

THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

Research Paper No. 9

How Factual is Your Counterfactual?

**Gary King
Lanche Zeng**

October 2001
ISSN 0000 0000

**HARVARD BURDEN OF DISEASE UNIT
NATIONAL INSTITUTE ON AGING GRANT 1-P01-AG17625**

THE GLOBAL BURDEN OF DISEASE 2000 IN AGING POPULATIONS

This working paper series reports on research supported by the National Institute on Aging program grant entitled The Global Burden of Disease 2000 in Aging Populations (1-P01-AG17625). The purpose of the grant is to strengthen the methodological and empirical bases for undertaking comparative assessments of health problems, their determinants and consequences in aging populations.

Since the publication of the Global Burden of Disease Study 1990, there has been increasing interest in comparative analyses of health outcomes, determinants and consequences. A major revision of the Global Burden of Disease Study has been launched for the year 2000 with the full commitment of the World Health Organization (WHO). The Global Programme on Evidence for Health Policy at WHO has developed a Global Burden of Disease Network, which operates in parallel to the research conducted as part of the program project. The program project will strengthen the scientific basis for the large-scale undertaking led by WHO at the global, regional and national level.

The purpose of this working paper series is to present original research that emerges from the various project components of this program grant. The views expressed in these working papers are those of the author(s) and do not necessarily reflect the views of the Harvard Burden of Disease Unit, the World Health Organization nor the National Institute on Aging.

THE HARVARD BURDEN OF DISEASE UNIT

The Harvard Burden of Disease Unit was established to design, test, and implement methodologies to aid in the effective allocation of health resources. To achieve this end, the Unit conducts research in collaboration with national governments, international agencies and other researchers and policy-makers. The Unit's research has two main foci:

- to forge the theory, design, and implementation of approaches to the combined measurement of mortality and non-fatal health outcomes, in order to develop valid, reliable, comparable and comprehensive measures of population health and comparative assessments of the burden of diseases, injuries and risk factors; and
- to investigate the costs, efficacy and effectiveness of major health interventions applied in diverse settings, toward the goal of establishing a broad database on cost-effectiveness.

Harvard Burden of Disease Unit
Center for Population and Development Studies
9 Bow Street
Cambridge, MA 02138
www.hsph.harvard.edu/organizations/bdu

How Factual is Your Counterfactual?¹

Gary King² and Lanche Zeng³

Abstract

Inferences about counterfactuals are essential for prediction, answering “what if” questions, and estimating causal effects. However, when the counterfactuals posed are too far from the data at hand, conclusions drawn from well-specified statistical analyses become based on speculation and convenient but indefensible model assumptions rather than empirical evidence. Yet, standard model outputs do not reveal the degree of model-dependence, and so this problem can be hard to detect. We develop easy-to-apply methods to evaluate counterfactuals that do not require sensitivity testing over specified classes of models. One analysis with these methods applies to the class of all models, for any smooth conditional expectation function, and to the set of all possible dependent variables, given only the choice of a set of explanatory variables. We use these methods to evaluate the extensive scholarly literatures on the effects of changes in the degree of democracy in a country (on any dependent variable), and find evidence that many scholars are inadvertently drawing conclusions about democracy based more on their hypotheses than on their empirical evidence.

¹Thanks to Jim Alt, Scott Ashworth, Neal Beck, Jack Goldstone, Orit Kedar, Walter Mebane, Maurizio Pisati, Jas Sekhon, Simon Jackman for helpful discussions and the National Science Foundation (IIS-9874747), the National Institute on Aging (PO1 AG17625-01), and the World Health Organization for research support. Software to implement the methods in this paper is available from <http://Gking.Harvard.Edu>.

² Professor of Government, Harvard University and Senior Science Advisor, Evidence and Information for Policy Cluster, World Health Organization

³ Associate Professor of Political Science, George Washington University

1 Introduction

Social science is about making inferences, using facts we know to learn about facts we do not know. Some inferential targets (the facts we do not know) are *factual*, which means that they exist even if we do not know them. Many want to know, for example, the proportion of Florida voters in the 2000 election who intended to vote for George W. Bush. Although events subsequent to the election demonstrated the difficulty of determining this number, no one seems to question its existence. In contrast, other inferential targets are *counterfactual*, and thus do not exist, at least not yet. Counterfactual inference is crucial for studying “what if” questions, such as what kind of president Al Gore would have made if the 2000 election had turned out differently. It is also crucial for making forecasts since the quantity of interest is not knowable at the time of the forecast, although it will eventually become known. Counterfactual inference is essential as well in making causal inferences, since causal effects are differences between factual and counterfactual inferences: for example, how much smaller would the 2001 federal tax reduction have been if Gore rather than Bush had been elected president.

As Lebow (2000: 558) explains, “Counterfactuals are an essential ingredient of scholarship. They help determine the research questions we deem important and the answers we find to them. They are necessary to evaluate the political, economic, and moral benefits of real-world outcomes. These evaluations in turn help drive future research.” Counterfactual inference has thus been popular topic in political science (Fearon, 1991, Thorson and Sylvan, 1982; Tetlock and Belkin, 1996; Tetlock and Lebow), psychology (Tetlock, 1999; Tetlock et al., 2000), history (Murphy, 1969, Gould, 1969; Dozois and Schmidt, 1998; Tally, 2000), philosophy (Lewis, 1973; Kvart, 1986), computer science (Pearl, 2000), statistics (Rubin, 1974; Holland, 1986), and other disciplines. As scholars have long recognized, however, some counterfactuals are more amenable to empirical analysis than others. In particular, some counterfactuals are more strained, farther from the data, or otherwise unrealistic.¹ With a sample of U.S. time series data, we could reasonably ask how much U.S. presidential approval would drop if inflation increased by two percentage points, and

¹For example, Fearon (1991) and Lebow (2000) distinguish between “miracle” and “plausible” counterfactuals and offer qualitative ways of judging the difference. Tetlock and Belkin (1996: ch. 1) also discuss criteria for judging counterfactuals (of which “historical consistency” may be of most relevance to our analysis). In this paper, we provide quantitative measures of these and related criteria that are meant to complement the ideas for qualitative research discussed by these authors.

we could generate a fairly certain answer. We should not expect to get a valid answer from the same data, given any model, if we asked how much approval would drop if inflation increased by 200 percentage points. Remarkably, however, whatever statistical model we used to compute the first counterfactual inference could also be used to compute the second. Our confidence interval for the second inference would be somewhat wider than the first, but the second inference would be far more uncertain than the confidence interval indicates. The confidence interval is not wrong: if the other assumptions of the model are correct, it accurately portrays the uncertainties, conditional on the model. The problem is that we have little reason to believe the model when it veers so far from the data. In other words, the second inference is far more *model dependent* than the first. Thus, for example, if the model used were a linear regression, the 2% inference would not depend very heavily on the exact linearity assumption as long as the model fits the observed data reasonably well and the underlying true conditional expectation function is reasonably smooth. However, the 200% inference is so far from historical experience that any conclusion will be very sensitive to most features of the model, no matter how good the fit of the model to the data. The linearity assumption in this context thus looms very large. Extrapolate using a slightly quadratic function instead of a linear one, even one that fits the in sample data only slightly differently, and inferences will differ dramatically, even if little evidence exists with which to distinguish between the two models.

