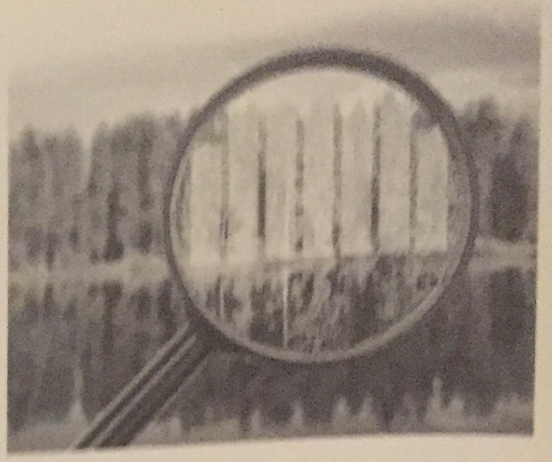
CHAPTER 10



Relationships Between Measurement Variables

Thought Questions

1. Judging from the scatterplot in Figure 9.5, there is a *positive correlation* between verbal SAT score and GPA. For used cars, there is a *negative correlation* between the age of the car and the selling price. Explain what it means for two variables to have a positive correlation or a negative correlation.
2. Suppose you were to make a scatterplot of (adult) sons' heights versus fathers' heights by collecting data on both from several of your male friends. You would now like to predict how tall your nephew will be when he grows up, based on his father's height. Could you use your scatterplot to help you make this prediction? Explain.
3. Do you think each of the following pairs of variables would have a positive correlation, a negative correlation, or no correlation?
   a. Calories eaten per day and weight.
   b. Calories eaten per day and IQ.
   c. Amount of alcohol consumed and accuracy on a manual dexterity test.
   d. Number of ministers and number of liquor stores in cities in Pennsylvania.
   e. Height of husband and height of wife.
4. An article in the *Sacramento Bee* (29 May 1998, p. A17) noted, "Americans are just too fat, researchers say, with 54 percent of all adults heavier than is healthy. If the trend continues, experts say that within a few generations virtually every U.S. adult will be overweight." This prediction is based on "extrapolating," which assumes the current rate of increase will continue indefinitely. Is that a reasonable assumption? Do you agree with the prediction? Explain.

10.1 Statistical Relationships

One of the interesting advances made possible by the use of statistical methods is the quantification and potential confirmation of relationships. In the first part of this book, we discussed relationships between aspirin and heart attacks, meditation and test scores, and smoking during pregnancy and child's IQ, to name just a few. In Chapter 9, we saw examples of relationships between two variables illustrated with pictures, such as the scatterplot of verbal SAT scores and college GPAs.

Although we have examined many relationships up to this point, we have not considered how those relationships could be expressed quantitatively. In this chapter, we discuss **correlation,** which measures the *strength* of a certain type of relationship between two measurement variables, and **regression,** which is a numerical method for trying to *predict* the value of one measurement variable from knowing the value of another one.

Statistical Relationships versus Deterministic Relationships

A **statistical relationship** differs from a **deterministic relationship** in that, in the latter case, if we know the value of one variable, we can determine the value of the other exactly. For example, the relationship between volume and weight of water is deterministic. The old saying, "A pint's a pound the world around," isn't quite true, but the deterministic relationship between volume and weight of water does hold. (A pint is actually closer to 1.04 pounds.) We can express the relationship by a formula, and if we know one value, we can solve for the other (weight in pounds $= 1.04 \times$ volume in pints).

Natural Variability in Statistical Relationships

In a statistical relationship, natural variability exists in the relationship between the two measurements. For example, we could describe the average relationship between height and weight for adult females, but very few women would fit that exact formula. If we knew a woman's height, we could predict the average weight for all women with that same height, but we could not predict her weight exactly. Similarly, we can say that, on average, taking aspirin every other day reduces one's chance of having a heart attack, but we cannot predict what will happen to one specific individual.

