value of *r*. For the data presented in Table 12-2, the calculation is as follows:

$$r = \frac{\sum(Z_X \bullet Z_Y)}{n - 1}$$

$$r = \frac{8.68}{9}$$

$$r = +0.96$$

Note that we use $n - 1$ in the denominator of the formula. You should be aware, though, that some presentations of the formula rely on *n* alone. The difference in the two approaches traces back to the manner in which the standard deviation for each distribution was calculated (recall that the standard deviation is a necessary ingredient for the calculation of a *Z* score). As you may recall from Chapter 2, the choice of using $n - 1$ versus *n* in the denominator when

Table 12-3 Cross Product Calculations for *X* and *Y* Variables (Negative Association)

(1) Case	(2) X	(3) Z_X	(4) Y	(5) Z_Y	(6) $Z_X \bullet Z_Y$
1	20	−1.49	190	1.39	−2.07
2	25	−1.16	177	0.94	−1.09
3	30	−0.83	184	1.18	−0.98
4	35	−0.50	153	0.10	−0.05
5	40	−0.17	159	0.31	−0.05
6	45	0.17	155	0.17	0.03
7	50	0.50	130	−0.70	−0.35
8	55	0.83	122	−0.98	−0.81
9	60	1.16	126	−0.84	−0.97
10	65	1.49	105	−1.57	−2.34
				Sum of Cross Products =	−8.68

Mean of *X* = 42.50
Standard Deviation of *X* = 15.14

Mean of *Y* = 150.10
Standard Deviation of *Y* = 28.65

calculating the standard deviation of a sample is somewhat discretionary. The assumption in this text is that $n - 1$ was used in the calculation of the standard deviations of Variable *X* and Variable *Y*.

Just to make certain that you're on track with the notion of cross products and how they're used to develop a value for *r*, consider the data presented in Table 12-3—another example similar to the one you encountered in Table 12-2. In the case of Table 12-3, however, you'll note that as the raw scores (and the corresponding *Z* scores) for Variable *X* increase, those for Variable *Y* decrease, indicating a negative association. As expected, the resulting *r* value reflects the negative or inverse direction of the relationship:

$$r = \frac{\sum(Z_X \bullet Z_Y)}{n - 1}$$

$$r = \frac{-8.68}{9}$$

$$r = -0.96$$

LEARNING CHECK

Question: In the context of Pearson's *r*, what is a cross product; that is, how is a cross product computed?

Answer: A cross product is the result of multiplying a Z_X score by a Z_Y score. To obtain the Z_X and Z_Y scores, individual *X* and *Y* scores must first be converted to *Z* scores.

Earlier I mentioned that the computational formula for r can be rather threatening if you don't know what it really represents. Now you know, however, that the calculation (based on the cross products of Z scores) is actually quite straightforward. Still, you deserve to have a look at a typical computational formula for r—if only to convince yourself that the business of statistical analysis isn't always as complex as it might appear. The formula that follows, for example, is typical of how a computational formula for r might be presented:

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

Such a formula can be very handy if you're using a calculator to compute the value of r. Given the increasing use of computers and statistical software, however, the real issue is likely to be whether or not you have a solid understanding of what lies behind a procedure and how to interpret the results. In the case of Pearson's r, my guess is that the conceptual formula, based on the cross product calculations, gives you a better understanding of what's really involved in the calculation.

Application

We've covered the necessary information to move forward with a typical research application, so let's begin with the example presented in Table 12-4. Assume that information has been collected from a sample of 20 people on two different variables: level of education (the number of school years completed) and number of memberships in voluntary associations (clubs and organizations). We'll treat the education variable as the X or independent variable; the variable related to the number of memberships will be treated as the Y or dependent variable. The raw scores are shown as Variable X and Variable Y.

A quick examination of the data suggests that there's likely to be a positive association between the two variables. After all, the general pattern is one in which high levels of education are associated with a high number of association memberships. By the same token, low levels of education are generally associated with low numbers of memberships. The question, of course, is just how strong the association is. Rather than relying on the more complex computational formula, we'll move forward on the basis of the conceptual formula outlined earlier. Recall that we'll be working toward developing a value for r as a function of the cross products.

The first step is the conversion of the raw score distributions into Z score distributions. The mean and standard deviation (which are necessary ingredients in the conversion of raw scores to Z scores) are given at the bottom of the table. The Z score transformations, along with the cross products, are listed

Table 12-4 Raw Data and Cross Product Calculations for Educational Level and Association Memberships

	X	Y	Z_X	Z_Y	$Z_X \bullet Z_Y$
Case	Educational Level	Association Memberships			
1	10	5	–0.16	0.51	–0.08
2	11	3	0.09	–0.56	–0.05
3	16	6	1.32	1.04	1.37
4	7	4	–0.90	–0.03	0.03
5	6	2	–1.15	–1.09	1.25
6	7	0	–0.90	–2.15	1.94
7	11	3	0.09	–0.56	–0.05
8	20	4	2.31	–0.03	–0.07
9	14	5	0.83	0.51	0.42
10	11	5	0.09	0.51	0.05
11	6	4	–1.15	–0.03	0.03
12	7	4	–0.90	–0.03	0.03
13	9	4	–0.41	–0.03	0.01
14	7	2	–0.90	–1.09	0.98
15	10	6	–0.16	1.04	–0.17
16	16	6	1.32	1.04	1.37
17	17	7	1.57	1.57	2.46
18	11	7	0.09	1.57	0.14
19	10	2	–0.16	–1.09	0.17
20	7	2	–0.90	–1.09	0.98
			Sum of Cross Products =		10.81

Mean of X = 10.65
Standard Deviation of X = 4.04

Mean of Y = 4.05
Standard Deviation of Y = 1.88

in the appropriate columns. The sum of the cross products (10.81) is shown at the bottom of the last column.