But how can we tell how model dependent are our inferences when the counterfactual is not so obviously extreme, or when it involves more than one explanatory variable? The answer to this question does not come from any of the model-based quantities we normally compute and our statistics programs typically report, such as standard errors, confidence intervals, coefficients, likelihood values, predicted values, test statistics, first differences, p-values, etc. To understand how far from the facts are our counterfactual inferences, and thus how model-dependent are our inferences, we need to look elsewhere. At present, scholars study model-dependence primarily via sensitivity analyses: changing the model and assessing how much our conclusions change. If the changes are substantively large for models in a particular class, then inferences are deemed model dependent. If the class of models examined are all a priori reasonable, and conclusions change a lot as the models within the class change, then the analyst may conclude that the data contain little or

no information about counterfactual question at hand. This is a fine approach, but it is insufficient in circumstances where the class of possible models cannot be easily formalized and identified, or where the models within a particular class cannot feasibly be enumerated and run, i.e., most of the time. In practice, the class of models chosen are those that are convenient — such as those with different control variables under the same functional form. In practice, the identified class of models rarely includes all possibilities, and it normally excludes at least some models that have a reasonable probability of returning different substantive conclusions. In contrast, our analysis applies to the class of nearly all models, whether or not they can be formalized, enumerated, and run, and for the class of all possible dependent variables.

In addition to the methodological contributions that would appear applicable to most empirical analyses, the substantive goal of this paper is to evaluate the counterfactuals used in the scholarly literatures on the effects of democracy. These effects (on any of the dependent variables used in the literature) have long been among the the most studied questions in comparative politics and international relations. Our results demonstrate that many scholars in these literatures are asking counterfactual questions that are far from their data, and are therefore inadvertently drawing conclusions about the effects of democracy based on indefensible model assumptions rather than empirical evidence. The results also show that almost all analyses about democracy include at least some counterfactuals with little empirical support.

Section 2 shows more specifically how to identify questions about the future and “what if” scenarios that cannot be answered well in given data sets. This section introduces several new approaches for assessing how based in factual evidence is a given counterfactual. Section 3 provides a new decomposition of the bias in estimating causal effects using observational data that is more suited to the problems most prevalent in political science. This decomposition enables us to identify causal questions without good causal answers in given data sets and shows how to narrow these questions in some cases to those that can be answered more decisively. We use each of our methods to evaluate counterfactuals regarding the effects of democracy. Section 4 concludes.

2 Forecasts and “What If” Questions

Although statistical technology sometimes differs for making forecasts and estimating the answers to “what if” questions (e.g., Gelman and King, 1994), the logic is sufficiently similar that we consider them together. Although our suggestions are general, we use aspects of the international conflict literature as a running example to fix ideas. Thus, let Y , our outcome variable, denote the degree of conflict initiated by a country, and X denote a vector of explanatory variables, including measures such as GDP and democracy.² In regression-type models, including least squares, logit, probit, event counts, duration models, and most others used in the social sciences, we usually compute forecasts and answers to “what if” questions using the model-based conditional expected value of Y given a chosen vector of values x of the explanatory variables, X .

The model typically includes a specification for (i.e., assumption about) the conditional expectation function (CEF):

$$E(Y|X) = g(X, \beta), \tag{1}$$

where $g(\cdot)$ is some specified functional form and β is a vector of parameters to be estimated. To make a forecast, we plug the specified vector of values x , and the point estimate of β , which we denote $\hat{\beta}$, into this CEF and compute the estimated CEF:

$$\hat{E}(Y|x) = g(x, \hat{\beta}). \tag{2}$$

The estimated CEF for the familiar linear regression model, for example, is $x\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \dots + \hat{\beta}_kx_k$; the logit model is $[1 + e^{x\hat{\beta}}]^{-1}$; the exponential duration model is $e^{-x\hat{\beta}}$; count models are usually specified as $e^{x\hat{\beta}}$, etc.³

Interestingly, each of these CEFs can be computed for *any* (real) values of x . The model never complains, and exactly the same calculation can be applied for any value of x . However, even if the model fits the data we have in our sample well, a vector x far from any rows in the matrix X is not likely to produce accurate forecasts. If a linear model indicates that one more year of education will earn you an extra \$1,000 in annual income,

²For notational convenience we also use Y and X to denote the observed data matrix of the dependent and explanatory variables when the context is clear.

³More generally, we are interested in the full conditional density, $P(Y|x) = \int P(Y|x, \beta)P(\beta|Y)d\beta$. All our methods apply in this situation as well, but for expository purposes we continue to focus on the CEF in Equation 1. Equation 2 is included only to fix ideas, since our methods do not require an estimate of the CEF or a specification of the model. Normally a better way of computing the estimated CEF recognizes the uncertainty in $\hat{\beta}$: $\hat{E}(Y|x) = \int g(x, \hat{\beta})P(\hat{\beta})d\hat{\beta}$, where $P(\hat{\beta})$ is the posterior density of β .

the model also implies that 10 more years of education will get you \$10,000 in extra annual income. In fact, it also says — with as straight a face as a statistical model ever offers — that 50 years more of education will raise your salary by \$50,000. Even though no statistical assumption may be violated as a result of your choice of any set of real numbers for x , the model obviously produces better forecasts (and “what if” evaluations) for some values of x than others. Predictive confidence intervals for forecasts farther from the data are usually larger, but confidence intervals computed in the usual way still assume the veracity of the model no matter how far the counterfactual is from the data.

If you are worrying about your choice of a model, that may be a good thing in general, but it will not help here. Other models will not do verifiably better with the same data; we simply cannot determine from the evidence which model is more appropriate. So searching for a better model, without better data, better theory, or a different counterfactual question, in this case is simply futile. We merely need to recognize that some questions cannot be answered from some data sets. Our linearity (or other functional form assumptions) are written globally — for any value of x — but in fact are relevant only locally — in or near our observed data. In this paper, we seek to understand where “local” ends and “global” begins. For forecasting and analyzing what if questions, our task comes down to seeing how “far” x is from the observed X .

We now offer several procedures that can be used to assess whether a question posed can be reliably answered from any statistical model. We will not assume knowledge of the functional form, model, estimator, or dependent variable, but we do assume that the CEF is fairly smooth (a concept we define formally below). Such is not always the case, but the idea comports with most statistical models and theoretical processes in the discipline.

2.1 Interpolation vs. Extrapolation

An important distinction in assessing the counterfactual question x is whether answering it by computing $E(Y|x)$ would involve interpolation or extrapolation (e.g., Kuo, 2001). Assuming some minimal smoothness in the CEF in the region near the data implies that the data would in general contain more information about counterfactual questions that require interpolation than those that require extrapolation. Hence answering a question involving extrapolation is normally far more model-dependent than one involving interpolation.

