CHAPTER 7

▲▲▲▲▲▲▲▲▲▲

INTRODUCTION TO STATISTICS

Measuring Relationships for Interval Data

In Chapter 6 we were concerned generally with "relationships" between independent and dependent variables. That is, we wanted to see whether the presence or absence of a test factor affected the value of the dependent variable. This was too simple, for two reasons.

First, as noted in footnote 5 in that chapter, the test factor in field research is frequently not something that is simply present or absent; rather, it takes on a variety of values. Our task then is not just to see whether the *presence or absence* of a test factor affects the value of the dependent variable, but instead, to see whether and how the *value* of the independent variable affects the value of the dependent variable. For example, in relating education to income, we do not treat people simply as "educated" or "not educated." They have varying amounts of education and our task is to see whether the amount of education a person has affects his or her income.

Second, "relationship" cannot be dichotomized, although I treated it as a dichotomy in Chapter 6, to ease the presentation there. Two variables are not simply related or not related. Relationships vary in two ways: first, in how *strongly* the independent variable affects the dependent variable. For instance, education might have only a minor effect on income. The average income of college graduates might be $33,000 and the average income of high school dropouts might be $30,000. Or, it might have a major effect. College graduates might average $50,000 while high school dropouts average $15,000.

Relationships also vary in how *completely* the independent variable determines scores on the dependent variable. College graduates might average $50,000 income and high school dropouts $15,000, for instance, yet there might still be much variation in incomes that could not be attributed to variation in people's education. Some college graduates might make only $10,000 a year and some high school dropouts might make $80,000 or $90,000 a year, even though the *average* income of the college graduates was higher than the *average* income of the dropouts. This would indicate that although education affected incomes sharply, it was relatively incomplete as an explanation of people's incomes. Because income still varied a good deal within each level of the independent variable, there must be other things affecting income in important ways, and we often would guess incorrectly if we tried to predict a person's income solely on the basis of education. This is what it means to say that education is not a very "complete" explanation of income.

Thus variables are not simply "related" or "not related." Their relationship may be such that the independent variable has a *greater or lesser effect* on the dependent variable; and it may be such that the independent variable determines the dependent variable *more or less completely*. Generally speaking, political research is not so much concerned with whether or not two variables are related but with whether or not they have a "strong" relationship (in one or both of the senses used above). This can be seen in our examples of research design in the preceding chapter. Although for the sake of simplicity these were presented as if we were interested only in whether or not a relationship existed, it is clear that what was of interest to the investigators in each case was finding out how *strong* a relationship existed. In "Presidential Lobbying," for instance, the president was not simply concerned with whether or not he was able to influence voting for the bill, but with *how many* votes he could swing.

Our task in evaluating the results of research, then, is to measure how strong a relationship exists between the independent variable(s) and the dependent variable. The tools we need to accomplish this task are found in the field of statistics.

STATISTICS

Although modern political scientists have begun to use statistics extensively only in the past few decades, it was actually political scientists of a sort who first developed the field, for statistics originally grew out of the need to keep records for the state. The name *statistics* derives from the Latin *statisticus,* "of state affairs."

Statistics includes two main activities: statistical inference and statistical measurement (including the measurement of relationships, with which we are concerned in this chapter). Statistical inference consists of estimating how likely it is that a particular result could be due to chance; it tells us how reliable the results of our research are. I discuss inference in Chapter 9. In this chapter and in Chapter 8, I introduce some statistical techniques for measuring the strength of relationships.

THE IMPORTANCE OF LEVELS OF MEASUREMENT

In Chapter 5 we saw that we have more information about a relationship between variables if we work at a higher level of measurement than if we work at a lower level of measurement. It should not be too surprising that methods of measuring relationships between variables are different depending on the level at which the variables were measured. If we know more about a relationship, we should be able to measure a greater variety of things about it. Just as higher levels of measurement yield relatively richer information about a variable, so techniques for measuring relationships at high levels of measurement give relatively richer information about those relationships.

Basically, there are two major types of techniques: those appropriate for data measured at the interval and those suited for lower-level measurements. Recall that we mentioned two ways to measure the "strength" of a relationship between two variables: (1) by how great a difference the independent variable makes in the dependent variable, that is, how greatly values of the dependent variable differ, given varying scores on the independent variable; or (2) by how completely the dependent variable is determined by the independent variable, that is, how *complete* an explanation of the dependent variable is provided by the independent variable. I shall call the first way of measuring a relationship *effect-descriptive,* and the second *correlational.* The critical difference between working with interval-measured data and working with data measured at a lower level is that effect-descriptive measurement can apply only to interval-scale data. Correlational measurement of one sort or another can apply to data measured at any level.

Ordinal and nominal measurement techniques do not tell us how great a difference in the dependent variable is produced by a given difference in the independent variable, although this is precisely what is required to measure the relationship in an effect-descriptive way. The whole point of nominal and ordinal measurement is that in neither do we have available a unit by which to measure the difference between two values of a variable. This means that we cannot measure how great a difference is induced in the dependent variable by a change in the independent variable. If we are using an ordinal-scale variable we know whether one value is higher than another, but we do not know how much higher it is. If we are using a nominal-scale variable, of course, all we know is whether the two values are distinct.

This becomes a particularly important distinction in political research, because under most circumstances, effect-descriptive ways of measuring the strength of a relationship are more useful than correlational ways. I demonstrate this in the next few sections.

WORKING WITH INTERVAL DATA

Regression Analysis

A convenient way to summarize data on two interval-scale variables so that we can easily see what is going on in the relationship between them is to plot all the observations on a *scattergram,* as in Figure 7–1. Each dot in the scattergram represents one observation (a person, state, or country, for example), placed

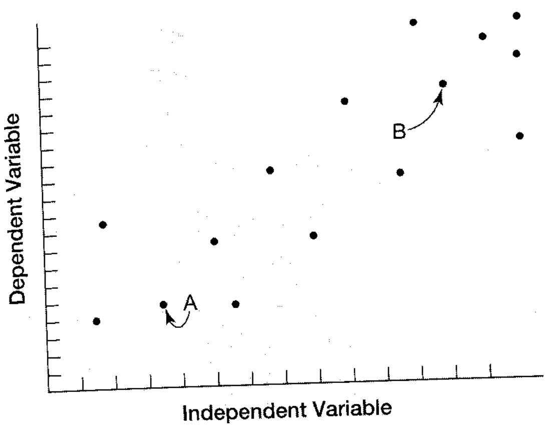


Figure 7–1 **Scattergram**

on the graph according to its scores on the two variables. For instance, dot A represents an observation that combines scores of 3.5 on the independent variable and 5 on the dependent variable. Dot B represents an observation with scores of 12 on the independent variable and 17 on the dependent variable.

