Case-Selection Techniques in Case Study Research:

A Menu of Qualitative and Quantitative Options

Jason Seawright

UC Berkeley
Department of Political Science
210 Barrows Hall
Berkeley, CA
seawright@berkeley.edu

John Gerring

Boston University
Department of Political Science
232 Bay State Road
Boston, MA 02215
617-353-2756
jgerring@bu.edu

Draft: August 26, 2005

In large-N analysis the problem of case selection has a straightforward solution: random sampling.[1] In case study analysis, however, selection procedures are not as well developed and are often problematic. This is usually understood as a problem of representativeness or sample bias: how to make sure that the chosen case(s) accurately represents the population of interest. Perhaps no other feature of case study research has been so vigorously interrogated in recent years (Achen and Snidal 1989; Collier and Mahoney 1996; Geddes 1990; Rohlfing 2004; Sekhon 2004).

However, the problem of case-selection is not limited to sample bias. Cases chosen for intensive analysis must also provide sufficient variation along key parameters.[2] This issue receives little attention in cross-case research because it is usually safe to assume that a large-N sample will automatically contain sufficient variation to provide leverage for causal analysis, an assumption that does not necessarily hold for small-N samples.

Achieving the twin goals of representativeness and variation is not a simple matter when the total sample size is sharply limited, as it is (by definition) in case study research. Unfortunately, the literature on this question -- sometimes known as "purposeful sampling" (Patton 2002: 230-46) – has not advanced very far. Early work by Eckstein (1975) and Lijphart (1971, 1975) is often cited, but rarely developed. Thus, the discipline has accumulated a panoply of case study types – "extreme," "deviant," "crucial," "heuristic," "plausibility probe," et al. But their defining characteristics, and their range of usage, remains obscure. Indeed, many of these case study types appear to possess overlapping traits or to be talking about different properties of the case study. For example, some case study types refer to the varying objectives of case study research; others refer to different kinds of cases (i.e., it is the properties of the case itself that is at issue). A few of these well-established case study types are not, in our opinion, very strong methods at all. Moreover, there has been little or no attempt to discern how these methods might be practiced in the context of a large population, i.e., when there are a great many candidates for intensive analysis. Recent work integrating case study research with large populations has focused mostly on how to analyze a previously identified case – "nested design" (Lieberman 2005) or "nested induction" (Coppedge 2002a) – not on problems of case selection.

In this paper we clarify the methodological issues involved in case selection where the procedure is deliberate rather than randomized and where the population is large. Note that because the total number of cases to be selected is small, statistical properties will not apply neatly to the sampling techniques discussed in this paper. Hence, it would be a mistake to attribute too much analytical leverage to any of these case-selection strategies, as compared with the leveraged derived from within-case analysis. In other words, it often may not matter how cases are chosen, so long as the scholar learns something from them. Yet cases must nonetheless be chosen in some way, so it is useful to ask how we might implement existing ideas about systematic case selection.

PRELIMINARIES

We define a "case study" as the intensive analysis of a single unit or a small number of units (the cases) where the purpose is to understand a larger class of similar units (a population of cases). Case-study research thus requires a sample of one or several cases, each of which is a single, spatially-delimited phenomenon observed either at a given point in time or over some period of time. Within the context of a particular study, the definition of a case is determined by the theoretical interests that motivate the study, for cases exemplify the principal unit of concern in that study. If the study is about the behavior of nation-states then cases are constituted by nation-states. If the study is about

[1] To be sure, this technique is difficult to implement in hard-to-enumerate populations or in populations whose boundaries are not well-established. Even so, the *procedures* for arriving at representative samples in large-N research are well-studied and relatively unproblematic (e.g., Scheaffer, Mendenhall, and Ott 1995).

[2] Sometimes, such variation can be achieved via comparisons with background cases.

individuals, then cases are constituted by individuals. And so forth. One's understanding of a case is thus dependent upon the central proposition an author is intending to prove or demonstrate (Gerring 2004, 2006). It follows that the question of case-selection is meaningless until the researcher has arrived at a specific hypothesis, or at least a research question.

In this paper we are concerned primarily with *causal* inference, rather than inferences that are descriptive or predictive in nature. Thus, all research questions involve at least one independent variable (X) and one dependent variable (Y). If the analyst seeks to explain a puzzling outcome but has no preconceptions about its causes, then the research may be described as *Y-centered*. If a researcher plans to investigate the effects of a particular cause with no preconceptions about what these effects might be, the research may be described as *X-centered*. *Y-centered* and *X-centered* research both begin with a research question but without a hypothesis. If, on the other hand, an analyst wants to explore a particular causal relationship the research may be described as *X/Y-centered*, for it connects a particular cause or set of causes with a particular outcome.[3] This will be known as an argument, hypothesis, or proposition -- terms that we use more or less interchangeably.[4]

Questions about case selection inevitably hinge upon the population of an inference. It is only by reference to this larger set of cases that one can begin to think about which cases might be most appropriate for in-depth analysis. Thus, researchers must specify the set of cases which the research question or a specific proposition might apply to -- the breadth, domain, scope, or population (all are near-synonyms) of the case study.[5]

In the following discussion, we disregard three important considerations: a) pragmatic, logistical issues, b) the theoretical prominence of a case in the literature on a topic, and c) within-case characteristics of a case. All of these factors influence a researcher's selection of cases, and rightly so. However, the first two factors are not methodological in character; they do not bear on the validity of an inference stemming from case study research. Moreover, there is not much that can be said about them that is not already self-evident to the researcher.

The third factor is methodological, properly speaking, and there is a great deal to be said about it (Gerring and McDermott forthcoming). If, for example, Case A provides an example of a change in one of the key variables of interest then it may be preferred over Case B for this reason alone. If that change is quasi-experimental in nature, then Case A is even more likely to provide a useful tool for causal analysis. In this study, however, we focus on factors of case-selection that rest on the *cross-case* characteristics of a case – how the case fits into the theoretically-specified population. This is how the term "case-selection" is typically understood, so we are simply following convention by dividing up the subject in this manner.

TECHNIQUES OF CASE-SELECTION

---

[3] This expands on Mill (1843/1872: 253), who wrote of scientific inquiry as twofold: "either inquiries into the cause of a given effect or into the effects or properties of a given cause."

[4] Note that to pursue an X/Y-centered analysis does *not* imply that the writer is attempting to prove or disprove a monocausal or deterministic argument. The presumed causal relationship between X and Y may be of any sort. X may explain only a small amount of variation in Y. The X/Y relationship may be probabilistic. X may refer either to a single variable or a vector of causal factors. The *only* distinguishing feature of X/Y-centered analysis is that a specific causal factor(s), a specific outcome, and some pattern of association between the two, are hypothesized.

[5] Caution is evidently required when specifying the population. One does not wish to claim too much. Nor does one wish to claim too little. Mistakes can be made in either direction. The important point is that when a researcher restricts an inference to a small population of cases, or to the population that she has studied (which may be quite large), without offering a compelling theoretical justification for that restriction, she is open to the charge of gerrymandering – establishing a domain on no other basis than that certain cases seem to fit the inference under study. Green and Shapiro (1994) call this an "arbitrary domain restriction." The breadth of an inference must make sense; there must be an explicable reason for including some cases and excluding others.

In large-sample research we have observed that the issue of case selection is usually handled by some version of randomization. The law of large numbers in statistics (Stone 1996: xx-xx) tells us that, if the sample consists of a large enough number of independent random draws, the selected cases are highly likely to be fairly representative of the overall population on any given variable. Furthermore, because a large number of cases are chosen, the researcher can usually be assured that the sample will provide sufficient variation on key variables. If cases in the population are distributed homogeneously across the ranges of the variables, then it is highly probable that some cases will be included from each important segment of those ranges. (For situations in which cases with theoretically relevant values of the variables are rare, a stratified sample, which oversamples some values, may be employed.)

A demonstration of the fact that random sampling is likely to produce a representative sample is shown in Figure 1, which shows a histogram of the mean values of five hundred random samples, each consisting of one thousand cases. For each case, one variable has been measured: a continuous variable that falls somewhere between zero and one. In the population, the mean value of this variable is 0.5. How representative are the random samples? A good way of judging this is to compare the means of each of the five hundred random samples with the population mean. As can be seen in the figure, all of the sample means are very close to the population mean. So random sampling was a success, and each of the five hundred samples turns out to be fairly representative of the population.

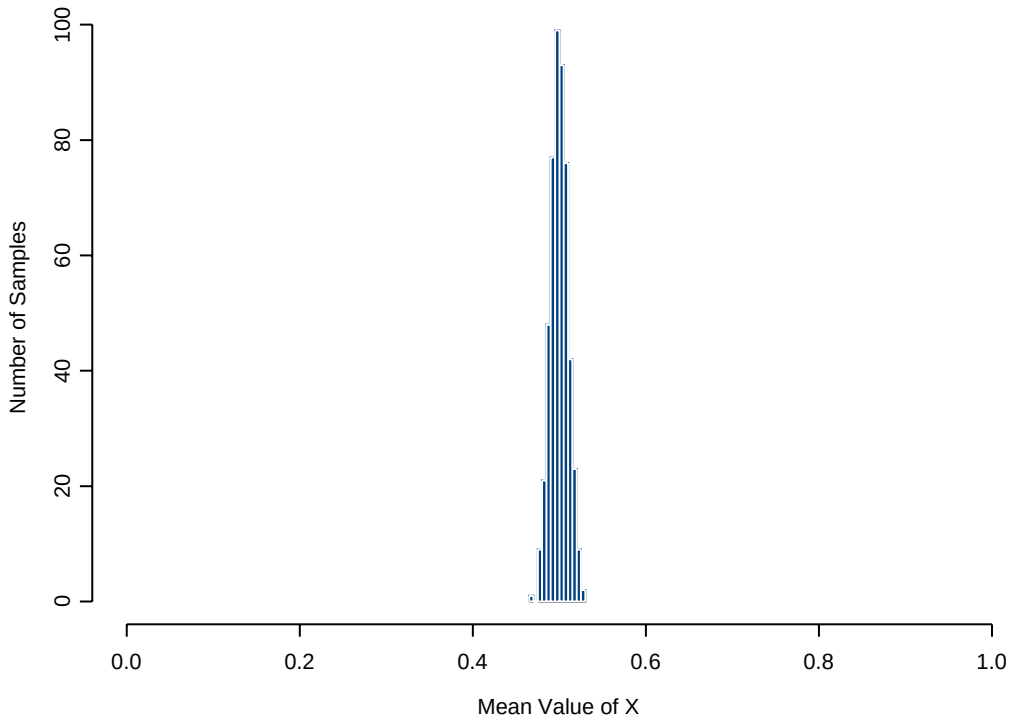

Figure 1: A histogram showing the mean values of one variable in 500 samples of 1000 cases each. The population mean is 0.5.