Recall that all we have to do now to obtain the value of r is divide the sum of the cross products by $n - 1$. Thus, we can calculate the value as follows:

$$r = \frac{\sum(Z_X \bullet Z_Y)}{n - 1}$$

$$r = \frac{10.81}{19}$$

$$r = +0.57$$

Interpretation

We now have a calculated *r* value of +0.57, but there's still the question of how we should interpret it. Statisticians actually use the information provided by the *r* value in two ways.

The value of *r* is referred to as the **correlation coefficient**. The sign (+ or –) in front of the *r* value indicates whether the association is positive (direct) or negative (inverse). The absolute value of *r* (the magnitude, without respect to the sign) is a measure of the strength of the relationship. The closer the value gets to 1.0 (either +1.0 or –1.0), the stronger the association.

☑ LEARNING CHECK

Question: What does the sign of the correlation coefficient tell us?
Answer: The sign of the correlation coefficient tells us the direction of the association (either positive or negative).

Question: What does the magnitude of the correlation coefficient tell us?
Answer: The magnitude of the correlation coefficient tells us the strength of the association.

As a rule, correlation coefficients, whether positive or negative, are interpreted as follows (Salkind, 2000):

.0 to .2	No relationship to very weak association
.2 to .4	Weak association
.4 to .6	Moderate association
.6 to .8	Strong association
.8 to 1.0	Very strong to perfect association

Although the value of *r* is important in its own right, the real utility of the measure is found in its squared value. When *r* is squared (to become r^2), it's referred to as the **coefficient of determination**. It's the coefficient of determination that's so meaningful in statistical analysis. Let me explain.

A quick look at the data tells us that each variable reflects some variation. People's level of education varies, and the number of associations to which they belong also varies. The question is, how much of the variation in one variable can be attributed to variation in the other variable? As it turns out, that's what the coefficient of determination is all about.

In other words, the coefficient of determination (r^2) is a measure of the explained variance—the amount of variation in one variable that is attributable to variation in the other variable. Having obtained a positive *r* value, we know that the association between the variables is in a positive direction. The coefficient of determination, though, allows us to go well beyond that in our statement about the relationship.

☑ LEARNING CHECK

Question: What is the coefficient of determination, and how is it symbolized?

Answer: The coefficient of determination is the amount of variation in one variable that is attributable to variation in another variable. It is symbolized as r^2.

Remember: The coefficient of determination is simply the value of r squared (r^2). The r^2 value is transformed into a percentage value as follows: $r^2 \times 100$. For example, starting with our r value of +0.57, we can derive the coefficient of determination as follows:

$$\begin{aligned} r &= +0.57 \\ r^2 &= 0.325 \\ r^2 \times 100 &= 32.5\% \end{aligned}$$

Interpretation: 32.5% of the variation in number of memberships is attributable to variation in level of education.

The interpretation of r^2 is really quite telling. In this example, it tells us that variation in level of education explains 32.5% of the variation in the number of memberships in voluntary organizations. In everyday terms, it allows us to make more quantitatively based statements about why some people have more memberships and some people have fewer memberships. It's certainly true that some portion of the variation remains unexplained or is not explained by the independent variable (in this case, 67.5%), but the information provided by the coefficient of determination tells us quite a bit about the relationship at hand.

That said, let me caution you about something. When working with Pearson's r, always bear in mind that the real interpretive power is found in the coefficient of determination. Accordingly, you should remind yourself that what might appear at first glance to be a strong association (an r value approaching ±1.0) has a way of decreasing in magnitude when the value is squared. For example, an r value of –0.70 might seem to be quite strong. When the value is squared, however, we find ourselves looking at an r^2 value of 0.49.

Finally, it's important to remember that r^2 is considered to be a symmetrical measure. That means that the interpretation of r^2 works both ways. We can think of r^2 as indicating the amount of variation in Y that is attributable to variation in X, or we can think of it as the amount of variation in X that is attributable to variation in Y.

Assuming you've digested the matter of r and r^2, there's one more element to consider—namely, the matter of a hypothesis test. So far in this chapter, our attention has been directed toward the question of the strength and direction of the association between two variables. We've yet to deal with the matter of a hypothesis test for the significance of r. That's where we'll turn now.

An Additional Step: Testing the Null

It's one thing to measure the apparent strength of association between two variables based on a pattern that's reflected in sample data. It's quite another matter, though, to deal with the question of whether or not the sample reflects what's occurring in the larger population. To deal with this second question—whether or not the pattern reflects the larger population—we turn to a hypothesis-testing procedure. In short, we test the significance of r.

As always, we'll start by stating an appropriate null hypothesis, and we'll select a level of significance in advance. Because our interest is in the question of whether or not the observed association departs from chance (the assumption of no association), we can express the null hypothesis as follows:

$$H_0: r = 0$$

Following the normal procedure of selecting a level of significance in advance, we'll set the level of significance at .05. It's possible to test the null hypothesis using t, but we can also simply compare the calculated value of r (the calculated test statistic) to a table of critical values such as the one in Appendix I. In short, the table shown in Appendix I takes all the work out of the process.