By looking at the pattern formed by the dots, we can tell a good deal about the relationship between the two variables. For instance, in Figure 7–1, we note that there are few dots in the lower-right and upper-left corners of the graph. This means that high scores on the dependent variable tend to coincide with high scores on the independent variable, and low scores on the dependent variable tend to coincide with low scores on the independent variable. Thus we know that the two variables are positively related. Furthermore, this relationship appears to be approximately linear. (See the discussion of "linear relationships" in the box on page 67.)

We have done two things so far. We have observed which way the dependent variable moves with changes in the independent variable, and we have observed that it moves at a steady rate at all values of the independent variable (a linear relationship) rather than at changing rates (a nonlinear relationship). These are both part of an effect-descriptive measurement of the relationship.

The scattergrams in Figure 7–2 illustrate various other patterns we might have observed. Graph A shows a nonlinear relationship (the dependent variable increases faster with increases in the independent variable if the independent variable has a high value). Graph B shows a linear relationship in which the dependent variable increases more gradually than in the graph in Figure 7–1. Graph C shows a negative linear relationship in which the dependent variable decreases as the independent variable increases. Graph D shows a pattern in which there is no relationship.

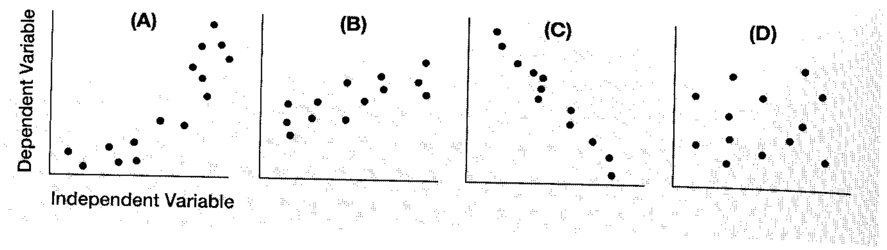


Figure 7–2 **Assorted Scattergrams**

Although the scattergram tells us a good deal about a relationship, it can be unwieldy to work with. It is not uncommon in a research report to discuss 30 or 40 separate relationships. It would be painful to read such a paper if each relationship were presented in the form of a scattergram. What is more, comparing two scattergrams gives us only an approximate idea of the differences between two relationships. Comparing the graphs in Figures 7–1 and 7–2(B), we can say that in the first graph the independent variable causes greater shifts in the dependent variable than in the second, but we cannot say precisely how much greater the shifts are. If the differences were more subtle, or if we were comparing several relationships, the job would become well-nigh impossible.

Finally and most important, we often measure the strength of a relationship between two variables while holding a third variable constant. (See the discussion of this topic on pages 90–92.) To do this using scattergrams may be extremely cumbersome.

For all of these reasons, it is useful to devise a precise numerical measure to summarize the relevant characteristics of a relationship shown in a scattergram. The measure commonly used to summarize the effect-descriptive characteristics of a scattergram is the *regression coefficient.*

The linear regression coefficient is derived in the following way. First, the pattern in the dots of a scattergram is summed up by the single line that best approximates it. For a linear relationship, the mathematically best procedure is to choose that line that minimizes the squared differences between observed values of the dependent variable and its idealized values as given by the simplifying line. This is illustrated in Figure 7–3, where a simplifying line has been drawn through a scattergram with observations on seven hypothetical countries to summarize the pattern across the countries. It has been drawn so as to minimize the squared differences between each of the observed points, such as A, and the point B at which a country having A's score on the independent variable would be expected to fall on the idealized simplifying line.

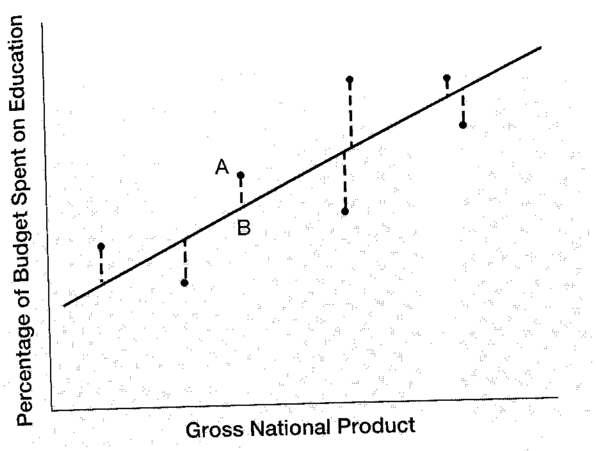


Figure 7–3 **The Regression Line**

The simplifying line may be thought of as a rule for predicting scores on the dependent variable from scores on the independent variables. Its usefulness as a predictor depends on keeping the average squared value of "deviant" scores on the dependent variable as low as possible. A single summarizing line can be described more easily than a pattern of dots. In particular, a straight line such as this can be fully described by the equation

$$y = a + bx$$

where *y* is the predicted value of the dependent variable, *x* is the value of the independent variable, and *a* and *b* are numbers relating *y* to *x*. The number *a*, the expected value of *y* when *x* equals zero, is called the *intercept* of the regression equation; it is the value of *y* where the regression line crosses the *y* axis, that is, where *x* equals zero (see Figure 7–4). The number *b*, or the *slope* of the regression equation, shows by how many units *y* increases as *x* increases one unit. (If *b* is negative, *y* decreases as *x* increases; there is a negative relationship between the variables.) In other words, to find the predicted value of the dependent variable for any specified value of the independent variable, you must add *a* (the predicted value of *y* when *x* equals zero) to *b* times the number of units by which *x* exceeds zero.

