1 What Is Multiple Regression?

In this chapter, we examine some of the basic characteristics of multiple regression. The aim is to give you enough information to begin to read and interpret results from multiple regression analysis. In later chapters, we'll revisit many of these questions and answers in greater detail.

1.1. What Is Multiple Regression?

Multiple regression is a statistical method for studying the relationship between a single *dependent* variable and one or more *independent* variables. It is unquestionably the most widely used statistical technique in the social sciences. It is also widely used in the biological and physical sciences.

1.2. What Is Multiple Regression Good For?

There are two major uses of multiple regression: prediction and causal analysis. In a prediction study, the goal is to develop a formula for making predictions about the dependent variable, based on the observed values of the independent variables. For example, an economist may want to predict next year's gross national product (GNP) based on such variables as last year's GNP, current interest rates, current levels of unemployment, and other variables. A criminologist may want to predict the likelihood that a released convict will be arrested, based on his age, the number of previous arrests, and the crime for which he was imprisoned.

In a causal analysis, the independent variables are regarded as causes of the dependent variable. The aim of the study is to determine

whether a particular independent variable *really* affects the dependent variable, and to estimate the magnitude of that effect, if any. For example, a criminologist may have data showing that prisoners who participate in educational programs are less likely to be re-arrested after they are released. She may perform a multiple regression to see if this apparent relationship is real or if it could be explained away by the fact that the prisoners who enroll in educational programs tend to be those with less serious criminal histories.

These two uses of multiple regression are not mutually exclusive. The criminologist whose main interest is in the effect of educational programs may also use the regression model to make predictions about future arrests.

1.3. Are There Other Names for Multiple Regression?

A more complete name is *ordinary least squares multiple linear regression*. *Least squares* is the method used to estimate the regression equation. *Ordinary* serves to distinguish the simplest method of least squares from more complicated methods such as weighted least squares, generalized least squares, and two-stage least squares. *Multiple* means that there are two or more independent variables. *Linear* describes the kind of equation that is estimated by the multiple regression method. You'll often see various combinations of these words, as in "linear regression" or "least squares regression" or "OLS regression." (OLS stands for ordinary least squares.)

The term *regression* is harder to explain. One of the early uses of regression was by the English scientist Sir Francis Galton (1822-1911), who was investigating the relationship between heights of fathers and sons. Galton used a linear equation to describe that relationship. He noticed, however, that very tall fathers tended to have sons who were shorter than they were, whereas very short fathers tended to have sons who were taller than they were. He called this phenomenon "regression to the mean," and somehow that name stuck to the entire method.

You'll also see other names for the variables used in a multiple regression analysis. The dependent variable is sometimes called the *response* variable or the *outcome* variable. The independent variables

may be referred to as *predictor* variables, *explanatory* variables, *regressor* variables, or *covariates*.

1.4. Why Is Multiple Regression So Popular?

Multiple regression does two things that are very desirable. For prediction studies, multiple regression makes it possible to *combine* many variables to produce optimal predictions of the dependent variable. For causal analysis, it *separates* the effects of independent variables on the dependent variable so that you can examine the unique contribution of each variable. In later sections, we'll look closely at how these two goals are accomplished.

In the last 30 years, statisticians have introduced a number of more sophisticated methods that achieve similar goals. These methods go by such names as logistic regression, Poisson regression, structural equation models, and survival analysis. Despite the arrival of these alternatives, multiple regression has retained its popularity, in part because it is easier to use and easier to understand.

1.5. Why Is Regression "Linear"?

To say that regression is linear means that it is based on a linear equation. In turn, a linear equation gets its name from the fact that if you graph the equation, you get a straight line. This is easy to see if there is a dependent variable and a single independent variable. Suppose that the dependent variable is a person's annual income, in dollars, and the independent variable is how many years of schooling that person has completed. Here's an example of a linear equation that predicts income, based on schooling:

$$\text{INCOME} = 8{,}000 + (1{,}000 \times \text{SCHOOLING}).$$

If you draw a graph of this equation, you get the straight line shown in Figure 1.1.