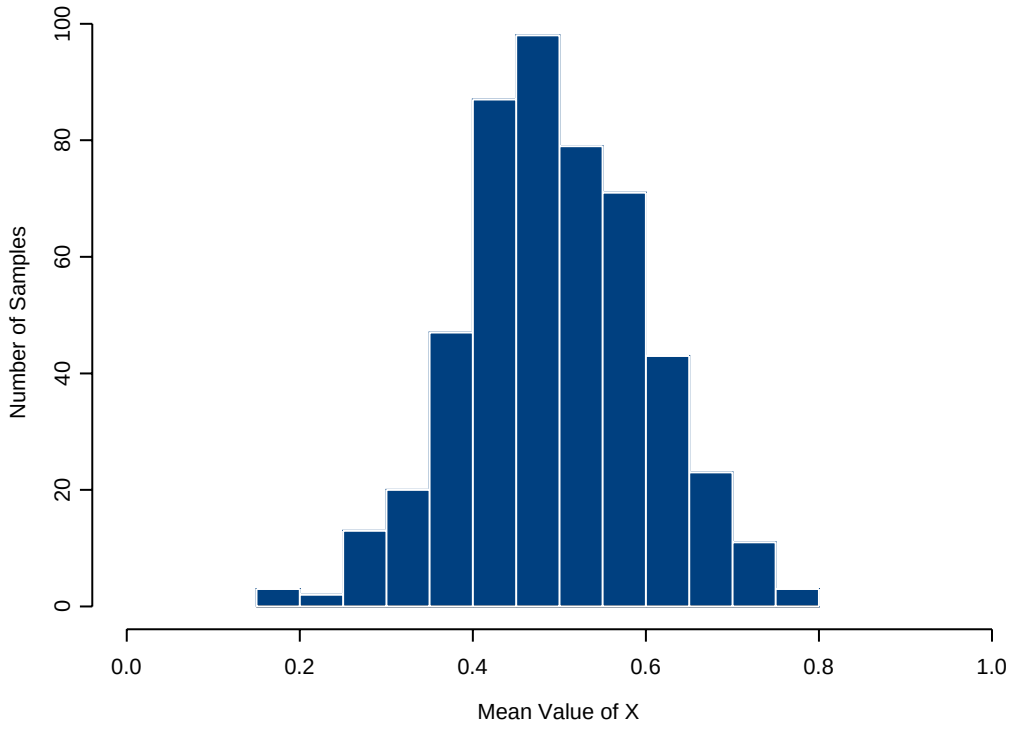

Figure 2: A histogram showing the mean values of one variable in 500 samples of 5 cases each. The population mean is, once again, 0.5.

However, in case-study research the sample is small (by definition) and this makes randomization problematic. Consider what would happen if the sample size was changed from one thousand cases to five. The results are shown in Figure 2. On average, these small-N random samples produce the right answer, so the procedure culminates in results that are unbiased. However, many of the sample means are rather far from the population mean, and some are quite far indeed. Hence, even though this case-selection technique produces representative samples *on average*, any given sample may be wildly unrepresentative. In statistical terms, the problem is that small sample sizes tend to produce estimates with a great deal of variance. Therefore, random sampling is unreliable in small-N research.

How, then, are we to choose a sample for case-study analysis? Case selection in case study research has the same twin objectives as random sampling. That is, one desires a) a representative sample and b) useful variation on the dimensions of theoretical interest.[6] One's choice of cases is therefore driven by the way a case is situated on such dimensions within the population of interest. It is from such cross-case characteristics that we derive the following seven case study types: *typical*, *diverse, extreme*, *deviant*, *influential*, *most-similar*, and *most-different*.

Table 1 summarizes each case study type, including its general definition, a technique for identifying it within a large-N population, its uses, and its probable representativeness. Note that most of these methods may be practiced on a single case, while three – the diverse, most-similar, and most-different – require at least two cases. However, all *may* employ additional cases, with the proviso that, at some point, they will no longer offer an opportunity for in-depth analysis and will thus no longer be "case studies" in the usual sense.

Our discussion of these techniques will follow a fairly straightforward procedure: we will briefly state an idea about case selection from the tradition of case-study research, we will specify the central issue involved in that approach to case selection, and then we will review available statistical tools for addressing this issue. It should be clear that the goal of this paper is not to develop new quantitative estimators but rather to show how existing estimators can be put to good use in case selection.

To be sure, it is not always useful -- or possible -- to employ quantitative case-selection techniques in small-N research. Several caveats must be satisfied. First, the inference must pertain to a reasonably large population. If the population of cases is less than several dozen, statistical techniques will not be applicable. Second, relevant data must be available for that population, or a sizeable sample of that population, on all of the key variables, and the researcher must feel reasonably confident in the accuracy and conceptual validity of these variables. Third, all the standard assumptions of statistical research (e.g., identification, specification, robustness) must be carefully considered. We shall not dilate further on these matters except to warn the researcher against the unthinking use of statistical techniques. Finally, it may be worthwhile to recall that case selection is often an iterative process; within-case research may suggest revisions to the statistical techniques used to select cases, potentially leading to a new sample and new opportunities for within-case analysis.

The exposition will be guided by an ongoing example, the (presumably causal) relationship between per capita GDP and level of democracy (Lipset 1959). Figure 3 displays the basic data about this relationship in the form of a scatterplot. The Democracy variable from the Polity IV dataset (Marshall and Jaggers 2000) is used as a measure of democracy; per capita GDP data is taken from the Penn World Tables dataset (Summers and Heston 1991). The classical result in this field is strikingly illustrated: wealthy countries are almost exclusively democratic. For heuristic purposes, certain unrealistic simplifying assumptions will be adopted. We shall assume, for example, that the Polity measure of democracy is continuous and unbounded (but see Trier and Jackman 2003). We shall assume, more importantly, that the true relationship between economic development and

[6] Where multiple cases are chosen the researcher must also be aware of problems of case-independence. However, these problems are in no sense unique to case study work (Gerring 2001: 178-81).

democracy is log-linear, positive, and causally asymmetric, with economic development treated as exogenous and democracy as endogenous (but see Gerring et al. forthcoming; Przeworski et al. 2000).

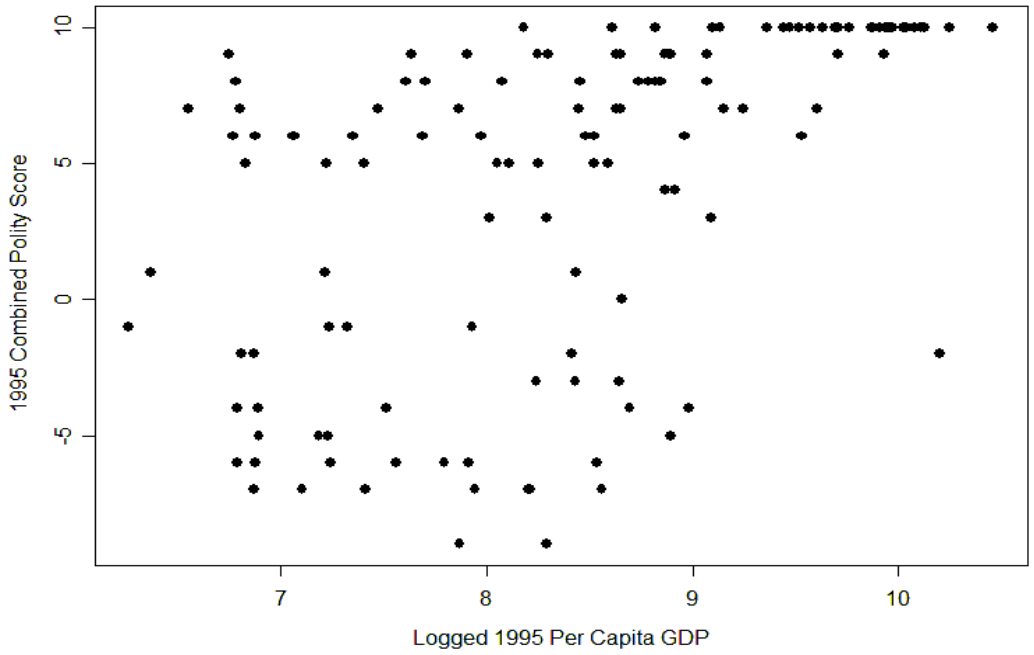

Figure 3: A scatterplot showing level of democracy (on the vertical axis) and level of wealth (on the horizontal axis) of all available countries in 1995. A total of 131 countries have 1995 scores in both data sets.

Table 1:

Cross-Case Methods of Case-Selection and Analysis

1. Typical
- *Definition:* Cases (1 or more) are typical examples of some cross-case relationship.
- *Large-N technique:* A low-residual case (on-lier).
- *Uses:* Confirmatory. To probe causal mechanisms that may either confirm or disconfirm a given theory.
- *Representativeness:* By definition, the typical case is representative, given the specified relationship.

2. Diverse
- *Definition:* Cases (2 or more) exemplify diverse values of X, Y, or X/Y.
- *Large-N technique:* Diversity may be calculated by a) categorical values of X or Y (e.g., Jewish, Catholic, Protestant), b) standard deviations of X or Y (if continuous), c) combinations of values (e.g., based on cross-tabulations, factor analysis, or discriminant analysis).
- *Uses:* Exploratory or confirmatory. Illuminates the full range of variation on X, Y, or X/Y.
- *Representativeness:* Diverse cases are likely to be representative in the minimal sense of representing the full variation of the population. (Of course, they may not mirror the *distribution* of that variation in the population.)

3. Extreme
- *Definition:* Cases (1 or more) exemplify extreme or unusual values on X or Y relative to some univariate distribution.
- *Large-N technique:* A case lying many standard deviations away from the mean of X or Y.
- *Uses:* Exploratory. Open-ended probe of X or Y.
- *Representativeness:* Achievable only in comparison with a larger sample of cases.

4. Deviant
- *Definition:* Cases (1 or more) deviate from some cross-case relationship.
- *Large-N technique:* A high-residual case (outlier).
- *Uses:* Exploratory or confirmatory. To probe new explanations for Y, to disconfirm a deterministic argument, or to confirm an existing explanation (rare).
- *Representativeness:* After the case study is conducted it may be corroborated by a cross-case test, which includes a general hypothesis (a new variable) based on the case study research. If the case is now an on-lier, it may be considered representative of the new relationship.

5. Influential
- *Definition:* Cases (1 or more) with influential configurations of the independent variables.
- *Large-N technique:* Hat matrix or Cook's Distance.
- *Uses:* Confirmatory. To double-check cases that influence the results of a cross-case analysis.
- *Representativeness:* An influential case is typically not representative. If it were typical of the sample as a whole, it would not have unusual influence on estimates of the overall relationship.

6. Most-similar
- *Definition:* Cases (2 or more) are similar on specified variables other than X_1 and/or Y.
- *Large-N technique:* Matching.
- *Uses:* Exploratory if the hypothesis is X- or Y-centered. Confirmatory if X/Y-centered.
- *Representativeness:* Most-similar cases that are broadly representative of the population will provide the strongest basis for generalization.

7. Most-different
- *Definition:* Cases (2 or more) are different on specified variables other than X_1 and Y.
- *Large-N technique:* The inverse of the most-similar method of large-N case selection (see above).
- *Uses:* Exploratory or confirmatory. To a) eliminate necessary causes (definitively), or to b) provide weak evidence of the existence of a causal relationship.
- *Representativeness:* Most-different cases that are broadly representative of the population will provide the strongest basis for generalization.

Note: X_1 refers to the causal factor of theoretical interest.

TYPICAL CASE

The typical case study focuses on a case that exemplifies a stable, cross-case relationship. By construction, the typical case may also be considered a *representative* case, according to the terms of whatever cross-case model is employed. Indeed, the latter term is often employed in the psychological literature (e.g., Hersen and Barlow 1976: 24).

Because the typical case is well-explained by an existing model, the puzzle of interest to the researcher lies *within* that case. Specifically, the researcher wants to find a typical case of some phenomenon so that she can better explore the causal mechanisms at work in a general, cross-case relationship. This exploration of causal mechanisms may lead toward several different conclusions. If the existing theory suggests a specific causal pathway, then the researcher may perform a pattern-matching investigation, in which the evidence at hand (in the case) is judged according to whether it validates the stipulated causal mechanisms or not. If it does not, the researcher may try to show that the causal mechanisms are other than previously stipulated. Or she may argue that there are *no* plausible causal mechanisms connecting this independent variable with this particular outcome. In the latter case, a typical-case research design may provide disconfirming evidence of a general causal proposition. But the usual employment of a typical case is to provide support for, or clarification of, an existing causal hypothesis.

Large-N Analysis. How does one identify a typical case from a large population of potential cases? Suppose that an arbitrary case in the population, denoted as case number *i*, has a known score on each of several relevant variables. For the sake of economy of language, let the variables involved in the relationship be labelled y_i and $x_{1,i}, \ldots x_{K,i}$, where y_i is the score of case *i* on one variable and each of the $x_{k,i}$'s is the score of case *i* on one of the *k* other variables under consideration. Thus, the relationship involves a total of $K + 1$ variables. K can be any integer greater than or equal to 1.

