relationships
differences

IN CHAPTER 7 WE DEALT WITH testing hypotheses concerning differences between sample means. In this chapter we will begin examining questions concerning relationships between variables. Although you should not make too much of the distinction between **relationships** and **differences** (if treatments have *different* means, then means are *related* to treatments), the distinction is useful in terms of the interests of the experimenter and the structure of the experiment. When we are concerned with differences between means, the experiment usually consists of a few quantitative or qualitative levels of the independent variable (e.g., Treatment A and Treatment B) and the experimenter is interested in showing that the dependent variable differs from one treatment to another. When we are concerned with relationships, however, the independent variable (X) usually has many quantitative levels and the experimenter is interested in showing that the dependent variable is some *function* of the independent variable.

correlation
regression

This chapter will deal with two interwoven topics: **correlation** and **regression.** Statisticians commonly make a distinction between these two techniques. Although the distinction is frequently not followed in practice, it is important enough to consider briefly. In problems of simple correlation and regression, the data consist of two observations from each of N subjects, one observation on each of the two variables under consideration. If we were interested in the correlation between running speed in a maze (Y) and number of trials to reach some criterion (X) (both common measures of learning), we would obtain a running-speed score and a trials-to-criterion score from each subject. Similarly, if we were interested in the regression of running speed (Y) on the number of food pellets per reinforcement (X), each subject would have scores corresponding to his speed and the number of pellets he received. The difference between these two situations illustrates the statistical distinction between correlation and regression. In both cases, Y (running speed) is a **random variable,** beyond the experimenter's control. We don't know what the rat's running speed will be until we carry out a trial and measure the speed. In the former case, X is also a random variable, since the number of trials to criterion depends on how fast the animal learns, and this is beyond the control of the experimenter. Put another way, a replication of the experiment would leave us with different values of both Y and X. In the food pellet example, however, X is a **fixed variable.** The number of pellets is determined by the experimenter (for example, 0, 1, 2, or 3 pellets) and would remain constant across replications.

random variable

fixed variable

To most statisticians, the word *regression* is reserved for those situations in which the value of X is *fixed* or specified by the experimenter before the data are collected. In these situations, no sampling error is involved in X, and repeated replications of the experiment will involve the same set of X values. The word *correlation* is used to describe the situation in which both X and Y are random variables. In this case, the Xs, as well as the Ys, vary from one replication to another and thus sampling error is involved in both variables. This distinction is basically the distinction between what are called **linear regression models** and **bivariate normal models.** We will consider the distinction between these two models in more detail in Section 9.7.

linear regression models
bivariate normal models

As mentioned earlier, the distinction between the two models, although appropriate on statistical grounds, tends to break down in practice. A more pragmatic distinction relies on the interest of the experimenter. If the purpose of the research is to

prediction

allow **prediction** of Y on the basis of knowledge about X, we will speak of regression. If, on the other hand, the purpose is merely to obtain a statistic expressing the degree of relationship between the two variables, we will speak of correlation. Although it is possible to raise legitimate objections to this distinction, it has the advantage of describing the different ways in which these two procedures are used in practice. We will see many instances of situations in which regression (rather than correlation) is the goal even when both variables are random.

Having differentiated between correlation and regression, we will now proceed to treat the two techniques together, since they are so closely related. The general problem then becomes one of developing an equation to predict one variable from knowledge of the other (regression) and of obtaining a measure of the degree of this relationship (correlation). The only restriction we will impose for the moment is that the relationship between X and Y be linear. Curvilinear relationships will not be considered, although in Chapter 15 we will see how they can be handled by closely related procedures.

9.1 Scatterplot

scatterplot
scatter diagram
scattergram

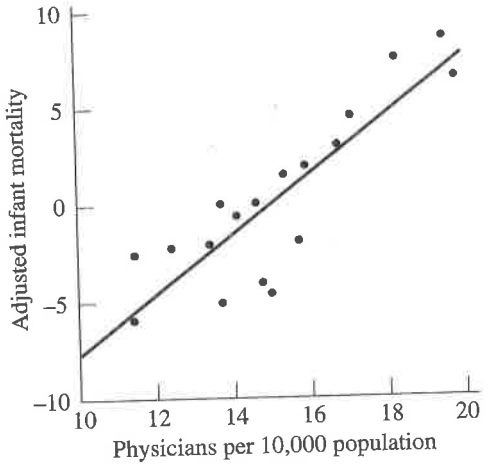
When we collect measures on two variables for the purpose of examining the relationship between these variables, one of the most useful techniques for gaining insight into this relationship is a **scatterplot** (also called a **scatter diagram** or **scattergram**). In a scatterplot, each experimental subject in the study is represented by a point in two-dimensional space. The coordinates of this point (X_i, Y_i) are the individual's (or object's) scores on variables X and Y, respectively. Examples of three such plots appear in Figure 9.1 on page 246.

predictor
criterion

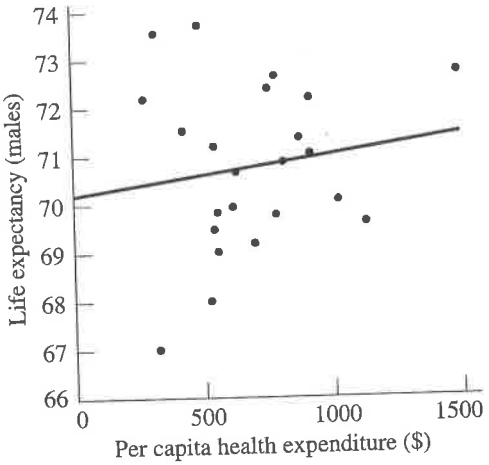
In a scatterplot, the **predictor** variable is traditionally represented on the abscissa, or x-axis, and the **criterion** variable on the ordinate, or y-axis. If the eventual purpose of the study is to predict one variable from knowledge of the other, the distinction is obvious; the criterion variable is the one to be predicted, whereas the predictor variable is the one from which the prediction is made. If the problem is simply one of obtaining a correlation coefficient, the distinction may be obvious (incidence of cancer would be dependent on amount smoked rather than the reverse, and thus incidence would appear on the ordinate), or it may not (neither running speed nor number of correct choices—common dependent variables in a learning study—is obviously in a dependent position relative to the other). Where the distinction is not obvious, it is irrelevant which variable is labeled X and which Y.

Consider the three scatter diagrams in Figure 9.1. Figure 9.1(a) is plotted from data reported by St. Leger, Cochrane, and Moore (1978) on the relationship between infant mortality, adjusted for gross national product, and the number of physicians per 10,000 population.[1] Notice the fascinating result that infant mortality *increases* with the number of physicians. That is certainly an unexpected result, but it is almost

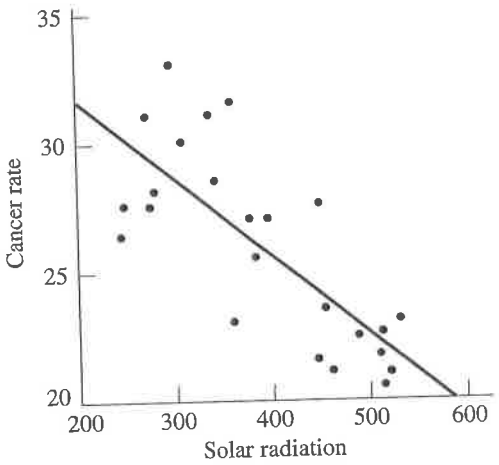
[1]Some people have asked how mortality can be negative. The answer is that this is the mortality rate *adjusted for* gross national product. After adjustment the rate can be negative.



(a) Infant mortality as a function of number of physicians



(b) Life expectancy as a function of health care expenditures



(c) Cancer rate as a function of solar radiation

FIGURE 9.1 *Three scatter diagrams*

certainly not due to chance. (As you look at these data and read the rest of the chapter you might think about possible explanations for this surprising result.)

The lines superimposed on Figures 9.1(a)–9.1(c) represent those straight lines that "best fit the data." How we determine that line will be the subject of much of this chapter. I have included the lines in each of these figures because they help to clarify the relationships. These lines are what we will call the **regression lines** of Y predicted on X (abbreviated "Y on X"), and they represent our best prediction of Y_i for a given value of X_i, for the ith subject or observation. Given any specified value of X, the cor-

regression lines

responding height of the regression line represents our best prediction of Y (designated $\hat{Y}$, and read "Y hat"). In other words, we can draw a vertical line from X_i to the regression line and then move horizontally to the y-axis and read $\hat{Y}_i$.

correlation (r)

The degree to which the points cluster around the regression line (in other words, the degree to which the actual values of Y agree with the predicted values) is related to the **correlation (r)** between X and Y. Correlation coefficients range between 1 and –1. For Figure 9.1(a) the points cluster very closely about the line, indicating that there is a strong linear relationship between the two variables. If the points fell exactly on the line, the correlation would be +1.00. As it is, the correlation is actually 0.81, which represents a high degree of relationship for real variables in the behavioral sciences.

In Figure 9.1(b) I have plotted data on the relationship between life expectancy (for males) and per capita expenditure on health care for 23 developed (mostly European) countries. These data are found in Cochrane, St. Leger, and Moore (1978). At a time when there is considerable discussion nationally about the cost of health care, these data give us pause. If we were to measure the health of a nation by life expectancy (admittedly not the only, and certainly not the best, measure), it would appear that the total amount of money we spend on health care bears no relationship to the resultant quality of health (assuming that different countries apportion their expenditures in similar ways). (Several hundred thousand dollars spent on transplanting an organ from a baboon into a 57-year-old male may increase *his* life expectancy by a few years, but it is not going to make a dent in the *nation's* life expectancy. A similar amount of money spent on prevention efforts with young children, however, may eventually have a very substantial effect—hence the inclusion of this example in a text primarily aimed at psychologists.) The two countries with the longest life expectancy (Iceland and Japan) spend nearly the same amount of money on health care as the country with the shortest life expectancy (Portugal). The United States has the second highest rate of expenditure but ranks near the bottom in life expectancy. Figure 9.1(b) represents a situation in which there is no apparent relationship between the two variables under consideration. If there were absolutely no relationship between the variables, the correlation would be 0. As it is, the correlation is only 0.14, and even that can be shown not to be reliably different from 0.

Finally, Figure 9.1(c) presents data from an article in *Newsweek* (1991) on the relationship between breast cancer and sunshine. For those of us who love the sun, it is encouraging to find that there may be at least some benefit from additional sunlight. Notice that as the amount of solar radiation increases, the incidence of deaths from breast cancer *decreases.* (It has been suggested that perhaps the higher rate of breast cancer with decreased sunlight is attributable to a Vitamin D deficiency.) This is a good illustration of a negative relationship, and the correlation here is –0.76.

It is important to note that the sign of the correlation coefficient has no meaning other than to denote the direction of the relationship. Correlations of 0.75 and –0.75 signify exactly the same *degree* of relationship. It is only the direction of that relationship that is different. Figures 9.1(a) and 9.1(c) illustrate this, because the two correlations are nearly the same except for their signs (0.81 versus –0.76).

An alternative approach to interpreting scatter diagrams can be seen in Figures 9.2 and 9.3 on page 248. These figures are based on data from the Achenbach Teacher

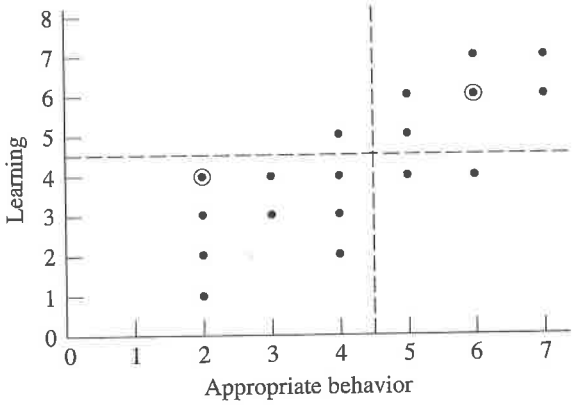


FIGURE 9.2 *Relationship between learning and appropriate behavior in normal boys, ages 12–16*

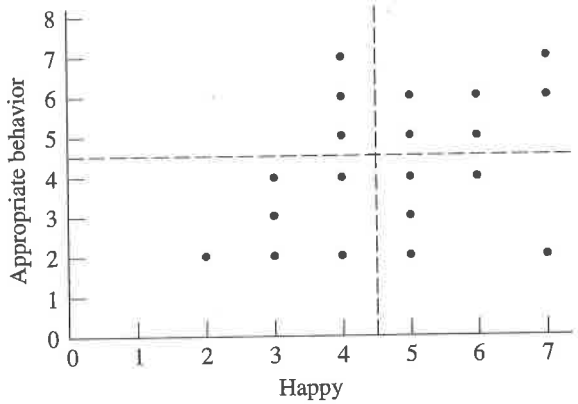


FIGURE 9.3 *Relationship between happiness and appropriate behavior in normal boys, ages 12–16*

Report form (Achenbach, 1991a), which is a rating form for behavior problems. (The data have been modified somewhat for purposes of this example.) Figure 9.2 plots the relationship between the teacher's rating of the degree to which the child exhibits behavior appropriate to the situation and the teacher's rating of how much the child is learning. All ratings are on 7-point scales. The dots with circles around them represent coincident points for two children. Figure 9.3 plots the degree to which the teacher rates the child as happy and the degree to which the child shows appropriate behavior.