This simple equation makes it possible to predict a person's income if we know how many years of schooling the person has completed. For example, if a person has 10 years of schooling, we get a predicted income of $18,000:

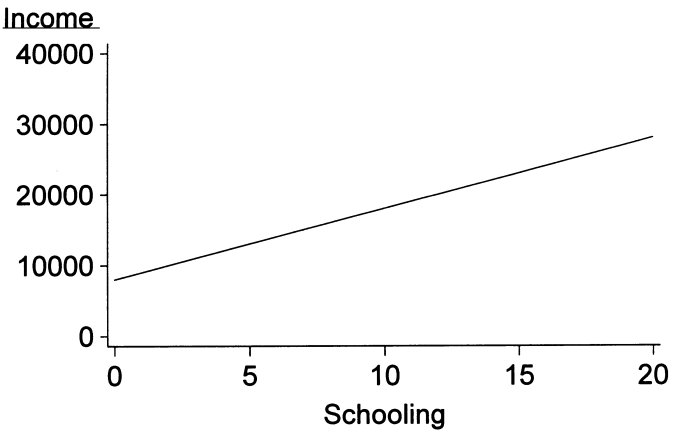

Figure 1.1. Graph of Regression Equation for Income on Schooling

$$18{,}000 = 8{,}000 + (1{,}000 \times 10).$$

Note that, according to the equation, a person with 0 years of schooling is predicted to earn $8,000. Also, each additional year of schooling increases the predicted income by $1,000.

Is it possible to get reasonable predictions from such a simple equation? Maybe and maybe not. A complete answer to this question involves many complex issues, some of which we will consider later. One issue is whether we could get better predictions with different numbers besides 8,000 and 1,000. Perhaps 9,452 and 1,789 would do better. The method of least squares is designed to find numbers that, in some sense, give us optimal predictions of the dependent variable.

We can write the two-variable linear equation in more general terms as

$$y = a + bx.$$

In this equation, y is the dependent variable and x is the independent variable. In our example, y is income and x is years of schooling. The letters a and b represent constant numbers. We call a the *intercept* and b the *slope*. These names refer to features of the graph in Figure 1.1. The intercept is the point on the vertical axis which "intercepts" the line. In other words, it is the value of y when x is 0. In this example, the intercept is 8,000. The slope tells us how big a change in y we get from a 1-unit increase in x. In this example, y goes up by $1,000 for

each 1-year increase in schooling. Clearly, a larger slope corresponds to a steeper line. If the slope is 0, on the other hand, the line is perfectly flat. If the slope is negative, then an increase in x results in a *decrease* in y.

1.6. What Does a Linear Equation Look Like With More Than Two Variables?

In most applications of regression analysis, the linear equation has more than one independent variable. For prediction purposes, you can usually get better predictions if you base them on more than one piece of information. For causal analysis, you want to be able to look at the effect of one variable while *controlling* for other variables. This is accomplished by putting the other variables in the regression equation.

Suppose, for example, that we want to include age as a predictor of income. This makes sense because most people's incomes increase with age, at least until they retire. A linear equation that incorporates age might look like this:

$$\text{INCOME} = 6{,}000 + (800 \times \text{SCHOOLING}) + (400 \times \text{AGE}).$$

This equation tells us that income goes up by $400 for each additional year of age; it also goes up by $800 for each additional year of schooling. For a person who is 40 years old and has 14 years of schooling, the predicted income is

$$33{,}200 = 6{,}000 + (800 \times 14) + (400 \times 40).$$

A more general way of writing an equation with two independent variables is

$$y = a + b_1x_1 + b_2x_2.$$

In our example, x_1 is schooling and x_2 is age. We still call this a linear equation, although it's more difficult to draw a graph that looks like a straight line. (It would have to be a 3-dimensional graph, and the equation would be represented by a plane rather than a line.) The b's are called *slope coefficients*, but often we just call them coefficients or slopes. The essence of a linear equation is this: We multiply each

variable by some number (the slope for that variable). We add those products together. Finally, we add another number (the intercept).

1.7. Why Does Multiple Regression Use Linear Equations?

We have just described the relationship between income, schooling, and age by a linear equation. Is this sensible? Maybe the real relationship is something highly nonlinear, like the following:

$$y = \left(\frac{a_1 + b_1 x_1}{a_2 + b_2 x_2} \right)^d$$

Such an equation is certainly possible. On the other hand, there is no reason to think that this nonlinear equation is any better than the linear equation. A useful general principle in science is that when you don't know the true form of a relationship, start with something simple. A linear equation is perhaps the simplest way to describe a relationship between two or more variables and still get reasonably accurate predictions. The simplicity of a linear equation also makes it much easier and faster to do the computations necessary to get good estimates of the slope coefficients and the intercept.