☑ LEARNING CHECK

Question: When testing the significance of r, what is the null hypothesis?

Answer: The null hypothesis is a statement that $r = 0$.

As with other tables of critical values you've used, Appendix I presents critical values on the basis of the appropriate number of degrees of freedom and level of significance. When looking at Appendix I, note that the critical values of r are presented without regard to the sign that may be associated with the r value (+ or –). For example, a critical value of 0.43 actually represents a critical value of +0.43 *or* –0.43.

In testing the significance of r, the number of degrees of freedom is defined as $n - 2$, where n equals the number of cases or paired observations under consideration. For example, the r value we just calculated was based on observations for 20 people or cases. Thus, the appropriate number of degrees of freedom is equal to 20 – 2, or 18. The degrees of freedom are listed in the first column of Appendix I.

LEARNING CHECK

Question: How do you determine the number of degrees of freedom when testing the significance of *r*?

Answer: The number of degrees of freedom is equal to the number of cases or paired observations minus 2 (*n* – 2).

Levels of significance are listed across the top of the table; we're working at the .05 level. Locating the column for the .05 level of significance and the row for 18 degrees of freedom, we note that the point of intersection reveals a critical value of 0.444. Comparing our calculated test statistic of 0.57 to this critical value, we determine that the calculated test statistic exceeds the critical value. Therefore, we reject the null hypothesis (with a .05 probability of making a Type I error).

Let me suggest you take a couple of moments to carefully examine the critical values in Appendix I, particularly the way they vary according to the number of degrees of freedom. For example, the critical value (at the .05 level of significance) for 28 degrees of freedom is shown as 0.361. In other words, when working with a sample of 30 cases (degrees of freedom = 28), it would take a calculated value of *r* (either + or –) equal to or greater than 0.361 to reject the null. When working with a sample of 10 cases, however (degrees of freedom = 8), it would take a calculated *r* value (+ or –) equal to or greater than 0.632 to reject the null hypothesis. This point should make intuitive sense to you by now. In terms of what it might take to convince you that there's an association between two variables, you'd probably demand more extreme evidence (a larger *r* value), so to speak, if you were working with a very small sample, as opposed to a larger sample.

Conclusion and Interpretation

A test of the null hypothesis gives us an important foundation for our results from a correlation analysis. It's one thing to determine that there appears to be a strong (positive or negative) association between two variables based on sample data, but there's still the issue of whether or not the pattern in the sample data reflects a similar pattern in the population. And that, of course, is much the same question that we've dealt with in other hypothesis-testing procedures.

The issue always comes back to the notion that the sample data, in one way or another, could be extreme—sample information that doesn't really mirror the population in question. Since the critical values of *r* are so dependent on sample size, a test for the significance of *r* (the test of the null hypothesis that *r* = 0) is really a second but very important step in correlation analysis.

Assuming a noteworthy value of r is obtained, it should always be viewed in the context of whether or not it's significant.

In retrospect, it's easy to see why Pearson's r is such a popular statistical measure. The simple multiplication (cross products) of the Z scores, divided by $n - 1$, produces an r value, and that r value, in turn, tells us something about the strength and direction of the association between two variables. We also know that squaring the value of r will produce the coefficient of determination (r^2)—a measure of the extent to which the variation in one variable is attributable to variation in the other variable. We also know that there is a simple procedure available to test the significance of r, assuming that the presumed association between two variables is strong enough to capture our attention.

By most standards, all of that would be quite enough benefit from one simple measure. As it turns out, though, the fun has just begun, so to speak. Armed with the value of r, along with the means and standard deviations of the raw scores in two distributions, we're actually in a position to make certain predictions. More specifically, we can make predictions about a Y value on the basis of a known or assumed X value. How we go about that falls under the topic of **regression analysis**, and that's what we take up next.

Regression Analysis

A central element in the calculation of r was the conversion of distributions of raw scores into distributions of Z scores. Indeed, it was the conversion of raw scores into Z scores that allowed us to look at the association between two variables originally measured on very different scales (for example, height measured in inches and weight measured in pounds).

In the case of regression analysis, we find ourselves working with results of the correlation analysis, but we also return to our distribution of raw scores. We go back to our raw scores because the aim of regression analysis is the prediction of one value from another. In a sense, you can think of regression analysis as a technique that allows you to use existing data to predict future values. To better understand all of that, let's turn to an example.

An Application

Let's say a university administrator is concerned about the number of graduate students who enter the university but fail to complete their degrees. Let's assume the administrator's goal is to get a better handle on the association between a student's performance on the GRE (the Graduate Record Examination, a standardized graduate school admission test) and graduate school GPA (grade point average in graduate school). Knowing something about the association between these two variables might put the administrator in a better position to predict the future performance of an applicant.

Let's assume the administrator has selected a random sample of 10 student files for analysis (including students who completed a graduate degree and those who dropped out or were dismissed). The stage is now set for a detailed analysis of the data presented in Table 12-5.

Following the convention outlined earlier, we designate the student GRE score as the independent or *X* variable and the GPA score as the dependent or *Y* variable. The *X* variable (the GRE score) is measured as the combined score on the quantitative and verbal portions of the test. Because each portion of the test has a score range from 200 to 800, a combined score could range from 400 to 1600. The *Y* variable (GPA) is measured on a scale that ranges from 0.00 to 4.00. Note that the mean and standard deviation for each distribution are given at the bottom of the table. Also shown are the calculated values of r (+0.92) and r^2 (0.85, or 85%).