In these figures, the vertical and horizontal grid lines divide both axes at their means. For example, the vertical line in Figure 9.2 divides those children who were below the mean on appropriate behavior from those who were above the mean. Similarly, the horizontal line separates those children who were below the mean on the learning variable from those who were above it. In both cases the variable we would

be most likely to think of as the dependent variable is plotted on the y-axis, and the likely independent variable on the x-axis. Admittedly, the choice is somewhat arbitrary.

If there were a strong positive relationship between learning and appropriate behavior, for example, we would expect that most of the children who were high (above the mean) on one variable would be high on the other. Likewise, most of those who were below the mean on one variable should be below it on the other. Such an idea can be represented by a simple table in which we count the number of children who were above the mean on both variables, the number below the mean on both variables, and the number above the mean on one and below it on the other. Such a table is shown in Table 9.1 for the data in Figures 9.2 and 9.3.

TABLE 9.1

(a) Appropriate Behavior versus Learning (Figure 9.2)

		Appropriate Behavior	
		Above Mean	Below Mean
Learning	Above Mean	7	1
	Below Mean	2	10

(b) Happy versus Appropriate Behavior (Figure 9.3)

		Happy	
		Above Mean	Below Mean
Appropriate Behavior	Above Mean	6	3
	Below Mean	5	6

With a strong positive relationship between the two variables, we would expect most of the data points in Table 9.1 to fall in the "Above–Above" and "Below–Below" cells, with only a smattering in the "Above–Below" and "Below–Above" cells. On the other hand, if the two variables are not related to each other, we would expect to see approximately equal numbers of data points in the four cells of the table (or quadrants of the scatter diagram). From Table 9.1(a) we see that for the relationship between appropriate behavior and learning, 17 out of the 20 children fall in the cells associated with a positive relationship between the variables. In other words, if they are below the mean on one variable, they are generally below the mean on the other, and vice versa. Only 3 of the children break this pattern. However, for the data plotted in Figure 9.3, Table 9.1(b) shows us that only 12 children are on the same side of the mean on both variables, whereas 5 children are below the mean on Appropriate Behavior but above it on Happy, and 3 children show the opposite pattern.

These two examples illustrate in a simple way the interpretation of scatter diagrams and the relationship between variables. In the first case there is a strong positive relationship between the variables. In the second case the relationship, although

positive, is considerably weaker. This result is reflected in the correlations between the variables. For Figure 9.2 the correlation is 0.78, whereas for Figure 9.3 it is 0.38. (Keep in mind that I have used small samples for ease of discussion, and these correlations might well be different if larger samples had been used. This is especially true because I hunted around to find somewhat extreme examples and may have managed to find unrepresentative ones. We will address the issue of the unreliability of sample results in later chapters.)

9.2 The Relationship between Stress and Health

Psychologists have long been interested in the relationship between stress and health, and have accumulated evidence to show that there are very real negative effects of stress on both the psychological and physical health of people. Wagner, Compas, and Howell (1988) investigated the relationship between stress and mental health in first-year college students. Using a scale they developed to measure the frequency, perceived importance, and desirability of recent life events, they created a measure of negative events weighted by the reported frequency and the respondent's subjective estimate of the impact of each event. This served as their measure of the subject's perceived social and environmental stress. They also asked students to complete the Hopkins Symptom Checklist, assessing the presence or absence of 57 psychological symptoms. The stem-and-leaf displays and boxplots for the stress and symptom measures are shown in Table 9.2.

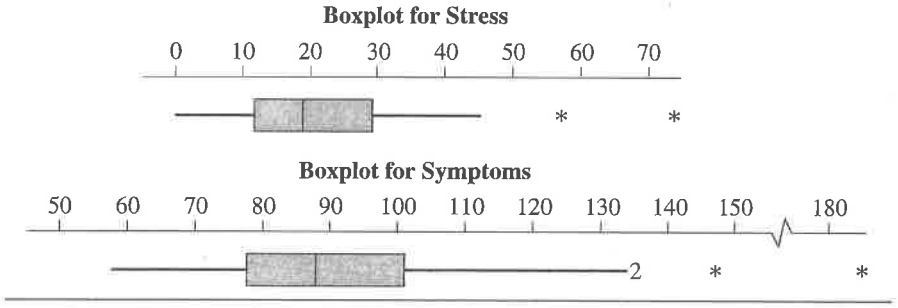
Before we consider the relationship between these variables, we need to study the variables individually. The stem-and-leaf displays for both variables show that the distributions are unimodal but are slightly positively skewed. Except for a few extreme values, there is nothing about either variable that should disturb us, such as extreme skewness or bimodality.[2] Note that there is a fair amount of variability in each variable. This variability is important, because if we want to show that different stress scores are associated with differences in symptoms, it is important to have these differences in the first place.

The boxplots in Table 9.2 reveal the presence of outliers on both variables. (The "2" is used to indicate the presence of two overlapping data points.) The existence of outliers should alert us to potential problems that these scores might cause. The first thing we might do is to check the data to see whether these few subjects were responding in unreasonable ways—for example, do they report the occurrence of all sorts of unlikely life events or symptoms, making us question the legitimacy of their responses? (Some subjects have been known to treat psychological experiments with something less than the respect and reverence they deserve! I'm sure that you find that hard to believe, but, sadly, it's true.) The second thing to check is whether the

[2]Bimodality does not make the use of correlation and regression inappropriate, but it is a signal that we should examine the data carefully and think about the interpretation.

TABLE 9.2 Description of data on the relationship between stress and mental health

Stem-and-Leaf for Stress		Stem-and-Leaf for Symptoms	
0*	01222234	5.	8
0.	556677888899	6*	112234
1*	01112222222333444444	6.	55668
1.	5556666777888999	7*	00012334444
2*	0001111223333334	7.	57788899
2.	556778999	8*	00011122233344
3*	012233444	8.	5666677888899
3.	56677778	9*	0111223344
4*	23444	9.	556679999
4.	55	10*	0001112224
		10.	567799
HI	57, 74	11*	112
		11.	78
		12*	11
Code:	2. | 5 = 25	12.	57
		13*	1
		HI	135, 135, 147, 186
		Code:	5. | 8 = 58



same subject produced outlying data points on both variables. This would suggest that this subject's data, although legitimate, might have a disproportionate influence on the resulting correlation. The third thing to do is to make a scatterplot of the data, again looking for the undue influence of particular extreme data points. (Such a scatterplot will appear later in Figure 9.4, p. 255) Finally, we can run our analyses including and excluding extreme points to see what differences appear in the results. If you carry out each of these steps on the data, you will find nothing to suggest that the outliers we have identified influenced the resulting correlation or regression equation in any important way. However, these steps are important precursors to any good analysis—if only because they give us greater faith in our final result. A more extensive discussion of techniques for examining data to be used in regression analyses will be found in Chapter 15 when we discuss multiple regression.

9.3 The Covariance

coveriance (cov_{XY} or s_{XY})

The correlation coefficient we seek to compute on the data[3] in Table 9.3 is itself based on a statistic called the **covariance** (cov_{XY} or s_{XY}). The covariance is basically a number that reflects the degree to which two variables vary together.

To define the covariance mathematically, we can write

$$cov_{XY} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N - 1}$$

From this equation it is apparent that the covariance is similar in form to the variance. If we changed all the Ys in the equation to Xs, we would have s_X^2; if we changed the Xs to Ys, we would have s_Y^2.

Some insight into the meaning of the covariance can be gained by first thinking back to what was said about the four quadrants in Figures 9.2 and 9.3, and then by considering what we would expect to find in the case of a positive correlation between stress and symptoms. In this situation, high stress scores will be paired with high symptom scores. Thus, for a stressed subject with many problems, both $(X - \overline{X})$ and $(Y - \overline{Y})$ will be positive and their product will be positive. For a subject experiencing

TABLE 9.3 Data on stress and symptoms for 10 representative participants

Participant	Stress (X)	Symptoms (Y)
1	30	99
2	27	94
3	9	80
4	20	70
5	3	100
6	15	109
7	5	62
8	10	81
9	23	74
10	34	121
⋮	⋮	⋮

$\sum X = 2297$ $\quad \sum Y = 9705$

$\sum X^2 = 67{,}489$ $\quad \sum Y^2 = 923{,}787$

$\overline{X} = 21.467$ $\quad \overline{Y} = 90.701$

$s_X = 13.096$ $\quad s_Y = 20.266$

$\sum XY = 222{,}576$

$N = 107$

[3] A copy of the complete data set is available on the CD in the file named Tab9-3.dat.

little stress and few problems, both $(X - \overline{X})$ and $(Y - \overline{Y})$ will be negative, but their product will again be positive. Thus, the sum of $(X - \overline{X})(Y - \overline{Y})$ will be large and positive, giving us a large positive covariance.

The reverse would be expected in the case of a strong negative relationship. Here, large positive values of $(X - \overline{X})$ most likely will be paired with large negative values of $(Y - \overline{Y})$, and vice versa. Thus, the sum of products of the deviations will be large and negative, indicating a strong negative relationship.

Finally, consider a situation in which there is no relationship between X and Y. In this case, a positive value of $(X - \overline{X})$ will sometimes be paired with a positive value and sometimes with a negative value of $(Y - \overline{Y})$. The result is that the products of the deviations will be positive about half the time and negative about half the time, producing a near-zero sum and indicating no relationship between the variables.

For a given set of data, it is possible to show that cov_{XY} will be at its positive maximum whenever X and Y are perfectly positively correlated ($r = 1.00$), and at its negative maximum whenever they are perfectly negatively correlated ($r = -1.00$). When the two variables are perfectly uncorrelated ($r = 0$), cov_{XY} will be zero.

For computational purposes a simple expression for the covariance is given by

$$\text{cov}_{XY} = \frac{\sum XY - \dfrac{\sum X \sum Y}{N}}{N - 1}$$

For the full data set represented in Table 9.3, the covariance is

$$\text{cov}_{XY} = \frac{222{,}576 - \dfrac{(2297)(9705)}{107}}{106} = \frac{222{,}576 - 208{,}340.05}{106}$$

$$= 134.301$$

9.4 The Pearson Product-Moment Correlation Coefficient (r)

What we said about the covariance might suggest that we could use it as a measure of the degree of relationship between two variables. An immediate difficulty arises, however, because the absolute value of cov_{XY} is also a function of the standard deviations of X and Y. Thus, a value of $\text{cov}_{XY} = 134$, for example, might reflect a high degree of correlation when the standard deviations are small, but a low degree of correlation when the standard deviations are high. To resolve this difficulty, we divide the covariance by the standard deviations and make this our estimate of correlation. Thus, we define

$$r = \frac{\text{cov}_{XY}}{s_X s_Y}$$

Since the maximum value of cov_{XY} can be shown to be $\pm s_X s_Y$, it follows that the limits on r are ± 1.00. One interpretation of r, then, is that it is a measure of the degree to which the covariance approaches its maximum.

From Table 9.3 and subsequent calculations, we know that $s_X = 13.096$, $s_Y = 20.266$, and $\text{cov}_{XY} = 134.301$. Then the correlation between X and Y is given by

$$r = \frac{\text{cov}_{XY}}{s_X s_Y}$$

$$r = \frac{134.301}{(13.096)(20.266)} = 0.506$$

This coefficient must be interpreted cautiously; do not attribute meaning to it that it does not possess. Specifically, $r = 0.506$ should *not* be interpreted to mean that there is 50.6% of a relationship (whatever that might mean) between stress and symptoms. The correlation coefficient is simply a point on the scale between –1 and 1, and the closer it is to either of those limits, the stronger is the relationship between the two variables. For a more specific interpretation, we can speak in terms of r^2, which will be discussed shortly. It is important to emphasize again that the sign of the correlation merely reflects the direction of the relationship and, possibly, the arbitrary nature of the scale. Changing a variable from "number of items correct" to "number of items incorrect" would reverse the sign of the correlation, but it would have no effect on its absolute value.