With these symbols, the established relationships among the variables can be expressed mathematically. The idea is to find a function, f(), such that the average score of y for cases with some specific set of scores on $x_1 \ldots x_K$ is equal to $f(x_1, \ldots x_K)$. Thus, the function f() should be chosen to capture the key ideas about the relationship of interest. A familiar example may make this discussion clearer.

Often, researchers choose an additive, linear function to play the role of f(). Using traditional statistical notation, in which the average score of y_i across infinite repetitions of case *i* is denoted by $E(y_i)$, an additive, linear function represents a relationship in which:

$$E(y_i) = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_K x_{K,i} \quad (1)$$

Each of the β_k's in this equation represents an unknown constant. Regression analysis allows researchers to use known information about the y and $x_1 \ldots x_K$ variables for a set of cases to estimate these unknown constants. Since the case selection procedures discussed here allow for the possibility that at least some cases may in fact deviate from the overall relationship, it is probably best to estimate the β_k's using a robust regression procedure (Rousseeuw and Leroy 2003), as is done in the examples below. Estimates of β_k will be denoted here as b_k.

Note that nothing in this discussion requires the relationship under consideration to be causal in nature. Nor is it necessary to think of the y variable as an outcome and the $x_1 \ldots x_K$ variables as causes. All we must assume is that the relationship under discussion is an established regularity among the cases in the population.

Using this terminology, we can now develop a formula for the degree to which a particular case is typical in light of a given relationship. A case is "typical" in the terms of small-N methodology to the extent that its score on the y variable is close to the average score on that variable for a case with the same scores on the $x_1 \ldots x_K$ variables, as given by Equation 1. That is,

$$\text{Typicality}(i) = -\text{abs}[y_i - E(y_i \mid x_{1,i},\ldots x_{K,i})] = -\text{abs}[y_i - b_0 + b_1 x_{1,i} + \ldots + b_K x_{K,i}] \quad (2)$$

According to this discussion, the typicality of a case with respect to a particular relationship in qualitative terms is just -1 times the absolute value of that case's error term in regression analysis. This measure of typicality ranges, in theory, from negative infinity to zero. When a case falls close to the regression line, and is therefore quite typical, the typicality will be just below zero. By contrast, when a case falls far from the regression line, and is therefore very atypical, the typicality will be far below zero. Hence, selecting typical cases involves selecting those cases with a typicality score close to zero.

For the specific case of an additive, linear function, this equation provides a mathematically precise expression of the underlying idea of typicality, as used in qualitative methodology. If the researcher is interested in choosing, for example, five typical cases, that can be accomplished by computing the typicality statistic for each case and then choosing the five cases with the least negative typicality scores, i.e., the scores closest to zero.

In a large-N sample there will often be many cases with almost identical high (i.e., near-zero) typicality scores. In such situations, researchers may elect not to focus on the cases with the highest estimated typicality, for such estimates may not be accurate enough to distinguish among several almost-identical cases. Instead, researchers may choose to randomly select from the set of cases with very high typicality, or even to choose from among these cases according to non-methodological criteria. However, scholars should try to avoid selecting from among the set of typical cases in a way that is correlated with relevant omitted variables; such selection procedures complicate the task of causal inference.

Example. Returning to the example introduced above, involving the relationship between per capita GDP and level of democracy, how might a set of typical cases be selected? In this instance, it is relatively easy to specify an appropriate function capturing the relevant relationship. After all, the y variable is simply the Polity democracy score, and there is only one independent variable: logged per capita GDP. Hence, the simplest relevant model is:

$$E(\text{Polity}_i) = \beta_0 + \beta_1 \text{GDP}_i \quad (3)$$

Scholars may also wish to include other nonlinear transformations of the logged per capita GDP variable, in order to allow a more flexible functional form. In the current example, we will add a quadratic term. Hence, the model to be considered is:

$$E(\text{Polity}_i) = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{GDP}_i^2 \quad (4)$$

For the purposes of selecting typical cases, the specific coefficient estimates are relatively unimportant, but we will report them, to two digits after the decimal, for the sake of completeness:

$$E(\text{Polity}_i) = 10.52 - 4.59\ \text{GDP}_i + 0.45\ \text{GDP}_i^2 \quad (5)$$

Much more important are the residuals for each case. Figure 4 shows a histogram of these residuals. Obviously, a fairly large number of cases have quite low residuals and are therefore considered typical. A higher proportion of cases fall far below the regression line than far above it, suggesting either that the model may be incomplete or that the error term does not have a normal distribution. Hopefully, within-case analysis will be able to shed light on the reasons for the assymmetry.[7]

[7] In this example, the asymmetry is probably due to the failure of the model to take into account the restricted range of the dependent variable, as discussed above.

Because of the large number of cases with quite small residuals, the researcher will have a range of options for selecting typical cases. In fact, in this example, 26 cases have a typicality score between 0 and -1. Any or all of these might reasonably be selected for within-case analysis as typical cases with respect to the model described in Equation 4.

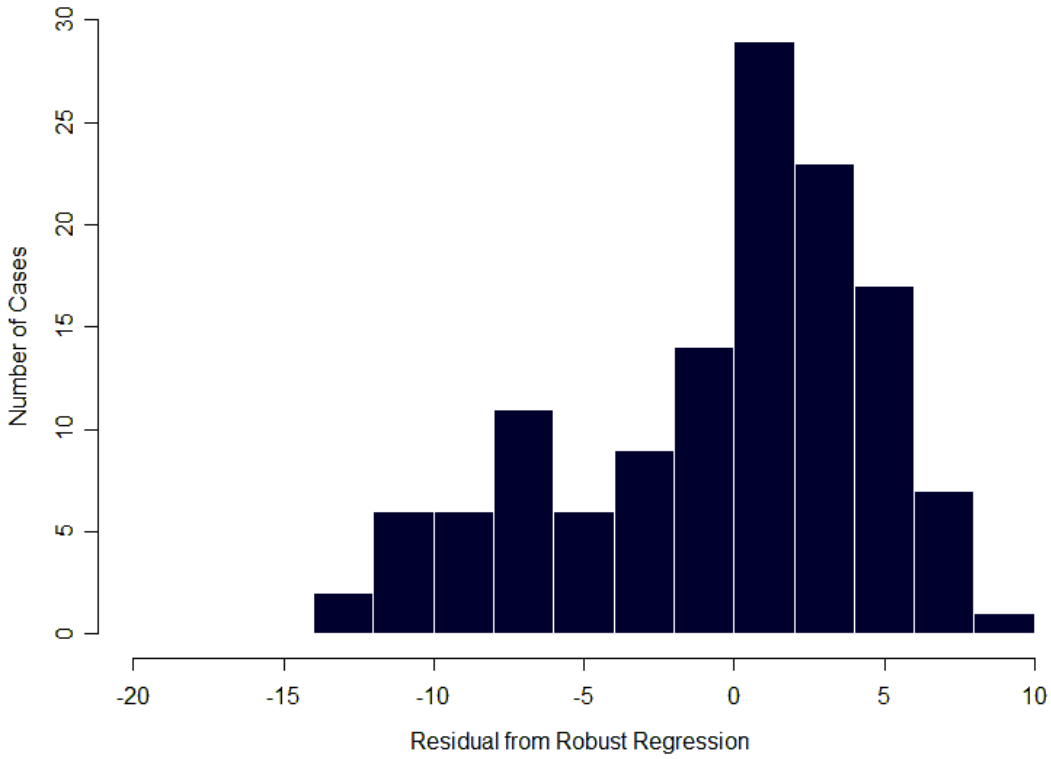

Figure 4: A histogram of the residuals from a robust regression of logged per capita GDP on level of democracy.

Conclusion. Typicality responds to the first desideratum of case selection, that the chosen case be representative of a population of cases (as defined by the primary inference). Even so, it is important to remind ourselves that the single-minded pursuit of representativeness does not ensure that it will be achieved. Note that the test of typicality introduced here, the size of a case's residual, can be misleading if the statistical model is misspecified. And it provides little insurance against errors that are purely stochastic. A case may lie directly on the regression line but still be, in some important respect, a-typical. For example, it might have an odd combination of values; the interaction of variables might be different from other cases; or additional causal mechanisms might be at work.

DIVERSE CASES

A second case-selection strategy has as its primary objective the achievement of maximum variance along relevant dimensions. We refer to this as a *diverse-case* method. For obvious reasons, this method requires the selection of a set of cases – at minimum, two – which are intended to represent the full range of values characterizing X_1, Y, or some particular X_1/Y relationship.[8] As previously, the investigation is understood to be exploratory when the researcher focuses on X or Y, and confirmatory when she focuses on a particular X_1/Y relationship (a specific hypothesis).

Where the individual variable of interest is categorical (on/off, red/black/blue, Jewish/Protestant/Catholic), the identification of diversity is readily apparent. The investigator simply chooses one case from each category. For a continuous variable, the choices are not so obvious. However, the researcher usually chooses both extreme values (high and low), and perhaps the mean or median as well. The researcher may also look for break-points in the distribution that seem to correspond to categorical differences among cases. Or she may follow a theoretical hunch about which threshold-values count, i.e., which are likely to produce different values on Y. Where the causal factor of interest is a *vector* of variables, and where these factors can be measured, the researcher may simply combine various causal factors into a series of cells, based upon cross-tabulations of factors deemed to have an effect on Y. Let us say that some outcome is thought to be affected by sex, race (black/white), and marital status. Here, a diverse-case strategy of case-selection would identify one case within each of these intersecting cells – a total of eight cases. Things become slightly more complicated when one or more of these factors is continuous, rather than dichotomous, since the researcher will have to arbitrarily re-define that variable as a categorical variable (as above). Where causal variables are continuous and the outcome is dichotomous, the researcher may employ discriminant analysis to identify diverse cases.

Diversity may also be understood in terms of various causal paths running from exogenous factors to a particular outcome. Perhaps X_1, X_2, and X_3 all cause Y, but they do so independently of each other and in different ways. Each is a sufficient cause of Y.[9] George and Smoke, for example, wish to explore different types of deterrence failure – by "fait accompli," by "limited probe," and by "controlled pressure." Consequently, they wish to find cases that exemplify each type of causal

[8] This method has not received much attention on the part of qualitative methodologists; hence, the absence of a generally recognized name. It bears some resemblance to J.S. Mill's Joint Method of Agreement and Difference (Mill 1834/1872), which is to say, a mixture of most-similar and most-different analysis, as discussed below. Patton (2002: 234) employs the concept of "maximum variation (heterogeneity) sampling."

[9] This sometimes referred to as causal equifinality (Elman forthcoming; George and Bennett 2004).

mechanism.[10] This may be identified by a traditional form of path analysis, by Qualitative Comparative Analysis (QCA), by sequence analysis, or by informal (non-quantitative) methods.[11]

"Diversity" may thus refer to a range of variation on X or Y, to a particular combination of causal factors, or to various causal pathways. In each instance, the goal of case-selection is to capture the full range of causal types along the dimension of interest -- either X_1, Y, or X_1/Y.

Large-N Analysis. Diverse-case selection is easily accommodated in a large-N context by using some version of stratified random sampling. In this approach, the researcher identifies the different substantive categories of interest, as well as the number of cases to be chosen from each category. Then, the needed cases may be randomly chosen from among those available in each category (Cochran 1977: xx-xx).

One assumes that the identification of diverse categories of cases will, at the same time, identify categories that are *internally* homogenous (in all respects that might affect the causal relationship of interest). Thus, the choice of cases within each category should be relatively unproblematic. However, if there is suspected diversity within each category, then measures should be taken to assure that the chosen cases are typical of each category. A case study should not focus on an atypical member of a sub-group.

Indeed, considerations of diversity and typicality often go together. Thus, in a study of globalization and social welfare systems, Duane Swank first identifies three distinctive groups of welfare states: "universalistic" (social democratic), "corporatist conservative," and "liberal." Next, he looks within each group to find the most-typical cases. He decides that the Nordic countries are more typical of the universalistic model than the Netherlands since the latter has "some characteristics of the occupationally based program structure and a political context of Christian Democratic-led governments typical of the corporatist conservative nations."[12]

Conclusions. Encompassing a full range of variation is likely to enhance the representativeness of the sample of cases chosen by the researcher. This is a distinct advantage. Of course, the inclusion of a full range of variation may distort the actual distribution of cases across this spectrum. If there are more "high" cases than "low" cases in a population and the researcher chooses only one high case and one low case, the resulting sample of two is not perfectly representative. Even so, the diverse-case method probably has stronger claims to representativeness than any other small-N sample (including the typical case).