A scatter plot of the data from Table 12-5 is shown in Figure 12-6. As we might have expected, the general pattern suggets a positive association between the two variables. As GRE scores increase, so do GPAs. The pattern, however, is far from perfect. By no means are the points in the scatter plot aligned in a straight line.

With a little imagination, we might conceive of a line that could pass through the distribution represented by the points—a line that reflects the general trend in the pattern of cases. But an imaginary line that was eyeballed, so to speak, might not be too useful. After all, different people are apt to come up with different imaginary lines, and some lines would more accurately represent the data than others. There's one line—a very precise line—however, that best fits

Table 12-5 Data for 10 Cases (GRE scores and GPAs)

	X	Y	Z_X	Z_Y	$Z_X \bullet Z_Y$
Case	GRE Score	GPA			
1	1378	3.55	1.01	0.72	0.73
2	956	2.65	–0.86	–0.86	0.74
3	1222	3.54	0.32	0.70	0.22
4	830	2.24	–1.42	–1.58	2.24
5	991	3.00	–0.71	–0.25	0.18
6	1300	3.77	0.67	1.11	0.74
7	1521	4.00	1.65	1.51	2.49
8	899	2.62	–1.11	–0.91	1.01
9	1254	3.07	0.46	–0.12	–0.06
10	1149	2.94	0.00	–0.35	0.00
				Sum of Cross Products =	8.29

Mean of X = 1150
Standard Deviation of X = 225.20

Mean of Y = 3.14
Standard Deviation of Y = 0.57

r = +0.92
r^2 = 0.85 (85%)

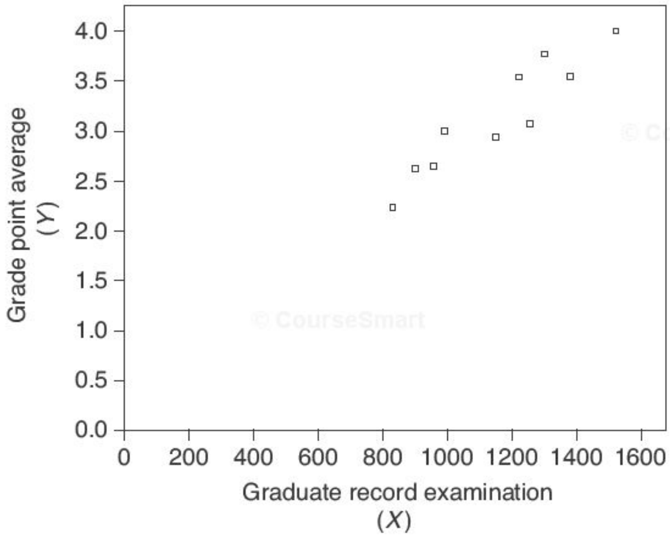

Figure 12-6 GRE and GPA Values

the data. It is known, appropriately enough, as the **line of best fit**. This line, also called the **regression line**, allows us to predict the value of *Y* on the basis of the value of *X*. To understand how all of that is done, let's take a look first at the logic behind the line and then at the equation for the line.

The Logic of Prediction and the Line of Best Fit

Returning to our example involving GRE scores and GPAs, recall that we already have the value of Pearson's *r* (the correlation coefficient) as +0.92 (see Table 12-5). Let's assume we have tested the null hypothesis at the .05 level of significance and found that we could reject the idea of no association between the variables. Armed with this information, we are ready to take the next step—attempting to predict an applicant's future success (measured in terms of GPA) on the basis of the applicant's GRE score.

Had we discovered that the association between the two variables was perfect, we would have obtained an *r* value of +1.0. Had we discovered a perfect association, we could have produced a scatter plot and easily drawn a straight line through the points. It would have been a rather simple task because in a scatter plot based on a perfect association, all the points would be aligned in a straight line. In a case like that, it would be easy to make a prediction about future success. All we'd have to do is locate a person's GRE score along the *X*-axis, draw a line up to the line passing through the cases, and then draw a line over to the axis representing future GPA values. The predicted GPA value would simply be the GPA associated with a given GRE score.

Unfortunately, however, we didn't find a perfect association between the GRE scores and future GPAs. Our r and r^2 values were high, but they certainly fell short of indicating a perfect association between the two variables. Thus, the prediction of students' future performance based on their GRE scores becomes a bit more problematic. As you're about to discover, however, the prediction may be a bit more problematic, but it's possible nonetheless. It begins with the same notion we've dealt with before—namely, the idea of a straight line passing through the distribution of points in a scatter plot.

As it turns out, we can mathematically determine the path of a straight line that best fits a scatter plot—one that passes through the various points in such a way that the line best represents the overall pattern of association between the values. It is the line that passes through the points in such a way that the squared distances of the points (cases) from the line (taken collectively) are at a minimum. In a sense, that's what the term *regression analysis* is all about. The term refers to the various elements involved in producing the line of best fit and making predictions based on that line.

Because the regression line (or line of best fit) is the line that passes through the distribution in such a manner that the squared distances of the points to the line is at a minimum, it's also often referred to as the **least squares line**. The regression line (or line of best fit or least squares line) for the data we're considering is shown in Figure 12-7. This is the same scatter plot as the one shown in Figure 12-6, but the regression line has been added. Let me suggest that you take a few moments to study the scatter plot, along with its associated line of best fit.