Even if the true relationship is *not* linear, a linear equation will often provide a good approximation. Furthermore, it's easy to modify the linear equation to represent certain kinds of nonlinearity, as we'll see in Chapter 8. Consider the relationship between age and income. Although income certainly increases with age, it probably increases more rapidly at earlier ages and more slowly at later ages, and it may eventually begin to decrease. We can represent this kind of relationship by including both age and the *square* of age in the equation:

$$\text{INCOME} = a + b_1\text{AGE} + b_2\text{AGE}^2.$$

Figure 1.2 shows a graph of this equation for certain values of the slopes and the intercept. Equations like this can easily be handled by any computer program that does ordinary multiple regression. Highly nonlinear equations—like the one shown earlier—require more specialized computer programs that are not so widely available.

Figure 1.2. Income as a Nonlinear Function of Age

1.8. What Kinds of Data Are Needed for Multiple Regression?

To do a regression analysis, you first need a set of cases (also called units of analysis or observations). In the social sciences, the cases are most often persons, but they could also be organizations, countries, or other groups. In economics, the cases are sometimes units of time, like years or quarters. For each case, you need measurements on all the variables in the regression equation.

Table 1.1 is an example of a data set that could be used to do a multiple regression predicting income from age and years of schooling. The data come from the General Social Survey, an annual survey based on a national probability sample of U.S. adults (Davis and Smith, 1997). Table 1.1 lists a small subset of these data, specifically, all white males living in New England. The data are arranged so that each row corresponds to a case (person) and each column corresponds to a variable. For example, the first row describes a person whose income was $48,000 who had 12 years of schooling and was 54 years old. Neither the cases nor the variables have to be in any particular order. Virtually any regression program could read the data in Table 1.1 exactly as they appear.

The 35 cases in Table 1.1 are sufficient to do the multiple regression but, as in any statistical analysis, the more cases the better. For the computation to work at all, you must have at least as many cases as variables (including the dependent variable). To do a decent job,

TABLE 1.1 Data on Income, Schooling, and Age, 1983 General Social Survey

Income	*Schooling*	*Age*
48,000	12	54
26,000	12	28
26,000	7	56
48,000	14	47
13,000	14	23
34,000	12	60
18,000	11	36
24,000	16	34
81,000	16	61
21,000	12	38
9,000	6	53
18,000	12	34
34,000	13	58
21,000	14	38
81,000	12	46
48,000	20	54
6,000	7	76
21,000	14	35
21,000	12	34
9,000	14	23
34,000	14	44
7,000	9	31
24,000	8	56
34,000	16	37
34,000	17	40
4,000	12	20
5,000	9	65
13,000	14	53
7,000	20	33
13,000	12	31
34,000	7	30
10,000	16	36
48,000	18	54
6,000	12	19
2,000	10	25

you need far more than that. Most regression analysts would be reluctant to do a regression with less than five cases per variable, although there are exceptional situations when fewer cases might be enough.

The most desirable data come from a probability sample from some well-defined population, as is the case with the data in Table 1.1. In practice, people often use whatever cases happen to be available. A medical researcher, for example, may use all the patients admitted to a particular hospital in a 1-year period. An educational researcher may use all the students enrolled in a particular school. Although it is acceptable to use such "convenience samples," you must be very cautious in generalizing the results to other populations. What you find in one school or hospital may not apply to any other. Convenience samples are also more likely to violate the assumptions that justify multiple regression (see Chapter 6).

1.9. What Kinds of Variables Can Be Used in Multiple Regression?

For the data in Table 1.1, all the variables were quantitative variables. Age, income, and years of schooling are all measured on some well-defined scale. For each of these scales, it's reasonable to claim that an increase of a specified amount means the same thing no matter where you start. Thus, an increase from $20,000 to $30,000 is, in some sense, equivalent to an increase from $30,000 to $40,000. An increase from 25 to 30 years of age is comparable to an increase from 30 to 35 years of age. Variables like this, called *interval scales*, are entirely appropriate for regression analysis.

Many variables in the social sciences don't have this property. Suppose, for example, that a questionnaire includes the statement "This country needs stronger gun control laws" and then asks people whether they strongly agree, agree, disagree, or strongly disagree. The researcher assigns the following scores:

1 = strongly disagree
2 = disagree
3 = agree
4 = strongly agree.

I think most people would accept the claim that higher scores represent stronger agreement with the statement, but it's not at all

clear that the distance between 1 and 2 is the same as the distance between 2 and 3, or between 3 and 4. Variables like this are called *ordinal scales*. The numbers tell you the order on some dimension of interest, but they don't tell you the magnitude of the difference between one value and another.

Strictly speaking, ordinal variables are inappropriate for multiple regression because the linear equation, to be meaningful, requires information on the magnitude of changes. In practice, however, ordinal variables are used quite often in regression analysis because there aren't good alternatives. If you use such variables, you are implicitly assuming that an increase (or a decrease) of one unit on the scale means the same no matter where you start. This might be a reasonable approximation in many cases.