For intuition, imagine we have observed data on income for those with high school de-

degrees and for those with four-year college degrees, and we wished to estimate what income one with a two-year college degree would have. This is a simple counterfactual “what if” question because we have no data on people with two-year degrees. The interpolation task, then, is to draw some curve from expected income among high school graduates to the expected income among four-year college graduates. Without any assumptions, this curve could go anywhere, and our inferred income for those with two-year degrees would not be constrained at all. Imposing the assumption that the CEF is “smooth” (i.e., that it contains no sharp changes of direction and that it not bend too fast or too many times between the two end points, a concept we formalize below) is quite reasonable for this example and indeed for most political science processes. The consequence of this smoothness assumption is to narrow greatly the range of income into which the interpolated value must fall. Even if it is higher than income for four-year colleges or lower than for high school educated students, it probably won’t be too much outside this range. However, now suppose we observed the same data but needed to extrapolate to the income for those with masters degrees. We could impose some smoothness again, but even one bend in the curve could make the extrapolation change a lot more than the interpolation. One way to look at this is that the same level of smoothness (say the number of changes of direction allowed) constrains the interpolated value more than the extrapolated value, since for interpolation any change in direction must be accompanied by a change back to intersect the other observed point. With extrapolation, one change need not be matched with a change in the opposite direction, since there exists no observed point on the other side of the counterfactual being estimated.

Thus, as a first step in evaluating a counterfactual question, we need to know whether the question requires interpolation or extrapolation. If it involves extrapolation, we still might wish to proceed if the question is sufficiently important, but we would be aware of how much more model dependent our answers will be. How to do this with one variable should now be obvious; fortunately, doing it for more than one requires only one additional concept: Indeed, our main message in this section is that, mathematically, questions that involve interpolation are values of the vector x which fall in the *convex hull* of X , and with this definition (and our explanations below) researchers can easily check whether a particular question requires extrapolation.⁴

⁴An alternative formal definition of extrapolation is an inference that occurs at x that is off the *sup*-

Formally, the convex hull of a set of points is the smallest convex set that contains them.⁵ This is easiest to see graphically, such as via the example in Figure 1 for one explanatory variable (on the left) and for two (on the right), given simulated data. The small vertical lines in the left graph denote data points on the one explanatory variable in that example. The convex hull for one variable is marked by the maximum and minimum data points: any counterfactual question between those points requires interpolation; points outside involve extrapolation. (The left graph also includes a nonparametric density estimate, a smooth version of a histogram, that gives another view of the same data.) For two explanatory variables, the convex hull is given by a polygon with extreme data points as vertices such that for any two points in the polygon, all points that are on the line connecting them are also in the polygon (i.e., the polygon is a convex set). A counterfactual question x that appears outside the polygon requires extrapolation. Anything inside involves interpolation.

Although Figure 1 only portrays convex hulls for one and two explanatory variables, the concept is well defined for any number of dimensions. For three explanatory variables, and thus three dimensions, the convex hull could be found by “shrink wrapping” the fixed points in three dimensional space. The shrink wrapped surface encloses counterfactual questions requiring interpolation. For four or more explanatory variables, the convex hull is more difficult to visualize, but we can still easily assess whether a point lies within the hull based on the mathematical property of the hull. Appendix A shows how to do this by solving a standard linear programming problem, meaning that generally available linear programming software can be employed to solve the problem quickly.

We can also describe the advantages of interpolation over extrapolation more formally.

port of the distribution that generated X , so that there is zero probability of having observations within some neighborhood of x in repeated sampling. Manski (1995: 16) uses this definition and shows that for continuous functions, interpolation enables nonparametric identification of the conditional distribution $P(Y|X = x)$ (and therefore the CEF at x) while extrapolation requires additional assumptions. Unfortunately, estimating the support of the distribution of X , $P(X)$, from sample data is difficult or infeasible for more than a few explanatory variables, and so in this section we focus on our definition using the convex hull which leads to easy verification. In Section 3, however, where we discuss counterfactuals in causal inference that have a more specific form, we show that checking whether a counterfactual falls on the support of the X distribution is also easily implementable. For finite samples, the convex hull of X and the support of $P(X)$ are closely but qualitatively related, there is no universal quantitative relationship. Asymptotically, the convex hull either equals the support or contains it as a subset and so can be seen as a conservative approach.

⁵A set is convex if for any two elements in the set, the convex combinations of them are also in the set. A point is a convex combination of two other points if it lies on the line segment between the two points, i.e., it is a linear combination of the two points with coefficients that are each between zero and one and together sum to one.

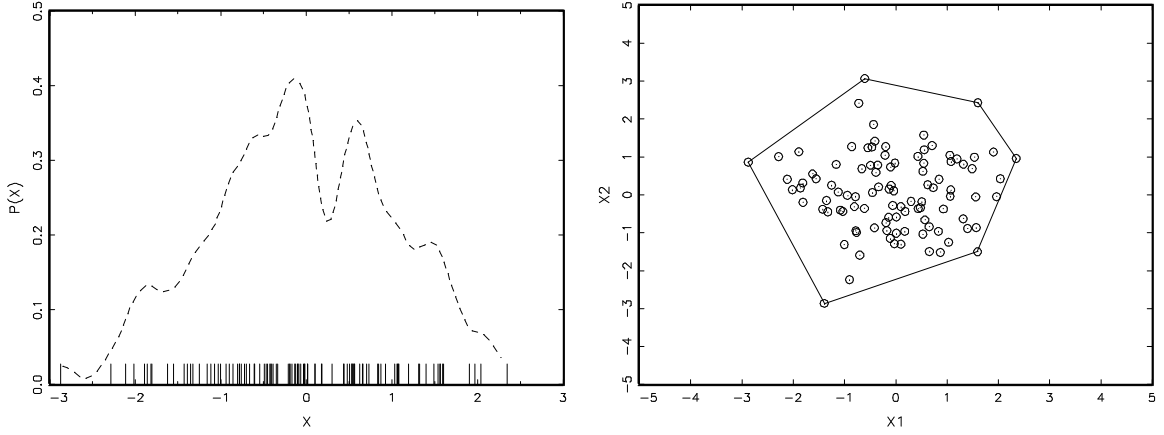


Figure 1: *Interpolation vs. Extrapolation: The convex hull of X is the smallest convex set that contains the data. Inference on points inside the convex hull requires interpolation, outside it requires extrapolation. With one explanatory variable, the convex hull is the closed interval between the minimum and the maximum values of the observed data (as portrayed on the left graph, along with a nonparametric density). With two explanatory variables, the convex hull is a polygon with vertices at the extreme points of the data (as in the right graph).*

If we make no assumptions about the true CEF, then we can learn nothing about counterfactuals other than those that are coincident with the data points X (the “factuals”). Counterfactual inference thus requires some type of additional assumption, which almost always involves a version of smoothness in the CEF. For our more general purposes, we shall only require that the CEF is smooth in the precise sense that it satisfies the Lipschitz condition on a convex set containing both the observed data X and the counterfactual x , so that for any two points in this convex set — in particular, the counterfactual x and an observed point $X_i \in X$ —

$$\|E(Y|x) - E(Y|X_i)\| \leq K\|x - X_i\|, \quad (3)$$

where K is some fixed positive number. Having bounded derivatives on the convex set would be sufficient for this condition to hold per the Mean Value Theorem, but not necessary. Under this condition, the “distance” (defined by the norm) between the CEF at the counterfactual $E(Y|x)$ and the observed data point corresponding to X_i , $Y_i = E(Y|X_i) + e(X_i)$, is bounded:

$$\begin{aligned} \|E(Y|x) - [E(Y|X_i) + e(X_i)]\| &\leq \|E(Y|x) - E(Y|X_i)\| + \|e(X_i)\| \\ &\leq K\|x - X_i\| + \|e(X_i)\|, \end{aligned} \quad (4)$$

where we denote $e(X_i)$ as the (bounded) random error in the outcome variable at point X_i , and $\|x - X_i\|$ as the distance between x and X_i . This result means that the error in using the data point at X_i (i.e. $Y|X_i$) to approximate the CEF at the counterfactual point x depends on the distance from the counterfactual x to the observed X_i . Equation 4 also implies that one such bound exists for every $X_i \in X$, and it is easy to see that collectively the bounds are more informative about $E(Y|x)$ if x is in the convex hull of X than if it is outside simply because, in most cases, more points are nearby. Of course, we might hesitate to assume anything about the CEF outside of the observed data domain that is the convex hull of X only, in which case no bound is available for extrapolation at all, meaning the observed data contain no information about the CEF at that point.

