CHAPTER 3

How the Cases You Choose Affect the Answers You Get

Selection Bias and Related Issues

Comparative politics, like other subfields in political science, has norms and conventions about what constitutes an appropriate research strategy and what kind of evidence makes an argument persuasive. Although the norm has begun to change, for many years one of our most durable conventions was the selection of cases for study from one end of the outcome continuum we wished to explain.[1] That is, if we want to understand something—for example, revolution—we select one or more occurrences and subject them to scrutiny to see if we can identify antecedent events or characteristics as causes.

Most graduate students learn in the statistics courses forced upon them that such selection on the dependent variable often leads to wrong answers, but few remember why, or what the implications of violating this rule might be for their own work. And so, comparativists often ignore or forget about it when undertaking or assessing nonquantitative research.

This chapter demonstrates the consequences of violating the rule. It does so by comparing the conclusions reached in several influential studies based on cases selected on the dependent variable with retests of the same arguments using samples not correlated with the outcome. All the studies discussed in this chapter are intelligent, plausible, insightful, and possibly correct in their knowledge claims. All have been advanced by highly respected social scientists. The effort here is not to discredit arguments or belittle authors—who are, after all, working within accepted conventions—but to demonstrate the deficiencies of the conventions themselves.

1. Comparative politics is not the only field bedeviled by problems with selection bias (see Achen and Snidal 1989).

These conventions affect not only authors but readers of comparative politics. Authors, including some of those discussed below, are frequently aware of the tentativeness of the evidence supporting their arguments and indicate their awareness in the caveats they attach to them. Readers, however, tend to ignore the caveats and give greater weight to unsystematic evidence than it deserves. Many studies in which authors have carefully hedged their explanatory claims are discussed in seminars, cited in literature reviews, and summarized in qualifying exams as though the tentative arguments advanced were supported by solid evidence. The purpose of this chapter is as much to decrease the credulity of readers as to increase the sophistication of researchers.

The message of the chapter is not that the examination of cases selected because they have experienced a particular outcome is never warranted, but rather that the analyst should understand what can and cannot be accomplished with cases selected for this reason. Some kinds of tests of conditions proposed as necessary or sufficient for explaining outcomes can be carried out using only cases that have experienced an outcome, although assessment of what Braumoeller and Goertz (2000) refer to as trivialness requires at least some information about the rest of the universe of cases.[2]

The close examination of an anomalous case with a particular outcome can also serve a useful role in either generating a proposed revision of current theory or suggesting domain conditions not previously understood. A test of the proposed revision or domain condition would require examining a wider range of cases, however. Although the proposal of a revision is a useful

2. As Dion (1998) has pointed out, selection on the dependent variable does not undermine tests of "necessary but not sufficient" or "necessary and sufficient" arguments. Braumoeller and Goertz (2000) propose a series of tests that, taken together, would increase confidence in a necessary or sufficient argument. Carrying out these tests requires: (1) being able to estimate the error in the measurement of both proposed causes and effects; (2) including enough cases selected to have the outcome so that an appropriate statistical test can reject the null hypothesis (with no measurement error, the minimum number is seven; as measurement error increases, so does the required number of cases); (3) collecting enough information about the full universe of cases to assure oneself that there is enough variation in both purported cause and outcome to avoid trivialness. The issue of trivialness is discussed below. It refers to proposed necessary conditions that are theoretically meaningless because they vary little if at all. Braumoeller and Goertz note, for example, that the argument that democratic dyads are necessary for peace is trivial before 1800 because there were no democratic dyads then.

contribution to knowledge building, the revision should not be accepted until it has been tested and confirmed on a representative sample of cases.

The Nature of the Problem

The adverse effects of selecting cases for study on the dependent variable stem from the logic of inference. When one sets out to explain why countries A and B have, say, developed more rapidly than countries C through I, one is implicitly looking for some antecedent factors *X* through *Z* that countries A and B possess in greater degree than do countries C through I. The crux of the difficulty that arises when cases are selected on the dependent variable is that if one studies only countries A and B, one can collect only part of the information needed, namely, the extent of factors *X* through *Z* in countries A and B . Unless one also studies countries C through I (or a sample of them) to make sure they have less of *X* through *Z,* one cannot know whether the factors identified really vary with the outcome under investigation.

The problem becomes more obvious when shown in graphs rather than expressed in words. Suppose a universe of developing countries A through I, where A and B are among the fastest growing. On the basis of an intensive study of A and B, one concludes that factor *X* is the cause of their success. In concluding this, one implicitly assumes that if countries C through I were examined, they would turn out to have less of factor *X* than do A and B, and that one would observe the relationship shown in figure 3.1.

Yet if one examines only countries A and B, it is possible that the full range of cases would look more like one of the scatterplots in figure 3.2. That is, it is possible that there is no relationship

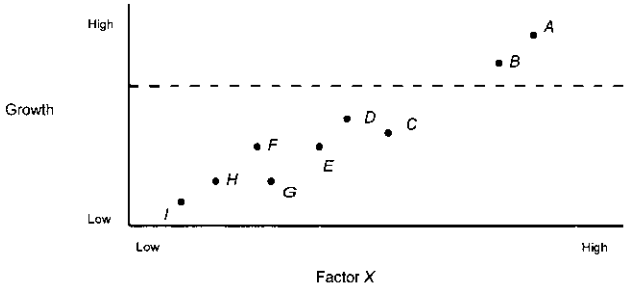

Fig. 3.1. Assumed relationship between factor *X* and growth

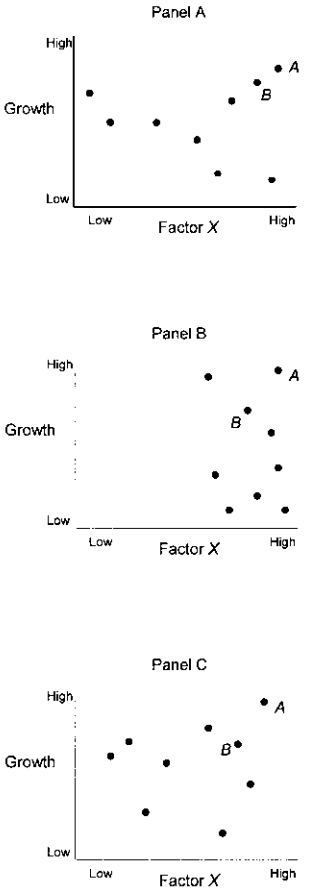

Fig. 3.2. Alternative possible relationships between factor *X* and growth

between *X* and the rate of development. The only things that can actually be explained using a sample selected on the dependent variable are differences among the selected cases.

When one looks only at the cases above the broken line in figure 3.1, two kinds of mistaken inference can occur. The first involves jumping to the conclusion that any characteristic that the selected cases share is a cause. The other involves inferring that relationships (or absence of relationships) between variables *within* the selected set of cases reflect relationships in the entire population of cases.

In the statistical literature, attention has focused on the second kind of faulty inference (Achen 1986; King 1989). If the true relationship between factor *X* and the dependent variable is that

shown in figure 3.1 but one selects cases in a manner that results in the examination only of cases located above the broken line, statistical procedures carried out on the selected cases may indicate that no relationship exists or even that the relationship is the opposite of the true one. Selection on the dependent variable biases statistical results toward finding no relationship even when one does, in fact, exist.

In nonquantitative work, however, the first kind of faulty inference is at least as common as the second. If the main causes of the dependent variable are factors *R* through *T,* not including *X,* and one selects cases from one end of the dependent variable, *X* may appear to be important in the selected sample either because of random variation or because it explains some of the differences among cases still remaining in the data set even after the selection has limited it (or because it is correlated with some other factor that explains the remaining differences). In the former situation, the true relationship might look like one of the panels in figure 3.2, but the analyst—on the basis of bits and pieces of information rather than a systematic check—assumes that cases C through I are located in the lower left quadrant and concludes that factor *X* causes the outcome of interest even though, in fact, no relationship exists. In the latter situation, factor *X* makes a minor contribution to the outcome, but the analyst overestimates its importance. An example should help to make these points clearer.

A Straightforward Case of Selection on the Dependent Variable

Analysts trying to explain why some developing countries have grown so much more rapidly than others regularly select a few successful new industrializing countries (NICs) for study. Prior to the debt crisis, which began in 1982, the cases most often examined were Taiwan, South Korea, Singapore, Brazil, and Mexico (see, e.g., Haggard 1990). In all these countries, during the periods of most rapid growth, governments exerted extensive control over labor and prevented most expressions of worker discontent. Having noted this similarity, analysts argue that the repression, co-optation, discipline, or weakness of labor contributes to high growth. Chalmers Johnson (1987, 149), for example, asserts that weak unions and "federations of unions devoid of all but token political power are real comparative advantages in international

economic competition." Frederic Deyo (1984, 1987) argues that an export-led growth strategy depends on cheap skilled labor and, consequently, a disciplined and quiescent labor force. Hagen Koo (1987) claims that labor control is needed in order to attract foreign investment.[3]

These claims draw additional plausibility from their convergence with arguments made in studies aimed not at explaining growth but at understanding authoritarian interventions in the more developed countries of Latin America. Among the best known of these is Guillermo O'Donnell's argument (1973) that the transition from the easy stage of import-substitution industrialization to a more capital-intensive stage creates a need for reduced consumption and, hence, a demand for the repression of labor.[4] In the same vein, Fernando Henrique Cardoso (1973a) and Peter Evans (1979) argue that labor repression helps attract foreign investment.

Whatever the details of the argument, many scholars who have studied the NICs seem to agree that repression or co-optation of the labor force contributes to growth. Taiwan, South Korea (especially between 1961 and 1986), Singapore (after 1968), Brazil (1967–81), and Mexico (before 1982) all had repressed and/or co-opted labor forces and relatively high growth rates. In other words, all have the outcome of interest and all exhibit another common trait—labor repression—so analysts conclude that labor repression has caused the outcome.

