CHAPTER 5

Understanding What to Avoid

In chapter 4, we discussed how to construct a study with a determinate research design in which observation selection procedures make valid inferences possible. Carrying out this task successfully is necessary but not sufficient if we are to make valid inferences: analytical errors later in the research process can destroy the good work we have done earlier. In this chapter, we discuss how, once we have selected observations for analysis, we can understand sources of inefficiency and bias and reduce them to manageable proportions. We will then consider how we can control the research in such a way as to deal effectively with these problems.

In discussing inefficiency and bias, let us recall our criteria that we introduced in sections 2.7 and 3.4 for judging inferences. If we have a determinate research design, we then need to concern ourselves with the two key problems that we will discuss in this chapter: *bias* and *inefficiency*. To understand these concepts, it is useful to think of any inference as an estimate of a particular point with an interval around it. For example, we might guess someone's age as forty years, plus or minus two years. Forty years is our best guess (the estimate) and the interval from thirty-eight to forty-two includes our best guess at the center, with an estimate of our uncertainty (the width of the interval). We wish to choose the interval so that the true age falls within it a large proportion of the time. *Unbiasedness refers to centering the interval around the right estimate whereas efficiency refers to narrowing an appropriately centered interval.*

These definitions of unbiasedness and efficiency apply regardless of whether we are seeking to make a descriptive inference, as in the example about age or a causal inference. If we were, for instance, to estimate the effect of education on income (the number of dollars in income received for each additional year of education), we would have a point estimate of the effect surrounded by an interval reflecting our uncertainty as to the exact amount. We would want an interval as narrow as possible (for efficiency) and centered around the right estimate (for unbiasedness). We also want the estimate of the width of the interval to be an honest representation of our uncertainty.

In this chapter, we focus on four sources of bias and inefficiency, beginning with the stage of research at which we seek to improve the

quality of information and proceeding through the making of causal inferences. In section 5.1, we discuss measurement error, which can bias our results as well as make them less efficient. We then consider in section 5.2 the bias in our causal inferences that can result when we have omitted explanatory variables that we should have included in the analysis. In section 5.3 we take up the inverse problem: controlling for irrelevant variables that reduce the efficiency of our analysis. Finally, we study the problem that results when our "dependent" variable affects our "explanatory" variables. This problem is known as endogeneity and is introduced in section 5.4. Finally, in sections 5.5 and 5.6 we discuss, respectively, random assignment of values of the explanatory variables and various methods of nonexperimental control.

5.1 Measurement Error

Once we have selected our observations, we have to measure the values of variables in which we are interested. Since all observation and measurement in the social sciences is imprecise, we are immediately confronted with issues of measurement error.

Much analysis in social science research attempts to estimate the amount of error and to reduce it as much as possible. Quantitative research produces more precise (numerical) measures, but not necessarily more accurate ones. Reliability—different measurements of the same phenomenon yield the same results—is sometimes purchased at the expense of validity—the measurements reflect what the investigator is trying to measure. Qualitative researchers try to achieve accurate measures, but they generally have somewhat less precision.

Quantitative measurement and qualitative observation are in essential respects very similar. To be sure, qualitative researchers typically label their categories with words, whereas quantitative researchers assign numerical values to their categories and measures. But both quantitative and qualitative researchers use nominal, ordinal, and interval measurements. With nominal categories, observations are grouped into a set of categories without the assumption that the categories are in any particular order. The relevant categories may be based on legal or institutional forms; for instance, students of comparative politics may be interested in patterns of presidential, parliamentary, and authoritarian rule across countries. Ordinal categories divide phenomena according to some ordering scheme. For example, a qualitative researcher might divide nations into three or four categories according to their degree of industrialization or the size of their military forces. Finally, interval measurement uses continuous variables, as in studies of transaction flows across national borders.

The differences between quantitative and qualitative measurement involve how data are represented, not the theoretical status of measurement. Qualitative researchers use words like "more" or "less," "larger" or "smaller," and "strong" or "weak" for measurements; quantitative researchers use numbers.

For example, most qualitative researchers in international relations are acutely aware that "number of battle deaths" is not necessarily a good index of how significant wars are for subsequent patterns of world politics. In balance-of-power theory, not the severity of war but a "consequential" change in the major actors is viewed as the relevant theoretical concept of instability to be measured (see Gulick 1967 and Waltz 1979:162). Yet in avoiding invalidity, the qualitative researcher often risks unreliability due to measurement error. How are we to know what counts as "consequential," if that term is not precisely defined? Indeed, the very language seems to imply that such a judgment will be made depending on the systemic outcome—which would bias subsequent estimates of the relationship in the direction of the hypothesis.

No formula can specify the tradeoffs between using quantitative indicators that may not validly reflect the underlying concepts in which we are interested, or qualitative judgments that are inherently imprecise and subject to unconscious biases. But both kinds of researchers should provide estimates of the uncertainty of their inferences. Quantitative researchers should provide standard errors along with their numerical measurements; qualitative researchers should offer uncertainty estimates in the form of carefully worded judgments about their observations. The difference between quantitative and qualitative measurement is in the style of representation of essentially the same ideas.

Qualitative and quantitative measurements are similar in another way. For each, the categories or measures used are usually artifacts created by the investigator and are not "given" in nature. The division of nations into democratic and autocratic regimes or into parliamentary and presidential regimes depends on categories that are intellectual constructs, as does the ordering of nations along such dimensions as more or less industrialized.

Obviously, a universally right answer does not exist: all measurement depends on the problem that the investigator seeks to understand. The closer the categorical scheme is to the investigator's original theoretical and empirical ideas, the better; however, this very fact emphasizes the point that the categories are artifacts of the investigator's purposes. The number of parliamentary regimes in which proportional representation is the principal system of representation depends on the investigator's classification of "parliamentary regimes" and of

what counts as a system of proportional representation. Researchers in international relations may seek to study recorded monetary flows across national borders, but their use of a continuous measure depends on decisions as to what kinds of transactions to count, on rules as to what constitutes a single transaction, and on definitions of national borders. Similarly, the proportion of the vote that is Democratic in a Congressional district is based on classifications made by the analyst assuming that the "Democratic" and "Republican" party labels have the same meaning, for his or her purposes, across all 435 congressional districts.

Even the categorization schemes we have used in this section for measurements (nominal, ordinal, and interval) depend upon the theoretical purpose for which a measure is used. For example, it might seem obvious that ethnicity is a prototypical nominal variable, which might be coded in the United States as black, white, Latino, Native American and Asian-American. However, there is great variation across nominal ethnic groups in how strongly members of such groups identify with their particular group. We could, therefore, categorize ethnic groups on an ordinal scale in terms of, for example, the proportion of a group's members who strongly identify with it. Or we might be interested in the size of an ethnic group, in which case ethnicity might be used as an interval-level measure. The key point is to *use the measure that is most appropriate to our theoretical purposes*.

Problems in measurement occur most often when we measure without explicit reference to any theoretical structure. For example, researchers sometimes take a naturally continuous variable that could be measured well, such as age, and categorize it into young, middle-aged, and old. For some purposes, these categories might be sufficient, but as a theoretical representation of a person's age, this is an unnecessarily imprecise procedure. The *grouping error* created here would be quite substantial and should be avoided. Avoiding grouping error is a special case of the principle: do not discard data unnecessarily.

However, we can make the opposite mistake—assigning continuous, interval-level numerical values to naturally discrete variables. Interval-level measurement is *not* generally better than ordinal or nominal measurement. For example, a survey question might ask for religious affiliation and also intensity of religious commitment. Intensity of religious commitment could—if the questions are asked properly—be measured as an ordinal variable, maybe even an interval one, depending on the nature of the measuring instrument. But it would make less sense to assign a numerical ranking to the particular religion to which an individual belonged. In such a case, an ordinal or continuous variable probably does not exist and measurement error would be created by such a procedure.

The choice between nominal categories, on one hand, and ordinal or interval ones, on the other, may involve a tradeoff between descriptive richness and facilitation of comparison. For example, consider the voting rules used by international organizations. The institutional rule governing voting is important because it reflects conceptions of state sovereignty, and because it has implications for the types of resolutions that can pass, for resources allocated to the organization, and for expectations of compliance with the organization's mandates.

A set of nominal categories could distinguish among systems in which a single member can veto any resolution (as in the League of Nations Council acting under the provisions of Article 15 of the Covenant); in which only certain members can veto resolutions (as in the Security Council of the United Nations); in which some form of supermajority voting prevails (as in decisions concerning the internal market of the European Community); and in which simple majority voting is the rule (as for many votes in the United Nations General Assembly). Each of these systems is likely to generate distinct bargaining dynamics, and if our purpose is to study the dynamics of one such system (such as a system in which any member can exercise a veto), it is essential to have our categories defined, so that we do not inappropriately include other types of systems in our analysis. Nominal categories would be appropriate for such a project.

However, we could also view these categories in an ordinal way, from most restrictive (unanimity required) to least (simple majority). Such a categorization would be necessary were we to test theoretical propositions about the relationship between the restrictiveness of a voting rule and patterns of bargaining or the distributive features of typical outcomes. However, at least two of our categories—vetoes by certain members and qualified majority voting—are rather indistinct because they include a range of different arrangements. The first category includes complete veto by only one member, which verges on dictatorship, and veto by all but a few inconsequential members; the second includes the rule in the European Community that prevents any two states from having a blocking minority on issues involving the internal market. The formula used in the International Monetary Fund is nominally a case of qualified majority voting, but it gives such a blocking minority both to the United States and, recently, to the European Community acting as a bloc. Hence, it seems to belong in both of these categories.

We might, therefore, wish to go a step further to generate an interval-level measure based on the proportion of states (or the proportion of resources, based on gross national product, contributions to the organization, or population represented by states) required for passage

of resolutions, measuring international organizations on a scale of voting restrictiveness.

However, different bases for such a measure—for example, whether population or gross national product were used as the measure of resources—would generate different results. Hence, the advantages of precision in such measurements might be countered by the liabilities either of arbitrariness in the basis for measurement or of the complexity of aggregate measures. Each category has advantages and limitations: the researcher's purpose must determine the choice that is made.

In the following two subsections, we will analyze the specific consequences of measurement error for qualitative research and reach some conclusions that may seem surprising. Few would disagree that *systematic* measurement error, such as a consistent overestimate of certain units, causes bias and, since the bias does not disappear with more error-laden observations, inconsistency. However, a closer analysis shows that only some types of systematic measurement error will bias our causal inferences. In addition, the consequences of *nonsystematic* measurement error may be less clear. We will discuss nonsystematic measurement error in two parts: in the dependent variable and then in the explanatory variable. As we will demonstrate, error in the dependent variable causes inefficiencies, which are likely to produce incorrect results in any one instance and make it difficult to find persistent evidence of systematic effects. In other words, nonsystematic measurement error in the dependent variable causes no bias but can increase inefficiency substantially. More interesting is nonsystematic error in the key causal variable, which unfailingly biases inferences in predictable ways. Understanding the nature of these biases will help ameliorate or possibly avoid them.

5.1.1 Systematic Measurement Error

In this section, we address the consequences of *systematic* measurement error. Systematic measurement error, such as a measure being a consistent overestimate for certain types of units, can sometimes cause bias and inconsistency in estimating causal effects. Our task is to find out what types of systematic measurement error result in which types of bias. In both quantitative and qualitative research, systematic error can derive from choices on the part of researchers that slant the data in favor of the researcher's prior expectations. In quantitative work, the researcher may use such biased data because it is the only numerical series available. In qualitative research, systematic measurement error can result from subjective evaluations made by investigators who have

already formed their hypotheses and who wish to demonstrate their correctness.

It should be obvious that *any systematic measurement error will bias descriptive inferences*.[1] Consider, for example, the simplest possible case in which we inadvertently overestimate the amount of annual income of every survey respondent by $1,000. Our estimate of the average annual income for the whole sample will obviously be overestimated by the same figure. If we were interested in estimating the causal effect of a college education on average annual income, the systematic measurement error would have no effect on our causal inference. If, for example, our college group really earns $30,000 on average, but our control group of people who did not go to college earn an average of $25,000, our estimate of the causal effect of a college education on annual income would be $5,000. If the income of every person in both groups was overestimated by the same amount (say $1,000 again), then our causal effect—now calculated as the difference between $31,000 and $26,000—would still be $5,000. Thus, *systematic measurement error which affects all units by the same constant amount causes no bias in causal inference*. (This is easiest to see by focusing on the constant effects version of the unit homogeneity assumption described in section 3.3.1.)

However, suppose there is a systematic error in one part of the sample: college graduates systematically overreport their income because they want to impress the interviewer, but the control group reports its income more accurately. In this case, both the descriptive inference *and* our inference about the causal effect of education on income would be biased. If we knew of the reporting problem, we might be able to ask better survey questions or elicit the information in other ways. If the information has already been collected and we have no opportunity to collect more, then we may at least be able to ascertain the direction of the bias to make a post hoc correction.

To reinforce this point, consider an example from the literature on regional integration in international relations. That literature sought, more than most work in international relations, to test specific hypotheses, sometimes with quantitative indicators. However, one of the most important concepts in the literature—the degree to which policy authority is transferred to an international organization from nation-states—is not easily amenable to valid quantitative measurement. Researchers therefore devised qualitative measurements of this variable, which they coded on the basis of their own detailed knowledge of

[1] An exception is when positive systematic errors cancel out negative systematic ones, but this odd case is more properly described as a type of nonsystematic measurement error.

the issues involved (e.g., Lindberg and Scheingold 1970:71, table 3.1). Their explanatory variables included subjective categorizations of such variables as "elite value complementarity" and "decision-making style" (see Nye 1971 or Lindberg and Sheingold 1971). They tried to examine associations between the explanatory and dependent variables, when the variables were measured in this manner.

This approach was a response to concerns about validity: expert researchers coded the information and could examine whether it was relevant to the concepts underlying their measurements. But the approach ran the risk of subjective measurement error. The researchers had to exercise great self-discipline in the process and refrain from coding their explanatory variables in light of their theoretical positions or expectations. In any given case, they may have done so, but it is difficult for their readers to know to what extent they were successful.

Our advice in these circumstances is, first, to try to use judgments made for entirely different purposes by *other researchers*. This element of arbitrariness in qualitative or quantitative measurement guarantees that the measures will not be influenced by your hypotheses, which presumably were not formed until later. This strategy is frequently followed in quantitative research—a researcher takes someone else's measures and applies them to his or her own purposes—but it is also an excellent strategy in qualitative research. For example, it may be possible to organize joint coding of key variables by informed observers with different preferred interpretations and explanations of the phenomena. Qualitative data banks having standard categories may be constructed on the basis of shared expertise and discussion. They can then be used for evaluating hypotheses. If you are the first person to use a set of variables, it is helpful to let *other informed people* code your variables without knowing your theory of the relationship you wish to evaluate. Show them your field notes and taped interviews, and see if their conclusions about measures are the same as yours. Since replicability in coding increases confidence in qualitative variables, the more highly qualified observers who cross-check your measures, the better.

5.1.2 Nonsystematic Measurement Error

Nonsystematic measurement error, whether quantitative or qualitative, is another problem faced by all researchers.[2] Nonsystematic error does not bias the variable's measurement. In the present context, we

[2] Whether this is due to our inability to measure the real world accurately or due to randomness in nature is a philosophical question to which different answers can be given. (section 2.6). Whichever position we accept, the consequence is the same.

define variables with nonsystematic, or random, measurement error as having values that are sometimes too high and sometimes too low, but correct on average. Random error obviously creates inefficiencies but not bias in making descriptive inferences. This point has already been discussed in section 2.7.1. Here, we go beyond the consequence of random measurement error for descriptive inference to its conseqence for causal inference.