Statistical relationships are useful for describing what happens to a population, or aggregate. The stronger the relationship, the more useful it is for predicting what will happen for an individual. When researchers make claims about statistical relationships, they are not claiming that the relationship will hold for everyone.

10.2 Strength versus Statistical Significance

To find out if a statistical relationship exists between two variables, researchers must usually rely on measurements from only a sample of individuals from a larger population. However, for any particular sample, a relationship may exist even if there is no relationship between the two variables in the population. It may be just the "luck of the draw" that that particular sample exhibited the relationship.

For example, suppose an observational study followed for 5 years a sample of 1000 owners of satellite dishes and a sample of 1000 nonowners and found that four of the satellite dish owners developed brain cancer, whereas only two of the nonowners did. Could the researcher legitimately claim that the rate of cancer among all satellite dish owners is twice that among nonowners? You would probably not be persuaded that the observed relationship was indicative of a problem in the larger population. The numbers are simply too small to be convincing.

Defining Statistical Significance

To overcome this problem, statisticians try to determine whether an observed relationship in a sample is **statistically significant.** To determine this, we ask what the chances are that a relationship that strong or stronger would have been observed in the sample if there really were nothing going on in the population. If those chances are small, we declare that the relationship is statistically significant and was not just a fluke. To be convincing, an observed relationship must also be statistically significant.

> Most researchers are willing to declare that a relationship is statistically significant if there is only a small chance of observing the relationship in the sample when actually nothing is going on in the population. A common criterion is to define a "small chance" to be 5%, but sometimes 10%, 1%, or some other value is used. In other words, a relationship observed in sample data is typically considered to be statistically significant if that relationship is stronger than 95% of the relationships we would expect to see just by chance.

Of course, this reasoning carries with it the implication that of all the relationships that do occur by chance alone, 5% of them will erroneously earn the title of statistical significance. However, this is the price we pay for not being able to measure the entire population—while still being able to determine that statistically significant relationships do exist. We will learn how to assess statistical significance in Chapters 13, 22, and 23.

Two Warnings about Statistical Significance

Two important points, which we will study in detail in Chapter 24, often lead people to misinterpret statistical significance. First, it is easier to rule out chance if the observed relationship is based on very large numbers of observations. Even a minor

relationship will achieve "statistical significance" if the sample is very large. However, earning that title does not necessarily imply that there is a *strong* relationship or even one of practical importance.

EXAMPLE 10.1 **Small but Significant Increase in Risk of Breast Cancer**

News Story 12 in the Appendix, "Working nights may increase breast cancer risk," contains the following quote by Francine Laden, one of the co-authors of the study: "The numbers in our study are small, but they are statistically significant." As a reader, what do you think that means? Reading further in the news story reveals the answer:

> *The study was based on more than 78,000 nurses from 1988 through 1998. It found that nurses who worked rotating night shifts at least three times a month for one to 29 years were 8 percent more likely to develop breast cancer. For those who worked the shifts for more than 30 years, the relative risk went up by 36 percent.*

The "small numbers" Dr. Laden referenced were the small increases in the risk of breast cancer, of 8 percent and 36 percent (especially the 8 percent). Because the study was based on over 78,000 women, even the small relationship observed in the sample probably reflects a real relationship in the population. In other words, the relationship in the sample, while not strong, is "statistically significant." ■

Second, a very strong relationship won't necessarily achieve "statistical significance" if the sample is very small. If you read about researchers who "failed to find a statistically significant relationship" between two variables, do not be confused into thinking that they have proven that there *isn't* a relationship. It may be that they simply didn't take enough measurements to rule out chance as an explanation.

