CHAPTER

5

INTRODUCTION TO CORRELATION AND REGRESSION (ORDINARY LEAST SQUARES)

WHAT THIS CHAPTER IS ABOUT

So far we have been dealing with procedures for analyzing categorical data. We now turn to a powerful body of techniques that can be applied when the dependent variable is an interval or ratio variable: ordinary least-squares regression and correlation analysis. In this chapter we deal with the two-variable case, where we have a dependent variable and a single independent variable, to illustrate the logic. In the following two chapters we deal with multiple regression, which is used when we want to explore the effects of several independent variables on a dependent variable, the typical case in social science research.

INTRODUCTION

Suppose we have a set of data arrayed like this:

Father's Years of Schooling	*Respondent's Years of Schooling*
2	4
12	10
4	8
13	13
6	9
6	4
8	13
4	6
8	6
10	11

What can we say about the relationship between father's education and respondent's education? Not much. Visual inspection of the two arrays is quite uninformative. However, if we *plot* the two variables in two-dimensional space, the nature of the relationship is revealed. When you inspect the plot (Figure 5.1), it is immediately evident that the children of highly educated fathers tend to be highly educated themselves. In this situation, we say that the father's and the respondent's education are *positively correlated.*

Although we can see that the father's and respondent's education are positively correlated, we want to quantify the relationship in two respects. First, we want a way to

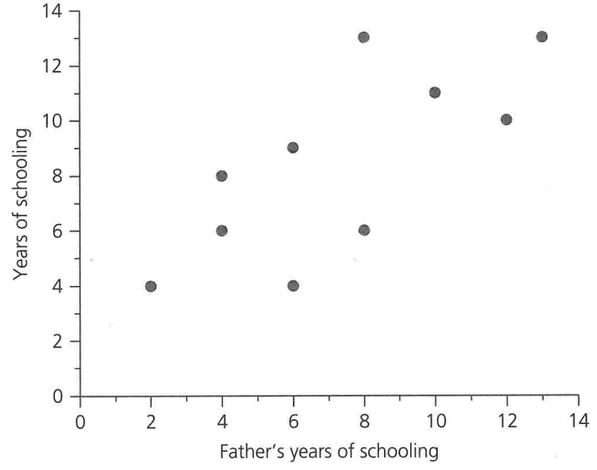

FIGURE 5.1. *Scatter Plot of Years of Schooling by Father's Years of Schooling (Hypothetical Data, N = 10).*

describe the *character* of the relationship between the father's and respondent's years of schooling. How large a difference in the dependent variable, years of schooling, would we expect on average for a person whose father's schooling (the independent variable) differs by one unit (one year)? What level of schooling would we expect, or predict, on average for each person, given that we know how much schooling his or her father has? Second, we want a way to characterize the *strength* of the co-relation, or *correlation*, between the respondent's and father's years of schooling. Can we get a precise prediction of the respondent's level of education from the father's level of education or only an approximate one?

QUANTIFYING THE SIZE OF A RELATIONSHIP: REGRESSION ANALYSIS

The conventional and simplest way to describe the character of the relationship between two variables is to put a straight line through the points that "best" summarizes the average relationship between the two variables. Recall from school algebra that straight lines are represented by an equation of the form

$$Y = a + b(X) \tag{5.1}$$

where a is the *intercept* (the value of Y when the value of X is zero) and b is the *slope* (the change in Y for each unit change in X). Figure 5.2 shows the coefficients a and b for our example involving years of education (Y) and father's years of education (X). The figure is a graphic representation of the equation:

$$\hat{E} = 3.38 + .687(E_F) \tag{5.2}$$

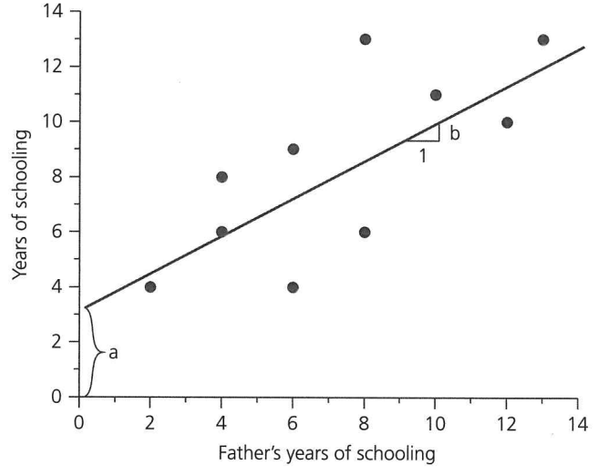

FIGURE 5.2. *Least-Squares Regression Line of the Relation Between Years of Schooling and Father's Years of Schooling.*