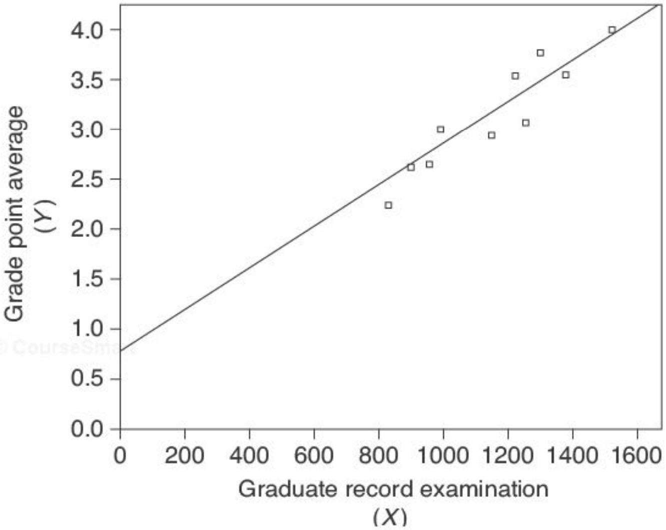

Figure 12-7 GRE and GPA Values (With Regression Line)

And that brings us back to the central purpose of regression analysis. Assuming we knew something about the regression line—assuming we knew the path of the line—it would be a simple matter to predict the value of one variable on the basis of the value of the other variable. In other words, given a value of X, we should be able to predict the value of Y. That, in short, is what regression analysis allows us to do. Given this ultimate goal, we now turn to the equation and related formulas.

☑ LEARNING CHECK

Question: What is the line of best fit? What are some other terms for the same line?

Answer: The line of best fit is the line that passes through the points in a scatter plot in such a way that it provides the best representation of the overall association between the variables. This is also known as the regression line or the least squares line.

The Regression Equation

Remember what our task is: We want to predict students' future performance (GPA) on the basis of their GRE scores. To do this, we'll rely on the **regression equation**—an equation that describes the least squares line (or line of best fit). The regression equation is defined as follows:

$$Y' = a + bX$$

The elements of the formula are as follows:

- $\boldsymbol{Y'}$ is a value that we're attempting to predict (in this case, a particular GPA); the symbol is read as **Y-prime**. The term Y' stands for a predicted value, as opposed to an actual value.
- X is a value that we are given (in this case, an applicant's GRE score).
- a is the point where the regression line (the line of best fit) crosses the Y-axis of a scatter plot. This is known as the Y-intercept (the point at which the line *intercepts* the Y-axis).
- b represents the slope of the regression line (the amount of change in Y that is associated with a unit change in X).

Assuming we can come up with the values of a and b (referred to as *constants*), we can predict a future GPA on the basis of a GRE score. Since the relationship between the two variables is less than perfect, we approach the analysis with the knowledge that our prediction will likely be less than perfect as well. On the other hand, it's safe to say that the use of the regression equation puts us in a position to make a very educated guess, even though it is apt to be less than perfect.

Returning to the elements in the equation, it's clear that now we need to know the values for a and b (the constants). Down the road, once we have a and b, we can substitute values for X to make predictions about Y values (or, more correctly, the Y' values). But first we have to have the values of a and b.

Calculation of the *b* Term (the Slope of the Line). As noted previously, the ***b* term in the regression equation ($Y' = a + bX$)**, is the slope of the line—that is, the amount of change in Y that accompanies a unit change in X. In our present example, the slope tells us the increase in a student's GPA that is expected to occur with a one-point increase in a student's GRE score.

Since we're interested in actual scores on the GRE and actual GPA values, we return to our raw scores for this portion of our analysis. The transformed Z scores came into play when we calculated the correlation coefficient, and the correlation coefficient comes into play in the regression procedure. But we have to deal with the raw scores to get an accurate picture of the unit changes that are going on in each distribution (that is, how a change in GRE score is accompanied by a change in GPA).

The situation is rendered far more comprehensible if we think in terms of standard deviation units—a change of a certain number of standard deviation units in one variable accompanied by a change of a certain number of standard deviation units in the other variable. As it turns out, this idea lies at the heart of the computation of the b term in the regression equation. In fact, the b term is simply an expression of the ratio of the standard deviation of the Y variable (s_Y) to the standard deviation of the X variable (s_X), taking into account the strength of association (the value of r) between the two variables.

We already have the standard deviation for each distribution of raw scores. Those, along with the means of each distribution, were shown in Table 12-5 as follows:

$$\begin{aligned} \overline{X} &= 1150 & s_X &= 225.20 \\ \overline{Y} &= 3.14 & s_Y &= 0.57 \end{aligned}$$

The next step is to express the relationship between the two standard deviations as follows: s_Y/s_X. The final step in the calculation of the b term is to multiply the ratio by the correlation coefficient (r). Note that we'll multiply the ratio by the value of r, not the value of r^2. The various steps for the calculation of the b term can be summarized as follows:

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

$$b = 0.92\left(\frac{0.57}{225.20}\right)$$

$$b = 0.002$$

The real meaning of the *b* term, or slope of the line, doesn't come to the forefront until we have calculated the *Y*-intercept (the *a* term). Therefore, we'll turn to that calculation next.