The slope, often simply called the *regression coefficient,* is the most valuable part of this equation for most purposes in the social sciences. By telling how great a shift we can expect in the dependent variable if the independent variable shifts by one unit, the slope provides a single, precise summary measure of how great an impact the independent variable has on the dependent variable. Let's assume, for instance, that the relationship between income and electoral participation is linear and can be summarized by the regression equation

$$\text{percent voting} = 50.5 + \tfrac{1}{2}\ (\text{income in thousands of dollars})$$

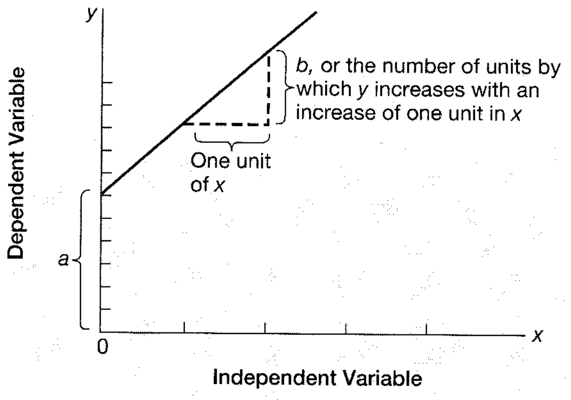

y
Dependent Variable
b, or the number of units by which y increases with an increase of one unit in x
One unit of x
a
0
x
Independent Variable

Figure 7–4 **The Regression Equation**
The equation of this line is $y = 6 + 3x$. The predicted value of y when x is 4, for instance, is $6 + (4 \times 3)$, or 18.

This means that with every additional thousand dollars of income, ½ percent more potential voters vote. For example, if income is $2,000, the predicted percent voting equals 50.5 + (½ × 2), or 51.5. If income is $3,000, the predicted percent voting equals 50.5 + (½ × 3), or 52.0.

Remember, however, that even though we work with neat, impersonal numbers, we do not escape the scholar's obligation to think. If we have guessed the direction of causation between *x* and *y* incorrectly, plugging in our data and getting numbers out will not make the results valid. If *y* causes *x* rather than vice versa, the formulas will still give us an *a* and a *b*, but the shift of one unit in *x* in the real world will *not* be followed by a shift of *b* units in *y*. Thus the arguments made in Chapter 6 apply even when we work with simple numbers like these.

The problem of comparing units. It may be seen from an examination of the concept *slope* (the number of units by which *y* changes with a change of one unit in *x*) that the slope has meaning only with regard to the units in which *x* and *y* are measured. For example, if there is a regression coefficient of –10.5 for nations' diplomatic involvement with the United States (measured by the number or magnitude of exchanges per year) predicted from their distance from the United States measured in thousands of miles, there would be a regression coefficient of –0.0105 for the same dependent variable if distance were measured in miles. That is, if diplomatic involvement could be expected to decrease by 10.5 with every thousand miles of distance from the United States, it would be expected to decrease by 0.0105 with every mile of distance.

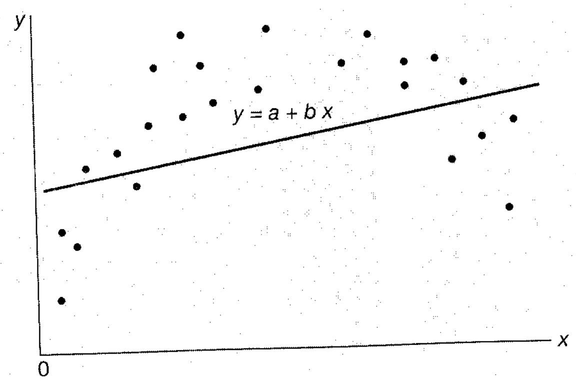
If we are working with just two variables, this poses no real difficulty. But often we may be interested in comparing the effects of two or more independent variables on a particular dependent variable. If the two independent variables are measured in different sorts of units, this can be difficult. Continuing with our example, we might want to know which variable—nations' distance from the United States or the volume of their trade with the United States—has a greater impact on their diplomatic interaction with the United States. If the regression coefficient for volume of trade, measured in millions of dollars, is +0.024, how can we tell whether it has a greater impact on diplomatic interaction than does geographic distance, with its slope of –10.5? The units—thousands of miles and millions of dollars—are not comparable; therefore, the coefficients based on those units are not comparable either.

The technique of *standardized regression coefficients,* or *beta weights,* has been developed to wash out the effect of varying units in regression analysis. To present it is beyond the technical level of this book, but you should at least be aware of its existence. It is no cure-all; essentially it transforms regression coefficients into a form of correlation coefficient (see pp. 106–110), which are independent of unit but involve problems of their own. Personally, I do not recommend the technique of standardized regression for most purposes, but it has its adherents. A good presentation is that of Blalock (1979, pp. 477–482).

Checking for linearity. We must be careful to make certain in using linear regression analysis that the data do fit a more or less linear pattern. Otherwise, the regression equation, will not summarize the pattern in the data, but will distort it. Figure 7–5 shows a linear regression equation passed through the scattergram of a nonlinear relationship. This regression line fits the data very badly and is not a useful summary of the relationship. You should always check your data before using linear regression analysis. The best way to do this is simply to draw a scattergram (or better yet, let the computer do it for you) and see whether it looks linear.

Many relationships in political science do turn out to be linear. But if the relationship you are investigating turns out to be nonlinear, that is no reason

Figure 7–5 **Linear Regression on a Nonlinear Relationship**



give up analyzing it. It merely means that the relationship is more complex an you anticipated—and probably more interesting. A nonlinear regression uation may be found to fit the pattern fairly well.

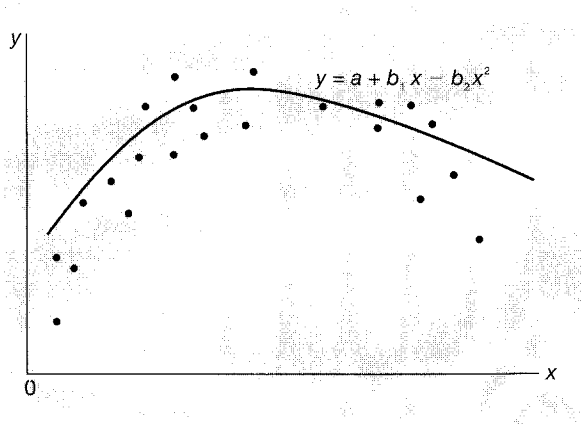
In Figure 7–6, a nonlinear regression equation $y = a + b_1x - b_2x^2$ has been ssed through the scattergram from Figure 7–5. It summarizes the pattern in e data more accurately. In this case, two coefficients, b_1 and b_2, are required express the impact that a change in x will have on y, since that change is not e same at all values of x. We can see that y increases with x but decreases with e square of x. The regression equation provides a handy summary descrip›n of the effect, now a bit more complicated, that x has on y.