But that conclusion is unwarranted. Perhaps there are other countries in which labor suffers at least as much repression as in the high-growth countries examined but that have failed to prosper. In order to establish the plausibility of the claim that labor repression contributes to development, it would be necessary to select a sample of cases without reference to their position on the dependent variable (growth), rate each case on its level of labor repression, and show that, on average, countries with higher levels of repression grow faster.

To be persuasive, theories must be tested on at least a few cases other than those examined in the initial development of the

3. Haggard (1986, 354–56) provides a careful and nuanced review of several of these arguments.

4. The dependent variable in O'Donnell's study is regime type, not growth, and its research design is exemplary. O'Donnell compared the two countries that had experienced military intervention with a set of other Latin American countries that, at the time he wrote, remained democratic.

idea. At the stage of theory development, it is virtually impossible to avoid "overfitting," that is, tailoring arguments to fit the circumstances found in particular cases. Testing arguments on other cases allows the analyst to discover which factors proposed as possible causes during the discovery stage of theory building really do have general causal influence and which should, in the context of a general argument, be thought of as part of the "error term." The "error term" contains all those serendipitous, conjunctural, and other kinds of factors that contribute to particular outcomes in particular cases but that do not *systematically* influence outcomes.

Domain of the Argument

To test this or any other hypothesis, one must first identify the universe of cases to which the hypothesis should apply and then find or develop measures of the hypothesized causes and effects. The theory or hypothesis being tested determines the appropriate unit of analysis and the universe of potential observations.

If a theory suggests a relationship between some cause and individual behavior, the test of hypotheses derived from that theory should be based on observations of individuals. Where the unit of analysis is the individual, valid inferences can often be made in studies of single countries or even single towns, because, unless the town has been chosen precisely because the particular kind of individual behavior to be explained prevails within it, observing a range of individuals within a town does not entail selection on the dependent variable. The full range of individual variation may well occur within a town. Thus, for example, the research design used in William Sheridan Allen's *The Nazi Seizure of Power* (1973) avoids selection bias by including both individuals who embraced Nazism and those who resisted, and also by including change in individual attitudes over time.[5]

If, however, the hypothesis predicts country-level outcomes, as those linking labor repression and growth usually do, one should test it on a set of countries that reflects a reasonable range

5. In his critique of King, Keohane, and Verba (1994), Rogowski (1995) has noted that Allen's thoughtful study of one town in which Nazism enjoyed an early and substantial success deepens and enriches our understanding of the rise of Nazism. In the comparative field, we are inclined to equate cases automatically with territorial entities, but the unit of analysis used by Allen is clearly the individual, not the town, and thus he did not select on the dependent variable.

of variation on the country-level outcome. In short, the cases on which an argument is tested should reflect the level of analysis at which the argument is posed.

In everyday language, a case is a single entity, most often a country, but possibly a city, region, agency, administration, social movement, party, revolution, election, policy decision, or virtually anything else that involves interacting human beings. The more technical definition of a case is a unit within which each variable measured takes on only one value or is classified in only one category (Eckstein 1975). Many everyday language case studies include multiple technical cases, otherwise known as observations. Much of the disagreement in the literature over the usefulness of case studies has arisen from a confusion between, on the one hand, the everyday usage of the word *case* to mean (usually) a country; and, on the other hand, the more technical usage of *case* to mean an observation—the sense intended by those who give methodological advice, such as King, Keohane, and Verba (1994).

The appropriate universe of observations on which to test a hypothesis depends on the domain implied by the hypothesis. In other words, the domain depends on the substantive content of the theory or hypothesis itself, not necessarily on the author's statements about where the argument should apply. If an analyst proposes a theory about the effects of industrialization on late developing democratic countries, then tests of the theory can and should be carried out on a sample of countries drawn from the universe of *all* late developing democratic countries. Theories can contain substantive elements that limit their domain to particular regions of the world or time periods, and, if so, those limitations should be kept in mind during testing. Theories are not, however, automatically limited to the domain within which they were first proposed. Authors sometimes fail to realize that their arguments might apply to countries with which they are unfamiliar.

Well-intentioned scholars can disagree about what constitutes the appropriate domain of a theory, but their disagreements should derive from different interpretations of the implications of the theory. Tests of hypotheses in controversial domains can be useful in establishing clearer limits to the domain, extending it, and suggesting new hypotheses about why the domain has the limits it does. It is also legitimate to test arguments in domains outside those implied by theories to see whether the theories

have greater generality than their creators realized, though negative results in such tests fail to disconfirm the argument within its original domain.

If the whole universe of cases is too large to study, examination of a random sample is usually recommended as a means of ensuring that the criteria of selection do not correlate with the dependent variable. One can, however, make valid inferences from any sample selected in a way that does not inadvertently result in a set of cases clustered at one end of the outcome continuum. Moreover, randomization does not guarantee the absence of correlation. If, at a particular time, the universe contains only cases that have passed a certain threshold of success because "nature" has in some fashion weeded out the others, then even random or total samples will, in effect, have been selected on the dependent variable. If, for example, potential states that failed to adopt a given military innovation in the fifteenth century were later defeated and incorporated into other states, one would not be able to find evidence of the importance of this innovation by examining a random sample of the states that existed in the eighteenth century. All surviving states would have the innovation.[6]

Some theories have implications that apply to only one end of the dependent variable. To test hypotheses based on these implications, the analyst must, of course, choose cases from the relevant part of the outcome continuum. This may appear at first glance to entail selection on the dependent variable, but it does not. The outcome relevant for the test of a particular implication is the outcome predicted by this hypothesis about that implication, not the outcome explained by the theory. The full range of variation in the outcome predicted by the hypothesis may be contained at one end of the outcome predicted by the theory. For example, one of the implications of the cadre-interests argument described in chapter 2 is that military governments are more likely to negotiate their extrication from power than are personalist regimes. One way to test this implication is to compare the incidence of negotiation by different kinds of dictatorship during the years in which breakdown occurs. In other words, only regimes that had experienced breakdown would be included in the test (one end of the breakdown versus

6. An extensive and thought-provoking discussion of selection by nature can be found in Przeworski and Limongi (1993).

persistence outcome continuum), but the hypothesis actually being tested is about the incidence of negotiation during transitions, not about the causes of breakdown. The outcome continuum relevant for testing this hypothesis is the negotiation versus no negotiation continuum, not the breakdown versus persistence continuum.

For the hypothesis that labor repression contributes to growth, different arguments about the specific reasons a weak labor force might have this effect imply different domains for the argument. One possibility is that the domain should simply include all developing countries. In one of the tests of the argument below, I have included all developing countries for which the Penn World Tables collected data between 1970 and 1982, except those with communist governments, those embroiled in civil war for more than a third of the period covered, and those that are extremely small (fewer than 500,000 inhabitants).[7] Communist countries are excluded because the various theories apply only to countries with capitalist or mixed economies. The other exclusions involve countries with characteristics not related to labor repression that could be expected to affect greatly their growth rates and thus might distort the apparent relationship between labor repression and growth. In the second test, I narrow the domain to conform to arguments associated with O'Donnell and others who expect labor repression to contribute to growth once a certain threshold of development has been reached.

Measurement

The outcome to be explained, growth rate, presents no measurement problems; various measures are readily available. For this test, I used the Penn World Tables to calculate growth in GDP per capita between 1970 and 1982, since most of the studies of development strategies focus on the period before the debt crisis. A further test of the hypothesis that included economic performance in the far more adverse post-1982 international economic environment would also be interesting and useful.

The hypothesized cause—labor repression, co-optation, or quiescence—is more difficult to measure. Standard indicators are not available, and labor repression can take different forms

7. Developing countries are defined as those with per capita income below $4,200 in 1979. This cut-off point excludes wealthy oil exporters (per capita income above $4,200), Saudi Arabia, Kuwait, Libya, and the United Arab Emirates.

in different contexts, for example, state co-optation in one country and private violence against workers in another. To deal with this difficulty, I developed criteria for ranking each country on labor repression, using the *Country Reports on Human Rights Practices* prepared for congressional committees on foreign relations (U.S. Department of State, 1979–83), *Amnesty International Annual Reports* (1973–83), and many studies of labor in particular countries.

Eighty-four developing countries were given scores between zero and one for every year between 1970 and 1981 on five factors expected to contribute to the ability of workers to defend their interests:

- The extent to which unions are legal and free to function
- The autonomy of unions from government or ruling-party control or manipulation
- The right to bargain collectively and to strike
- The degree of political participation allowed to workers and the organizations that represent them
- Freedom from violence, arbitrary arrest, and other forms of repression

When these factors are combined, possible scores range from zero to five, with high scores indicating extreme control and repression and low scores reflecting freedom to organize, independence from ruling parties, legal protection of the right to bargain and strike, freedom to participate in politics, and protection from violence and repression. Countries with very low scores include Fiji, Mauritius, and Jamaica. The highest scorers are Uganda, Haiti, and Iraq. The countries included and their average labor repression scores are shown in appendix B.

In countries that experienced regime changes, policies toward labor usually changed along with the government. Yearly scoring of each country allowed those changes to be tracked. The coding sheet that was used to keep track of information while consulting multiple sources is shown in appendix B, along with the coding scheme. The coding scheme gives careful rules for translating the information gathered into numbers.