EXTREME CASE

The extreme-case method selects a case because of its extreme value on the independent (X) *or* dependent (Y) variable of interest. Thus, studies of domestic violence may focus on extreme instances of abuse (Browne 1987). Studies of altruism may focus on those who risked their lives to help others (e.g., Holocaust resisters; Monroe 1996). Studies of ethnic politics may focus on the most heterogeneous societies in order to better understand the role of ethnicity in a democratic setting (e.g., Papua New Guinea; Reilly 2000/2001). Studies of industrial policy focus on the most

---

[10] More precisely, George and Smoke (1974: 534, 522-36, ch 18; see also discussion in Collier and Mahoney 1996: 78) set out to investigate causal pathways and discovered, through the course of their investigation of many cases, these three causal types. Yet, for our purposes what is important is that the final sample include at least one representative of each "type."

[11] Path analysis is discussed in most introductory statistics texts. QCA is discussed in Drass and Ragin (1992), Hicks (1999: 69-73), Hicks, Misra, Ng (1995), Ragin (1987, 2000), and several chapters by Ragin in Janoski and Hicks (1993). Sequence analysis is explained in Abbott and Tsay (2000).

[12] Swank (2002: 11). See also Esping-Andersen (1990).

successful countries (i.e., the NICs; Deyo 1987). And so forth (for further examples see Collier and Mahoney 1996; Geddes 1990).

The notion of "extreme" may now be defined more precisely. An extreme value is an observation that lies far away from the mean of a given distribution. That is to say, it is *unusual*. If most cases are positive along a given dimension, then a negative case constitutes an extreme case. If most cases are negative, then a positive case constitutes an extreme case. Evidently, one is not simply concerned with cases where something "happened," but also with cases where something did not. For case-study analysis, it is often the rareness of the value that makes a case valuable, not its positive or negative value (contrast Emigh 1997; Mahoney and Goertz 2004; Ragin 2000: 60; Ragin 2004: 126).

Large-N Analysis. As we have said, extreme cases lie far from the mean of a variable. Extremity (E) for the *i*th case, can be defined in terms of the sample mean ($\bar{X}$) and the standard deviation (s) for that variable:

$$E_i = \left| \frac{X_i - \bar{X}}{s} \right|$$

This definition of extremity is the absolute value of the Z-score (Stone 1996: xx-xx) for the *i*th case. Cases with a large E_i qualify as extreme. Decisions about how large the extreme-ness needs to be in order for cases to count as extreme are, to some extent, arbitrary. However, some general guidelines can be offered. In keeping with statistical tradition, cases with an extremeness score smaller than 2 would generally not be considered extreme, even if they are the most extreme cases in the sample. If the researcher wishes to be somewhat conservative in classifying cases as extreme, a higher threshold such as 3, can be used. In general, the choice of threshold is left to the researcher, to be made in a way that is appropriate to the research problem at hand.

Example. The mean of the democracy measure is 2.76, suggesting that, on average, the countries in the 1995 data set tend to be somewhat democratic. The standard deviation is 6.92, implying that there is a fair amount of scatter around the mean in these data.

Figure 5 shows a histogram of the extremeness scores for all countries on level of democracy. As can easily be seen, no cases have extremeness scores greater than two. Hence, some flexibility is required in choosing extreme cases on this variable. The two countries with the largest extremeness scores are Qatar and Saudi Arabia, both of which have an extremeness of 1.84. These countries, which both have a democracy score of -10 for 1995, are probably the two best candidates for extreme cases.

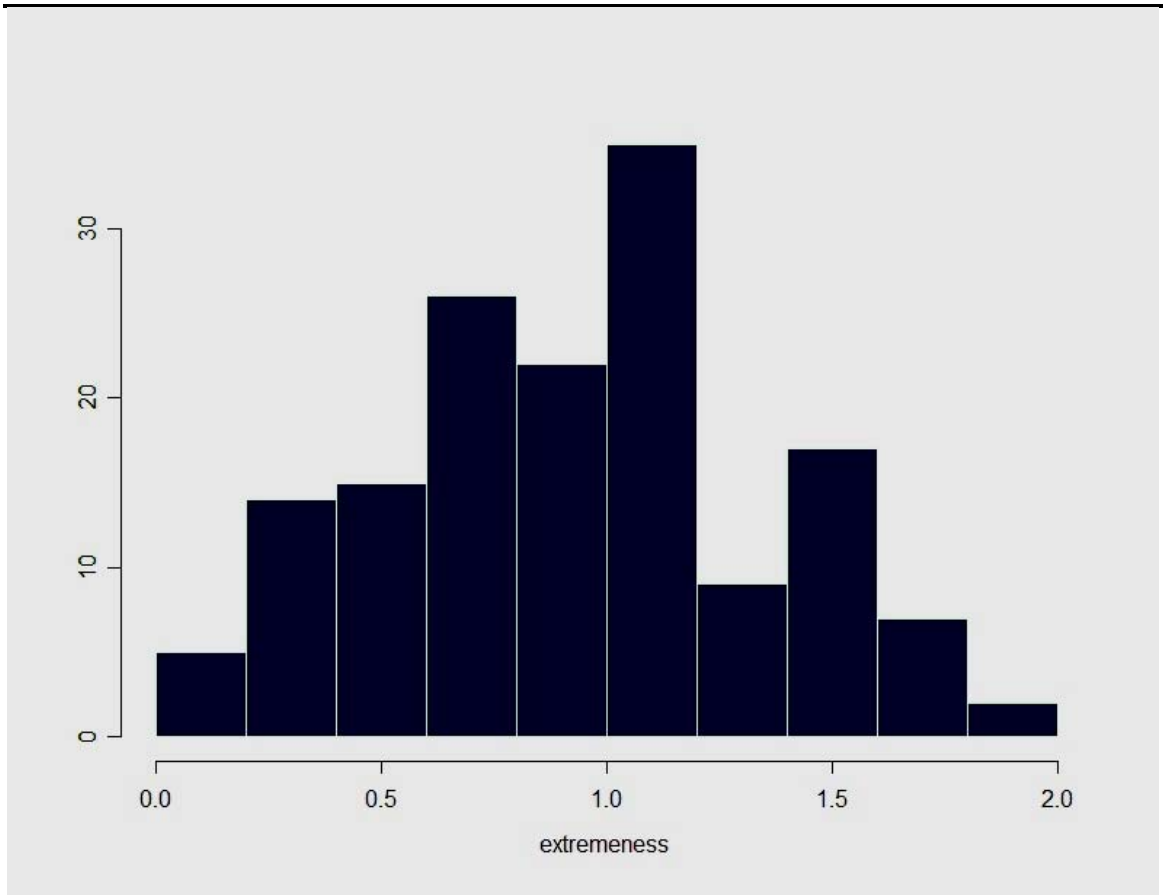

Figure 5: A histogram of the extremeness scores for all countries in the data set on level of democracy.

Conclusion. The extreme-case method appears to violate the social science folk wisdom warning us not to "select on the dependent variable."[13] Selecting cases on the dependent variable is indeed problematic if a number of cases are chosen, all of which lie on one end of a variable's spectrum (they are all positive *or* negative), and if the researcher then subjects this sample to cross-case analysis as if it were representative of a population.[14] Results for this sort of analysis would almost assuredly be biased. Moreover, there will be little variation to explain since the contrasting values of each case are explicitly constrained.

However, this is not the proper employment of the extreme-case method. (It is more appropriately labeled an extreme-*sample* method.) The extreme-case method refers back to a larger sample of cases that lie in the background of the analysis and provide a full range of variation as well as a more representative picture of the population. It is a conscious attempt to *maximize* variance on the dimension of interest, not to minimize it. If this population of cases is well understood -- either through the author's own cross-case analysis, through the work of others, or through common sense – then a researcher may justify the selection of a single case exemplifying an extreme value for within-case analysis. If not, the researcher may be well-advised to follow a diverse-case method (see below).

By way of conclusion, let us return to the problem of representativeness. In the context of causal analysis, representativeness refers to a case that exemplifies values on X and Y that conform to a general pattern. In a cross-case model, the representativeness of an individual case is gauged by the size of its residual. The representative case is therefore a typical case (as discussed above), not a deviant case (as discussed below). It will be seen that an extreme case may be typical *or* deviant. There is simply no way to tell because the researcher has not yet specified an X/Y causal proposition. Once such a causal proposition has been specified we may then ask whether the case in question is similar – in all respects that might affect the X/Y relationship of interest – to some population of cases. It is at this point that it becomes possible to say (within the context of a cross-case statistical model) whether a case lies near to, or far from, the regression line. However, this sort of analysis means that the researcher is no longer pursuing an extreme-case method. The extreme case method is purely exploratory -- a way of probing possible causes of Y, or possible effects of X, in an open-ended fashion. If the researcher has some notion of what additional factors might affect the outcome of interest, or of what relationship the causal factor of interest might have on Y, then she ought to pursue one of the other methods explored below. This also implies that an extreme-case method might morph into a different kind of approach as a study evolves, that is, as a more specific hypothesis comes to light.

DEVIANT CASE

The *deviant-case* method selects that case(s) which, by reference to some general understanding of a topic (either a specific theory or common sense), demonstrates a surprising value. Barbara Geddes notes the importance of deviant cases in medical science, where researchers are habitually focused on that which is "pathological" (according to standard theory and practice).[15] Likewise, in psychology and sociology case studies may be comprised of deviant (in the social sense) persons or groups. In economics, case studies may consist of countries or businesses that over-

[13] Geddes (1990), King, Keohane, and Verba (1994). See also discussion in Brady and Collier (2004), Collier and Mahoney (1996), Rogowski (1995).

[14] The exception would be a circumstance in which the researcher intends to disprove a deterministic argument (Dion 1998).

[15] Geddes (2003: 131). For other examples of case work from the annals of medicine see "Clinical Reports" in *The Lancet*, "Case Studies" in *The Canadian Medical Association Journal*, and various issues of the *Journal of Obstetrics and Gynecology*, often devoted to clinical cases (discussed in Jenicek 2001: 7).

perform (e.g., Botswana; Microsoft) or under-perform (e.g., Britain through most of the twentieth century; Sears in recent decades) relative to some set of expectations. In political science, case studies may focus on countries where the welfare state is more developed (e.g., Sweden) or less developed (e.g., the United States) than one would expect, given a set of general expectations about welfare state development. The deviant case is closely linked to the investigation of theoretical anomalies. Indeed, to say "deviant" is to imply "anomalous."[16]

Note that while extreme cases are judged relative to the mean of a single distribution (the distribution of values along a single variable), deviant cases are judged relative to some general model of causal relations. The deviant-case method selects cases which, by reference to some general cross-case relationship, demonstrate a surprising value. They are "deviant" in that they are poorly explained by the multivariate model. The important point is that deviant-ness can only be assessed relative to the general (quantitative or qualitative) model employed. This means that the relative deviant-ness of a case is likely to change whenever the general model is altered. For example, the US is a deviant welfare state when this outcome is gauged relative to societal wealth. But it is less deviant – and perhaps not deviant at all -- when certain additional (political and societal) factors are included in the model.[17] Deviance is model-dependent. Thus, when discussing the concept of the deviant case it is helpful to ask the following question: *relative to what general model* (or set of background factors) is Case A deviant?

The purpose of a deviant-case analysis is usually to probe for new – but as yet unspecified - explanations. In this circumstance, the deviant-case method is only slightly more bounded than the extreme-case method. It, too, is an exploratory form of research. The researcher hopes that causal processes within the deviant case will illustrate some causal factor that is applicable to other (deviant) cases. This means that, in most circumstances, a deviant-case study culminates in a general proposition – one that may be applied to other cases in the population.

