Symposium: Perfecting Methodology, or Methodological Perfectionism?

Large-N Observational Data Analysis (aka Messy Data): A Modest Defense

John Gerring
Boston University
jgerring@bu.edu

A specter is haunting political science. It is the specter of methodological perfectionism. This dogma places methods before substance and imposes a narrow spectrum of acceptable methods on the discipline.

There are of course many varieties of methodological perfectionism, and political science has experienced more than a few over the past century. Methodologists from the “statistical” side of the discipline (aka PolMeth, quantitative methods) generally prize causal knowledge over descriptive knowledge, theory appraisal over theory discovery, micro-theory (aka micro-mechanisms) over macro-theory, and internal validity over external validity.

Methodologists from the “qualitative” (aka case study, QCA) side of the discipline share most of these preferences, with the notable exception of theory appraisal. (They value work that is exploratory rather than confirmatory.) I will not address qualitative work in the interpretive tradition, as this tradition—which certainly has its own version of perfection—is not as influential at the present time.

So defined, both the quant and qual side of the methodological divide agree on the nature of the problem we are facing today. Broadly stated, the most common method of drawing causal inferences in social science—based on large samples with no pretense of a randomized treatment (“regression”)—does not work. Consequently, we need to re-think the traditional approach to causality.

For those trained in statistics, the way forward is to be found in experimental or quasi-experimental evidence. Any imperfections in the assignment process or in the post-treatment period should be handled by appropriate statistical procedures—via matching, instrumental variables, regression-discontinuity models, and the like (Angrist and Pischke 2010; Gerber, Green, and Kaplan 2004; Imbens and Wooldridge 2009; Morgan and Winship 2007; Rubin 2005).

For those trained in qualitative methods, the way forward lies in case-based methods, e.g., (a) tracing a discrete process from the purported cause, through its various mechanisms, to an effect, (b) causal-process observations, (c) counterfactual thought-experiments, (d) well-matched cases that satisfy the strictures of a most-similar or most-different case design, or (e) comparative-historical work, which combines elements of the foregoing (Bennett 2010; Bennett and Elman 2006; Brady and Collier 2010; George and Bennett 2005; Gerring 2007; Mahoney 2000, 2010; Mahoney and Rueschemeyer 2003).

Many methodologists embrace both of these solutions (e.g., Freedman 2008; Seawright 2010), as do I. Indeed, qualitative evidence is often quite important for conducting a strong experimental or quasi-experimental design (Cook et al. 2010; Dunning 2008b; Paluck 2010; Rosenbaum 2010: 323–24). I do not want to portray these two solutions as necessarily in conflict with one another. Nonetheless, they are quite distinct approaches to causal inference, and I shall treat them as such.

I want to propose that both the experimental and case-based approaches to causal inference are valid, but—here are the crucial caveats—only if understood as ideals rather than uniform thresholds of adequacy, and only if understood within the larger context of methodological objectives and tools that have traditionally been applied to questions of social science (and new tools that are now entering the lexicon, such as randomization inference and extreme bounds analysis). The costs of adopting stricter methodological standards must be reckoned along with the benefits. Tradeoffs—e.g., between causal and descriptive knowledge, theory appraisal and theory discovery, micro-theory and macro-theory, internal and external validity—are inescapable.

Let us consider these costs in somewhat more detail. (This ground has been covered many times, but some mention of these issues is important in order to properly frame the main argument.) Although experiments usually achieve a high degree of internal validity, there is often a sacrifice in external validity or in the type of problems that can be addressed. Natural experiments are wonderful tools but are limited to circumstances of extraordinary serendipity, and are not always easy to generalize from (Dunning 2008a). The current raft of econometric tools are fairly easy to apply (given handy statistical software packages), but do not always rectify the problems they are designed to rectify. A prime example in point is the instrumental variables approach to causal inference, where one often finds a facile use of a technique without any real acknowledgment of its potential problems (e.g., violation of the exclusion restriction).
Likewise, case-based causal inference is easy to practice but hard to practice well. Confounders are generally legion, even in carefully matched cases (the most-similar form of comparison), and counterfactual thought-experiments are not always sufficient to eliminate them. Strong process tracing research usually depends on strong and specific theoretical predictions about the causal mechanisms at work. However, most social-science theories do not issue highly specific predictions about process and outcome, or they specify a number of possible mechanisms, none of which is necessary for $Y$ (causal equifinality). Likewise, strong process-tracing is usually possible only with proximal causal relationships; distal causes, which compose a large share of social science theories, are difficult to process-trace. The question, to adopt the vocabulary employed by Bennett (2010), is how often Hoop tests, Smoking Gun tests, or Doubly Decisive tests can be applied to case study evidence. Is this a realistic standard of evidence for most case studies? (Bennett [2010: 219] specifies, “The evidence must strongly discriminate between alternative hypotheses.”) Likewise, do case studies claiming to have applied these tests do so with integrity? Just as it is important to question dubious assertions about “natural experiments,” it is important to question assertions of slam-dunk process-tracing. And even when internal validity can be established, it is often difficult to generalize from a case study. 

What worries me is that, insofar as our discipline’s current experimental and case-based standards are taken seriously (i.e., interpreted strictly), a very high bar is being set for admission into the ranks of political science, a standard that only a small minority of work currently satisfies (and that much of my own work certainly does not). This sort of methodological perfectionism puts researchers in a situation where they may (a) pretend that they have attained methodological purity when they have not (in order to convince colleagues and reviewers), (b) privately feel morose about their work, and/or (c) abandon their substantive/theoretical goals in favor of things that might be studied in a way that satisfies current standards. None of these developments are especially healthy for the discipline—not to mention for the individuals involved.

Of course, one might argue that a surge of methodological extremism is precisely what is needed at the present time, as a tonic. If lots of what we currently do does not pass the more exacting standards that the discipline has adopted, it could be a sign that we need to buckle up and work harder, and learn some new tricks—not that we should relax our standards. And if current methods are not working very well because the nature of the data on certain subjects is recalcitrant, perhaps it makes sense to reallocate our energies towards problems that are more tractable. To this extent, methods may legitimately drive substance (i.e., in situations where little progress can be made on Subject A and lots of progress on Subject B).

The advent of experimental and quasi-experimental standards has had the salutary effect of raising methodological consciousness in the discipline (a good thing, from most perspectives) and has resulted in some superb work (a good thing, from all perspectives). And now that incentives of political scientists are strongly aligned to promote first-class research designs, the ingenuity of the discipline should be unleashed. At the end of a decade or so we should be able to answer the question that has dogged proponents of these methods all along: what is the possible purview of such methods (i.e., experiments, natural experiments, and strong case-based knowledge) for political science questions? And what is the value-added for the discipline of having these first-class studies? Do they cumulate into broader theoretical insights?

Perhaps the current evangelical spirit represents a temporary swing of the pendulum—an entirely appropriate reaction against the traditional, regression-based species of political science that became the hallmark of the behavioralist movement. In a few years the fever may pass and we will find some reasonable middle ground.

If this is what the future holds, we need to start thinking about what that middle ground might consist of. And if not, i.e., if the current methods juggernaut continues, we need to figure out what the end-result of our collective move toward higher quantitative and qualitative standards might be. Either way, our subject matter is highly consequential.

I must introduce one note of clarification before proceeding. The spirit of methodological perfectionism that I am describing is most noticeable among methodologists and among those who follow current developments in the methodological literature. Recent PhDs are likely to be more aware of these trends than older scholars. Those who attempt to publish in the very top mainstream journals (e.g., APSR, AJPS, JOP, BJPS, WP) are more likely to be aware of these trends than those who write for subfield journals or who publish books rather than articles. Research universities are more prone to these developments than liberal arts colleges.

For some scholars, this debate may seem like a debate from nowhere. However, insofar as the spirit of methodological perfectionism pervades the discipline’s top journals and top departments, it affects the direction of political science at large. Consequently, those who feel distant from these currents may still be affected by the tide.

**Messy Data**

Despite frequent espousals of pluralism, methodologists seem to agree that there is one category of endeavor that does not deserve forbearance. This excluded category may be described as large-N observational research where the treatment bears no resemblance to randomized assignment, where the assignment principle is not known, where there are no good instruments, and where there is no opportunity for convincing process-tracing.

This mouthful is sometimes subsumed under the rubric of “regression.” However, regression methods are often used for the analysis of experimental data and quasi-experimental data. Moreover, other (non-regression) estimators—e.g., difference of means tests, matching, randomization inference—can be used to analyze large-N samples of the sort described above.

In truth, the blighted patch of desert described by the passage in italics is a residual category. It includes any research design that doesn’t pass either the quantitative (experimental/quasi-experimental) or qualitative (case-based) thresh-
olds of adequacy, and whose purpose is causal inference (not merely description). I shall refer to this nebulous area henceforth as large-N observational data, aka messy data. Assume by this phrase that violations of stable unit treatment value assumption (SUTVA) and ignorability are serious concerns; confounders cannot be ruled out. Therefore, the internal validity of the study is seriously challenged. Perhaps most worrisome of all, the researcher has no way of determining whether the assumptions necessary for causal inference have been met unless a parallel experiment has been carried out. Thus, the only time we can be pretty sure that messy large-N observational research is accurate is when it is redundant.\textsuperscript{6}

In the face of such damning conclusions, it is easy to see why there has been a flight by methodologists from large-N observational data to other approaches. David Freedman (1997: 114; emphasis added) states boldly, “I see no cases in which regression equations, let alone the more complex methods, have succeeded as engines for discovering causal relationships.” While this may be viewed as somewhat extreme, Freedman articulates a widely-held skepticism toward causal analyses based on ex post statistical adjustments. Of late, regression has become the whipping-boy of methodologists, both quant and qual. It is perhaps the only area of agreement one finds today that stretches across the entire methodological spectrum—from positivism (so-called) to interpretivism (so-called).\textsuperscript{7} A sign of this new consensus is that it has become a sign of sophistication to scoff at “correlations” and to describe them as “descriptive” or as “stylized facts.”

Let me be clear. I have no argument with these critiques. However, I want to argue that Freedman’s conclusion is wrong. More important, I want to argue that it misframes the question we ought to be asking.

Let me now utter a few platitudes that I assume most readers will be willing to entertain, at least as a point of departure.

1. We learn about the world in myriad ways—including common sense, personal experience, secondary research (conducted by others), theoretical suppositions, deductive logic, exploratory data analysis, and (last and perhaps least) formally devised research designs. Of these, some conform to the current template of acceptable research, while most do not.

2. Different problems demand different approaches, and political science encompasses an extraordinarily broad set of theoretical frameworks and empirical data.

3. Where others have been to some extent determines where it might be fruitful to go. A good research design is understandable only in the context of a particular research tradition, where triangulation on a common problem is often useful.

If readers are willing to accept these truisms, it follows that no single methodological standard is likely to be applicable to all political science work (unless that standard is extremely abstract, i.e., on the level of philosophy of science). It also suggests a more flexible methodological standard for work in political science, one that might include large-N observational data analysis.

Note that many practicing political scientists continue to employ messy observational data as their empirical workhorse. Thus, my argument may be understood as a qualified defense of the status quo—qualified, in the following manner.

I want to argue that the question we ought to be asking is not whether method A or B is adequate—but rather (a) whether it adds to our knowledge of a subject, (b) whether it is the best method (or one of several equally good methods) for the job, and (c) whether an accurate assessment of overall uncertainty (not simply statistical uncertainty) is attached to the conclusions. If these criteria are met, then the study ought to be considered methodologically adequate, even if far from ideal. It follows that a study may be extremely shaky but still adequate, so long as it allows us to update our priors, beats the alternatives, and presents a plausible uncertainty estimate. The slogan is best-possible—rather than best (Gerring 2011a).