Calculation of the *a* Term (the *Y*-Intercept). The ***a* term in the regression equation ($Y' = a + bX$)** was previously defined as the point where the regression line (the line of best fit in a scatter plot) crosses the *Y*-axis. It is mathematically defined as follows:

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

or the mean of *Y*, minus the slope (*b*) times the mean of *X*

It should make intuitive sense to you that the regression line will, in some way, reflect the means of both distributions. It's true that there is a certain amount of variation in the GRE scores, but there is also an average GRE score (a mean for the distribution of GRE scores). By the same token, there's a certain amount of variation in the GPAs, but there's also an average GPA (a mean for the distribution of GPAs). In essence, the formula for the calculation of the *Y*-intercept (the *a* term) takes into account both means—the mean of *X* (the GRE scores) and the mean of *Y* (the GPAs). By the same token, it should make intuitive sense that the formula would also take into account the slope of the line (the *b* term), inasmuch as the slope of the line will, in part, influence where the line crosses the *Y*-axis.

The means for our raw score distributions (GRE and GPA raw scores) were given in Table 12-5 as 1150 and 3.14, respectively. Using that information, along with our calculated *b* term (for the slope of the line, equal to .002), we can easily determine the *a* term (the *Y*-intercept) as follows:

$$\begin{aligned}
a &= \overline{Y} - b\overline{X} \\
a &= 3.14 - b(1150) \\
a &= 3.14 - 0.002(1150) \\
a &= 3.14 - 2.30 \\
a &= 0.84
\end{aligned}$$

Making a Prediction. To make a prediction (to calculate a Y'), all that's necessary is to return to the formula for the regression equation: $Y' = a + bX$. For example, let's say we are reviewing an application for admission to graduate school, and the student's GRE score equals 1000. Since we now know the values of *a* and *b*, it's a simple matter to make the prediction. Using the regression equation, the prediction would move forward as follows:

$$\begin{aligned}
Y' &= a + bX \\
Y' &= 0.84 + b(1000) \\
Y' &= 0.84 + 0.002(1000) \\
Y' &= 0.84 + 2.00 \\
Y' &= 2.84
\end{aligned}$$

Given a GRE score of 1000, we predict the student will achieve a GPA of 2.84. Does this mean that we know, without question, that the student will ultimately achieve a GPA of 2.84? No, we can't make a prediction like that with total certainty. The regression line would only yield a perfect prediction if we were dealing with an underlying perfect association. On the other hand, the regression procedure (with its line of best fit and associated equation) does give us a decided advantage over a pure guess. Yes, it may amount to a guess, but it's an educated guess.

Additionally, you should note that the prediction, in this case, would be a prediction of Y' on the basis of a value of X. In statistical jargon, we say that we have regressed Y on X. Unlike the correlation procedure that produces a symmetrical measure (one that will produce the same result regardless of which variable is designated as X and which is designated as Y), the results of a regression analysis are very much a function of how you designate the variables. Some thoughtful consideration of how the constants are determined should convince you of why that is the case.

Now, what about our prediction that a GRE score of 1000 will result in a GPA of 2.84? You may have a few doubts about all this. True, the regression equation allows us to make an educated guess. But, you may well ask, just how "educated" is that guess likely to be.

The Standard Error of the Estimate

If you think back to what you learned earlier when you learned how to estimate the mean of a population or a proportion—when you learned how to construct confidence intervals—you probably sense where this is going. You're probably already thinking about the fact that your estimate (your prediction, as it were) is subject to a certain amount of error. If that's where your thinking has taken you, let me congratulate you—you're definitely on the right track.

Remember: Any prediction you make (short of one based on a perfect association or an r value of ± 1.0) will be subject to some amount of error. In regression terms, this overall expression of potential error in an estimate of Y' is referred to as the **standard error of the estimate** (s_e). Conceptually, it's an overall measure of the extent to which the predicted Y' values deviate from the actual Y values. Since the standard error of the estimate is an overall measure of deviation (deviation between the predicted and actual values of Y), you can think of it as a type of standard deviation.

The formula for the calculation of the standard error of the estimate is as follows:

$$s_e = \sqrt{\frac{\sum\left(Y - Y'\right)^2}{n - 2}}$$

Note that this formula is remarkably similar to the formula for the standard deviation that you first encountered in Chapter 2. First, it involves the summation of deviations—in this case, the deviations between the predicted values (the Y' values) and the actual values (the Y values). Next, the sum of the deviations is divided by $n - 2$, to yield the error variance. The final step merely involves taking the square root of the error variance.

If we return to our example involving the GRE scores and the GPAs, we could calculate the standard error of the estimate (s_e) according to the steps outlined in Table 12-6. Assuming we carried out these calculations, we'd eventually discover that the standard error of the estimate is equal to 0.22. We'd then be in a position to make a more grounded statement about a predicted value.

For example, we could think back to the 1-2-3 Rule that we encountered earlier and easily make use of it in connection with our prediction. As you'll recall, we learned from the 1-2-3 Rule that approximately 95% of the cases under a normal distribution will be found ±2 standard deviations from the mean.