Outliers in X would make this extrapolation-detection method inefficient. That is, some questions x would be identified as requiring interpolation even though they would really involve extrapolation. However, if this method identified a question as requiring extrapolation, then the suspicion of outliers in X would only make the finding more solid. Of course, outliers need to be identified before, or as part of, any estimation strategy or the model applied will not be reasonably estimated, even if the counterfactual question asked is close to the data. As such, the problem of outliers is not unique to the issue at hand.

While we do not need to specify any further details about the CEF (i.e., except for smoothness in the form of meeting Equation 3) to establish the bound in Equation 4, and thus the relative informational advantage of interpolation over extrapolation everywhere in this very general class of CEFs, additional knowledge of the CEF can inform us more about the absolute quality of any inference. In particular, if assumptions can be added to narrow further the range of functions the CEF can take on, then this information can be used to narrow the bound on the error in approximating the true CEF from the specific models (see, for example, Wu and Schaback, 1993; Madych and Nelson, 1992; and Shaback, 1996). To our knowledge, all such model-specific bounds are related to the distance between the counterfactual and the observed data, too, which reinforces our model-independent conclusion about the advantage of interpolation over extrapolation.

Similarly, if specific forms of nonlinearity and interactions (such as squared terms or products of existing variables) are known to be present in the CEF, explicitly using them

as additional input variables will result in a CEF that is a smoother function of the larger set of inputs. As a result, counterfactual inference within the convex hull of this larger data matrix X will be more accurate, and so the results of the test we propose will be more informative about the approximation error in making interpolations.⁶ However, researchers should not include these extra terms in their input data X unless they know they belong in the CEF: putting them in when they do not belong could cause one to conclude incorrectly that a counterfactual requires extrapolation.

While the informational advantage for interpolation holds in general, if we are willing to assume that the CEF is smooth in the neighborhood of observed data, then points just outside the convex hull are arguably “less of an extrapolation” than those farther outside. Another related issue is that it is possible for a point just inside the extrapolation region (i.e., just outside the convex hull of X) to be closer to a large amount of data points than one inside the hull that occupies a large empty region away from most of the data. These qualifications — as well as the need to evaluate counterfactuals within the interpolation region — point to the need for other methods that could be used simultaneously, a problem to which we now turn. But even with the need for other methods, we view the distinction between interpolation and extrapolation as sufficiently fundamental that we think readers of all articles making counterfactual inferences have the right to know which type is involved.

2.2 How Far Is The Counterfactual from the Data?

In addition to assessing whether a counterfactual question requires interpolation or extrapolation, we can further examine how “far” it is from the data by computing various distance or closeness measures. For example, one such measure could be derived based on parametric approaches. If we assume that X is generated from some super-population with multivariate probability density $P(X)$, then we could ascertain whether there exists sufficient hyper-volume in the region near x to base a reasonably reliable inference. If R is the region near x defined in some way, then we might calculate the fraction of the probability density $P(X)$ that falls within R as the integral $\int_{x \in R} P(X) dX$. If this probability mass around x is too small, then we might conclude that there does not exist enough

⁶Identification of “features” of the original data such as squared terms or interactions that may be present in the CEF is part of the routine data “pre-processing” step that political scientists often perform. For a rigorous treatment of this topic see, for example, Bishop (1995: Chapter 8).

empirical evidence nearby x to make a reasonable inference.

The question then is what distribution we should assume for $P(X)$ and how to estimate it. Since X can be quite high dimensional (i.e., there might be many explanatory variables), and X may have scattered missingness, a simple parametric approach would be to use the same model for missing data as King, Honaker, Joseph, and Scheve (2001), which is multivariate normal on the fully observed data. If data are missing, the multivariate normal is estimated using all available information, by integrating out the missing cell values. One advantage of this approach is that we can use the software written for this purpose, Amelia (Honaker, Joseph, King, and Scheve, 2001), to accomplish our task. Amelia can be used to estimate the mean vector μ and the covariance matrix Σ of the multivariate normal distribution, and then the integral above can be computed.⁷

Although as with imputation problems, this parametric approach may be a reasonable approximation in many situations, its disadvantage is that a multivariate normal may be too much of a simplification, especially with many dichotomous variables. Also, sampling designs that stratify on some variables in X may contain no information on aspects of $P(X)$ without adding auxiliary information. We now address both issues by using a distance measure based on Gower’s (1971) nonparametric measure of similarity, which allows mixed data types. Gower’s measure of similarity between two points x_i and x_j in K dimensional space is:

$$g_{ij} = \frac{1}{K} \sum_{k=1}^K \left(1 - \frac{|x_{ik} - x_{jk}|}{r_k} \right) \quad (5)$$

where x_{ik} is element k of x_i , x_{jk} is element k of x_j , and r_k is the range of the k th element and, applied to our problem, is

$$r_k = \max(X_{.k} \cup x_{.k}) - \min(X_{.k} \cup x_{.k}). \quad (6)$$

where the min and max functions return the smallest and largest elements respectively in the set including the k th elements of the counterfactual x and explanatory variables X . Thus, the elements of the measure are normalized for each variable to range between zero and one, and then averaged. The measure is designed to apply to all types of variables, including both quantitative and qualitative data.

⁷Or, instead of the integral, which can be complicated to compute, we could use the log posterior density evaluated at the observed value x . A simple version of this would be to plug in the estimates, $\ln P(x|\hat{\mu}, \hat{\Sigma})$. A somewhat better version would be to average over the uncertainty in estimating the parameters $\ln \int P(x|\mu, \Sigma)P(\mu, \Sigma)d\{\mu, \Sigma\}$.

The distance measure between the points x_i and x_j we use is $G_{ij} = \sqrt{1 - g_{ij}}$, and thus

$$G_{ij}^2 = \frac{1}{K} \sum_{k=1}^K \frac{|x_{ik} - x_{jk}|}{r_k}. \quad (7)$$

Gower (1971) shows that this measure also satisfies the triangle inequality

$$G_{ij} + G_{ik} \geq G_{jk}, \quad (8)$$

and hence has a metric interpretation.⁸

The problem with the usual uses of the Gower distance measure in our context is that if data are collected by stratifying on X , the minimum and maximum values in (6) may be misleading. For example, suppose education is one of the variables and is only measured for 1, 2, 3, or 4 years of college education, and assume for the moment that x is coded within those years. In this case, r_k will be observed at 3. But the full range of years of education might instead be seen as having a range of 15. The fact that the range of X was artificially restricted does not bias any estimates conditional on the model, but it can have a large affect on our measure of closeness of x to X . In fact, if the data collection design restricts education to this small range, then *any* distance from x to X in this situation should be a good deal narrower than the full range of 15. Our measure of similarity should reflect this: in this example, we should be normalizing by dividing by 15 instead of 3. In that way, the maximum distance would be 3/15 instead of 1, as is appropriate. Although few have followed up on this suggestion, Gower (1966: 859) recognizes this problem and allows r_k to be defined based on external information and so, although our particular empirical example in Section 2.3 is unaffected, we suggest using available sample design information.