Formulas are available to calculate equations for regression lines satisfyg the least-squares criteria. This is particularly true for linear regression; the rmulas for a and b are found in every standard statistics text, including the ıes cited at the end of this chapter. But there are no set "formulas" for nonıear regression equations, for there is an infinite variety of nonlinear equa›ns that you might fit to any set of data. It usually is necessary to play around th alternative nonlinear equations for a while. But these, too, can be worked ıt readily enough.

To the extent that your research is based on a well-thought-out theory, is will help you to design an appropriate nonlinear equation. For instance, if ur theory predicts that the dependent variable will always increase with ineases in the independent variable, but at a constantly diminishing rate (a iminishing marginal returns" model), the equation depicted in Figure 7–6 ›uld be inappropriate because it must inevitably reverse direction at some ›int. An equation such as $y = a + b \log x$, as depicted in Figure 7–7, would be ›propriate.

One important warning: Remember that the presentation I have made ›re is only a broad, introductory overview. Competence in using measures

ȷure 7–6 **Nonlinear Regression Equation**



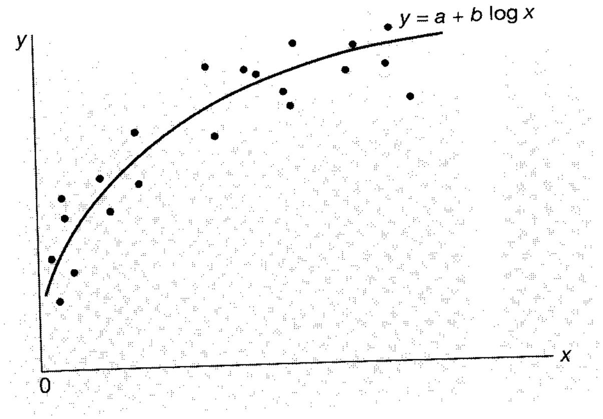


Figure 7–7 **Model of Diminishing Marginal Returns**

like those presented here requires more thorough training than is within the scope of this book.

Examining the residuals. Any regression line is actually a reflection of the stage that we have reached in developing a theory. Because our theory anticipates a relationship between two variables, we measure the relationship between them by calculating the regression equation for the line that best summarizes the pattern of the relationship. In this sense, regression analysis is an expression of what we already think about the subject. It may surprise us; where we expected a relationship there may be none, or it may be nonlinear, and so on. But it is an expression of what we already have been thinking about the subject.

However, the regression line can serve a further important function; by pointing out important independent variables that had not yet occurred to us, it can help refine our theory. Looking back at Figure 7–3, note that the regression line does not provide perfect prediction of the values on the dependent variable for the cases in the scattergram. This means that there is still unexplained variation in the values of the dependent variable. Some of the observations are higher on the dependent variable than we would expect from their value on the independent variable, and some are lower. *Something else, beyond the independent variable, is affecting the dependent variable.*

This difference between the observed value and the predicted value is called the *residual*. Examining these residuals points out to us those cases in which the "something else" has the effect of raising (or, conversely, lowering) the dependent variable. In Figure 7–3, for instance, a case such as A is one in which the effect of the "something else" is to raise the value of the dependent variable; the actual value for case A is higher than the value B, which would have been predicted from the regression line.

Now, once the cases have been sorted out in this way, we may notice that cases with similar residuals have some additional characteristic in common; this may then suggest an additional variable that may be brought into our theory. Consider Figure 7–3. On examining the residuals in that figure, you might notice that all of the countries for which educational spending was higher than predicted were not democracies, and all of the countries for which it was unexpectedly low were democracies. In this way you might have discovered the identity of the "something else," beyond GNP, that acts as an independent variable. Notice also that it would have been difficult to identify the presence of another factor if you had not first regressed educational spending on GNP. The richest democracies in Figure 7–3 show a higher rate of educational spending than do the poorest nondemocracies, so that it might not have been at all obvious that "democracy" was a variable you should use to explain levels of educational spending.

The technique of examining residuals is illustrated in Figure 7–8, adapted from V. O. Key, Jr.'s *Southern Politics* (1950, p. 48). Key wanted to measure the impact of factions in Alabama primaries, so he related counties' votes for Folsom, a progressive candidate in the 1946 gubernatorial primary, to their votes for Sparkman, a progressive candidate in the 1946 senatorial primary. He found a moderately strong relationship between the two. Because this meant that counties tended to lean the same way in both elections, it indicated the presence of conservative and progressive factions structuring the vote, as Key had expected. Had such factions not existed, there would have been no particular reason to expect a county to vote in much the same way in the two primaries. But the relationship was not very tight. Many counties, such as the one that gave 90 percent of its vote to Folsom but only 45 percent to Sparkman, voted far differently from what one would have expected simply on the basis of conservative and progressive factions. This indicated the presence of other variables, which were causing additional variations in the vote.

By examining the residuals around the regression line, Key got some idea of what those variables might be. In this case it turned out that the residuals could be explained in part by the effect of "friends and neighbors." Counties in Folsom's home part of the state voted for him more enthusiastically than would have been expected on the basis of their vote for Sparkman. Counties in Sparkman's home part of the state voted less enthusiastically for Folsom than would have been expected from their vote for Sparkman (which presumably was high because he was a local boy). Similarly, the "home" counties of Folsom's opponent went less heavily for Folsom than would have been expected on the basis of their vote for Sparkman. This pointed out to Key the importance of local solidarity, one of the major forces retarding the development of stable statewide factions in Alabama politics at that time. Much of the looseness in the relationship between the votes for two candidates of the same faction was shown to be a result of people's tendency to vote for a candidate on the basis of where he came from in the state rather than the faction with which he was identified.