In the estimation of causal effects, random measurement error has a different effect when the error is in an explanatory variable than when the error is in the dependent variable. Random measurement error in the dependent variable reduces the efficiency of the causal estimate but does not bias it. It can lead to estimates of causal relationships that are at times too high and at times too low. However, the estimate will be, on average, correct. Indeed, random measurement error in a dependent variable is not different or even generally distinguishable from the usual random error present in the world as reflected in the dependent variable.

Random error in an explanatory variable can also produce inefficiencies that lead to estimates that are uncertainly high or low. But it also has an effect very different from random error in the dependent variable: random error in an explanatory variable produces bias in the estimate of the relationship between the explanatory and the dependent variable. That bias takes a particular form: it results in the estimation of a weaker causal relationship than is the case. If the true relationship is positive, random error in the explanatory variable will bias the estimate downwards towards a smaller or zero relationship. If the relationship is negative it will bias the relationship upwards towards zero.

Since this difference between the effect of random error in an explanatory variable and random error in a dependent variable is not intuitively obvious, we present formal proofs of each effect as well as a graphic presentation and an illustrative example. We begin with the effect of random error in a dependent variable.

5.1.2.1 NONSYSTEMATIC MEASUREMENT ERROR IN THE DEPENDENT VARIABLE

Nonsystematic or random measurement error in a dependent variable does not bias the usual estimate of the causal effect, but it does make the estimate less efficient. In any one application, this inefficiency will yield unpredictable results, sometimes giving causal inferences that are too large and sometimes too small. Measurement error in the dependent variable thus increases the uncertainty of our inferences. In other words, random measurement error in a dependent

variable creates a problem similar to that created by a small number of observations; in both cases, the amount of information we can bring to bear on a problem is less than we would like. The result is that *random measurement error in the dependent variable produces estimates of causal effects that are less efficient and more uncertain.*

When we use several data sets, as we should when feasible, estimates based on dependent variables with random measurement error will be unstable. Some data sets will produce evidence of strong relationships while others will yield nonexistent or negative effects, even if the true relationship has not changed at all. This inefficiency makes it harder, sometimes considerably harder, to find systematic descriptive or causal features in one data set or (perhaps more obviously) across different data sets. Estimates of uncertainty will often be larger than the estimated size of relationships among our variables. Thus, we may have insufficient information to conclude that a causal effect exists when it may actually be present but masked by random error in the dependent variable (and represented in increased uncertainty of an inference). Qualitative and quantitative researchers who are aware of this general result will have no additional tools to deal with measurement error—except a stronger impetus to improve the measurements of the observations they have or collect new observations with the same (or lower) levels of measurement error. Understanding these results with a fixed amount of data will enable scholars to more appropriately qualify their conclusions. Such an explicit recognition of uncertainty may motivate these investigators or others to conduct follow-up studies with more carefully measured dependent variables (or with larger numbers of observations). It should be of even more help in designing research, since scholars frequently face a trade-off between attaining additional precision for each measurement and obtaining more observations. The goal is more information relevant to our hypothesis: we need to make judgments as to whether this information can best be obtained by more observations within existing cases or collecting more data.

Consider the following example of random measurement error in the dependent variuable. In studying the effects of economic performance on violent crime in developing countries or across the regions of a single developing country, we may measure the dependent variable (illegal violence) by observing each community for a short period of time. Of course, these observations will be relatively poor measurements: correct on average, but, in some communities, we will miss much crime and underestimate the average violence; in other communities, we will see a lot of crime and will overestimate average violence.

Suppose our measurement of our explanatory variable—the state of

the economy—is the percentage unemployed in the community and we measure that quite well (perhaps from good government data). If we studied the effect of the economy as indicated by the percentage unemployed on the average amount of violent crime, we would expect very uncertain results—results that are also unstable across several applications—precisely because the dependent variable was measured imperfectly, even though the measurement technique was correct on average. Our awareness that this was the source of the problem, combined with a continuing belief that there should be a strong relationship, provides a good justification for a new study in which we might observe community crime at more sites or for longer periods of time. Once again, we see that measurement error and few observations lead to similar problems. We could improve efficiency either by increasing the accuracy of our observations (perhaps by using good police records and, thus, reducing measurement error) or by increasing the number of imperfectly measured observations in different communities. In either case, the solution is to increase the amount of information that we bring to bear on this inference problem. This is another example of why the amount of *information* we bring to bear on a problem is more important than the raw number of observations we have (the number of observations being our measure of information).

To show why this is the case, we use a simplified version of this example first in a graphic presentation and then offer a more formal proof. In figure 5.1, the horizontal axis represents unemployment. We imagine that the two categories ("4 percent" and "7 percent") are perfectly measured. The vertical axis is a measure of violent crime.

In figure 5.1, the two solid circles can be viewed as representing an example of a simple study with no measurement error in either variable. We can imagine that we have a large number of observations, all of which happen to fall exactly on the two solid dots, so that we know the position of each dot quite well. Alternatively, we can imagine that we have only two observations, but they have very little nonsystematic error of any kind. Of course, neither of these cases will likely occur in reality, but this model highlights the essential problems of measurement error in a dependent variable for the more general and complicated case. Note how the solid line fits these two points.

Now imagine another study where violent crime was measured with nonsystematic error. To emphasize that these measures are correct on average, we plot the four open circles, each symmetrically above and below the original solid circles.[3] A new line fit to all six data

[3] We imagine again that the open circles are either a large number of observations that happen to fall exactly on these four points or that there happens to be little stochastic variability.

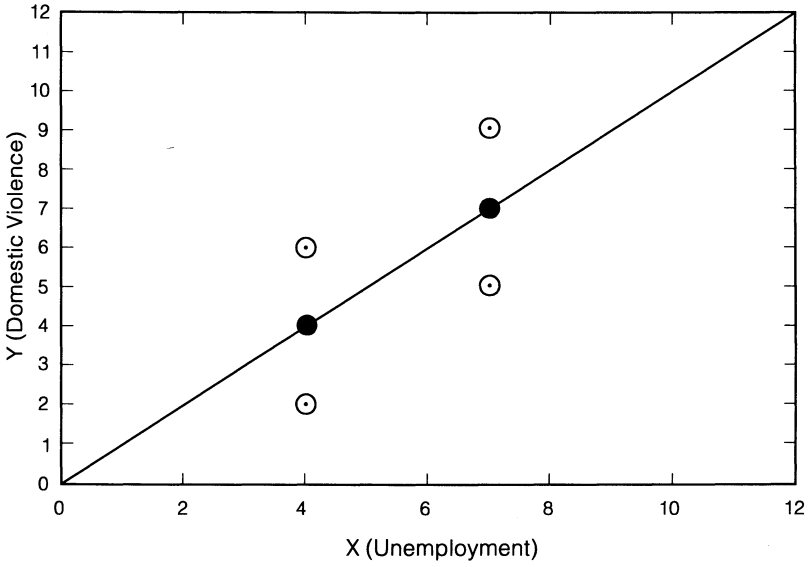

Figure 5.1 Measurement Error in the Dependent Variable

points is exactly the same line as originally plotted. Note again that this line is drawn by minimizing the prediction errors, the *vertical* deviations from the line.

However, the new line is more uncertain in several ways. For example, a line with a moderately steeper or flatter slope would fit these points almost as well. In addition, the vertical position of the line is also more uncertain, and the line itself provides worse predictions of where the individual data points should lie. The result is that measurement error in the dependent variable produces more inefficient estimates. Even though they are still unbiased—that is, on average across numerous similar studies—they might be far off in any one study.

A Formal Analysis of Measurement Error in y. Consider a simple linear model with a dependent variable measured with error and one errorless explanatory variable. We are interested in estimating the effect parameter β:

$$E(Y^*) = \beta X$$

We also specify a second feature of the random variables, the variance:

$$V(Y_i^*) = \sigma^2$$

which we assume to be the same for all units $i = 1, \ldots, n$.[4]

Although these equations define our model, we unfortunately do not observe Y^* but instead Y, where

$$Y = Y^* + U$$

That is, the observed dependent variable Y is equal to the true dependent variable Y^* plus some random measurement error U. To formalize the idea that U contains only *nonsystematic* measurement error, we require that the error cancels on average across hypothetical replications, $E(U) = 0$, and that it is uncorrelated with the true dependent variable, $C(U,Y^*) = 0$, and with the explanatory variable, $C(U,X) = 0$.[5] We further assume that the measurement error has variance $V(U_i) = \tau^2$ for each and every unit i. If τ^2 is zero, Y contains no measurement error and is equal to Y^*; the larger this variance, the more error our measure Y contains.

How does random measurement error in the dependent variable affect one's estimates of β? To see, we use our usual estimator but with Y instead of Y^*:

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

and then calculate the average across hypothetical replications:

$$E(b) = E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right)$$

$$= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2}$$

$$= \frac{\sum_{i=1}^{n} X_i E(Y_i + U)}{\sum_{i=1}^{n} X_i^2}$$

[4] Statistical readers will recognize this as the property of homoskedasticity, or constant variance.

[5] These error assumptions imply that the expected value of the observed dependent variable is the same as the expected value of the true dependent variable:

$$E(Y) = E(Y^* + U) = E(Y^*) + E(U) = E(Y^*) = \beta X$$

$$= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2}$$

$$= \beta$$

This analysis demonstrates that even with measurement error in the dependent variable, the standard estimator will be unbiased (equal to β on average), just as we showed for a dependent variable without measurement error in equation (3.8).

However, to complete this analysis, we must assess the efficiency of our estimator in the presence of a dependent variable measured with error. We use the usual procedure:

$$V(b) = V\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \tag{5.1}$$

$$= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \sum_{i=1}^{n} X_i^2 V(Y_i^* + U)$$

$$= \frac{\sigma^2 + \tau^2}{\sum_{i=1}^{n} X_i^2}$$

Note that this estimator is *less* efficient than the same estimator applied to data without measurement error in the dependent variable (compare equation [3.9]) by the amount of the measurement error in the dependent variable τ^2.

5.1.2.2 NONSYSTEMATIC MEASUREMENT ERROR IN AN EXPLANATORY VARIABLE

As we pointed out above, nonsystematic error in the explanatory variable has the same consequences for estimates of the value of that variable—for descriptive inferences—as it has for estimates of the value of the dependent variable: the measures will sometimes be too high, sometimes too low, but on average they will be right. As with nonsystematic error in the dependent variable, random error in the explanatory variable can also make estimates of causal effects uncertain and inefficient. But the random error in the explanatory variable has another, quite different consequence from the case in which the random error is in the dependent variable. When it is the explanatory

variable that is measured with random error, there is a systematic bias in the estimates of the causal relationship, a bias in the direction of zero or no relationship. In other words, when there is a true causal connection between an explanatory variable and a dependent variable, random error in the former can serve to mask that fact by depressing the relationship. If we were to test our hypothesis across several data sets we would not only find great variation in the results, as with random error in the dependent variable, we would also encounter a systematic bias across the several data sets towards a weaker relationship than is in fact the case.

Just as with measurement error in the dependent variable, even if we recognize the presence of measurement error in the explanatory variables, more carefully analyzing the variables measured with error will not ameliorate the *consequences* of this measurement error unless we follow the advice given here. Better measurements would of course improve the situation.

Consider again our study of the effects of unemployment on crime in various communities of an underdeveloped country. However, suppose the data situation is the opposite of that mentioned above: in the country we are studying, crime reports are accurate and easy to obtain from government offices, but unemployment is a political issue and hence not accurately measurable. Since systematic sample surveys are not permitted, we decide to measure unemployment by direct observation (just as in our earlier example, where we measured crime by direct observation). We infer the rate of unemployment from the number of people standing idle in the center of various villages as we drive through. Since the hour and day when we observe the villages would vary, as would the weather, we would have a lot of random error in our estimates of the degree of unemployment. Across a large number of villages, our estimates would not be systematically high or low. An estimate based on any pair of villages would be quite inefficient: any pair might be based on observations on Sunday (when many people may linger outside) or on a rainy day (when few would). But many observations of pairs of villages at different times on different days, in rain or shine, would produce, on average, correct estimates of the effect. However, as indicated above, the consequence will be very different from the consequence of similar error in our measure of the dependent variable, violent crime.

Figure 5.2 illustrates this situation. The two solid dots represent one study with no measurement error in either variable.[6] The slope of the

[6] We also continue to assume that each point represents data either with almost no stochastic variation or numerous points that happen to fall in the same place. As in section 5.1, the purpose of this assumption is to keep the focus on the problem.

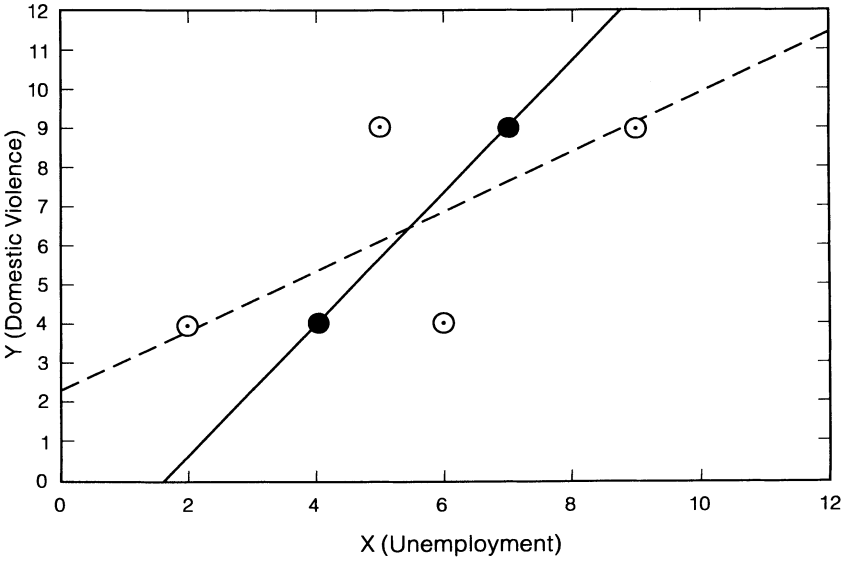

Figure 5.2 Measurement Error in the Explanatory Variable

solid line is then the correct estimate of the causal effect of unemployment on crime. To show the consequences of measurement error, we add two additional points (open circles) to the right and the left of each of the solid dots, to represent measurement error in the explanatory variable that is correct on average (that is, equal to the filled dot on average). The dashed line is fit to the open circles, and the difference between the two lines is the bias due to random measurement error in the explanatory variable. We emphasize again that the lines are drawn so as to minimize the errors in predicting the dependent variable (the errors appear in the figure as *vertical* deviations from the line being fit), given each value of the explanatory variables.

Thus, the estimated effect of unemployment, made here with considerable random measurement error, will be much smaller (since the dashed line is flatter) than the true effect. We could infer from our knowledge of the existence of measurement error in the explanatory variable that the true effect of unemployment on crime is larger than the observed correlation found in this research project.