However, there is also a second, less common, reason for choosing a deviant case. If the researcher is interested in *disconfirming a deterministic proposition*, then any deviant case will do so long as it lies within the specified population of the inference.[18] Deterministic arguments may be framed as necessary, sufficient, or necessary and sufficient. They usually apply to variables that are understood to take dichotomous, rather than continuous, outcomes (or where scalar outcomes are dichotomized as high/low, big/small and so forth). Evidently, if a presumed necessary cause is not present when its effect is present, or a sufficient cause is not accompanied by its presumed effect, then the causal proposition has been disconfirmed. Note that while simply identifying a disconfirming case is enough in some situations, a careful case-study may still be important as a means of persuading readers that the relevant variables have been measured correctly and therefore that the case is, in fact, disconfirming.[19]

---

[16] For a discussion of the important role of anomalies in the development of scientific theorizing see Elman (2003), Lakatos (1978). For examples of deviant-case research designs in the social sciences see Amenta (1991), Coppedge (2004), Eckstein (1975), Emigh (1997), Kendall, Wolf (1949/1955).

[17] Alesina and Glaeser (2004).

[18] Dion (1998).

[19] Of course, there may be some ambiguity over the proper scope of the proposition. Defenders of the hypothesis may respond that the chosen case lies outside the population of expected cases. They may also question whether the researcher has reached the correct conclusion regarding the deviant case under study. However, as a general rule, deterministic arguments are vulnerable to single-case examples. If such cases are found, and if the researcher can make plausible arguments about their inclusion in the population of an inference and their correct classification as deviant cases, this constitutes very strong evidence against the plausibility of the argument in question. Or, alternatively, it may prompt the researcher to suggest a reframing of that hypothesis. She might suggest that the argument, rather than being deterministic, is in fact probabilistic; there are exceptions. She might suggest a change in the scope conditions (the population) of the inference. She might suggest that the argument requires further qualifying conditions. X may still be a condition of Y, but only under certain circumstances (this may be understood as a scope condition or as an argument about the

Large-N Analysis. In statistical terms, deviant-case selection is the opposite of typical-case selection. Where a typical case is as close as possible to the prediction of a formal, mathematical representation of the hypothesis at hand, a deviant cases is as far as possible from that prediction. Hence, referring back to the model developed in Equation 1, we can define the extent to which a case deviates from the predicted relationship as follows:[20]

$$\text{Deviant-ness } (i) = \text{abs}[y_i - E(y_i \mid x_{1,i},\ldots x_{K,i})] = \text{abs}[y_i - b_0 + b_1 x_{1,i} + \ldots + b_K x_{K,i}] \qquad (6)$$

Deviant-ness ranges from 0, for cases exactly on the regression line, to a theoretical limit of positive infinity. Researchers will typically be interested in selecting from the cases with the highest overall estimated deviant-ness.

Note that when the purpose of a deviant-case analysis is exploratory – i.e., when an author is searching for new causal factors that will be relevant across the broader set of cases – then missing variables are usually not problematic. The caveat is that the effort to identify these missing variables must be the focus of an author's research.

Example. In our running example, the most deviant cases fall below the regression line, as can be seen in Figure 4. In fact, all eight of the cases with a deviant-ness score of more than 10 are below the regression line. Those eight cases are: Croatia, Cuba, Indonesia, Iran, Morocco, Singapore, Syria, and Uzbekistan. An analysis focused on deviant cases might well select a subset of these.

Conclusion. As we have noted, the deviant-case method is usually an exploratory form of analysis. As soon as a researcher's exploration of a particular case has identified a factor to explain that case, it is no longer (by definition) deviant. (The exception would be a circumstance in which a case's odd combination of values is deemed to be "accidental," and therefore unexplainable by any general model.) If the new explanation can be accurately measured as a single variable (or set of variables) across a larger sample of cases, then a new cross-case model is in order. In this fashion, a case study initially framed as "deviant-case" may transform into some other sort of analysis.

This feature of the deviant-case study also helps to resolve questions about its representativeness. Evidently, the representativeness of a deviant case is problematic since the case in question is, by construction, a-typical. However, doubts about representativeness can be mitigated if the researcher generalizes whatever proposition is provided by the case study to other cases. In a statistical model, this is accomplished by the creation of a new variable, as discussed. In a small-N setting, it may be accomplished by the coding of adjacent cases so as to determine whether they confirm (or at least do not openly contradict) the hypothesis. In each scenario, the outcome of this new cross-case analysis should pull the deviant case towards the expected value predicted by the general model. The deviant case is no longer deviant; in statistical terms, its residual has shrunk. It is now typical, or at least more typical.

INFLUENTIAL CASE

Sometimes, the choice of a case is motivated solely by the need to check the assumptions behind some general model of causal relations. In this circumstance the extent to which a case fits the overall model is important only insofar as it might affect the overall set of findings for the whole population. Once cases that do influence overall findings have been identified, it is important to

---

inclusion of proper controls). For further discussion of necessary conditions see Dion (1998), Goertz, Starr (2003).

[20] We use the somewhat awkward term "deviant-ness" rather than the more natural "deviance" because deviance already has a somewhat different meaning in statistics.

decide whether or not they genuinely fit in the sample (and whether they might give clues about important missing variables). Because the techniques for identifying this sort of case are slightly different than those used to identify the previous sort of deviant case, we apply a new term to this method – the *influential case*. This is consistent with the goal of this style of case study, to explore cases that may be influential vis-à-vis some larger cross-case theory.

Large-N Analysis. Influential cases in regression are those cases that, if counterfactually assigned a different value on the dependent variable, would most substantially change the resulting estimates. Two quantitative measures of influence are commonly applied in statistical analysis. The first, often referred to as the leverage of a case, derives from what is called the "hat matrix."[21] Suppose that the scores on the independent variables for all of the cases in a regression are represented by the matrix **X**, which has N rows (representing each of the N cases) and $K + 1$ columns (representing the K independent variables and allowing for a constant). Further, allow **Y** to represent the scores on the dependent variable for all of the cases. Obviously, **Y** will have N rows and only 1 column.

Using these symbols, the formula for the hat matrix, **H**, is as follows:

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{X}^{\mathrm{T}} \qquad (6)$$

In this equation, the symbol T represents a matrix transpose operation, and the symbol "-1" represents a matrix inverse operation (see Greene 2002 for a brief review). A measure of the leverage of each case can be derived from the diagonal of the hat matrix. Specifically, the leverage of case i is given by the number in the (i,i) position in the hat matrix, or $\mathbf{H}_{i,i}$.[22]

For any **X** matrix, the diagonal entries in the hat matrix will automatically total up to 1. Hence, interpretations of the leverage scores for different cases will necessarily depend on the overall number of cases. Clearly, any case with a score near 1 is a case with a great deal of leverage. In most regression situations, however, no case has a score near 1. A standard rule of thumb is to pay close attention to cases with a leverage score higher than $2(K + 1)/N$. Cases with a leverage score above this value are good candidates for influential-case selection.

An interesting feature of the hat matrix is that it does not depend on the values of the dependent variable. Indeed, the **Y** vector does not appear in Equation 6. This means that the measure of leverage derived from the hat matrix is, in effect, a measure of *potential* influence. It tells us how much difference the case would make in the final estimate if it were to have an unusual score on the dependent variable, but it does not tell us how much difference each case actually made in the final estimate.

Analysts involved in selecting influential cases will sometimes be interested in measures of potential influence, because such measures are relevant in selecting cases when there may be some *a priori* uncertainty about scores on the dependent variable. Much of the information in such case studies comes from a careful, in-depth measurement of the dependent variable – which may sometimes be unknown, or only approximately known, before the case study begins. The measure of leverage derived from the hat matrix is appropriate for such situations because it does not require actual scores for the dependent variable.

---

[21] This somewhat curious name derives from the fact that, if the hat matrix is multiplied with the vector containing values of the dependent variable, the result is the vector of fitted values for each case. Typically, the vector of fitted values for the dependent variable is distinguished from the actual vector of values on the dependent variable by the use of the "^" or "hat" symbol. Hence, the hat matrix, which produces the fitted values, can be said to put the hat on the dependent variable.

[22] The discussion here involves the use of the hat matrix in linear regression. Analysts may also be interested in situations that do not resemble linear regression problems, e.g., where the dependent variable is dichotomous or categorical. Sometimes, these situations can be accommodated within the framework of generalized linear models, which includes its own generalization of the hat matrix (McCullagh and Nelder 1989: **xxxx**).

A second commonly-discussed measure of influence in statistics is Cook's distance. This statistic is a measure of the extent to which the estimates of the β_i parameters would change if a given case were omitted from the analysis. Because regression analysis typically includes more than one β_i parameter, a measure of influence requires some method of combining the differences in each parameter to produce an overall measure of a case's influence. The Cook's distance statistic resolves this dilemma by taking a weighted sum of the squared parameter differences associated with deleting a specific case. Specifically, the formula for Cook's distance is:

$$\frac{(b_{-i} - b)^T \mathbf{X}^T \mathbf{X}(b_{-i} - b)}{(K+1)MSE} \tag{7}$$

In this formula, *b* represents all of the parameter estimates from the regression using the whole set of cases. b_{-i} represents the parameter estimates from the regression that excludes the *i*th case. **X**, as above, represents the matrix of independent variables. *K* is the total number of independent variables (not including the constant, which is allowed for in the formula by the use of $K + 1$). Finally, *MSE* stands for the mean squared error, which is a measure of the amount of variation in the dependent variable not linearly associated with the independent variables.[23]

This somewhat intimidating mathematical notation gives precise expression to the intuitive idea, discussed above, of measuring influence as a weighted sum of the differences that result in each parameter estimate when a single case is deleted from the data set. One disadvantage to this formula, however, is that it requires a number of extra regressions to be run in order to compute measures of influence for each case. The overall regression must of course be computed–but then, an additional regression, with one case deleted, is required for each case.

Fortunately, matrix-algebraic manipulation demonstrates that the expression for Cook's Distance given in Equation 7 is equivalent to the following, computationally much easier expression:

$$\frac{r_i^2 \mathbf{H}_{i,i}}{(K+1)(1 - \mathbf{H}_{i,i})} \tag{8}$$

In this expression, $\mathbf{H}_{i,i}$ refers to the measure of leverage for the *i*th case, taken from the diagonal of the hat matrix, as discussed above. *K* once again represents the number of independent variables. Finally, r_i^2 is a special, modified version of the *i*th case's regression residual, known as the Studentized residual, which needs to be separately computed.

The Studentized residual is designed so that the residuals for all cases will have the same variance. If the standard regression residual for case *i* is denoted by ε_i, then the Studentized residual, r_i^2 can be computed as follows. (All symbols in this expression are as previously defined.)

$$r_i = \frac{\varepsilon_i}{\sqrt{MSE(1 - \mathbf{H}_{i,i})}} \tag{9}$$

As can be seen from an inspection of Equations 8 and 9, Cook's distance for a case depends primarily on two quantities: the size of the regression residual for that case and the leverage for that case. The most influential cases are those with substantial leverage that lie significantly off of the regression line.

[23] Specifically, the MSE is found by summing the squared residuals from the full regression and then dividing by $N - K - 1$, where *N* is the number of cases and *K* is the number of independent variables.

Cook's distance for a given case provides a summary of the overall difference that the decision to include that case makes for the parameter estimates. Cases with a large Cook's distance contribute quite a lot to the inferences drawn from the analysis. In this sense, such cases are vital for maintaining analytic conclusions. Discovering a significant measurement error on the dependent variable or an important omitted variable for such a case may dramatically revise estimates of the overall relationships. Hence, it may be reasonable to select influential cases for in-depth study.

To conclude, three statistical concepts have been introduced in this section. The hat matrix provides a measure of leverage, or potential influence. Based solely on each case's scores on the independent variables, the hat matrix tells us how much a change in (or a measurement error on) the dependent variable for that case would affect the overall regression line.