The purpose of coding sheets and coding rules, discussed at greater length in chapter 4, is to help make sure that the same factors are assessed in every case and that they are all judged

using the same criteria. While it is obvious that some such aid to memory is required for dealing with eighty-four cases, explicit coding rules also increase the precision of studies that focus on only a few cases. I would urge getting into the habit of writing down explicit coding rules, no matter what the number of observations. It helps the analyst stick to the same rules across countries and time, and it also helps readers understand exactly what the analyst means when she makes assessments of key causal factors. The phrase *labor repression* no doubt has somewhat different connotations for scholars with different areas of expertise, but the person who has read the coding scheme in appendix B will have a very clear idea of what is meant by the term here.

Although the indicator of labor repression created in this way is an imperfect measure of a complex set of phenomena, and experts might have small disagreements about the placement of a few cases, this measure is at least as precise as the verbal descriptions available in the literature. It seems, therefore, adequate to the present task of demonstrating a methodological point.

Tests of the hypothesis linking labor repression to growth using these data are shown in figures 3.3, 3.4, 3.5, and 3.6. Figure 3.3 shows the relationship between average labor repression and average growth from 1970 to 1981 for the sample of NICs most frequently studied (Taiwan, South Korea, Singapore, Brazil, and Mexico). This scatterplot reflects the most commonly chosen research strategy for studying the NICs in the 1970s and 1980s. It shows that repression and growth were both relatively high in all five countries. Analysts assumed, without checking carefully, that most of the cases they had not examined would lie in the lower left quadrant of the figure. From data like these—but in verbal form—researchers have concluded that labor repression contributes to growth. The plot shown here actually lends some plausibility to the argument because, using the quantitative measure of labor repression I created, it is possible to show small differences in labor repression that are not discernible in the verbal descriptions. Original statements of the argument did not distinguish levels of repressiveness among these cases of relatively high repression.

Note that the faulty inference expressed in the literature on the NICs is the opposite of the one that a thoughtless analyst using statistical methods would have drawn. A number cruncher might have concluded, on the basis of these data points, that

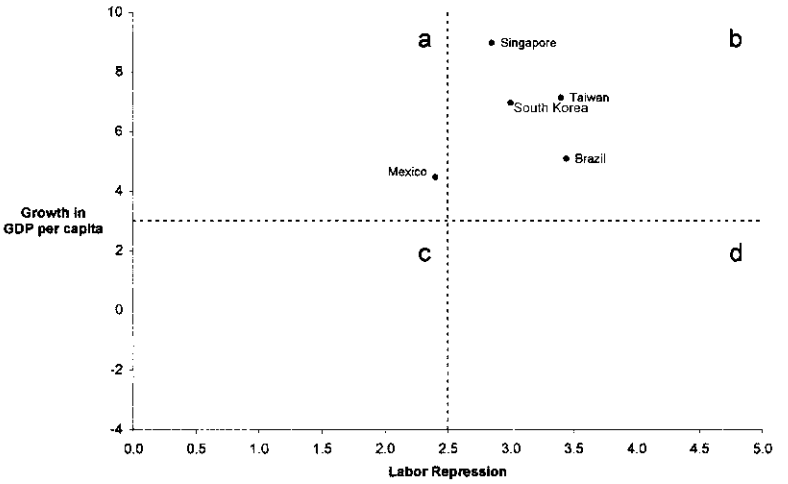

Fig. 3.3. Labor repression and growth in the most frequently studied cases, 1970–81. (GDP per capita from Penn World Tables.)

repression did not cause growth, because the variance in repression explained little of the variance in growth rate within this high-growth sample; on the other hand, the nonquantitative comparativist would conclude that since all cases are high on both growth and repression, repression must be a cause of growth. But, in fact, no conclusion can be drawn from figure 3.3. It simply contains too little information.

Scholars working on East Asia, where the fastest-growing NICs have historically been located, played an important role in developing the argument linking labor repression to growth. If, rather than selecting the five industrializing countries most frequently described in the literature, we examine the cases most familiar to East Asia specialists, it appears that repression does indeed contribute to growth, as shown in figure 3.4.

Based on an image of the world drawn from a few countries in one part of the world, some analysts advanced general arguments about the role of labor repression in growth, implying that the relationship that seemed apparent in Asia would also characterize the entire developing world. Such an inference cannot be justified, because the selection of cases by virtue of their location in East Asia biases the sample just as surely as would selection explicitly based on growth rates. This is so because, on average, growth rates in East Asia are unusually high. (See table 3.1.) Geographical area is correlated with growth, and consequently

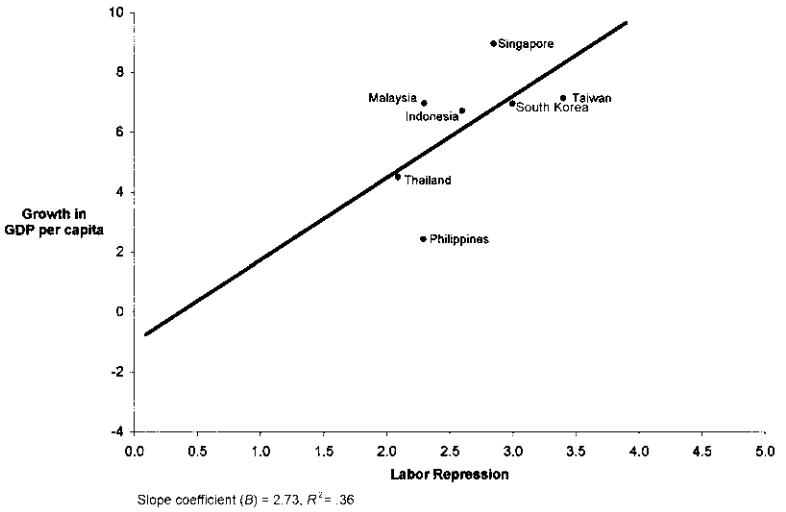

Fig. 3.4. Labor repression and growth in the Asian cases, 1970–81. (GDP per capita from Penn World Tables; for Thailand, from World Bank 1984.)

the selection of cases by geographical location amounts to selection on the dependent variable.

When one looks at the relationship between average labor repression and average growth for a larger sample of countries that includes slow-growing as well as fast-growing ones, the apparent relationship shown in figure 3.4 disappears. As figure 3.5 shows, the slope is approximately flat, and the R^2 is near zero. In other words, level of labor repression has no discernible effect on growth in the larger sample.

It might be objected that several of the arguments linking labor repression to growth were never intended to apply to the entire Third World. Rather, their logic depends on tensions that develop only after industrialization has progressed to a certain

TABLE 3.1. Average Country Growth Rates by Region

	1960–82 (% per capita)	1965–86 (% per capita)
East Asia	5.2	5.1
South Asia	1.4	1.5
Africa	1.0	0.5
Latin America	2.2	1.2
Middle East and North Africa	4.7	3.6

Source: Calculated from data in World Bank (1984, 1988).

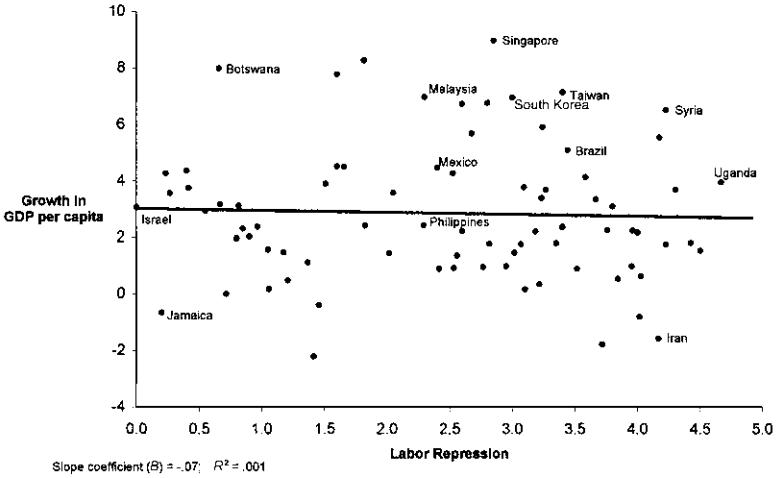

Fig. 3.5. Labor repression and growth in the full universe of developing countries, 1970–81. The countries included, and their labor repression scores, appear in appendix B. (GDP per capita from Penn World Tables.)

stage. Since the literature is unclear about exactly what level of industrialization countries would need to achieve before labor repression would be expected to contribute to growth, I had to decide on a reasonable cutoff point. I used the level of development in South Korea at the beginning of the 1970s as the threshold, since South Korea was the least developed of the countries often discussed as successful examples of labor repression and growth. Figure 3.6 shows the relationship between average labor repression and average growth in the subset of countries that were at least as developed as South Korea in 1970. As figure 3.6 shows, there is no linear relationship between labor repression and growth, even in this subset of cases. The slope is only slightly positive, and the R^2 remains near zero.

In this set of cases, the country with the lowest average growth is Iran, which also scores very high on labor repression. Since Iran's growth rate was depressed toward the end of this period by the revolution, it could be argued that it should be removed from the data set, even though its civil war did not last very long. If this is done, the slope coefficient rises to 0.27, but it remains far from statistical significance, and the R^2 remains near zero. Thus, even with Iran removed, the analysis fails to support the claim that labor repression contributes to growth.