The foregoing statement pertains to methodology considered in its narrowest sense, i.e., pertaining to internal validity and precision. I would also argue that we need to find a way to incorporate other goals into our understanding of methodology. This includes, most importantly, the theoretical contribution of a study (its breadth of application and commensurability with other work). I don’t imagine that there will be much dissent from this argument; I raise it only so that readers can keep it in view, and because it sometimes militates toward large-N observational data and away from some of the more rarified methods, which are often limited in external validity and/or theoretical fecundity.

Messy Data at Work: Democracy as a Dependent Variable

The way to prove my point—and to prove that Freedman is wrong—is to not engage the question at the abstract level of philosophy or methodological theory, where complex issues are usually difficult to resolve. Rather, following Freedman (who was fond of dissecting published work), I must show that for some questions of importance to political science, large-N observational data provides the best (or a best) approach to the problem, there is some value-added to our understanding of a subject, and reasonable estimates of uncertainty can be arrived at (through statistical procedures such as Bayesian inference, randomization inference, or extreme bounds and/or through qualitative reasoning).

I shall begin with a research area that provides a tough test for my argument: crossnational regressions with institutions on the left and right side of the model. It is a tough test because such analyses are characterized by many of the features that lead methodologists to despair of ever reaching causal inferences based on observational data. Institutions are broad, abstract phenomena—democracy, development, rule of law, property rights, veto points—that are difficult to conceptualize, much less to measure. Even where they can be measured, they are difficult to interpret causally because they are usually not manipulable (perhaps not even in principle, although this may be debated). Institutions tend to be slow-moving and thus provide little change over observable periods, usually
limited to the past half-century. Institutions are highly correlated—good things go together, as do bad things—so it is difficult to tease apart the signal from the background noise and from potential confounders. One rarely finds quasi-experimental assignment to treatment; endogeneity is the norm. Good instruments (ones that are strongly correlated with the treatment and do not affect the outcome) are rare. And the units of interest—nation-states—are remarkably heterogeneous. (Some might even wonder if they belong in the same sample.)

Recall, however, the burden of the argument: not that good inferences result from such bad data but that better inferences may result from large-N observational data analysis than from other types of analyses. Indeed, many of the characteristics of the typical crossnational regression also pose problems for experimental or quasi-experimental analysis or for convincing case-based analysis (Coppedge forthcoming: ch 5; Gerring and Thacker 2008: ch 7; Seawright 2010). It is not clear that there is a good alternative.

Now, let us turn to a specific example. I shall focus on an area well known to most readers (and certainly to all comparativists) by virtue of the volume of work that has been devoted to it and the prominence of the theory—modernization. This refers to the relationship between development and democracy, a topic first tackled in a serious way by Seymour Martin Lipset (1959), and since explored by myriad studies, generally in a regression framework (e.g., Boix and Stokes 2003; Casper and Tufis 2003; Epstein et al. 2006; Przeworski and Limongi 1997).

Two findings have emerged from this body of work. The first is that development (as measured by GDP per capita) helps democracies consolidate. The richer a democracy, the less likely it is to relapse into authoritarian rule. The second is that development has a very small—perhaps even nonexistent—effect on democratization. Rich countries are only slightly more likely to democratize than similarly situated poor countries. This second finding is still contested, but the boundaries of the possible relationship seem clear. Development has a small or null effect on democratization.

Our questions of interest are whether regression (i.e., large-N observational data analysis) has added to our knowledge of the subject, whether it is the best method (or one of several equally good methods) for the job, and whether an accurate assessment of overall uncertainty has been attached to the conclusions of the cited studies. I think the answers are: yes, yes, and probably not. Thus, in two respects messy-data methods proved their worth, though perhaps not in the third.

However, the one failing is by no means irredeemable (see Glynn, below). Indeed, it is partly a product of the methodological perfectionism that I am complaining about. Authors feel compelled to present their findings as if they arose from experimental or quasi-experimental evidence, even though they probably know (in their heart of hearts) that their t statistics do not encompass many threats to inference situated in the research design. If I am correct in these conclusions, then a modest case for large-N observational data has been established, at least in one instance.

Michael Coppedge’s magisterial survey, Approaching

Democracy: Research Methods in Comparative Politics (forthcoming), examines what we know about democratization, and how we know it. Here, a much broader survey of the methodological and substantive ground is covered. Because it bears directly on our question, I will paraphrase at some length.

Coppedge does not discuss any experimental or quasi-experimental work, suggesting that this method has yet to be applied (or has not been successfully applied) to the topic. Case study work on democratization is voluminous, and Coppedge is impressed by the rich narrative histories of specific cases as well as the number of potentially fruitful general hypotheses that might be garnered from this corpus. However, he also notes that there are many threats to inference, even with respect to the cases under study (problems of internal validity). Moreover, the broader hypotheses that might be gleaned from the case study literature have been difficult to generalize from because they are developed in the context of specific cases and are not always couched in ways that would lend themselves to broader application. Coppedge concludes that theory generation, not theory testing, is the province of case studies in the area of democratization. “Histories and case studies are great ways to develop ideas about things that may matter generally, but cannot show that they do matter generally” (Coppedge forthcoming: 18–19, chap. 5).

For theory testing, Coppedge concludes that crossnational statistical analysis is required—even though it may not always be sufficient, for all the reasons we have discussed (and which Coppedge discusses in much greater detail). Likewise, Coppedge criticizes crossnational analyses for their lack of theoretical integration: researchers settle on slightly different operationalizations, samples, and/or estimators and, as a result, their findings do not cumulate. Nonetheless, he hazardsthe following conclusions, based on his reading of the literature (and his own analyses). Factors that seem to have little impact on any democracy outcome include “land area, population, age of the country, the rule of law (as currently measured), colonial rule (without differentiating among colonial powers), and linguistic fragmentation.” Factors that have some impact on at least one outcome measure of democracy, and have been confirmed by multiple studies and extensive robustness tests, are summarized as follows:

Income (the log of per capita GDP) is associated with higher cross-national levels of democracy; income and economic growth are both associated with a higher probability of survival as a democracy and a lower probability of transition. Greater absolute changes in level and a higher probability of breakdown are found in the more unequal societies. Rentier states tend to be less democratic. And religiously fragmented societies are less stable: more likely to experience both transitions and breakdowns. (chap. 9: 56–57, draft version)

In addition, Coppedge identifies a number of factors whose effects are robust, though difficult to interpret. This includes “the core vs. periphery distinction, the proportion of democratic neighbors, the distinction between capitalist and communist economies, the number of past regime transitions, a
it is difficult to do so in a definitive fashion (because proving a
thesis is often just as useful as proving a hypothesis, even though
I believe the thesis. Indeed, failing to disprove a null hypoth-
esis warrants by extant research—as summarized by Coppedge’s
painstaking and comprehensive review—it is still something
of an achievement, given the difficulties of the subject matter.
More to the point, it is a lot more than we have learned from
other approaches. To this extent, and with all the usual cave-
ats, messy data analysis is vindicated.

**Messy Data at Work:**

**Democracy as an Independent Variable**

Let us now consider a slightly different question, where
democracy lies on the right side of a causal model. Seawright
(2010) claims that regression has made no contribution to the
question of democracy’s relationship to growth. He says (echo-
ing the common wisdom) that no hypothesis is robust to all
plausible estimators and specifications and all “positive” (sta-
tistically significant) findings are subject to potential confound-
ers.

My own view is that democracy shows a strongly posi-
tive relationship to growth if measured in a non-dichotomous
and historical fashion (Gerring, Bond, Barndt, and Moreno
2005; Persson and Tabellini 2009). However, let us lay that
argument aside, as it is not addressed by Seawright’s other-
wise admirably comprehensive review. I agree with Seawright’s
and others’ conclusions, so far as they go. As convention-
ally operationalized, there is no robust and plausible relation-
ship between democracy and growth. Is this not useful knowl-
edge? Have we not updated our priors? To put the point some-
what differently, do we know anything more about the theo-
retical question of interest after having looked at the cross-
national data? I submit that we have learned quite a lot.

I want to bang on this drum a little while longer. Suppose
a policymaker is interested to know whether there might be a
relationship between democracy and growth. He comes to you,
the resident comparativist, for recommendations of studies
that you should take a look at. Seawright’s advice would seem
to be the following: “Read only case studies, avoid all large-N
crossnational studies, and hope that someone, someday, fig-
ures out a way to study this in an experimental or quasi-exper-
imental fashion.” I doubt that this is sage advice.

Note that if there were a reasonably strong (and therefore
practically and theoretically relevant) causal relationship be-
tween democracy and growth one would expect it to appear in
crossnational empirical tests and to be at least somewhat stable
across various (plausible) robustness tests. The fact that it
does not (when democracy is operationalized in the conven-
tional fashion) is informative. One is much less inclined to
believe the thesis. Indeed, failing to disprove a null hypoth-
esis is often just as useful as proving a hypothesis, even though
it is difficult to do so in a definitive fashion (because proving a
null hypothesis means, in effect, disproving any possible rela-
tionship between X and Y, an argument that must take into
account all possible forms that a relationship might take).

Now, let us approach the question from another angle.
The critique of large-N observational data would be trenchant
if a viable alternative were available. However, Seawright
equivocates on this point. He suggests scaling down our theo-
retical ambitions to examine causal mechanisms—factors pre-
dicted to lie in between democracy and growth. Yet, the only
empirical example of this sort of work that is cited (Baum and
Lake 2003) is also a regression analysis—a fact that Seawright
notes, disapprovingly. Moreover, Seawright lays out a number
of methodological difficulties that such mechanistic studies
are likely to face. One is left to wonder whether or not there is
a viable alternative to the crossnational regression in this par-
ticular instance.

It so happens that I have participated in a case-study
endeavor to show causal mechanisms lying within the democ-
acy (stock) and growth relationship (Gerring, Kingstone,
Lange, and Sinha 2011). We found it a useful exercise. But it
was certainly not without its methodological difficulties, and it
certainly did not meet the test of inferential validity that case
study researchers aspire to. So, again, I found that an available
method added to the sum total of human knowledge on our
subject but lay very far from the methodological standards
currently being advocated.

**Robustness Tests**

Critics of messy observational data point out that results
are generally unstable when slight changes are made in the
measurement of key variables, the specification of a model, or
the chosen estimator (including corrections for autocorrelation
and the like). Robustness tests show few robust results, and
virtually no stable results (where the coefficient on a key vari-
able of interest remains stable). Under the circumstances,
it seems clear that coefficients and standard errors are not to be
taken literally. This is especially true for the cursed format of
the crossnational regression, for all the reasons we have dis-
cussed (Kittel 2006; Rodrik 2005; Seawright 2010; Summers
1991; Treisman 2007). Because we don’t know which (if any)
operationalization, specification, and estimator correctly mod-
els the data generation process (DGP), we are at sea.

Critics are right to be skeptical of studies that show only
one empirical test for an argument. Appeals to “theory” are
generally not very convincing. (Note that if the theory is strong
there is little point in testing; we already know what’s out
there. If the theory is weak, we are not strengthening our faith
in assumptions by appealing to it.) Usually, there are a variety
of plausible ways to model the DGP. However, the researcher
typically only shows a subset of these possibilities (one sus-
psects that alternative models have been suppressed, by virtue
of their non-corroborating results). This is a serious problem.
Consequently, consumers are not in a good position to judge
the veracity of an argument based on messy observational
data, unless they have played with the data themselves.

Yet, this final clause suggests something important. Im-
partial examinations of the data generally reveal that some rela-
tionships are more robust than others. Of course, this could be the product of persistent X/Y endogeneity, unmeasured confounders, a biased sample, and so forth. Robustness tests will never offer the explicit demonstration of causality that is provided by experimental or superb case-based analysis. And they will never provide a precise estimate of X’s impact on Y. The purpose, rather, is to test whether a very generally stated hypothesis—conceptualized vaguely as “positive” or “negative”—is likely to be true or not.