Adjusted r

correlation coefficient in the population, ρ (rho)

Although the correlation we have just computed is the one we normally report, it is not an unbiased estimate of the **correlation coefficient in the population,** denoted **(ρ) rho.** To see why this would be the case, imagine two randomly selected pairs of points—for example, (23, 18) and (40, 66). (I pulled those numbers out of the air.) If you plot these points and fit a line to them, the line will fit perfectly because, as you most likely learned in elementary school, two points determine a straight line. Since the line fits perfectly, the correlation will be 1.00, even though the points were chosen at random. Clearly, that correlation of 1.00 does not mean that the correlation in the population from which those points were drawn is 1.00 or anywhere near it. When the number of observations is small, the sample correlation will be a biased estimate of the population correlation coefficient. To correct for this we can compute what is known as the **adjusted correlation coefficient** (r_{adj}):

adjusted correlation coefficient (r_{adj})

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

This is a relatively unbiased estimate of the population correlation coefficient.

In the example we have been using, the sample size is reasonably large ($N = 107$). Therefore we would not expect a great difference between r and r_{adj}.

$$r_{adj} = \sqrt{1 - \frac{(1 - 0.506)^2(106)}{105}} = 0.499$$

which is very close to $r = 0.506$. This agreement will not be the case, however, for very small samples.

When we discuss multiple regression, which involves multiple predictors of Y, in Chapter 15, we will see that this equation for the adjusted correlation will continue

to hold. The only difference will be that the denominator will be $N - p - 1$, where p stands for the number of predictors. (That is where the $N - 2$ came from in this equation.)

We could draw a parallel between the adjusted r and the way we calculate a sample variance. As I explained earlier, in calculating the sample variance, we divide the sum of squared deviations by $N - 1$ to create an unbiased estimate of the population variance. That is comparable to what we do when we compute an adjusted r. The odd thing is that no one would seriously consider reporting anything but the unbiased estimate of the population variance, whereas we think nothing of reporting a biased estimate of the population correlation coefficient. I don't know why we behave inconsistently like that—we just do. The only reason I even discuss the adjusted value is that most computer software presents both statistics, and students are likely to wonder about the difference and which one they should care about.

9.5 The Regression Line

We have just seen that there is a reasonable degree of relationship between stress and psychological symptoms ($r = 0.506$). We can obtain a better idea of what this relationship is by looking at a scatterplot of the two variables and the regression line for predicting symptoms (Y) on the basis of stress (X). The scatterplot is shown in Figure 9.4, where the best-fitting line for predicting Y on the basis of X has been superimposed. We will see shortly where this line came from, but notice first the way in which the symptom scores increase linearly with increases in stress scores. Our

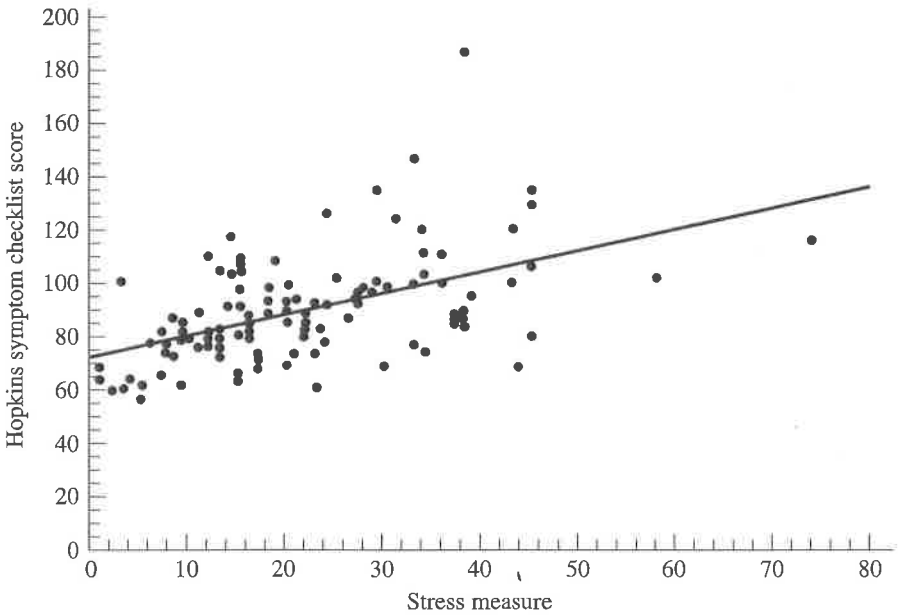


FIGURE 9.4 *Scatterplot of symptoms as a function of stress*

correlation coefficient told us that such a relationship existed, but it is easier to appreciate just what it means when you see it presented graphically. Notice also that the degree of scatter of points about the regression line remains about the same as you move from low values of stress to high values, although, with a correlation of approximately 0.50, the scatter is fairly wide. We will discuss scatter in more detail when we consider the assumptions on which our procedures are based.

As you may remember from high school, the equation of a straight line is an equation of the form $Y = bX + a$. For our purposes, we will write the equation as

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

where

$\hat{Y}$ = the predicted value of Y

slope

b = the **slope** of the regression line (the amount of difference in $\hat{Y}$ associated with a one-unit difference in X)

intercept

a = the **intercept** (the value of $\hat{Y}$ when $X = 0$)

X = the value of the predictor variable

Our task will be to solve for those values of a and b that will produce the best-fitting linear function. In other words, we want to use our existing data to solve for the values of a and b such that the regression line (the values of $\hat{Y}$ for different values of X) will come as close as possible to the actual obtained values of Y. But how are we to define the phrase "best fitting"? A logical way would be in terms of **errors of prediction**—that is, in terms of the $(Y - \hat{Y})$ deviations. Since $\hat{Y}$ is the value of the symptom variable that our equation would *predict* for a given level of stress, and Y is a value that we actually *obtained*, $(Y - \hat{Y})$ is the error of prediction, usually called the **residual.** We want to find the line (the set of $\hat{Y}$s) that minimizes such errors. We cannot just minimize the *sum* of the errors, however, because for an infinite variety of lines—any line that goes through the point $(\overline{X}, \overline{Y})$—that sum will always be zero. (We will overshoot some and undershoot others.) Instead, we will look for that line that minimizes the sum of the *squared* errors—that minimizes $\sum (Y - \hat{Y})^2$. (Note that I said much the same thing in Chapter 2 when I was discussing the variance. There I was discussing deviations from the mean, and here I am discussing deviations from the regression line—sort of a floating or changing mean. These two concepts—errors of prediction and variance—have much in common, as we shall see.)[4]

errors of prediction

residual

The optimal values of a and b can be obtained by solving for those values of a and b that minimize $\sum (Y - \hat{Y})^2$. The solution is not difficult, and those who wish can find it in earlier editions of this book or in Draper and Smith (1981, p. 13). The solution to the problem yields what are often called the **normal equations:**

normal equations

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

$$b = \frac{\text{cov}_{XY}}{s_X^2}$$

[4]For those who are interested, Rousseeuw and Leroy (1987) present a good discussion of alternative criteria that could be minimized, often to good advantage.

We now have solutions for a and b that will minimize $\sum (Y - \hat{Y})^2$.[5] To indicate that our solution was designed to minimize errors in predicting Y from X (rather than the other way around), the constants are sometimes denoted $a_{Y \cdot X}$ and $b_{Y \cdot X}$. When no confusion would arise, the subscripts are usually omitted. [When your purpose is to predict X on the basis of Y (i.e., X *on* Y), then you can simply reverse X and Y in the previous equations.]

As an example of the calculation of regression coefficients, consider the data in Table 9.3. From that table we know that $\overline{X} = 21.467$, $\overline{Y} = 90.701$, and $s_X = 13.096$. We also know that $\text{cov}_{XY} = 134.301$. Thus,

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{134.301}{13.096^2} = 0.7831$$

$$a = \overline{Y} - b\overline{X} = 90.701 - (0.7831)(21.467) = 73.891$$

$$\hat{Y} = bX + a = (0.7831)(X) + 73.891$$

We have already seen the scatter diagram with the regression line for Y on X superimposed in Figure 9.4. This is the equation of that line.[6]

A word is in order about actually plotting the regression line. To plot the line, you can simply take any two values of X (preferably at opposite ends of the scale), calculate $\hat{Y}$ for each, mark these coordinates on the figure, and connect them with a straight line. For our data, we have

$$\hat{Y}_i = (0.7831)(X_i) + 73.891$$

When $X_i = 0$,

$$\hat{Y}_i = (0.7831)(0) + 73.891 = 73.891$$

and when $X_i = 50$,

$$\hat{Y}_i = (0.7831)(50) + 73.891 = 113.046$$

The line then passes through the points ($X = 0$, $Y = 73.891$) and ($X = 50$, $Y = 113.046$), as shown in Figure 9.4. The regression line will also pass through the points $(0, a)$ and $(\overline{X}, \overline{Y})$, which provides a quick check on accuracy.

If you calculate both regression lines (Y on X and X on Y), it will be apparent that the two are not coincident. They do intersect at the point $(\overline{X}, \overline{Y})$, but they have different slopes. The fact that they are different lines reflects the fact that they were designed for different purposes—one minimizes $\sum (Y - \hat{Y})^2$ and the other minimizes $\sum (X - \hat{X})^2$. They both go through the point $(\overline{X}, \overline{Y})$ because a person who is *average* on one variable would be expected to be *average* on the other, but only when the correlation between the two variables is ± 1.00 will the lines be coincident.

---

[5] An interesting alternative formula for b can be written as $b = r(s_y/s_x)$. This shows explicitly the relationship between the correlation coefficient and the slope of the regression line. Note that when $s_y = s_x$, b will equal r. Can you think of a case where this would happen? (*Answer:* When both variables have a standard deviation of 1, which happens when the variables are standardized.)

[6] An excellent Java applet that allows you to enter individual data points and see their effect on the regression line is available at http://www.math.csusb.edu/faculty/stanton/m262/regress/regress.html.

Interpretations of Regression

In certain situations the regression line is useful in its own right. For example, a college admissions officer might be interested in an equation for predicting college performance on the basis of high-school grade point average (although she would most likely want to include multiple predictors in ways to be discussed in Chapter 15). Similarly, a farm manager might be interested in predicting milk production based on the nutrient value of the feed she is recommending. But these examples are somewhat unusual. In most applications of regression in psychology, we are not particularly interested in making an actual prediction. Although we might be interested in knowing the relationship between family income and educational achievement, it is unlikely that we would take any particular child's family-income measure and use that to predict his educational achievement. We are usually much more interested in general principles than in individual predictions. A regression equation, however, can in fact tell us something meaningful about these general principles, even though we may never actually use it to form a prediction for a specific case.

Intercept

We have defined the intercept as that value of $\hat{Y}$ when X equals zero. As such, it has meaning in some situations and not in others, primarily depending on whether or not $X = 0$ has meaning and is near or within the range of values of X used to derive the estimate of the intercept. If, for example, we took a group of overweight people and looked at the relationship between self-esteem (Y) and weight loss (X) (assuming that it is linear), the intercept would tell us what level of self-esteem to expect for an individual who lost 0 pounds. Often, however, there is no meaningful interpretation of the intercept other than a mathematical one. If we are looking at the relationship between self-esteem (Y) and actual weight (X) for adults, it is obviously foolish to ask what someone's self-esteem would be if he weighed 0 pounds. The intercept would appear to tell us this, but it represents such an extreme extrapolation from available data as to be meaningless. (In this case, a nonzero intercept would suggest a lack of linearity over the wider range of weight from 0 to 300 pounds, but we probably are not interested in nonlinearity in the extremes anyway.)

Slope

We have defined the slope as the change in $\hat{Y}$ for a one-unit change in X. As such it is a measure of the predicted *rate of change* in Y. By definition, then, the slope is often a meaningful measure. If we are looking at the regression of income on years of schooling, the slope will tell us how much of a difference in income would be associated with each additional year of school. Similarly, if an engineer knows that the slope relating fuel economy in miles per gallon (mpg) to weight of the automobile is 0.01, and if she can assume a causal relationship between mpg and weight, then she knows that for every pound that she can reduce the weight of the car she will increase

its fuel economy by 0.01 mpg. Thus, if the manufacturer replaces a 30-pound spare tire with one of those annoying 20-pound temporary ones, the car will gain 0.1 mpg.