Then there are variables that don't have any order at all. What do you do with a variable like gender (male or female) or marital status (never married, married, divorced, widowed)? Variables like this are called *nominal scales*. If the variable has only two categories, like gender, the solution is easy. Just assign a score of 1 to one of the categories and a score of 0 to the other category. It doesn't matter which one you choose, as long as you remember which is which. Such 1-0 variables are called *dummy variables* or *indicator variables*. Later on we'll discuss how to interpret the slope coefficients for dummy variables. We'll also see how the method of dummy variables can be extended to nominal variables with more than two categories. Dummy variables are perfectly OK as *independent* variables in a multiple regression. Although it's not fatal to use a dummy variable as a *dependent* variable in a regression analysis, there are much better methods available. The most popular alternative—known as logit analysis or logistic regression—will be briefly discussed in Chapter 9.

1.10. What Is Ordinary Least Squares?

Ordinary least squares is the method most often used to get values for the regression coefficients (the slopes and the intercept). The basic idea of least squares is pretty simple, although the computations can get quite complicated if there are many independent variables.

If we knew the values of the regression coefficients, we could use the linear equation to produce a predicted value of the dependent variable for each case in the sample. We usually don't know the true values of the coefficients, but we can try out different guesses and see which ones produce the "best" predicted values. Suppose, for example, that we make some guesses for the data in Table 1.1. Let's guess a value of zero for the intercept, $1,000 for the slope for schooling, and $500 for the slope for age. For the first case in the sample, aged 54 with 12 years of schooling, we get a predicted value of

$$0 + (1{,}000 \times 12) + (500 \times 54) = 39{,}000.$$

The observed income for this person is $48,000, so our prediction is $9,000 too low. Still, it's not bad for just guessing the coefficients. Now let's try the second person, who had 12 years of education and was 28 years old:

$$0 + (1{,}000 \times 12) + (500 \times 28) = 26{,}000.$$

Our predicted value is now identical to the observed value of $26,000. For the third person, aged 56 with 7 years of schooling, we have

$$0 + (1{,}000 \times 7) + (500 \times 56) = 35{,}000.$$

This person's observed income is $26,000, so our prediction is $9,000 too high.

We could continue generating predictions for each case in the sample. For each case we could then calculate a *prediction error*:

$$\text{Error} = \text{Observed Value} - \text{Predicted Value}.$$

Some of these errors will be positive (observed higher than predicted), and some will be negative (observed less than predicted). Some will be large and others will be small. Clearly, we would like to find coefficients that make these errors as small as possible.

For data in the social sciences, it's virtually impossible to find coefficients that make all the errors equal to zero. In our example, that would mean that we could perfectly predict every person's income just by knowing his or her age and years of schooling. Not likely. We'll have to settle for making the errors as small as we can. The problem is that if we tried out different sets of coefficients, we would find that errors for some people get larger while errors for

other people get smaller. How do we balance one error against another?

There's no universally accepted answer to this question, but the most widely used method is the least squares criterion. This criterion says to

> *choose coefficients that make the sum of the squared prediction errors as small as possible.*

For the income example, the first error was $9,000, so its square is 81,000,000. The second error was $0 so its square is 0. The third error was –$9,000 and its square is 81,000,000. Notice that squaring this third error turns a negative quantity into a positive quantity. That means that the least squares criterion doesn't care about the direction of the error, just the magnitude. On the other hand, the least squares criterion hates big errors. It would much rather have a lot of small errors than one big error.

If we continue in this way for all 35 cases, we find that the total sum of squared errors for our initial guess of the coefficients is 12,296,499,985. This may seem like a very large number, but there's no way to evaluate it without comparing it to something else. The important question is this: Can we find another set of coefficients that gives us a smaller sum of squared errors?

One way to answer this question is by trial and error. We could make a lot of different guesses for the coefficients and compute the sum of squared errors for each set of guesses. Then we could pick the one with the smallest sum. Unfortunately, the number of possible guesses is infinite, so this would not be a very efficient way to get the optimal set of coefficients. Fortunately, the best set of coefficients can be obtained directly by some straightforward but tedious calculations, best left to a computer. For the data in Table 1.1, the linear equation that minimizes the sum of squared errors is

$$\text{INCOME} = -25{,}965 + (2{,}057 \times \text{SCHOOLING}) + (600 \times \text{AGE}).$$

For this set of coefficients, the sum of squared errors is 9,364,695,694, which is 24% lower than the sum of squared errors for our original guesses.