Cook's distance goes further, considering scores on both the independent and the dependent variables in order to actually tell us how much the overall regression estimates would be affected if each case were to be dropped completely from the analysis. This produces a measure of how much actual—and not potential—influence each case has on the overall regression.

Either the hat matrix or Cook's distance may serve as an acceptable measure of influence for selecting case studies, although the differences just discussed must be kept in mind. In the examples below, Cook's distance will be used as the primary measure of influence because our interest is in whether any particular cases might be influencing the coefficient estimates in our democracy-and-development regression.

A third concept, the Studentized residual, was introduced as a necessary element in computing Cook's distance. (The hat matrix is, of course, also a necessary ingredient in Cook's distance.)

Example. Figure 5 shows the Cook's distance scores for each of the countries in the 1995 per capita GDP and democracy data set. Most countries have quite low Cook's distances. The three most serious exceptions to this generalization are the numbered lines in the figure: Jamaica (74), Japan (75), and Nepal (105). Of these three, Nepal is clearly the most influential by a wide margin. Hence, any case study of influential cases with respect to the relationship modeled in Equation 4 would have to start with an in-depth consideration of Nepal.

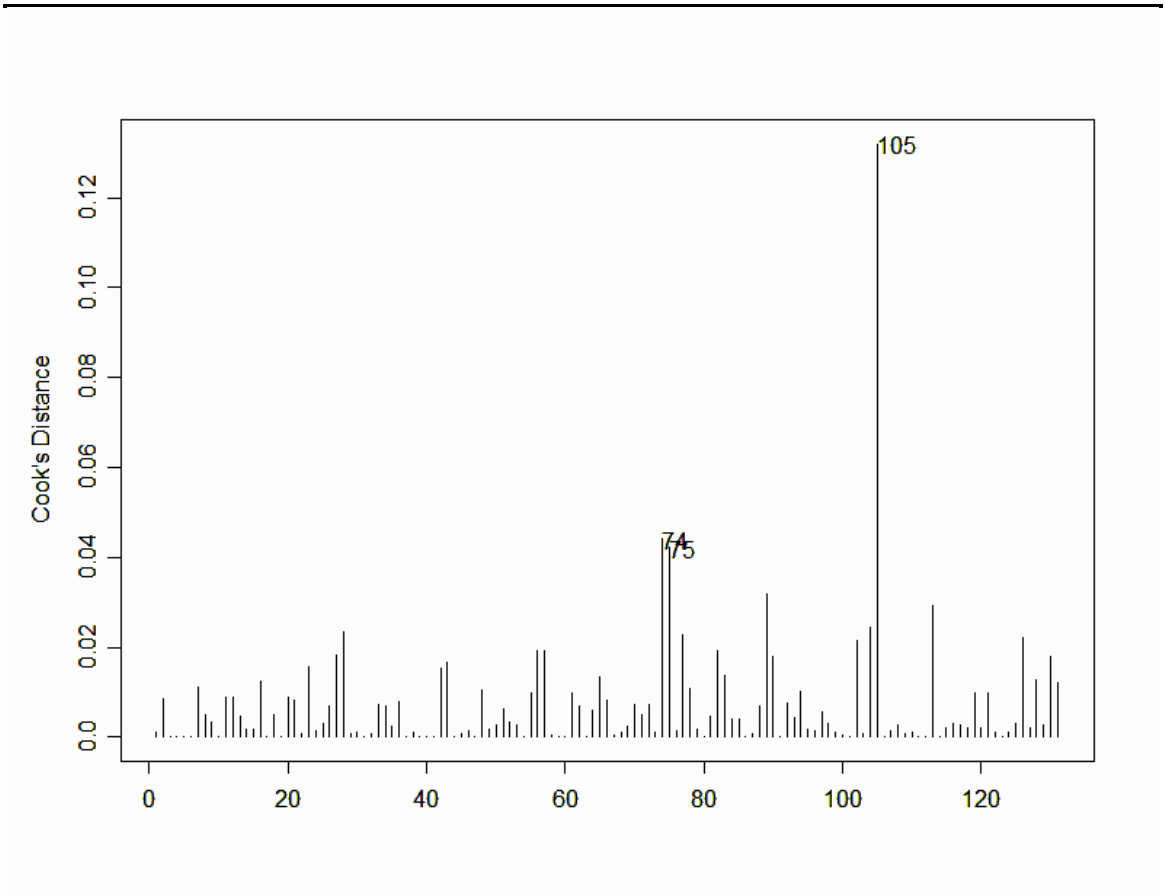

Figure 5. The Cook's distance scores for an OLS regression of democracy on logged per capita GDP. The three numbered cases have high Cook's distance scores.

Conclusions. The use of an influential-case strategy of case selection is limited to instances in which a researcher has reason to be concerned that her results are being driven by one or a few cases. This is most likely to be true in small-to-moderate sized samples. If Ertman had been interested in explaining statebuilding on a global scale he would have had access to well over a hundred country-cases (depending upon the historical period). In this context, the possibly influential status of a single case may be conveniently ignored – assuming, that is, that the larger covariational pattern is clear enough.

Where N is very large -- greater than 1000, let us say – it is extremely unlikely that a small set of cases (much less an individual case) will play an "influential" role. Of course, there may be influential *sets* of cases, e.g., countries within a particular continent or cultural region, or persons of Irish extraction. Sets of influential observations are often problematic in a time-series cross-section dataset where each unit (e.g., country) contains multiple observations (through time), and hence may have a strong influence on aggregate results. Still, the general rule is: the larger the sample, the less important individual cases are likely to be and, hence, the less likely a researcher is to use hat matrix and Cook's distance statistics for purposes of case selection. In these instances, it may not matter very much what values individual cases display. (It may of course matter for the purpose of investigating causal mechanisms; however, for this purpose one would not employ influential statistics to choose cases.)

MOST-SIMILAR CASES

The most-similar method, like the diverse-case method, employs a minimum of two cases.[24] In its purest form, the chosen pair of cases is similar on all the measured independent variables *except* the independent variable of interest. Table 3 offers a stylized example of the simplest sort of most-similar analysis, with only two cases and with all variables measured dichotomously. Here, the two cases are similar in various respects that might be relevant to the outcome of interest, as signified by X_2. This is the vector of control variables, and they are constant across the cases. The cases differ, however, on one causal variable – X_1 – and on the outcome. Thus, it is presumed that the presence or absence of this particular factor causes variation on the outcome.

[24] Lijphart (1971, 1975), Meckstroth (1975), Przeworski and Teune (1970), Skocpol and Somers (1980). Sometimes, the most-similar method is known as the "method of difference" (Mill 1843/1872).

Table 3:
Most-Similar Analysis with Two Cases

Cases	*Variables*		
	X_1	X_2	Y
#1	+	+	+
#2	–	+	–

+/–: the score demonstrated by a case on a particular dimension (variable), coded dichotomously. X_1: the variable of theoretical interest. X_2: the background/control variable or vector. Y: the outcome. Note that in this example all variables are positively correlated with the outcome.

Leon Epstein's study of party cohesion in legislatures focuses on two most-similar countries, the United States and Canada. Canada has highly disciplined parties whose members vote together on the floor of the House of Commons while the US has weak, undisciplined parties, whose members often defect on floor votes in Congress. In explaining these divergent outcomes, persistent over many years, Epstein first discusses possible causal factors that are held more or less constant by the research design. Both countries inherited an English political culture, both have large territories and heterogeneous populations, both are federal, and both have fairly loose party structures with strong regional bases and a weak center. These are the control variables. Epstein highlights one difference: Canada is parliamentary while the US is presidential. If this is the only relevant difference between the US and Canada, then it is reasonable to conclude, with Epstein, that it is the cause of the party system difference.[25] The concern that other relevant differences may also exist inclines the researcher to bolster the most-similar analysis with a large-N cross-case study or within-case analysis.[26]

Large-N Analysis. Having outlined the most-similar research design, we turn to the question of how to identify such cases from within a large-N cross-case dataset. For heuristic purposes, we shall focus on two-case comparisons. Readers should be aware that this can, and often should, be adapted to more complex comparisons.

The most useful statistical tool for identifying cases for in-depth analysis in a most-similar setting is probably some variety of "matching" strategy.[27] Statistical estimates of causal effects based on matching techniques have been a major topic in quantitative methodology over the last twenty-five years, first in statistics (Rosenbaum and Rubin 1983; Rosenbaum 2002) and subsequently in econometrics (Hahn 1998; Hirano, Imbens, and Ridder 2003) and political science (Ho, Imai, King and Stewart 2004; Imai forthcoming). Matching techniques are based on an extension of experimental logic. In a randomized experiment, elaborate statistical models are unnecessary for causal inference — because, for a large enough selection of cases, the treatment group and the control group have a very high probability of being quite similar, on both measured and unmeasured variables (other than the independent variable and its effects). Hence, quite simple statistical treatments (e.g., a difference of means test) may be sufficient to demonstrate a causal inference.

In observational studies, it is quite unusual to find situations in which the cases with a high score on the independent variable (which roughly correspond to the treatment group in an experiment) are similar on all measured and unmeasured variables—once again, other than the independent variable and its effects—to the cases with a lower score on the independent variable (corresponding to the control group). Typically, the treatment group in an observational study will differ in many ways from the control group. For example, in studies that seek to estimate the effects of democracy on wages to labor (e.g., Rodrik 1998), it is a problem that democracies are, on average, wealthier than other countries. After all, wealth may condition different approaches to economic distribution, thus introducing a confounding variable into the analysis.

One common approach to this problem is to introduce a variable for each potential confounder (e.g., wealth) in a general analysis of causal relationships (e.g., a regression model). Matching techniques have been developed as an explicit alternative to this control-variable approach. This alternative approach begins by identifying a set of variables (other than the dependent variable or the main independent variable) on which the cases are to be matched. Then, for each case in the treatment group, the researcher identifies as many cases as possible from the control group with the

---

[25] For further examples of the most-similar method see Brenner (1976), Hamilton (1977), Lipset (1968), Moulder (1977), Posner (2004).

[26] Of course, the previous methods discussed above are also subject to these difficulties. However, because they rest – implicitly or explicitly – on cross-case variance, and hence incorporate a larger set of cases, they are less liable to problems stemming from insufficient variation and probabilistic causation.

[27] For good introductions see Ho et al. (2004), Morgan and Harding (2005), Rosenbaum (2004), Rosenbaum and Silber (2001). For a discussion of matching procedures in Stata see Abadie et al. (2001).

exact same scores on the matching variables (the covariates). Finally, the scholar looks at the difference on the dependent variable between the cases in the treatment group and the matching cases in the control group. If the set of matching variables is broad enough to include all confounders, the average difference between the treatment group and the matching control cases should provide a good estimate of the causal effect. Even in a situation in which the set of matching variables includes some, but not all, confounders, matching may produce better causal inferences than regression and related models because cases that match on a set of explicitly selected variables may also be more likely to be similar on unmeasured confounders.[28]

Unfortunately, in many observational studies the matching procedure described above – known as exact matching -- is impossible. This procedure typically fails for continuous variables such as wealth, age, or distance, since there may be no two cases with the same score on a continuous variable. For example, there is no undemocratic country with the exact same per capita GDP as the United States. Moreover, the larger the number of matching variables employed (either dichotomous or continuous), the lower the likelihood of finding exact matches.

In situations where exact matching is infeasible, researchers may instead employ approximate matching, in which cases from the control group that are "close enough" to matching cases from the treatment group are accepted as matches. Major weaknesses of this approach include the fact that the definition of "close enough" is inevitably arbitrary, as well as the fact that, for large enough sets of matching variables, few if any treatment cases will have even approximate matches.