It has been suggested that not labor repression per se but the repression of a previously well organized and mobilized working

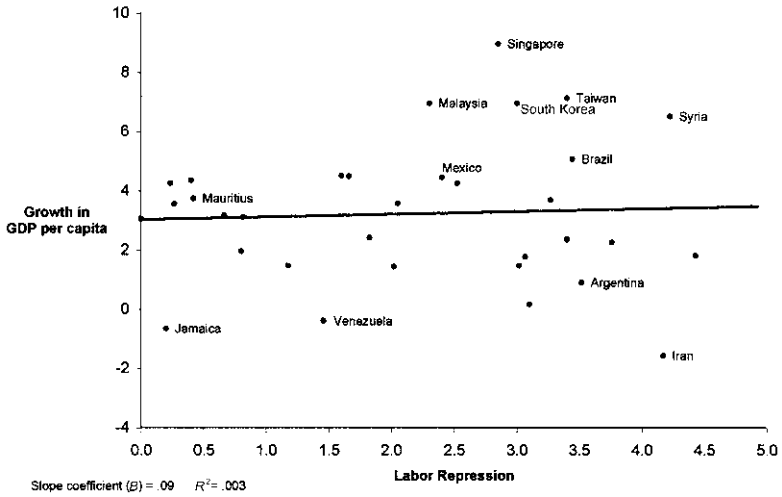

Fig. 3.6. Labor repression and growth in higher-income developing countries, 1970–81. The countries included are those from appendix B whose GDP per capita in 1970 was greater than that of South Korea. (GDP per capita from Penn World Tables.)

class would improve economic performance (O'Donnell 1973; Collier 1979). To test whether increasing repression increases growth, I have estimated time-series models of the effect of yearly labor repression on growth in the following year. In one of the models, two factors that might also be expected to affect growth—oil exports and level of development at the beginning of the period—are controlled for. In the other model, instead of trying to identify the various things that might be expected to affect growth, country fixed effects estimators are used to hold constant all the various country-specific factors that could affect growth rates. When country fixed effects estimators are used, coefficients can be interpreted as reflecting the effect of changes in the variable of interest—here, labor repression—within each country, rather than cross-country differences.

Table 3.2 shows the results of these two regressions. In the model with control variables, the effect of labor repression on growth is both minuscule and statistically insignificant. In the model using fixed effects, the coefficient for labor repression is positive, but not statistically significant. If the coefficient were reliable, it would indicate that for each unit of increase in the labor repression score, a little under a third of a percentage point of extra growth could be expected. The low R^2 for the model

TABLE 3.2. The Effect of Changes in Labor Repression on Growth

Dependent Variable: Annual Growth in GDP Per Capita

	OLS with Control Variables[a]		OLS with Fixed Effects[a]	
	Coefficient	$P > \|Z\|$	Coefficient	$P > \|Z\|$
Labor repression (range 0–5)	.018	.917	.288	.286
Oil exports	.008	.850		
Development level	−.000	.751		
R^2	.000		.099	

[a]Panel corrected standard errors.

shows that even with the inclusion of seventy-nine country fixed effects,[8] the regression explains almost none of the variance in growth.[9]

The point of this exercise is not to demonstrate that the hypothesis that labor repression contributes to growth is false. It may have a small positive effect. It might be that the addition of appropriate control variables or an elaborate lag structure would make clear a relationship that does not show in the simple tests done here. These tests do show, however, that the strong relationship that seems to exist when the analyst examines only the most rapidly growing countries is hard to find when a more representative sample of cases is examined. If analysts interested in the success of the NICs had examined a more representative sample, they would probably have reached different conclusions about the relationship between labor repression and growth. As figures 3.5 and 3.6 show, labor is as often repressed in slow-growing Third World countries as in fast-growing ones.

To sum up, the first example above (fig. 3.3) demonstrates selection bias in its simplest form: the cases are selected precisely because they share the trait one wants to explain. In the second example (fig. 3.4), cases are selected on the basis of a characteristic — geographical region — that is correlated with the dependent variable. In both instances, the hypothesized relationship was a simple, direct one: a higher level of X (labor repression) seemed to result in a higher level of Y (growth).

Not all causal arguments are so simple. Researchers sometimes posit arguments with complicated structures of prior and

8. Four countries had to be excluded because of missing data.

9. Other models were tried using different error specifications and corrections for autocorrelation, even though the regressions reported in table 3.2 disclosed no autocorrelation. In none did the coefficient for repression reach statistical significance.

intervening variables that are more difficult to test rigorously. The consequences of selection on the dependent variable, however, are the same no matter how complicated the argument. The next section will consider another frequently encountered variation on this theme: selection on the dependent variable in a complicated, contingent historical argument.

Selection on the Dependent Variable in a Complicated Historical Argument

Theda Skocpol's stimulating and thoughtful book *States and Social Revolutions* (1979) combines selection on the dependent variable with a complex historical argument. She wants to explain why revolutions occur, so she picks the three most well known instances—the French, Russian, and Chinese revolutions—to examine in detail. She also examines a few cases in which revolution failed to occur, using them as contrasts at strategic points in her chain of argument. The use of cases selected from both ends of the dependent variable makes this a more sophisticated design than the studies of the NICs.

Skocpol argues that external military threats cause state officials to initiate reforms opposed by the dominant class. If the dominant class has an independent economic base and a share of political power, its opposition will be effective and will cause a split in the elite. If, in addition, peasants live in solidary communities autonomous from day-to-day landlord supervision, they will take advantage of the elite split and rebel, which will lead to revolution. (This argument is schematized in fig. 3.7.) This explanation, according to Skocpol, mirrors the historical record in

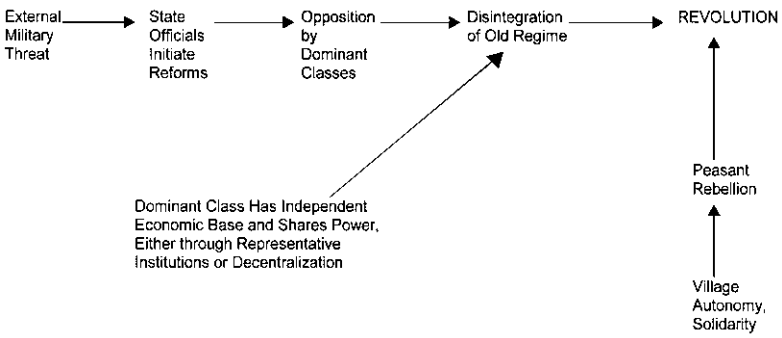

Fig. 3.7. Schematization of Skocpol's argument

France and in the parts of China controlled by the Communists during Japanese occupation. The Russian case differs from the other two in that the upper class lacked the independent economic base necessary to impede state-sponsored reforms. Consequently, the elite remained unified, and revolution failed to occur after the Crimean War. Nevertheless, defeat in World War I caused elite disintegration, which opened the way for revolution in 1917.

At two points in the chain of argument, Skocpol introduces contrasting cases to strengthen her contention that structural features identified as causes in these three cases have general significance. In an examination of Prussia during the late eighteenth to early nineteenth century and Japan during the late nineteenth century, she finds that dominant classes lacked the independent economic base necessary to obstruct state reforms. Both Prussia and Japan faced military threats at least as severe as that facing France, but elites remained unified, and revolution failed to occur. She also looks at Britain during the Civil War and Germany in 1848 and finds levels of village autonomy low. In both cases, elites split, but peasants failed to take advantage of the situation; as a result, revolutions did not occur. These comparisons are summarized in figures 3.8 and 3.9. As the figures show, the cases she examines appear to provide strong support for the argument.

There is no question that the examination of contrasting cases makes the argument more persuasive than it would otherwise be, though an assessment of the argument based on a few cases selected from the other end of the dependent variable carries less weight than would a test based on more cases selected without reference to the dependent variable. Nevertheless, examination of contrasting cases is a solid step in the right

	Elite Split	Elite Cohesive
Dominant Class Economically Independent, Shares Power	France China, after Taiping Rebellion	
Dominant Class Dependent, Excluded from Power	Russia, World War I	Prussia Japan China, before Taiping Rebellion Russia, before World War I

Fig. 3.8. Given external threat, the effect of dominant-class power on the likelihood of an elite split

	Revolution	No Revolution
Village Autonomy	Russia France China, in area controlled by Communists	
Village Dependence		Britain, 1640-60 Germany, 1848 China, before Communists

Fig. 3.9. Given an elite split, the effect of village autonomy on the likelihood of revolution

direction and one of the reasons that Skocpol's study has been considered so persuasive.

Skocpol makes no effort, however, to test other links in the chain of argument with comparable care. In particular, she offers no contrasting cases to strengthen her claim that

> developments within the international states system as such—especially defeats in wars or threats of invasion and struggles over colonial controls—have directly contributed to virtually all outbreaks of revolutionary crises. (23)[10]

This claim, which looms large in the overall thesis, seems especially problematic if we accept her implicit definition of "threatened," that is, as threatened as late-eighteenth-century France. France—arguably the most powerful country in the world at the time—was certainly less threatened than its neighbors.

Most countries in the world have suffered foreign pressures as great as those suffered by prerevolutionary France, and yet revolutions occur infrequently. This raises the question, Are revolutions infrequent because of the absence of appropriate structural conditions, as Skocpol's argument implies, or because foreign threats have less causal impact than Skocpol believes? To distinguish between these two possibilities, one would need to choose a set of cases in which the structural conditions identified by Skocpol did in fact exist (in effect, holding the structural condi-

10. Note that "contributed to virtually all" is a probabilistic statement, not a statement that foreign threat is necessary but not sufficient to explain revolution. Other statements of this argument, however, can be interpreted as meaning that external threats are necessary but not sufficient causes of revolution (Dion 1998).

tions constant). Within this set of countries, one would then need to assess the relationship between level of threat and revolutionary outcome. If threat and occurrence of revolution tended to go together in this set of cases, we would have greater faith in the correctness of Skocpol's argument. If, however, high levels of threat did not seem to increase the likelihood of revolution within this set of cases, we would feel more skeptical about it.