Here $\hat{E}$ indicates the *expected* number of years of school completed by people with each level of father's years of schooling (E_F) on the assumption that the relationship is *linear*, that is, that each increase in the father's education produces a given increase in the respondent's education regardless of the initial level; 3.38 is the *intercept*, that is, the expected years of schooling for people whose fathers had no schooling at all; and .687 is the *slope*, that is, the expected increase in years of schooling for each one-year increase in the father's schooling. From this equation, we would predict that those whose fathers have 10 years of schooling would have 10.25 years of schooling because 3.38 + 10*.687 = 10.25. Similarly, we would predict that the children of university graduates would have 2.75 more years of schooling, on average, than the children of high school graduates because .687*(16 – 12) = 2.75. Estimating the value of the dependent variable in a regression equation for given values of the independent variable is known as *evaluating* the equation.

So far we have said nothing about how we derive the values for the coefficients shown in Equation 5.2. The criterion for putting a line through a set of points is that we minimize the sum of the squared errors of prediction—that is, we minimize the sum of the squared differences between the observed and predicted values. Lines derived in this way are known as *ordinary least-squares regression* lines. Figure 5.3 illustrates this criterion. The term e_i ($= E_i - \hat{E}_i$, or the actual number of years of schooling for the *i*th person minus the expected number of years of schooling for that person given his or her father's years of schooling) shown in the figure is the error of prediction between the specified point and the regression line. If we square each of these errors of prediction (which are also called *residuals*) and sum them, there is one and only one line for which this sum of squares is smallest. This is the ordinary least-squares (OLS) regression line.

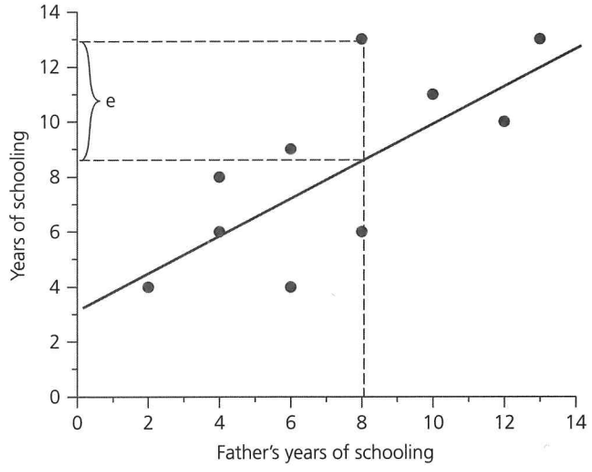

FIGURE 5.3. *Least-Squares Regression Line of the Relation Between Years of Schooling and Father's Years of Schooling, Showing How the "Error of Prediction" or "Residual" Is Defined.*

WHY USE THE "LEAST SQUARES" CRITERION TO DETERMINE THE BEST-FITTING LINE?

Note that "least squares" is not the only plausible criterion of "best fit." An intuitively more appealing criterion is to minimize the sum of the absolute deviations of observed values from expected values. Absolute values are mathematically intractable, however, whereas sums of squares have convenient algebraic properties, which is probably why the inventors of regression analysis hit upon the criterion of minimizing the sum of squared errors. The consequence is that observations with unusually large deviations from the typical pattern of association can strongly affect regression estimates; because the deviations are squared, such observations have the greatest weight. The presence of atypical observations, known in this context as high leverage points, can therefore produce quite misleading results. We will discuss this point further in the upcoming paragraphs and in Chapter Ten.

It can be shown, via algebra or calculus, that the following formulas for the slope and intercept satisfy the least squares criterion:

$$b = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X^2)} \tag{5.3}$$

and

$$a = \bar{Y} - b(\bar{X}) = \frac{\sum Y}{N} - b\frac{\sum X}{N} \tag{5.4}$$

ASSESSING THE STRENGTH OF A RELATIONSHIP: CORRELATION ANALYSIS

Now that we have seen how regression lines are derived and how they are interpreted, we need to assess how good the prediction is. Our criterion for goodness of prediction or *goodness of fit* is the fraction or proportion of the variance in the dependent variable that can be attributed to variance in the independent variable. We define