EXAMPLE 10.2 **Do Younger Drivers Eat and Drink More while Driving?**

News Story 5 (summarized in the Appendix), "Driving while distracted is common, researchers say" contains the following quote:

> *Stutts' team had to reduce the sample size from 144 people to 70 when they ran into budget and time constraints while minutely cataloging hundreds of hours of video. The reduced sample size does not compromise the researchers' findings, Stutts said, although it does make analyzing population subsets difficult.*

What does this mean? Consulting Original Source 5 on the companion website, one example explicitly stated is when the researchers tried to compare behavior across age groups. For instance, in Table 7 of the report (p. 36), it is shown that 92.9 percent of 18- to 29-year-old drivers were eating or drinking while driving. Middle-aged drivers weren't as bad, with 71.4 percent of drivers in their 30s and 40s and 78.6 percent of drivers in their 50s eating or drinking. And a mere 42.9 percent of drivers 60 and over were observed eating or drinking while driving. It would seem that these reflect real differences in behavior in the population of all drivers, and not just in the drivers observed in this study. But because there were only 14 drivers observed in each age group, the observed relationship between age and eating behavior is not statistically

significant. It is impossible to know whether or not the relationship exists in the population. The authors of the report wrote:

> *Compared to older drivers, younger drivers appeared more likely to eat or drink while driving. . . . Sample sizes within age groups, however, were small, prohibiting valid statistical testing.* (pp. 61–62)

Notice that in this example, the authors of the original report and the journalist who wrote the news story interpreted the problem correctly. An incorrect, and not uncommon, interpretation would be to say that "no significant difference was found in eating and drinking behavior across age groups." While technically true, this language would lead readers to believe that there is no difference in these behaviors in the population, when in fact the sample was just too small to decide one way or the other. Occasionally a completely misleading statement will be made by saying that "no difference" was found, when the author means that no statistically significant difference was found. ■

10.3 Measuring Strength Through Correlation

A Linear Relationship

It is convenient to have a single number to measure the strength of the relationship between two measurement variables and to have that number be independent of the units used to make the measurements. For instance, if height is reported in inches instead of centimeters (and not rounded off in either case), the strength of the relationship between height and weight should not change.

Many types of relationships can occur between measurement variables, but in this chapter we consider only the most common one. The **correlation** between two measurement variables is an indicator of *how closely their values fall to a straight line* on a scatterplot. Sometimes this measure is called the *Pearson product–moment correlation* or the *correlation coefficient* or is simply represented by the letter *r*.

Notice that the statistical definition of *correlation* is more restricted than its common usage. For example, if the value of one measurement variable is always the square of the value of the other variable, they have a perfect relationship but may still have no statistical correlation. As used in statistics, correlation measures *linear relationships* only; that is, it measures how close the individual points in a scatterplot are to a straight line.

Other Features of Correlations

Here are some other features of correlations:

1. A correlation of +1 (or 100%) indicates that there is a perfect linear relationship between the two variables. As one increases, so does the other. All individuals fall on the same straight line, just as when two variables have a deterministic linear relationship.

2. A correlation of -1 also indicates that there is a perfect linear relationship between the two variables. However, as one increases, the other *decreases*. The individuals all fall on a straight line that slopes *downward*.
3. A correlation of zero could indicate that there is no linear relationship between the two variables. It could also indicate that the best straight line through the data on a scatterplot is exactly horizontal.
4. A *positive correlation* indicates that the variables increase together.
5. A *negative correlation* indicates that as one variable increases, the other decreases.
6. Correlations are unaffected if the units of measurement are changed. For example, the correlation between weight and height remains the same regardless of whether height is expressed in inches, feet, or millimeters.

Examples of Positive, Negative, and No Linear Relationship

Following are some examples of both positive and negative relationships. Notice how the closeness of the points to a straight line determines the *magnitude* of the correlation, whereas whether the line slopes up or down determines if the correlation is positive or negative.

EXAMPLE 10.3 Verbal SAT and GPA

In Chapter 9, we saw a scatterplot showing the relationship between the two variables verbal SAT and GPA for a sample of college students. The correlation for the data in the scatterplot is .485, indicating a moderate positive relationship. In other words, students with higher verbal SAT scores tend to have higher GPAs as well, but the relationship is nowhere close to being exact. ■

EXAMPLE 10.4 Husbands' and Wives' Ages and Heights

Marsh (1988, p. 315) and Hand et al. (1994, pp. 179–183) reported data on the ages and heights of a random sample of 200 married couples in Britain, collected in 1980 by the Office of Population Census and Surveys. Figures 10.1 and 10.2 (see next page) show scatterplots for the ages and the heights, respectively, of the couples. Notice that the ages fall much closer to a straight line than do the heights. In other words, husbands' and wives' ages are likely to be closely related, whereas their heights are less closely related. The correlation between husbands' and wives' ages is .94, whereas the correlation between their heights is only .36. Thus, the values for the correlations confirm what we see from looking at the scatterplots.