To carry out this test, as with the prior one, we first need to establish the appropriate domain of the argument. The question of what would constitute an appropriate domain for testing Skocpol's argument is controversial. Skocpol herself is extremely modest about the domain for her argument, stating at one point: "Can [the arguments presented in this book] be applied beyond the French, Russian, and Chinese cases? In a sense, the answer is unequivocally 'no'. . . . [T]he causes of revolutions . . . necessarily vary according to the historical and international circumstances of the countries involved" (288). Skocpol does not eschew generalizability entirely, however, since she evidently considers seventeenth-century England, eighteenth- and nineteenth-century Prussia, and mid-nineteenth-century Germany and Japan within the domain of her argument. But she does explicitly limit her argument to "agrarian states," which I take as including countries in the early stages of industrialization (since all the cases included in her study had begun to industrialize) but excluding fully industrialized countries and preagrarian primitive societies. She also limits the argument to countries that have never been colonized; wealthy, "historically autonomous and well-established imperial states" (288); and countries "whose state and class structures had not been recently created" (40).

In the face of such modesty, the rest of the scholarly community has two options in assessing the study. One is to accept the self-imposed limitations suggested by the author and try to test the argument on the set of cases implied by them. The broadest interpretation of these limiting criteria suggests that the appropriate universe thus defined would include, besides some (but not all) of those actually used by Skocpol, only the larger and wealthier pre–World War I states of Europe: Belgium, the Netherlands, Spain, Portugal, Sweden, Lithuania before 1795, Poland before partition, Austria, the Austro-Hungarian Empire, and the Ottoman Empire. This universe includes a fair number of nonrevolutions, so it would be quite possible to retest the argument on this set of cases.

Limiting the argument to this domain does, however, restrict its interest, since twentieth-century revolutions, with the exception of the Russian revolution, have occurred in poor countries that had been at least partly colonized and would thus be outside the domain of Skocpol's argument. Moreover, Skocpol's own selection of cases casts some doubt on the appropriateness of the domain she describes. Japan was not a "well-established imperial state" in the mid-nineteenth century. Nor was China in the twentieth. Germany's state structure, though not affected by colonialism, had been recently created in 1848. China, Japan, and arguably Russia were poor. The time period and geographical location identified by Skocpol as those in which the Chinese peasantry had the autonomy to rebel were precisely the time period and area of Japanese colonization. In short, many of Skocpol's cases violate her own criteria for limiting the domain of her argument.

The alternative approach is to derive the domain of the argument directly from the substantive claims of the argument itself. If we do this, the appropriate domain would seem to include all independent, not fully industrialized states (and possibly empires). These restrictions are necessary because the argument seems to require (1) the existence of an indigenous state elite and dominant class; and (2) a peasantry. Skocpol herself is most adamant about excluding colonized nations from the domain (288–90). This seems a reasonable exclusion during the period of colonization (when the state elite and often the dominant class as well are not indigenous) and perhaps for some limited time—a decade or two—after independence.[11] The claim that any country that has ever been colonized should be forever excluded does not seem to flow from anything in the argument itself, however, and also seems to ignore the role of conquest in the development of the states included in the original argument. After all, England was once colonized by the Normans, large parts of Russia by the Mongols and Tatars, and China by the Mongols. In all three, aspects of subsequent state organization and development are commonly traced to the effects of these conquests.

11. If we think of the domain as derived from the argument itself rather than from Skocpol's somewhat ad hoc comments about it, then her inclusion of China during the time that much of the country was colonized by Japan seems less puzzling. Throughout the Japanese occupation, an indigenous Chinese state elite and dominant class continued to exist in southern China, and it was they whom the Chinese Communists eventually defeated, not the Japanese.

Skocpol also argues that small countries should be excluded because revolutions in them may be caused or prevented by outside intervention (289). This is a legitimate concern. It should not lead to the blanket exclusion of all small countries, however, since we know that outside intervention has failed to prevent revolution in a number of them. Yet it might be reasonable to eliminate some cases in which a persuasive argument can be made about the decisiveness of intervention.

Ideally, a test of Skocpol's hypothesis about the effects of military competition would examine all independent, not fully industrialized states characterized by the structural features—village autonomy and a dominant class with an independent economic base and access to political power—that she identifies as necessary to complete the sequence from military threat to revolution. Then one could determine whether revolutions occur more frequently in countries that have faced military threats.

In practice, identifying the universe of cases that meet these structural criteria is probably impossible. It would require extensive knowledge about every country in the world from the English Civil War to the present. Nonetheless, moderately serious tests of her argument are possible, and one is shown below.

As it happens, several Latin American countries (Mexico, Guatemala, El Salvador, Honduras, Nicaragua, Ecuador, Peru, Bolivia, and Paraguay) have the structural characteristics she identifies and so can be used as a set of cases on which to test the hypothesis linking military threat to revolution. These cases are obviously not selected at random, but since their geographical location is not correlated with revolution, geography does not serve as a proxy for the dependent variable (as occurred in the test of the relationship between labor repression and growth among the East Asian NICs).

In all these countries, dominant classes had an independent economic base in land and/or mining from the nineteenth century until well into the twentieth. They also shared political power. Thus, they had the economic and political resources that Skocpol identifies as necessary to oppose state-sponsored reforms successfully and so pave the way for revolution.

These countries also all contained (and most still contain) large, severely exploited indigenous and mestizo populations, many of whom lived in autonomous, solidary villages. Spanish colonial policy reinforced, and in some areas imposed, corporate village structure. After independence, changes in property rights

reduced village control over land, but this reduction in the functions that had contributed to building village autonomy and solidarity was at least partially offset by the increase in absentee landlordism that accompanied increasing commercialization.

Much of the land in these countries was held in large tracts. Some peasants lived on the haciendas, but many lived in traditional villages, owned tiny parcels of land or had use rights to communal land, and worked seasonally on the haciendas. These villages often had long histories of conflict with large landowners over land ownership, water rights, and grazing rights. Villages governed themselves in traditional ways. Landlords have rarely lived in villages in these countries. In short, the rural areas of these Latin American countries approximate Skocpol's description of the autonomous, solidary village structure that makes possible peasants' participation in revolution. Differences of opinion are, of course, possible about whether peasants in these countries were really autonomous enough from day-to-day landlord control to enable them to play the role Skocpol allots to peasants in bringing about social revolutions. Perhaps the best evidence that they were is that revolutions have in fact occurred in several of these countries, and peasant rebellions have occurred in most of them.

With these structural features on which the outcome is contingent held constant, it becomes possible to test the relationship between external threat and revolution. In the test below, I have used a higher level of threat than that experienced by France in the late eighteenth century. I wanted to choose a criterion for assessing threat that would eliminate arguments about whether a country was "really" threatened enough, and I found it hard to establish an unambiguous criterion that corresponded to the "France threshold." Consequently, the criterion used here is loss of a war, accompanied by invasion and/or loss of territory to the opponent. With such a high threat threshold, finding cases of revolution in the absence of threat will not disconfirm Skocpol's argument, since the countries may have experienced external pressures sufficient to meet her criteria even though they did not lose wars. If, however, several countries did lose wars (and the structural conditions identified as necessary by Skocpol are present) but have not had revolutions, this test will cast doubt on her argument.

Figure 3.10 shows eight instances of extreme military threat that failed to lead to revolution, two revolutions (if the Cuban

	Revolution	No Revolution
Defeated and Invaded or Lost Territory	Bolivia (1935), revolution 1952	Peru (1839) Bolivia (1839) Mexico (1848) Mexico (1862-66) Paraguay (1869) Peru (1883) Bolivia (1883) Colombia (1903)
Not Defeated within 20 Years	Mexico, revolution 1910–17 Nicaragua, revolution 1979	All Others

Note: The Cuban Revolution is not, in Skocpol's terms, a social revolution because it did not entail massive uprisings of the lower classes.

Fig. 3.10. Relationship between military defeat and revolution in Latin America (with Skocpol's structural variables held constant)

revolution of 1959 is not counted, because it does not fit Skocpol's definition of a social revolution as entailing massive uprisings of the lower classes) that were not preceded by any unusual degree of external competition or threat, and one revolution, the Bolivian, that fits Skocpol's argument. I do not think any foreign power deserves credit or blame for any of the revolutions that have occurred, and thus the finding that two revolutions occurred without unusual foreign threat is not undermined by foreign influences on revolutionary outcome. The United States may deserve credit or blame for the nonoccurrence of revolution in El Salvador and Guatemala, but if these revolutions had been successful, they would have increased the number of cases in which revolutions occurred in the absence of unusual foreign threat and thus added to the evidence undermining Skocpol's argument. In short, among these cases there is little support for the claim that foreign threat increases the likelihood of revolution. If we accept the idea that the domain depends on the argument itself, then these findings suggest that if Skocpol had selected a broader range of cases to examine, rather than selecting on the dependent variable, she would have reached different conclusions.

This test does not constitute a definitive disconfirmation of Skocpol's argument. Competing interpretations of all the concepts used in the argument—village autonomy, dominant-class independence, military pressure—exist, and different operationalizations might lead to different results. In particular, my

operationalization of threat fails to capture the complexity of Skocpol's idea, and a different operationalization might put Nicaragua and Mexico in the threat-revolution cell. Any indicator of threat that identified Nicaragua in 1979 and Mexico in 1910 as threatened, however, would add hundreds of other country-years to the threat–no revolution cell, because the amount of U.S. pressure experienced by these countries at these times was not at all unusual in the region. In short, despite some deficiencies in operationalization, this cursory examination of cases not selected on the dependent variable does cast doubt on the original argument.