$$r^2 = 1 - \frac{\sum(Y-\hat{Y})^2/N}{\sum(Y-\bar{Y})^2/N} \tag{5.5}$$

That is, r^2, which is just the square of the Pearson correlation coefficient, is equal to 1 minus the ratio of the variance around the regression line to the variance around the mean of the dependent variable. (The Pearson correlation coefficient is, of course, the correlation coefficient you have encountered in introductory statistics courses. It has the

advantage of ranging from −1 to +1 depending on whether two variables move together or in opposite directions. But it is not as readily interpretable as its square.) When the variance around the regression line is just as large as the variance around the mean of the dependent variable—that is, when knowing the value of the independent variable does not help us predict the value of the dependent variable (in which case, the mean of the dependent variable is the least squares prediction of each value)—the ratio is 1 and $r^2 = 0$; this case is illustrated in (a) of Figure 5.4. When knowledge of the value of the independent variable allows perfect prediction of the value of the dependent variable, the ratio is 0, and hence $r^2 = 1$; this case is illustrated in (b) of Figure 5.4.

Note that OLS regression finds the best *linear* relationship between two variables, even when the actual functional form of the relationship is nonlinear. For example, the correlation between X and Y in (c) of Figure 5.4 is zero, even though it is obvious that the two variables are perfectly (curvilinearly) related. See also Figure 10.1, which reproduces a set of graphs constructed by Anscombe (1973) to show that a given correlation may be associated with very different relationships between two variables. Linear regression provides an adequate summary of a relationship only when it correctly represents the

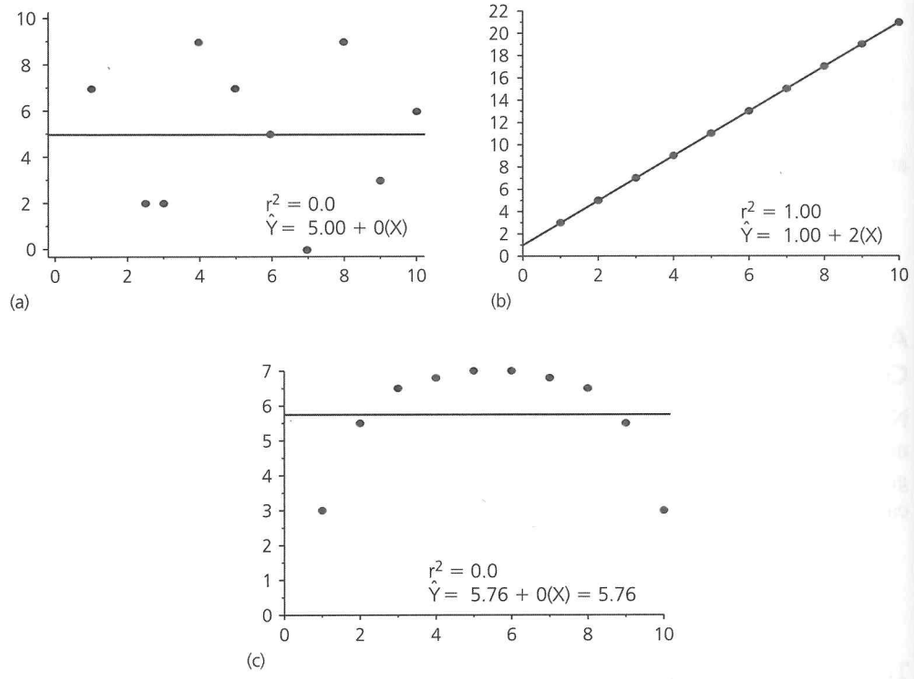

FIGURE 5.4. *Least-Squares Regression Lines for Three Configurations of Data: (a) Perfect Independence, (b) Perfect Correlation, and (c) Perfect Curvilinear Correlation—a Parabola Symmetrical to the x-Axis.*