Another reason for modifying the standard definition of r_k , of course, is that for our purposes the min and max functions in (6) are meant to summarize the data. As such, in order to avoid having the nature of the question constraining the answer, they should not include x . Thus, we modify Gower's measure by removing x from the definition and choosing r_k accordingly — from the design if X was selected by stratification or as in Equation 6 without x if X was randomly sampled.

⁸Ordinal explanatory variables are typically assumed interval or coded as a set of dichotomous variables, and Gower's measure follows that practice. Some versions also include weights, but we we exclude those here.

Since x may be outside the convex hull of X , and because we allow r_k to be set by the user from design information, the modified version of G^2 may range anywhere from zero on up. Thus, if $G^2 = 0$, then x and the row in question of X are identical, and the larger G^2_{ij} , the more different the two rows are. (If $G^2 > 1$ then x lies outside the convex hull of X , but the reverse does not necessarily hold.)

With G^2 applied to our problem, we need to summarize n numbers, the distances between x and each row in X . If space permits, we suggest presenting a cumulative frequency plot of G^2 horizontally by the fraction of rows in X with G^2 values less than the given value on the horizontal axis. If space is short, such as would happen if many counterfactuals need to be evaluated, any fixed point on this graph could be used as a one-number summary. For example, one possibility is to report the fraction of rows with G^2 values less than some specific number, such as the “generalized variance” of X .⁹ We illustrate in Section 2.3.

In concluding, we note that the methods discussed here do not exhaust the possible methods for evaluating how far x is from X . The parametric distance method analogous to G^2 is Mahalanobis distance, which underlies the normal distribution and the approach outlined earlier in this section. Any method of outlier detection could be used to see whether x is an outlier with respect to X (although note that the usual outlier detection scheme uses rows of $\{Y, X\}$, whereas we only want to look at X). Methods of multivariate cluster analysis or other high dimensional data techniques would also be useful for identifying how close x is to portions of X .

2.3 Counterfactuals About Democracy

We now apply these methods of evaluating counterfactuals to address one of the most asked questions in political science: what is the effect of a democratic form of government (as compared to less democratic forms). We study counterfactuals relating to the degree of democracy using data collected by the State Failure Task Force (Esty et al., 1995, 1998a, 1998b, 1999). The State Failure Task Force is a group of academics and consultants,

⁹The generalized variance, or squared generalized standard deviation, is the geometric variability (Cuadras and Fortiana, 1995; Cuadras, Fortina, and Oliva, 1997). The geometric variability is a generalized version of the usual variance formula in that for Euclidean distances (which are inappropriate with binary data, for example), it equals the regular variance for one explanatory variable, or in general the trace of the covariance matrix of X . For other measures of distance, such as used in Equation 7, the geometric variability is a generalized measure of dispersion of X .

selected and funded by the U.S. Central Intelligence Agency, whose task is to collect data and develop methods for forecasting state failure. State failure is the collapse of the central authority of the state, so that it can no longer impose the rule of law on its citizens. Examples include Somalia and Bosnia.

Our independent evaluation of the task force's efforts revealed some (fixable) methodological errors but a superb data set (King and Zeng, 2002). Indeed, this example is an especially good one for our present purposes because the task force authors had the resources of the federal government to marshal for their data collection efforts, and so the usual scarcity of time, resources, expertise, etc., that affect most data collection efforts are not constraints here. The only limitation on types and especially combinations of data the task force could collect, in addition to the authors' research design, was the world: that is, countries can be found with only a finite number of bundles of characteristics, and this constraint affects everyone studying counterfactuals about democracy. Thus, to the extent that we find that certain counterfactual questions of interest are unanswerable with task force data, we know that remedying the problem will require something more than merely additional resources.

After extensive searches, the task force settled on a logistic regression model to forecast state failure as a function of trade openness, (as a proxy for economic conditions and government effectiveness) the infant mortality rate, and democracy. Democracy is coded as two dummy variables representing autocracy, partial democracy, and full democracy. King and Zeng (2002) added to these the fraction of the population in the military, population density, and legislative effectiveness and showed that a prior-corrected version of a committee neural network model could forecast uniformly better than the task force's approach. The details of these alternative statistical models are not relevant here, since the methods developed in this section hold for almost any statistical model, and as such we only need a list of explanatory variables and a given counterfactual question. Thus, for present purposes, we use all these explanatory variables, and focus on a counterfactual of central importance: what would happen to the probability of state failure if the degree of democracy changed.

Since the dependent variable need not be specified for any of the techniques introduced in this section, the conclusions from one analysis applied to a given set of explanatory

variables holds for any dependent variable that might in the future be used with these variables. This is fortunate in the present application, since something close to the same counterfactual also accounts for much of the research on the democratic peace hypothesis in international relations. Although the dependent variable there is inter-state conflict rather than intra-state failure, our methods and this same analysis apply there too without modification. “What would happen if more of the world were democratic” is a question that also underlies much other work in comparative politics and international relations over the last half century as well as a good deal of American foreign policy, and so our analysis here may be of some interest to a variety of readers.

We begin a description of our empirical analyses with four clear examples, the first two obviously extrapolations and the second two obviously interpolations, and then move to averages of many other cases of more substantive interest. For example, what would happen if Canada in 1996 had become an autocracy, but its values on other variables remained at their actual values? We find that this extreme counterfactual is outside the convex hull of the observed data and therefore requires extrapolation. Similarly, if we ask what would have happened if Saudi Arabia in 1996 had become a full democracy, we would also be required to make an extrapolation, since it too falls outside the convex hull.

Conversely, suppose we ask what would happen if Poland had become an autocracy in 1990 (i.e., just after it became a democracy). From qualitative information available about Poland, this counterfactual is quite plausible, and many even thought (and worried) about it actually occurring at the time. Our analysis confirms the plausibility of this suspicion since this question falls within the convex hull; analyzing it would require interpolation and thus not much model dependence. Another reasonable counterfactual is to ask what would have happened had Hungary become a full democracy in 1989 (i.e., just before it actually did become a democracy). This question is also in the convex hull and would also require only interpolation to produce specific estimates.