With this modest objective in mind robustness tests are more than just window dressing. Note that although few results withstand all possible and plausible robustness tests, some are more robust than others. These results deserve to be taken very seriously—provided they are plausible (knowing what we know about the world). As an example, one might return to the relationship between development and democratic consolidation. Here is a result that seems unlikely to go away, no matter how much researchers torture the data. Likewise, weaker results—those that are fragile in the face of robustness tests, such as the relationship between development and democratization—deserve to be treated with greater skepticism. This does not mean they should be dismissed; it means simply that the estimate must be surrounded by very large confidence intervals.

This is more—much more—than nothing. Note that all causal inference is based on assumptions. (Even experiments rest on assumptions, though they are much fewer and usually less problematic.) In the words of Donald Rubin,

Causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical. There is always a trade-off between assumptions and data—both bring information. With better data, fewer assumptions are needed. But in the causal inference setting, assumptions are always needed, and it is imperative that they be explicated and justified. One reason for providing this detail is so that readers can understand the basis of conclusions. A related reason is that such understanding should lead to scrutiny of the assumptions, investigation of them, and, ideally, improvements. Sadly, this stating of assumptions is typically absent in many analyses purporting to be causal and replaced by a statement of what computer programs were run.\(^8\)

Robustness tests are tests of assumptions, usually understood by reference to a benchmark model (which the researcher considers to represent the most plausible rendering of the DGP). The purpose of each test is to verify the main finding under slightly different assumptions. If the finding holds, it is considered robust.\(^9\)

Conclusions

This short opinion-piece has discussed only one type of large-N observational data inference, where countries serve as units of analysis. Evidently, I have not offered anything like a comprehensive review of this gargantuan subject. Yet, if messy data offers a viable strategy in crossnational analysis, which might be considered the worst-case scenario for causal inference, it ought to be viable in other settings.

Thus, I submit that large-N observational research where the treatment bears no resemblance to randomized assignment and where there is no opportunity for process-tracing or strong causal-process observations has made a fundamental contribution to some areas of political science research.\(^10\) Plausibly, it may continue to do so. But it will do so in a productive fashion only if its achievements are recognized and if reasonable standards for publication are accepted by the discipline. If not, I fear that broad questions like the relationship of development to democracy will go unanswered—or will be answered only by journalists and amateur prognosticators. And if this occurs, the cause of truth will be set back immeasurably. Perhaps social science will be purer, more scientific (from a certain angle). But it will be less consequential. And society will not be well-served.

Methodologists who are depressed about the uncertainty of knowledge in political science would do well to contemplate the field of archaeology. Here, researchers are in a much worse position vis-à-vis the things they want to find out (presumably, exactly the same sort of things that social scientists wish to find out about contemporary society). Their subjects are long-departed, leaving few remains. All is conjecture. Yet, this does not stop archaeologists from drawing conclusions—however tentative—about their subject. And these conclusions are generally regarded as an advance over popular myths about the past (though they may incorporate myths as a form of evidence).\(^11\)

Likewise, the difficulties presented by observational data should not prevent political scientists from drawing conclusions about causality—with the critical caveat that those conclusions be framed with appropriate confidence intervals. As Christopher Achen (1982: 77–78) has observed, all evidence is descriptive, for causation is an inferential form of knowledge. Even experiments don’t speak for themselves.

It is true, of course, that drawing inferences based on weak data is perilous. High uncertainty means that conclusions will often be wrong—less than 50% of the time, one would hope, but a lot of the time nonetheless. One must ponder carefully the ramifications of giving bad policy advice based on messy data analysis. Bad policies may be pursued, lives may be lost, and the credibility of social science may suffer accordingly. This fear prompts some researchers to identify scientific virtue with reticence. Professors should speak only when they are pretty certain of an answer. Otherwise, they should keep mum.

Sometimes it is important to resist the temptation to prognosticate, i.e., to insist that we do not know the answer to a question, however important that question might be. The flip side of the coin is that by refusing to engage questions of public concern, members of the academy withdraw from the debate. The questions do not go away, nor do the—quite possibly faulty—answers. Likewise, the policies persist, based on those faulty answers.

Suppose that political scientists, as a group, decide to
take a principled stand on the question of democratization by resolutely insisting on our ignorance. That is, we do not know why countries democratize, much less how to promote this process. We are prepared to tell you why extant studies are faulty, or at least highly uncertain. But we are not prepared to say which policies the U.S., or any other country, might pursue in order to foster political freedoms in the world because there is no secure causal knowledge on this question. Is this a responsible position to take?

Moral philosophers sometimes distinguish between “negative” and “positive” duties. The first is our duty to avoid inflicting harm on others. The second is our duty to do good (e.g., to alleviate suffering where we can do so). If one subscribes to this distinction, the harm caused by doing harm is much greater than the corresponding harm caused by not doing good. From this perspective, the idea of a social-scientific Hippocratic oath—i.e., pronounce only on issues where there are high levels of certainty—is appealing.

Yet, in the policy sphere the distinction between negative and positive duties is difficult to sustain. It is not clear, for example, that withdrawing assistance to democratization processes around the world would be a virtuous act. It is not even clear what such a withdrawal would consist of. Countries must have foreign policies, unless they are to withdraw entirely from the world, and any foreign policy will presumably have some effect on the pace and progress of democratization. It is difficult to construct a neutral foreign policy because in not taking a position on democratization a country still has a causal effect on that outcome. This means that for a policy science such as political science, it would be difficult to define and maintain a principled stance of “doing no harm.” Likewise, we are not a sect of priests whose moral purity is more important than the well being of society at large, so a deontological (person-centered) morality is in the end difficult to justify.

Of course, there is an argument for reticence if uncertainty is too complicated a notion to communicate to the general public or to policymakers. These consumers of political science look to academics for certainty and are illprepared, either psychologically or professionally, for confidence intervals. No matter how carefully advice is tendered, no matter how many caveats are attached, the message will be transmitted in the popular media as “Professor A says X causes Y, and we should follow policy Z.”

And yet, I do not see a way around it. We cannot, in good conscience, avoid communicating the knowledge that we possess about pressing issues of the day if they touch upon our area of expertise, even if our knowledge is based on messy data, and hence highly uncertain, and even if there is a risk that it might be misinterpreted and thereby lead to bad policies.

Let us return to the main argument briefly, so as to recapitulate and to clear up any misconceptions. Experiments, quasi-experiments, and slam-dunk case-based evidence are strongly preferred wherever viable, as they generally provide superior internal validity. It is my hope that political scientists will find ingenious ways to widen the applicability of these methods to questions that animate the field.

But where they are (a) impossible to implement, (b) of questionable internal or external validity, or (c) irrelevant for building general theory or addressing questions of public concern, these Grade A methods must be supplemented or replaced by other methods, crude those they may be. This large class of Grade B methods may be categorized broadly as large-N observational research, aka messy data analysis.

Thus, I propose one cheer—perhaps even two cheers—for this much-maligned but hardy breed of causal strategies. Messy data is often the least-bad of all feasible alternatives. And for this, it should be honored.

The problem is that political scientists have generally assumed that there is, or ought to be, one standard of causal inference with a very high level of certainty (say, 90% or 95%) applying to all work in the field. This is an unrealistic standard if we are to continue to pursue the panoply of diverse causal questions that have traditionally motivated the field—and that seem to have great policy and practical significance. Lower standards of certainty are required for some questions that are not amenable to Grade A methods.

Likewise, those who work with messy data need to muster the courage to state honestly and forthrightly the high level of uncertainty that usually accompanies their causal inferences. Do not simply recite the t statistic and p value. Messy data calls for grappling with research design issues that are not summarizable with asterisks.

Notes

1 Many of the points in this short essay are dealt with in a more detailed (and more nuanced) manner in Gerring (2011b). My thanks to Taylor Boas, Jake Bowers, Michael Coppsedge, Adam Glynn, Evan Lieberman, Jay Seawright, and David Waldner for their feedback on earlier versions of this polemic. Needless to say, they are not to be implicated in the argument.

2 The subordination of substantive arguments to methodological considerations is discussed in Mead (2010), Shapiro (2005), and Smith (2003).

3 I have written a book on case study methods (Gerring 2007) and am a strong proponent of experimental and quasi-experimental methods (Gerring 2011a).

4 These and other issues are addressed in Deaton (2010), Harrington (2000), Heckman (2010), Humphreys and Weinstein (2009), Learner (2010), Lieberson and Horwich (2008), Scriven (2008).


6 Various studies comparing analyses of the same phenomenon with experimental and non-experimental data show significant disparities in results, offering direct evidence that observational research is flawed (e.g., Benson and Hartz 2000; Friedlander and Robins 1995; Glazerman, Levy, and Myers 2003; LaLonde 1986). Cook, Shadish, and Wong (2008) offer a more optimistic appraisal.


9 Bartels (1997), Imbens (2003), Learner (1983), Levine and Renelt (1992), Montgomery and Nyhan (2010), Rosenbaum (2012), Rosen-
baum and Rubin (1983), Sala-i-Martin (1997), Sims (1988), Western (1995), Young (2009). Note that my understanding of a “robustness” test incorporates both alternative specifications (the usual focus of extreme bounds analysis and sensitivity testing) and alternative estimators (as discussed in most econometrics texts), as well as various operationalizations for key variables (X, Y, Z). Any important assumption underlying an empirical model that can be tested by altering some element of that model ought to be included in a set of robustness tests.

11 Here, my conclusions echo the traditional wisdom, as articulated by Laitin (2002) and Fearon and Laitin (2008).

11 Consider if an archaeologist were to discover a cross-cultural dataset of the sort that comparativists wrestle with—a first-century World Development Indicators. Presumably, he or she would be thrilled, and our knowledge of the first-century AD would be dramatically advanced.

References


Regression’s Weaknesses and Strengths:  
A Reply to Gerring

Jason Seawright  
Northwestern University  
j-seawright@northwestern.edu

Assumptions are the rule, not the exception, in both descriptive and causal inference in the social sciences. This fact has long been used as a defense of the specific families of assumptions used to make causal inferences on the basis of regression-type models (Freedman 2004: 195). Yet the defense is weak. Inferences differ in terms of the strength, complexity, plausibility, and testability of the assumptions they require. On all of these fronts, regression-type analysis of observational data often performs so poorly that it is difficult to give the results a persuasive causal interpretation. In what follows, I will make this argument by showing how hard it can be to assign causal interpretations to regression models that show either unstable or stable results across the range of models that the discipline considers plausible, as well as the challenges involved with drawing causal conclusions from either the unconditional or the conditional analysis of quantitative observational data. For these reasons, I disagree with Gerring’s argument that the regression analysis of messy data is a good default option for social scientists; instead, it is a weak default, an independent, additive, approximately normally distributed parameters and the independent variables, and which feature historical reasons, additive models which are linear in both the selection of plausible models also includes less salutary forms of selection. Some of these reflect ossified convention. For historical reasons, additive models which are linear in both the parameters and the independent variables, and which feature an independent, additive, approximately normally distributed error term, are our collective default for the analysis of continuous dependent variables (Stigler 1990).

Second, the set of models which are currently regarded by the scholarly community as plausible and which can be estimated using existing data comprise a quite unusual sample from the population of possible models for a given relationship. The distinctiveness of this sample is in part healthy: presumably, knowledge of cases and substance rules out a range of specifications that are statistically possible but in some sense foolish. Thus, we rarely estimate models in which the positions of planets, for example, are taken to predict economic performance or political institutions.

However, the extreme winnowing that produces our collection of plausible models also includes less salutary forms of selection. Some of these reflect ossified convention. For historical reasons, additive models which are linear in both the parameters and the independent variables, and which feature an independent, additive, approximately normally distributed error term, are our collective default for the analysis of continuous dependent variables (Stigler 1990).

