Smith, Rogers M. 2003. "Reconnecting Political Theory to Empirical Inquiry, or A Return to the Cave?" In *The Evolution of Political Knowledge: Theory and Inquiry in American Politics*. Edward D. Mansfield and Richard Sisson, eds. (Ohio State University Press), 60–88.

Summers, Lawrence H. 1991. "The Scientific Illusion in Empirical Macroeconomics." *The Scandinavian Journal of Economics* 93:2 (June), 129–148.

Treisman, Daniel. 2007. "What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research?" *Annual Review of Political Science* 10, 211–244.

Western, Bruce. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39:3, 786–817.

Young, Cristobal. 2009. "Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth." *American Sociological Review* 74:3 (June), 380–397.

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, and one far more suited to describing initial facts, discovering puzzles, and characterizing patterns than to causal inference. I then argue that the importance of difficult-to-research questions is a weak defense of the status quo, and I conclude by briefly sketching the valuable but carefully delimited role that regression should play in our research.

The Trouble with Unstable Results

Gerring helpfully discusses the range of studies showing that cross-national regressions, in particular, routinely show unstable patterns of results across plausible specifications regarding the most important relationships. While this difficulty is by now well known, two points are nonetheless worth emphasizing.

First, instability invites the tempting but unjustified inference that the true relationship of interest is weak. Gerring's commentary provides an example of this interpretation:

> …if there were a reasonably strong (and therefore practically and theoretically relevant) causal relationship between democracy and growth, one would expect it to appear in cross-national empirical tests and to be at least somewhat stable across various (plausible) robustness tests.

While this expectation makes intuitive sense, it is nonetheless unreliable. With respect to the relationship between democracy and growth, the range of models which scholars have regarded as potentially credible produce results ranging from substantively and statistically significant negative effects to similarly significant positive effects (Seawright 2010). Within this range of results, there is no special reason to believe that the truth lies in the middle. It might instead be the case that the largest negative estimate produced to date in fact reflects the causal truth; or, perhaps, a very positive estimate corresponds with the correct model. If one model captures the structure of the data-generating process, or one estimate is correct, then all the others are incorrect and irrelevant. Instability across accepted specifications thus should not be seen as providing evidence that the true relationship is weak. Such instability only provides evidence that our consensus about how to write down regression models is weak.

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

Table 1: Two Simple Models of Democracy and Development

	Model 1 Estimates (P Values)	Model 2 Estimates (P Values)
Intercept	-46.6 (< 0.01)	-13.3 (< 0.01)
Logged Per Capita GDP	5.6 (< 0.01)	
GDP Rank		0.2 (< 0.01)
GDP Residual		8.5 (0.59)
R^2	0.16	0.16
N	128	128

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

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-

ables 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 part 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;[1] 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

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared. 2008. "Income and Democracy." *American Economic Review* 98:3, 808–842.

Berk, Richard A. 2003. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

Bhavnani, Rikhil R. 2009. "Do Electoral Quotas Work After They are Withdrawn? Evidence from a Natural Experiment in India." *American Political Science Review* 103:1, 23–35.

Brady, Henry E. and John E. McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Presented at the Annual Meeting of the Society for Political Methodology, Palo Alto, CA.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Journal of Conflict Resolution* 47:1, 72–93.

Cole, Stephen R., Robert W. Platt, Enrique F. Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. 2010. "Illustrating Bias Due to Conditioning on a Collider." *International Journal of Epidemiology* 39:2, 417–420.

Corstange, Daniel. 2010. "Vote Buying Under Competition and Monopsony: Evidence from a List Experiment in Lebanon." Presented at the Annual Conference of the American Political Science Association, Washington, D.C.

Dunning, Thad. 2010. "Design-Based Inference: Beyond the Pitfalls of Regression Analysis?" In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 273–311.

Dunning, Thad and Lauren Harrison. 2010. "Cross-Cutting Cleavages and Ethnic Voting: An Experimental Study of Cousinage in Mali." *American Political Science Review* 104:1, 21–39.

Freedman, David A. 1997. "From Association to Causation via Regression." *Advances in Applied Mathematics* 18:1, 59–110.

Freedman, David A. 1999. "From Association to Causation: Some Remarks on the History of Statistics." *Statistical Science* 14:3, 243–258.

Freedman, David A. 2004. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40:2, 180–193.

Humphreys, Paul and David A. Freedman. 1996. "The Grand Leap." *British Journal for the Philosophy of Science* 47:1, 113–123.

Humphreys, Macartan and Jeremy M. Weinstein. 2009. "Field Experiments and the Political Economy of Development." *Annual Review of Political Science* 12, 367–378.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14:2, 131–159.

Mahoney, James. 2010. "After KKV: The New Methodology of Qualitative Research." *World Politics* 62:1 (January), 120–147.

Morgan, Stephen L. and Christopher Winship. 2006. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer.

Seawright, Jason. 2010. "Regression-Based Inference: A Case Study in Failed Causal Assessment." In *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd ed. Henry E. Brady and David Collier, eds. (Lanham, MD: Rowman & Littlefield), 247–271.

Stigler, Stephen M. 1990. *The History of Statistics: The Measurement of Uncertainty Before 1990*. Cambridge: Belknap Press of Harvard University Press.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55:3, 399–422.

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