■

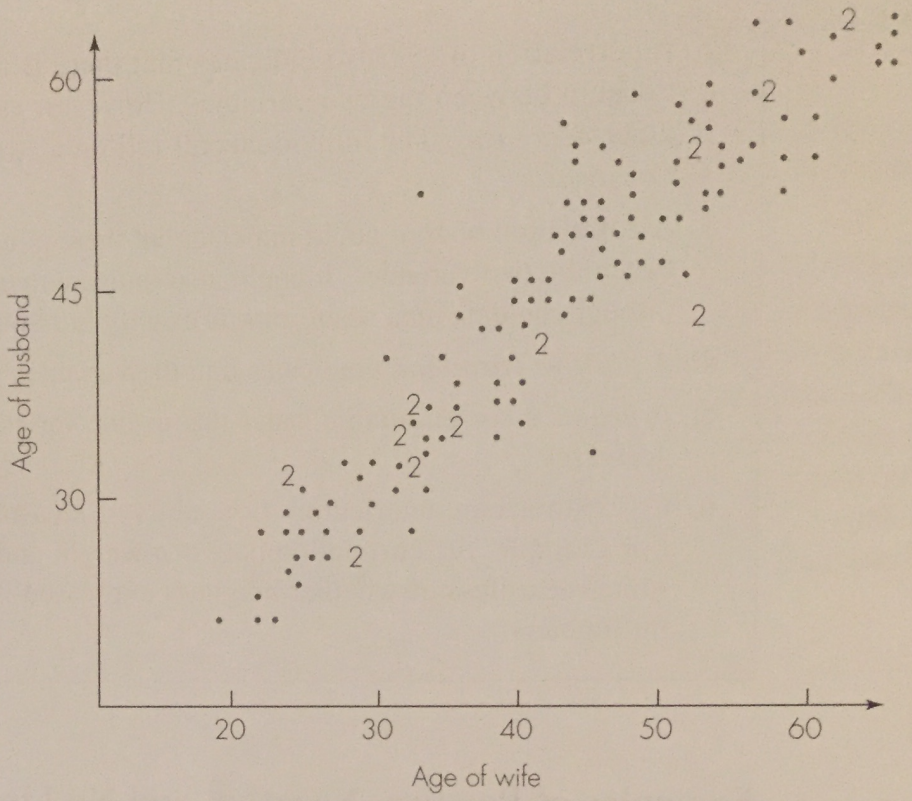


Figure 10.1
Scatterplot of British husbands' and wives' ages; correlation = .94
Source: Hand et al., 1994.