Our sample of plausible models is further constrained by the set of available indicators. While scholars sometimes create new indicators to capture novel hypotheses that lie at the center of their explanatory agendas, they rarely go to the same amount of work to measure potential confounding variables. Instead, the control variables in our plausible models are generally some subset of the current collective stock of data. Some subset of that stock of variables becomes defined as the core control variables, without which a model is inherently implausible; this process of definition, I think, reflects in part an accumulation of past findings and arguments and in part a process of social consensus. But, regardless of the mix of these two...
components, such norms certainly further constrain the range of plausible models.

Last but obviously not least, the set of plausible models is limited by our contemporary repertoire of concepts and indicators. Scholars working before the development of systematic conceptualizations of, and survey measures for, the ideas of retrospective economic evaluations or strategic voting pressures would have an obvious excuse for failing to include those variables in their models of vote choice—but, good excuse or no, the models remain misspecified. The variables that will be discovered or invented over the next century quite evidently cannot be included in today’s models, even though they may be necessary for causal inference.

The net result of these and the other constraints listed above is that the range of results found in today’s set of published plausible models cannot even be taken as providing logical upper and lower bounds for the true causal effect. Some scholars might be tempted to argue that, while it is possible for the true causal effect to fall outside the range of contemporary statistical estimates, it is unlikely. This argument is not an implication of regression theory and is not even universally supported by tests that compare observational regression estimates with experimental benchmarks.

To sum up, unstable regression results on observational data simply do not teach us about the direction, magnitude, practical relevance, or theoretical importance of the underlying causal relations. We may tend to believe less in causal effects that cannot be consistently demonstrated using messy data, but such disbelief is not well grounded and should probably be resisted. That is to say, “we do not know” does not imply “it is not so.”

The Trouble with Stable Results

While researchers are likely to be broadly familiar with the argument that instability in statistical results demonstrates significant uncertainty in our knowledge about causal relations, it is much less widely discussed but nonetheless true that stable statistical results can also be compatible with uncertainty in causal knowledge. To see this point, let us consider one of the most stable findings in comparative politics: that GDP per capita is significantly associated with democracy. Some scholars make much of the distinction between predicting transitions to democracy and predicting democratic breakdown; for the moment, I will disregard this distinction, for reasons to be discussed below.

It is true that democracy and development are strongly related, for a wide variety of measures of democracy, a range of operationalizations of development, and a broad class of statistical models. Yet it nonetheless remains uncertain whether development in fact causes democracy.

While most models reproduce the widely accepted result that development increases the probability of democracy, some do not. In particular, Acemoglu, Johnson, Robinson, and Yared (2008) show that including country fixed effects in an analysis almost completely removes this relationship. A convergent finding can be shown using two simple cross-sectional regression models, shown in Table 1.

Model 1 in the table shows a bivariate regression predicting democracy on the basis of per capita GDP (logged, as is often the case in this literature, to deal with the skewness of the variable). The analysis is carried out using 1985 data, although the year is not important and similar findings can be produced for a wide range of years. Here we find the standard result: wealth strongly and positively predicts democracy.

Model 2 refines this finding, partitioning the democracy variable into two orthogonal components. The first component is a country’s rank in the global 1985 distribution of wealth, while the second is that country’s residual in a regression predicting logged GDP using GDP rank as an explanation. In other words, the rank variable shows countries’ relative order in the global economic hierarchy but not the fine detail of their level of wealth, while the residual shows the component of the level of wealth that cannot be predicted by rank order. The two components can be linearly combined to recover the original GDP variable.

This model allows us to ask which aspect of wealth—relative position in the world hierarchy or absolute resources—is in fact correlated with level of democracy. The question is crucial given that most theorizing about this relationship, from the days of modernization theory to the present, has treated the absolute level of economic resources as the cause of interest. Hence, if relative rather than absolute wealth is key, most theoretical work on this central issue has been misdirected in important ways.

If wealth per se is a cause of democracy, then both components in this partition of GDP should be associated with level

<table>
<thead>
<tr>
<th>Components</th>
<th>Model 1 Estimates (P Values)</th>
<th>Model 2 Estimates (P Values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-46.6 (&lt; 0.01)</td>
<td>-13.3 (&lt; 0.01)</td>
</tr>
<tr>
<td>Logged Per Capita GDP</td>
<td>5.6 (&lt;0.01)</td>
<td>0.2 (&lt;0.01)</td>
</tr>
<tr>
<td>GDP Rank</td>
<td></td>
<td>8.5 (0.59)</td>
</tr>
<tr>
<td>GDP Residual</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>N</td>
<td>128</td>
<td>128</td>
</tr>
</tbody>
</table>
of democracy. Moving up the rank order should help because it generally involves a gain in level of wealth, but increases in level of wealth that are not quite large enough to produce a change in rank order should also help. But in fact, as Model 2 shows, virtually all of the predictive power of the GDP variable is captured by the rank component; the coefficient for the residual component is not even close to achieving statistical significance.

The distinction between rank order and level of GDP is crucial because, while levels of GDP change substantially over time, rank orders do not. Between 1960 and 1990, for example, the correlation in GDP rank orders is 0.88. For this reason, the 1990 GDP rank order is almost as good a predictor of a country’s 1960 level of democracy as is that country’s 1960 level of GDP. In my judgment, these findings are consistent with the hypothesis that both long-term development trajectories and long-term regime trajectories are caused by decisions or institutional patterns at critical junctures well before the 20th century, an idea that is supported by much more robust case-study research (e.g., Mahoney 2010).

To the extent that these findings imply path dependence, most panel analyses of wealth and regime type are statistically problematic because they omit the critical historical events that set countries on one path or another (whatever those might be). Furthermore, findings relating wealth and democratic consolidation become causally ambiguous. Consolidation may be a consequence of a country’s wealth, in absolute or relative terms, or alternatively may be a component of an institutional package that helps propel high levels of long-term economic performance.

As this example shows, stable results across specifications may simply mean that all of those specifications omit the same key confounder. These issues do not arise in the same way for experiments and other strong research designs. Because of their reliance on randomization or detailed case information, findings from these kinds of studies are, in comparison with the regression analysis of observational data, much less fragile to alternative model specifications.

The Trouble with Unconditional Inference

If neither stable results nor unstable results, with reference to the regression-type analysis of observational studies, can be logically taken to have clear implications for causal inference, the reader may begin to doubt that we could ever be confident that we have found causal knowledge with such a model. This doubt is, I think, healthy. To further nourish it, let us consider the same dilemma along the lines of another dichotomy, that between unconditional and conditional inference.

Unconditional inference involves a simple bivariate analysis of the relationship between the hypothesized cause and the outcome. For experiments, and many natural experiments, unconditional inference should be seen as the gold standard for causal inference (Freedman 2008, Dunning 2010). However, for observational studies, scholars have long been taught to regard unconditional inferences as entirely suspect.

The reason for this suspicion is the very real possibility of confounders, i.e., variables which belong in the model but are excluded from it and that distort the relationship between the independent and dependent variables. Experiments greatly reduce the problem of confounding by randomly assigning cases to treatment groups; successful natural experiments similarly abate confounding through a randomization, albeit one not controlled by the scholar. In regression-type observational studies, however, there is no randomization. Instead, there is every reason to believe that cases take on their observed scores on the independent variable because of complex social, economic, and political processes that may well also directly affect the outcome. Confounding, we anticipate, is therefore ubiquitous.

This does not necessarily mean that an unconditional inference is incorrect—there may by some miracle be no confounding in this particular analytic instance, or it might by extreme coincidence be the case that the various biases brought about by confounders happen to more or less cancel out. But it is nonetheless clear that confounding will usually be a problem, that we have no tools for identifying the handful of instances in which it might not be a problem, and therefore that unconditional analysis will rarely provide reliable causal inference.

The Trouble with Conditional Inference

The conclusion that unconditional inferences are unreliable for observational studies should not surprise. The following argument may be more surprising: conditional inferences, i.e., inferences that introduce control variables, are typically no more reliable than unconditional inferences in observational studies. I will develop this argument in two stages. First, there are some variables that, when added to an otherwise correct model as controls, distort causal inference. Second, even variables which appear as controls in the correct model may often, in imperfect real-world models, make causal inference worse, not better.

For decades, the literature on causal inference has warned against conditioning on post-treatment variables, i.e., variables that are caused by the independent variable (for useful recent discussions, see Rosenbaum 2002, King and Zeng 2006, and Morgan and Winship 2006). When a scholar conditions on a post-treatment variable, she inadvertently subtracts the effect of any causal pathway from the main independent variable, through that post-treatment variable, and to the outcome. If this subtraction is not taken account of analytically, the result will be a biased estimate of the overall causal effect of the independent variable of interest. It is somewhat less widely known that other categories of impermissible control variables exist; in particular, conditioning on “collider” variables can create new problems of confounding even when none existed before (Pearl 2000: 17–18, Cole et al. 2010). What happens if a variable is a confounder but also meets the criteria for post-treatment or collider status? If we are to follow the standard advice for achieving unbiased causal inference, such variables must be simultaneously included and excluded from our models. In a typical observational study, we lack the ability to identify with confidence which of the potential control vari-
variables belong in any of these categories, so it is hard to be sure whether we are making things better or worse by conditioning.

Suppose that, for some potential control variable, we are somehow entirely confident that the variable is a confounder and is neither a collider nor in any post-treatment. Surely inference conditional on such a control variable is more reliable and closer to the causal truth than unconditional inference?

In fact, there is no certainty about this at all. The problem is that, while we may have identified a confounder, we are almost never certain that we have identified the last confounder. Thus, it remains probable that other omitted variables bias the inference even when conditioning on the known confounder. If the net bias produced by the set of remaining confounders is zero or points in the same direction as the bias connected with our known confounder, then the conditional inference will be superior to unconditional inference. However, the net remaining bias can point in the opposite direction, in which case conditional inference will often be worse than unconditional inference—a circumstance which, in some simulation studies, holds for 50% of potential control variables (Clarke 2005).

So, as every introductory methods text will tell us, in observational studies we cannot trust unconditional inferences. Yet barring unusual sorts of a priori causal knowledge, we also cannot trust that our conditional inferences will be closer to, rather than farther from, the truth than the unconditional inference. The value added by control variables can be obscure.

When the Stakes are High

The above arguments, together with the preference I and other scholars express against regression-type analysis and for in-depth case-based arguments, on the one hand, and experimental or natural-experimental designs, on the other, are sometimes seen, by Gerring and others, as an unhelpful form of “methodological perfectionism.” Are there important questions that cannot be studied using these stronger designs? For such questions, does regression not offer a best-available approach?

I am unsure. It is true that there are many important substantive domains in political science that have been dominated by regression-type studies of observational data. Such designs have been the stock-in-trade of our discipline and the centerpiece of our methodological training for decades, so their dominance should not surprise us. Nor should we take the de facto dominance of these techniques as an indication that other approaches cannot work. Until relatively recently, experimental and natural experimental research had peripheral status in most political science subfields, and powerful voices made arguments denigrating the inferential value of case studies vis-à-vis regression.

What is certain is that political scientists have already, over the last decade, found ways of using these techniques to address questions at both macro and micro levels that have long been central to our discipline (e.g., Wantchekon 2003, Brady and McNulty 2004, Bhavnani 2009, Humphreys and Weinstein 2009, Corstange 2010, Dunning and Harrison 2010). It seems at least possible that an ongoing emphasis on the importance of research design and the relative inferential weakness of regression-type studies will motivate the hard work and ingenuity necessary to bring these techniques into full engagement with a broader range of issues.