Users of regression analysis in political science far too rarely go on to the creative and exploratory labor of examining the residuals to see what addi-

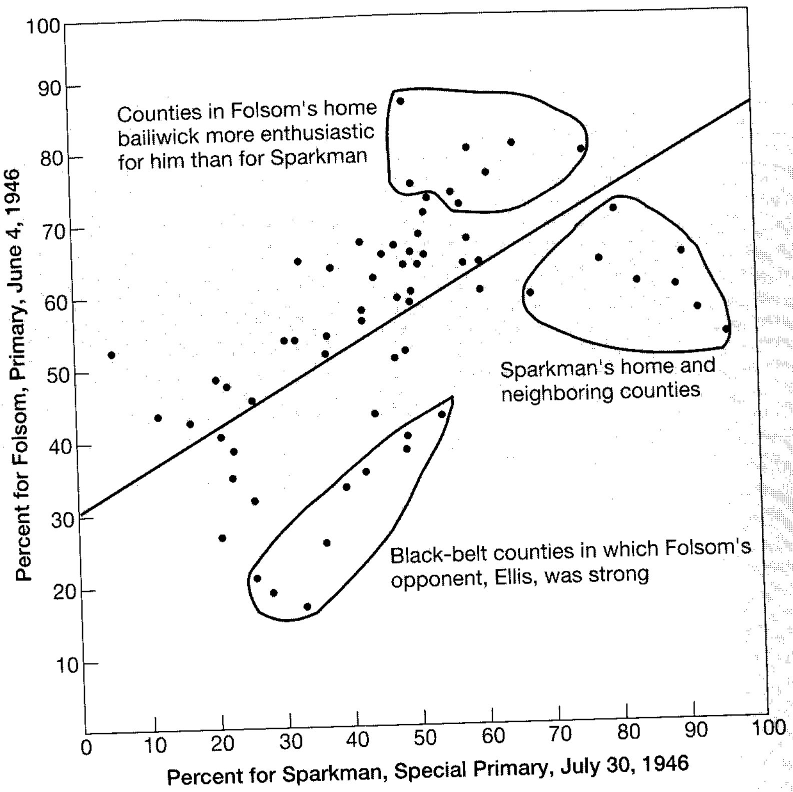


Figure 7–8 **Example of Residual Analysis**

Source: V. O. Key, Jr., *Southern Politics* (New York: Alfred A. Knopf, Inc., 1950), p. 48.

tional variables affect the dependent variable. Usually, the spread of dots around the regression line is treated as an act of God, or as a measure of the basic uncertainty of human affairs. On the contrary, it is a trove in which new variables lie waiting to be discovered. I suspect the reason most of us do not go on to examine this trove is that we have developed a proprietorial sense toward our theories before we ever get to the point of testing them. There is a certain completeness about one's own theory, and it does not occur to us to use our theory as a "mere" starting point in the search for explanations.

Correlation Analysis

At the beginning of this chapter I pointed out that there are two ways t measure the strength of a relationship: by measuring how much difference th independent variable makes in the dependent variable, and by measuring ho completely the independent variable determines the dependent variable. For interval-scale data, the regression coefficient accomplishes the first of these; the correlation coefficient accomplishes the second.

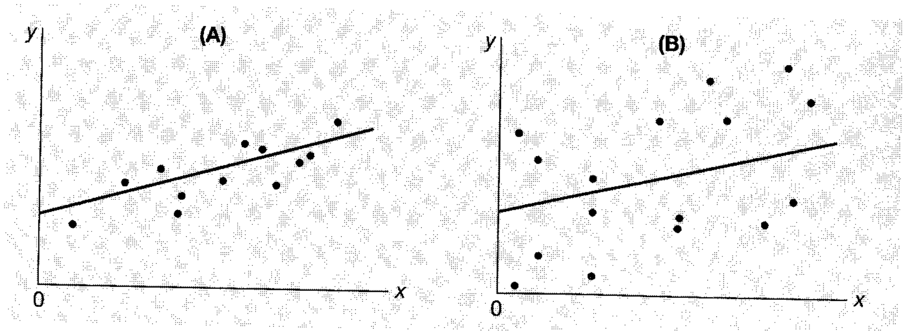
Consider the graphs in Figure 7–9. Both relationships can be summarized by the same regression line, but the value of the dependent variable in graph B is less closely determined by the independent variable than in graph A. A change in the independent variable tends to produce the same change in the dependent variable, on the average, in both graphs. But this tendency is weaker, and more likely to be disturbed by "other factors," in graph B; in other words, the residuals tend to be larger in graph B than in graph A. In one sense, then, the relationship in graph B is weaker than that in graph A. The dependent variable is less a result of the independent variable, compared to "other factors" (the unknown things that cause the residuals to exist), in B than in A.

The *product-moment correlation coefficient*, *r*, measures how widely such a body of data spreads around a regression line. This coefficient compares a set of data with ideal models of a perfect relationship and a perfect lack of relationship, and assigns to the relationship a score ranging in absolute value from zero to 1, depending on how closely the data approximate a perfect relationship. The two extreme models are illustrated in Figure 7–10.

In graph A of Figure 7–10, the data all fall on a straight line through the scattergram. A regression line passed through them would leave no residual variation at all in the dependent variable. Thus the independent variable determines the dependent variable completely. The correlation coefficient for this type has an absolute value of 1.

In graph B, on the other hand, values of the dependent variable are combined randomly with values of the independent variable, so that any given value of the dependent variable is as likely to coincide with a low value on the independent variable as with a high one. Thus there is no pattern to the relationship. This indicates that the independent variable has no effect on the dependent variable. The correlation coefficient for this type has absolute value zero.

Figure 7–9 **Two Correlations**



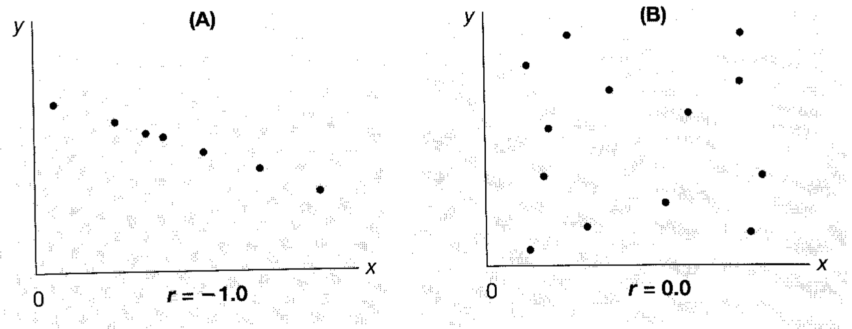


Figure 7–10 **Extreme Models of Correlation**

Most relationships, of course, fall somewhere between these two extremes; the closer a relationship approaches the situation in graph A, the higher its correlation coefficient will be in absolute value. Thus the correlation coefficient provides a measure by which the strengths of various relationships can be compared, in the *correlational* sense of "strength of relationship."