This equation says that each additional year of schooling increases the predicted annual income by $2,057 and each additional year of age increases the predicted income by $600. The intercept (–$25,965) could be interpreted as the predicted income for a person

who is 0 years old with 0 years of schooling. Of course, no one in this sample has anywhere near these values for age and schooling. In general, the intercept in a regression equation doesn't have a very useful interpretation, especially when values of 0 on the independent variables are far from the values that are actually observed. Still, the intercept is necessary to make the predictions come out right.

1.11. How Can We Judge How Good the Predictions Are?

Least squares regression always produces the "best" set of linear predictions for a given set of data. But sometimes the best isn't very good. It would be nice to have some way of gauging just how good the predictions are. The most common statistic for doing this is something called the *coefficient of determination*. Most people know this statistic by its symbol, R^2, pronounced *r-squared*.

The basic idea behind R^2 is to compare two quantities:

- The sum of squared errors produced by the least squares equation that you're evaluating and
- The sum of squared errors for a least squares equation with *no* independent variables (just the intercept).

When an equation has no independent variables, the least squares estimate for the intercept is just the mean (average) of the dependent variable. That implies that our predicted value for every case is the mean.

For our income example, we subtract the mean income ($25,200) from each observed income, square the result, and sum over all 35 persons. This produces a sum of squared errors of 12,947,600,000. Recall that the sum of squared errors from our equation with age and schooling was 9,364,695,694. Now divide the smaller sum by the larger sum and subtract the result from 1.0, which gives us an R^2 of .28, or 28%. To write this in symbols, let *SSE* be the sum of squared errors. The formula is

$$R^2 = 1 - \frac{SSE\,(\text{regression})}{SSE(\text{mean only})}$$

We say, then, that using age and schooling to predict income yields a 28% reduction in the (sum of squared) prediction errors, compared

with using only the mean. That's not bad for data in the social sciences. Alternatively, we can say that age and schooling "explain" 28% of the variation in income.

1.12. How Do We Judge How Good the Coefficient Estimates Are?

In any regression analysis, we typically want to know something about the accuracy of the numbers we get when we calculate estimates of the regression coefficients. In our income example, we got a least squares coefficient estimate of $600 for the age variable. That may be the best estimate we can get with the data, but it's unlikely to be exactly "right." As in most applications of regression analysis, there are three possible sources of error:

- Measurement error: Very few variables can be measured with perfect accuracy, especially in the social sciences.
- Sampling error: In many cases, our data are only a sample from some larger population, and the sample will never be exactly like the population.
- Uncontrolled variation: Age and schooling are surely not the only variables that affect a person's income, and these uncontrolled variables may "disturb" the relationship between age and income.

To make any headway, we have to make some assumptions about how these sources of error operate. Without going into details, the basic assumption is that the errors occur in a random, unsystematic fashion (see Chapter 6 for more about these assumptions). The result is random variation in our coefficient estimates. We evaluate the extent and importance of this random variation by calculating *confidence intervals* or *hypothesis tests*.

Confidence intervals give us a range of possible values for the coefficients. Although we may not be *certain* that the true value falls in the calculated range, we can be reasonably confident. Hypothesis tests are used to answer the question of whether or not the true coefficient is zero. Again, we never get a definitive answer, but can calculate the probability of being wrong. If you've already had a statistics course, you'll recall that the first step in getting confidence intervals and hypothesis tests is usually the calculation of the standard error (a kind of standard deviation). If you haven't had a statistics course, what follows may seem a little obscure. Still, it's essential for interpreting results from a multiple regression analysis.

Every computer program that does multiple regression will automatically calculate and report a standard error for each regression coefficient. For our income example, the coefficient for age is $600, and its standard error is $210. For the schooling coefficient ($2,057), the standard error is $849. To get a 95% confidence interval, we follow the same procedure used to construct confidence intervals around a mean: We add two standard errors to the coefficient, and then we subtract two standard errors from the coefficient. (For these data, a more precise multiplier is 2.037, but 2 is close enough unless the sample is quite small.) For age, we have $600 + (2 \times 210) = 1{,}020$ and $600 - (2 \times 210) = 180$. We then say that we are 95% confident that the true coefficient lies somewhere between $180 and $1,020. Similar calculations for the schooling coefficient produce a 95% confidence interval of $359 to $3,755.

In published research using regression analysis, you're more likely to see hypothesis tests than confidence intervals. Usually, the kind of question people most want answered is "Does this particular variable really affect the dependent variable?" If a variable has no effect, then its true coefficient is zero. When you calculate a multiple regression, you virtually never get a coefficient that's exactly zero, though you may find some that are very small. Small coefficients could easily be produced by the three kinds of random error mentioned earlier, even when the true coefficient is zero. But how small is very small, and how big does an estimated coefficient have to be for us to conclude that the true coefficient is something other than zero? That's the job of hypothesis tests—to tell us whether nonzero coefficients could have been produced by random error.