KARL PEARSON (1857–1936) established the discipline of mathematical statistics and was the principal developer of linear regression and correlation; in recognition of this, the product moment, or ordinary least squares, correlation coefficient, r, is also known as Pearson's r. Pearson's work on classifying probability distributions forms the basis for classical (frequentist) statistical theory and underlies the general linear model. But his contributions are very extensive—for example, he invented the standard deviation and the x2 test. He founded the journal Biometrika in 1901 and edited it until his death; he also founded the journal Annals of Eugenics (now Annals of Human Genetics) in 1925. Pearson was born in London to a family of religious dissenters. He studied mathematics at Cambridge but then studied medieval and sixteenth-century German literature at the Universities of Berlin and Heidelberg and became enough of an expert to be offered a Germanics post at Kings College, Cambridge, which he declined. Instead, he read law (his father was a barrister) but never practiced, returning to mathematics. In his youth he also became a feminist and a socialist (the transformation of his birth name, Carl, to Karl, was said to have resulted initially from the way his name was spelled by a clerk when he enrolled at Heidelberg but supposedly was adopted by him in tribute to Karl Marx, whom he apparently had met). He eventually became universally known as KP. In 1884 he was appointed to the Goldsmid Chair of Applied Mathematics and Mechanics at University College, London, and, in 1891, to the chair in Geometry. There he met W.F.R. Weldon, a zoologist interested in evolutionary theory who posed a number of research problems that stimulated Pearson to think about statistical distributions; their collaboration lasted until Weldon's untimely death in 1906.

character of the relationship. When it fails to do so, additional variables need to be included in the model. You will see how to do this in the next chapter.

Returning to our example about intergenerational continuity in educational attainment, we note that $r^2 = .536$, which tells us that the variance around the regression line is about half the size of variance around the mean of the dependent variable, and therefore that about half of the variance in educational attainment is explained by the corresponding variability in father's education. As social science results go, this is a very high correlation.

A USEFUL COMPUTATIONAL FORMULA FOR *r*

The following is a useful computational formula for the correlation coefficient, r, which comes in handy when you have to do hand calculations:

$$r = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

THE RELATIONSHIP BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

Suppose we were to *standardize* our variables before computing the regression of Y on X, by, for each variable, subtracting the mean from the value of each observation and dividing by the standard deviation. Doing this produces new variables with mean = 0 and standard deviation = 1. Then we would have a regression equation of the form

$$\hat{y} = \beta(x) \tag{5.6}$$

(The convention adopted here, which is widely but not universally used, is to represent standardized variables by lowercase Latin symbols and the coefficients of standardized variables by Greek rather than Latin symbols.) There is no intercept because the regression line must necessarily pass through the mean of each variable, which for standardized variables is the (0,0) point. We interpret β as indicating the number of *standard deviations* by which we would expect two observations to differ on Y that differ by one standard deviation on X. (This follows directly from the fact that for standardized variables, the standard deviation is one. Thus, one standard deviation on X is one unit on x; and the same for Y and y.) It can be shown, through a simple manipulation of the algebraic computational formulas for the coefficients, that in the two-variable case, $r = \beta$. It is also true that r is invariant under linear transformations. (A linear transformation is one in which a variable is multiplied [or divided] by a constant and/or a constant is added [or subtracted] . Consider two variables, Y and Y', with $Y' = a + b(Y)$. In this case, $r_{xy} = r_{xy}'$.) So the correlation between standardized variables and unstandardized variables is necessarily perfect.

A convenient pair of formulas for moving between b and β (which also holds for *multiple regression* coefficients) is

$$\beta = b\left(\frac{s_X}{s_Y}\right) \Rightarrow \quad b = \beta\left(\frac{s_X}{s_Y}\right) \tag{5.7}$$

$$a = \overline{Y} - b(\overline{X}) \tag{5.8}$$

where s_X and s_Y are the the standard deviations of X and Y, respectively.

FACTORS AFFECTING THE SIZE OF CORRELATION (AND REGRESSION) COEFFICIENTS

Now that we see how to interpret correlation and regression coefficients, we need to consider potential troubles—factors that affect the size of coefficients in ways that may lead to incorrect interpretation and false inferences by the unwary.

Outliers and Leverage Points

As noted, correlation and regression statistics are very sensitive to observations that deviate substantially from the typical pattern. This is a consequence of the least squares criterion—because "errors" (differences between observed and predicted values on the