The analysis of the consequences of measurement error in an explanatory variable leads to two practical guidelines:

1. If an analysis suggests no effect to begin with, then the true effect is difficult to ascertain since the direction of bias is unknown; the analysis will then be largely indeterminate and should be described as such. The true

effect may be zero, negative, or positive, and nothing in the data will provide an indication of which it is.

2. However, if an analysis suggests that the explanatory variable with random measurement error has a small positive effect, then we should use the results in this section as justification for concluding that the true effect is probably even larger than we found. Similarly, if we find a small negative effect, the results in this section can be used as evidence that the true effect is probably an even larger negative relationship.

Since measurement error is a fundamental characteristic of all qualitative research, these guidelines should be widely applicable.

We must qualify these conclusions somewhat so that researchers know exactly when they do and do not apply. First, the analysis in the box below, on which our advice is based, applies to models with only a single explanatory variable. Similar results do apply to many situations with multiple explanatory variables, but not to all. The analysis applies just the same if a researcher has many explanatory variables, but only one with substantial random measurement error. However, if one has multiple explanatory variables and is simultaneously analyzing their effects, and if each has different kinds of measurement error, we can only ascertain the kinds of biases likely to arise by extending the formal analysis below. It turns out that although qualitative researchers often have many explanatory variables, they most frequently study the effect of each variable sequentially rather than simultaneously. Unfortunately, as we describe in section 5.2, this procedure can cause other problems, such as omitted variable bias, but it does mean that results similar to those analyzed here apply quite widely in qualitative research.

A Formal Analysis of Random Measurement Error in *X*. We first define a model as follows:

$$E(Y) = \beta X^*$$

where we do not observe the true explanatory variable X^* but instead observe X where

$$X = X^* + U$$

and the random measurement error U has similar properties as before: it is zero on average, $E(U) = 0$, and is uncorrelated with the true explanatory variable, $C(U,X^*) = 0$, and with the dependent variable, $C(U,Y) = 0$.

What happens when we use the standard estimator for β with the error-ridden X, instead of the unobserved X^*? This situation corresponds to the usual one in qualitative research in which we have measurement error but do not make any special adjustment for the results that follow. To analyze the consequences of this procedure, we evaluate bias, which will turn out to be the primary consequence of this sort of measurement problem. We thus begin with the standard estimator in equation (3.7) applied to the observed X and Y for the model above.

$$
\begin{aligned}
b &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad (5.2)\\
&= \frac{\sum_{i=1}^{n} (X_i^* + U_i) Y_i}{\sum_{i=1}^{n} (X_i^* + U_i)^2}\\
&= \frac{\sum_{i=1}^{n} X_i^* Y_i + \left(\sum_{i=1}^{n} U_i Y_i\right)}{\sum_{i=1}^{n} X_i^{*2} + \sum_{i=1}^{n} U_i^2 + \left(2\sum_{i=1}^{n} X_i^* U_i\right)}
\end{aligned}
$$

It should be clear that b will be biased, $E(b) \neq \beta$. Furthermore, the two parenthetical terms in the last line of equation (5.2) will be zero on average because we have assumed that U and Y, and U and X^*, are uncorrelated (that is, $C(U_i,Y_i) = E(U_i,Y_i) = 0$). This equation therefore reduces to approximately[7]

$$
b \approx \frac{\sum_{i=1}^{n} X_i^* Y_i}{\sum_{i=1}^{n} X_i^{*2} + \sum_{i=1}^{n} U_i^2}
$$

This equation for the estimator of β in the model above is the same as the standard one, except for the extra term in the denominator, $\sum_{i=1}^{n} U_i^2$ (compare equation [3.7]). This term represents the amount of measurement error in X, the sample variance of the error U. In the absence of measurement error, this term is zero, and the equation reduces to the standard estimator in equation (3.7), since we would have actually observed the true values of the explanatory variable.

In the general case with some measurement error, $\sum_{i=1}^{n} U_i^2$ is a sum of squared terms and so will always be positive. Since this term is added to the denominator, b will approach zero. If the correct esti-

[7] Since this equation holds exactly only in large samples, we are really analyzing consistency instead of unbiasedness (section 2.7.1). More precisely, the parenthetical terms in equation (5.2), when divided by n, vanish as n approaches infinity.

mator would produce a large positive number, random measurement error in the explanatory variable would incorrectly cause the researcher to think b was positive but smaller. If the estimate based on X^* were a large negative number, a researcher analyzing data with random measurement error would think the estimate was a smaller negative number.

It would be straightforward to use this formal analysis to show that random measurement error in the explanatory variables also causes inefficiencies, but bias is generally a more serious problem, and we will deal with it first.

5.2 Excluding Relevant Variables: Bias

Most qualitative social scientists appreciate the importance of controlling for the possibly spurious effects of other variables when estimating the effect of one variable on another. Ways to effect this control include, among others, John Stuart Mill's (1843) methods of difference and similarity (which, ironically, are referred to by Przeworski and Teune (1982) as most similar and most different systems designs, respectively), Verba's (1967) "disciplined-configurative case comparisons," (which are similar to George's [1982] "structured-focused comparisons"), and diverse ways of using ceteris paribus assumptions and similar counterfactuals. These phrases are frequently invoked, but researchers often have difficulty applying them effectively. Unfortunately, qualitative researchers have few tools for expressing the precise consequences of failing to take into account additional variables in particular research situations: that is, of "omitted variable bias." We provide these tools in this section.

We begin our discussion of this issue with a verbal analysis of the consequences of omitted variable bias and follow it with a formal analysis of this problem. Then we will turn to broader questions of research design raised by omitted variable bias.

5.2.1 Gauging the Bias from Omitted Variables

Suppose we wish to estimate the causal effect of our explanatory variable X_1 on our dependent variable Y. If we are undertaking a quantitative analysis, we denote this causal effect of X_1 on Y as β_1. One way of estimating β_1 is by running a regression equation or another form of analysis, which yields an estimate b_1 of β_1. If we are carrying out qualitative research, we will also seek to make such an estimate of the

causal effect; however, this estimate will depend on verbal argument and the investigator's assessment, based on experience and judgment.

Suppose that after we have made these estimates (quantitatively or qualitatively) a colleague takes a look at our analysis and objects that we have omitted an important control variable, X_2. We have been estimating the effect of campaign spending on the proportion of the votes received by a congressional candidate. Our colleague conjectures that our finding is spurious due to "omitted variable bias." That is, she suggests that our estimate b_1 of β_1 is incorrect since we have failed to take into account another explanatory variable X_2 (such as a measure of whether or not the candidate is an incumbent). The true model should presumably control for the effect of the new variable.

How are we to evaluate her claim? In particular, under what conditions would our omission of the variable measuring incumbency affect our estimate of the effect of spending on votes and under what conditions would it have no effect? Clearly, the omission of a term measuring incumbency will not matter if incumbency has no effect on the dependent variable; that is, if X_2 is irrelevant, because it has no effect on Y, it will not cause bias. This is the first special case: irrelevant omitted variables cause no bias. Thus, if incumbency had no electoral consequences we could ignore the fact that it was omitted.

The second special case, which also produces no bias, occurs when the omitted variable is uncorrelated with the included explanatory variable. Thus, there is also no bias if incumbency status is uncorrelated with our explanatory variable, campaign spending. Intuitively, when an omitted variable is uncorrelated with the main explanatory variable of interest, controlling for it would not change our estimate of the causal effect of our main variable, since we control for the portion of the variation that the two variables have in common, if any. Thus, *we can safely omit control variables, even if they have a strong influence on the dependent variable, as long as they do not vary with the included explanatory variable*.[8]

[8] Note the difference between the two cases in which omitting a variable is acceptable. In the first case, in which the omitted variable is unrelated to the dependent variable, there is no bias and we lose no power in predicting future values of the dependent variable. In the latter case, in which the omitted variable is unrelated to the independent variable though related to the dependent variable, we have no bias in our estimate of the *relationship* of the included explanatory variable and the dependent variable, but we lose some accuracy in forecasting future values of the dependent variable. Thus, if incumbency were unrelated to campaign spending, omitting it would not bias our estimate of the relationship of campaign spending to votes. But if our goal were forecasting, we would wish to map all of the systematic variation in the dependent variable, and omitting incumbency would prevent that since we are leaving out an important causal variable. However, even if our long-term goal were the fullest systematic explanation of the

If these special cases do not hold for some omitted variable (i.e., this variable is correlated with the included explanatory variable and has an effect on the dependent variable), then failure to control for it will bias our estimate (or perception) of the effect of the included variable. In the case at hand, our colleague would be right in her criticism since incumbency is related to both the dependent variable and the independent variable: incumbents get more votes and they spend more.

This insight can be put in formal terms by focusing on the last line of equation (5.5) from the box below:

$$E(b_1) = \beta_1 + F\beta_2 \tag{5.3}$$

This is the equation used to calculate the bias in the estimate of the effect of X_1 on the dependent variable Y. In this equation, F represents the degree of correlation between the two explanatory variables X_1 and X_2.[9] If the estimator calculated by using only X_1 as an explanatory variable (that is b_1) was unbiased, it would equal β_1 on average; that is, it would be true that $E(b_1) = \beta_1$. This estimator is unbiased in the two special cases where the *bias term* $F\beta_2$ equals zero. It is easy to see that this formalizes the conditions for unbiasedness that we stated above. That is, we can omit a control variable if either

- The omitted variable has no causal effect on the dependent variable (that is, $\beta_2 = 0$, regardless of the nature of the relationship between the included and excluded variables F); or
- The omitted variable is uncorrelated with the included variable (that is, $F = 0$, regardless of the value of β_2.)

If we discover an omitted variable that we suspect might be biasing our results, our analysis should *not* end here. If possible, we should control for the omitted variable. And even if we cannot, because we have no good source of data about the omitted variable, our model can help us to ascertain the direction of bias, which can be extremely helpful. Having an underestimate or an overestimate may substantially bolster or weaken an existing argument.

For example, suppose we study a few sub-Saharan African states and find that coups d'etat appear more frequently in politically repressive regimes—that β_1 (the effect of repression on the likelihood of a coup) is positive. That is, the explanatory variable is the degree of po-

---

vote, it might prove difficult to be very confident of several causal effects within the framework of a single study. Thus, it might pay to focus on one causal effect (or just a few), whatever our long-term goal.

[9] More precisely, F is the coefficient estimate produced when X_1 is regressed on X_2.

litical repression, and the dependent variable is the likelihood of a coup. The unit of analysis is the sub-Saharan African countries. We might even expand the sample to other African states and come to the same conclusion. However, suppose that we did not consider the possible effects of economic conditions on coups. Although we might have no data on economic conditions, it is reasonable to hypothesize that unemployment would probably increase the probability of a coup d'etat ($\beta_2 > 0$), and it also seems likely that unemployment is positively correlated with political repression ($F > 0$). We also assume, for the purposes of this illustration that economic conditions are prior to our key causal variable, the degree of political repression. If this is the case, the degree of bias in our analysis could be severe. Since unemployment has a positive correlation with both the dependent variable and the explanatory variable ($F\beta_2 > 0$ in this case), excluding that variable would mean that we were inadvertently estimating the effect of repression and unemployment on the likelihood of a coup instead of just repression ($\beta_1 + F\beta_2$ instead of β_1). Furthermore, because the joint impact of repression and unemployment is greater than the effect of repression alone ($\beta_1 + F\beta_2$ is greater than β_1), the estimate of the effect of repression (b_1) will be too large on average. Therefore, this analysis shows that by excluding the effects of unemployment, we overestimated the effects of political repression. (This is different from the consequences of measurement error in the explanatory variables since omitted variable bias can sometimes cause a negative relationship to be estimated as a positive one.)

Omitting relevant variables does not always result in overestimates of causal effects. For example, we could reasonably hypothesize that in some other countries (perhaps the subject of a new study), political repression and unemployment were inversely related (that F is negative). In these countries, political repression might enable the government to control warring factions, impose peace from above, and put most people to work. This in turn means that the effect of bias introduced by the negative relationship of unemployment and repression ($F\beta_2$) will also be negative, so long as we are still willing to assume that more unemployment will increase the probability of a coup in these countries. The substantive consequence is that the estimated effect of repression on the likelihood of a coup ($E(b_1)$) will now be less than the true effect (β_1). Thus, if economic conditions are excluded, b_1 will generally be an *underestimate* of the effect of political repression. If F is sufficiently negative and β_2 is sufficiently large, then we might routinely estimate a positive β_1 to be negative and incorrectly conclude that more political repression decreases the probability of a coup d'etat! Even if we had insufficient information on unemployment rates

to include it in the original study, an analysis like this can still help us generate reasonable substantive conclusions.

As these examples should make clear, we need not actually run a regression to estimate parameters, to assess the degrees and directions of bias, or to arrive at such conclusions. Qualitative and intuitive estimates are subject to the same kinds of biases as are strictly quantitative ones. This section shows that in both situations, information outside the existing data can help substantially in estimating the degree and direction of bias.

If we know that our research design might suffer from omitted variables but do not know what those variables are, then we may very well have flawed conclusions (and some future researcher is likely to find them). The incentives to find out more are obvious. Fortunately, in most cases, researchers have considerable information about variables outside their analysis. Sometimes this information is detailed but available for only some subunits, or partial but widely applicable, or even from previous research studies. Whatever the source, even incomplete information can help one focus on the likely degree and direction of bias in our causal effects.

Of course, even scholars who understand the consequences of omitted variable bias may encounter difficulties in identifying variables that might be omitted from their analysis. No formula can be provided to deal with this problem, but we do advise that all researchers, quantitative and qualitative, systematically look for omitted control variables and consider whether they should be included in the analysis. We suggest some guidelines for such a review in this section.

Omitted variables can cause difficulties even when we have adequate information on all relevant variables. Scholars sometimes have such information, and believing the several variables to be positively related to the dependent variable, they estimate the causal effects of these variables sequentially, in separate "bivariate" analyses. It is particularly tempting to use this approach in studies with a small number of observations, since including many explanatory variables simultaneously creates very imprecise estimates or even an indeterminate research design, as discussed in section 4.1. Unfortunately, however, each analysis excludes the other relevant variables, and this omission leads to omitted variable bias in each estimation. The ideal solution is not merely to collect information on all relevant variables, but explicitly and *simultaneously* to control for all relevant variables. The qualitative researcher must recognize that failure to take into account all relevant variables at the same time leads to biased inferences. Recognition of the sources of bias is valuable, even if small numbers of observations make it impossible to remove them.

Concern for omitted variable bias, however, should *not* lead us automatically to include every variable whose omission might cause bias because it is correlated with the independent variable and has an effect on the dependent variable. In general, *we should not control for an explanatory variable that is in part a consequence of our key causal variable.*

Consider the following example. Suppose we are interested in the causal effect of an additional $10,000 in income (our treatment variable) on the probability that a citizen will vote for the Democratic candidate (our dependent variable). Should we control for whether this citizen reports planning to vote Democratic in an interview five minutes before he arrives at the polls? This control variable certainly affects the dependent variable and is probably correlated with the explanatory variable. Intuitively, the answer is no. If we did control for it, the estimated effect of income on voting Democratic would be almost entirely attributed to the control variable, which in this case is hardly an alternative causal explanation. A blind application of the omitted variable bias rules, above, might incorrectly lead one to control for this variable. After all, this possible control variable certainly has an effect on the dependent variable—voting Democratic—and it is correlated with the key explanatory variable—income. But including this variable would attribute part of the causal effect of our key explanatory variable to the control variable.