Actually, what the tests give us is not a simple yes or no answer but a *probability* or *p value*. If the *p* value is small, it's taken as evidence that the coefficient is not zero. The test is calculated by first dividing each coefficient by its standard error, producing something called a *t* statistic. Then you consult a *t* table (or the computer does this for you) to calculate the associated *p* value. Most multiple regression programs do these calculations for you, but occasionally you may find one that only reports the coefficient and its standard error, in which case you'll have to determine the *p* value yourself by referring to a *t* table, which is displayed and explained in any introductory statistics text.

For our income example, if you divide the coefficient for age by its standard error (600 / 210), you get a *t* statistic of 2.86. This has an associated *p* value of .007, which has the following interpretation: If

the true coefficient for age is 0, the probability of estimating a coefficient larger than 600 (or smaller than −600) would be about .007, or about one in 150. For schooling, the *t* statistic has a value of 2.42, with an associated *p* value of .02. Again, we can say that if the true coefficient for schooling is zero, the probability of estimating a coefficient this much different from zero is .02. In general, the smaller the *p* value, the stronger the evidence that the coefficient is *not* zero.

Are these *p* values small enough to conclude that the true coefficients for age and schooling are something other than zero? There's no absolute rule about this: It depends on the situation, especially on the cost of making an error of one sort or another. Because these costs are difficult to quantify, social researchers usually don't think much about this. Instead, they rely on the customary standards of .05 and .01. If the *p* value is less than .05, we say that the coefficient is "significantly different from zero" and conclude that there is evidence for a nonzero coefficient. If the *p* value is less than .01, we say that the coefficient is "highly significant" and conclude that there is strong evidence for a nonzero coefficient. Even though there's probably too much reliance on these rote standards, you won't go far wrong by using them.

1.13. How Does Multiple Regression "Control" for Variables?

It's easy to see why multiple regression might be good for making predictions: It's explicitly designed to make errors of prediction as small as possible (using the least squares criterion for overall smallness). Earlier, I claimed that another major use of multiple regression is to examine the effects of some independent variables on the dependent variable while "controlling" for other independent variables. In our income example, the coefficient for years of schooling can be interpreted as the effect of schooling on income while controlling for age or "holding age constant." Similarly, the coefficient for age can be interpreted as the effect of age on income while controlling for years of schooling.

In what sense does multiple regression control for variables, and how does it do it? This is a complex and subtle question, and I can only hope to scratch the surface in this chapter. Many practicing

researchers just accept it on faith that multiple regression controls for variables. The issue is also somewhat controversial. There are some statisticians who take the conservative position that only a randomized experiment can *really* control for extraneous variables, and that multiple regression is, in most cases, only a poor approximation.

With that in mind, let's first see how an experiment controls for extraneous variables. Suppose, for example, that we want to evaluate the effectiveness of a training course to raise SAT scores (the dependent variable). We give the course to some people and withhold it from others. Then we compare how well they do on the test. For this to be a valid experiment, we have to make sure that people in the two groups are treated exactly the same, except for the training course. For example, it wouldn't be a fair comparison if people who didn't get the training took the SAT exam in a hot, noisy room while those who got the training took their exams in a comfortable room.

All the conditions that affect performance must be equalized for the two groups. Even if we were successful at doing that, there's still the problem that people in the two groups might be different. Maybe those who got the training are smarter than those who didn't. We certainly can't force people to be equally intelligent, but what we *can* do is randomly assign people to one group or the other. For each person recruited into the study, we could flip a coin to decide if they should be in the treatment group or the control group. With random assignment, the two groups will, on average, be the same on all possible characteristics: age, sex, race, intelligence, ambition, anxiety, and so on.

With observational (nonexperimental) studies, we don't have those options for ensuring that our comparison groups are comparable. Suppose, for example, that we do a survey of 10,000 college freshmen and we ask them for their SAT scores. We also ask them if they ever took an SAT training course. At this point, the damage is already done. Those who took SAT training may be very different in other respects from those who did not. A simple comparison of average SAT scores for those who did and did not take SAT training could easily give us very misleading results.