Standardized Regression Coefficients

Although we rarely work with standardized data (data that have been transformed so as to have a mean of 0 and a standard deviation of 1 on each variable), it is worth considering what b would represent if the data for each variable were standardized separately. In that case, a difference of one unit in X or Y would represent a difference of one standard deviation. Thus, if the slope were 0.75, for standardized data, we would be able to say that a one standard deviation increase in X will be reflected in three-quarters of a standard deviation increase in $\hat{Y}$. When speaking of the slope coefficient for standardized data, we often refer to the **standardized regression coefficient** as β **(beta)** to differentiate it from the coefficient for nonstandardized data (b). We will return to the idea of standardized variables when we discuss multiple regression. (What would the intercept be if the variables were standardized? *Hint:* The line goes through the means.)

standardized regression coefficient β (beta)

Correlation and Beta

What we have just seen with respect to the slope for standardized variables is directly applicable to the correlation coefficient. Recall that r is defined as $\text{cov}_{XY}/s_X s_Y$, whereas b is defined as cov_{XY}/s_X^2. If the data are standardized, $s_X = s_Y = s_X^2 = 1$, and the slope and the correlation coefficient will be equal. Thus, one interpretation of the correlation coefficient is that it is equal to what the slope would be if the variables were standardized. That suggests that a derivative interpretation of $r = 0.80$, for example, is that one standard deviation difference in X is associated *on the average* with an eight-tenths of a standard deviation difference in Y. In some situations such an interpretation can be meaningfully applied.

A Note of Caution

What has just been said about the interpretation of b and r must be tempered with a bit of caution. To say that a one-unit difference in family income is associated with a 0.75 unit difference in academic achievement is not to be interpreted to mean that raising family income for Mary Smith will automatically raise her academic achievement. In other words, we are not speaking about cause and effect. We can say that people who score higher on the income variable also score higher on the achievement variable without in any way implying causation or suggesting what would happen to a given individual if her family income were to increase. Family income is associated (in a correlational sense) with a host of other variables (e.g., attitudes toward education, number of books in the home, access to a variety of environments) and there is

no reason to expect all of these to change merely because income changes. Those who argue that eradicating poverty will lead to a wide variety of changes in people's lives often fall into such a cause-and-effect trap. Eradicating poverty is certainly a worthwhile and important goal, but the correlations between income and educational achievement *may* be totally irrelevant to the issue.

9.6 The Accuracy of Prediction

The fact that we can fit a regression line to a set of data does not mean that our problems are solved. On the contrary, they have only begun. The important point is not whether a straight line can be drawn through the data (you can always do that) but whether that line represents a reasonable fit to the data—in other words, whether our effort was worthwhile.

In beginning a discussion of errors of prediction, it is instructive to consider the situation in which we wish to predict Y without any knowledge of the value of X.

The Standard Deviation as a Measure of Error

As mentioned earlier, the data plotted in Figure 9.4 represent the number of symptoms shown by students (Y) as a function of the number of stressful life events (X). Assume that you are now given the task of predicting the number of symptoms that will be shown by a particular individual, but that you have no knowledge of the number of stressful life events he or she has experienced. Your best prediction in this case would be the mean number of symptoms ($\overline{Y}$) (averaged across all subjects), and the error associated with your prediction would be the standard deviation of Y (i.e., s_Y), since your prediction is the mean and s_Y deals with deviations around the mean. We know that s_Y is defined as

$$s_Y = \sqrt{\frac{\sum (Y - \overline{Y})^2}{N - 1}}$$

or, in terms of the variance,

$$s_Y^2 = \frac{\sum (Y - \overline{Y})^2}{N - 1}$$

sum of squares (SS_Y)

The numerator is the sum of squared deviations from $\overline{Y}$ (the point you would have predicted in this example) and is what we will refer to as the **sum of squares of Y (SS_Y).** The denominator is simply the degrees of freedom. Thus, we can write

$$s_Y^2 = \frac{SS_Y}{df}$$

The Standard Error of Estimate

Now suppose we wish to make a prediction about symptoms for a student who has a specified number of stressful life events. If we had an infinitely large sample of data,

our prediction for symptoms would be the mean of those values of symptoms (Y) that were obtained by all students who had that particular value of stress. In other words, it would be a conditional mean—conditioned on that value of X. We do not have an infinite sample, however, so we will use the regression line. (If all of the assumptions that we will discuss shortly are met, the expected value of the Y scores associated with each specific value of X would lie on the regression line.) In our case, we know the relevant value of X and the regression equation, and our best prediction would be $\hat{Y}$. In line with our previous measure of error (the standard deviation), the error associated with the present prediction will again be a function of the deviations of Y about the predicted point, but in this case the predicted point is $\hat{Y}$ rather than $\overline{Y}$. Specifically, a measure of error can now be defined as

$$s_{Y \cdot X} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N - 2}} = \sqrt{\frac{SS_{residual}}{df}}$$

and again the sum of squared deviations is taken about the prediction ($\hat{Y}$). The sum of squared deviations about $\hat{Y}$ is often denoted $SS_{residual}$ because it represents variability that remains *after* we use X to predict Y.[7] The statistic $s_{Y \cdot X}$ is called the **standard error of estimate.** It is denoted as $s_{Y \cdot X}$ to indicate that it is the standard deviation of Y predicted from X. It is the most common (although not always the best) measure of the error of prediction. Its square, $s^2_{Y \cdot X}$, is called the **residual variance** or **error variance,** and it can be shown to be an unbiased estimate of the corresponding parameter ($\sigma^2_{Y \cdot X}$) in the population. We have $N - 2$ df because we lost 2 degrees of freedom in estimating our regression line. (Both a and b were estimated from sample data.)

standard error of estimate

residual variance

error variance

I have suggested that if we had an infinite number of observations, our prediction for a given value of X would be the mean of the Ys associated with that value of X. This idea helps us appreciate what $s_{Y \cdot X}$ is. If we had the infinite sample and calculated the variances for the Ys at each value of X, the average of those variances would be the residual variance, and its square root would be $s_{Y \cdot X}$. The set of Ys corresponding to a specific X is called a **conditional distribution** of Y because it is the distribution of Y scores for those cases that meet a certain condition with respect to X. We say that these standard deviations are conditional on X because we calculate them from Y values corresponding to specific values of X. On the other hand, our usual standard deviation of Y (s_Y) is not conditional on X because we calculate it using all values of Y, regardless of their corresponding X values.

conditional distribution

One way to obtain the standard error of estimate would be to calculate $\hat{Y}$ for each observation and then to find $s_{Y \cdot X}$ directly, as has been done in Table 9.4 on page 262. Finding the standard error using this technique is hardly the most enjoyable way to spend a winter evening. Fortunately, a much simpler procedure exists. It not only provides a way of obtaining the standard error of estimate, but also leads directly into even more important matters.

[7] It is also frequently denoted SS_{error} because it is a sum of squared errors of prediction.

TABLE 9.4 Direct calculation of the standard error of estimate

Subject	Stress (X)	Symptoms (Y)	$\hat{Y}$	$Y - \hat{Y}$
1	30	99	97.383	1.617
2	27	94	95.034	–1.034
3	9	80	80.938	–0.938
4	20	70	89.552	–19.552
5	3	100	76.239	23.761
6	15	109	85.636	23.364
7	5	62	77.806	–15.806
8	10	81	81.721	–0.721
9	23	74	91.902	–17.902
10	34	121	100.515	20.485
⋮	⋮	⋮	⋮	⋮

$$\sum (Y - \hat{Y}) = 0$$

$$\sum (Y - \hat{Y})^2 = 32{,}388.049$$

$$s^2_{Y \cdot X} = \frac{\sum (Y - \hat{Y})^2}{N - 2} = \frac{32{,}388.049}{105} = 308.458 \qquad s_{Y \cdot X} = \sqrt{308.458} = 17.563$$

r^2 and the Standard Error of Estimate

In much of what follows, we will abandon the term *variance* in favor of sums o squares (SS). As you should recall, a variance is a sum of squared deviations from th mean (generally known as a sum of squares) divided by the degrees of freedom. Th problem with variances is that they are not additive unless they are based on the sam *df*. Sums of squares are additive regardless of the degrees of freedom and thus ar much easier measures to use.[8]

We earlier defined the residual or error variance as

$$s^2_{Y \cdot X} = \frac{\sum (Y - \hat{Y})^2}{N - 2} = \frac{\text{SS}_{\text{residual}}}{N - 2}$$

With considerable algebraic manipulation, it is possible to show

$$s_{Y \cdot X} = s_Y \sqrt{(1 - r^2)\frac{N - 1}{N - 2}}$$

For large samples the fraction $(N - 1)/(N - 2)$ is essentially 1, and we can thus wr the equation as it is often found in statistics texts:

$$s^2_{Y \cdot X} = s^2_Y (1 - r^2)$$

[8]Later in the book when I wish to speak about a variance-type measure but do not want to specify whethe is a variance, a sum of squares, or something similar, I will use the vague, wishy-washy term *variation*.

or

$$s_{X \cdot Y} = s_Y \sqrt{(1 - r^2)}$$

Keep in mind, however, that for small samples these equations are only an approximation and $s^2_{Y \cdot X}$ will underestimate the error variance by the fraction $(N - 1)/(N - 2)$. For samples of any size, however, $SS_{residual} = SS_Y(1 - r^2)$. This particular formula is going to play a role throughout the rest of the book, especially in Chapters 15 and 16.

Errors of Prediction as a Function of r

Now that we have obtained an expression for the standard error of estimate in terms of r, it is instructive to consider how this error decreases as r increases. In Table 9.5, we see the magnitude of the standard error relative to the standard deviation of Y (the error to be expected when X is unknown) for selected values of r.

TABLE 9.5 The standard error of estimate as a function of r

r	$s_{Y \cdot X}$	r	$s_{Y \cdot X}$
0	s_Y	0.60	$0.800s_Y$
0.10	$0.995s_Y$	0.70	$0.714s_Y$
0.20	$0.980s_Y$	0.80	$0.600s_Y$
0.30	$0.954s_Y$	0.866	$0.500s_Y$
0.40	$0.917s_Y$	0.90	$0.436s_Y$
0.50	$0.866s_Y$	0.95	$0.312s_Y$

The values in Table 9.5 are somewhat sobering in their implications. With a correlation of 0.20, the standard error of our estimate is fully 98% of what it would be if X were unknown. This means that if the correlation is 0.20, using $\hat{Y}$ as our prediction rather than $\overline{Y}$ (i.e., taking X into account) reduces the standard error by only 2%. Even more discouraging is that if r is 0.50, as it is in our example, the standard error of estimate is still 87% of the standard deviation. To reduce our error to one-half of what it would be without knowledge of X requires a correlation of 0.866, and even a correlation of 0.95 reduces the error by only about two-thirds. All of this is not to say that there is nothing to be gained by using a regression equation as the basis of prediction, only that the predictions should be interpreted with a certain degree of caution. All is not lost, however, because it is often the kinds of relationships we see, rather than their absolute magnitudes, that are of interest to us.

r^2 as a Measure of Predictable Variability

From the preceding equation expressing residual error in terms of r^2, it is possible to derive an extremely important interpretation of the correlation coefficient. We have already seen that

$$SS_{residual} = SS_Y(1 - r^2)$$

Expanding and rearranging, we have

$$SS_{residual} = SS_Y - SS_Y(r^2)$$

$$r^2 = \frac{SS_Y - SS_{residual}}{SS_Y}$$

In this equation, SS_Y, which you know to be equal to $\sum (Y - \overline{Y})^2$, is the sum of squares of Y and represents the totals of

1. The part of the sum of squares of Y that is related to X [i.e., $SS_Y(r^2)$]
2. The part of the sum of squares of Y that is independent of X [i.e., $SS_{residual}$]

In the context of our example, we are talking about that part of the number of symptoms people exhibited that is related to how many stressful life events they had experienced, and that part that is related to other things. The quantity $SS_{residual}$ is the sum of squares of Y that is independent of X and is a measure of the amount of error remaining even after we use X to predict Y. These concepts can be made clearer with a second example.

Suppose we were interested in studying the relationship between amount of cigarette smoking (X) and age at death (Y). As we watch people die over time, we notice several things. First, we see that not all die at precisely the same age. There is variability in age at death regardless of smoking behavior, and this variability is measured by $SS_Y = \sum (Y - \overline{Y})^2$. We also notice that some people smoke more than others. This variability in smoking regardless of age at death is measured by $SS_X = \sum (X - \overline{X})^2$. We further find that cigarette smokers tend to die earlier than nonsmokers, and heavy smokers earlier than light smokers. Thus, we write a regression equation to predict Y from X. Since people differ in their smoking behavior, they will also differ in their *predicted* life expectancy ($\hat{Y}$), and we will label this variability $SS_{\hat{Y}} = \sum (\hat{Y} - \overline{Y})^2$. This last measure is variability in Y that is directly attributable to variability in X, since different values of $\hat{Y}$ arise from different values of X and the same values of $\hat{Y}$ arise from the same value of X—that is, $\hat{Y}$ does not vary unless X varies.

We have one last source of variability: the variability in the life expectancy of those people who smoke exactly the same amount. This is measured by $SS_{residual}$ and is the variability in Y that cannot be explained by the variability in X (since these people do not differ in the amount they smoke). These several sources of variability (sums of squares) are summarized in Table 9.6.