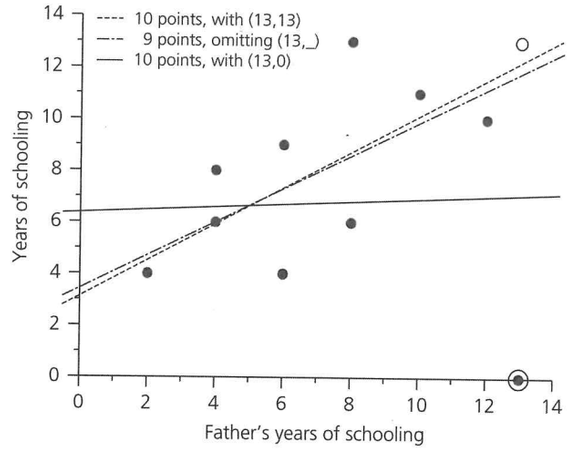

FIGURE 5.5. *The Effect of a Single-Deviant Case (High Leverage Point).*

dependent variable) are squared, the larger the error, the more it will contribute to the sum of squared errors relative to its absolute size. Thus, correlation coefficients can be substantially affected by a few deviant observations, with regression slopes pulled strongly toward them, producing misleading results. To see this, consider the following example, illustrated in Figure 5.5. Suppose that in our example about intergenerational educational transmission, the fourth case had values (13,0) (shown as a solid circle surrounded by an open circle) instead of (13,13) (shown as an open circle). That is, suppose that in the fourth case the child of a man with thirteen years of schooling had no education instead of thirteen years of schooling—perhaps because the child was mentally impaired. The alteration of just one point, from (13,13) to (13,0), dramatically changes the regression line and misrepresents the typical relationship between the father's and respondent's education, making it appear that there is no relationship at all (the regression equation for the ten points with (13,0) as the fourth value is $\hat{E} = 6.74 + .0491(E_F)$; $r^2 = .002$).

This example illustrates the condition under which deviant cases are influential—that is, have high "leverage." This is when points are far away from the center of the multivariate distribution. Outliers close to the center of the distribution, for example, the (8,13) point in Figure 5.5, have less influence because, although they can pull the regression line up or down, they have relatively little effect on the slope. We will consider this distinction further in Chapter Ten.

The most straightforward solution is to omit the offending case. When this is done, the regression line through the remaining nine points is very close to the regression line through ten points with (13,13). However, this generally is an undesirable practice because it creates the temptation to start "cleaning up" the data by omitting whatever cases tend to fall far from the regression surface. Two better strategies, which will be elaborated in Chapters Seven and Ten, are (1) to think carefully about whether the outliers

might have been generated by a different process from the remainder of the data and, when you suspect that possibility, to explicitly model the process; or (2) to use a robust regression procedure that downweights large outliers. Fortunately, the damage done by outliers diminishes as sample sizes increase. However, even with large samples extreme outliers can be distorting—for example, incomes in the millions of dollars. One simple way to deal with extreme values on univariate distributions is to truncate the distribution, for example, in the United States in 2006 by specifying $150,000 for incomes of $150,000 or above (this is what the GSS does; in 2006, just over 2 percent of the GSS sample had incomes this high); but this creates its own problems, as we will see next. A better way, which you will see in Chapter Fourteen, is to use interval regression (an elaboration of tobit regression) to correctly specify the category values.

Truncation

Analysts are sometimes tempted to divide their study population into subgroups on the basis of values on the independent or dependent variable or on variables substantially correlated with the independent or dependent variable. For example, an analyst who suspects that income depends more heavily on education among those with nonmanual occupations than among those with manual occupations might attempt to test this hypothesis by correlating income with education separately for nonmanual and manual workers. This is a bad idea because income is correlated with occupational status; thus, dividing the population on the basis of occupational status will truncate the distribution of the dependent variable, which, all else equal, will reduce the size of the correlation. Moreover, if one subgroup, say manual workers, has a smaller variance with respect to income than does the other subgroup, say nonmanual workers (and this is likely to be true in most societies), the size of the correlation will be more substantially reduced for manual than for nonmanual workers, thus leading the analyst to—mistakenly—believe that the hypothesis is confirmed.

To see this, consider a highly stylized example, shown as Figure 5.6. To keep the example simple, imagine that all manual workers in the sample have less than seven years of schooling and that all nonmanual workers have more than seven years of schooling. Note that in the example, there is exactly the same income return to an additional year of education for nonmanual and manual workers. Note further that each point is an equal distance from the regression line. Now, suppose the correlation between income and education were computed separately for manual and nonmanual workers. The correlation for both groups would be smaller than the correlation computed over the total sample, and the correlation would be smaller for manual than for nonmanual workers. This follows directly from Equation 5.5 because, from the way the example was constructed, the variance around the regression line is identical in all three cases, but the variance around the mean of the dependent variable is smaller for nonmanual workers than for the total sample and smaller for manual workers than for nonmanual workers. Although, for the sake of clarity, the example is highly stylized, the principle holds generally: when distributions are truncated the correlation tends to be reduced. This, by the way, is the main reason GRE scores are weak predictors of grades in graduate school courses: graduate