I have referred here only to the absolute value of the correlation coefficient. In addition to showing how closely a relationship approaches the situation in graph A, the correlation coefficient indicates by its sign whether the relationship is positive or negative. The coefficient ranges from –1.0 (a perfect negative relationship, such as the one in graph A of Figure 7–10) through 0.0 (graph B) to +1.0 (a perfect positive relationship, similar to the one in graph A but tilted up). In a positive relationship, increases in the dependent variable produce increases in the independent variable; in a negative relationship, increases in the independent variable produce decreases in the dependent variable.

Interpreting the correlation coefficient. Although it is clear enough what correlation coefficients of –1, 0, or +1 mean, there is no easy way of interpreting what coefficients between these values mean. It is true that the higher the absolute value of the coefficient, the closer it approaches the model in graph A of Figure 7–10, so that if we wish to compare two different relationships, we can say which is stronger. But it is not easy to see what the difference between them means. It is *not* true, for instance, that the difference between $r = .8$ and $r = .6$ is the same as the difference between $r = .4$ and $r = .2$. And it is also not true that $r = -.6$ is twice as strong as $r = -.3$. This is reminiscent of the difference between ordinal and interval measurement: We know that the higher the absolute value of *r*, the stronger the relationship; but we do not know *how much* stronger one relationship is than another.

Fortunately, the *square* of the correlation coefficient (sometimes called the *coefficient of determination*, but more often just r^2) does have a usable interpretation at all values of r. Before we can consider this, I must first introduce the concept of *variance*.

Variance. The variance of a variable is the average squared deviation of values of that variable from their own mean.[1] For instance, if there are just three cases, with scores of –1, 4, and 5 for a variable, their mean is $(-1 + 4 + 5)/3 = 8/3$, and their variance is

$$\frac{(-1-{}^{8}\!/\!_{3})^2 + (4-{}^{8}\!/\!_{3})^2 + (5-{}^{8}\!/\!_{3})^2}{3}$$

$$= \frac{(-{}^{11}\!/\!_{3})^2 + ({}^{4}\!/\!_{3})^2 + ({}^{7}\!/\!_{3})^2}{3}$$

$$= \frac{{}^{121}\!/\!_{9} + {}^{16}\!/\!_{9} + {}^{49}\!/\!_{9}}{3} = 6.89$$

The formula for the variance of any variable x is

$$\text{variance}_x = \frac{\Sigma(x-\bar{x})^2}{N}$$

where $\bar{x}$ is the mean of x and N is the number of observations we wish to average. The Σ sign simply means that for all the observations, we are to calculate $(x-\bar{x})^2$ and then add these results together. The expression $\Sigma(x-\bar{x})^2$ is equivalent to writing $(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_N-\bar{x})^2$, where x_1 is the observation for the first case, x_2 the observation for the second case, and so on.

The variance is a measure of how widely the observed values of a variable vary among themselves. If they do not vary at all but each has the same value, the variance will be zero. This is true because each value will equal the mean, and thus the sum of the squared deviations from the mean will equal zero. The more the values vary among themselves, the further each will be from the mean of them all, and consequently, the greater the sum of squared deviations from the mean. Thus, the more that values vary among themselves, the higher their variance will be.

The variance of the dependent variable y can be depicted in a scattergram by drawing a horizontal line at $y = \bar{y}$, and drawing in the residuals from this line, as in Figure 7–11. The average squared value of the residuals, because it is the average squared deviation of the values of y from their own mean, is the variance of y.

One way to view our goal in theoretical social science research is to note that our task is usually to account for the variance in a dependent variable. It is the variance in something that puzzles us and challenges us to produce an explanation. Why is it that some people make more money than other people? That some nations are frequently involved in wars and others are not? That some members of Congress vote for a bill and others oppose it? That some people are more politically alienated than others? All of these questions simply ask: To what can the variance in this variable be attributed—*why does this variable vary?*

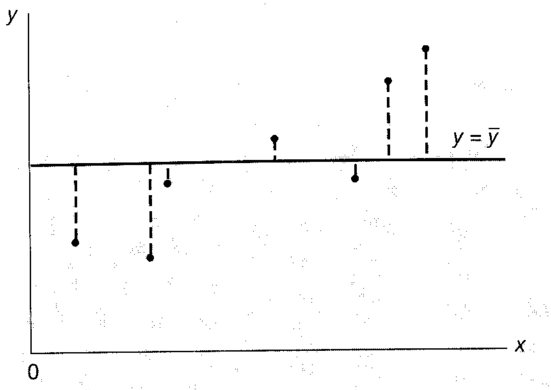


Figure 7–11 **The Variance of *y***

Comparing Figures 7–3 and 7–11, we should see at least a superficial similarity between the residuals around the least-squares line and the variance of the dependent variable. The least-squares line is a line passed through the scattergram in any direction such that the squared deviations of values of the dependent variable from that line are minimized. The line $y = \bar{y}$ is a *horizontal* line passed through the data in the scattergram, and inspection will suggest what is, in fact, mathematically true: This line is the one *horizontal* line that minimizes the squared deviations of values of the dependent variable from itself (see Blalock, 1979, p. 58). That is, any other horizontal line passed through the data would yield a greater sum of squared deviations in the dependent variable.

This similarity suggests that the squared deviations around the regression line may be treated as the variance of the dependent variable around the values predicted for it by the regression equation. As already noted, this is the variance in the dependent variable that is still left unaccounted for after the effect of the independent variable has been estimated.

Thus we have two variances for the dependent variable: its variance around its own mean (the "total variance") and its variance around the regression line ("variance left unexplained by the independent variable"). To the extent that the dependent variable is determined by the independent variable, this unexplained variance will be small compared to the total variance. If the dependent variable can be predicted perfectly from the independent

[1] The mean of any variable is its arithmetic mean, or average: the sum of all the values divided by the number of cases.

variable, as in Figure 7–10(A), the unexplained variance will be zero. If the dependent variable is unrelated to the independent variable, as in Figure 7–10(B), the regression line will be horizontal, indicating that the same value of the dependent variable is predicted at all values of the independent variable; inasmuch as the line $y = \bar{y}$ is the horizontal line that minimizes squared deviations around itself, the regression line will equal the line $y = \bar{y}$ in this case. Thus the unexplained variance will equal the total variance.