TABLE 9.6 Sources of variance in regression for the study of smoking and life expectancy

SS_X	= variability in amount smoked = $\sum (X - \overline{X})^2$
SS_Y	= variability in life expectancy = $\sum (Y - \overline{Y})^2$
$SS_{\hat{Y}}$	= variability in life expectancy directly attributable to variability in smoking behavior = $\sum (\hat{Y} - \overline{Y})^2$
$SS_{residual}$	= variability in life expectancy that cannot be attributed to variability in smoking behavior = $\sum (Y - \hat{Y})^2 = SS_Y - SS_{\hat{Y}}$

If we considered the absurd extreme in which all of the nonsmokers die at exactly age 72 and all of the smokers smoke precisely the same amount and die at exactly age 68, then all of the variability in life expectancy is directly predictable from variability in smoking behavior. If you smoke you will die at 68, and if you don't you will die at 72. Here $SS_{\hat{Y}} = SS_Y$, and $SS_{residual} = 0$.

As a more realistic example, assume smokers tend to die earlier than nonsmokers, but within each group there is a certain amount of variability in life expectancy. This is a situation in which some of SS_Y is attributable to smoking ($SS_{\hat{Y}}$) and some is not ($SS_{residual}$). What we want to be able to do is to specify what *percentage* of the overall variability in life expectancy is attributable to variability in smoking behavior. In other words, we want a measure that represents

$$\frac{SS_{\hat{Y}}}{SS_Y} = \frac{SS_Y - SS_{residual}}{SS_Y}$$

As we have seen, that measure is r^2. In other words,

$$r^2 = \frac{SS_{\hat{Y}}}{SS_Y}$$

This interpretation of r^2 is extremely useful. If, for example, the correlation between amount smoked and life expectancy were an unrealistically high 0.80, we could say that $0.80^2 = 64\%$ of the variability in life expectancy is directly predictable from the variability in smoking behavior. (Obviously, this is an outrageous exaggeration of the real world.) If the correlation were a more likely $r = 0.10$, we would say that $0.10^2 = 1\%$ of the variability in life expectancy is related to smoking behavior, whereas the other 99% is related to other factors.

Phrases such as "accounted for by," "attributable to," "predictable from," and "associated with" are *not* to be interpreted as statements of cause and effect. Thus, you could say, "I can predict 10% of the variability of the weather by paying attention to twinges in the ankle that I broke last year—when it aches we are likely to have rain, and when it feels fine the weather is likely to be clear." This does not imply that sore ankles cause rain, or even that rain itself causes sore ankles. For example, it might be that your ankle hurts when it rains because low barometric pressure, which is often associated with rain, somehow affects ankles.

From this discussion it should be apparent that r^2 is easier to interpret as a measure of correlation than is r, since it represents the degree to which the variability in one measure is attributable to variability in the other measure. I recommend that you always square correlation coefficients to get some idea of whether you are talking about anything important. In our symptoms-and-stress example, $r^2 = 0.506^2 = 0.256$. Thus, about one-quarter of the variability in symptoms can be predicted from variability in stress. That strikes me as an impressive level of prediction, given all the other factors that influence psychological symptoms.

There is not universal agreement that r^2 is our best measure of the contribution of one variable to the prediction of another, although that is certainly the most popular measure. Judd and McClelland (1989) strongly endorse r^2 because, when we index error in terms of the sum of squared errors, it is the **proportional reduction in error**

proportional reduction in error (PRE)

(PRE). In other words, when we do not use X to predict Y, our error is SS_Y. When we use X as the predictor, the error is $SS_{residual}$. Since

$$r^2 = \frac{SS_Y - SS_{\hat{Y}}}{SS_Y}$$

the value of r^2 can be seen to be the percentage by which error is reduced when X is used as the predictor.[9]

proportional improvement in prediction (PIP)

Others, however, have suggested the **proportional improvement in prediction (PIP)** as a better measure.

$$\text{PIP} = 1 - \sqrt{(1 - r^2)}$$

For large sample sizes this statistic is the *reduction* in the size of the standard error of estimate (see Table 9.5). Similarly, as we shall see shortly, it is a measure of the reduction in the width of the confidence interval on our prediction.

The choice between r^2 and PIP is really dependent on how you wish to measure error. When we focus on r^2 we are focusing on measuring error in terms of sums of squares. When we focus on PIP we are measuring error in standard deviation units.

Darlington (1990) has argued for the use of r instead of r^2 as representing the magnitude of an effect. A strong argument in this direction was also made by Ozer (1985), whose paper is well worth reading. In addition, Rosenthal and Rubin (1982) have shown that even small values of r^2 (or almost any other measure of the magnitude of an effect) can be associated with powerful effects, regardless of how you measure that effect (see Chapter 10).

I have discussed r^2 as an index of percentage of variation for a particular reason. There is a very strong movement, at least in psychology, toward more frequent reporting of the magnitude of an effect, rather than just a test statistic and a p value. As I mention in Chapter 7, there are two major types of magnitude measures. One type is called effect size, and is represented by Cohen's **d,** which is most appropriate when we have means of two or more groups. The second type of measure is the "percentage of variation," of which r^2 is the most common representative. We first saw this measure in Chapter 7, where we were looking at the percentage of variation in sexual arousal scores that was associated with the presence or absence of homophobia. We saw it in this chapter, where we found that 25.6% of the variation in psychological symptoms is associated with variation in stress. We will see it again in Chapter 10 when we cover the point-biserial correlation. It will come back again in the analysis of variance chapters (especially Chapters 11 and 13), where it will be disguised as eta-squared and related measures. Finally, it will appear in important ways when we talk about multiple regression. The common thread through all of this is that we want some measure of how much of the variation in a dependent variable is attributable to variation in an independent variable, whether that independent variable be categorical or continuous.

[9]It is interesting to note that r^2_{adj} (defined on p. 254) is nearly equivalent to the ratio of the *variance* terms corresponding to the sums of squares in the equation. (Well, it is interesting to *some* people.)

9.7 Assumptions Underlying Regression and Correlation

We have derived the standard error of estimate and other statistics without making any assumptions concerning the population(s) from which the data were drawn. Nor do we need such assumptions to use $s_{Y \cdot X}$ as an unbiased estimator of $\sigma_{Y \cdot X}$. If we are to use $s_{Y \cdot X}$ in any meaningful way, however, we will have to introduce certain parametric assumptions. To understand why, consider the data plotted in Figure 9.5. Notice the four statistics labeled $s^2_{Y \cdot 1}$, $s^2_{Y \cdot 2}$, $s^2_{Y \cdot 3}$, and $s^2_{Y \cdot 4}$. Each represents the variance of the points around the regression line in an **array** of X (the residual variance of Y conditional on a specific X). As mentioned earlier, the average of these variances, weighted by the degrees of freedom for each array, would be $s^2_{Y \cdot X}$, the residual or error variance. If $s^2_{Y \cdot X}$ is to have any practical meaning, it must be representative of the various terms of which it is an average. This leads us to the assumption of **homogeneity of variance in arrays,** which is nothing but the assumption that the variance of Y for each value of X is constant (in the population). This assumption will become important when we apply tests of significance using $s^2_{Y \cdot X}$.

array

homogeneity of variance in arrays

One further assumption that will be necessary when we come to testing hypotheses is that of **normality in arrays.** We will assume that in the population the values of Y corresponding to any specified value of X—that is, the **conditional array** of Y for X_i—are normally distributed around $\hat{Y}$. This assumption is directly analogous to the normality assumption we made with the t test—that each treatment population was normally distributed around its own mean—and we make it for similar reasons.

normality in arrays

conditional array

To anticipate what we will discuss in Chapter 11, note that our assumptions of homogeneity of variance and normality in arrays are equivalent to the assumptions of homogeneity of variance and normality of populations that we will make in discussing the analysis of variance. In Chapter 11 we will assume that the treatment populations from which data were drawn are normally distributed and all have the same variance.

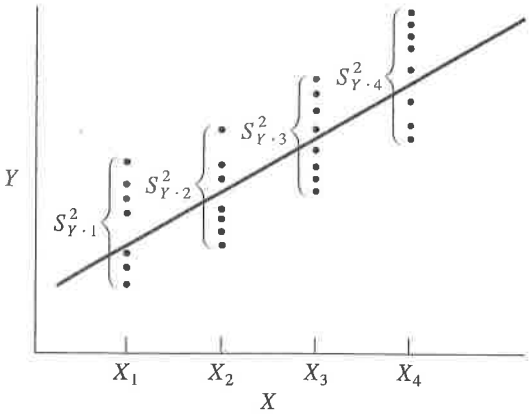


FIGURE 9.5 *Scatter diagram illustrating regression assumptions*

If you think of the levels of X in Figure 9.5 as representing different experimental conditions, you can see the relationship between the regression and analysis of variance assumptions.

The assumptions of normality and homogeneity of variance in arrays are associated with the regression model, where we are dealing with fixed values of X. On the other hand, when our interest is centered on the correlation between X and Y, we are dealing with the bivariate model, in which X and Y are both random variables. In this case, we are primarily concerned with using the sample correlation (r) as an estimate of the correlation coefficient in the population (ρ). Here we will replace the regression model assumptions with the assumption that we are sampling from a bivariate normal distribution.

The bivariate normal distribution looks roughly like the pictures you see each fall of surplus wheat piled in the main street of some midwestern town. The way the grain pile falls off on all sides resembles a normal distribution. (If there were no correlation between X and Y, the pile would look as though all the grain were dropped in the center of the pile and spread out symmetrically in all directions. When X and Y are correlated the pile is elongated, as when grain is dumped along a street and spreads out to the sides and down the ends.) Imagine that the pile had actually been dumped on top of a huge scattergram, with the main axis of the pile oriented along the regression line. If you sliced the pile on a line corresponding to any given value of X, you would see that the cut end is a normal distribution. You would also have a normal distribution if you sliced the pile along a line corresponding to any given value of Y. These are called **conditional distributions** because the first represents the distribution of Y given (conditional on) a specific value of X, whereas the second represents the distribution of X conditional on a specific value of Y. If, instead, we looked at *all* the values of Y regardless of X (or all values of X regardless of Y), we would have what is called the **marginal distribution** of Y (or X). For a bivariate normal distribution, both the conditional and the marginal distributions will be normally distributed. (Recall that for the regression model we assumed only normality of Y in the arrays of X—what we now know as conditional normality of Y. For the regression model, there is no assumption of normality of the conditional distribution of X or of the marginal distributions.)

conditional distributions

marginal distribution

9.8 Confidence Limits on Y

Although the standard error of estimate is useful as an overall measure of error, it is not a good estimate of the error associated with any single prediction. When we wish to predict a value of Y for a given subject, the error in our estimate will be smaller when X is near $\overline{X}$ than when X is far from $\overline{X}$. (For an intuitive understanding of this, consider what would happen to the predictions for different values of X if we rotated the regression line slightly around the point $(\overline{X}, \overline{Y})$. There would be negligible changes near the means, but there would be substantial changes in the extremes.) If we wish to predict Y on the basis of X for a new member of the population (someone

who was not included in the original sample), the standard error of our prediction is given by

$$s'_{Y \cdot X} = s_{Y \cdot X}\sqrt{1 + \frac{1}{N} + \frac{(X_i - \overline{X})^2}{(N-1)s_X^2}}$$

where $X_i - \overline{X}$ is the deviation of the individual's X score from the mean of X. This leads to the following confidence limits on Y:

$$\text{CI}(Y) = \hat{Y} \pm (t_{\alpha/2})(s'_{Y \cdot X})$$

This equation will lead to elliptical confidence limits around the regression line, which are narrowest for $X = \overline{X}$ and become wider as $| X - \overline{X} |$ increases.

To take a specific example, assume that we wanted to set confidence limits on the number of symptoms (Y) experienced by a student with a stress score of 10—a fairly low level of stress. We know that

$$\begin{aligned} s_{Y \cdot X} &= 17.563 \\ s_X^2 &= 171.505 \\ \overline{X} &= 21.467 \\ \hat{Y} &= 0.7831(10) + 73.891 = 81.722 \\ t_{0.025} &= 1.984 \\ N &= 107 \end{aligned}$$

Then

$$s'_{Y \cdot X} = s_{Y \cdot X}\sqrt{1 + \frac{1}{N} + \frac{(X_i - \overline{X})^2}{(N-1)s_X^2}}$$

$$s'_{Y \cdot X} = 17.563\sqrt{1 + \frac{1}{107} + \frac{(10 - 21.467)^2}{(106)171.505}}$$

$$= 17.563\sqrt{1.0166} = 17.708$$

Then

$$\begin{aligned} \text{CI}(Y) &= \hat{Y} \pm (t_{\alpha/2})(s'_{Y \cdot X}) \\ &= 81.722 \pm 1.984(17.708) \\ &= 81.722 \pm 35.133 \\ 46.589 &\leq Y \leq 116.855 \end{aligned}$$

The confidence interval is 46.589 to 116.855, and the probability is 0.95 that an interval computed in this way will include the level of symptoms reported by an individual whose stress score is 10. That interval is wide, but it is not as large as the 95% confidence interval of $50.5 \leq Y \leq 130.9$ that we would have had if we had not used X—that is, if we had just based our confidence interval on the obtained values of Y (and s_Y) rather than making it conditional on X.