To take another example, suppose we are interested in the causal effect of a sharp increase in crude-oil prices on public opinion about the existence of an energy shortage. We could obtain measures of oil prices (our key causal variable) from newspapers and use opinion polls as our dependent variable to gauge the public's perception of whether there is an energy shortage. But we might ask whether we should control for the effects of television coverage of energy problems. Certainly television coverage of energy problems is correlated with both the included explanatory variable (crude oil prices) and the dependent variable (public opinion about an energy shortage). However, since television coverage is in part a consequence of real-world oil prices, we should not control for that coverage in assessing the causal influence of oil prices on public opinion about an energy shortage. If instead we were interested in the causal effect of television coverage, we would control for oil prices, since these prices come *before* the key explanatory variable (which is now coverage).[10]

[10] It is worth considering just what it means to look at the estimated causal effect of crude-oil prices on public opinion about an energy shortage, while controlling for the amount of television coverage about energy shortages. Consider two descriptions, both of which are important in that they enable us to further analyze and study the causal processes in greater depth. First, this estimated effect is just the effect of that aspect of oil

Thus, to estimate the total effect of an explanatory variable, we should list all variables that, according to our theoretical model, could cause the dependent variable. To repeat the point made above: in general, we should not control for an explanatory variable that is in part a consequence of our key explanatory variable. Having eliminated these possible explanatory variables, we should then control for other potential explanatory variables that would otherwise cause omitted variable bias—those that are correlated with both the dependent variable and with the included explanatory variables.[11]

The argument that we should not control for explanatory variables that are consequences of our key explanatory variables has a very important implication for the role of theory in research design. Thinking about this issue, we can see why we should begin with or at least work towards a theoretically-motivated model rather than "data-mining": running regressions or qualitative analyses with whatever explanatory variables we can think of. Without a theoretical model, we cannot decide which potential explanatory variables should be included in our analysis. Indeed, in the absence of a model, we might get the strongest results by using a trivial explanatory variable—such as intention to vote Democratic five minutes before entering the polling place—and controlling for all other factors correlated with it. We cannot determine whether to control for or ignore possible explanatory variables that are correlated with each other without a theoretically motivated model, without which we have serious dangers either of omitted variable bias or triviality in research design.

Choosing when to add additional explanatory variables to our analysis is by no means simple. The number of additional variables is always unlimited, our resources are limited, and, above all, the more

---

prices that *directly* affects public opinion about an energy shortage, apart from the aspect of the causal effect that affects public opinion indirectly with changing television coverage. That is, it is the direct and not the indirect effect of oil on opinion. The total effect can be found by not controlling for the extent of television coverage of energy shortages at all. An alternative description of this effect is the effect of energy prices on the variable "public opinion about energy shortages given a fixed degree of television coverage about energy shortages." As an example of the latter, imagine the experiment in which we controlled network television coverage of oil shortages and forced it to remain at the same level while crude oil prices varied naturally. Since coverage is a constant in this experiment, it is controlled for without any other explicit procedure. Even if we could not do an experiment, we could still estimate this conditional effect of oil prices on public opinion about energy shortages by controlling for television coverage.

[11] In addition, we might be interested in just the direct or indirect effect of a variable, or even in the causal effect of some other variable in an equation. In this situation, a perfectly reasonable procedure is to run several different analyses on the same data, as long as we understand the differences in interpretation.

explanatory variables we include, the less leverage we have for estimating any of the individual causal effects. Avoiding omitted variable bias is one reason to add additional explanatory variables. If relevant variables are omitted, our ability to estimate causal inferences correctly is limited.

A Formal Analysis of Omitted Variable Bias. Let us begin with a simple model with two explanatory variables

$$E(Y) = X_1\beta_1 + X_2\beta_2 \tag{5.4}$$

Suppose now that we came upon an important analysis which reported the effect of X_1 on Y without controlling for X_2. Under what circumstances would we have grounds for criticizing this work or justification for seeking funds to redo the study? To answer this question, we formally evaluate the estimator with the omitted control variable.

The estimator of β_1 where we omit X_2 is

$$b_1 = \frac{\sum_{i=1}^{n} X_{1i}Y_i}{\sum_{i=1}^{n} X_{1i}^2}$$

To evaluate this estimator, we take the expectation of b_1 across hypothetical replications under the model in equation (5.4):

$$\begin{aligned} E(b_1) &= E\left(\frac{\sum_{i=1}^{n} X_{1i}Y_i}{\sum_{i=1}^{n} X_{1i}^2}\right) \qquad (5.5)\\ &= \frac{\sum_{i=1}^{n} X_{1i}E(Y_i)}{\sum_{i=1}^{n} X_{1i}^2} \\ &= \frac{\sum_{i=1}^{n} X_{1i}(X_{1i}\beta_1 + X_{2i}\beta_2)}{\sum_{i=1}^{n} X_{1i}^2} \\ &= \frac{\sum_{i=1}^{n} X_{1i}^2\beta_1 + \sum_{i=1}^{n} X_{1i}X_{2i}\beta_2)}{\sum_{i=1}^{n} X_{1i}^2} \\ &= \beta_1 + F\beta_2 \end{aligned}$$

where $F = \frac{\sum_{i=1}^{n} X_{1i} X_{2i}}{\sum_{i=1}^{n} X_{1i}^2}$, the slope coefficient from the regression of X_1 on X_2. The last line of this equation is reproduced in the text in equation (5.3) and is discussed in some detail above.

5.2.2 Examples of Omitted Variable Bias

In this section, we consider several quantitative and qualitative examples, some hypothetical and some from actual research. For example, educational level is one of the best predictors of political participation. Those who have higher levels of education are more likely to vote and more likely to take part in politics in a number of other ways. Suppose we find this to be the case in a new data set but want to go further and see whether the relationship between the two variables is causal and, if so, how education leads to participation.

The first thing we might do would be to see whether there are omitted variables antecedent to education that are correlated with education and at the same time cause participation. Two examples might be the political involvement of the individual's parents and the race of the individual. Parents active in politics might inculcate an interest in participation in their children and at the same time be the kind of parents who foster educational attainment in their children. If we did not include this variable, we might have a spurious relationship between education and political activity or an estimate of the relationship that was too strong.

Race might play the same role. In a racially discriminatory society, blacks might be barred from both educational opportunities and political participation. In such a case, the apparent effect of education on participation would not be real. Ideally, we would want to eliminate all possible omitted variables that might explain away part or all of the relationship between education and participation.

But the fact that the relationship between education and participation diminishes or disappears when we control for an antecedent variable does not necessarily mean that education is irrelevant. Suppose we found that the education-participation link diminished when we controlled for race. One reason might be, as in the example above, that discrimination against blacks meant that race was associated separately with both educational attainment and participation. Under these conditions, no real causal link between education and participation would exist. On the other hand, race might affect political participa-

tion *through* education. Racial discrimination might reduce the access of blacks to education. Education might, in turn, be the main factor leading to participation. In this case, the reduction in the relationship between education and participation that is introduced when the investigator adds race to the analysis does not diminish the importance of education. Rather, it explains how race and education interact to affect participation.

Note that these two situations are fundamentally different. If lower participation on the part of blacks was due to a lack of education, we might expect participation to increase if their average level of education increased. But if the reason for lower participation was direct political discrimination that prevented the participation of blacks as citizens, educational improvement would be irrelevant to changes in patterns of participation.

We might also look for variables that are simultaneous with education or that followed it. We might look for omitted variables that show the relationship between education and participation to be spurious. Or we might look for variables that help explain how education works to foster participation. In the former category might be such a variable as the general intelligence level of the individual (which might lead to doing well in school and to political activity). In the latter category might be variables measuring aspects of education such as exposure to civics courses, opportunities to take part in student government, and learning of basic communications skills. If it were found that one or more of the latter, when included in the analysis, reduced the relationship between educational attainment and participation (when we controlled for communications skills, there was no independent effect of educational attainment on participation), this finding would not mean that education was irrelevant. The requisite communications skills were learned in school and there would be a difference in such skills across educational levels. What the analysis would tell us would be how education influenced participation.

All of these examples illustrate once again why it is necessary to have a theoretical model in mind to evaluate. There is no other way to choose what variables to use in our analysis. A theory of how education affected civic activity would guide us to the variables to include. Though we do not add additional variables to a regression equation in qualitative research, the logic is much the same when we decide what other factors to take into account. Consider the research question we raised earlier: the impact of summit meetings on cooperation between the superpowers. Suppose we find that cooperation between the United States and the USSR was higher in years following a summit

than preceding one. How would we know that the effect is real and not the result of some omitted variable? And if we are convinced it is real, can we explicate further how it works?

We might want to consider antecedent variables that would be related to the likelihood of a summit and might also be direct causes of cooperation. Perhaps when leaders in each country have confidence in each other, they meet frequently and their countries cooperate. Or perhaps when the geopolitical ambitions of both sides are limited for domestic political reasons, they schedule meetings and they cooperate. In such circumstances, summits themselves would play no direct role in fostering cooperation, though the scheduling of a summit might be a good indicator that things were going well between the superpowers. It is also possible that summits would be part of a causal sequence, just as race might have affected educational level which in turn affected participation. When the superpower leaders have confidence in one another, they call a summit to reinforce that mutual confidence. This, in turn, leads to cooperation. In this case, the summit is far from irrelevant. Without it, there would be less cooperation. Confidence and summits interact to create cooperation. Suppose we take such factors into account and find that summits seem to play an independent role—i.e., when we control for the previous mutual confidence of the leaders and their geopolitical ambitions, the conclusion is that a summit seems to lead to more cooperation. We might still go further and ask how that happens. We might compare among summits in terms of characteristics that might make them more or less successful and see if such factors are related to the degree of cooperation that follows. Again we have to select factors to consider, and these might include: the degree of preparation, whether the issues were economic rather than security, the degree of domestic harmony in each nation, the weather at the summit, and the food. Theory would have to guide us; that is, we would need a view of concepts and relationships that would point to relevant explanatory variables and would propose hypotheses consistent with logic and experience about their effects.

For researchers with a small number of observations, omitted variable bias is very difficult to avoid. In this situation, inefficiency is very costly; including too many irrelevant control variables may make a research design indeterminate (section 4.1). But omitting relevant control variables can introduce bias. And a priori the researcher may not know whether a candidate variable is relevant or not.

We may be tempted at this point to conclude that causal inference is impossible with small numbers of observations. In our view, however, the lessons to be learned are more limited and more optimistic. Understanding the difficulty of making valid causal inferences with few ob-

servations should make us cautious about making causal assertions. As indicated in chapter 2, good description and descriptive inference are more valuable than faulty causal inference. Much qualitative research would indeed be improved if there were more attention to valid descriptive inference and less impulse to make causal assertions on the basis of inadequate evidence with incorrect assessments of their uncertainty. However, limited progress in understanding causal issues is nevertheless possible, *if* the theoretical issues with which we are concerned are posed with sufficient clarity and linked to appropriate observable implications. A recent example from international relations research may help make this point.

Helen Milner's study, *Resisting Protectionism* (1988), was motivated by a puzzle: why was U.S. trade policy more protectionist in the 1920s than in the 1970s despite the numerous similarities between the two periods? Her hypothesis was that international interdependence increased between the 1920s and 1970s and helped to account for the difference in U.S. behavior. At this aggregate level of analysis, however, she had only the two observations that had motivated her puzzle which could not help her distinguish her hypothesis from many other possible explanations of this observed variation. The level of uncertainty in her theory would therefore have been much too high had she stopped here. Hence she had to look elsewhere for additional observable implications of her theory.

Milner's approach was to elaborate the process by which her causal effect was thought to take place. She hypothesized that economic interdependence between capitalist democracies affects national preferences by influencing the preferences of industries and firms, which successfully lobby for their preferred policies. Milner therefore studied a variety of U.S. industries in the 1920s and 1970s and French industries in the 1970s and found that those with large multinational investments and more export dependence were the least protectionist. These findings helped confirm her broader theory of the differences in overall U.S. policy between the 1920s and 1970s. Her procedures were therefore consistent with a key part of our methodological advice: specify the observable implications of the theory, even if they are not the objects of principal concern, and design the research so that inferences can be made about these implications and used to evaluate the theory. Hence Milner's study is exemplary in many ways.

The most serious problem of research design that Milner faced involved potential omitted variables. The most obvious control variable is the degree of competition from imports, since more intense competition from foreign imports tends to produce more protectionist firm preferences. That is, import competition is likely to be correlated with

Milner's dependent variable, and it is in most cases antecedent to or simultaneous with her explanatory variables. If this control variable were also correlated with her key causal explanatory variables, multinational investment and export dependence, her results would be biased. Indeed, a negative correlation between import competition and export dependence would have seemed likely on the principles of comparative advantage, so this hypothetical bias would have become real if import competition were not included as a control.

Milner dealt with this problem by selecting for study only industries that were severely affected by foreign competition. Hence, she held constant the severity of import competition and eliminated, or at least greatly reduced, this problem of omitted variable bias. She could have held this key control variable constant at a different level—such as only industries with moderately high levels of import penetration—so long as it was indeed constant for her observations.

Having controlled for import competition, however, Milner still faced other questions of omitted variables. The two major candidates that she considered most seriously, based on a review of the theoretical and empirical literature in her field, were (1) that changes in U.S. power would account for the differences between outcomes in the 1920s and 1970s, and (2) that changes in the domestic political processes of the United States would do so. Her attempt to control for the first factor was built into her original research design: since the proportion of world trade involving the United States in the 1970s was roughly similar to its trade involvement in the 1920s, she controlled for this dimension of American power at the aggregate level of U.S. policy, as well as at the industry and firm level. However, she did not control for the differences between the political isolationism of the United States in the 1920s and its hegemonic position as alliance leader in the 1970s; these factors could be analyzed further to ascertain their potentially biasing effects.

Milner controlled for domestic political processes by comparing industries and firms within the 1920s and within the 1970s, since all firms within these groups faced the same governmental structures and political processes. Her additional study of six import-competing industries in France during the 1970s obviously did not help her hold domestic political processes constant, but it did help her discover that the causal effect of export dependence on preferences for protectionism did not vary with changes in domestic political processes. By carefully considering several potential sources of omitted variable bias and designing her study accordingly, Milner greatly reduced the potential for bias.

However, Milner did not explicitly control for several other possible omitted variables. Her study focused "on corporate trade preferences and does not examine directly the influence of public opinion, ideology, organized labor, domestic political structure, or other possible factors" (1988: 15–16). Her decision not to control for these variables could have been justified on the theoretical grounds that these omitted variables are unrelated to, or are in part consequences of, the key causal variables (export dependence and multinational investment), or have no effect on the dependent variable (preferences for protectionism at the level of the firm, aggregated to industries). However, if these omitted variables were plausibly linked to both her explanatory and dependent variables and were causally prior to her explanatory variable, she would have had to design her study explicitly to control for them.[12]

Finally, Milner's procedure for selecting industries risked making her causal inferences inefficient. As we have noted, her case-selection procedure enabled her to control for the most serious potential source of omitted variable bias by holding import competition constant, which on theoretical grounds was expected to be causally prior to and correlated with her key causal variable and to influence her dependent variables. She selected those industries that had the highest levels of import competition and did not stratify by any other variable. She then studied the preferences of each industry in her sample, and of many firms, for protectionism preferences (her dependent variable) and researched the degree of international economic dependence (her explanatory variable).