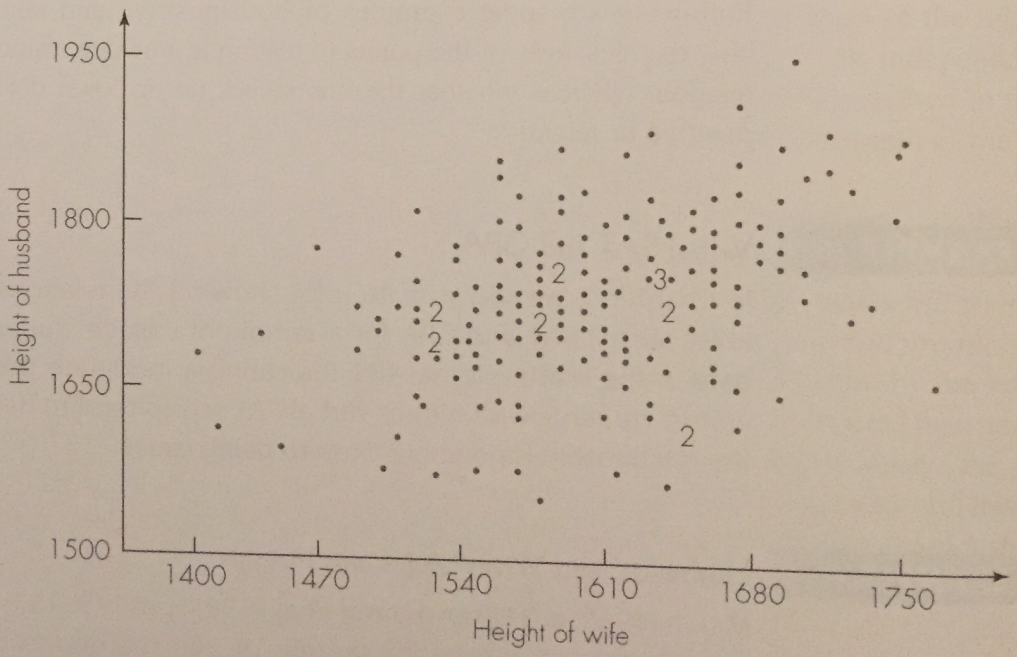


Figure 10.2
Scatterplot of British husbands' and wives' heights (in millimeters); correlation = .36
Source: Hand et al., 1994.

EXAMPLE 10.5 **Facebook Friends and Studying for Statistics Class**

In early 2011, the social media network Facebook had about 700 million users across the world and was widely used by college students. At that time, was there a relationship between the numbers of Facebook "friends" students had and how much time they spent studying? Were those with more friends likely to spend more time on Facebook and less time studying? Or perhaps those with more friends were more