9.9 A Computer Example Showing the Role of Test-Taking Skills

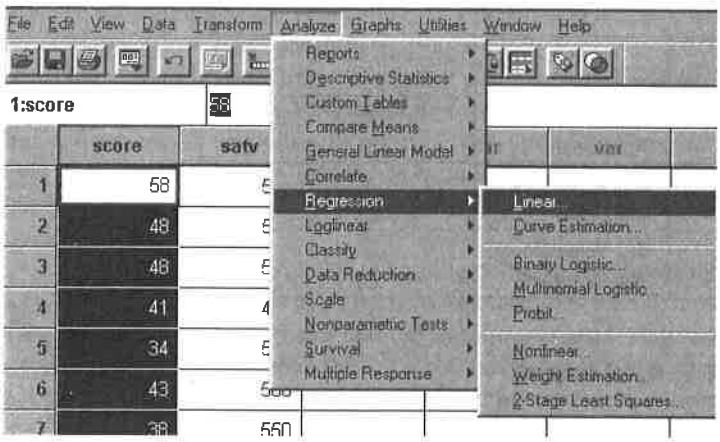
Most of us can do reasonably well if we study a body of material and then take an exam on that material. But how would we do if we just took the exam without even looking at the material? Katz, Lautenschlager, Blackburn, and Harris (1990) examined that question by asking some students to read a passage and then answer a series of multiple-choice questions, and asking others to answer the questions without having seen the passage. We will concentrate on the second group. The task described here is very much like the task that North American students face when they take the SAT exams for admission to University. This led the researchers to suspect that students who did well on the SAT would also do well on this task, since they both involve test-taking skills such as eliminating unlikely alternatives.

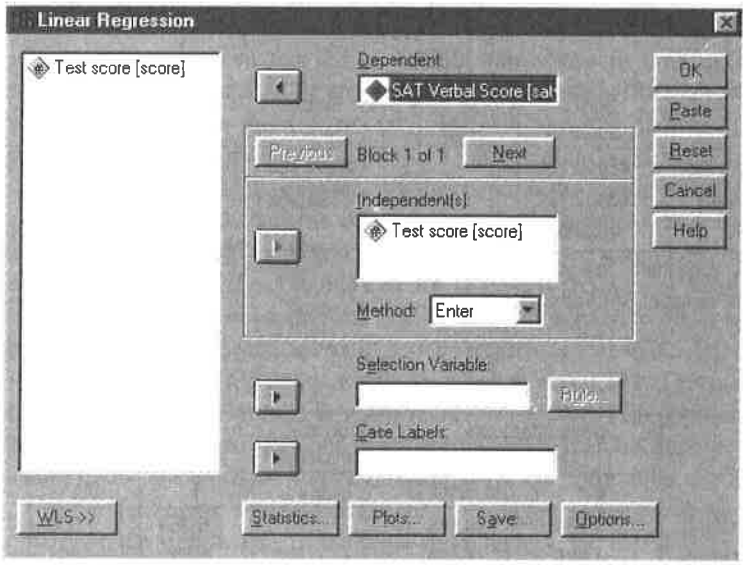
Data with the same sample characteristics as the data obtained by Katz et al. are given in Table 9.7. The variable Score represents the percentage of items answered correctly when the student has not seen the passage, and the variable SATV is the student's verbal SAT score from his or her college application.

Exhibit 9.1 illustrates the analysis using SPSS regression. There are a number of things here to point out. First, we must decide which is the dependent variable and which is the independent variable. This would make no difference if we just wanted to compute the correlation between the variables, but it is important in regression. In this case I have made a relatively arbitrary decision that my interest lies primarily in seeing whether people who do well at making intelligent guesses also do well on the

TABLE 9.7 Data based on Katz et al. (1990) for the group that did not read the passage

Score	SATV	Score	SATV
58	590	48	590
48	580	41	490
34	550	43	580
38	550	53	700
41	560	60	690
55	800	44	600
43	650	49	580
47	660	33	590
47	600	40	540
46	610	53	580
40	620	45	600
39	560	47	560
50	570	53	630
46	510	53	620





Descriptive Statistics

	Mean	Std. Deviation	N
SAT Verbal Score	598.57	61.57	28
Test Score	46.21	6.73	28

(continued)

EXHIBIT 9.1 SPSS output on Katz et al. (1990) study of test-taking behavior

Correlations

		SAT Verbal Score	Test Score
Pearson Correlation	SAT Verbal Score	1.000	.532
	Test Score	.532	1.000
Sig. (1-tailed)	SAT Verbal Score	.	.002
	Test Score	.002	.
N	SAT Verbal Score	28	28
	Test Score	28	28

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.532	.283	.255	53.13

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	28940.123	1	28940.123	10.251	.004[a]
	Residual	73402.734	26	2823.182		
	Total	102342.9	27			

a. Predictors: (Constant), Test score

b. Dependent Variable: SAT Verbal Score

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	373.736	70.938		5.269	.000
	Test score	4.865	1.520	.532	3.202	.004

a. Dependent Variable: SAT Verbal Score

EXHIBIT 9.1 *(continued)*

SAT. Therefore I am using SATV as the dependent variable, even though it was actually taken prior to the experiment. The first two panels of Exhibit 9.1 illustrate the menu selections required for SPSS. The means and standard deviations are found in the middle of the output, and you can see that we are dealing with a group that has high achievement scores (the mean is almost 600, with a standard deviation of about 60. This puts them about 100 points above the average for the SAT. They also do quite well on Katz' test, getting nearly 50% of the items correct. Below these statistics you

see the correlation between Score and SATV, which is 0.532. We will test this correlation for significance in a moment.

In the section labeled Model Summary you see both R and R^2. The "R" here is capitalized because if there were multiple predictors it would be a multiple correlation, and we always capitalize that symbol. One thing to note is that R here is calculated as the square root of R^2, and as such it will always be positive, even if the relationship is negative. This is a result of the fact that the procedure is applicable for multiple predictors.

The ANOVA table is a test of the null hypothesis that the correlation is 0 in the population. We will discuss hypothesis testing next, but what is most important here is that the test statistic is F, and that the significance level associated with that F is $p = 0.004$. Since p is less than 0.05, we will reject the null hypothesis and conclude that the variables are not linearly independent. In other words, there is a linear relationship between how well students score on a test that reflects test-taking skills, and how well they perform on the SAT. The exact nature of this relationship is shown in the next part of the printout. Here we have a table labeled Coefficients, and this table gives us the intercept and the slope. The intercept is labeled here as Constant, because it is the constant that you add to every prediction. In this case it is 373.736. Technically it means that if a student answered 0 questions correctly on Katz' test, we would expect them to have an SAT of approximately 370. Since a score of 0 would be so far from the scores these students actually obtained (and it is hard to imagine anyone earning a 0 even by guessing), I would not pay very much attention to that value.

In this table the slope is labeled by the name of the predictor variable. (All software solutions do this, because if there were multiple predictors we would have to know which variable goes with which slope. The easiest way to do this is to use the variable name as the label.) In this case the slope is 4.865, which means that two students who differ by 1 point on Katz' test would be predicted to differ by 4.865 on the SAT. Our regression equation would now be written as $\hat{Y} = 4.865(\text{Score}) + 373.736$.

The standardized regression coefficient is shown as 0.532. This means that a one standard deviation difference in test scores is associated with approximately a one-half standard deviation difference in SAT scores. Note that, because we have only one predictor, this standardized coefficient is equal to the correlation coefficient.

To the right of the standardized regression coefficient you will see t and p values for tests on the significance of the slope and intercept. We will discuss the test on the slope shortly. The test on the intercept is rarely of interest, but its interpretation should be evident from what I say about testing the slope.

9.10 Hypothesis Testing

In this chapter we have seen how to calculate r as an estimate of the relationship between two variables and how to calculate the slope (b) as a measure of the rate of change of Y as a function of X. In addition to estimating r and b, we often wish to perform a significance test on the null hypothesis that the corresponding population parameters equal 0. The fact that a value of r or b calculated from a sample is not 0 is not in itself evidence that the corresponding parameters in the population are also nonzero.

Testing the Significance of r

The most common hypothesis that we test for a sample correlation is that the correlation between X and Y in the population, denoted ρ (rho), is 0. This is a meaningful test because the null hypothesis being tested is really the hypothesis that X and Y are linearly independent. Rejection of this hypothesis leads to the conclusion that they are not independent and that there is some linear relationship between them.

It can be shown that when $\rho = 0$, for large N, r will be approximately normally distributed around 0.

A legitimate t test can be formed from the ratio

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

which is distributed as t on $N-2$ df. Returning to the example in Exhibit 9.1, $r = 0.532$ and $N = 28$. Thus,

$$t = \frac{0.532\sqrt{26}}{\sqrt{1-0.532^2}} = \frac{0.532\sqrt{26}}{\sqrt{0.717}} = 3.202$$

This value of t is significant at $\alpha = 0.05$ (two-tailed), and we can thus conclude that there is a significant relationship between SAT scores and scores on Katz' test. In other words, we can conclude that differences in SAT are associated with differences in test scores, although this does not necessarily imply a causal association.

In Chapter 7 we saw a brief mention of the F statistic, about which we will have much more to say in Chapters 11–16. You should know that any t statistic on d degrees of freedom can be squared to produce an F statistic on 1 and d degrees of freedom. Many statistical packages use the F statistic instead of t to test hypotheses. In this case you simply take the square root of that F to obtain the t statistics we are discussing here. (From Exhibit 9.1 we find an F of 10.251. The square root of this is 3.202, which agrees with the t we have just computed for this test.)

As a second example, if we go back to our data on stress and psychological symptoms in Table 9.2, and the accompanying text, we find $r = 0.506$ and $N = 107$. Thus,

$$t = \frac{0.506\sqrt{105}}{\sqrt{1-0.506^2}} = \frac{0.506\sqrt{105}}{\sqrt{0.744}} = 6.011$$

Here again we will reject H_0: $\rho = 0$. We will conclude that there is a significant relationship between stress and symptoms. Differences in stress are associated with differences in reported psychological symptoms.

Testing the Significance of b

If you think about the problem for a moment, you will realize that a test on b is equivalent to a test on r. If it is true that X and Y are related, then it must also be true that Y varies with X—that is, that the slope is nonzero. This suggests that a test on b will produce the same answer as a test on r, and we could dispense with a test for b altogether. However, since regression coefficients play an important role in multiple regression, and since in multiple regression a significant correlation does not neces-

sarily imply a significant slope for each predictor variable, the exact form of the test will be given here.

We will represent the parametric equivalent of b (the slope we would compute if we had X and Y measures on the whole population) as b^*.[10] It can be shown that b is normally distributed about b^* with a standard error approximated by[11]

$$s_b = \frac{s_{Y \cdot X}}{s_X\sqrt{N-1}}$$

Thus, if we wish to test the hypothesis that the true slope of the regression line in the population is zero (H_0: $b^* = 0$), we can simply form the ratio

$$t = \frac{b - b^*}{s_b} = \frac{b}{\dfrac{s_{Y \cdot X}}{s_X\sqrt{N-1}}} = \frac{(b)(s_X)(\sqrt{N-1})}{s_{Y \cdot X}}$$

which is distributed as t on $N - 2$ df.

For our sample data on SAT performance and test-taking ability, $b = 4.865$, $s_X = 6.73$, and $s_{Y \cdot X} = 53.127$. Thus

$$t = \frac{(4.865)(6.73)(\sqrt{27})}{53.127} = 3.202$$

which is the same answer we obtained when we tested r. Since $t_{obt} = 3.202$ and $t_{0.025}(26) = 2.056$, we will reject H_0 and conclude that our regression line has a nonzero slope. In other words, higher levels of test-taking skills are associated with higher predicted SAT scores.

From what we know about the sampling distribution of b, it is possible to set up confidence limits on b^*:

$$\text{CI}(b^*) = b \pm (t_{\alpha/2})\left[\frac{s_{Y \cdot X}}{s_X\sqrt{N-1}}\right]$$

where $t_{\alpha/2}$ is the two-tailed critical value of t on $N - 2$ df.