To better deal with situations in which exact matching is impossible, methodologists have offered an alternative procedure, known as "propensity score matching." This approach suggests a different definition of similarity than the previous two. Rather than focusing on sharing scores on the matching variables, propensity-score matching focuses on sharing a similar estimated probability of having been in the treatment group, conditional on the matching variables. In other words, when looking for a match for a specific case in the treatment group, researchers look for cases in the control group that—before the score on the independent variable was known—would have been as likely to be in the treatment group as the other case. This is accomplished by a two-stage analysis, the first stage of which approaches the key independent variable, X_1 (understood as the "treatment"), as a *dependent* variable and the *matching* variables as *independent* variables. Once this model has been estimated, usually using nonparametric regression techniques that replace the assumption of linearity with a looser assumption of smoothness, any resulting coefficient estimates are disregarded. Instead, the second stage of the analysis employs the fitted values for each case, which tell us the probability of that case being assigned to the treatment group, conditional on its scores on the matching variables. These fitted values are referred to as propensity scores. The final step in the process is to choose matches for each case in the treatment group. This is accomplished by selecting cases from the control group with similar propensity scores. The end result of this procedure is a set of matched-cases that can be compared in whatever way the researcher deems appropriate. These are the "most-similar" cases, returning to the qualitative terminology. Rosenbaum and Silber (2001: 223) summarize:

> Unlike model-based adjustments, where patients vanish and are replaced by the coefficients of a model, in matching, ostensibly comparable patterns are compared directly, one by one. Modern matching methods involve statistical modeling and combinatorial algorithms, but the end result is a collection of pairs or sets of people who look comparable, at least on average. In matching, people retain their integrity as people, so they can be examined and their stories can be told individually.

Matching, conclude the authors, "facilitates, rather than inhibits, thick description" (Ibid.). The same matching techniques that have been used successfully in observational studies of medical treatments

[28] However, matching is clearly inferior to a well-designed and well-executed randomized experiment. The benefits of matching extend only so far as equivalence on the variables explicitly included and any unmeasured variables that fortuitously happen to be similar across the cases. By contrast, proper randomization handles all unmeasured variables.

could also be adopted to the study of nation-states, political parties, cities, or indeed any traditional paired cases in the social sciences.

Example. Suppose that, in order to study the relationship between wealth and democracy, the researcher wishes to select cases that are as similar as possible to India and Costa Rica in background variables—while being as different as possible on per capita GDP.

In order to select most-similar cases for the study of the relationship between wealth and democracy, we will need a statistical model of the causes of a country's wealth. Obviously, such a proposition is complex; since this is simply an illustrative example, we will be satisfied with a cartoon model that only includes two independent variables. Specifically, a country's wealth will be assumed to be a function of the origin of its legal system (including the British legal heritage variable discussed in the previous example, as well as dummy variables for French legal heritage, socialist legal heritage, German legal heritage, and Scandinavian legal heritage), and a variable measuring the latitude of the country's capital.

The first step in selecting most-similar cases is to run a nonparametric regression with these two variables as the independent variables and logged per capita GDP (the independent variable of theoretical interest) as the dependent variable. The fitted values from this regression serve as propensity scores, and cases with similar propensity scores are interpreted as matching. It is important to keep in mind that the quality of the match depends on the quality of the statistical model used to generate the propensity scores; a superficial model, like the one used here, obviously produces superficial matches.

The propensity scores for our two focus cases are: Costa Rica, 7.63, and India, 8.02. Examining the propensity score data, we see that Benin has a propensity score of 7.58—quite similar to Costa Rica's—and a per capita GDP of $1163, which is substantially different from Costa Rica's $5486. Hence, Benin and Costa Rica may be seen as most-similar cases for testing the relationship between wealth and democracy. Similarly, Singapore's propensity score of 7.99 is a very nearly exact match for India's, in spite of a noticeable difference between Singapore's per capita GDP of $27,020 and India's $2066. These two pairs of cases thus meet the criteria for most-similar-case comparison.

Conclusion. The most-similar method is one of the oldest recognized techniques of qualitative analysis, harking back to J.S. Mill's classic study, *System of Logic* (first published in 1834). By contrast, "matching" statistics are a relatively new technique in the arsenal of the social sciences, and have rarely been employed for the purpose of selecting cases for in-depth analysis. Yet, we believe that there may be a fruitful interchange between the two approaches. Indeed, the current popularity of matching among statisticians – relative, that is, to garden-variety regression models – rests upon what qualitative researchers would recognize as a "case-based" approach to causal analysis. Hence, we think it perfectly reasonable to appropriate this large-N method of analysis for case study purposes.

MOST-DIFFERENT CASES

The most-different method of case selection is the reverse-image of the previous research design. Here, variation on independent variables is prized, while variation on the outcome is eschewed. Rather than looking for cases that are most-similar, one looks for cases that are most-different. Specifically, the researcher tries to identify cases where just one independent variable, as well as the dependent variable, co-vary, and all other plausible independent variables show different values. These are deemed "most different" cases, though they are similar in two essential respects -- the causal variable of interest (X_1) and the outcome (Y).[29] Analysts have long argued that this research design is a weaker tool for causal inference than the most-similar method (Gerring 2006).

[29] Traditionally, outcomes are understood as "positive" outcomes; things that happened. However, as we have argued above, they could just as well be composed of negative outcomes; things that did not occur.

COMPLICATIONS

The seven-part typology summarized in Table 1 is intended to provide a comprehensive menu of options for researchers seeking to identify useful cases for in-depth research, a means of implementing these options in large-N settings, and useful advice for how to maximize variation on key dimensions while maintaining claims to case-representativeness within a broader population. In this final section we summarize and extend our comments with respect to several complications that may arise in the course of implementing these procedures.

Some case studies follow only one strategy of case selection. They may be neatly classified as *typical*, *diverse, extreme*, *deviant*, *influential*, *most-similar*, or *most-different*, as discussed. However, it is important to recognize that many case studies also mix and match among these case selection strategies. We have noted that insofar as all case studies seek representative samples, they are always in search of "typical" cases. Thus, it is common for writers to declare that their case is both extreme and typical; it has an extreme value on the X or Y of interest but is not, in other respects, idiosyncratic. There is not much that we can say about these combinations of strategies except that, where the cases allow for a variety of empirical strategies, there is no reason not to pursue them. And where the same cases can serve several functions at once (without further effort on the researcher's part), there is little cost to a multi-pronged approach to case analysis.

The second complication that deserves emphasis is the changing status of a case during the course of a researcher's investigation – which may last for years, if not decades. Often, a researcher begins in an exploratory mode and proceeds to a confirmatory mode -- that is, she develops a specific X/Y hypothesis. The goal of research, after all, is discovery, not simply the verification or falsification of existing hypotheses.

Unfortunately, research strategies that are ideal for exploration are not always ideal for confirmation. The extreme-case method, for example, is inherently exploratory; the researcher is concerned merely to explore variation on a single dimension (X or Y). Once a specific hypothesis is adopted, the researcher must shift to a different research design. This transformation of research designs is quite common to case study research. One cannot construct the perfect research design until a) one has a specific hypothesis and b) one is reasonably certain about what one is going to find out there in the empirical world. In short, the perfect case study research design is often apparent to the researcher only ex post facto.

There are three ways to handle this. One can explain, straightforwardly, that the initial research was undertaken in an exploratory fashion, and therefore not constructed to test the specific hypothesis that is – now – the primary argument. Alternatively, one can try to re-design the study after the new (or revised) hypothesis has been formulated. This may require additional field research or perhaps the integration of additional cases or variables that can be obtained through secondary sources or through consultation of experts. A final approach is to simply jettison, or deemphasize, the portion of research that no longer addresses the (revised) key hypothesis. A three-case study may become a two-case study, and so forth. The lost time and effort are the only costs of this down-sizing. In the event, practical considerations will probably determine which of these three strategies, or combinations of strategies, is to be followed. (They are not mutually exclusive.) The point to remember, and our point of conclusion in this article, is that revision of one's cross-case research design is *normal* and perhaps to be expected. Not all twists and turns on the meandering trail of truth can be reasonably anticipated. Where one starts is not always where one ends up.

Indeed, outcomes that are continuous in nature such as welfare spending or growth are not readily understandable in this positive/negative fashion. The most-different method is also sometimes referred to as the "method of agreement," following its inventor, J.S. Mill (1843/1872). See also DeFelice (1986), Gerring (2001: 212-4), Lijphart (1971, 1975), Meckstroth (1975), Przeworski and Teune (1970), Skocpol and Somers (1980).

A final complication, which we have noted in each section of the paper, is that of representativeness. There are only two situations in which a case study researcher need not be concerned with the representativeness of her chosen case. The first is the influential case research design, where a case is chosen because of its possible influence on a cross-case model, and hence is not expected to be representative of a larger sample. The second is the deviant case method, where the chosen case is employed to confirm a broader cross-case argument to which the case stands as an apparent exception. Here, the researcher's task may be simply to show that deviant-case A is poorly explained by theory B because of idiosyncratic feature C. In this instance the case is also not intended to represent a broader class of cases.

In all other circumstances, cases must be representative of the population of interest in whatever ways might be relevant to the proposition in question. Note that where a researcher is attempting to disconfirm a deterministic proposition the question of representativeness is perhaps more appropriately understood as a question of classification: is the chosen case appropriately classified as a member of the designated population? If so, then it is fodder for a disconfirming case study.

If the researcher is attempting to confirm a deterministic proposition, or to make probabilistic arguments about a causal relationship, then the problem of representativeness is of the more usual sort: is Case A unit-homogenous relative to other cases in the population? This is not an easy matter to test. However, in a large-N context the residual for that case (in whatever model the researcher has greatest confidence in) is a reasonable place to start. Of course, this test is only as good as the model at hand. Any incorrect specifications or incorrect modeling procedures will likely bias the results and give an incorrect assessment of each case's "typicality." In addition, there is the possibility of stochastic error, errors that cannot be modeled in a general framework. Given the explanatory weight that individual cases are asked to bear in a case study analysis, it is wise to consider more than just the residual test of representativeness. Deductive logic – expectations about the causal relationships of interest and the case of choice – are sometimes more useful than purely inductive exercises, whether qualitative or quantitative.

In any case, there is no dispensing with the question. Case studies (with the two exceptions already noted) rest upon an assumed synecdoche: the case should stand for a population. If this is not true, or if there is reason to doubt this assumption, then the utility of the case study is brought severely into question.

REFERENCES

Abadie, Alberto, David Drukker, Jane Leber Herr, Guido W. Imbens. 2001. "Implementing Matching Estimators for Average Treatment Effects in Stata." *The Stata Journal* 1:1-18.

Abadie, Alberto, Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* (March) 113-32.

Abbott, Andrew, Angela Tsay. 2000. "Sequence Analysis and Optimal Matching Methods in Sociology." *Sociological Methods and Research* 29, 3-33.

Achen, Christopher H., Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41 (January) 143-69.

Alesina, Alberto, Sule Ozler, Nouriel Roubini, Phillip Swagel. 1996. "Political Instability and Economic Growth." *Journal of Economic Growth* 1:2.

Amenta, Edwin. 1991. "Making the Most of a Case Study: Theories of the Welfare State and the American Experience." In Charles C. Ragin (ed), *Issues and Alternatives in Comparative Social Research* (Leiden: E.J. Brill) 172-94.

Barro, Robert J. 1999. "Determinants of Democracy." *Journal of Political Economy* 107:6.

Brady, Henry, David Collier (eds). 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards.* Roman and Littlefield.

Brenner, Robert. 1976. "Agrarian Class Structure and Economic Development in Pre-Industrial Europe." *Past and Present* 70 (February) 30-75.

Browne, Angela. 1987. *When Battered Women Kill.* New York: Free Press.

Cao Y, L. Qin, L. Zhang, J. Safrit, D.D. Ho. 1995. "Virologic and Immunologic Characterization of Long-term Survivors of HIV Type I Infection." *New England Journal of Medicine* 332, 201–8.

Cochran, William G. 1977. *Sampling Techniques.* John Wiley and Sons.

Cohen, Morris R. and Ernest Nagel. 1934. *An Introduction to Logic and Scientific Method.* New York: Harcourt, Brace and Company.

Collier, David, James Mahoney. 1996. "Insights and Pitfalls: Selection Bias in Qualitative Research." *World Politics* 49 (October) 56-91.

Collier, Ruth Berins, David Collier. 1991/2002. *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America.* Notre Dame, IN: University of Notre Dame Press.