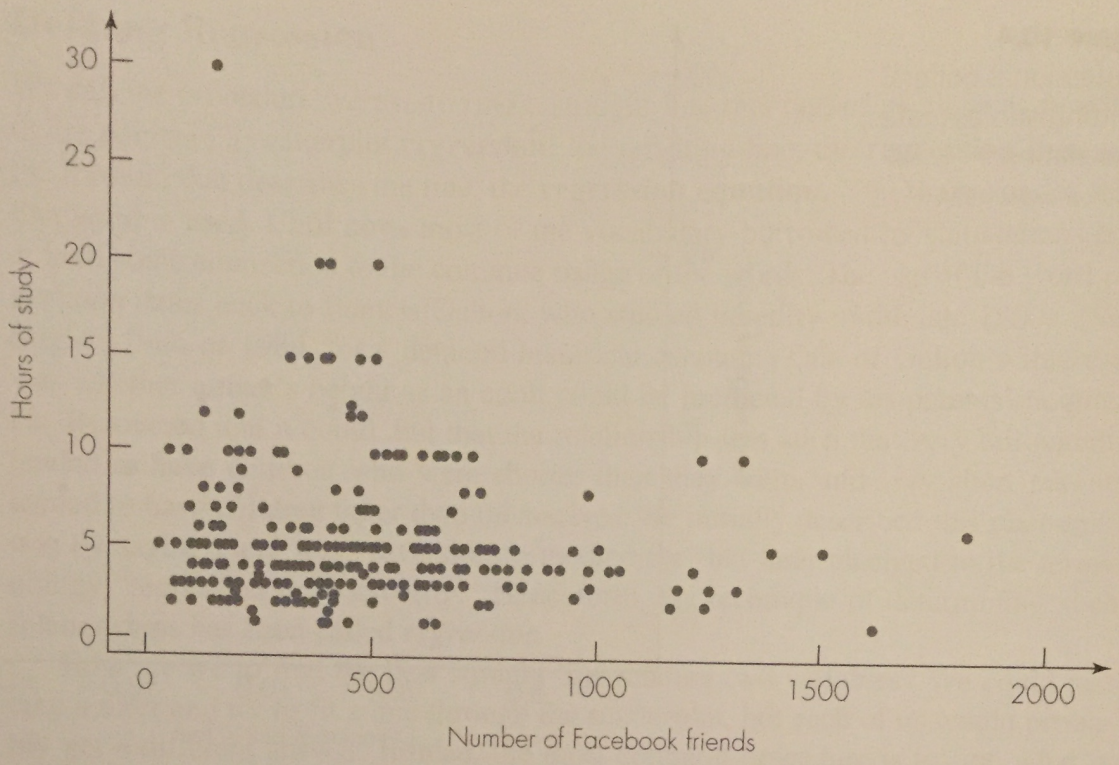


Figure 10.3
Hours spent studying for a Statistics class and number of Facebook friends; correlation = −.057
Source: The author's students.

ambitious in general and studied more? Two of the questions on a survey of students in two introductory statistics classes in the 2011 Winter quarter were:

- How many Facebook friends do you have?
- How many hours per week on average do you study for this class? Include time spent studying, doing homework, and in office hours, but not time in class or discussion.

Of the 277 students who responded, 267 were Facebook users. Figure 10.3 displays a scatterplot of their responses to these two questions. (Two outliers with unrealistic values on one or the other of the questions were removed because they were clearly facetious answers.) There does not appear to be much of a relationship between these two variables, and the near-zero correlation of –.057 confirms this fact. The outlier at the top of the plot is a person who studied 30 hours a week and had 150 Facebook friends. That point is somewhat responsible for the negative correlation, and in fact, if that individual is removed the correlation is even closer to 0, at –.032. There does not appear to be a linear relationship between number of Facebook friends and study hours. ■

EXAMPLE 10.6

Professional Golfers' Putting Success

Iman (1994, p. 507) reported on a study conducted by *Sports Illustrated* magazine in which the magazine studied success rates at putting for professional golfers. Using data from 15 tournaments, the researchers determined the percentage of successful putts at distances from 2 feet to 20 feet. We have restricted our attention to the part of the data that follows a linear relationship, which includes putting distances from 5 feet to 15 feet. Figure 10.4 (next page) illustrates this relationship. The correlation between distance and rate of success is −.94. Notice the negative sign, which indicates that as distance goes up, success rate goes down. ■