For our data, the 95% confidence limits are

$$\text{CI}(b^*) = 4.865 \pm 2.056\left[\frac{53.127}{6.73\sqrt{27}}\right]$$

$$= 4.865 \pm 3.123 = 1.742 \leq b^* \leq 7.988$$

Thus, the chances are 95 out of 100 that the limits 1.742 and 7.988 encompass the true value of b^*. [As you may recall, I pointed out in Chapter 7 that many statisticians would object to this last statement as phrased, because a set of specific limits (1.742

[10] Many textbooks use β instead of b^*, but that would lead to confusion with the standardized regression coefficient.

[11] There is a surprising disagreement concerning the best approximation for the standard error of b. Its denominator is variously given as $s_X\sqrt{N}$, $s_X\sqrt{N-1}$, $s_X\sqrt{N-2}$. The solution given here can be found in Neter and Wasserman (1974, pp. 58–59).

to 7.988) is not a random variable and therefore does not really have a probability associated with it. It either encloses the parameter or it does not. However, I also pointed out that as a statement of subjective probability, this declaration is a reasonable one.] Note that the confidence limits do not include 0. This is in line with the results of our t test, which rejected H_0: $b^* = 0$.

Testing the Difference between Two Independent *b*s

This test is less common than the test on a single slope, but the question that it is designed to ask is often a very meaningful question. Suppose we have two sets of data on the relationship between the amount that a person smokes and life expectancy. One set is made up of females, and the other of males. We have two separate data sets rather than one large one because we do not want our results to be contaminated by normal differences in life expectancy between males and females. Suppose further that we obtained the following data:

	Males	Females
b	–0.40	–0.20
$s_{Y \cdot X}$	2.10	2.30
s_X^2	2.50	2.80
N	101	101

It is apparent that for our data the regression line for males is steeper than the regression line for females. If this difference is significant, it means that males decrease their life expectancy more than do females for any given increment in the amount they smoke. If this were true, it would be an important finding, and we are therefore interested in testing the difference between b_1 and b_2.

The t test for differences between two independent regression coefficients is directly analogous to the test of the difference between two independent means. If H_0 is true (H_0: $b_1^* = b_2^*$), the sampling distribution of $b_1 - b_2$ is normal with a mean of 0 and a standard error of

$$s_{b_1 - b_2} = \sqrt{s_{b_1}^2 + s_{b_2}^2}$$

This means that the ratio

$$t = \frac{b_1 - b_2}{\sqrt{s_{b_1}^2 + s_{b_2}^2}}$$

is distributed as t on $N_1 + N_2 - 4$ df. We already know that the standard error of b can be estimated by

$$s_b = \frac{s_{Y \cdot X}}{s_X \sqrt{N - 1}}$$

and therefore can write

$$s_{b_1-b_2} = \sqrt{\frac{s^2_{Y\cdot X_1}}{s^2_{X_1}(N_1 - 1)} + \frac{s^2_{Y\cdot X_2}}{s^2_{X_2}(N_2 - 1)}}$$

where $s^2_{Y\cdot X_1}$ and $s^2_{Y\cdot X_2}$ are the error variances for the two samples. As was the case with means, if we assume homogeneity of error variances, we can pool these two estimates, weighting each by its degrees of freedom:

$$s^2_{Y\cdot X} = \frac{(N_1 - 2)s^2_{Y\cdot X_1} + (N_2 - 2)s^2_{Y\cdot X_2}}{N_1 + N_2 - 4}$$

For our data,

$$s^2_{Y\cdot X} = \frac{99(2.10^2) + 99(2.30^2)}{101 + 101 - 4} = 4.85$$

Substituting this pooled estimate into the equation, we obtain

$$s_{b_1-b_2} = \sqrt{\frac{s^2_{Y\cdot X_1}}{s^2_{X_1}(N_1 - 1)} + \frac{s^2_{Y\cdot X_2}}{s^2_{X_2}(N_2 - 1)}}$$

$$= \sqrt{\frac{4.85}{(2.5)(100)} + \frac{4.85}{(2.8)(100)}} = 0.192$$

Given $s_{b_1-b_2}$, we can now solve for t:

$$t = \frac{b_1 - b_2}{s_{b_1-b_2}} = \frac{(-0.40) - (-0.20)}{0.192} = -1.04$$

on 198 *df*. Since $t_{0.025}(198) = \pm 1.97$, we would fail to reject H_0 and would therefore conclude that we have no reason to doubt that life expectancy decreases as a function of smoking at the same rate for males as for females.

It is worth noting that although H_0: $b^* = 0$ is equivalent to H_0: $\rho = 0$, it does not follow that H_0: $b_1^* - b_2^* = 0$ is equivalent to H_0: $\rho_1 - \rho_2 = 0$. If you think about it for a moment, it should be apparent that two scatter diagrams could have the same regression line ($b_1^* = b_2^*$) but different degrees of scatter around that line (hence $\rho_1 \neq \rho_2$). The reverse also holds—two different regression lines could fit their respective sets of data equally well.

Testing the Difference between Two Independent *r*s

When we test the difference between two independent *r*s, a minor difficulty arises. When $\rho \neq 0$, the sampling distribution of r is not approximately normal (it becomes more and more skewed as $\rho \Rightarrow \pm 1.00$), and its standard error is not easily estimated. The same holds for the difference $r_1 - r_2$. This raises an obvious problem, because, as you can imagine, we will need to know the standard error of a difference between slopes if we are to create a t test on that difference. Fortunately, the solution was provided by R. A. Fisher.

Fisher (1921) showed that if we transform r to

$$r' = (0.5)\log_e\left|\frac{1+r}{1-r}\right|$$

then r' is approximately normally distributed around ρ' (the transformed value of ρ) with standard error

$$s_{r'} = \frac{1}{\sqrt{N-3}}$$

(Fisher labeled his statistic "z," but "r'" is often used to avoid confusion with the standard normal deviate.) Because we know the standard error, we can now test the null hypothesis that $\rho_1 - \rho_2 = 0$ by converting each r to r' and solving for

$$z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}}$$

Note that our test statistic is z rather than t, since our standard error does not rely on statistics computed from the sample (other than N) and is therefore a parameter.

Appendix r' tabulates the values of r' for different values of r, which eliminates the need to solve the equation for r'.

To take a simple example, assume that for a sample of 53 males, the correlation between number of packs of cigarettes smoked per day and life expectancy was 0.50. For a sample of 53 females, the correlation was 0.40. (These are unrealistically high values for r, but they better illustrate the effects of the transformation.) The question of interest is, Are these two coefficients significantly different, or are the differences in line with what we would expect when sampling from the same bivariate population of (X, Y) pairs?

	Males	Females
r	0.50	0.40
r'	0.549	0.424
N	53	53

$$z = \frac{0.549 - 0.424}{\sqrt{\frac{1}{53-3} + \frac{1}{53-3}}} = \frac{0.125}{\sqrt{\frac{2}{50}}} = \frac{0.125}{\frac{1}{5}} = 0.625$$

Since $z_{obt} = 0.625$ is less than $z_{0.025} = \pm 1.96$, we fail to reject H_0 and conclude that, with a two-tailed test at $\alpha = 0.05$, we have no reason to doubt that the correlation between smoking and life expectancy is the same for males as it is for females.

Testing the Hypothesis That ρ Equals Any Specified Value

Now that we have discussed the concept of r', we are in a position to test the null hypothesis that ρ is equal to any value, not just to zero. You probably can't think of many situations in which you would like to do that, and neither can I. But the ability to do so allows us to establish confidence limits on ρ, a more interesting procedure.

As we have seen, for any value of ρ, the sampling distribution of r' is approximately normally distributed around ρ' (the transformed value of ρ) with a standard error of $\frac{1}{\sqrt{N-3}}$. From this it follows that

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{N-3}}}$$

is a standard normal deviate. Thus, if we want to test the null hypothesis that a sample r of 0.30 (with $N = 103$) came from a population where $\rho = 0.50$, we proceed as follows:

$$r = 0.30 \qquad r' = 0.310$$

$$\rho = 0.50 \qquad \rho' = 0.549$$

$$N = 103 \qquad s_{r'} = 1/\sqrt{N-3} = 0.10$$

$$z = \frac{0.310 - 0.549}{0.10} = \frac{-0.239}{0.10} = -2.39$$

Since $z_{obt} = -2.39$ is more extreme than $z_{0.025} = \pm 1.96$, we reject H_0 at $\alpha = 0.05$ (two-tailed) and conclude that our sample did not come from a population where $\rho = 0.50$.

Confidence Limits on ρ

We can easily establish confidence limits on ρ by solving the previous equation for ρ instead of z. To do this, we first solve for confidence limits on ρ', and then convert ρ' to ρ.

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{N-3}}}$$

therefore

$$\pm z\sqrt{\frac{1}{N-3}} = r' - \rho'$$

and thus

$$\text{CI}(\rho') = r' \pm z_{\alpha/2}\sqrt{\frac{1}{N-3}}$$

For our stress example, $r = 0.506$ ($r' = .557$) and $N = 107$, so the 95% confidence limits are

$$\begin{aligned}\text{CI}(\rho') &= 0.557 \pm 1.96\sqrt{\frac{1}{104}} \\ &= 0.557 \pm 1.96(0.098) = 0.557 \pm 0.192 \\ &= 0.365 \le \rho' \le 0.749\end{aligned}$$

Converting from ρ' back to ρ and rounding, we obtain

$$0.350 \leq \rho \leq 0.635$$

Thus, the limits are $\rho = 0.350$ and $\rho = 0.635$. The probability is 0.95 that limits obtained in this way encompass the true value of ρ. Note that $\rho = 0$ is not included within our limits, thus offering a simultaneous test of H_0: $\rho = 0$, should we be interested in that information.

Confidence Limits versus Tests of Significance

At least in the behavioral sciences, most textbooks, courses, and published research have focused on tests of significance and paid scant attention to confidence limits. In some cases that really is probably appropriate, but in other cases it leaves the reader short.

In this chapter we have repeatedly referred to an example on stress and psychological symptoms. For the first few people who investigated this issue, it really was an important question whether there was a significant relationship between these two variables. But now that everyone believes it, a more appropriate question becomes how large the relationship is. And for that question, a suitable answer is provided by a statement such as the correlation between the two variables was 0.506, with a 95% confidence interval of $0.350 \leq \rho \leq 0.635$. (A comparable statement from the public opinion polling field would be something like $r = 0.506$ with a 95% margin of error of ± 0.15 (approx.).[12]

Testing the Difference between Two Nonindependent *r*s

Occasionally we come across a situation in which we wish to test the difference between two correlations that are not independent. (In fact, I am probably asked this question a couple of times per year.) Such a case arises when we correlate two variables at Time 1 and then again at some later point (Time 2), and we want to ask whether there has been a significant change in the correlation over time. As an example, Reilly, Drudge, Rosen, Loew, and Fischer (1985) administered two intelligence tests (the WISC-R and the McCarthy) to first-grade children, and then administered the Wide Range Achievement Test (WRAT) to those same children 2 years later. They obtained, among other findings, the following correlations:

	WRAT	WISC-R	McCarthy
WRAT	1.00	0.80	0.72
WISC-R		1.00	0.89
McCarthy			1.00

[12] I had to insert the label "approx." here because the limits, as we saw above, are not exactly symmetrical around *r*.

Note that the WISC-R and the McCarthy are highly correlated but that the WISC-R correlates somewhat more highly with the WRAT (reading) than does the McCarthy. It is of interest to ask whether this difference between the WISC-R–WRAT correlation (0.80) and the McCarthy–WRAT correlation (0.72) is significant, but to answer that question requires a test on nonindependent correlations because they both have the WRAT in common and they are based on the same sample.

When we have two correlations that are not independent—as these are not, because the tests were based on the same 26 children—we must take into account this lack of independence. Specifically, we must incorporate a term representing the degree to which the two tests are themselves correlated. Hotelling (1931) proposed the traditional solution, but a better test was devised by Williams (1959) and endorsed by Steiger (1980). This latter test takes the form

$$t = (r_{12} - r_{13}) \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}$$

where

$$|R| = (1 - r_{12}^2 - r_{13}^2 - r_{23}^2) + (2r_{12}r_{13}r_{23})$$

This ratio is distributed as t on $N - 3$ df. In this equation, r_{12} and r_{13} refer to the correlation coefficients whose difference is to be tested, and r_{23} refers to the correlation between the two predictors. $|R|$ is the determinant of the 3×3 matrix of intercorrelations, but you can calculate it as shown without knowing anything about determinants.