In the end, however, I expect it to be the case that some important questions remain inaccessible for these methods. Of course, one might remark, there are always important questions that remain beyond the scope of all scientific methods; that a question matters does not guarantee that we can answer it well. And, for the most important questions, is it not true that the quality of our answers is unusually important?

Where Regression Shines

None of this should be taken as an attack on regression analysis, or a call for a ban on the technique. What regression does well, it does very well indeed—in fact, sometimes optimally well, as statistical theory can show. Trouble arises when we push regression too far outside its domain of competence.

What, then, are the strengths of regression? The technique is a powerful tool for the summary of complex cross-tabulations and scatter plots. Regression can sometimes make consistent but small descriptive relationships among variables more visible and can often dramatically aid comprehension of central themes in data by replacing an overwhelming mass of numbers or dots with a few key estimates (Berk 2003).

When scholars move beyond the tasks of summarizing and clarifying which constitute the key area of regression’s strength in the social sciences, trouble can arise. It is important to understand that, in terms of inferential logic, regression is no different from the (potentially multidimensional) scatter plot or cross-tabulation that it summarizes. Matrix algebra simply does not convert observational data into causal laws (Humphreys and Freedman 1996, Freedman 1997, Freedman 1999).

When regression is used with careful attention to its real strengths, it can be a powerful tool, along with difference-in-means tests, graphs, cross-tabulations, and other such techniques, in the analyst’s arsenal for descriptive and exploratory analysis. Furthermore, there are certainly moments when one or another piece of descriptive knowledge has strong causal implications; in such instances, regression may sometimes play a pivotal role in a causal argument.

However, we must accept that regression analysis of observational data will usually leave a great deal of causal uncertainty in its wake. Indeed, we cannot know in general whether regression analysis of messy data moves us closer to, or farther from, causal understanding. Our theorems cannot help us here; those which show regression-type analysis in a positive causal light do not apply to messy data, and those that do apply for messy data usually lack causal implications. So any defense of regression analysis of messy data must be pragmatic: the technique has to be shown to work for some important goal. That demonstration of efficacy has to be specific to the subject matter at hand and independent of the regression analysis itself. An example is regression work on forecasting election results; here the regression analysis of messy data has been shown to have some practical (predictive, although
not causal) value through out-of-sample prediction. However, we rarely produce such demonstrations of practical value for our regression research. As such, we simply cannot say whether we are better off with or without regression-type research in these contexts.

To the extent that our discipline values causal over descriptive knowledge, we must consider the possibility that regression-type studies of observational data have been significantly overvalued and overrepresented in our history over the last several decades. It may be time to shift some portion of resources such as funding, training, institutional support, and pages in our journals away from regression-type studies and toward case studies, experiments, natural experiments, and related approaches.

Note

1 See, e.g., a symposium of ten articles on U.S. election forecasting in the October, 2008, issue of PS: Political Science & Politics.

References


Messy Data, Messy Conclusions: A Response to Gerring

Adam Glynn
Harvard University
aglynn@fas.harvard.edu

What can we learn from the analysis of a large-N observational data set (aka “messy” data)? Gerring argues that despite warnings from a number of critics, such an analysis may be deemed adequate as long as

... it allows us to update our priors, it beats the alternatives, and it presents a plausible uncertainty estimate.

This seems a rigorous benchmark. Depending on our definition of plausible, even randomized trials may fail this standard when issues of treatment compliance, treatment heterogeneity, experimenter effects, or interference between units muddy the interpretation of results. Of course, observational studies may fail this standard even when such issues are not a concern. Regression results from observational studies have two additional sources of uncertainty when compared to randomized studies.

The first is due to a lack of specificity about the manipulation of explanatory variables. In an experiment, the researcher controls the explanatory variable, and hence it is manipulable by definition. In observational studies, explanatory variables might hypothetically be manipulated in a number of ways, and these different manipulations can imply different effects. In the democratic consolidation example cited by Gerring, the “effect” of income (on the likelihood of a relapse into authoritarianism) depends on exactly how one intends to manipulate income. If income is increased by the discovery of oil, this may have different consequences than if income is increased by...
improved education.

The second source of uncertainty that is special to observational studies is the instability of regression results across different sets of pre-treatment conditioning variables. Randomized experiments do not exhibit this instability because in large samples, different treatment groups are guaranteed to have similar distributions for all possible pre-treatment variables.

However, despite these two complications, it is certainly the case that observational data allow us to “update our priors” regarding effects. To see this, it is helpful to more closely consider a single observation (country) from the democratic consolidation example. Suppose we only know that this country is poor, the “effect” of income is a comparison between the outcome for this country (relapse or consolidation) under their current state of income (poor), and the outcome we would have observed (relapse or consolidation) if we had somehow increased income (rich). There are three possible effects of increasing income from poor to rich: negative (consolidation to relapse), positive (relapse to consolidation), no effect (relapse to relapse or consolidation to consolidation). Our priors in this case might represent probabilities over these three effects.

Now suppose that we observe the outcome variable for this country, and in fact this poor democracy relapsed into authoritarianism. For this country, we now know that increasing income (from poor to rich) would not have had a “negative effect” (from consolidation to relapse), because we observed relapse when poor. Increasing income could only have had either a positive effect (from relapse to consolidation) or no effect (from relapse to relapse). Therefore, if our prior beliefs put any positive probability on a negative effect, we must update because we now know that there is no probability of a negative effect for this country.

Furthermore, note that our method for ruling out the “negative effects” relied only on our knowledge that this country relapsed when poor—we did not impose any constraints on the outcome we would have observed if this country had been rich. Therefore, we can rule out negative effects regardless of the exact method of manipulation. To be specific, for this country we know that the effect of increasing income due to discovering oil, or the effect of increasing income due to improving education, or the effect of increasing income through any other means, could not have had a negative effect—it is logically impossible to move from consolidation to relapse if we know that the starting point is relapse.

While these sorts of logical arguments are straightforward when considering individual cases, they can also be used to consider the limits of instability for regression results. These limits provide conservative estimates of uncertainty. In the next section, I provide a stylized example.

The Limits of Instability for Regression Slopes with Observational Data: A Stylized Example

While the potential instability of regression results with observational data is well known, it is helpful to consider a stylized example in order to explore the source and the limits of this instability. Figure 1 (a) presents an example of a simple regression. In this case, both X and Y are dichotomous and only take on the values 0 and 1, therefore this plot can also be thought of as a 2x2 table (the points in the plot have been jittered so all 12 of them can be seen). Notice that the arrangement of points in the plot implies a positive regression slope of 1/3. Two thirds of the X=1 points have Y=1, while one third of the X=0 points have Y=1, so the difference is 1/3.

Of course, the regression slope of 1/3 that we see in Figure 1 (a) merely represents a simple regression result. We could get different regression results by conditioning on a third variable. For example, consider a hypothetical dichotomous conditioning variable that takes the values “a” and “b.” Figure 1 (b) presents the same data that were presented in Figure 1 (a), except now the values of this conditioning variable are represented on the plot. The conditional regression analysis (based on this dichotomous conditioning variable) works by fitting two separate regression lines—one to the “a” points and one to the “b” points. For this example, we see that the slope is zero for both of these lines, and therefore our conditional regression estimate is zero (quite different from the 1/3 we got in the simple regression).

This instability of regression results is well known to most practitioners, and we can get many different answers by considering alternative conditioning variables. For example, Figure 1 (c) demonstrates how we can get a very small regression estimate. This plot presents the same data again with a different conditioning variable that is trichotomous (taking the values “a”, “b”, and “c”). In Figure 1 (c), the regression slope is zero for both the points that take the conditioning value “a” and for the points that take the conditioning value “c.” However, the regression slope is -1 for the points taking the conditioning value “b.” In a model with different conditional slopes for different observations (i.e., a model with interactions), we obtain the overall estimate by averaging across the conditional slopes according to the proportion of observations associated with each slope (Cochrane 1968). Using weights according to the proportion of the “a,” “b,” and “c” observations, the conditional estimate is the average slope, which is -2/12.

Similarly, Figure 1 (d) demonstrates how we can get a very large regression estimate. This plot presents the same data again with a different trichotomous conditioning variable. In this plot, the regression slope is zero for both the points that take the conditioning value “a” and for the points that take the conditioning value “c.” However, the regression slope is 1 for the points taking the conditioning value “b.” Using weights according to the proportion of observations taking each value of the conditioning variable, the average conditional estimate is the average slope, which is 6/12.

Again, it is not remarkable that Figures 1 (a), (b), (c), and (d) demonstrate the instability of regression results. What is remarkable, or at least is less well known, is that the plots in Figure 1 demonstrate a limit to this instability.

To see this, first imagine all possible conditioning variables that are not collinear with X. That is, if you are allowed to label the 12 points in any manner you wish (“a,” “b,” “c,” “d,” ...), so long as each label has a representative in the X=0 and
Unconditional Estimate = $\frac{1}{3}$

Average Conditional Estimate = 0

Average Conditional Estimate = $-\frac{2}{12}$

Average Conditional Estimate = $\frac{6}{12}$

X=1 groups, there is no labeling that will produce an average slope smaller than $-\frac{2}{12}$ (as in Figure 1 (c)) or greater than $\frac{6}{12}$ (as in Figure 1 (d)). In other words, if we restrict ourselves to conditioning variables that are not collinear with X, there is no set of conditioning variables that would produce an average slope estimate that is smaller than $-\frac{2}{12}$ or greater than $\frac{6}{12}$—there is a limit to the instability of the regression results when we assume an absence of collinearity.

In fact, even if we allow conditioning variables that are collinear with X (i.e., we are allowed to label the 12 points in any manner we wish), the average slope must at least be $-\frac{1}{3}$ and can at most be $\frac{2}{3}$—there is a limit to the instability of the regression results. This limit is often known as the Manski bounds (Manski 1990, 2003). In the language of the previous section, we can rule out average slopes smaller than $-\frac{1}{3}$ or larger than $\frac{2}{3}$.

**Development, Democracy, and Gerring’s Proposed Standards?**

If the Manski bounds demonstrate limits to the instability of regression results, then messy observational data may add to our knowledge of a given subject area—even if we cannot provide a single answer, we can rule out some answers. However, while we learn something from messy data, we may be unsatisfied by the wide range of answers we obtain. Without additional information or assumptions, we cannot state that any particular slope estimate (or set of slope estimates) in the bounds is more likely than any other. Perhaps most importantly, without additional information or assumptions, we can never rule out a slope of zero, and therefore we can never establish the presence of an effect.

As an example, consider the previously discussed result on democratic consolidation. Has this finding added to our knowledge on the subject? In light of the stylized example above, the answer seems to be yes. A positive correlation between development and retained democracy does not conclusively indicate a positive causal effect; however, it does rule out highly negative effects.\(^5\)

However, while the Manski bounds demonstrate conclusively that we learn something from this observational data (because they represent all the results we might get from any conceivable set of conditioning variables), they are likely to be...
conservative because they only utilize data on the explanatory variable and the outcome variable. By utilizing other variables, we may be able to tighten these bounds, but only when the information from these variables is combined with additional assumptions.

As an example, consider the typical practice of running regressions with different sets of conditioning variables, and using the range of results we get from these regressions as bounds. These bounds will be valid only if we assume that one of our sets of conditioning variables is the correct set, or if we assume that the answer we would get from the correct set of variables is contained within these bounds. In other words, the conditioning variables may allow us to tighten the Manski bounds, but only if we are willing to make assumptions on the basis of these variables. 

Unfortunately, as demonstrated by many of the regression critics cited by Gerring, these alternative bounds (the range of estimates produced by running different regression specifications) are likely to be anti-conservative. Because regression results can, in practice, only utilize non-collinear sets of measured confounders, we will often find it untenantable to believe that any of the regressions we have run use the correct set of conditioning variables.