We now further analyze these four counterfactuals questions using our modified Gower distance measure. The question is how far the counterfactual x is from each row in the observed data set X , and so the distance measure applied to the entire data set gives n numbers. We summarize these numbers, without loss of information, in the cumulative frequency plots in Figure 2. The left plot includes counterfactuals that change from autoc-

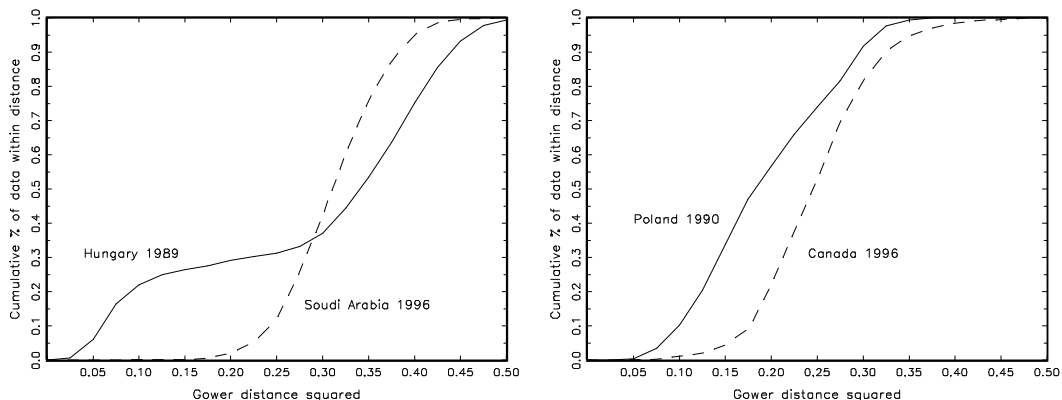


Figure 2: *Distance to Four Counterfactuals: Cumulative frequencies of modified Gower distances. Countries in the left graph are changed from autocracies to democracies, and the reverse in the right graph. Dashed lines are outside the convex hull of observed data; solid lines are within it.*

racies to democracies, and the right plot is the reverse. The dark line in each graph refers to a counterfactual within the convex hull and the dashed line is for a counterfactual outside the hull. Each line gives the cumulative distribution of our modified Gower distance square measures. Take, for example, the value of G^2 (given horizontally) of 0.1 (which is an average distance of 10% from the minimum to the maximum values on each variable in X).¹⁰ Essentially no real country-years are within 0.10 or less of this counterfactual for changing Saudi Arabia to a democracy, but about 25% of the data are within this distance for Hungary. Similarly, just a few observations in the data are within even 0.15 of Canada changing to an autocracy, although about a quarter of the country-years are within this distance for Poland. Of course, the full cumulative densities in the figure provide a lot more information than this one point.

We now examine a larger sets of counterfactuals all at once with more convenient numerical summaries along the lines of our verbal description of Figure 2. We leave all variables set at their actual values and first ask what would happen to all autocracies if they became full democracies, and to all full democracies if they became autocracies. This analysis includes 5,814 country-years, including 1,775 tracking full democracies and 4,039 representing autocracies. What we found was that only 28.4% of the country-years in this widely examined counterfactual fell within the convex hull of the observed data. This means that to analyze this counterfactual in practice, 71.6% of the country-years

¹⁰This number is equivalent to a “generalized standard deviation” for X of $\sqrt{0.1} = 0.32$.

Table 1: How Factual Are Counterfactuals About Democracy?

Counterfactuals	N	% in Hull	Average % of Data “Nearby”	
			All	in Hull only
Entire World				
—Full Dem. to Autoc.	1775	53.1%	5.5%	8.4%
—Autoc. to Full Dem.	4039	17.6	2.4	8.2
—Part. Dem. to Autoc.	1376	80.5	12.3	14.7
—Autoc. to Part. Dem.	4039	61.8	4.2	6.0
Europe and Former USSR				
—Full Dem. to Autoc.	961	54.2%	4.0%	5.8%
—Autoc. to Full Dem.	863	23.3	3.8	10.7
—Part. Dem. to Autoc.	493	86.0	11.2	12.7
—Autoc. to Part. Dem.	863	76.6	5.3	6.5
Canada and Latin America				
—Full Dem. to Autoc.	383	64.5	8.6	11.7
—Autoc. to Full Dem.	604	30.5	3.4	8.1
—Part. Dem. to Autoc.	328	81.7	11.9	13.9
—Autoc. to Part. Dem.	604	69.5	5.4	7.3
Other Regions:				
—Full Dem. to Autoc.	431	40.4	5.9	11.6
—Autoc. to Full Dem.	2572	12.8	1.7	6.4
—Part. Dem. to Autoc.	555	74.6	13.6	17.3
—Autoc. to Part. Dem.	2572	55.0	3.5	5.4

would require extrapolation and would thus be highly model-dependent regardless of the model applied or dependent variable analyzed. As Table 1 summarizes, the result is not symmetric: Among the full democracies switched to autocracies, 53% require interpolation, whereas among the autocracies switched to full democracies, only 17% are interpolation problems. Unfortunately, little discussion in the literature reflects these facts.

The first few columns of Table 1 break down this average result for counterfactuals from three different regions. The rest of the table provides the fraction of countries within a modified Gower distance of 0.1 of a counterfactual, averaged over all counterfactuals within a given region and type of change in democracy. For example, across the 4,039 country-years where we could hypothetically change autocracies to partial democracies, an average of only 4.2% of the data points are this close to the counterfactual.

The overall picture in this table is striking. Studying the effects of changes in democracy have been a major project within comparative politics and international relations for at least half a century. This table applies to almost every such analysis with democracy as an explanatory variable in any field, regardless of the choice of dependent variable. Some areas and counterfactuals are less strained than others, but the results here show that most inferences in these fields (or most countries within each analysis) are highly model dependent, based much more on unverifiable assumptions about the model than on empirical data. A large fraction are highly model-dependent extrapolations, but even those which are interpolations are fairly distant from available data. This result varies by region and by counterfactual, and it would of course vary more if we changed the set of explanatory variables, but no matter how you look at it, the warning in Table 1 about interpreting the political science literature comes through clearly.

Numerous interesting case studies also emerge from analyses like these. For example, public policy makers and the media spent considerable time debating what would happen if Haiti became more of a democracy. In the early to mid-1990s, we find that the counterfactual of moving Haiti from a partial to a full democracy was in the convex hull, and hence a question that had a chance of being answered accurately with the available data. By 1996, conditions had worsened in the country, and this counterfactual became more counter to the facts, moving well out of the hull and thus required extrapolation.

3 Causal Effects

We now turn to counterfactual evaluation as part of causal inference in general and about the effects of democracy in particular. We begin with a version of the democratic peace hypothesis, which states that democracies are less conflictual than nondemocracies and later generalize to all possible dependent variables. Let D denote the “treatment” (or “key causal”) variable where $D = 1$ denotes democracy and $D = 0$ denotes nondemocracy.¹¹ The dependent variable is Y , the degree of conflict.

To define the causal effect of democracy on conflict, we denote Y_1 as the degree of conflict that would be observed if the country were a democracy and Y_0 as the degree of conflict if the country were a nondemocracy. Obviously, only either Y_0 or Y_1 are observed

¹¹The analysis of treatments with more than two levels is similar (e.g., Lechner 1999, Imbens n.d.). For expository purposes we focus on the binary treatment case in this paper.

for any one country at any given time, but not both, since (in our present simplified formulation) a country either is or is not a democracy. That is, we observe only $Y = Y_0(1 - D) + Y_1(D)$.

In principle, the democracy variable can have a different causal effect for every country in the sample. We can then define the causal effect of democracy by averaging over the whole world, or for the democracies and nondemocracies separately (or for any other subset of countries). For democracies, it is the so-called “average causal effect among the treated,” which we define as follows:

$$\begin{aligned}\theta &= E(Y_1|D = 1) - E(Y_0|D = 1) \\ &= \text{“Factual”} - \text{“Counterfactual”}\end{aligned}\tag{9}$$

We call the first term factual since Y_1 is observable when $D = 1$, although the expected value still may need to be estimated. We refer to the second as counterfactual because Y_0 (the degree of conflict that would be initiated by a country if it were not a democracy) is not observed and indeed is unobservable in democratic countries ($D = 1$). The causal effect for nondemocracies ($D = 0$) is directly analogous and also involves factual and counterfactual terms.