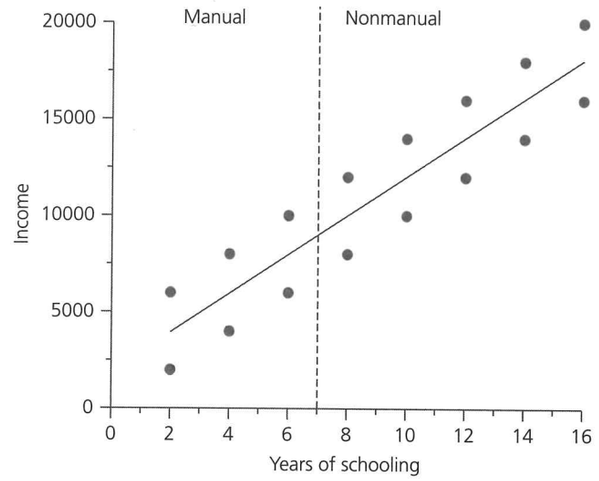

FIGURE 5.6. *Truncating Distributions Reduces Correlations.*

departments do not admit people with low GREs, thereby truncating the distribution of GRE scores. But this does not imply that GRE scores should be ignored in the admissions process, as statistically illiterate professors argue from time to time.

A "REAL DATA" EXAMPLE OF THE EFFECT OF TRUNCATING THE DISTRIBUTION Analyzing the U.S. sample for the *Political Action: An Eight Nation Study, 1973–1976* (Barnes and Kaase 1979) some years ago, I was puzzled to discover an extremely low correlation between education and income (less than .1, whereas in U.S. surveys the typical correlation between these two variables is on the order of .3). Further investigation revealed that the low end of both the education and income distributions were severely truncated, presumably as a result of inadequacies in either the sampling or the field work procedures. When the data were weighted to reproduce the bivariate distribution of education and income observed in the U.S. census for 1980 (the year closest to the survey), the estimated correlation approximated that typically found in U.S. surveys.

Regression Toward the Mean

The consequences of truncation actually are worse than just suggested, because of a phenomenon known as "regression toward the mean." When two measurements are made at different points in time, for example, pre-test and post-test measurements in a randomized experiment or scores on the GRE, it is typical to observe that those cases with high values on the first observation tend, on average, to have lower values on the second observation, and that those cases with low values on the first observation tend to have higher

values on the second observation. That is, both the high and the low values move toward (or "regress toward") the mean. This is true even when there is no change in the *true* value between the two measurements.

The reason for this is that observed measurements consist of two components: a true score and a component representing error in measurement of the underlying true score. For example, consider the GRE. The observed score for each individual can be thought of as consisting of a component measuring the candidate's "true" (or underlying or constant) ability to do the kind of work measured by the test and a random component comprised of variations in the exact questions asked in that administration of the test, the candidate's level of energy and mental acuity, level of confidence (Steele 1997), and so on. It then follows that those who have high scores in any given administration of the test will disproportionately include those who have high positive random components, and those who have low scores will disproportionately include those who have low random components. But because the second component *is* random, those who have high random components on the first test will tend, on average, to have lower random components on the second test and those who have low random components on the first test, will tend, on average, to have higher random components on the second test. The result is that the correlation between the two tests will be less than perfect and also that the regression coefficient relating the second to the first test will be less than 1.0. This is true even if the means and standard deviations of the two tests are identical.

An important implication of this result is that a researcher who targets for special intervention a low-scoring group (those who did poorly on a practice GRE, those with low grade point averages, and so on) will be bound to conclude, incorrectly, that the intervention was successful. Of course, if that same researcher chose the high-scoring group for the same intervention, he or she would be forced to conclude that the intervention was completely unsuccessful—indeed, that it was counterproductive. All of this is a simple consequence of analyzing a nonrandom subset of the original sample.

Exactly the same phenomenon—measurement error—has the effect of lowering the correlation between separate phenomena, for example, education and income, the heights of fathers and sons, and so on. This kind of observation is what led Francis Galton, one of the founders of correlation and regression analysis, to conclude in the late nineteenth century that a natural phenomenon of intergenerational transmission was a "reversion" (or "regression") toward "mediocrity"—hence the term "regression analysis" to describe the linear prediction procedure discussed here. But what Galton failed to notice is that there is also, and for exactly the same reason, a tendency for values near the mean to move *away* from the mean. The result is that the variance of the *predicted* (but not the *observed*) values and the slope of the regression line are reduced in proportion to the complement of the correlation between the variables. (For a book-length treatment of this topic, see Campbell and Kenney [1999].)