This leaves the analyst in a quandary. Given our data, we have a conservative answer (the Manski bounds) and an anti-conservative answer (the bounds on regression results), but we may not have a “plausible” answer. In order to move forward, the analyst must be willing to make and defend assumptions. In some cases, the bounds on regression results are expanded by making assumptions about the effects of unobserved conditioning variables (Brumback et al. 2004, Lin et al. 1998, Rosenbaum 1987, 2002, Rosenbaum and Rubin 1983). In others, the Manski bounds are tightened by making a variety of assumptions (Manski 2003), sometimes in a Bayesian framework (Quinn 2011).

Specifying and defending these assumptions is hard work, and we may never arrive at a set of assumptions that will be agreeable to all readers—messy data lead to messy conclusions. Still, this sort of work with messy observational data can lead to useful conclusions—see for example the Cornfield et al. (1959) study on smoking and lung cancer. It is unfortunate that typical practice eschews this hard work and presents only the anti-conservative bounds implicit in regression results.

Notes

1 More typically, we might propose that these probabilities follow a Dirichlet distribution, and specify our priors on the parameters of this distribution.

2 Linearity is justified in this case due to the binary treatment.

3 To avoid concerns about a lack of significance, we can interpret the points in this plot to merely represent the proportions in the cells of a 2x2 table. Therefore, the sample size might actually be quite large.

4 This process of averaging over the slopes of regression lines in an interactive model is sometimes known as direct adjustment (Cochran 1968), and is the most straightforward way to estimate an average effect using an interactive regression model.

5 With a continuous explanatory variable, the causal question of interest must be stated more precisely in order to estimate effects or specify the regression bounds. Often, parametric assumptions are made that obviates the need for this precision.

6 It is worth noting that this is also true for post-treatment variables, although standard regression analysis cannot be used when conditioning on post-treatment variables. Glynn and Quinn (2011) provides an example on the use of post-treatment variables to tighten the Manski bounds.

References


Quinn, Kevin M. 2011. “What Can We Learn From a 2x2 Table?” Unpublished Manuscript.

Qualitative Research: Progress Despite Imperfection

Andrew Bennett
Georgetown University
bennetta@georgetown.edu

I have encountered what John Gerring aptly describes as a fear of “the specter of methodological perfectionism” in my students and colleagues. In my view this fear imputes to methodologists more optimism on the perfectibility of research methods and more pessimism on the contributions of imperfect methods than most of us actually hold, but like any phobia, this fear is sufficiently real in the minds of those who hold it that it deserves remediation.

My own view, similar to that in Gerring’s *Social Science*
Methodology: A Criterial Framework (2001), is that there are no perfect research methods for observational data or even for the experimental and quasi-experimental approaches that are generating renewed interest, so we must make methodological choices among imperfect alternatives that represent different tradeoffs of desiderata such as external and internal validity. Alexander George and I tried to make this clear in our 2005 book by discussing the limits as well as the comparative advantages of case study methods (2005: 22–34) and explicitly rejecting perfectionism in research (2005: 10–11, fn. 14). Soon after we published our book, however, I noticed that when they presented their work, graduate students, visiting speakers, and even colleagues would cast a nervous glance in my direction whenever the issue of qualitative methods arose, apparently concerned that I would accuse them of grievous methodological mistakes.

Of course I do critique other scholars’ research methods, just as I expect them to critically evaluate my own research methods and findings. My critiques, however, assume that all methods are imperfect, that methodological choices hinge in part on prior theoretical and empirical knowledge that other scholars usually possess to a greater degree than I do regarding their particular research topics, and that reasonable people can disagree on the complex question of which methodological choices are better than others in a particular project with particular research objectives (theory generation, theory testing, etc.). The fear of methodological perfectionism should thus be relatively easy to lay to rest, even if we must all continuously concern ourselves with how our methods might be better.

In addressing the specter of methodological perfectionism, however, we must also address the opposite worry of methodological fallibilism: if our methods are inevitably imperfect, and if it is difficult to get inter-subjective agreement on what methods to use in a particular project, how can we make any claims that our research leads to cumulative, progressive, and general knowledge that is useful to policymakers?

The challenge, then, is to prevent the “best” from being the enemy of the “good,” or to ensure that our aspiration to optimize our methods does not paralyze us and keep us from pursuing work that is imperfect but that makes genuine contributions worthy of the time and effort invested. In this essay, I address the questions of both perfectionism and fallibilism by exploring an example from my own work that I consider to be methodologically imperfect but theoretically and empirically progressive and policy relevant. I proceed as follows: after a very brief discussion of the general issues that Gerring raises, I re-examine research that colleagues and I did in the 1990s on burden-sharing in security coalitions (Bennett, Lepgold, and Unger, 1994, 1997). I use this example to demonstrate how even imperfect work such as ours, together with subsequent imperfect research by other scholars, can contribute to cumulative, generalizable, and policy-relevant insights. After summarizing our research, which focused on burden-sharing in the 1990–1991 Gulf War coalition, I review some of the subsequent data and research on burden-sharing in later security coalitions, focusing on those involved in the Balkans in the 1990s, Afghanistan since 2001, and Iraq since 2003. This discussion shows the strengths of our research in anticipating burden-sharing outcomes in later security coalitions as well as the weaknesses of our work in overlooking important phenomena. My analysis builds upon and updates the discussion of this example in George and Bennett (2005: 255–61), and I present it in such a way that colleagues can teach this example Socratically to get students to think about issues of case selection, theory-generation, typological theorizing, policy relevance, and theoretical progress.

Methodological Perfectionism

John Gerring’s essay argues that “the current vision of perfection prizes causal knowledge over descriptive knowledge, theory appraisal over theory discovery, micro-theory (aka micro-mechanisms) over macro-theory, and internal validity over external validity.” Gerring does not cite any specific works in this passage as embodying these four perspectives, but as a prelude to re-analyzing my earlier research I briefly summarize my views on them here.

On the first point, I support the view, nicely expressed by King, Keohane, and Verba (1994: Chapter 2), that both causal and descriptive knowledge are important and that we can’t make useful causal inferences unless we get our descriptions right. Much of the value of the historical explanation of cases lies in providing detailed descriptions of the historical events that we deem relevant to building and testing theories, and many important contributions to the social sciences are empirical rather than theoretical.

Regarding the appraisal and discovery of theories, George and I (2005:12) critiqued King, Keohane, and Verba (1994: 14) for putting too much emphasis on theory appraisal relative to theory discovery. In contrast, we included the term “theory development” in the title of our book to embrace both the discovery and appraisal of theories.

On the micro-macro issue, George and I have argued that explanation via reference to causal mechanisms requires researchers to commit in principle to make their theories consistent with the finest level of detail that they can observe. We also explicitly stated, however, that this commitment “does not mean that the explanatory weight or meaningful variation occurs at this level” and we added that “macrosocial mechanisms can be posited and tested at the macrolevel” (141–42).

Finally, although some might read George’s and my book as emphasizing internal validity over external validity, this misses the crux of our argument. Naturally, we all want both external and internal validity, but the point George and I made is that statistical methods have some advantages at the former and case studies are in some ways better at achieving the latter (hence our encouragement of multi-method research). We did not privilege internal over external validity (see pp. 109–124 on generalizing from case studies), but we warned against doing the reverse, critiquing King, Keohane, and Verba for offering advice on ways of increasing sample sizes without noting the tradeoffs this can involve regarding internal validity and conceptual stretching (George and Bennett 2005: 172–176).

In short, I see the dichotomies Gerring lays out as being
“both-and” desiderata rather than “either-or” propositions, and I welcome research projects that make widely different tradeoffs among these desirable features of social science. If I understand Gerring correctly, this is his view, as well.

**Progress Despite Imperfection:**
**Burden Sharing in Security Coalitions**

In the early 1990s, with my colleagues Joseph Lepgold and Danny Unger, I set out to study the puzzle of why coalition contributions to the Desert Storm coalition that ousted Iraqi forces from Kuwait in 1991, totaling $70 billion dollars and 240,000 troops, were far more extensive than would have been predicted by collective action theory, which at that time was the dominant theoretical approach to alliance behavior. By itself, collective action theory should have predicted free-riding from all states except the United States in response to Iraq’s invasion of Kuwait, as the United States appeared to be capable of and willing to evict Iraq from Kuwait all by itself if necessary (President Bush publicly vowed in August 1990 that the Iraqi invasion of Kuwait would not stand even before he had concrete commitments of support from U.S. allies).

To address the surprising lack of free riding, we developed a typological theory that integrated collective action theory with three other explanations. First, all other things equal, balance of threat theory should have predicted contributions from states like Kuwait and Saudi Arabia whose occupation by or proximity to Iraq made them feel threatened and who feared the United States would not provide the public good of security against Iraq to a sufficient degree without assistance. Second, alliance security dilemma theory should have predicted contributions from states which were at that time heavily dependent on the United States for their security, such as Germany and Japan. Finally, theories of domestic politics suggested that a state would contribute to the Desert Storm coalition if public opinion in that state favored a contribution and were powerful enough to override public opinion.²

Our typological theory on how combinations of the variables in each of these theories would interact in producing outcomes is summarized in Table 1. Note that the columns in Table 1 indicate how the variables in one theory, acting in isolation from those of other theories, would predict a contri-

<table>
<thead>
<tr>
<th></th>
<th>Collective Action</th>
<th>Balance of Threat</th>
<th>Alliance Dilemma</th>
<th>Domestic Politics</th>
<th>Expected Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saudi Arabia</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td>Turkey</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Contribute*</td>
</tr>
<tr>
<td>United States</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td>Great Britain</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td>Egypt</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Contribute</td>
</tr>
<tr>
<td>France, Canada, Australia</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td>Germany, Japan</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Contribute*</td>
</tr>
<tr>
<td>Iran, Syria</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>No Contrib.</td>
</tr>
<tr>
<td>China, USSR</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>No Contrib.</td>
</tr>
</tbody>
</table>

*In types marked with an asterisk, countries are expected to contribute only if strong state leaders override domestic opposition, as happened in Egypt and Turkey. In Japan and Germany, security dependence on the United States was so high that domestic opposition was muted.
bution from a state. For example, collective action theory would look at several variables, including the relative capabilities of a state, the expected costs of forcing Iraqi troops out of Kuwait, and whether a state was needed as a basing ground (as in Saudi Arabia and Turkey), to determine whether collective action theory should have predicted a contribution from or free riding by that state. The rows of Table 1 indicate how states with the specified combinations of predictions from the four individual theories should be expected to behave.

Our typological theory did quite well in fitting and explaining the contributions of the cases we studied, including the United States, Britain, Egypt, France, Germany, and Japan, and the process tracing evidence in these cases largely validated the theory (Syria’s contribution was an anomaly for our theory but the rest of the cases listed largely fit the theory [Bennett, Lepgold, and Unger, 1997]). In particular, the study highlighted the fact that alliance dependence was in itself sufficient to engender large economic contributions from Japan and Germany even though essentially all the other variables in these cases created incentives to try to free ride on American efforts.

Yet the first Gulf War was not an entirely independent test of our theory, as we had constructed our theory with some preliminary knowledge of the outcomes of the cases (the process tracing evidence, of which we were largely ignorant before actually carrying out the case studies, provided more independent corroboration of the theory’s hypothesized mechanisms).

A tougher test concerns contributions and non-contributions to subsequent security coalitions. Table 2 presents some of the cases that do and do not fit our initial theory well from security coalitions in Bosnia (first the United Nations Protection force from 1992–1995 [UNRPOFOR] and then the NATO implementation force 1995–1996 [IFOR]), Iraq 2003–2011, and Afghanistan 2001–2011. Cases are identified by the name of the (non)contributor, the year the coalition started (not necessarily the year of a country’s contribution), and the number of troops contributed (in bold). The cases that arguably fit our theory are in plain type, while those that do not are in italics. Table 2 includes only countries that contributed more than 500 troops, except for the case of the small U.S. contribution to UNPROFOR to allow a comparison to the large U.S. contribution to IFOR. I also include a few instances of countries that might have been expected to contribute more than 500 troops but did not. Each coalition also had a large number of contributions of fewer than 500 troops.