This selection procedure is inefficient with respect to her causal inferences because her key causal variables varied less than would have been desirable (Milner 1988:39–42). Although this inefficiency turned out not to be a severe problem in her case, it did mean that she had to do more case studies than were necessary to reach the same level of certainty about her conclusions (see section 6.2). Put differently, with the same number of cases, chosen so that they varied widely on her explanatory variable, she could have produced more certain causal in-

[12] Milner addresses the potential for omitted variable bias, but her reasoning is flawed: "By looking at different industries, at different times, and in different countries, [the research design] allows these [omitted control variables] to vary, while showing that the basic argument still holds" (1988:15). In fact, the only way "to hold control variables constant" is actually to hold them constant, not to let them vary. If plausible competing theories had identified these variables as important, she could have looked at a set of observations which differed on her key explanatory variable (degree of international economic dependence of the country, industry, or firm) but not on these control variables.

ferences. That is, her design would have been more efficient had she chosen some industries and firms with no foreign ties and some with high levels of foreign involvement, all of which suffered from constant levels of economic distress and import penetration.

Researchers can never conclusively reject the hypothesis that omitted variables have biased their analyses. However, Milner was able to make a stronger, more convincing case for her hypothesis than she could have done had she not tried to control for some evident sources of omitted variable bias. Milner's rigorous study indicates that social scientists who work with qualitative material need not despair of making limited causal inferences. Perfection is unattainable, perhaps even undefinable; but careful linking of theory and method can enable studies to be designed in a way that will improve the plausibility of our arguments and reduce the uncertainty of our causal inferences.

5.3 Including Irrelevant Variables: Inefficiency

Because of the potential problems with omitted variable bias described in section 5.2, we might naively think that it is essential to collect and simultaneously estimate the causal effects of all possible explanatory variables. At the outset, we should remember that this is not the implication of section 5.2. We showed there that omitting an explanatory variable that is uncorrelated with the included explanatory variables does not create bias, even if the variable has a strong causal impact on the dependent variable, and that controlling for variables that are the consequences of explanatory variables is a mistake. Hence, *our argument should not lead researchers to collect information on every possible causal influence or to criticize research which fails to do so*.

Of course, a researcher might still be uncertain about which antecedent control variables have causal impact or are correlated with the included variables. In this situation, some researchers might attempt to include all control variables that are conceivably correlated with the included explanatory variables as well as all those that might be expected on theoretical grounds to affect the dependent variable. This is likely to be a very long list of variables, many of which may be irrelevant. Such an approach, which appears at first glance to be a cautious and prudent means of avoiding omitted variable bias, would, in fact, risk producing a research design that could only produce indeterminate results. In research with relatively few observations, indeterminacy, as discussed in section 4.1, is a particularly serious problem, and such a "cautious" design would actually be detrimental. This section discusses the costs of including irrelevant explanatory variables and provides essential qualifications to the "include everything" approach.

The inclusion of irrelevant variables can be very costly. Our key point is that even if the control variable has no causal effect on the dependent variable, *the more correlated the main explanatory variable is with the irrelevant control variable, the less efficient is the estimate of the main causal effect.*

To illustrate, let us focus on two different procedures (or "estimators") for calculating an estimate of the causal effect of an appropriately included explanatory variable. The first estimate of this effect is from an analysis with no irrelevant control variables; the second includes one irrelevant control variable. The formal analysis in the box below provides the following conclusions about the relative worth of these two procedures, in addition to the one already mentioned. First, *both estimators are unbiased*. That is, even when controlling for an irrelevant explanatory variable, the usual estimator still gives the right answer on average. Second, *if the irrelevant control variable is uncorrelated with the main explanatory variable, the estimate of the causal effect of the latter is not only unbiased, but it is as efficient as if the irrelevant variable had not been included.* Indeed, if these variables are uncorrelated, precisely the same inference will result. However, if the irrelevant control variable is highly correlated with the main explanatory variable, substantial inefficiency will occur.

The costs of controlling for irrelevant variables are therefore high. When we do so, each study we conduct is much more likely to yield estimates far from the true causal effects. When we replicate a study in a new data set in which there is a high correlation between the key explanatory variable and an irrelevant included control variable, we will be likely to find different results, which would suggest different causal inferences. Thus, even if we control for all irrelevant explanatory variables (and make no other mistakes), we will get the right answer on average, but we may be far from the right answer in any single project and possibly every one. On average, the reanalysis will produce the same effect but the irrelevant variable will increase the inefficiency, just as if we had discarded some of our observations. The implication should be clear: by including an irrelevant variable, we are putting more demands on our finite data set, resulting in less information available for each inference.

As an example, consider again the study of coups d'etat in African states. A preliminary study indicated that the degree of political repression, the main explanatory variable of interest, increased the frequency of coups. Suppose another scholar argued that the original study was flawed because it did not control for whether the state won independence in a violent or negotiated break from colonial rule. Suppose we believe this second scholar is wrong and that the nature of the

break from colonial rule had no effect on the dependent variable—the frequency of coups (after the main explanatory variable, political repression, is controlled for). What would be the consequences of controlling for this irrelevant, additional variable?

The answer depends on the relationship between the irrelevant variable, which measures the nature of the break from colonial rule, and the main explanatory variable, which measures political repression. If the correlation between these variables is high—as seems plausible—then including these control variables would produce quite inefficient estimates of the effect of political repression. To understand this, notice that to control for how independence was achieved, the researcher might divide his categories of repressive and nonrepressive regimes according to whether they broke from colonial rule violently or by negotiation. The frequency of coups in each category could be counted to assess the causal effects of political repression, while the means of breaking from colonial rule is controlled. Although this sort of design is a reasonable way to avoid omitted variable bias, it can have high costs: when the additional control variable has no effect on the dependent variable but is correlated with an included explanatory variable, the number of observations in each category is reduced and the main causal effect is estimated much less efficiently. This result means that much of the hard work the researcher has put in was wasted, since unnecessarily reducing efficiency is equivalent to discarding observations. The best solution is to always collect more observations, but if this is not possible, researchers are well-advised to identify irrelevant variables and not control for them.

A Formal Analysis of Included Variable Inefficiencies. Suppose the true model is $E(Y) = X_1\beta$ and $V(Y) = \sigma^2$. However, we incorrectly think that a second explanatory variable X_2 also belongs in the equation. So we estimate

$$E(Y) = X_1\beta_1 + X_2\beta_2 \tag{5.6}$$

not knowing that in fact $\beta_2 = 0$. What consequence does a simultaneous estimation of both parameters have for our estimate of β_1?

Define b_1 as the correct estimator, based only on a regression of Y on X_1, and $\hat{\beta}_1$ as the first coefficient on X_i from a regression of Y on X_1 and X_2. It is easy to show that we cannot distinguish between these two estimators on the basis of unbiasedness (being correct on average across many hypothetical experiments), since both are unbiased:

$$E(b_1) = E(\hat{\beta}_1) = \beta_1 \tag{5.7}$$

The estimators do differ, however, with respect to efficiency. The correct estimator has a variance (calculated in equation [3.9]) of

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} X_{1i}^2} \tag{5.8}$$

whereas the other estimator has variance

$$\begin{aligned} V(\hat{\beta}_1) &= \frac{\sigma^2}{(1 - r_{12}^2)\sum_{i=1}^{n} X_{1i}^2} \\ &= \frac{V(b_1)}{(1 - r_{12}^2)} \end{aligned} \tag{5.9}$$

where the correlation between X_1 and X_2 is r_{12} (see Goldberger 1991:245).

From the last line in equation (5.9), we can see the precise relationship between the variances of the two estimators. If the correlation between the two explanatory variables is zero, then it makes no difference whether you include the irrelevant variable or not, since both estimators have the same variance. However, the more correlated two variables are, the higher the variance, and thus lower the efficiency, of $\hat{\beta}_1$.

5.4 Endogeneity

Political science research is rarely experimental. We do not usually have the opportunity to manipulate the explanatory variables; we just observe them. One consequence of this lack of control is endogeneity—that the values our explanatory variables take on are sometimes a consequence, rather than a cause, of our dependent variable. With true experimental manipulation, the direction of causality is unambiguous. But for many areas of qualitative and quantitative research, endogeneity is a common and serious problem.[13]

[13] Qualitative researchers do sometimes manipulate explanatory variables through participant observation. Even in-depth interviews can be a form of experiment if different questions are asked systematically or other conditions are changed in different interviews. In fact, it can even be a problem even for in-depth interviews, since a researcher might feel more comfortable applying experimental "treatments" (asking certain ques-

In the absence of investigator control over the values of the explanatory variables, the direction of causality is always a difficult issue. In nonexperimental research—quantitative or qualitative—explanatory and dependent variables vary because of factors out of the control (and often out of sight) of the researcher. States invade; army officers plot coups; inflation drops; government policies are enacted; candidates decide to run for office; voters choose among candidates. A scholar must try to piece together an argument about what is causing what.

An example is provided by the literature on U.S. congressional elections. Many scholars have argued that the dramatic rise of the electoral advantage of incumbency during the late 1960s was due in large part to the increase in constituency service performed by members of Congress. That is, the franking privilege, budgets for travel to the district, staff in the district to handle specific constituent requests, pork-barrel projects, and other perquisites of office have allowed congressional incumbents to build up support in their districts. Many citizens vote for incumbent candidates on these grounds.

This constituency-service hypothesis seems perfectly reasonable, but does the evidence support it? Numerous scholars have attempted to provide such evidence (for a review of this literature, see Cain, Ferejohn, and Fiorina 1987), but the positive evidence is scarce. The modal study of this question is based on measures of the constituency service performed by a sample of members of Congress and of the proportion of the vote for the incumbent candidate. The researchers then estimate the causal impact of service on the vote through regression analysis. Surprisingly, many of these estimates indicate that the effect is zero or even negative.

It seems likely that the problem of endogeneity accounts for these paradoxical results. In other words, members at highest risk of losing the next election (perhaps because of a scandal or hard times in their district) do extra constituency service. Incumbents who feel secure about being reelected probably focus on other aspects of their jobs, such as policy-making in Washington. The result is that those incumbents who do the most service receive the fewest votes. This does not mean that constituency service reduces the vote, only that a strong expected vote reduces service. By ignoring the feedback effect, one's inferences will be strongly biased.

David Laitin outlines an example of an endogeneity problem in one of the classics of early twentieth century social science, Max Weber's *The Protestant Ethic and the Spirit of Capitalism.* "Weber attempted to

---

tions) to certain, nonrandomly selected, respondents. Experimenters have numerous problems of their own, but endogeneity is not usually one of them.

demonstrate that a specific type of economic behavior—the capitalist spirit — was (inadvertently) induced by Protestant teachings and doctrines. But . . . Weber and his followers could not answer one objection that was raised to their thesis: namely that the Europeans who already had an interest in breaking the bonds of precapitalist spirit might well have left the church precisely for that purpose. In other words, the economic interests of certain groups could be seen as inducing the development of the Protestant ethic. Without a better controlled study, Weber's line of causation could be turned the other way." (Laitin 1986:187; see also R. H. Tawney 1935 who originated the criticism).

In the remainder of this section, we will discuss five methods of coping with the difficult problem of endogeneity:

- Correcting a biased inference (section 5.4.1);
- Parsing the dependent variable and studying only those parts that are consequences, rather than causes, of the explanatory variable (section 5.4.2);
- Transforming an endogeneity problem into bias due to an omitted variable, and controlling for this variable (section 5.4.3);
- Carefully selecting at least some observations without endogeneity problems (section 5.4.4); and
- Parsing the explanatory variables to ensure that only those parts which are truly exogenous are in the analysis (section 5.4.5).

Each of these five procedures can be viewed as a method of avoiding endogeneity problems, but each can also be seen as a way of clarifying a causal hypothesis. For a causal hypothesis that ignores an endogeneity problem is, in the end, a theoretical problem, requiring respecification so that it is at least possible that the explanatory variables could influence the dependent variable. We will discuss the first two solutions to endogeneity in the context of our quantitative constituency service example and the remaining three with the help of extended examples from qualitative research.

5.4.1 Correcting Biased Inferences

The last line of equation (5.13) in the box below provides a procedure for assessing the exact direction and degree of bias due to endogeneity. For convenience, we reproduce equation (5.13) here:

$$E(b) = \beta + \text{Bias}$$

This equation implies that if endogeneity is present, we are not making the causal inference we desire. That is, if the bias term is zero, our method of inference (or estimator b) will be unbiased on average (that

is, equal to β). But if we have endogeneity bias, we are estimating the correct inference plus a bias factor. Endogeneity is a problem because we are generally unaware of the size or direction of the bias. This bias factor will be large or small, negative or positive, depending on the specific empirical example. Fortunately, even if we cannot avoid endogeneity bias in the first place, we can sometimes correct for it after the fact by ascertaining the direction and perhaps the degree of the bias.

Equation (5.13) demonstrates that the bias factor depends on the correlation between the explanatory variable and the error term—the part of the dependent variable unexplained by the explanatory variable. For example, if the constituency-service hypothesis is correct, then the causal effect of constituency service on the vote (β in the equation) is positive. If, in addition, the expected vote affects the level of constituency service we observe, then the bias term will be negative. That is, even after the effect of constituency service on the vote is taken into account, constituency service will inversely correlate with the error term because incumbents who have lower expected votes will perform more service. The result is that the bias term is negative, and uncorrected inferences in this case are biased estimates of the causal effect β (or, equivalently, unbiased estimates of [β + bias]). Thus, even if the constituency-service hypothesis is true, endogeneity bias would cause us to estimate the effect of service as a smaller positive number than it should be, as zero, or even as negative, depending on the size of the bias factor. Hence, we can conclude that the correct estimate of the effect of service on the vote is larger than we estimated in an analysis conducted with no endogeneity correction. As a result, our uncorrected analysis yields a lower bound on the effect of service, making the constituency-service hypothesis more plausible.

Thus, even if we cannot avoid endogeneity bias, we can sometimes improve our inferences after the fact by estimating the degree of bias. At a minimum, this enables us to determine the direction of bias, perhaps providing an upper or lower bound on the correct estimate. At best, we can use this technique to produce fully unbiased inferences.

5.4.2 Parsing the Dependent Variable

One way to avoid endogeneity bias is to reconceptualize the dependent variable as itself containing a dependent and an explanatory component. The explanatory component of the dependent variable interferes with our analysis through a feedback mechanism, that is, by influencing our key causal (explanatory) variable. The other component of our dependent variable is truly dependent, a function, and not

a cause, of our explanatory variable. The goal of this method of avoiding endogeneity bias is to identify and measure only the dependent component of our dependent variable.

For example, in a study of the constituency-service hypothesis, King (1991a) separated from the total vote for a member of congress the portion due solely to incumbency status. In recent years, the electoral advantage of incumbency status is about 8–10 percentage points of the vote, as compared to a base for many incumbents of roughly 52 percent of the two-party vote. Through a statistical procedure, King then estimated the incumbency advantage, which was a solely dependent component of the dependent variable, and he used this figure in place of the raw vote to estimate the effects of constituency service. Since the incumbent's vote advantage, being such a small portion of the entire vote, would not have much of an effect on the propensity for incumbent legislators to engage in constituency service, he avoided endogeneity bias. His results indicated that an extra $10,000 added to the budget of the average state legislator for constituency service (among other things) gives this incumbent an additional 1.54 percentage point advantage (plus or minus about 0.4 percent) in the next election, hence providing the first empirical support for the constituency-service hypothesis.