The situation is not hopeless, however, because we can still do *statistical* controls. For example, there is plenty of evidence that people with a strong high school grade point average (GPA) do better on the SAT than those with a lower high school GPA. If people

with high GPAs were more likely to take SAT training, it could look like people with the training did better even though the training itself had no effect. If we know the students' high school GPAs, we can restrict our comparisons to those who have the same, or nearly the same, GPA. Suppose, for example, that our sample contained 500 students with high school GPAs that were exactly 3.00, and that 100 of these students took an SAT training course. Then, if we compare the SAT scores of these 100 students with the 400 students who did not get SAT training, we could be reasonably confident that our comparison was not contaminated by differences in high school GPA.

We could do the same sort of statistical control for any other variable. We know, for example, that males tend to do better than females on the math component of the SAT. For a valid comparison of SAT training versus no training, therefore, we should restrict our sample to males only or females only.

Although this is a very useful method, it has several rather obvious limitations. First, unless the sample is very large, it may be difficult to find a substantial number of people who are identical, or even very similar, on the control variable. It may not be too hard for a variable like gender that has only two values, but it can be very difficult for variables measured on a continuum, like GPA or income. In our SAT study, for example, if we only had 500 cases to start with, we might find only 10 people who had GPAs of exactly 3.00. Of those 10, perhaps only 2 got SAT training. That's hardly enough cases to make a reliable comparison. Instead of requiring that people have *exactly* the same GPAs, we could instead take all people who have, say, between 3.0 and 3.5, but that reduces the effectiveness of the control and could still allow for some contaminating effect of GPA differences on SAT performance.

The second problem is an extension of the first. If we only control for sex, the SAT scores could still be contaminated by differences in GPA. If we only control for GPA, the results could be due to sex differences. It's not good enough to control one variable at a time. We really need to control simultaneously for gender, GPA, socioeconomic status, and everything else that might affect SAT performance. To do this statistically, we would need to find a group of people who have the same gender, the same GPA, the same socioeconomic status, and so on. Even if we had a sample of 1 million college

freshmen, it would be hard to isolate any group that was exactly alike on all these variables.

There is a third problem with this method of statistical control, but one that can also be seen as an advantage. If we restrict our sample to those with GPAs of 3.0, we might find only a trivial difference between the average SAT scores of those who got training and those who did not. If we look only at those with GPAs of 2.0, we might find a substantial difference in the SAT performance of those with and without training. Now maybe this simply reflects reality. It's quite plausible that SAT training might be more useful for poorer students than for better students. On the other hand, it could also be just random variation, in which case our interpretation of the results has been needlessly complicated. If we divide the sample into many different groups, with the people in each group being approximately the same on the control variables, we may end up with a very complicated pattern of results. Fortunately, this problem has a simple solution: Within each group of people who are nearly the same on the control variables, we compute the "effect" of SAT training (the difference between the average SAT scores of those who did and those who did not get training); then we average those effects across the groups (possibly weighting by the size of the group). Thus, if men show an average gain of 20 points from SAT training and women show an average gain of 10 points from SAT training, and if there are equal numbers of men and women, the overall estimate for the effectiveness of training would be a 15 point gain. Of course, you may not want to compute such an average if you have reason to think that the effects are really different for men and women.

What does this long-winded digression have to do with multiple regression? Well, multiple regression can be seen as just an extension of the basic logic of statistical control, but one that solves the three problems discussed above. It enables us to control for variables like GPA even though no two people in the sample have exactly the same GPA. It allows for the simultaneous control of many variables even though no two people are exactly alike on all the variables. And it only gives us a single estimate for the "effect" of each variable, which is analogous to the weighted average of effects in different subgroups. Exactly how it does all these things is beyond the scope of this chapter. See Chapter 5 for more details.

1.14. Is Multiple Regression as Good as an Experiment?

Compared to the cruder methods of statistical control (finding homogeneous subgroups and making comparisons within subgroups), multiple regression has clear advantages. Those advantages are purchased at some cost, however. To solve the three problems of the crude methods, we have to make some assumptions about the form of the relationship between the independent variables and the dependent variable. Specifically, as we saw earlier in the chapter, we have to assume that those relationships can be described, at least approximately, by a linear equation. If that assumption is incorrect, multiple regression could give us misleading results.

Multiple regression shares an additional problem with *all* methods of statistical control, a problem that is the major focus of those who claim that multiple regression will never be a good substitute for the randomized experiment. To statistically control for a variable, you have to be able to *measure* that variable so that you can explicitly build it into the data analysis, either by putting it in the regression equation or by using it to form homogeneous subgroups. Unfortunately, there's no way that we can measure all the variables that might conceivably affect the dependent variable. No matter how many variables we include in a regression equation, someone can always come along and say, "Yes, but you neglected to control for variable X and I feel certain that your results would have been different if you had done so."