Arguments about Necessary Causes

Some have interpreted Skocpol's statements as meaning that she sees external threat as a necessary but not sufficient cause of revolution. As Douglas Dion (1998) and others have noted, the logic underlying tests of arguments about necessary causes differs from that described above. Methods for testing arguments about necessary causes have only begun to be developed, but Dion suggests a Bayesian approach.[12] Bayesian analysis provides a way of assessing the impact of new information on one's prior beliefs about the likelihood that a particular theory is true. If, in order to keep things simple, we set aside the possibility of measurement error and think of only one rival hypothesis to the one being tested, Bayes rule can be expressed as:

$$P_{osterior}(WH|D) = \frac{P_{rior}(WH)\, P(D|WH)}{P_{rior}(WH)P(D|WH) + P_{rior}(RH)P(D|RH)},$$

where

- $P_{osterior}$ (WH|D) is the probability that the working hypothesis (the one being tested) is true, in light of the new evidence collected in a study.
- P_{rior}(WH) is the analyst's belief about whether the working hypothesis is true before conducting the study.
- P(D|WH) is the probability that the data uncovered in this study would turn up *if* the working hypothesis were true.
- P_{rior}(RH) is the analyst's belief about the likelihood that the rival hypothesis (the most likely alternative to the working hypothesis) is true prior to conducting the study.

12. See Baumoeller and Goertz (2000) for a careful non-Bayesian approach.

P(D|RH) is the probability that the data uncovered in this study would have emerged *if* the rival hypothesis were true.

If we interpret the Skocpol argument as one about the necessity of external threat, then the appropriate initial research strategy is to choose cases that have experienced revolutions and then check to see if an external threat preceded the revolution. The information about these external threats is the new data that will be used to update the assessment of the likelihood that the working hypothesis is true. Note, however, that the only way to assess the likelihood of observing the new data given that the *rival* hypothesis is true is to know enough about the whole relevant universe of cases (not just those that experienced revolution) to be able to estimate the probability of observing these events (in this case, external threats) if the rival hypothesis better describes reality. In other words, we need to know something about the frequency of the hypothesized preceding event in the universe as a whole.

In order to use Bayes' rule, it is also necessary to state a level of prior belief that the working hypothesis is true. These prior beliefs come from prior research on a subject. When little prior research has been done on a subject, it has become conventional to treat prior beliefs as neutral between the two competing hypotheses, that is, to set P(WH) = P(RH) = 0.5.

To return to the Skocpol example, if we use "as threatened as France" as the appropriate threat threshold, then I would estimate that 95 percent of all countries that would otherwise fit within the domain of the theory have experienced such a threat at some time, many of them repeatedly. With this estimate, the probability of observing the data (external threat) in any particular country between 1600 and the present, given the rival hypothesis that external threat does not cause revolution, can be calculated.

Skocpol examined three cases and found external threats in all three. The probability of seeing these data, *if* the working hypothesis is true (and there is no measurement error) equals one. If the rival hypothesis is true and 95 percent of the countries in the domain of the argument have experienced similar levels of threat at some time, then the probability of observing three instances of threat if the rival hypothesis is true equals $0.95 \times 0.95 \times 0.95 = 0.857$. Plugging these numbers into Bayes' rule, we get:

$$P_{osterior}\,(WH|D) = [0.5(1)] / [0.5(1) + 0.5(0.857)] = 0.539$$

In other words, when the hypothesized necessary cause is very common in the world, the increase in one's level of belief in the argument is increased only very modestly (from 0.5 to 0.539) when a few new cases are examined. If the hypothesized necessary cause only occurred in ten percent of the cases in the appropriate universe, then examining three cases and finding the data expected would increase our prior belief in the argument to above 99 percent. Thus the number of observations needed to affect posterior beliefs about a hypothesis depends very dramatically on the general distribution of the hypothesized necessary cause.[13]

From the point of view of research design, this discussion of Bayesian inference leads to two conclusions that have not been much emphasized in the literature on testing arguments about necessary causes. First, the Bayesian approach requires that the data used to assess the likelihood that the theory is true be newly observed. It must come from cases observed for the purpose of testing the argument, not from the cases from which the hypothesis was induced. The original cases, along with other research and general knowledge about the world, influence the observer's prior beliefs about whether the argument is true. Bayes' rule provides a way of judging how much more convinced by an argument we should be after seeing new data, not how much faith we should put in a plausible but untested argument.

Second, although arguments about necessary conditions can be tested using only cases selected on the dependent variable, the use of Bayesian logic to assess how much has been learned from the test requires gathering enough information about how often the hypothesized necessary cause occurs in the world more generally in order to estimate the probability that the data that were actually observed would have been observed if the rival hypothesis were true. In the non-Bayesian approach suggested by Braumoeller and Goertz (2000), this issue has been addressed under the label trivialness. Braumoeller and Goertz argue that to be non-trivially necessary, the hypothesized necessary cause must be shown to vary more than a little in the full

13. Dion (1998) provides a chart showing how many cases would be needed to reach 95 percent confidence that an argument is true, given different prior levels of belief and different estimates of the likelihood of observing the data if the rival hypothesis is true.

relevant universe. Whatever amount of variation is deemed sufficient for non-trivialness, we cannot discover it without examining at least some cases not selected on the dependent variable. If the hypothesized necessary cause occurs infrequently, then examining only a few additional cases would suffice to meet their condition.

Time Series, Case Studies, and Selection Bias

Case studies, perhaps the most common form of research in the comparative field, can often be thought of as nonquantitative time-series research designs. They usually examine a single country over a period of time, often for the purpose of explaining some outcome at the end or showing the effects of some change that occurred during the time examined. Case studies are often criticized as single data points and hence incapable of revealing anything about cause-and-effect relationships, but most can be more reasonably thought of as a series of observations of the same case at different times. In fact, most of what are called case studies actually include unsystematic observations at multiple levels of analysis (for example, individuals, government administrations, and parties) and observations of multiple entities at the same level of analysis (for example, several parties in one country) as well as observations over time. For now, however, let us focus on the simplest kind of case study—say, a study of the evolution of one party over time.

Such case studies are subject to several methodological pitfalls, solutions to which are discussed at greater length in chapter 4. Here I want to note the methodological issues related to selection bias that can arise in the context of case studies and single-case quantitative time series. In the typical study of a single case, a country, organization, or group is chosen for examination because it has experienced something unusual, sometimes because it is considered typical of a group of cases that have experienced the unusual. The variation on the dependent variable is supplied by observations of the same case at other times (when it was not experiencing the unusual). Whether such a research design involves actual selection bias depends on whether the variation over time within the case reflects the full range of outcomes in the relevant universe. Often it does not. When the selection of multiple observations of a single case results in a truncated sample relative to the appropriate universe, the result is inadvertent

selection on the dependent variable, and selection bias can be expected to have the same results noted above.

The key concern for the researcher, then, is identifying the universe relevant to the question being asked or hypothesis being tested. Only when that has been done can he assess whether outcomes vary widely enough within the single case to avoid selection bias. Here, as elsewhere, the question under examination determines the appropriate universe. Sometimes one wants to understand the effect of a particular policy change in a particular setting. In this situation, one is not asking "What caused outcome *Y?*" but rather "What was the effect of cause *X?*" If cause *X* occurred only in one setting, then a one-country time series or case study is the appropriate research design (Campbell and Ross 1968). If *X* occurred in multiple places, the analyst would be wise to examine its effects in all of those places or a sample of them. Otherwise, he risks the possibility of attributing to *X* anything that might have happened in the chosen country during the time following *X*. When the analyst wants to know what caused outcome *Y,* it is always risky to examine only one entity in which *Y* occurred, even if he cares only about why it happened in a particular place and time. A case study of the particular place and time of interest may not provide the answer.

The reason it might not is that it is quite possible that selecting multiple observations of the same case will have the effect of holding constant or near constant some of the true causes of the outcome of interest, even if the dependent variable spans a considerable range. At the same time, whatever potential causal factors do vary within the single case over time will seem to explain differences in the outcome. These causes of the within-case variation can be less important causal variables that belong in a complete explanation, or they can be idiosyncratic factors that affect this case but not others and therefore do not belong in a general explanation. The analyst has no way of knowing. Either way, he will be tricked into focusing on these factors while giving short shrift to causal factors that may be changing slowly and not very noticeably during the time under study but that nevertheless explain the general trend in the outcome.

This problem is caused by inadvertent selection on one or more causal variables. Case studies are highly vulnerable to inadvertent selection from one end of the continuum of potential causes because so many factors remain constant or change slowly over time in a single entity. In statistical work, selection on the

dependent variable leads to biased estimates of the effects of causal factors, but the practical result is usually a failure to demonstrate a causal relationship that actually exists, because within the truncated sample there is little variation in the outcome for differences in the causal factors to explain. Selection on the independent variable, as often happens in case studies, does not lead to biased estimates in statistical work. Nevertheless, in practice it is hard to show that a relationship between a cause and effect exists if the cause varies little within the sample. Whether observations are quantitative does not affect the logic of the research design. In either case, it is quite possible to overlook factors of real causal importance, because they do not vary much over time or follow an incremental trend that country observers take for granted.

If one knows quite a bit about the underlying causal model, a single-case time-series design can be a good way to assess the effect of one potential cause while holding many other things constant (because they do not vary within the single case), but it will be less useful in the more typical situation where the analyst does not know the underlying model. The analyst will then fail to identify any causal factor that varies little within the case and will tend to overemphasize serendipitous contributors to the outcome.