5.4.3 Transforming Endogeneity into an Omitted Variable Problem

We can always think of endogeneity as a case of omitted variable bias, as the following famous example from the study of comparative electoral systems demonstrates. One of the great puzzles of political analysis for an earlier generation of political scientists was the fall of the Weimar Republic and its replacement by the Nazi regime in the early 1930s. One explanation, supported by some close and compelling case studies of Weimar Germany, was that the main cause was the imposition of proportional representation as the mode of election in the Weimar Constitution. The argument, briefly stated, is that proportional representation allows small parties representing specific ideological, interest, or religious groups to achieve representation in parliament. Under such an electoral system, there is no need for a candidate to compromise his or her position in order to achieve electoral success such as there is under a single-member-district, winner-take-all electoral system. Hence parliament will be filled with small ideological groups unwilling and unable to work together. The stalemate and frustration would make it possible for one of those groups—in this case the National Socialists—to seize power. (For the classic statement of this theory, see Hermens 1941).

The argument in the above paragraph was elaborated in several important case studies of the fall of the Weimar Republic. Historians and political scientists traced the collapse of Weimar to the electoral success of small ideological parties and their unwillingness to compromise in the Reichstag. There are many problems with the explanation, as of course there would be for an explanation of a complex outcome that is based on a single instance, but let us look only at the problem of endogeneity. The underlying explanation involved a causal mechanism with the following links in the causal chain: proportional representation was introduced and enabled small parties with narrow electoral bases to gain seats in the Reichstag (including parties dedicated to its overthrow, like the National Socialists). As a result, the Reichstag was stalemated and the populace was frustrated. This, in turn, led to a coup by one of the parties.

But further study—of Germany as well as of other observable implications—indicated that party fragmentation was not merely the result of proportional representation. Scholars reasoned that if party fragmentation led to adoption of proportional representation, it would also be the cause. By applying the same explanatory variable to other observations (following our rule from chapter 1 that evidence should be sought for hypotheses in data other than that in which they were generated), scholars found that societies with a large number of groups with narrow and intense views in opposition to other groups—minority, ethnic, or religious groups, for instance—are more likely to adopt proportional representation, since it is the only electoral system that the various factions in society can agree on. A closer look at German politics before the introduction of proportional representation confirmed this idea by locating many small factions. Proportional representation did not create these factions, although it may have facilitated their parliamentary expression. Nor were the factions the sole cause of proportional representation; however, both the adoption of proportional representation and parliamentary fragmentation seem to have been effects of social fragmentation. (See Lakeman and Lambert 1955:155 for an early explication of this argument.)

Thus, we have transformed an endogeneity problem into omitted variable bias. That is, prior social fragmentation is an omitted variable that causes proportional representation, is causally prior to it, and led in part to the fall of Weimar. By transforming the problem in this way, scholars were able to get a better handle on the problem since they could explicitly measure this omitted variable and control for it in subsequent studies. In this example, once the omitted variable was included and controlled for, scholars found that there was a reasonable

probability that the apparent causal relationship between proportional representation and the fall of the Weimar Republic was almost entirely spurious.

The subject of the relationship between electoral systems and democracy is still highly contested, although study of it has progressed greatly since these early studies. Scholars have expanded the study from one of concentrated case studies without much concern for the logic of explanation to one of studies based on many observations of given implications and gradually resolved some aspects of measurement and ultimately of inference. In so doing, they have been able to separate the exogenous from the endogenous effects more systematically.

5.4.4 Selecting Observations to Avoid Endogeneity

Endogeneity is a very common problem in much work on the impact of ideas on policy (Hall 1989; Goldstein and Keohane 1993). Insofar as the ideas *reflect the conditions* under which political actors operate—for instance, their material circumstances, which generate their material interests—analysis of the ideas' impact on policy is subject to omitted variable bias: actors' ideas are correlated with a causally prior omitted variable—material interests—which affects the dependent variable—political strategy (See section 5.4.3). And insofar as ideas serve as *rationalizations* of policies pursued on other grounds, the ideas can be mere *consequences* rather than causes of policy. Under these circumstances, ideas are endogenous: they may appear to explain actors' strategies, but in fact they result from these strategies.

The most difficult methodological task in studying the impact of ideas on policy is compensating for the closely related problems of omitted variable bias and endogeneity as they affect a given research problem. To show that ideas are causally important, it must be demonstrated that a given set of ideas held by policymakers, or some aspect of them, affect policies pursued and do not simply reflect those policies or their prior material interests. Researchers in this field must be especially careful in defining the causal effect of interest. In particular, the observed dependent variable (policies) and explanatory variable (ideas held by individuals) must be compared with a precisely defined counterfactual situation in which the explanatory variable takes on a different value: the relevant individuals had different ideas.

Comparative analysis is a good way to determine whether a given set of ideas is exogenous or endogenous. For instance, in a recent study of the role of ideas in the adoption of Stalinist economic policies in

other socialist countries, Nina Halpern (1993) engages in such an analysis. Her hypothesis is that Stalinist planning doctrine—ideas in which Eastern European and Chinese leaders believed—helps to explain their economic policies when they took power after World War II. This hypothesis is consistent with the fact that these leaders held Stalinist ideas and implemented Stalinist policy, but a mere correlation does not demonstrate causality. Indeed, endogeneity may be at work: Stalinist policies could have generated ideas justifying those policies, or anticipation that Stalinist policies would have to be followed could have generated such ideas.

Although Halpern does not use this language, she proceeds in a manner similar to that discussed in section 5.4.3, by transforming endogeneity into omitted variable bias. The principal alternative hypothesis that she considers is that Eastern Europe and Asian Communist states developed command economies after World War II solely as a result of Soviet military might and political influence. The counterfactual claim of this hypothesis is that even if Eastern Europeans and Chinese had not believed in Stalinist ideas about the desirability of planned economies, command economies would still have been implemented in their countries, and ideas justifying them would have appeared.

Halpern then argues that in the Eastern European countries occupied by the Red Army, Soviet power rather than ideas about the superiority of Stalinist doctrines may well have accounted for their adoption of command economies: "the alternative explanation that the choices were purely a response to Stalin's commands is impossible to disprove" (1993:89). Hence she searches for potential observations to which this source of omitted variable bias does not apply and finds the policies followed in China and Yugoslavia, the two largest socialist countries not occupied by Soviet troops after World War II. Since China was a huge country that had an indigenous revolution, Stalin could not dictate policy to it. The Communists in Yugoslavia also achieved power without the aid of the Red Army, and Marshall Tito demonstrated his independence from Moscow's orders from the end of World War II onward.

China instituted a command economy without being under the political or military domination of the Soviet Union; and in Yugoslavia, Stalinist measures were adopted *despite* Soviet policy. Halpern infers from such evidence that in these cases Soviet power alone does not explain policy change. Furthermore, with respect to China, she also considers and rejects another alternative hypothesis by which ideas would be endogenous: that similar economic situations made it appropriate to transplant Stalinist planning methods to China.

Having considered and rejected the alternative hypotheses which hold ideas as endogenous either to Soviet power or economic conditions, Halpern is then able to make her argument that Chinese (and to some extent and for a shorter time, Yugoslav) adoption of Stalinist doctrine provided a basis for agreement and the resolution of uncertainty for these postrevolutionary regimes. Although such an analysis remains quite tentative because of the small number of her theory's implications that she observed, it provides reasons for believing that ideas were not entirely endogenous in this situation—that they played a causal role.

This example illustrates how we can first translate a general concern about endogeneity into specific potential sources of omitted variable bias and then search for a subset of observations in which these sources of bias could not apply. In this case, by transforming the problem to one of omitted variable bias, Halpern was able to compare alternative explanatory hypotheses in an especially productive manner for her substantive hypothesis. She considered several alternative explanatory hypotheses to account for the adoption of command-economy policies and found that only in China, and to some extent Yugoslavia, was it reasonable to consider Stalinist doctrine (the ideas in question) to be largely exogenous. Hence she focused her research on China and Yugoslavia. Had she not carefully designed her study to deal with the problem of endogeneity, her conclusions would be much less convincing—consider, for instance, if she had tried to prove her case with the examples of Poland and Bulgaria!

5.4.5 Parsing the Explanatory Variable

In this section, we introduce a fifth and final method for eliminating the bias due to endogeneity. The goal of this method is to divide a potentially endogenous explanatory variable into two components: one that is clearly exogenous and one that is at least partly endogenous. The researcher then uses only the exogenous portion of the explanatory variable in a causal analysis.

An example of this solution to endogeneity comes from a study of voluntary participation in politics by Verba, Schlozman, and Brady (in progress). These authors were interested in explaining why African-Americans are much more politically active than Latinos, given that the two groups are similarly disadvantaged. The authors find that a variety of factors contribute to the difference, including recency of immigration to the United States and linguistic abilities. One of their key explanatory variables was attendance at religious services (church, synagogue, etc.). The investigators obviously had no control over

whether individuals attended these services, and so the potential for endogeneity could not be ruled out. In fact, they suspected that some Latinos and many more African-Americans attended religious services because they were politically active. Someone who was interested in being politically active might join a church because it offered a chance to learn such skills or was highly politicized. A politicized clergy might train congregants for political activity or provide them with political stimuli. In other words, the causal arrow might run from politics to nonpolitical experiences rather than vice versa.

Verba et al. solved this problem by parsing their key explanatory variable. They did this by arguing that religious institutions affect political participation in two ways. First, individuals learn civic skills in these institutions (for instance, how to make a speech or how to conduct a meeting). The acquisition of such skills, in turn, makes the citizen more competent to take part in political life and more willing to do so. Second, citizens are exposed to political stimulation (for instance, discussion of political matters or direct requests to become politically active from others associated with the institution). And this exposure, too, should affect political activity. The authors argued that the first component is largely exogenous, whereas the second is at least partly endogenous: that is, it is partly due to the extent to which individuals are politically active (the dependent variable).

The authors then conducted an auxiliary study to evaluate this hypothesis about exogenous and endogenous components of participation at religious services. They began by recognizing that the likelihood that an individual acquires civic skills in church depends on the organizational structure of the church. A church that is organized in a hierarchical manner, where clergy are appointed by central church officials and where congregants play little role in church governance, provides fewer opportunities for the individual church member to learn participatory civic skills than does a church organized on a congregational basis where the congregants play a significant role in church governance. Most African-Americans belong to Protestant churches organized on a congregational basis while most Latinos belong to Catholic churches organized on a hierarchical basis. The authors showed that it is this difference in church affiliation that explains the likelihood of acquiring civic skills. They showed, for instance, that for both groups as well as for Anglo-white Americans, it is the nature of the denomination that affects the acquisition of civic skills, not ethnicity, other social characteristics, or, especially, political participation.

Having convinced themselves that the acquisition of civic skills really was exogenous to political participation, Verba et al. measured the acquisition of civic skills at religious services and used this vari-

able, rather than attendance at religious services, as their explanatory variable. This approach solved the endogeneity problem, since they had now parsed their explanatory variable to include only its exogenous component.

This auxiliary study provided further supporting evidence that they had solved their endogeneity problem, since church affiliation of Latinos and African-Americans cannot plausibly be explained by their particular political involvements; church affiliation is in most cases acquired as a child through the family. The reasons why African-Americans are mostly Protestant are found in the histories of American slavery and the institutions that developed on Southern plantations. The reasons why Latinos are Catholic are rooted in the Spanish conquest of Latin America. Nor can the difference between the institutional structure of the Catholic and Protestant churches be attributed to the interests of church officials in involvement in current American politics. Rather, one has to go back to the Reformation to find the source of the difference in organizational structure.

A Formal Analysis of Endogeneity. This formal model demonstrates the bias created if a research design is afflicted by endogeneity, and nothing is done about it. Suppose we have one explanatory variable X and one dependent variable Y. We are interested in the causal effect of X on Y, and we use the following equation:

$$E(Y) = X\beta \tag{5.10}$$

This can also be written as $Y = X\beta + \varepsilon$, where $\varepsilon = Y - E(Y)$ is called the error or disturbance term. Suppose further that there is endogeneity; that is, X also depends on Y:

$$E(X) = Y\gamma \tag{5.11}$$

What happens if we ignore the reciprocal part of the relationship in equation (5.11) and estimate β as if only equation (5.10) were true? In other words, we estimate β (incorrectly assuming that $\gamma = 0$) with the usual equation:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \tag{3.7}$$

To evaluate this estimator, we use the property of unbiasedness and therefore calculate its expected value:

$$
\begin{aligned}
E(b) &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \qquad (5.13)\\
&= E\left(\frac{\sum_{i=1}^{n} X_i (X_i\beta + \varepsilon_i)}{\sum_{i=1}^{n} X_i^2}\right)\\
&= \beta + \frac{\sum_{i=1}^{n} C(X_i,\varepsilon_i)}{\sum_{i=1}^{n} V(X_i)}\\
&= \beta + \text{Bias}
\end{aligned}
$$

where Bias = $\sum_{i=1}^{n} C(X_i,\varepsilon_i)/\sum_{i=1}^{n} V(X_i)$. Normally, the covariance of X_i and the disturbance term ε_i, $C(X_i,\varepsilon_i)$, is zero so that the bias term is zero. Thus the expected value of b is β and therefore unbiased. It is usually true that after we take into account X in predicting Y, the portion we have remaining (ε) is not correlated with X. However, in the present situation, after we take into account the effect of X, there is still some variation left over due to feedback from the causal effect of Y on X. Thus, endogeneity means that the second term in the last line of equation (5.13) will not generally be zero, and the estimate will be biased.

The direction of the bias depends on the covariance, since the variance of X is always positive. However, in the unusual cases where the variance of X is extremely large, it will overwhelm the covariance and make the bias term negligible. The text gives an example with a substantive interpretation of this bias term.

5.5 Assigning values of the Explanatory Variable

We pointed out in section 4.4 that the best controlled experiments have two advantages: control over the selection of observations and control over the assignment of values of the explanatory variables to units. We only discussed selection at that point. Now that we have analyzed omitted variable bias and the other methodological pitfalls in this chapter, we can address the issue of control over assignment.

In a medical experiment, a drug being tested and a placebo constitute the treatments, which are randomly assigned to patients. Basically the same situation exists here as with random selection of observations: random assignment is very useful with large numbers of obser-

Let us use the model in equation (4.1), but this time we have a very large number of observations and our two explanatory variables are perfect linear combinations of one another. In fact, to make the problem even more transparent, suppose that the two variables are the same, so that $X_1 = X_2$. We might have coded X_1 and X_2 as two substantively different variables (like gender and pregnancy), but in a sample of data they might turn out to be the same (if all women surveyed happened to be pregnant). Can we distinguish the causal effects of these different variables?

Note that equation (4.1) can be written as follows:

$$E(Y) = X_1\beta_1 + X_2\beta_2, \tag{4.3}$$

$$= X_1(\beta_1 + \beta_2)$$

As should be obvious from the second line of this equation, regardless of what $E(Y)$ and X_1 are, numerous values of β_1 and β_2 can satisfy it. (For example, if $\beta_1 = 5$ and $\beta_2 = -20$ satisfy equation (4.3), then so does $\beta_1 = -20$ and $\beta_2 = 5$.) Thus, although we now have many more observations than parameters, multicollinearity leaves us with the same problem as when we had more parameters than units: no estimation method can give us unique estimates of the parameters.