Dividing the unexplained variance by the total variance tells us what proportion of the total variance is left after we have allowed the independent variable to explain as much as it can explain. As it happens, r^2 equals 1 minus this proportion, or

$$1 - \frac{\text{unexplained variance}}{\text{total variance}}$$

hat is, the proportion of the total variance in the dependent variable that can be ascribed to the independent variable.[2] The explained variance, therefore, gives us a useful interpretation of r at all values. If you read that an author has ound a correlation of –.30 between two variables, you should mentally square he correlation and interpret that statement: "the two variables are negatively elated, and 9 percent of the variance in one is due to the other."

Another helpful way to look at this interpretation is to think in terms of prediction. Operating without any knowledge of the independent variable, our best strategy in trying to predict values of the dependent variable for particular cases would be to guess that the value in any given case is the mean. We would be less wrong more of the time than with any other guess we could make.[3] If we now add knowledge of the independent variable, our best guess becomes the value predicted from the regression equation. The magnitude of he mistakes in each case is now represented by squared deviations around the predictions. The value r^2 measures the proportion by which we have reduced ur mistakes in predicting the dependent variable by introducing knowledge f the independent variable.

Correlation and Regression Compared

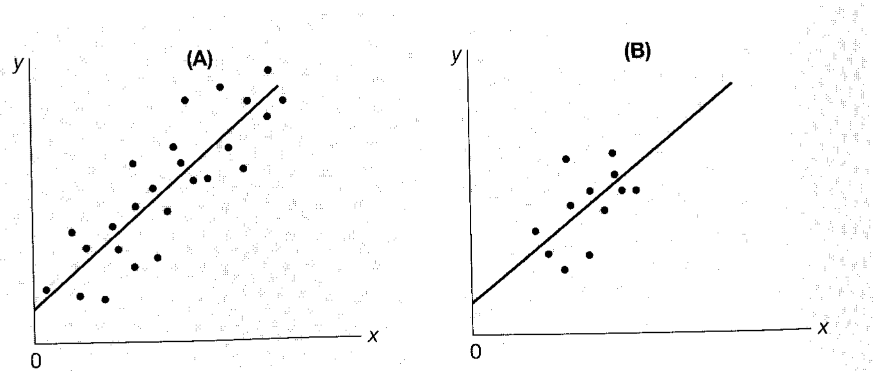
Our discussion so far leaves us with the question, "Is correlation or reression analysis the better way to measure the strength of a relationship?" Obviously, the answer must be, "Sometimes one is, sometimes the other." The key o deciding when to use each measure lies in the fact that the correlation coefficient reflects the variability of the independent variable directly, whereas the regression coefficient does not.

Consider the two scattergrams in Figure 7–12. The scattergram from graph A has been reproduced in graph B, except that all observations for which the independent variable is less than 2 or greater than 4 have been eliminated. The effect is to reduce the variability of the independent variable while leaving the basic relationship between the independent and dependent variables unchanged. *Under these circumstances, the regression coefficient in* B *will be approximately the same as that in* A, *but the correlation coefficient will be sharply lowered in* B.

Let us see why this should be so. The regression line which minimized squared deviations in the full set of data in graph A should continue to minimize squared deviations in the partial set of data in graph B. Therefore, we would expect the regression coefficient to be about the same in both graphs. On the other hand, because the variability of the independent variable has been reduced in graph B, the extent of its possible effects on the dependent variable are reduced. Relative to other causes of the dependent variable, which are just as free to operate as they were in graph A, the importance of the independent variable as a cause of the dependent variable must decline. But this is simply to say that the proportion of the total variance which is attributable to the independent variable declines. In other words, r^2 is reduced.

This becomes a matter of considerable importance in field research, because generally the variability of independent variables is beyond the investigator's control. For instance, a researcher might be interested in knowing whether sex or race had more to do with whether a person voted. The researcher might use census tract data on a large city, correlating percent black with percent voting and percent male with percent voting. But the distribution of people in most cities is such that the percent black would vary greatly while the percent male would not. (Blacks are concentrated highly in some tracts

Figure 7–12 **Regression and Correlation, with the Independent Variable Attenuated**



[2] For a good presentation of this interpretation, including the proof that

$$r^2 = 1 - \frac{\text{unexplained variance}}{\text{total variance}}$$

e Blalock (1979, pp. 405–409).

[3] At least this is true if we think of "mistakes" as squared deviations from the true value.

and almost absent in others; men are spread more or less evenly across all the tracts.)

The fact that percent male scarcely varies from one census tract to another guarantees that this researcher would find practically no correlation between percent male and percent voting. As indicated in the hypothetical scattergram in Figure 7–13, there is near zero variance in percent male; hence very little of the variance in percent voting can be due to it. On the other hand, if residential patterns were such that percent male could vary as much as percent black does, it might be that gender would show up as a major determinant of voter turnout.[4] There is limited usefulness to a measure that would have us conclude from this that race is a more important cause of participation than sex.

In a similar example, it frequently is asserted that academic departments should drop the Graduate Record Examination test scores as a factor in graduate admission. Among other—and better—reasons, this argument sometimes is based on the fact that in looking over the records of students in a department, it frequently turns out that how they did once they were admitted to the department is almost totally unrelated to their graduate record examination (GRE) scores. On the face of it, this is startling. But it ignores the fact that because GRE scores were an important factor in the initial selection of applicants, the variability among GRE scores in the department is low. In other words, if the department cuts off applicants at the 80th percentile, it should not expect its students' scores to be related to how they do in the department. By choosing them on that basis, it has ensured that they all have virtually the same, or very close, scores. If the department stopped using the exam scores as a basis for admission and developed a graduate student body with a greater variance in scores, it would not necessarily still be true that scores would be unrelated to performance.

The problem with using correlation coefficients in both of these cases is that while the coefficient is affected by the particular variance of the independent variable in the data at hand, the investigator clearly intends to extrapolate from these particular data to a more general case. In the census tract study, the researcher wants to make a statement about the impact of race or sex on voting, regardless of how people are located in the city. In the GRE study, the researcher wants to make a judgment about how to treat new applicants (among whom the variability of GRE scores would be higher) on the basis of the relationship between GRE scores and grades among the students currently in the department.

A good rule is that in any situation in which you wish to extrapolate from a particular set of data to a more general case, you should use regression coefficients. Because theoretical research almost always is interested in generalizing, regression analysis usually will be superior to correlation analysis. (This advice holds despite the fact, which I mentioned earlier, that in regression analysis there is always some difficulty in handling varying units.)