Table 12-6 Calculation of the Standard Error of the Estimate

Case	X	Y	Y'	$(Y - Y')$	$(Y - Y')^2$
1	1378	3.55	3.60	–0.05	0.00
2	956	2.65	2.75	–0.10	0.01
3	1222	3.54	3.28	0.26	0.07
4	830	2.24	2.50	–0.26	0.07
5	991	3.00	2.82	0.18	0.03
6	1300	3.77	3.44	0.33	0.11
7	1521	4.00	3.88	0.12	0.01
8	899	2.62	2.64	–0.02	0.00
9	1254	3.07	3.35	–0.28	0.08
10	1149	2.94	3.14	–0.20	0.04

$$\sum(Y - Y')^2 = 0.42$$

Calculation of Standard Error of the Estimate

$$s_e = \sqrt{\frac{\sum\left(Y - Y'\right)^2}{n - 2}}$$

$$s_e = \sqrt{\frac{0.42}{8}}$$

$$s_e = \sqrt{0.05}$$

$$s_e = 0.22$$

Our predicted GPA (based on a GRE score of 1000) is 2.84, and we know that the standard error of the estimate (which is really a standard deviation) is equal to 0.22. Suppose we subtract two standard error units from our predicted value and add two standard error units to our predicted value:

$$2.84 - 2(0.22) = 2.40$$
$$2.84 + 2(0.22) = 3.28$$

We're now back in somewhat familiar territory. Indeed, we're in a position to estimate that a GRE score of 1000 will be associated with a GPA that ranges between 2.40 and 3.28 and that we have used a method that will generate a correct estimate approximately 95 times out of 100.

Can we be 100% certain about our estimate? No, we can't be 100% certain that our estimate is correct. On the other hand, our estimate amounts to a very educated guess. And, as a friend of mine is fond of saying, an educated guess always beats a shot in the dark.

At the conclusion of this chapter, you'll find a variety of questions and problems, all designed to enhance your understanding of both correlation and regression. As in previous chapters, you'll find an emphasis on conceptual as opposed to computational elements, but you'll have a chance to sharpen your skills on both fronts. Let me suggest that you give the required time to the question/problem section. Correlation and regression are part of the statistical shelf of staples, so to speak. A solid understanding of the concepts will serve you well.

Chapter Summary

In this chapter, you were introduced to the topics of correlation and regression analysis, two of the more popular statistical procedures. Along the way, you extended your understanding of the notion of a relationship or association between variables as you added the concepts of strength and direction to your storehouse of knowledge. You were introduced to the notions of the correlation coefficient and the coefficient of determination—two concepts that allow us to say quite a lot about the association between two variables. What's more, you learned that it's an easy matter to test a null hypothesis involving a correlation coefficient.

In your exploration of regression analysis, you learned about the regression line and the regression equation. You also learned that it's possible to predict the value of one variable, given the value of another variable (provided the variables are associated in a linear pattern). You also learned, however, that such a prediction will not be perfect (unless, of course, the underlying association between the variables is perfect).

In short, you learned quite a lot. However, a full exploration of correlation and regression analysis is impossible in just a few pages. For this reason, I urge you to pay close attention to the next section ("Some Other Things You Should Know"). Think of it as an invitation to further exploration into the world of statistical analysis.

Some Other Things You Should Know

In many ways, a complete chapter on the topics of correlation and regression could approach a near limitless length. We've only scratched the surface in this presentation—touching on the very basic principles involved in the most fundamental applications. The topics of correlation and regression are so substantial that entire texts (often of considerable length) have been devoted to them. It's always difficult to draw the line when it comes to the matter of an introductory voyage into the world of statistics, and just how much should be given over to the topics of correlation and regression is a case in point.

More advanced texts, for example, often deal with the techniques of multiple and partial correlation. Similarly, other texts may present material on more advanced regression techniques. For example, see Ramsey and Shafer (2002) for an excellent treatment, should you want to explore the more advanced procedures.

As to the material presented here, there are still a few things you should take into consideration before you launch into a simple correlation or regression analysis. With the widespread availability of computer-based statistical software, correlation or regression analysis can be tempting, particularly if you're faced with a mountain of data that's crying out for analysis. On the other hand, certain assumptions should be met before embarking on the procedures, and you should always approach your interpretation of results with some caution.

As to the fundamental assumptions that underlie simple correlation analysis, you're already aware that the procedure rests on the availability of interval/ratio level data—two variables, each expressed in terms of an interval/ratio scale of measurement. Moreover, there is an assumption that each variable under consideration is normally distributed and that the variances of each distribution are roughly equal.

As to the caution that should be exercised in the interpretation of results, keep in mind that a prediction made on the basis of regression analysis is always subject to error. Just as the estimate of a population mean or proportion is always accompanied by some margin of error, the same applies in the case of a prediction of Y' on the basis of the regression equation.

Finally, you should always remember that your analysis, more than likely, will involve sample data, and with that go certain limitations and assumptions. Now, however, you're armed to deal with them. *Welcome to the world of statistical analysis!*

Key Terms

- *a* term in the regression equation $(Y' = a + bX)$
- *b* term in the regression equation $(Y' = a + bX)$
- coefficient of determination
- correlation
- correlation coefficient
- curvilinear association
- dependent variable
- independent variable
- least squares line
- line of best fit
- linear association
- negative (inverse) association
- Pearson's *r*
- perfect association
- positive (direct) association
- regression analysis
- regression equation
- regression line
- scatter plot
- standard error of the estimate
- strength of association
- *Y* prime (Y')

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. An *r* value of __________ would be interpreted as a perfect negative association.
2. An *r* value of __________ would be interpreted as a perfect positive association.
3. The value of *r* has a range from _________ to ___________.
4. An *r* value of 0 would be interpreted as ___________ association.
5. A __________ is a visual representation of the values of two variables on a case-by-case basis.
6. The ___________ coefficient, or *r*, reveals the strength and direction of an association between two variables.
7. The coefficient of _____________, or r^2, tells the amount of variation in one variable that is associated with or explained by variation in the other variable.
8. The regression line is the ______ of best _____; it is also referred to as the ____ squares line.
9. The equation for the regression line is $Y' =$ __________.
10. In the regression equation, ______ is the value that we are attempting to predict.
11. In the regression equation, ______ is the *Y*-intercept or the point where the regression line crosses the *Y*-axis.
12. In the regression equation, ______ represents the slope of the regression line.