Aggregation

Students who have spent some time studying the behavior of populations of individuals usually conclude that we live in a stochastic world in which nothing is very strongly related to anything else. For example, in the United States, typically about 10 percent of

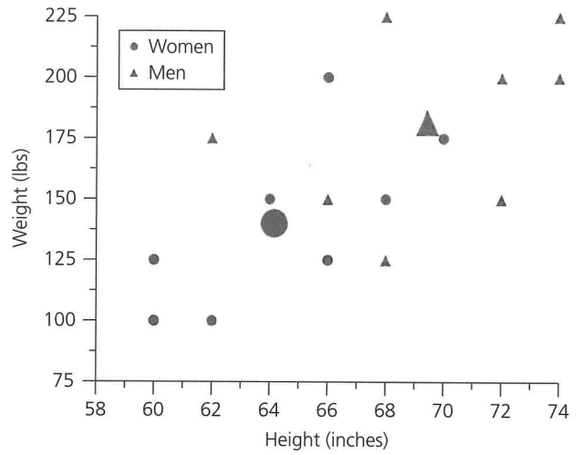

FIGURE 5.7. *The Effect of Aggregation on Correlations.*

the variance in income can be attributed to variance in education ($r \approx .3 \Rightarrow r^2 \approx .09$). Students are then puzzled when they discover that seemingly comparable correlations computed over aggregates, for example, the correlation between mean education and mean income for the detailed occupational categories used by the U.S. Bureau of the Census, tend to be far larger (in the present example, $r \approx .7 \Rightarrow r^2 \approx .49$). Why is this so? The explanation is simple. When correlations are computed over averages or other summary measures, a great deal of individual variability tends to "average out." In the extreme case, where there are only two aggregate categories, the correlation between the means for the two categories will necessarily be 1.0, as you can see in Figure 5.7 (where the large circle represents the mean height and weight for women, and the large triangle represents the mean height and weight for men); but the principle holds for more than two categories as well.

CORRELATION RATIOS

So far we have been discussing cases where we have two interval or ratio variables. Sometimes, however, we want to assess the strength of the association between a categorical variable and an interval or ratio variable. For example, we might be interested in whether religious groups differ in their acceptance of abortion. Or we might be interested in whether ethnic groups differ in their average income. The obvious way to answer these questions is to compute the mean score on an abortion attitudes index for each religious group or the mean income for each ethnic group. But if we discover that the means differ substantially enough to be of interest, we still are left with the question of how strong the relationship is. To determine this we can compute an analog to the (squared) correlation coefficient, known as the (squared) correlation ratio, η^2 (eta squared). η^2 is defined as

$$\eta^2 = 1 - \frac{\textit{Variance around the subgroup means}}{\textit{Variance around the grand mean}}$$
$$= 1 - \frac{\textit{Within group sum of squares}}{\textit{Total sum of squares}} \qquad (5.9)$$
$$= 1 - \frac{\sum_j \sum_i (Y_{ij} - \bar{Y}_{.j})^2}{\sum_j \sum_i (Y_{ij} - \bar{Y}_{..})^2}$$

where Y is the dependent variable, there are j groups, and i cases within each group. Thus, $\bar{Y}_{.j}$ is the mean of Y for group j, and $\bar{Y}_{..}$ is the grand mean of Y. From Equation 5.9, it is evident that if all the groups have the same mean on the dependent variable, knowing which group a case falls into explains nothing; the variance around the subgroup means equals the variance around the grand mean, and $\eta^2 = 0$. At the other extreme, if the groups differ in their means, and if all cases within each group have the same value on the dependent variable—that is, there is no within-group variance—then the ratio of the within-group sum of squares to the total sum of squares is 0, and $\eta^2 = 1$. From this we see that η^2, like r^2, is a *proportional reduction in variance* measure.

Let us explore the religion and abortion acceptance example with some actual data. In 2006 (and for most years since 1972) the GSS asked seven questions about the acceptability of abortion under various circumstances:

. . . should [it] be possible for a woman to obtain a legal abortion . . .