4.2 The Limits of Random Selection

We avoid selection bias in large-n studies if observations are randomly selected, because a random rule is uncorrelated with all possible explanatory or dependent variables.[2] Randomness is a powerful approach because it provides a selection procedure that is *automatically* uncorrelated with all variables. That is, with a large n, the odds of a selection rule correlating with any observed variable are extremely small. As a result, random selection of observations automatically eliminates selection bias in large-n studies. In a world in which there are many potential confounding variables, some of them unknown, randomness has many virtues for social scientists. If we have to abandon randomness, as is usually the case in political science research, we must do so with caution.

[2] We emphasize again that we should not confuse randomness with haphazardness. Random selection in this context means that every potential unit has an equal probability of selection into our sample and successive choices are independent, just as when names are picked out of a hat with replacements. This is only the simplest version of randomness, but all require specific probabilistic processes.

Controlled experiments are only occasionally constructed in the social sciences.[3] However, they provide a useful model for understanding certain aspects of the design of nonexperimental research. The best experiments usually combine random selection of observations and random assignments of values of the explanatory variables with a large number of observations (or experimental trials). Even though no experiment can solve the Fundamental Problem of Causal Inference, experimenters are often able to select their observations (rather than having them provided through social processes) and can assign treatments (values of the explanatory variables) to units. Hence it is worthwhile to focus on these two advantages of experiments: control over *selection of observations* and *assignment of values of the explanatory variables to units*. In practice, experimenters often do not select randomly, choosing instead from a convenient population such as college sophomores, but here we focus on the ideal situation. We discuss selection here, postponing our discussion of assignment of values of the explanatory variables until the end of chapter 5.

In qualitative research, and indeed in much quantitative research, random selection may not be feasible because the universe of cases is not clearly specified. For instance, if we wanted a random sample of foreign policy elites in the United States, we would not find an available list of all elites comparable to the list of congressional districts. We could put together lists from various sources, but there would always be the danger that these lists would have built-in biases. For instance, the universe for selection might be based on government lists of citizens who have been consulted on foreign policy issues. Surely such citizens could be considered to be members of a foreign policy elite. But if the research problem had to do with the relationship between social background and policy preferences, we might have a list that was biased toward high-status individuals who are generally supportive of government policy. In addition, we might not be able to study a sample of elites chosen at random from a list because travel costs might be too high. We might have to select only those who lived in the local region—thus possibly introducing other biases.

Even when random selection is feasible, it is not necessarily a wise technique to use. Qualitative researchers often balk (appropriately) at the notion of random selection, refusing to risk missing important cases that might not have been chosen by random selection. (Why study revolutions if we don't include the French Revolution?) Indeed, if we have only a small number of observations, random selection may not solve the problem of selection bias but may even be worse than

[3] For some examples, see Roth (1988), Iyengar and Kinder (1987), Fiorina and Plott (1978), Plott and Levine (1978), and Palfrey (1991).

other methods of selection. We believe that many qualitative researchers understand this point intuitively when they complain about what they perceive as the misguided preaching of some quantitative researchers about the virtues of randomness. In fact, using a very simple formal model of qualitative research, we will now prove that random selection of observations in small-*n* research will often cause very serious biases.

Suppose we have three units that have observations on the dependent variable of (High, Medium, Low), but only two of these three are to be selected into the analysis ($n = 2$). We now need a selection rule. If we let 1 denote a unit selected into the analysis and 0 denote an omitted unit, then only three selection rules are possible: (1,1,0), which means that we select the High and Medium choices but not the Low case, (0,1,1), and (1,0,1). The problem is that only the last selection rule, in which the second unit is omitted, is uncorrelated with the dependent variable.[4] Since random selection of observations is equivalent to a random choice of one of these three possible selection rules, random selection of units in this small-*n* example will produce selection bias with two-thirds probability! More careful selection of observations using a priori knowledge of the likely values of the dependent variable might be able to choose the third selection rule with much higher probability and thus avoid bias.

Qualitative researchers rarely resort explicitly to randomness as a selection rule, but they must be careful to ensure that the selection criteria actually employed do not have similar effects. Suppose, for example, that a researcher is interested in those East European countries with Catholic heritage that were dominated by the Soviet Union after World War II: Czechoslovakia, Hungary, and Poland. This researcher observes substantial variation in their politics during the 1970s and 1980s: in Poland, a well-organized antigovernment movement (Solidarity) emerged; in Czechoslovakia a much smaller group of intellectuals was active (Charter 77); while in Hungary, no such large national movement developed. The problem is to explain this discrepancy.

Exploring the nature of antigovernment movements requires close analysis of newspapers, recently declassified Communist Party documents, and many interviews with participants—hence, knowledge of the language. Furthermore, the difficulty of doing research in contemporary Eastern Europe means that a year of research will be required to study each country. It seems feasible, therefore, to study only two

[4] The (1,1,0) selection rule omits the low end of the scale (the Low unit), and the second (0,1,1) omits the unit at the high end (the High unit). Only the third case, in which "Medium" is not selected, is uncorrelated with the dependent variable.

countries for this work. Fortunately, for reasons unconnected with this project, the researcher already knows Czech and Polish, so she decides to study Charter 77 in Czechoslovakia and Solidarity in Poland. This is obviously different from random assignment, but at least the reason for selecting these countries is probably unrelated to the dependent variable. However, in our example it turns out that her selection rule (linguistic knowledge) *is* correlated with her dependent variable and that she will therefore encounter selection bias. In this case, a nonrandom, informed selection might have been better—if it were not for the linguistic requirement.

This researcher could avoid selection bias by forgetting her knowledge of Czech and learning Hungarian instead. But this solution will hardly seem an attractive option! In this observation, the more realistic alternative is that she use her awareness of selection bias to judge the direction of bias, at least partially correct for it, and qualify her conclusions appropriately. At the outset, she knows that she has reduced the degree of variance on her dependent variable in a systematic manner, which should tend to cause her to underestimate her causal estimates, at least on average (although other problems with the same research might change this).

Furthermore she should at least do enough secondary research on Hungary to know, for any plausible explanatory variable, whether the direction of selection bias will be in favor of, or against, her hypothesis. For example, she might hypothesize on the basis of the Czech and Polish cases that mass-based antigovernment movements arise under lenient, relatively nonrepressive communist regimes but not under strong, repressive ones. She should know that although Hungary had the most lenient of the East European communist governments, it lacked a mass-based antigovernment movement. Thus, if possible, the researcher should expand the number of observations to avoid selection bias; but even if more observations cannot be studied thoroughly, some knowledge of additional observations can at least mitigate the problem. A very productive strategy would be to supplement these two detailed case studies with a few much less detailed cases based on secondary data and, perhaps, a much more aggregate (and necessarily superficial) analysis of a large number of cases. If the detailed case studies produce a clear causal hypothesis, it may be much easier to collect information on just those few variables identified as important for a much larger number of observations across countries. (See section 4.3 for an analogous discussion and more formal treatment.) Another solution might be to reorganize the massive information collected in each of the two case studies into numerous observable implications of the theory. For example, if the theory that government repression suc-

cessfully inhibited the growth of antigovernment movements was correct, such movements should have done poorly in cities or regions where the secret police were zealous and efficient, as compared to those ares in which the secret police were more lax—controlling for the country involved.

4.3 Selection Bias

How should we select observations for inclusion in a study? If we are interviewing city officials, which ones should we interview? If we are doing comparative case studies of major wars, which wars should we select? If we are interested in presidential vetoes, should we select all vetoes, all since World War II, a random sample, or only those overridden by Congress? No issue is so ubiquitous early in the design phase of a research project as the question: which cases (or more precisely, which observations) should we select for study? In qualitative research, the decision as to which observations to select is crucial for the outcome of the research and the degree to which it can produce determinate and reliable results.

As we have seen in section 4.2, random selection is not generally appropriate in small-*n* research. But abandoning randomness opens the door to many sources of bias. The most obvious example is when we, knowing what we want to see as the outcome of the research (the confirmation of a favorite hypothesis), subtly or not so subtly select observations on the basis of combinations of the independent and dependent variables that support the desired conclusion. Suppose we believe that American investment in third world countries is a prime cause of internal violence, and then we select a set of nations with major U.S. investments in which there has been a good deal of internal violence and another set of nations where there is neither investment nor violence. There are other observations that illustrate the other combinations (large investment and no violence, or no small investment and large violence) but they are "conveniently" left out. Most selection bias is not as blatant as this, but since selection criteria in qualitative research are often implicit and selection is often made without any self-conscious attempt to evaluate potential biases, there are many opportunities to allow bias subtly to intrude on our selection procedures.[5]

[5] This example is a good illustration of what makes science distinctive. When we introduce this bias in order to support the conclusion we want, we are not behaving as social scientists ought to behave, but rather the way many of us behave when we are in political arguments in which we are defending a political position we cherish. We often select examples that prove our point. When we engage in research, we should try to get all

4.3.1 Selection on the Dependent Variable

Random selection with a large-*n* allows us to ignore the relationship between the selection criteria and other variables in our analysis. Once we move away from random selection, we should consider how the criteria used relate to each variable. That brings us to a basic and obvious rule: *selection should allow for the possibility of at least some variation on the dependent variable*. This point seems so obvious that we would think it hardly needs to be mentioned. How can we explain variations on a dependent variable if it does not vary? Unfortunately, the literature is full of work that makes just this mistake of failing to let the dependent variable vary; for example, research that tries to explain the outbreak of war with studies only of wars, the onset of revolutions with studies only of revolutions, or patterns of voter turnout with interviews only of nonvoters.[6]

We said in chapter 1 that good social scientists frequently thrive on anomalies that need to be explained. One consequence of this orientation is that investigators, particularly qualitative researchers, may select observations having a common, puzzling outcome, such as the social revolutions that occurred in France in the eighteenth century and those that occurred in France and China in the twentieth (Skocpol 1979). Such a choice of observations represents selection on the dependent variable, and therefore risks the selection bias discussed in this section. When observations are selected on the basis of a particular value of the dependent variable, nothing whatsoever can be learned about the causes of the dependent variable without taking into account other instances when the dependent variable takes on other values. For example, Theda Skocpol (1979) partially solves this problem in her research by explicitly including some limited information about "moments of revolutionary crisis" (Skocpol 1984:380) in seventeenth-century England, nineteenth-century Prussia/Germany, and nineteenth-century Japan. She views these observations as "control cases," although they are discussed in much less detail than her principal cases. The bias induced by selecting on the dependent variable does not imply that we should never take into account values of the dependent variable when designing research. What it does mean, as we

---

observations if possible. If selection is required, we should attempt to get those observations which are pivotal in deciding the question of interest, not those which merely support our position.

[6] In this section, we do not consider the possibility that a specific research project that is designed not to let the dependent variable change at all is part of a larger research program and therefore can provide useful information about causal hypotheses. We explain this point in section 4.4.

discuss below and in chapter 6, is that we must be aware of the biases introduced by such selection on the dependent variable and seek insofar as possible to correct for these biases.

There is also a milder and more common version of the problem of selection on the dependent variable. In some instances, the research design does allow variation on the dependent variable but that variation is truncated: that is, we limit our observations to less than the full range of variation on the dependent variable that exists in the real world. In these cases, something can be said about the causes of the dependent variable; but the inferences are likely to be biased since, if the explanatory variables do not take into account the selection rule, *any selection rule correlated with the dependent variable attenuates estimates of causal effects on average* (see Achen, 1986; King 1989: chapter 9). In quantitative research, this result means that numerical estimates of causal effects will be closer to zero than they really are. In qualitative research, selection bias will mean that the true causal effect is larger than the qualitative researcher is led to believe (unless of course the researcher is aware of our argument and adjusts his or her estimates accordingly). If we know selection bias exists and have no way to get around it by drawing a better sample, these results indicate that our estimate at least gives, on average, a lower bound to the true causal effect. The extent to which we underestimate the causal effect depends on the severity of the selection bias (the extent to which the selection rule is correlated with the dependent variable), about which we should have at least some idea, if not detailed evidence.

The cases of extreme selection bias—where there is by design no variation on the dependent variable—are easy to deal with: avoid them! We will not learn about causal effects from them. The modified form of selection bias, in which observations are selected in a manner related to the dependent variable, may be harder to avoid since we may not have access to all the observations we want. But fortunately the effects of this bias are not as devastating since we can learn something; our inferences might be biased but they will be so in a predictable way that we *can* compensate for. The following examples illustrate this point.

Given that we will often be forced to choose observations in a manner correlated with the dependent variable, and we therefore have selection bias, it is worthwhile to see whether we can still extract some useful information. Figure 4.1, a simple pictorial model of selection bias, shows that we can. Each dot is an observation (a person, for example). The horizontal axis is the explanatory variable (for example, number of accounting courses taken in business school). The vertical axis is the dependent variable (for example, starting salary in the first

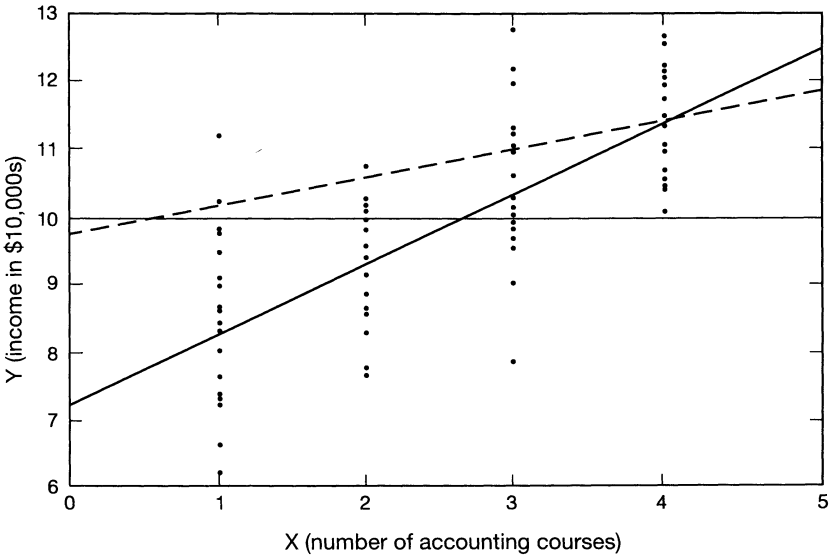

Figure 4.1 Selection Bias

full-time job, in units of $10,000). The regression line showing the relationship between these two variables is the solid line fit to the scatter of points. Each additional accounting course is worth on average about an additional $10,000 in starting salary. The scatter of points around this line indicates that, as usual, the regression line does not fit each student's situation perfectly. In figures like these, the *vertical* deviations between the points and the line represent the errors in predictions (given particular values of the explanatory variables) and are therefore minimized in fitting a line to the points.

Now suppose an incoming business-school student were interested in studying how he could increase his starting salary upon graduation. Not having learned about selection bias, this student decides to choose for study a sample of previous students composed only of those who did well in their first job—the ones who received jobs he would like. It may *seem* that if he wants to learn about how to earn more money it would be best to focus only on those with high earnings, but this reasoning is fallacious. For simplicity, suppose the choice included only those making at least $100,000. This sample selection rule is portrayed in figure 4.1 by a solid horizontal line at $Y = 10$, where only the points above the line are included in this student's study. Now, instead of fitting a regression line to all the points, he fits a line (the dashed line) only to the points in his sample. Selection bias exerts its effect by decreasing this line's slope compared to that of the solid line.

As a result of the selection bias, this student would incorrectly conclude that each additional accounting course is worth only about $5,000.