Application Questions/Problems

1. Consider the following set of data:

Case	*X*	*Y*
1	3	4
2	5	7
3	8	7
4	2	2
5	6	5
6	5	4
7	5	5
8	7	8
9	9	8
10	10	9
11	12	10
12	8	7
13	3	2

a. What is the value of the mean of *X*?
b. What is the value of the standard deviation of *X*?
c. What is the value of the mean of *Y*?
d. What is the value of the standard deviation of *Y*?
e. What is the value of the sum of the cross-products?
f. What is the value of *r*?

2. Consider the following set of data:

Case	*X*	*Y*
1	8	39
2	10	42
3	9	51
4	10	59
5	12	84
6	12	38
7	8	48
8	9	59
9	12	63
10	10	77

a. What is the value of the mean of *X*?
b. What is the value of the standard deviation of *X*?
c. What is the value of the mean of *Y*?
d. What is the value of the standard deviation of *Y*?
e. What is the value of the sum of the cross-products?
f. What is the value of *r*?

3. Two variables, Variable *X* and Variable *Y*, are the focus of a study. The study involves 14 research participants. The sum of the cross products ($Z_X \bullet Z_Y$) for the 14 cases is –11.62.
a. Calculate and interpret *r*.

b. Calculate and interpret r^2.

c. Assuming you were to test the significance of r at the .05 level of significance, state an appropriate null hypothesis. What would you conclude?

4. Two variables, Variable X and Variable Y, are the focus of a study. The study involves 50 research participants. The sum of the cross products $(Z_X \bullet Z_Y)$ for the 50 cases is 15.66.

a. Calculate and interpret r.

b. Calculate and interpret r^2.

c. Assuming you were to test the significance of r at the .05 level of significance, state an appropriate null hypothesis. What would you conclude?

5. Two variables, Variable X and Variable Y, are the focus of a study. The study involves 25 research participants. The sum of the cross products $(Z_X \bullet Z_Y)$ for the 25 cases is 21.58.

a. Calculate and interpret r.

b. Calculate and interpret r^2.

c. Assuming you were to test the significance of r at the .05 level of significance, state an appropriate null hypothesis. What would you conclude?

6. A researcher discovers the following information about the association between Variable X and Variable Y:

Mean of X = 50.49	Standard deviation of X = 12.83
Mean of Y = 18.30	Standard deviation of Y = 4.11

$r = -0.71$

Calculate a and b in the regression equation $(Y' = a + bX)$.

7. A researcher discovers the following information about the association between Variable X and Variable Y:

Mean of X = 20	Standard deviation of X = 6
Mean of Y = 100	Standard deviation of Y = 30

$r = +0.83$

Calculate a and b in the regression equation $(Y' = a + bX)$.

8. Assume you've collected information from 6 students as to how many hours they work each week and their grade point averages (GPAs). The information is shown below.

Student	X Hrs. Worked	Y GPA	Z_X	Z_Y	$Z_X \bullet Z_Y$
1	10	3.80	–0.66	0.76	–0.50
2	20	3.44	0.00	0.14	0.00
3	40	2.50	1.32	–1.48	–1.95
4	35	2.81	0.99	–0.95	–0.94
5	0	4.00	–1.32	1.10	–1.45
6	15	3.62	–0.33	0.45	–0.15

Mean of $X = 20$	Standard deviation of $X = 15.17$
Mean of $Y = 3.36$	Standard deviation of $Y = 0.58$

Note that the two variables, or number of hours worked each week (X) and GPA (Y), have already been transformed into Z scores.

a. Calculate and interpret r.

b. Calculate and interpret r^2.

c. Calculate a and b in the regression equation ($Y' = a + bX$).

9. Assume you've collected information from customers at a local bookstore. More specifically, for 10 customers, you have the following information on their levels of education and expenditures on book purchases.

Mean of $X = 13$	Standard deviation of $X = 3.46$
Mean of $Y = 15.70$	Standard deviation of $Y = 15.28$

Note that the two variables, years of education (X) and dollar amount of expenditure (Y), have already been transformed into Z scores

	X Yrs. Education	Y $ Expenditure	Z_X	Z_Y	$Z_X \bullet Z_Y$
1	13	5	0.00	−0.70	0.00
2	17	45	1.16	1.92	2.23
3	9	0	−1.16	−1.03	1.19
4	11	18	−0.58	0.15	−0.09
5	15	25	0.58	0.61	0.35
6	8	4	−1.45	−0.77	1.12
7	13	10	0.00	−0.37	0.00
8	18	35	1.45	1.26	1.83
9	16	15	0.87	−0.05	−0.04
10	10	0	−0.87	−1.03	0.90

a. Calculate and interpret r.

b. Calculate and interpret r^2.

c. Calculate a and b in the regression equation ($Y' = a + bX$).

d. Calculate the standard error of the estimate.

10. Using the information that you developed in your responses to Question 9 (related to level of education and expenditures at a book store), predict the amount of expenditure for someone with 20 years of education.