Converse, Philip E., G. Dupeux. 1962. "Politicization of the Electorate in France and the United States." *Public Opinion Quarterly* 16 (Spring).

Coppedge, Michael J. 2002a. "Nested Inference: How to Combine the Benefits of Large-Sample Comparisons and Case Studies." Presented at the annual meetings of the American Political Science Association, Boston (August-September).

Coppedge, Michael J. 2004. "The Conditional Impact of the Economy on Democracy in Latin America." Ms.

DeFelice, E. Gene. 1986. "Causal Inference and Comparative Methods." *Comparative Political Studies* 19:3 (October) 415-37.

Deyo, Frederic (ed). 1987. *The Political Economy of the New Asian Industrialism.* Ithaca: Cornell University Press.

Dion, Douglas. 1998. "Evidence and Inference in the Comparative Case Study." *Comparative Politics* 30:2 (January) 127-45.

Drass, Kriss, Charles C. Ragin. 1992. *QCA: Qualitative Comparative Analysis.* Evanston: Institute for Policy Research, Northwestern University.

Eckstein, Harry. 1975. "Case Studies and Theory in Political Science." In Fred I. Greenstein and Nelson W. Polsby (eds), *Handbook of Political Science, vol. 7. Political Science: Scope and Theory* (Reading, MA: Addison-Wesley).

Elman, Colin. 2003. "Lessons from Lakatos." In Colin Elman and Mirium Fendius Elman (eds), *Progress in International Relations Theory: Appraising the Field* (Cambridge: MIT Press).

Elman, Colin. Forthcoming. "Theoretical Typologies in the Qualitative Study of International Politics." *International Organization*.
Emigh, Rebecca. 1997. "The Power of Negative Thinking: the Use of Negative Case Methodology in the Development of Sociological Theory." *Theory and Society* 26, 649-84.
Ertman, Thomas. 1997. *Birth of the Leviathan: Building States and Regimes in Medieval and Early Modern Europe*. Cambridge: Cambridge University Press.
Esping-Andersen, Gosta. 1990. *The Three Worlds of Welfare Capitalism*. Princeton: Princeton University Press.
Fenno, Richard F., Jr. 1978. *Home Style: House Members in their Districts*. Boston: Little, Brown.
Freedman, David A., Robert Pisani, Roger Purves, Ani Adhikari. 1991. *Statistics, 2d ed.*. New York: W.W. Norton.
Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." In James A. Stimson (ed), *Political Analysis, vol 2* (Ann Arbor: University of Michigan Press).
Geddes, Barbara. 2003. *Paradigms and Sand Castles: Theory Building and Research Design in Comparative Politics*. Ann Arbor: University of Michigan Press.
George, Alexander L., Andrew Bennett. 2004. *Case Studies and Theory Development*. Cambridge: MIT Press.
George, Alexander L., Richard Smoke. 1974. *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.
Goertz, Gary, Harvey Starr (eds). 2003. *Necessary Conditions: Theory, Methodology and Applications*. New York: Rowman and Littlefield.
Gerring, John. 2001. *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.
Gerring, John. 2004. "What is a Case Study and What is it Good For?" *American Political Science Review* 98:2 (May) 341-54.
Gerring, John. 2006. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.
Gerring, John, Philip Bond, William Barndt, Carola Moreno. Forthcoming. "Democracy and Growth: A Historical Perspective." Boston University, Department of Political Science.
Gerring, John, Rose McDermott. Forthcoming. "Experiments and Quasi-Experiments: Towards a Unified Framework of Research Design." Boston University, Department of Political Science.
Green, Donald P. and Ian Shapiro. 1994. *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven: Yale University Press.
Greene, William H. 2002. *Econometric Analysis*. Prentice Hall.
Hahn, J. 1998. "On the Role of the Propensity Score in Efficient Estimation of Average Treatment Effects." *Econometrica* 66 (March).
Hamilton, Gary G. 1977. "Chinese Consumption of Foreign Commodities: A Comparative Perspective." *American Sociological Review* 42:6 (December) 877-91.
Hersen, Michel and David H. Barlow. 1976. *Single-Case Experimental Designs: Strategies for Studying Behavior Change*. Oxford: Pergamon Press.
Hicks, Alexander. 1999. *Social Democracy and Welfare Capitalism: A Century of Income Security Politics*. Ithaca: Cornell University Press.
Hicks, Alexander, Toya Misra, Tang Hah Ng. 1995. "The Programmatic Emergence of the Social Security State." *American Sociological Review* 60 (June) 329-49.
Ho, Daniel E., Kosuke Imai, Gary King, Elizabeth A. Stuart. 2004. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Ms.
Howard, Marc Morjé. 2003. The Weakness of Civil Society in Post-Communist Europe. Cambridge: Cambridge University Press.
Houser, Daniel, John Freeman. 2001. "Economic Consequences of Political Approval Management in Comparative Perspective." Ms.

Hirano, Keisuke, Guido Imbens, Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:4, 1161-89.
Imai, Kosuke. Forthcoming. "Do Get-Out-the-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review*.
Janoski, Thomas, Alexander Hicks (eds). 1993. *Methodological Advances in Comparative Political Economy*. Cambridge: Cambridge University Press.
Kalyvas, Stathis N. 1996. *The Rise of Christian Democracy in Europe*. Ithaca: Cornell University Press.
Karl, Terry Lynn. 1997. *The Paradox of Plenty: Oil Booms and Petro-States*. Berkeley: University of California Press.
Kendall, Patricia L., Katherine M. Wolf. 1949/1955. "The Analysis of Deviant Cases in Communications Research." In Paul F. Lazarsfeld and Frank N. Stanton (eds), *Communications Research, 1948-1949* (New York: Harper and Brothers). Reprinted in Paul F. Lazarsfeld and Morris Rosenberg (eds), *The Language of Social Research* (New York: Free Press) 167-70.
King, Gary, Robert O. Keohane, Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
Lakatos, Imre. 1978. *The Methodology of Scientific Research Programmes*. Cambridge University Press.
La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, Robert W. Vishny. 1999. "The Quality of Government." *Journal of Economics, Law and Organization* 15:1, 222-79.
Lieberman, Evan S. 2005. "Nested Analysis as a Mixed-Method Strategy for Comparative Research," *American Political Science Review* (August?).
Lijphart, Arend. 1971. "Comparative Politics and the Comparative Method." *American Political Science Review* 65:3 (September).
Lijphart, Arend. 1975. "The Comparable Cases Strategy in Comparative Research." *Comparative Political Studies* 8, 158-77.
Lipset, Seymour Martin. 1959. "Some Social Requisites of Democracy: Economic Development and Political Development." *American Political Science Review* 53 (March) 69-105.
Lipset, Seymour Martin. 1968. *Agrarian Socialism: The Cooperative Commonwealth Federation in Saskatchewan. A Study in Political Sociology*. Garden City, NY: Doubleday and Co.
Lipset, Seymour Martin, Martin A. Trow, James S. Coleman. 1956. *Union Democracy: The Internal Politics of the International Typographical Union*. New York: Free Press.
Lynd, Robert Staughton, Helen Merrell Lynd. 1929/1956. *Middletown: A Study in American Culture*. New York: Harcourt, Brace.
MacIntyre, Andrew. 2003. *Power of Institutions: Political Architecture and Governance*. Ithaca: Cornell University Press.
Mahoney, James, Gary Goertz. 2004. "The Possibility Principle: Choosing Negative Cases in Comparative Research." *American Political Science Review* 98:4 (November) 653-69.
Marshall, Monty G., Keith Jaggers. 2000. "Polity IV Dataset Project: Political Regime Characteristics and Transitions, 1800-1999." (http://www.bsos.umd.edu/cidcm/polity).
McCullagh, Peter, J. A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall/CRC.
Meckstroth, Theodore. 1975. "'Most Different Systems' and 'Most Similar Systems': A Study in the Logic of Comparative Inquiry." *Comparative Political Studies* 8:2 (July) 133-177.
Mill, John Stuart. 1843/1872. *System of Logic, 8th ed.* London: Longmans, Green.
Monroe, Kristen Renwick. 1996. *The Heart of Altruism: Perceptions of a Common Humanity*. Princeton: Princeton University Press.
Moore, Barrington, Jr. 1966. *Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World*. Boston: Beacon Press.
Morgan, Kimberly. 2003. "The Politics of Mothers' Employment: France in Comparative Perspective." *World Politics* 55:2 (January) 259-89.
Morgan, Stephen L., David J. Harding. 2005. "Matching Estimators of Causal Effects: From Stratification and Weighting to Practical Data Analysis Routines." Ms.
Moulder, Frances V. 1977. *Japan, China and the Modern World Economy: Toward a Reinterpretation of East Asian Develoment ca. 1600 to ca. 1918*. Cambridge: Cambridge University Press.

Munck, Gerardo L. 2004. "Tools for Qualitative Research." In Henry E. Brady and David Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Lanham: Rowman & Littlefield) 105-21.
Norris, Pippa. 2004. *Electoral Engineering: Voting Rules and Political Behavior*. Cambridge: Cambridge University Press.
Patton, Michael Quinn. 2002. *Qualitative Evaluation and Research Methods*. Thousand Oaks, California: Sage Publications.
Posner, Daniel. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98:4 (November) 529-46.
Przeworski, Adam, Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: John Wiley.
Przeworski, Adam, Michael Alvarez, Jose Antonio Cheibub, Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Material Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.
Ragin, Charles C. 1987. *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. Berkeley: University of California.
Ragin, Charles C. 2000. *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.
Ragin, Charles C. 2004. "Turning the Tables." In Henry E. Brady and David Collier (eds), *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Lanham: Rowman & Littlefield) 123-38.
Reilly, Ben. 2000/2001. "Democracy, Ethnic Fragmentation, and Internal Conflict: Confused Theories, Faulty Data, and the 'Crucial Case' of Papua New Guinea." *International Security* 25:3, 162-85.
Rodrik, Dani. 1998. "Democracies Pay Higher Wages." Ms.
Rogowski, Ronald. 1995. "The Role of Theory and Anomaly in Social-Scientific Inference," *American Political Science Review* 89:2 (June) 461-474.
Rohlfing, Ingo. 2004. "Have You Chosen the Right Case?: Uncertainty in Case Selection for Single Case Studies." Working paper, International University, Bremen, Germany.
Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag.
Rosenbaum, Paul R. 2004. "Matching in Observational Studies." In A. Gelman and X-L. Meng (eds), *Applied Bayesian Modeling and Causal Inference from an Incomplete-Data Perspective* (John Wiley).
Rosenbaum, Paul R., Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 40-51.
Rosenbaum, Paul R., Jeffrey H. Silber. 2001. "Matching and Thick Description in an Observational Study of Mortality after Surgery." *Biostatistics* 2:2, 217-32.
Rousseeuw, Peter J., Annick M. Leroy. 2003. *Robust Regression and Outlier Detection*. Wiley-Interscience.
Scheaffer, Richard L., William Mendenhall, Lyman Ott. 1995. *Elementary Survey Sampling*. Duxbury.
Sekhon, Jasjeet S. 2004. "Quality Meets Quantity: Case Studies, Conditional Probability and Counterfactuals." *Perspectives in Politics* 2:2 (June) 281-93.
Skocpol, Theda. 1979. *States and Social Revolutions: A Comparative Analysis of France, Russia, and China*. Cambridge: Cambridge University Press.
Skocpol, Theda and Margaret Somers. 1980. "The Uses of Comparative History in Macrosocial Inquiry." *Comparative Studies in Society and History* 22:2 (April) 147-97.
Stone, Charles J. 1996. *A Course in Probability and Statistics*. Belmont: Duxbury Press.
Summers, Robert, Alan Heston. 1991. "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988." *Quarterly Journal of Economics* 106:2 (May) 327-68.
Swank, Duane H. 2002. *Global Capital, Political Institutions, and Policy Change in Developed Welfare States*. Cambridge: Cambridge University Press.
Treier, Shawn, Simon Jackman. 2003. "Democracy as a Latent Variable." Ms.
Yin, Robert K. 2004. *Case Study Anthology*. Thousand Oaks, California: Sage Publications.