The UNPROFOR and IFOR cases suggest our theory got a lot right (Lord and Lord, 1997): the more powerful and proximate countries were the biggest contributors (collective action and balance of threat); coalition actions were more effective once the United States, which had stayed out of UNPROFOR, got involved in IFOR (collective action); the United States felt only a modest threat from the Balkan crises and was reluctant to get involved until the credibility of NATO was at stake (domestic politics, balance of threat); and once the United States did become involved through IFOR, many smaller states seeking to join NATO made contributions to win favor from the U.S. (alliance security dilemma).

Yet the high contributions of Jordan, Pakistan, and others to the UNPROFOR coalition do not fit our theory well. Our theory would have predicted modest symbolic contributions from each of these countries. One possible explanation is the Muslim countries’ religious affinity for the Bosnian Muslims. A more likely interpretation, however, is evident in the comparison with troop contributions to NATO’s IFOR operations: U.N. peacekeeping payments of about $1,000 per soldier per month, well more than what many developing countries paid their soldiers, may have helped motivate some countries’ contributions. This is evident in the fact that nearly all the developing countries that were big UNPROFOR contributors dropped out of the coalition when it transitioned to the pay-your-own-way IFOR force (Turkey, aspiring to be a member of the EU, stayed in the coalition). The role of such side payments does not contravene the logic of our original theory—collective action theory allows for private goods and side payments, and our case study of Egypt noted that one of the motivating factors behind its contribution to the 1991 Gulf War was that the United States forgave Egypt’s debts in return. Yet our initial research could have put more emphasis on the potential role of side payments (Tago 2008).

A second interesting phenomenon not captured in Table 2 is the large number of countries giving token contributions of a few soldiers. We understood during our initial research that in order to limit coordination costs the U.S. military wanted only a few large contributions from other countries, and niche contributions of capabilities, like mine-sweeping, that it lacked. The State Department, meanwhile, wanted to count as many countries as possible as contributors, regardless of the size of their contributions, to show the legitimacy of U.S. operations. Clearly, the latter concern often won out, but how, when, and at what cost are interesting and underexplored questions. In particular, when the number of coalition contributors with a say over key issues like prioritization of bombing targets grows larger, the transactions costs of achieving agreement on military tactics and strategies grows (Auerswald 2004). Also, the motives of small contributors may have extended beyond alliance dependence to a desire to share the spoils (or contracts) of U.S. coalitions and perhaps even a desire to learn from (or spy upon) U.S. military operations. Here again, more research is warranted.

The data on contributions in Iraq and Afghanistan similarly show the value of our theory while at the same time suggesting several other researchable puzzles that we failed to anticipate or overlooked. The combination of collective action theory, balance of threat, alliance dependence, and domestic politics clearly remained powerful as an explanation of states’ contributions. It is not surprising that the 2003 Iraq coalition was much smaller than that in 1991, for example, due to the lower sense of a shared threat from a weakened Iraq in 2003 and the decline in alliance dependence on the U.S. as the Cold War receded. Our domestic politics hypothesis helps explain why the newly elected government of Turkey refused in 2003 to allow the United States to use Turkey to stage a northern prong of its invasion of Iraq (Baltrusitis 2010).
Table 2: Contributions and Non-Contributions
That Do and Do Not Fit the 1994 Theory

<table>
<thead>
<tr>
<th>Year</th>
<th>Country</th>
<th>Collective Action</th>
<th>Balance of Threat</th>
<th>Alliance Dilemma</th>
<th>Domestic Politics</th>
<th>Expected Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>France</td>
<td>4,600</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td></td>
<td>UK</td>
<td>3,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>UK</td>
<td>7,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>France</td>
<td>7,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995</td>
<td>Italy</td>
<td>2,500</td>
<td></td>
<td></td>
<td></td>
<td>Contribute*</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>20,000</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Contribute</td>
</tr>
<tr>
<td>2001</td>
<td>US</td>
<td>23,600</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td>2003</td>
<td>US</td>
<td>250,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>UK</td>
<td>46,000</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Contribute*</td>
</tr>
<tr>
<td>1992</td>
<td>Canada</td>
<td>2,000</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Contribute</td>
</tr>
<tr>
<td></td>
<td>Neth.</td>
<td>1,700</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Turkey</td>
<td>1,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>1,400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Denmark</td>
<td>1,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td>1,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Poland</td>
<td>1,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ukraine</td>
<td>1,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Czech.</td>
<td>1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Belgium</td>
<td>900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Norway</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Slovak Rep.</td>
<td>600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neth.</td>
<td>2,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Turkey</td>
<td>1,600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Norway</td>
<td>1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sweden</td>
<td>900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Czech.</td>
<td>900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Denmark</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Poland</td>
<td>700</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UK</td>
<td>8,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>France</td>
<td>1,700</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>2,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Italy</td>
<td>2,400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neth.</td>
<td>1,800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Poland</td>
<td>1,100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spain</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Turkey</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Denmark</td>
<td>700</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>Norway</td>
<td>600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Romania</td>
<td>600</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Italy</td>
<td>3,200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Poland</td>
<td>2,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Georgia</td>
<td>2,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Notes on Table 2: 1992 is UNPROFOR deployments as of 2005; 1995 is IFOR deployments as of September 1996; 2001 is Afghanistan deployments as of June 2008; 2003 is peak deployments in Iraq between 2003 and 2007. Anomalous cases are in italics; troop numbers are in bold (troop numbers often changed over time and these numbers are indicative rather than definitive). NC is for surprising non-contributors. Pakistan is not listed as a contributor in the Afghanistan coalition because it does not have troops in Afghanistan, but Pakistani soldiers have contributed to the coalition effort by fighting Islamic militants in northern Pakistan.

One interesting and researchable phenomenon that our initial research only partly anticipated is that when two coalition operations (Afghanistan and Iraq) were ongoing simultaneously, the operation in which the alliance leader (in this case, the United States) was having more difficulty rounding up contributions was the one in which the leader called in favors from its most dependent allies. This is evident in the fact that Korea, Poland, Georgia, and Ukraine, all feeling dependent on the United States for their security, were disproportionately large contributors in Iraq (see Baltrusitis 2010 on the Korean case). At the same time, large U.S. NATO allies were bigger contributors in the U.S. led coalition in Afghanistan, which had more legitimacy in the views of these contributors. This is consistent with our original theory’s emphasis on the alliance dependence variable, but it is a rather novel example in that we did not anticipate how such dependence might play out when two operations of unequal global legitimacy were ongoing.

In addition, the British contribution in Iraq and the Australian contributions in Iraq and Afghanistan may attest to the power of individual leaders (see Dyson 2007 on the British case). This could be encompassed by our domestic politics variables but it accords a bigger role to specific individuals than we gave in our original articulation of these variables.

Probably the biggest missed opportunity in our original research, however, concerns our decision to set aside an anomalous case that we thought was a one-off phenomenon unlikely to be repeated. In our research on the 1991 Gulf War, we understood the case of Israel to be anomalous. Israel deeply shared the U.S. goal of evicting Iraq from Kuwait and reducing Iraq’s military capabilities, but as Israel understood any contribution it made would disrupt the participation of Saudi Arabia, Syria, Egypt, and others in the Desert Storm coalition, Israel contributed by not contributing. Even more interesting, Israel endured SCUD missile strikes from Iraq during the war but did not retaliate directly in order to help keep the Desert Storm coalition intact. Israeli leaders made it clear at the time that in exchange they expected the U.S. to support Israel’s position strongly in any post-war negotiations with the Palestinians.
I always tell my students to look for anomalies as sites for potential theory development, and here we had a whopper: a state that contributes by not contributing, and then uses its non-contribution to try to extract concessions from the coalition leader. Yet we did not do a case study of Israel’s behavior or theorize about it more deeply because we thought the case was so unlikely, idiosyncratic, and well-explained by quirky facts of which we were already aware that it did not merit further study.

On this we were woefully wrong. Fast forward to the U.S.-led coalition in Afghanistan and the next question I ask my students when I teach this example: Who is the “Israel” in the Afghan coalition, the country whose contribution would have most disrupted the contributions of another key member or members of the coalition? The answer, clearly, is India, as any contribution of troops from India would have vastly complicated the already problematic Pakistani assistance that is vital to U.S. efforts in Afghanistan. Like Israel, India used its non-contribution to try to extract concessions from the United States. India rather unsubtly let it be known that if the U.S. was not able to successfully pressure Pakistan to sharply reduce its supports of militants in the disputed Kashmir region, India would feel compelled to move more troops into Kashmir as part of the regional push against Islamic militants. This would drive Pakistan to move its troops from the tribal areas in Pakistan’s north, where they were fighting against Al Qaeda and Taliban fighters in areas bordering Afghanistan, to Pakistan’s southern border along the Kashmir region (Cody 2001; Coll 2008). Such a redeployment of Pakistan’s forces, of course, would have made it harder for the U.S. to establish security in Afghanistan. For this reason, the U.S. responded to India’s threatened “contribution” by pressuring Pakistan to cut off its support for militants in Kashmir.

India’s behavior shows that the case of Israel in our initial study was exactly the kind of potentially fruitful anomaly that I urge my students to prize. On first glance, the case of Israel looks explicable but so unusual that it is unlikely to generalize to many other cases. On closer inspection, however, the case of Israel reveals a counterintuitive outcome that is explicable in terms of a perfectly logical and highly generalizable mechanism: In many coalitions, there are one or more potential coalition partners who would cause a net loss to the strength, size, unity, or decisiveness of the coalition if they joined it because their disagreements or adversarial relations with existing coalition members would cause these members to drop out of the coalition or create debilitating arguments within it. This applies not only to international coalitions, but to domestic political coalitions, where it helps explain the awkward relations between, for example, the Tea Party and the Republican Party. The Tea Party wants to leverage its power by winning policy concessions from the Republican Party, and it has demonstrated its ability to win Republican primary nominations for Tea Party candidates who have then gone on to lose general elections in districts or states that Republicans hoped to win. Republicans, on the other hand, want to co-opt Tea Party supporters but worry that reaching out to or formally endorsing the Tea Party will alienate voters and elites who currently support the Republican Party. By failing to identify and theorize about the mechanisms related to such problematic potential coalition members, we missed a theory-building opportunity worthy of a dissertation or book.

The subsequent development of the security coalition literature reveals a number of additional improvements on our work and raises interesting questions that we had not thought to ask. These include: detailed and nuanced studies of the domestic politics of coalition contributions (Baltrusitis 2010; Spiezo 2008; Tago 2005); studies of how an increase in coalition members with a voice in decision-making complicates military operations (Auerswald, 2004); analysis of how security threats and national prestige affected allied contributions to every major U.S. led coalition from Vietnam in the 1960s to Iraq in 2003 (Davidson 2011); exploration of why states leave security coalitions (Tago 2009); examination of why states channel coercion through international organizations (Thompson 2009); and a study of why Britain joined the U.S. coalition in Iraq but not Vietnam (Dyson 2007). Many, but not all, of the authors of these works cited our study and were perhaps partly inspired by it to pose the questions they raised, but of course I only wish we had the perfect foresight to identify all these questions as the interesting next steps in the research program on security coalitions.

Still, despite these and other imperfections, I am satisfied that our earlier research improved upon the literature extant at that time, provided policy-relevant generalizations that got most of the big contributions (and non-contributions) to subsequent coalitions mostly right, and helped inspire subsequent research into related questions. These important and worthwhile contributions are not perfect, but they are good enough for me.