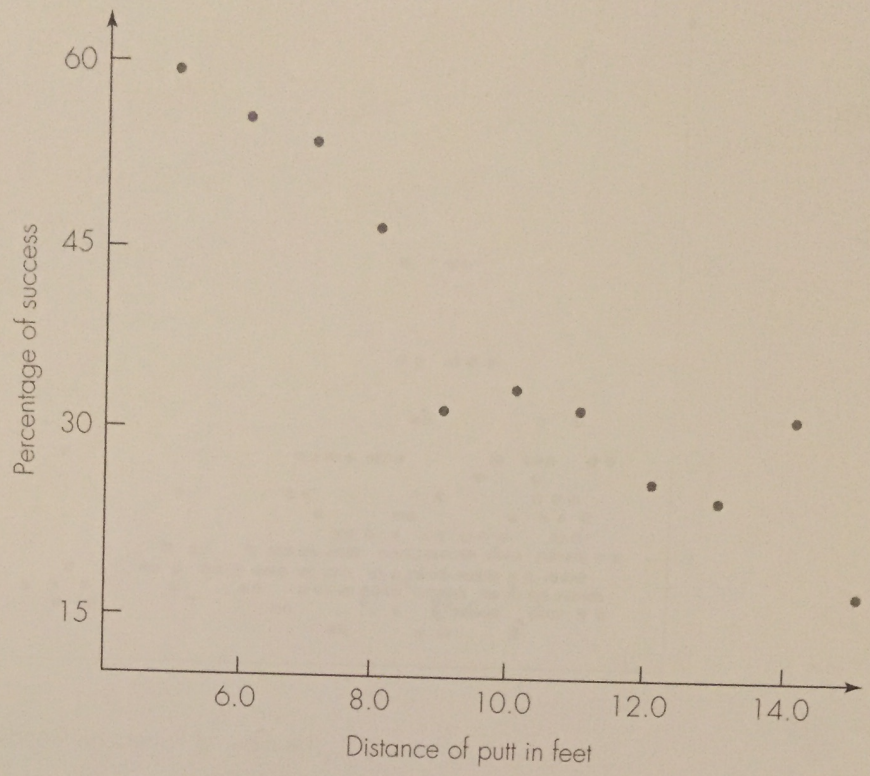


Figure 10.4
Professional golfers' putting success rates; correlation = −.94
Source: Iman, 1994.

Using Computers to Find Correlation

There are many calculators, computer programs and websites that will calculate correlation for you. Here is how to do it in Excel:

1. List the values of one of the variables in one column. Let's call the cells used "Array 1". For instance, if you use Column A, rows 1 to 10, then Array 1 is A1:A10.
2. List the values of the other variable in the same order in another column. Let's call the cells used "Array 2." Array 1 and Array 2 must be the same length.
3. The Excel function CORREL(Array 1, Array 2) gives the correlation.

10.4 Specifying Linear Relationships with Regression

Sometimes, in addition to knowing the strength of the connection between two variables, we would like a *formula* for the relationship. For example, it might be useful for colleges to have a formula for the connection between verbal SAT score and college GPA. They could use it to predict the potential GPAs of future students. Some colleges do that kind of prediction to decide who to admit, but they use a collection of variables instead of just one. The simplest kind of relationship between two variables is a straight line, and that's the only type we discuss here. Our goal is to find a straight line that comes as close as possible to the points in a scatterplot.

Defining Regression

We call the procedure we use to find a straight line that comes as close as possible to the points in a scatterplot **regression;** the resulting line, the **regression line;** and the formula that describes the line, the **regression equation.** You may wonder why that word is used. Until now, most of the vocabulary borrowed by statisticians had at least some connection to the common usage of the words. The use of the word *regression* dates back to Francis Galton, who studied heredity in the late 1800s. (See Stigler, 1986 or 1989, for a detailed historical account.) One of Galton's interests was whether a man's height as an adult could be predicted by his parents' heights. He discovered that it could, but that the relationship was such that very tall parents tended to have children who were shorter than they were, and very short parents tended to have children taller than themselves. He initially described this phenomenon by saying there was "reversion to mediocrity" but later changed to the terminology "regression to mediocrity." Henceforth, the technique of determining such relationships has been called *regression.*

How are we to find the best straight line relating two variables? We could just take a ruler and try to fit a line through the scatterplot, but each of us would probably get a different answer. Instead, the most common procedure is to find what is called the **least squares line.** In determining the least squares line, priority is given to how close the points fall to the line for the variable represented by the vertical axis. Those distances are squared and added up for all of the points in the sample. For the least squares line, that sum is smaller than it would be for any other line.

The vertical distances are chosen because the equation is often used to predict that variable when the one on the horizontal axis is known. Therefore, we want to minimize how far off the prediction would be in that direction. In other words, the horizontal axis usually represents an explanatory variable, and the vertical axis represents a response variable. We want to predict the value of the response variable from knowing the value of the explanatory variable. The line we use is the one that minimizes the sum of the squared errors resulting from this prediction for the individuals in the sample. The reasoning is that if the line is good at predicting the response for those in the sample, when the response is already known, then it will work well for predicting the response in the future when only the explanatory variable is known.

The Equation for the Line

All straight lines can be expressed by the same formula. Using standard conventions, we call the variable on the vertical axis y and the variable on the horizontal axis x. We can then write the equation for the line relating them as

$$y = a + bx$$

where for any given situation, a and b would be replaced by numbers. We call the number represented by a the **intercept** and the number represented by b the **slope.** The intercept simply tells us at what particular point the line crosses the vertical axis when the horizontal axis is at zero. The slope tells us how much of an increase there is for one variable (the one on the vertical axis) when the other (on the horizontal