The average causal effect for the entire set of observations is

$$\gamma = E(Y_1) - E(Y_0),\tag{10}$$

where both terms have a counterfactual element, since each expectation is taken over all countries, but Y_1 is only observed for democracies and Y_0 only for nondemocracies. These definitions of causal effects are widely used in a variety of literatures (Rubin, 1974; Holland, 1986; Robins, 1999, 1999b; King, Keohane, and Verba, 1994; Pearl, 2000).

A counterfactual x in this context therefore takes the form of some observed data with only *one* element changed — for example, Mexico with all its attributes fixed but with the regime type changed to autocracy. Fortunately, we can easily evaluate how factual is this counterfactual with the methods already introduced in Section 2: by checking whether x falls in the convex hull of the observed X and computing the Gower distance from x to X . In addition, since x has only one counterfactual element we show below that we can also easily consult another criterion, whether x falls on the *support* of $P(X)$.¹²

¹²The support of $P(X)$ is the range of values of X that are possible (i.e., have positive density) whether or not they occur in our data (see also Footnote 4).

In most cases, the true causal effect, θ or γ , is unknown and needs to be estimated, often from observational data since social experiments are costly and, in many cases, such as our running example of democracy, infeasible. In section 3.1, we discuss the sources of potential problems in using observational data to estimate the causal effects. We focus on θ there for expository purposes as is usual in the statistical literature but, as our proofs in Appendix B show, our results hold for the effect on nondemocracies and the overall effect γ as well. Our empirical examples analyze the overall average causal effect, which is the usual parameter of interest in political science. In addition to illuminating sources of potential problems in causal inference, the decomposition of the estimation bias shows that inference involving counterfactuals not on the common support is a critical source of bias.

3.1 Decomposition of Causal Effect Estimation Bias

We begin with the simplest estimator of θ using observational data, the difference in means (or, equivalently, the coefficient on D from a regression of Y on D):

$$\begin{aligned} d &= \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0) \\ &= \text{mean}(Y_1|D = 1) - \text{mean}(Y_0|D = 0), \end{aligned} \tag{11}$$

where $\text{mean}(a) = \sum_i a_i/n$. The first line is the data-based analogue to (9), whereas the second recognizes that, for example, when $D = 1$, $Y = Y_1$. To identify the sources of potential problems using observational data in causal inference, we now present a new decomposition of the bias $E(d - \theta)$ of d as an estimator of the causal effect θ . This decomposition generalizes Heckman et al.'s (1998b) three-part decomposition. Their decomposition was applied to a simpler problem that does not adequately represent the full range of issues in causal inference in political science. Our new version helps to identify and understand the threats to causal inference in our discipline, as well as to focus in on

where counterfactual inference is most at issue. Thus, we show that,

$$\begin{aligned}
\text{bias} &\equiv E(d - \theta) \\
&= E[\text{mean}(Y_1|D = 1) - \text{mean}(Y_0|D = 0) - \theta] \\
&= [E(Y_1|D = 1) - E(Y_0|D = 0)] - [E(Y_1|D = 1) - E(Y_0|D = 1)] \\
&= E(Y_0|D = 1) - E(Y_0|D = 0) \tag{12}
\end{aligned}$$

$$= \Delta_x + \Delta_z + \Delta_n + \Delta_d \tag{13}$$

We derive the last equality and give the mathematical definition of the terms Δ_x , Δ_z , Δ_n , and Δ_d in Appendix B. These four terms denote the four sources of bias in using observational data, with the subscripts being mnemonics for the components. They are due to, respectively, omitting relevant control variables (Δ_x), controlling for variables that have been affected by the key causal variable (Δ_z), insufficient overlap in the distributions of the control variables for the democracies and non-democracies (Δ_n), and density differences in the control variable distributions (Δ_d). We now explain and interpret each of these components.

The absence of all bias in estimating θ with d would be assured if we knew (from Equation 12) that

$$E(Y_0|D = 1) = E(Y_0|D = 0). \tag{14}$$

Assumption (14) says that it is safe to use the observed control group outcome ($Y_0|D = 0$, the level of conflict initiated by nondemocracies) in place of the unobserved counterfactual ($Y_0|D = 1$, the level of conflict initiated by democracies, if they were actually nondemocracies.) Since this is rarely the case, we introduce control variables: Let Z denote a vector of control variables (explanatory variables aside from D), such that $X = \{D, Z\}$. If, after conditioning on Z , treatment assignment is random — that is, if we measure and control for the right variables (those that are related to D and affect Y), so that

$$E(Y_0|D = 1, Z) = E(Y_0|D = 0, Z) \tag{15}$$

holds, then from (25) in Appendix B, the first component of bias vanishes: $\Delta_x = 0$. Thus, this first component of bias, Δ_x , is due to pertinent control variables being omitted from X so that (15) is violated. This is the familiar omitted variable bias, which can plague any model.

The second component of bias in our decomposition, Δ_z , deviates from zero when some of the control variables Z are consequences of the key causal variable D . If Z includes post-treatment variables, then when the key causal variable D changes, the post-treatment variables may change too. Thus, if we denote Z_1 and Z_0 as the values that Z take on when $D = 1$ and $D = 0$, respectively (components of Z that are strictly pre-treatment do not change between Z_0 and Z_1), then, as with Y , either Z_1 or Z_0 , but not both, are observed for any one observation, and the observed $Z = Z_0(1 - D) + Z_1(D)$ will be different from the counterfactuals Z_0 and Z_1 , resulting in a non-zero Δ_z in (22).

As an example that illustrates the bias of controlling for post-treatment variables, suppose we are predicting the vote with partisan identification. If we control for the intended vote five minutes before walking into the voting booth, our estimate of the effect of partisan identification would be nearly zero. The reason is that we are inappropriately controlling for the consequences of our key causal variable, and for most of the effects of it, thus biasing the overall effect. Yet, we certainly should control for a pre-treatment variable like race which cannot be a consequence of partisan identification but may be a confounding variable that needs to be controlled. Thus, causal models require separating out the pre- and post-treatment variables and controlling only for the pre-treatment, background characteristics.

Post-treatment variable bias may well be the largest overlooked component of bias in estimating causal effects in political science (see King, Keohane, and Verba, 1994: 173ff). It is known in the statistical literature, but is assumed away in most models and decompositions (Frangakis and Rubin, 2001). This decision may be reasonable in other fields, where the distinction between pre- and post-treatment variables is easier to recognize and avoid, but in political science, especially comparative politics and international relations, the problem is often severe. For example, is GDP a consequence or cause of democracy? How about education levels? Fertility rates? Infant mortality? Trade levels? Are international institutions causes or consequences of international cooperation? Many or possibly even most variables in these literatures are both causes and consequences of whatever is regarded as the treatment (or key causal) variable. As Lebow (2000: 575) explains “Scholars not infrequently assume that one aspect of the past can be changed and everything else kept constant, . . . [but these] ‘Surgical’ counterfactuals are no more realistic than sur-

gical air strikes.” This is especially easy to see in quantitative research when each of the variables in an estimation takes its turn in different paragraphs of an article playing the role of the “treatment.” However, only in rare statistical models, and only under stringent assumptions, is it possible to estimate more than one causal effect from a single model.