Notes

1 When I teach this example, I assign students to read Bennett, Lepgold, and Unger (1994), and then in class I present the tabular data herein from the Balkan, Iraqi, and Afghan security coalitions and ask them what the 1994 article got right, what it missed, and how it could be improved upon in light of subsequent evidence. Of course, assigning the present article could discourage students from thinking hard about their own conclusions to these questions, so the present article should be seen as the beginnings of an “answer key” for teachers to use or to distribute after a class discussion of this example.

2 Bennett, Lepgold, and Unger (1994). Our later study (Bennett, Lepgold, and Unger, 1997) added a fifth theory, which argued that lessons learned from previous burden sharing episodes would affect behavior in subsequent coalitions. I leave this argument aside here for the sake of simplicity.

3 For teaching purposes, a web link at: http://www.maxwell.syr.edu/moynihan/cqrm/Newsletters/ provides more complete data, including small coalition contributions, on each coalition. These data were compiled from news accounts and web sites over a number of years for teaching purposes but they are not authoritative and I do not list the sources as would be necessary to use the data for actual research. Also, the data are snapshots of frequently changing contributions and do not represent the totality of states’ contributions over time. Nonetheless, the data work well for the pedagogical purpose of asking students to identify which cases they think do and do not fit our original theory, and why. When I teach this example, I present each
I want to take this opportunity to thank the contributors to this symposium for their astute comments, which have informed my views of these topics and—I suspect—will also inform the views of many readers. I shall not attempt to address every issue that has arisen, but rather will address only those issues that seem to bear centrally upon my thesis. These comments will be divided across three areas: case study research, large-N observational research, and the larger topic of uncertainty (encompassing both).

Case Study Research (aka Causal Process Observations, Process Tracing)

Seawright seems to be arguing that case-based research is a strong form of causal inference, on par with experimental and quasi-experimental research. Yet, he does not provide any clarifying discussion of this point or any examples to back up this claim—which, to my mind, remains dubious.

Bennett seems to agree with my general argument—that causal inference is problematic in non-experimental settings and that case-based knowledge is no more immune to confounders than messy large-N observational data.

Bennett does not share my impression of a creeping methodological perfectionism in the field of qualitative methods. One might conclude that the intellectual movement has made less headway among quals than among quants. Perhaps so. But whether enforced by members of one camp or the other, the continued marginalization of qualitative methods in the discipline (e.g., their under-representation in top general-interest journals) would seem to provide evidence of the phenomenon of which I speak. Someone is judging that case studies do not surpass the bar of causal inference.

Large-N Observational Research

Seawright, Glynn, Bennett, and I agree that lots of problems attend any attempt to infer causation from large-N observational data, and that regression is not the deus ex machina that some members of the discipline once thought it was (or seemed to think it was). We agree that problems may be due to under-specified models as well as over-specified models. We agree that observational studies often suffer from ambiguity due to the non-manipulable nature of the treatment (as Glynn discusses). We agree that observational studies depend on a great many—often quite heroic—assumptions about the data generating process, assumptions that cannot always be directly tested but that nonetheless must be justified. We agree that observational studies often over-state the certainty of their findings. We agree—to quote from Seawright’s concluding sentence—that “regression-type studies of observational data have been significantly overvalued and overrepresented in our history over the last several decades,” or at least given

References


obtain a strong causal inference. But I realize that this is a
be causal——rather than the number of assumptions required to
descriptive relationships or relationships that are presumed to
according to the role that it plays——either as evidence of
animal? My personal preference would be to classify data
ambition to infer causality. What do we call that species of
guities, since some research really is descriptive and has no
odological world then I will happily place large-N observa-
tion)? I believe this residual category is quite large. Seawright
believes it is quite small, or at least that it will become so over
time, as political scientists apply their ingenuity to recalcitrant
questions.

Now we come to another ambiguity. Seawright finds that
regression’s strengths are descriptive and exploratory. He be-
lieves it to be…

a powerful tool for the summary of complex cross-tabula-
tions and scatter plots. Regression can sometimes make
consistent but small descriptive relationships among vari-
ables more visible and can often dramatically aid compre-
hension of central themes in data by replacing an over-
whelming mass of numbers or dots with a few key esti-
mates… Furthermore, there are certain moments when
one or another piece of descriptive knowledge has strong
causal implications; in such instances, regression may
sometimes play a pivotal role in a causal argument.

In one way of thinking, all evidence is descriptive, from which
the researcher must infer causality (Achen 1982: 77–78). How-
ever, some writers prefer to distinguish between evidence that
is causal—in the sense that one can infer causality from it
without making too many assumptions (e.g., from a well-de-
signed experiment)—and evidence that is descriptive or ex-
ploratory, in the sense that many assumptions are required to
infer causality. If this is the way one likes to dice up the meth-
odological world then I will happily place large-N observa-
tional data analysis into the latter category.

Of course, this way of defining things creates other ambi-
guities, since some research really is descriptive and has no
ambition to infer causality. What do we call that species of
animal? My personal preference would be to classify data
according to the role that it plays——either as evidence of
descriptive relationships or relationships that are presumed to
be causal——rather than the number of assumptions required to
obtain a strong causal inference. But I realize that this is a
matter of taste.

The point of importance is that large-N observational
data serves as the primary empirical basis for many extant
causal inferences. Calling this data descriptive or exploratory
cannot take that away; it merely calls attention to the ques-
tionable status of the evidence, a message I fully endorse. In
any case, Seawright seems to move closer to my own position
in the concluding section of his essay.

With respect to the specific question of development and
democracy, I don’t claim that this relationship is robust to
every possible gyration of the data, i.e., every plausible rob-
bustness test. Seawright, explicating Acemoglu, Johnson,
Robinson, and Yared (2008), offers the intriguing speculation
that the time-honored relationship reflects not economic de-
velopment per se but rather each country’s long-term histori-
cal position in the global economic hierarchy. It is not clear
what causal mechanisms are operative here, or how they would
differ from those associated with economic development. In
any case, it’s an interesting idea and certainly proves the point
that observational data about macro-historical phenomena are
open to a wide range of interpretations. (Indeed, even those
who agree that development causes democracy often disagree
on their explanation for the relationship.) I still find the stan-
dard explanations more plausible.2 But I must quickly add that
this is not a question I have worked with myself, so I don’t
have an intimate feel of the data (as Seawright does). Nor do I
have a dog in the fight, so to speak.

However, the methodological point at stake is the follow-
ing: if we want to know about the relationship between devel-
opment and democracy, isn’t regression (or some other large-
N observational tool) a useful tool—and quite possibly the
most useful of all available tools? It is telling that Seawright
(following Acemoglu et al 2008) resorts to regression in his
attempt to deconstruct the standard wisdom and to suggest a
new hypothesis. This suggests that he is learning something
from the regression analysis in Table 1 (and he admits as much
in his later discussion, quoted above).

Uncertainty

There are many possible answers to what Popper called
the demarcation problem—distinguishing scientific from non-
scientific arguments. Among these possible answers, I would
propose that validity and certainty are not nearly as important
as reasonable estimates of uncertainty.

To be sure, there is a common impression that scientific
studies ought to be valid more often than journalistic reports
and that science ought to adhere to a higher (more certain)
standard of truth. However, I cannot say if this is actually true
or not (in the sense of describing an empirical reality in which
the work of scientists [so-called] is compared with the work of
non-scientists). Lots of what scientists say turns out to be
wrong and a good deal of it is highly uncertain at the time of
publication. Indeed, this is what defines a frontier of knowl-
dge.

By contrast, I would propose that what distinguishes sci-
ence from non-science is the serious attention paid to uncer-
tainty estimates. We expect scientists to tell us not only what
their best guess is but also how good a guess it is likely to be. Journalists and other popular prognosticators are unlikely to dwell on this aspect of a question, at least not in a self-conscious and explicit fashion.

Let me press further on this issue, for I suspect that there is also an important distinction with respect to uncertainty between natural science and social science.

Within most natural-science fields, my sense is that there is usually general agreement about the relative (un)certainty of a finding. This stems from the fact that the methods of analysis in use within a field are fairly limited. Whether the field is experimental or non-experimental, whether it is dominated by data or by mathematical models, there are usually a few common methods that all practitioners employ, or are at least intimately familiar with. Consequently, it is usually fairly clear what standards ought to apply and what levels of uncertainty each finding implies.

Within the social sciences, our predicament is that the toolkit of available methods is, well, virtually limitless—including laboratory experiments and messy observational data, large-N and small-N samples, and so forth. Consequently, in assessing uncertainty one must span a great swath of diverse approaches to social knowledge. This constrains our ability to arrive at—and agree upon—uncertainty estimates. We have no commonly recognized metric to appeal to.

At the same time, arriving at reasonable estimates of overall uncertainty—ones that all practitioners can agree upon—may offer the only hope of attaining greater methodological unity across the diverse methods and disciplines of the social sciences.

Consider the following experience, which I expect many readers have shared. You are asked to review a paper or book that addresses some problem of importance to the discipline. You find the argument and evidence ingenious—the work clearly qualifies as a contribution to knowledge—but also highly dubious. Worse, in his/her effort to convince, the researcher has assumed the guise of a lawyer arguing a case, offering all the reasons why his/her argument might be true and none of the reasons why it might be false. In this situation, my willingness to endorse publication of the manuscript rests on an explicit and reasonable estimate of uncertainty. If the author is willing to oblige, I am willing to open the door. If not, then not.

This brings me to a central point of critique. The discipline is too focused on getting estimates right and not focused enough on getting estimates of uncertainty right. Note that we may never agree on whether democracy causes growth. But we should be able to agree that any causal inferences on this matter are highly uncertain. This, itself, is vital for the progress of the field. And it should allow for new studies of democracy and growth to appear (in top journals), without giving a misleading impression of certainty to unwary policymakers.

Of course, arriving at reasonable uncertainty estimates is not an easy task. Quantitative work with messy observational data is rightly criticized for equating t statistics and confidence intervals with the real (overall) uncertainty of an argument. Qualitative work is not prone to this error. But it is prone to an equally grave problem: uncertainties are often un-addressed, or not explicitly addressed, and there is no standard format for assessing and comparing uncertainty across studies.

Extreme bounds offers an intriguing possibility. However, as Glynn points out, the bounds identified as extreme may be too broad to be of much use. Further assumptions may be required in order to narrow the bounds of a causal proposition to levels that are informative, as discussed at the end of Glynn's essay. And yet doing so brings us back to the central problematic: causal inference often requires assumptions that are not directly testable.

Notes

1 Bennett et al (2003) find qualitative work well represented in the pages of some leading journals, such as Comparative Political Studies, International Organization, and World Politics, but their data indicate that qualitative work is infrequent in the top general-interest journals, including the American Political Science Review, the American Journal of Political Science, and the Journal of Politics, and in articles in the American Government sub-field.

2 With respect to the models shown in Table 1, I wonder about what is being tested here. Recall that in order to be meaningful, a regression model must replicate some actual data-generating process. What, then, is the process by which a change in (a) GDP rank and (b) GDP residual leads to regime-change (or stasis)? If this question cannot be answered, then the regression model is nonsensical. And if these independent variables do not resemble actual (real-life) interventions then there is no way of interpreting them in a causal fashion. By contrast, the attempt to distinguish between various elements of GDPpc, or elements that the GDPpc term might be proxying for, makes a lot of sense—so long as each element is interpretable in a causal fashion. Thus, scholars have examined the relationship between GDP (without the per capita denominator), population density, urbanization, education, infrastructure, and other modernization variables on the one hand, and regime-type on the other. But these efforts have not, to my knowledge, dethroned the empirical status of GDPpc; they have merely clarified some of the elements within the modernization rubric that might be carrying the causal burden.

References