For our example, let

r_{12} = correlation between the WISC-R and the WRAT = 0.80

r_{13} = correlation between the McCarthy and the WRAT = 0.72

r_{23} = correlation between the WISC-R and the McCarthy = 0.89

$N = 26$

then

$$|R| = (1 - 0.80^2 - 0.72^2 - 0.89^2) + (2)(0.80)(0.72)(0.89) = 0.075$$

$$t = (0.80 - 0.72) \sqrt{\frac{(25)(1+0.89)}{2\left(\frac{25}{23}\right)(0.075) + \frac{(0.80+0.72)^2}{4}(1-0.89)^3}}$$

$$= 1.36$$

A value of $t_{obt} = 1.36$ on 23 df is not significant. Although this does not prove the argument that the tests are equally effective in predicting third-grade children's performance on the reading scale of the WRAT, because you cannot prove the null hypothesis, it is consistent with that argument and thus supports it.

9.11 The Role of Assumptions in Correlation and Regression

There is considerable confusion in the literature concerning the assumptions underlying the use of correlation and regression techniques. Much of the confusion stems from the fact that the correlation and regression models, although they lead to many of the same results, are based on different assumptions. Confusion also arises because statisticians tend to make all their assumptions at the beginning and fail to point out that some of these assumptions are not required for certain purposes.

linearity of regression

curvilinear

The major assumption that underlies both the linear regression and bivariate normal models and all our interpretations is that of **linearity of regression.** We assume that whatever the relationship between X and Y, it is a linear one—meaning that the line that best fits the data is a straight one. We will later refer to measures of **curvilinear** (nonlinear) relationships, but standard discussions of correlation and regression assume linearity unless otherwise stated. (We do occasionally fit straight lines to curvilinear data, but we do so on the assumption that the line will be sufficiently accurate for our purpose—although the standard error of prediction might be poorly estimated. There are other forms of regression besides linear regression, but we will not discuss them here.)

As mentioned earlier, whether or not we make various assumptions depends on what we wish to do. If our purpose is simply to describe data, no assumptions are necessary. The regression line and r best describe the data at hand, without the necessity of any assumptions about the population from which the data were sampled.

If our purpose is to assess the degree to which variance in Y is linearly attributable to variance in X, we again need make no assumptions. This is true because s_Y^2 and $s_{Y \cdot X}^2$ are both unbiased estimators of their corresponding parameters, independent of any underlying assumptions, and

$$\frac{SS_Y - SS_{residual}}{SS_Y}$$

is algebraically equivalent to r^2.

If we want to set confidence limits on b or Y, or if we want to test hypotheses about b^*, we will need to make the conditional assumptions of homogeneity of variance and normality in arrays of Y. The assumption of homogeneity of variance is necessary to ensure that $s_{Y \cdot X}^2$ is representative of the variance of each array, and the assumption of normality is necessary because we use the standard normal distribution.

If we want to use r to test the hypothesis that $\rho = 0$, or if we wish to establish confidence limits on ρ, we will have to assume that the (X, Y) pairs are a random sample from a bivariate normal distribution.

9.12 Factors That Affect the Correlation

The correlation coefficient can be substantially affected by characteristics of the sample. Two such characteristics are the restriction of the range (or variance) of X and/or Y and the use of heterogeneous subsamples.

The Effect of Range Restrictions

range restrictions

A common problem concerns restrictions on the range over which X and Y vary. The effect of such **range restrictions** is to alter the correlation between X and Y from what it would have been if the range had not been so restricted. Depending on the nature of the data, the correlation may either rise or fall as a result of such restriction, although most commonly r is reduced.

With the exception of very unusual circumstances, restricting the range of X will increase r only when the restriction results in eliminating some curvilinear relationship. For example, if we correlated reading ability with age, where age ran from 0 to 70 years, the data would be decidedly curvilinear (flat to about age 4, rising to about 17 years of age, and then leveling off) and the correlation, which measures *linear* relationships, would be relatively low. If however, we restricted the range of ages to 5 to 17 years, the correlation would be quite high, since we would have eliminated those values of Y that were not varying linearly as a function of X.

The more usual effect of restricting the range of X or Y is to reduce the correlation. This problem is especially pertinent in the area of test construction, since here criterion measures (Y) may be available for only the higher values of X. Consider the hypothetical data in Figure 9.6. This figure represents the relation between college GPAs and scores on some standard achievement test (such as the SAT) for a hypothetical sample of students. In the ideal world of the test constructor, all people who took the exam would then be sent on to college and earn a GPA, and the correlation between achievement test scores and GPAs would be computed. As can be seen from Figure 9.6, this correlation would be reasonably high. In the real world, however, not everyone is admitted to college. Colleges take only the more able students, whether this classification be based on achievement test scores, high-school performance, or whatever. This means that GPAs are available mainly for students who had relatively high scores on the standardized test. Suppose that this has the effect of allowing us to evaluate the relationship between X and Y for only those values of X that are greater than 400. For the data in Figure 9.6, the correlation will be relatively low, not because the test is worthless, but because the range has been restricted. In other words, when we use the entire sample of points in Figure 9.6, the correlation is 0.65. However,

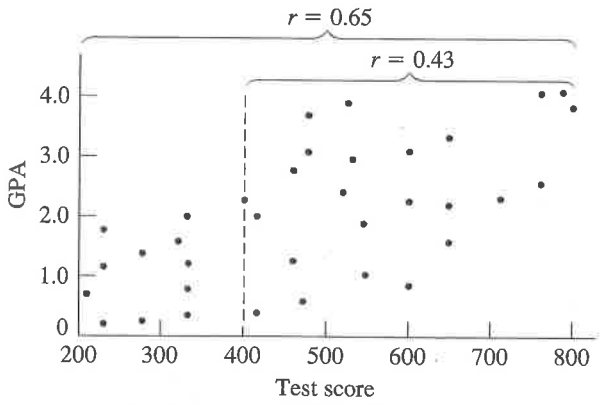


FIGURE 9.6 *Hypothetical data illustrating the effect of restricted range*

when we restrict the sample to those students having test scores of at least 400, the correlation drops to only 0.43. (This is easier to see if you cover up all data points for $X < 400$.)

We must take into account the effect of range restrictions whenever we see a correlation coefficient based on a restricted sample. The coefficient might be inappropriate for the question at hand. Essentially, what we have done is to ask how well a standardized test predicts a person's suitability for college, but we have answered that question by referring only to those people who were actually admitted to college.

The Effect of Heterogeneous Subsamples

heterogeneous subsamples

Another important consideration in evaluating the results of correlational analyses deals with **heterogeneous subsamples.** This point can be illustrated with a simple example involving the relationship between height and weight in male and female subjects. These variables may appear to have little to do with psychology, but considering the important role both variables play in the development of people's images of themselves, the example is not as far afield as you might expect. The data plotted in Figure 9.7, using Minitab, come from sample data from the Minitab manual (Ryan et al., 1985). These are actual data from ninety-two college students who were asked to report height, weight, gender, and several other variables. (Keep in mind that these are self-report data, and there may be systematic reporting biases.)

When we combine the data from both males and females, the relationship is strikingly good, with a correlation of 0.78. When you look at the data from the two genders separately, however, the correlations fall to 0.60 for males and 0.49 for females. (Males and females have been plotted using different symbols, with data from females primarily in the lower left.) The important point is that the high correlation we

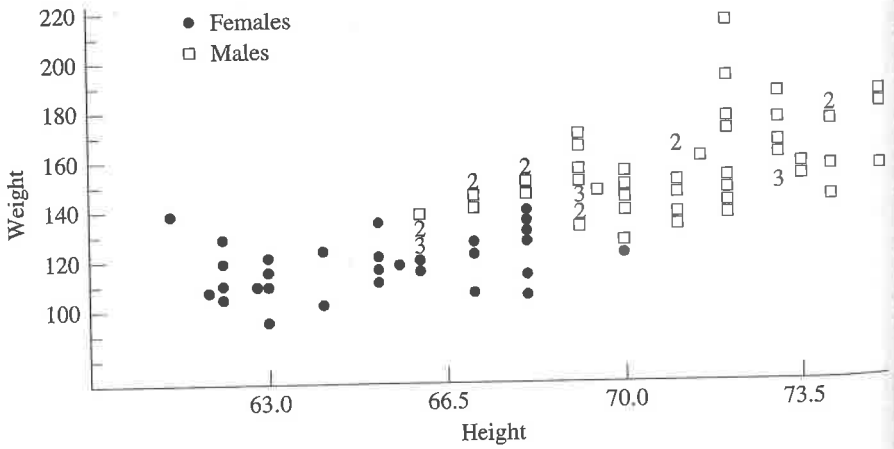


FIGURE 9.7 *Relationship between height and weight for males and females combined*

found when we combined genders is not due purely to the relation between height and weight. It is also due largely to the fact that men are, on average, taller and heavier than women. In fact, a little doodling on a sheet of paper will show that you could create artificial, and improbable, data where within each gender, weight is negatively related to height, while the relationship is positive when you collapse across gender. The point I am making here is that experimenters must be careful when they combine data from several sources. The relationship between two variables may be obscured or enhanced by the presence of a third variable. Such a finding is important in its own right.

A second example of heterogeneous subsamples that makes a similar point is the relationship between cholesterol consumption and cardiovascular disease in men and women. If you collapse across both genders, the relationship is not impressive. But when you separate the data by male and female, there is a distinct trend for cardiovascular disease to increase with increased consumption of cholesterol. This relationship is obscured in the combined data because men, regardless of cholesterol level, have an elevated level of cardiovascular disease compared to women.

9.13 Power Calculation for Pearson's *r*

Consider the problem of the individual who wishes to demonstrate a relationship between television violence and aggressive behavior. Assume that he has surmounted all the very real problems associated with designing this study and has devised a way to obtain a correlation between the two variables. He believes that the correlation coefficient in the population (ρ) is approximately 0.30. (This correlation may seem small, but it is impressive when you consider all the variables involved in aggressive behavior. This value is in line with the correlation obtained in a study by Eron, Huesmann, Lefkowitz, and Walden, 1972.) Our experimenter wants to conduct a study to find such a correlation but wants to know something about the power of his study before proceeding. Power calculations are easy to make in this situation.

As you should recall, when we calculate power we first define an effect size (**d**). We then introduce the sample size and compute δ, and finally we use δ to compute the power of our design from Appendix Power.

We begin by defining

$$\mathbf{d} = \rho_1 - \rho_0 = \rho_1 - 0 = \rho_1$$

where ρ_1 is the correlation in the population defined by H_1—in this case, 0.30. We next define

$$\delta = \mathbf{d}\sqrt{N-1} = \rho_1\sqrt{N-1}$$

For a sample of size 50,

$$\delta = 0.30\sqrt{50-1} = 2.1$$

From Appendix Power, for $\delta = 2.1$, and $\alpha = 0.05$ (two-tailed), power = 0.56.

A power coefficient of 0.56 does not please the experimenter, so he casts around for a way to increase power. He wants power = 0.80. From Appendix Power, we see that this will require $\delta = 2.8$. Therefore,

$$\delta = \rho_1\sqrt{N-1}$$

$$2.8 = 0.30\sqrt{N-1}$$

Squaring both sides yields

$$2.8^2 = 0.30^2(N-1)$$

$$\left(\frac{2.8}{0.30}\right)^2 + 1 = N = 88$$

Thus, to obtain power = 0.80, the experimenter will have to collect data on nearly 90 subjects.

Key Terms

Relationships (Introduction)
Differences (Introduction)
Correlation (Introduction)
Regression (Introduction)
Random variable (Introduction)
Fixed variable (Introduction)
Linear-regression models (Introduction)
Bivariate normal models (Introduction)
Prediction (Introduction)
Scatterplot (9.1)
Scatter diagram (9.1)
Scattergram (9.1)
Predictor (9.1)
Criterion (9.1)
Regression lines (9.1)
Correlation (r) (9.1)
Covariance (cov_{XY} or s_{XY}) (9.3)
Correlation coefficient in the population, ρ (rho) (9.4)
Adjusted correlation coefficient (r_{adj}) (9.4)
Slope (9.5)
Intercept (9.5)
Errors of prediction (9.5)
Residual (9.5)
Normal equations (9.5)
Standardized regression coefficient β (beta) (9.5)
Sum of squares (SS_Y) (9.6)
Standard error of estimate (9.6)
Residual variance (9.6)
Error variance (9.6)
Conditional distribution (9.6)
Proportional reduction in error (PRE) (9.6)
Proportional improvement in prediction (PIP) (9.6)
Array (9.7)
Homogeneity of variance in arrays (9.7)
Normality in arrays (9.7)
Conditional array (9.7)
Conditional distributions (9.7)
Marginal distribution (9.7)
Linearity of regression (9.11)
Curvilinear (9.11)
Range restrictions (9.12)
Heterogeneous subsamples (9.12)

Exercises

9.1 The State of Vermont is divided into 10 Health Planning Districts, which correspond roughly to counties. The following data for 1980 represent the percentage of births of babies under 2500 grams (Y), the fertility rate for females younger than 18