As an example, let us contemplate Albert Hirschman's careful and insightful study of inflation in Chile (1973). In this essay, Hirschman reviews Chile's major bouts of inflation between the nineteenth century and 1961. He reconsiders the role of foreign experts in Chilean policy formulation and shows the importance of dogmatic economic ideologies and policy mistakes in causing inflationary episodes. Hirschman argues in this study, and in the book of which it is a part, that policymakers gradually learn to resolve persistent problems, that the search for solutions has positive externalities in that it brings hitherto unnoticed issues to policymakers' attention, and that reformism, though messy and often emotionally unsatisfying, leads over time to significant improvements. Hirschman describes first the intermittent difficulties with inflation that Chile experienced between 1870 and 1939, and then the persistent and worsening inflation of 1940 to 1959. At each point in the story, Hirschman, with his customary flair and sensitivity to detail and context, discusses the policy mistakes and other factors that increased inflation. During the early period, inflationary episodes were caused by wars and civil wars, serious policy mistakes, and business expansions, all of which

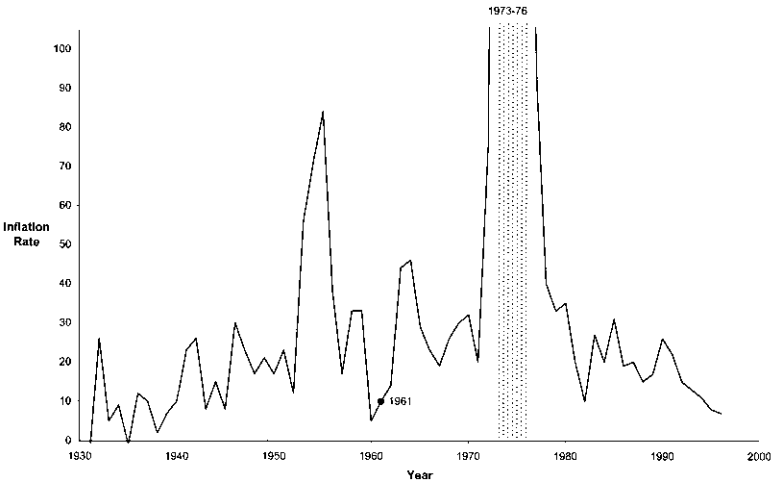

Fig. 3.11. Inflation in Chile, 1930–96. (Data for 1930–61 from Hirschman 1973; 1962–63, Corbo Loi 1974; 1964–96, IMF 1997.)

seem to have been largely self-correcting. Beginning in 1940, however, inflation became more persistent and more serious. It no longer returned to normal between episodes, and the trend line, though masked by zigzagging, began a determined upward slope (see fig. 3.11). Between 1940 and 1959, inflation averaged 28 percent per year (Hirschman 1973, 160). Chilean inflation rates are shown in table 3.3.

Hirschman attributes this worsening mostly to the effects of specific policies, especially the failure to restrict credit to the private sector and the routinization of wage adjustments. The emphasis throughout the essay is on the details of policy and the political context that influenced them. In explaining the control of inflation, which appeared to have been achieved in Chile in 1960–61, Hirschman stresses the intense political struggle over ending automatic wage adjustments—accomplished during an inflationary peak in 1956 (203–5)—and strengthening the system of credit control under President Jorge Alessandri in 1959 (219).

Although Hirschman mentions general economic factors such as fiscal deficits and exchange rates, the reader is left with the impression that Chilean inflation was caused by some fairly discrete policy mistakes. This impression is strengthened by the low inflation rates of 1960 and 1961, caused by specific policy changes introduced by the Alessandri administration. The reader never sees the bigger and more general picture: that the policy strategy

of state-sponsored import-substitution industrialization (ISI), which was initiated in Chile in 1939 and varied little during the next thirty-five years, caused increasingly serious inflation.

Although the general policy strategy remained stable and was not subjected to the kind of intense political debate that accompanied the policy changes emphasized by Hirschman, its implementation over time entailed increasingly distortionary tariff and exchange rate policies. These policies led to the same problems with balance of payments crises and inflation that afflicted so many other developing countries. Inflation plagued all developing

TABLE 3.3. Chilean Inflation, 1930–96

Year	Inflation Rate (%)	Year	Inflation Rate (%)
1930	−5	1964	46
1931	−4	1965	29
1932	26	1966	23
1933	5	1967	19
1934	9	1968	26
1935	−1	1969	30
1936	12	1970	32
1937	10	1971	20
1938	2	1972	75
1939	7	1973	361
1940	10	1974	505
1941	23	1975	375
1942	26	1976	212
1943	8	1977	92
1944	15	1978	40
1945	8	1979	33
1946	30	1980	35
1947	23	1981	20
1948	17	1982	10
1949	21	1983	27
1950	17	1984	20
1951	23	1985	31
1952	12	1986	19
1953	56	1987	20
1954	71	1988	15
1955	84	1989	17
1956	38	1990	26
1957	17	1991	22
1958	33	1992	15
1959	33	1993	13
1960	5	1994	11
1961	10	1995	8
1962	14	1996	7
1963	44		

Source: 1930–61, Hirschman (1973); 1962–63, Corbo Loi (1974); 1964–96, World Bank (2002).

countries that followed state-led import-substituting development strategies. The average inflation rate for low- and middle-income countries between 1970 and 1978, a time when virtually all but the most backward were following state-led ISI strategies, was 18 percent per year, compared to about 9 percent per year for industrialized countries (World Bank 1980, 110–11). Hirschman is, of course, entirely correct in noting that discrete policy mistakes worsen inflation; they account for some of the zigzags so apparent in figure 3.11. Nevertheless, the reader interested in understanding why Chile suffered from recurrent and worsening bouts of inflation for nearly four decades after 1939 will not have recognized the main underlying cause after reading this essay.[14]

In fairness, let me note that Hirschman did not aim to explain inflation in this essay. Rather, he sought to show that inflation, like other seemingly intractable problems, could be gradually conquered as policymakers learned to understand it and took advantage of occasionally propitious political circumstances to initiate reforms. When the primary underlying cause of something has not been identified, however, identifying and "fixing" less important and less systematic causes may not result in long-term improvement. The last measure of inflation in Hirschman's study is for 1961, when it appeared that policymakers had at last brought inflation under control. The apparent cure, however, turned out to be a very brief remission.

As is apparent in table 3.3 and figure 3.11, Chilean inflation did not begin its long-term downward trend until the abandonment of state-sponsored import-substitution development policies during the Pinochet administration. The extremely high inflation rates of the Allende years and their immediate aftermath cannot be blamed on development strategy, but if those years are excluded, it is still clear that the conquest of inflation in Chile began in the late 1970s. Current Chilean economic policy-making—and low inflation—demonstrates that Hirschman was correct in believing that human beings, including policymakers, learn. But in Chile, inflation was not finally conquered by reformist muddling through and discrete policy changes, as Hirschman had hoped. It was conquered by traumatic policy changes that reversed four decades of basic economic policy strategy.

14. In another essay, written a few years before the one on Chile and drawing on the experiences of several Latin American countries rather than only one, Hirschman (1968) was one of the first to identify a number of the systematic ill effects of the import-substitution strategy of industrialization, including its tendency to cause inflation.

The methodological point is that even if one cares only about what caused inflation in Chile, the best research strategy for discovering these causes may require examining other cases. Important causes, such as the basic thrust of development strategy, may not change very much within a few decades in a single country. Consequently, analysts may overlook their importance and instead concentrate attention on less important causes or on conjunctural factors that turn out to have no general causal effect. Case studies generally help to explain zigzags in the trend line, but they sometimes offer little leverage for explaining the trend itself.

Case Studies, Time Series, and Regression to the Mean

The remainder of this chapter focuses on some less obvious pitfalls that face the researcher who must choose not only which cases to examine but also the beginning and end points of the study. If either the starting or ending dates of a case study or time series are chosen because of their extreme scores, the analyst must be concerned about the effects of regression to the mean, in addition to the other possibilities for mistaken inference associated with selection bias. Because extreme outcomes typically result from a combination of extremes in their systematic causes and extreme unsystematic influences (what would be called the error term in quantitative work), terrible conditions at the initiation of a study are likely to improve with the passage of time, and wonderful situations are likely to deteriorate—even if there has been no change at all in the systematic factors causing them. Such changes in the unsystematic influences on outcomes lead unwary analysts to attribute improvement or deterioration to their favorite hero or villain among intervening events, even though the only real change that has occurred is in the random factors that influence everything in social science and the rest of the world.

Regression to the mean is the name given to the tendency of any extreme situation, score, outcome, or event to be followed by one that is less extreme simply because fewer extreme random factors happened to influence things the second time. Regression to the mean causes the mismeasure of systematic change in outcomes over time. After having misunderstood the amount of change that has actually occurred, the analyst often then compounds the mistake by building an argument to explain the changes that never occurred.

Regression to the mean has been most fully analyzed in the context of educational research. The classic example involves researchers trying to assess the usefulness of a new technique for teaching remedial reading. Students' reading ability is tested. Those who score below some threshold are selected to receive special help, using the new technique. After some time has elapsed, they are tested again, and the rest of the class is also retested as a control, since all students are expected to be increasing their skill over time. The students who received special help always make greater gains than the group that did not receive help, no matter what technique is tried. Illiteracy has not disappeared, however, because at least some and perhaps all of the gain demonstrated by these students is an artifact caused by regression to the mean, not a genuine effect of the remedial reading techniques. The students who scored lowest on the first test did so for two reasons: they read less well than others; and, for unsystematic reasons such as being sick or tired, they did especially badly on the first test. The second test, like the first, measures both the systematic component of reading ability and also random factors such as sickness and tiredness. Since it is unlikely that the same children would be sick or tired during both tests, on average the scores of the remedial group would not include extreme unsystematic elements the second time, and thus they would score higher even if their reading ability had not improved.

Regression to the mean has two sources, one conceptually trivial but practically very important in the social sciences, and the other both conceptually and practically important. The first is that every measurement contains an element of error. For simple physical measurements of things, such as temperature and length, the small element of error in the measure is usually of no practical consequence, but in social science most of the things we want to explain can be "measured" only very inadequately. One of the unsystematic contributors to every outcome is thus measurement error—simply the inaccuracy of all measures, whether quantitative or not. Every outcome is also affected by happenstance, by events that will never occur again and have no theoretical importance, by luck, by the particular skills, failings, and longevity of certain individuals, and so on. These unsystematic contributors to outcomes are also, if the research goal is systematic explanation, part of the "error term." That is, both real measurement error and serendipitous factors contribute to every assessment of an

outcome. They are the causes of the statistical artifact, regression to the mean.