This is a specific example of the way in which we can underestimate a causal effect when we have selection on the dependent variable. Luckily, there *is* something our student can do about his problem. Suppose after this student completes business school, he gets bored with making money and goes to graduate school in one of the social sciences where he learns about selection bias. He is very busy preparing for comprehensive examinations, so he does not have the time to redo his study properly. Nevertheless, he does know that his starting salary would have increased by some amount significantly *more* than his estimate of $5,000 for each additional accounting class. Since his selection rule was quite severe (indeed it was deterministic), he concludes that he would have made more money in business if he had taken additional accounting classes—but having decided not to maximize his income (who would enter graduate school with that in mind?)—he is thankful that he did not learn about selection bias until his values had changed.

4.3.1.1 EXAMPLES OF INVESTIGATOR-INDUCED SELECTION BIAS

The problem just described is common in qualitative research (see Geddes 1990). It can arise from a procedure as apparently innocuous as selecting cases based on available data, if data availability is related to the dependent variable. For instance, suppose we are interested in the determinants of presidential involvement in significant foreign policy decisions during recent years and that we propose to study those decisions on which information about the president's participation in meetings is available. The problem with this research design is that the selection rule (information availability) is probably correlated with relatively low levels of presidential involvement (the dependent variable) since the more secret meetings, which will not be available to us, are likely to have involved the president more fully than those whose deliberations have become public. Hence the set of observations on which information is available will overrepresent events with lower presidential involvement, thus biasing our inferences about the determinants of presidential involvement.

The reasoning used in our business-school example can help us learn about the consequences of unavoidable selection bias in qualitative research. Suppose, in the study just mentioned, we were interested in whether presidents are more involved when the events entail threats of force than when no such threats were made. Suppose also that existing evidence, based on perhaps two dozen observations, indi-

cates that such a relationship does exist, but that its magnitude is surprisingly small. To assess the degree of selection bias in this research, we would first compile a list of foreign policy situations in which the president took action or made public pronouncements, regardless of whether we had any information on decision-making processes. This list would avoid one source of selection bias that we had identified: greater secrecy with respect to decision-making involving threats of force. Our new list would not be a complete census of issues in which the president was engaged, since it would miss covert operations and those on which no actions were taken, but it would be a larger list than our original one, which required information about decision-making. We could then compare the two lists to ascertain whether (as we suspect) cases on which we had decision-making information were biased against those in which force was used or threatened. If so, we could reasonably infer that the true relationship was probably even stronger than it seemed from our original analysis.

The problem of selection bias appears often in comparative politics when researchers need to travel to particular places to study their subject matter. They often have limited options when it comes to choosing what units to study since some governments restrict access by foreign scholars. Unfortunately, the refusal to allow access may be correlated with the dependent variable in which the scholar is interested. A researcher who wanted to explain the liberalization of authoritarian regimes on the basis of the tactics used by dissident groups might produce biased results, especially if she only studied those places that allowed her to enter, since the factors that led the regime to allow her in would probably be correlated with the dependent variable, liberalization. We obviously do not advise clandestine research in inhospitable places. But we do advise self-conscious awareness of these problems and imagination in finding alternative data sources when on-site data are unavailable. Recognition of these difficulties could also lead to revision of our research designs to deal with the realities of scholarly access around the world. If no data solution is available, then we might be able to use these results on selection bias at least to learn in which direction our results will be biased—and thus perhaps provide a partial correction to the inevitable selection bias in a study like this. That is, if selection bias is unavoidable, we should analyze the problem and ascertain the direction and, if possible, the magnitude of the bias, then use this information to adjust our original estimates in the right direction.

Selection bias is such an endemic problem that it may be useful to consider some more examples. Consider a recent work by Michael Porter (1990). Porter was interested in the sources of what he called

"competitive advantage" for contemporary industries and firms. He designed a large-scale research project with ten national teams to study the subject. In selecting the ten nations for analysis, he chose, in his words, "ones that already compete successfully in a range of such industries, or, in the case of Korea and Singapore, show signs of an improving ability to do so" (Porter 1990:22). In his eagerness to explore the puzzle that interested him, Porter intentionally selected on his dependent variable, making his observed dependent variable nearly constant. As a result, any attempts by Porter, or anyone else using these data at this level of analysis, to explain variations in success among his ten countries will produce seriously biased causal effects.

But what Porter did—try to determine the circumstances and policies associated with competitive success—was somewhat related to Mill's method of agreement. This method is not a bad first attempt at the problem, in that it enabled Porter to develop some hypotheses about the causes of competitive advantage by seeing what these nations have in common; however, his research design made it impossible to evaluate any individual causal effect.

More serious is the logical flaw in the method: without a control group of nations (that is, with his explanatory variable set to other values), he cannot determine whether the absence of the hypothesized causal variables is associated with competitive failure. Thus, he has no way of knowing whether the conditions he has associated with success are not also associated with failure. In his provocative work, Porter has presented a fascinating set of *hypotheses* based on his cases of success, but without a range of competitive successes and failures (or selection based on something other than his dependent variable) he has no way of knowing whether he is totally right, completely wrong, or somewhere in between.[7]

A striking example of selection bias is found in the foreign policy literature dealing with deterrence: that is, "the use of threats to induce the opponents to behave in desirable ways" (Achen and Snidal 1989: 151). Students of deterrence have often examined "acute crises"—that is, those that have not been deterred at an earlier stage in the process of political calculation, signalling, and action. For descriptive pur-

[7] Porter claims to have numerous examples of countries which were not successful; however, these are introduced in his analyses by way of selectively chosen anecdotes and are not studied with similar methods as his original ten. When nonsystematically selecting supporting examples from the infinite range of supporting and nonsupporting possibilities, it is much too easy to fool ourselves into finding a relationship when none exists. We take no position on whether Porter's hypotheses are correct and only wish to point out that the information needed to make this decision must be collected more systematically.

poses, there is much to be said for such a focus, at least initially: as in Porter's emphasis on competitive success, the observer is able to describe the most significant episodes of interest and may be enabled to formulate hypotheses about the causes of observed outcomes. But as a basis for inference (and without appropriate corrections), such a biased set of observations is seriously flawed because instances in which deterrence has worked (at earlier stages in the process) have been systematically excluded from the set of observations to be analyzed. "When the cases are then misused to estimate the success rate of deterrence, the design induces a 'selection bias' of the sort familiar from policy-evaluation research" (Achen and Snidal 1989:162).

4.3.1.2 EXAMPLES OF SELECTION BIAS INDUCED BY THE WORLD

Does choosing a census of observations, instead of a sample, enable us to avoid selection bias? We might think so since there was apparently no selection at all, but this is not always correct. For example, suppose we wish to make a descriptive inference by estimating the strength of support for the Liberal party in New York State. Our dependent variable is the percent of the vote in New York State Assembly districts cast for the candidate (or candidates) endorsed by the Liberal party. The problem here is that the party often chooses not to endorse candidates in many electoral districts. If they do not endorse candidates in districts where they feel sure that they will lose (which seems to be the case), then we will have selection bias even if we choose every district in which the Liberal party made an endorsement. *The selection process in this example is performed as part of the political process we are studying, but it can have precisely the same consequences for our study as if we caused the problem ourselves.*

This problem of bias when the selection of cases is correlated with the dependent variable is one of the most general difficulties faced by those scholars who use the historical record as the source of their evidence, and they include virtually all of us. The reason is that the processes of "history" differentially select that which remains to be observed according to a set of rules that are not always clear from the record. However, it is *essential* to discover the process by which these data are produced. Let us take an example from another field: some cultures have created sculptures in stone, others in wood. Over time, the former survive, the latter decay. This pattern led some European scholars of art to underestimate the quality and sophistication of early African art, which tended to be made of wood, because the "history" had selectively eliminated some examples of sculpture while maintaining others. The careful scholar must always evaluate the possible selection biases in the evidence that is available: what kinds of events are

likely to have been recorded; what kinds of events are likely to have been ignored?

Consider another example. Social scientists often begin with an end point that they wish to "explain"—for example, the peculiar organizational configurations of modern states. The investigator observes that at an early point in time (say, A.D. 1500) a wide variety of organizational units existed in Europe, but at a later time (say, A.D. 1900), all, or almost all, important units were national states. What the researcher should do is begin with units in 1500 and explain later organizational forms in terms of a limited number of variables. Many of the units of analysis would have disappeared in the interim, because they lost wars or were otherwise amalgamated into larger entities; others would have survived. Careful categorization could thus yield a dependent variable that would index whether the entity that became a national state is still in existence in 1900; or if not, when it disappeared.

However, what many historical researchers inadvertently do is quite different. They begin, as Charles Tilly (1975: 15) has observed, by doing *retrospective* research: selecting "a small number of West European states still existing in the nineteenth and twentieth centuries for comparison." Unfortunately for such investigators, "England, France, and even Spain are *survivors* of a ruthless competition in which most contenders lost." The Europe of 1500 included some five hundred more or less independent political units, the Europe of 1900 about twenty-five. The German state did not exist in 1500, or even 1800. Comparing the histories of France, Germany, Spain, Belgium, and England (or, for that matter, any other set of modern Western European countries) for illumination on the processes of state-making weights the whole inquiry toward a certain kind of outcome which was, in fact, quite rare.

Such a procedure therefore selects on the basis of one value of the dependent variable—survival in the year 1900. It will bias the investigator's results, on average reducing the attributed effects of explanatory variables that distinguish the surviving states from their less durable counterparts. Tilly and his colleagues (1975), recognizing the selection bias problem, moved from a *retrospective* toward a *prospective* formulation of their research problem. Suppose, however, that such a huge effort had not been possible, or suppose they wished to collect the best available evidence in preparation for their larger study. They could have reanalyzed the available retrospective studies, inferring that those studies' estimates of causal effects were in most observations biased downward. They would need to remember that, even if the criteria described above do apply exactly, any one application might overestimate or underestimate the causal effect. The best

guess of the true causal effect, based on the flawed retrospective studies, however, would be that the causal effects were underestimated at least on average—if we assume that the rules above do apply and the criteria for selection were correlated with the dependent variable.

4.3.2 Selection on an Explanatory Variable

Selecting observations for inclusion in a study according to the categories of the key causal explanatory variable causes no inference problems. The reason is that our selection procedure does not predetermine the outcome of our study, since we have not restricted the degree of possible variation in the dependent variable. By limiting the range of our key causal variable, we may limit the generality of our conclusion or the certainty with which we can legitimately hold it, but we do not introduce bias. By selecting cases on the basis of values of this variable, we can control for that variable in our case selection. Bias is not introduced even if the causal variable is correlated with the dependent variable since we have already controlled for this explanatory variable.[8] Thus, it is possible to avoid bias while selecting on a variable that is correlated with the dependent variable, so long as we control for that variable in the analysis.

It is easy to see that selection on an explanatory variable causes no bias by referring again to figure 4.1. If we restricted this figure to exclude all the observations for which the explanatory variable equaled one, the logic of this figure would remain unchanged, and the correct line fit to the points would not change. The line would be somewhat less certain, since we now have fewer observations and less information to bear on the inference problem, but on average there would be no bias.[9]

Thus, one can avoid bias by selecting cases based on the key causal variable, but we can also achieve the same objective by selecting according to the categories of a control variable (so long as it is causally prior to the key causal variable, as all control variables should be). Experiments almost always select on the explanatory variables. Units are created when we manipulate the explanatory variables (administering a drug, for example) and watch what happens to the dependent variable (whether the patient's health improves). It would be difficult to select on the dependent variable in this case, since its value is not even

[8] In general, selection bias occurs when selecting on the dependent variable, after taking into account (or controlling for) the explanatory variables. Since one of these explanatory variables is the method of selection, we control for it and do not introduce bias.

[9] The inference would also be less certain if the range of values of the explanatory variables were limited through this selection. See section 6.2.

known until after the experiment. However, most experiments are far from perfect, and we can make the mistake of selecting on the dependent variable by inadvertently giving some treatments to patients based on their expected response.

For another example, if we are researching the effect of racial discrimination on black children's grades in school, it would be quite reasonable to select several schools with little discrimination and some with a lot of discrimination. Even though our selection rule will be correlated with the dependent variable (blacks get lower grades in schools with more discrimination), it will not be correlated with the dependent variable *after* taking into account the effect of the explanatory variables, since the selection rule is determined by the values of one of the explanatory variables.

We can also avoid bias by selecting on an explanatory variable that is irrelevant to our study (and has no effect on our dependent variable). For example, to study the effects of discrimination on grades, suppose someone chose all schools whose names begin with the letter "A." This, of course, is not recommended, but it would cause no bias as long as this irrelevant variable is not a proxy for some other variable that is correlated with the dependent variable.

One situation in which selection by an irrelevant variable can be very useful involves secondary analysis of existing data. For example, suppose we are interested in what makes for a successful coup d'etat. Our key hypothesis is that coups are more often successful when led by a military leader rather than a civilian one. Suppose we find a study of attempted coups that selected cases based on the extent to which the country had a hierarchical bureaucracy before a coup. We could use these data even if hierarchical bureaucratization is irrelevant to our research. To be safe, however, it would be easy enough to include this variable as a control in our analysis of the effects of military versus civilian leaders. We would include this control by studying the frequency of coup success for military versus civilian leaders in countries with and then without hierarchical bureaucratization. The presence of this control will help us avoid selection bias and its causal effect will indicate some possibly relevant information about the process by which the observations were really selected.

4.3.3 Other Types of Selection Bias

In all of the above examples, selection bias was introduced when the units were chosen according to some rule correlated with the dependent variable or correlated with the dependent variable after the ex-

planatory variables were taken into account. With this type of selection effect, estimated causal effects are always underestimates. This is by far the most common type of selection bias in both qualitative and quantitative research. However, it is worth mentioning another type of selection bias, since its effects can be precisely the opposite and cause *overestimation* of a causal effect.

Suppose the causal effect of some variable varies over the observations. Although we have not focused on this possibility, it is a real one. In section 3.1, we defined a causal effect for a single unit and allowed the effect to differ across units. For example, suppose we were interested in the causal effect of poverty on political violence in Latin American countries. This relationship might be stronger in some countries, such as those with a recent history of political violence, than in others. In this situation, where causal effects vary over the units, a selection rule correlated with the size of the causal effect would induce bias in estimates of *average* causal effects. Hence if we conducted our study only in countries with recent histories of political violence but sought to generalize from our findings to Latin America as a whole, we would be likely to overestimate the causal effect under investigation. If we selected units with large causal effects and averaged these effects during estimation, we would get an overestimate of the average causal effect. Similarly, if we selected units with small effects, the estimate of the average causal effect would be smaller than it should be.

4.4 Intentional Selection of Observations

In political science research, we typically have no control over the values of our explanatory variables; they are assigned by "nature" or "history" rather than by us. In this common situation, the main influence we can have at this stage of research design is in selecting cases and observations. As we have seen in section 4.2, when we are able to focus on only a small number of observations, we should rarely resort to random selection of observations. Usually, selection must be done in an *intentional* fashion, consistent with our research objectives and strategy.

Intentional selection of observations implies that we know in advance the values of at least some of the relevant variables, and that random selection of observations is ruled out. We are least likely to be fooled when cases are selected based on categories of the explanatory variables. The research itself, then, involves finding out the values of the dependent variable. However, in practice, we often have fragmentary evidence about the values of many of our variables, even before