That's not the case with randomization in an experimental setting. Randomization controls for *all* characteristics of the experimental subjects, regardless of whether those characteristics can be measured. Thus, with randomization there's no need to worry about whether those in the treatment group are smarter, more popular, more achievement oriented, or more alienated than those in the control group (assuming, of course, that there are enough subjects in the experiment to allow randomization to do its job effectively).

There's a more subtle aspect to this problem of statistical control: It's not enough to be able to measure all the variables that we want to control. We also have to measure them *well*. That means that if two people get the same score on some variable, they should really be the same on the underlying characteristic that we're trying to measure. If they're not the same, then we're not really holding that

variable constant when we include it in a regression model or create what we think are homogeneous subgroups. That may not be a serious problem when we're dealing with variables like gender or age (based on official records), but there are lots of "fuzzy" variables in the social sciences that we can measure only crudely, at best, among them intelligence, depression, need for achievement, marital conflict, and job satisfaction. Moreover, even those variables that we can measure precisely are often only "proxies" for variables that are much more subtle and difficult to measure. Thus, gender may be a proxy for cumulative differences in socialization between men and women.

Of course, the quality of measurement is always a matter of degree. No variable is ever measured perfectly, but some variables are measured much more accurately than others. As the quality of the measurement gets worse, the effectiveness of statistical controls deteriorates.

Does this mean, as some critics claim, that multiple regression is worthless for drawing conclusions about causal relationships? I think that's much too strong a reaction to these problems. Randomized experiments have been around only for the last century, but human beings have been making causal inferences from nonexperimental data for as long as there have been human beings. Although there have been plenty of mistaken conclusions, there have also been lots of valid conclusions. Multiple regression (and other forms of statistical control) can be seen as ways of improving on the informal and intuitive kinds of causal reasoning that go on in everyday life. There are simply too many areas in which randomized experiments are infeasible or unethical for us to reject nonexperimental data as a source of causal inference.

The most important component of any causal reasoning is the process of ruling out alternative explanations. Multiple regression is certainly very helpful in this process. Maybe it can't rule out *all* alternative explanations, but science—like life itself—is a matter of incremental improvements (punctuated by occasional radical leaps forward). When critics come up with a persuasive argument as to why some particular relationship might be spurious, it then becomes the task of the researcher to measure that potentially confounding variable and include it in the regression model. There *are* limits to the number of persuasive counterarguments that critics come up with.

Chapter Highlights

1. Multiple regression is used both for predicting outcomes and for investigating the causes of outcomes.
2. The most popular kind of regression is ordinary least squares, but there are other, more complicated regression methods.
3. Ordinary multiple regression is called linear because it can be represented graphically by a straight line.
4. A linear relationship between two variables is usually described by two numbers, the slope and the intercept.
5. Researchers typically assume that relationships are linear because it's the simplest kind of relationship and there's usually no good reason to consider something more complicated.
6. To do a regression, you need more cases than variables, ideally lots more.
7. Ordinal variables are not well represented by linear regression equations.
8. Ordinary least squares chooses the regression coefficients (slopes and intercept) to minimize the sum of the squared prediction errors.
9. The R^2 is the statistic most often used to measure how well the dependent variable can be predicted from knowledge of the independent variables.
10. To evaluate the least squares estimates of the regression coefficients, we usually rely on confidence intervals and hypothesis tests.
11. Multiple regression allows us to statistically control for measured variables, but this control is never as good as a randomized experiment.

Questions to Think About

1. For a sample of 13 hospitals, a researcher measured 100 different variables describing each hospital. How many of these variables can be put in a regression equation?
2. Which is more important in describing the relationship between two variables, the slope or the intercept?

3. Suppose you want to use regression to describe the relationship between people's age and how many hours a week they watch television. Which one should be the dependent variable and which one the independent variable? Is the relationship likely to be linear?
4. For a sample of 200 U.S. cities, a linear regression is estimated with percentage of people unemployed as the dependent variable and the percentage foreign born as the independent variable. The regression slope is .20. How would you interpret this number?
5. A researcher in a college admissions department runs a regression to predict college GPA based on information available on students' applications for admission (e.g., SAT scores, high school GPA, number of advanced placement courses). The R^2 for this regression is .15. Do you think this regression model would be useful for admission decisions?
6. Based on survey data, a psychologist runs a regression in which the dependent variable is a measure of depression and independent variables include marital status, employment status, income, gender, and body mass index (weight / height^2). He finds that people with higher body mass index are significantly more depressed, controlling for the other variables. Has he proven that being overweight causes depression? Why or why not?