Any time that cases are selected for study or "treatment" on the basis of high (or low) scores on some variable, the analyst unintentionally selects a sample with unusually positive (or negative) "error terms," in both senses discussed above. When the selected cases are measured a second time, the inaccuracy in their measurement is no more likely to be positive than negative, and the serendipitous events that affect the outcome then are also no more likely to be positive than negative. In consequence, cases with especially positive outcomes in a first measurement will look as though they are doing less well in a second one, and cases with especially negative outcomes in a first assessment will seem to improve over time—even if *nothing* systematic has changed.

Because most of the work on regression to the mean has occurred in the context of educational and psychological testing, students of comparative politics are sometimes unaware of its implications for their own work. Consider the following hypothetical study. The analyst wants to know what effect structural adjustment loans from international financial institutions have on the economic performance of developing countries. To answer this question, she compares the growth rate per capita in the countries that have received such loans with the growth rates of developing countries that have not. She needs to compare the countries of interest with another set of countries during the same time period as a means of controlling for international factors that affect growth all over the world. Since, however, structural adjustment loans went to countries with economies in crisis, and since these crises are caused by bad luck, bad weather, and unrepeatable events as well as by more systematic factors, their economic performance can be expected to improve, on average, whether or not the loans help. Thus, the unwary researcher who simply compares changes in the performance of countries receiving loans with changes in those that did not may be misled about the effect of the loans, since a certain amount of improvement in the countries with the worst performance could have been expected in any case.

Whenever research focuses on comparison of growth rates over time, the analyst needs to be attentive to the possibility of regression to the mean. The underlying causes of economic performance, such as resource endowment, human capital, savings

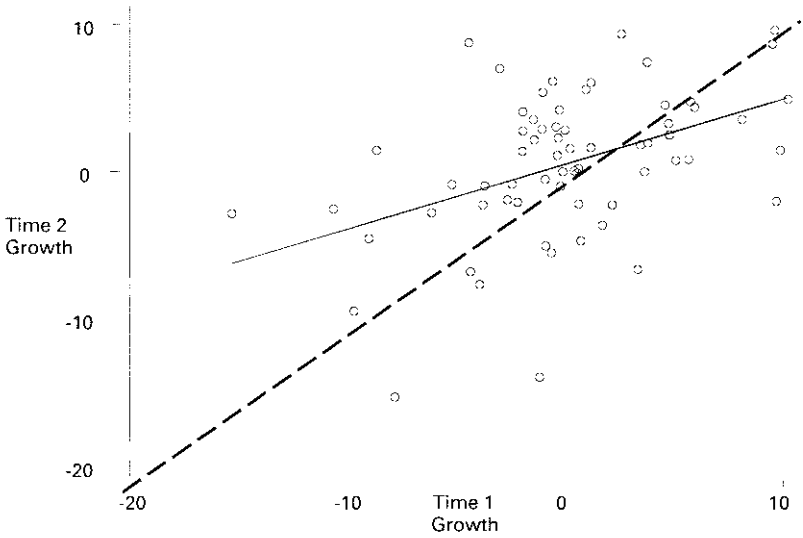

Fig. 3.12. Regression of growth in GDP per capita for 1991 on growth in GDP per capita for 1990 for developing countries. (GDP per capita from Penn World Tables.)

rate, competence of policymakers, and basic thrust of economic policy, do not change much from year to year in most countries. Consequently, we would expect a high correlation between rates of growth from one year to the next, and our casual observation that most Asian countries grow rapidly almost every year and most African countries do not supports that expectation. At the same time, unsystematic factors and measurement error contribute to the observed growth rate in every country every year. This unsystematic component of measured growth always leads to the appearance of faster growth in countries with the worst performance in an earlier time period and slower growth in countries with the best performance.

Figure 3.12 shows the relationship between growth in 1990 and growth in 1991 for developing countries.[15] The solid line is the regression line, which shows the estimated growth rate in 1991 given any particular growth rate in 1990. The dashed diagonal line is the hypothetical relationship we would expect to exist

15. The data set used to construct this scatterplot includes all low- and middle-income countries with more than a million inhabitants for which data were available from the Penn World Tables. Countries with fewer than a million inhabitants were excluded because their economies tend to be unusually volatile, and I did not want the results shown here to depend on unusual cases. The countries for which data are not available include most of those engaged in civil war during these years.

if all the causes of economic performance remained stable from year to year and therefore growth, on average, remained the same from year to year. The part of the regression line that reflects the performance of the countries with the highest growth rates in 1990 lies below the diagonal, showing that they tended to grow less rapidly in 1991 than they had in 1990. Meanwhile, the part of the regression line for countries with the lowest growth rates in 1990 lies above the diagonal, indicating that they grew more rapidly in 1991. Countries growing at above 5 percent per capita in 1990 grew, on average, only 3.7 percent in 1991. At the other extreme, countries with growth declining at 5 percent or more in 1990 were declining at only 2.8 percent per capita, on average, in 1991.

These tendencies were not caused by some vicissitude in the international economy that for once advantaged the poor and disadvantaged the rich. (The reader who suspects that this might be the case is urged to try this regression on other years. In every single pair, the fast-growing countries will do a little less well in the second year, and the slow-growing ones a little better. This result does not mean that growth rates are gradually evening out among countries.)

These tendencies are not caused by anything systematic, but rather by changes in the "error term." The countries with the highest scores at any time are those with not only good systematic economic performance but also, on average, those with positive error terms—either real measurement errors or serendipitous events and luck that cannot be expected again the following year. In the subsequent measurement, economic performance is, on average, still good, but the unsystematic component of the outcome is, on average, neither positive nor negative, and thus the overall score is lower. As a consequence, any time one selects cases for study because they are doing especially well, one can expect that their subsequent performance will decline a bit, and the inverse is true for cases selected because they are doing especially badly.

Regression to the mean is especially likely to interfere with reaching correct conclusions when one is trying to assess the effect of some "treatment," such as structural adjustment loans or aid programs aimed at meeting basic needs. This is because "treatments" are often provided only for those who donors think need them, usually those experiencing some sort of crisis. In such situations, the problem is not that the analyst selects cases from

one end of the continuum, but that the agencies supplying the "treatment" do.

Regression to the mean can also affect one's ability to assess the effect of spontaneous "treatments" such as military interventions. If democratic breakdown usually occurs during economic crises, then a research design that compares economic performance before and after military interventions is likely to overestimate the beneficial effects of military rule on the economy, for exactly the same reasons that educational researchers might be tempted to overestimate the beneficial effects of a remedial reading technique. On average, the poor economic performance of the pre-breakdown period was caused by both systematic and unfortunate serendipitous factors, but the serendipitous factors that affect performance during the later period under military rule will, on average, be average. If, for example, one compares the growth rate in Argentina, Brazil, Chile, and Uruguay during the year prior to the most recent breakdown of democracy with the average during the first five years of military rule — as a number of authors attempting to assess the effects of bureaucratic authoritarianism did, though usually not quantitatively — one is tempted to conclude that military rulers handle the economy better than do elected politicians. On average, per capita income declined by 1.5 percent during the year prior to breakdown in these countries, but it grew 0.8 percent per year, on average, during the first five years under military rule (not including the breakdown year itself).[16]

One cannot, however, conclude from these figures that military regimes perform better. To assess that question, one would need to model the regression to the mean that would be expected in the relevant years and then compare economic performance under the military with that predicted by the model. Alternatively, one might compare growth during military rule with long-term growth in the same countries, since the ups and downs in the error term would be evened out by averaging over many years. Average growth in these four countries from 1951 to the year before military intervention ranged from 0.9 percent for Uruguay to 3.2 percent for Brazil, all higher than the average during the first five years of military rule.[17] A more careful test could certainly be done, but this simple one is sufficient to sug-

16. These percentages, as well as those in the following paragraph, were calculated from the Penn World Tables.

17. Years included for Argentina are 1951 to 1965, because the military ruled for most of the time after that.

gest that these military governments were not especially successful at delivering rapid growth.

Conclusion

The examples in this chapter have shown that choosing cases for study from among those that cluster at one end of the outcome to be explained can lead to the wrong answers. Apparent causes that all the selected cases have in common may turn out to occur just as frequently among cases in which the effect they are supposed to have caused has not occurred. Relationships that seem to exist between causes and effects in a small sample selected on the dependent variable may disappear or be reversed when cases that span the full range of the dependent variable are examined. Arguments that seem plausible if a historical study or time series begins or ends at a particular date may seem less persuasive if the dates of the study are changed. Regression to the mean can lead the unwary researcher into explaining changes that did not occur. In short, selecting cases without giving careful thought to the logical implications of the selection entails a serious risk of reaching false conclusions.

This is not to say that studies of cases selected on the dependent variable have no place at all in comparative politics. They are useful for digging into the details of how phenomena come about and for developing insights. They identify plausible causal variables. They bring to light anomalies that current theories cannot accommodate. In so doing, they contribute to the creation and revision of proposed theories. By themselves, however, they cannot test the theories they propose (cf. Achen and Snidal 1989). To test theories, one must select cases in a way that does not undermine the logic of inference.

If we want to begin accumulating a body of theoretical knowledge in comparative politics, we need to change the conventions governing the kinds of evidence we regard as theoretically relevant. Conjectures based on cases selected on the dependent variable have a long and distinguished history in the subfield, and they will continue to be important as generators of insights and hypotheses. Regardless of how plausible such conjectures are, however, they retain probationary status as accumulated knowledge until they have been tested, and testing them usually requires the thoughtful selection of cases from the full range of possible outcomes.