- if there is a strong chance of serious defect in the baby?
- if she is married and does not want any more children?
- if the woman's own health is seriously endangered by the pregnancy?
- if the family has a very low income and cannot afford any more children?
- if she became pregnant as a result of rape?
- if she is not married and does not want to marry the man?
- if the woman wants it for any reason?

From these items I constructed a scale by counting the positive responses, excluding all cases with any missing data. The scale thus ranges from 0 to 7. Table 5.1 shows the mean number of positive responses by religion. All those who specified religions other than Protestant, Catholic, or Jewish or said they had no religion were included in the "Other and None" category. From the table, it is evident that Jews and other non-Christians are much more accepting of abortion than are Christians (Protestants and Catholics). But how important is religion in accounting for acceptance of abortion? To see this, we compute $\eta^2 = .070$. (The Stata computations to create Table 5.1 and to obtain η^2 are shown in the downloadable -do- and -log- files for the chapter.)

TABLE 5.1. Mean Number of Positive Responses to an Acceptance of Abortion Scale (Range: 0–7), by Religion, U.S. Adults, 2006.

Religion	Mean Number of Positive Responses	Standard Deviation	N
Protestants	3.7	2.5	(923)
Catholics	3.8	2.5	(420)
Jews	5.6	2.5	(26)
Other or none	5.3	2.2	(395)
Total	**4.1**	**2.5**	**(1,764)**

Clearly, religious affiliation does not explain much of the variance in abortion attitudes. How can this be, given the substantial size of the mean differences? The answer is simple. Jews and "Others" differ substantially from Protestants and, especially, Catholics in their acceptance of abortion. But these groups are quite small, especially Jews. Hence, no matter how deviant they are from the overall average, they are unlikely to have much impact; when more than half of the population is included in one group, as is the case here with Protestants, a large fraction of the variance in abortion acceptance is bound to be within-group variance rather than between-group variance.

A second use of the correlation ratio is to test assumptions of linearity. We will take this up in Chapter Seven.

A USEFUL COMPUTATIONAL FORMULA FOR η^2

A good formula to compute η^2 by hand from frequency or percentage distributions is

$$\eta^2 = \sum_j \sum_i f_{ij} X_{ij}^2 - \frac{\left(\sum_j \sum_i f_{ij} X_{ij}\right)^2}{\sum_j \sum_i f_{ij}}$$

where there are j groups and i categories of the dependent variable, which in this case is designated by X. So X_{ij} is the score for the ith category (of the jth group, although the category scores are the same for all groups), and f_{ij} is the number of cases in the ith category among members of the jth group. Notice the difference from Equation 5.9, where the i refers to individuals rather than to categories of the dependent variable.

WHAT THIS CHAPTER HAS SHOWN

In this chapter we have considered simple (two-variable) ordinary least-squares (OLS) correlation and regression, as a way of seeing the conceptual basis of OLS regression, the workhorse of modern statistical analysis. We also considered how the size of correlation and regression coefficients is affected by the bivariate distribution of cases—specifically, how results are affected by high-leverage outliers, by truncation, by regression to the mean, and by aggregation. It is important that you understand these effects thoroughly because many confused claims are made by those who fail to understand them. We then considered a variant on correlation coefficients, the squared correlation ratio, which is an analog to correlation when we have an interval or continuous dependent variable but a categorical independent variable. In the next chapter we extend our discussion to multiple correlation and regression, the analogous OLS technique when we have two or more independent variables.

CHAPTER

6

INTRODUCTION TO MULTIPLE CORRELATION AND REGRESSION (ORDINARY LEAST SQUARES)

WHAT THIS CHAPTER IS ABOUT

In this chapter we consider the central technique for dealing with the most typical social science problem—understanding how some outcome is affected by several determining variables that are correlated with each other. We begin with a conceptual overview of multiple correlation and regression, and then continue with a worked example to illustrate how to interpret regression coefficients. We then turn to consideration of the special properties of categorical independent variables, which can be included in multiple regression equations as a set of dichotomous ("dummy") variables, one for each category of the original variable (except that to enable estimation of the equation, one category must be represented only implicitly). In the course of our discussion of dummy variables, we develop a strategy for comparing groups that enables us to determine whether whatever social process we are investigating operates in the same way for two or more subsegments of the population—males and females, ethnic categories, and so on. We conclude with an alternative way of choosing a preferred model, the Bayesian Information Coefficient (*BIC*).