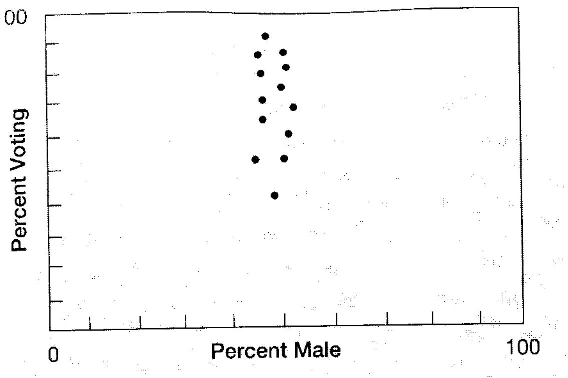


Figure 7–13 **Predicting Percent Voting from Percent Male**

There are circumstances, however, in which you may not intend to generalize, but only to describe a particular situation. For instance, someone might want to describe how a particular Congress, say the 80th or the 92nd, operated. It would then be appropriate to note that in that particular Congress, there was a negligible correlation between members' race and their votes on various bills. This would help establish what were the important factors influencing outcomes *in that Congress.* However, this would not deny the existence of a relationship between race and voting record for Congresses in general.

The Problem of Measurement Error

Our understanding of relationships between variables is intimately bound up with the extent to which we have been able to measure the variables validly. To the extent that there is *nonrandom error* in our variables, we are simply unable to state accurately what the relationship between those variables is. This should be clear from our earlier discussion of measurement error in Chapter 4.

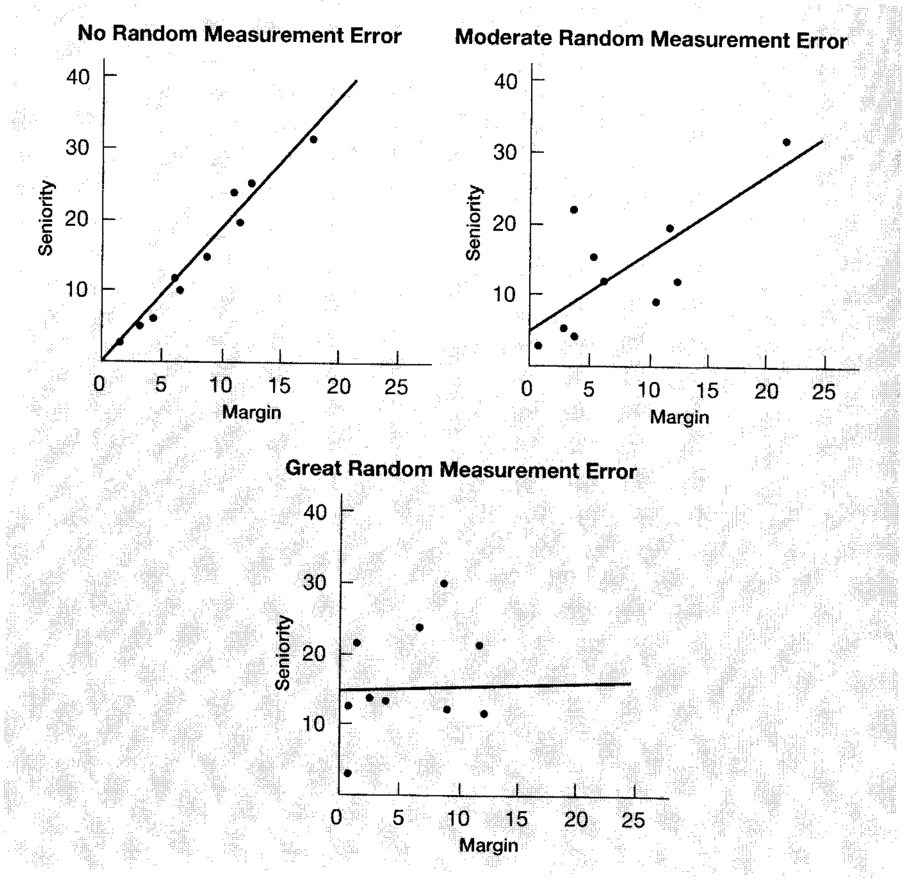
Random measurement error also distorts the relationship between variables. Refer back to Table 4–1 (p. 50). The relationship between the two variables is increasingly attenuated from left to right, as it might appear under conditions of increasing levels of random measurement error. The scattergrams for the

[4] This example also incorporates a major statistical problem in some correlation and regression analyses, the *ecological fallacy.* This can occur when data on aggregate units (such as percent black, median income, and so on, for census tracts, counties, or states) are used to infer how variables are related among individuals living in those aggregate units. See Robinson (1950), Stokes (1969), and Achen and Shively (1995).

three sets of data drawn from Table 4–1 are presented in Figure 7–14. As progressively greater amounts of random measurement error are present in two variables, the true form of the relationship between the variables is more and more lost to us. Not only is it lost to us, but it is systematically distorted. To the extent that there is random error in two variables, they will appear to us to be unrelated. Thus there is a real danger that the only relationships we will be able to perceive are those between variables that are easy to measure accurately, a possibility that bodes ill for social science.

If we are willing to make some assumptions regarding the nature of the random error, it is possible to reconstruct the true relationship between two variables, despite a considerable amount of random error. A good example of this technique is seen in Bartels (1993). (Unfortunately, this article and most other discussions of measurement error are difficult for the untrained reader.)

Figure 7–14 **Relationship Under Varying Degrees of Random Measurement Error**



FURTHER DISCUSSION

In this chapter I have drawn only the broad outlines of correlation and regression analysis for interval data. In fact, I have purposely refrained from giving formulas for calculating these measures, so that anyone wishing to put these techniques to use would be forced to go into them more thoroughly elsewhere. A good standard text in statistics for social scientists is Hubert M. Blalock, *Social Statistics* (1979). David Knoke and George W. Bohrnstedt's *Statistics for Social Data Analysis* (1994) is a more complete, somewhat more advanced, introduction than Blalock.

One question you might consider is this: What would happen in a regression analysis if the independent variable did not vary at all—if, say, we wanted to relate voting turnout to education, but everyone in our study had the same amount of education? This question looks ahead to Chapter 9.