To avoid this component of bias, Δ_z , we need to ensure that we control for no post-treatment variables, or that the distribution of our post-treatment variables does not vary with D :

$$P(Z_0|D = 1) = P(Z|D = 1). \quad (16)$$

If this assumption holds, then $\Delta_z = 0$ in (22) vanishes. Of course, there may be situations where it is difficult to ensure that both the first and second components of bias are eliminated. If a variable that is partially pre- and partially post-treatment is not controlled for, we might have omitted variable bias ($\Delta_x \neq 0$) because Equation 15 will no longer be satisfied; if this variable is controlled for, we might instead have post-treatment bias ($\Delta_z \neq 0$) (Rosenbaum 1984).

One way to avoid both Δ_z and Δ_x in the presence of variables that are partially post-treatment, aside from choosing better research designs in the first place, is to study *multiple-variable causal effects*. If we cannot study the effects of democracy controlling for GDP because higher GDP is in part a consequence of democracy, we may be able to study the joint causal effect of a change from nondemocracy to democracy *and* a simultaneous increase in GDP. This counterfactual is more realistic, i.e., closer to the data, because it reflects changes that actually occur in the world and does not require us to imagine holding variables constant that do not stay constant in nature. If this alternative formulation provides an interesting research question, then it can be studied without bias due to Δ_z since the joint causal effect will not be affected by post-treatment bias. Moreover, the multiple-variable causal effect might also have no omitted variable bias Δ_x , since both variables would be part of the treatment and could not be potential confounders. Of course, if this question is not of interest, and we need to stick with the original question, then no easy solution exists at present. At that point, we should recognize that the counterfactual question being posed is too unrealistic and too strained to provide a reasonable answer using the given data with any statistical model.

The last two components of bias — nonoverlap, Δ_n , and density differences, Δ_d —

are both affected by aspects of the distribution of the control variables, Z . For clarity, during our discussion here we assume that the two components of bias we have previously discussed are not an issue, so that we have no post-treatment bias, and we have the right pre-treatment variables Z to control for. From (23) and (24), we see that what contributes to the remaining two components of bias is the difference in the distribution of Z across the treatment (democracies) and control (nondemocracies) groups. These differences are precisely what create bias: For example, before assigning a drug or a placebo to two groups of patients, we first need to ensure that the two groups are the same on all relevant pre-treatment characteristics (i.e., Z). If, before treatment, the control group is on average healthier than the treatment group, then the effect of the drug will not be estimated correctly.

These group differences take two forms. First, the *support* of the distributions might differ, that is, there may be certain values of Z that some members of one group take on but no members of the other group have (for example, we might observe no full democracies with infant mortality rates as high as in some of the autocracies). This part of the difference constitutes non-overlap bias Δ_n . Intuitively, in regions of nonoverlap, the treatment and control groups are simply incomparable, and thus the counterfactuals underlying the causal inference are not sufficiently factual to draw valid inferences. The second form of difference is in the unequal distribution of Z across the two groups in the region of common support (for example, we may have observations on both democracies and autocracies that have a relatively low infant mortality rate, but a lot of more such observations are democracies than autocracies.) This component, Δ_d , contributes to the bias but through weighting or some other appropriate nonparametric or parametric method, we may be able to eliminate it.

Thus for the last two components of bias to vanish we need to ensure that the distribution of Z in the treatment group is the same as that in the control group. That is,

$$P(Z|D = 1) = P(Z|D = 0). \tag{17}$$

If we restrict inference to the overlap region only, then we do not have the problem of comparing the incomparable, and the nonoverlap bias Δ_n is eliminated. This would change the question being asked, but at least the result would be a question that has an answer,

which of course would be useful only if the resulting question is of interest. And if in the region of valid inference (overlap) we correctly adjust, the density differences across the two groups, then Δ_d is eliminated. Together, (17) holds.

So the question is *how* to ensure (17), that is, how to identify regions of density overlap, and how to adjust for differences in this region. It is easy to see, for the counterfactual x implied in the second term of (9), that being on the *common support* of $P(Z|D = 1)$ and $P(Z|D = 0)$ is in fact equivalent to being on the support of $P(X)$, so identifying the regions of density overlap is also a direct check of the quality of the key counterfactual required for causal inference. In the simple case when Z contains just one variable, such as GDP, we could plot the histogram of GDP for democracies and compare it to that for nondemocracies. Nonoverlap can easily be identified just by looking at these histograms (or nonparametric density estimates), which also readily reveals the nature of density differences across the two groups.

To avoid inference on the nonoverlap region, we would simply restrict our analysis to data with the same range of observations on GDP for both groups. To adjust for the density differences, some form of nonparametric matching, such as subclassification on GDP, can be easily applied as there is only one variable to match on. Or alternatively a parametric model of adequate functional flexibility can be used to control for GDP.

In most real applications, of course, Z contains many control variables, and so checking assumption (17) would involve estimating and comparing two multidimensional densities in hyperspace. For more than a few explanatory variables, this is a formidable task (essentially impossible without stringent assumptions). Adjusting the density differences by matching would be extremely difficult or infeasible since with many matching variables matches are hard to find; similarly, parametric models suffer from the curse of dimensionality as correct functional forms are difficult to identify, or to estimate efficiently due to the large number of parameters.

Fortunately, this curse of dimensionality problem has been solved with the use of the *propensity score*, $\pi \equiv \Pr(D = 1|Z)$, the probability of $D = 1$ given the control variables Z . The propensity score summarizes relevant aspects of the multidimensional Z with a scalar π . Rosenbaum and Rubin (1983) show that conditioning on π provides a substitute

for Equation 17 that is true without assumptions:

$$P(Z|D = 1, \pi) = P(Z|D = 0, \pi). \quad (18)$$

This statement is interesting in and of itself, since it implies that if we can estimate π , and control for it, we will be able to eliminate bias even if assumption (17) does not hold. For present purposes, Equation (18) enables us to show that assumption (17) can be expressed in an equivalent and more easy-to-verify form. That is, the assumption sufficient to eliminate the two remaining components of bias (so that $\Delta_n = \Delta_d = 0$) is that the distribution of π for democracies is identical to the distribution of π for nondemocracies:¹³

$$P(\pi|D = 1) = P(\pi|D = 0). \quad (19)$$

This expression solves the curse of dimensionality problem in Equation 17, and so is easier to use, because it only requires comparison of two unidimensional densities of π (one for democracies and one for nondemocracies), and adjusting the density differences based only on one control variable.¹⁴

3.2 The Causal Effect of Democracy

In this section we use the ideas discussed this section to analyze the literature studying causal effect of democracy, focusing mainly on the Δ_n component that is most relevant to our theme of counterfactual evaluation. We take the state failure data and consider the causal effect of partial democracy vs. autocracy, and then separately of the effect of full democracy vs. autocracy. (The analysis here applies to the causal effect of these changes on any dependent variable.) To identify regions of nonoverlap — i.e., areas in which one type of country does not have a comparable counterpart in the other, and therefore where the counterfactuals underlying the causal inference are not reasonably factual — we compute the propensity score and plot the density estimates (smooth versions

¹³By definition, $P(Z|D = 0) = \int P(Z|D = 0, \pi)P(\pi|D = 0)d\pi$, which upon substituting in Equation 18, gives $P(Z|D = 0) = \int P(Z|D = 1, \pi)P(\pi|D = 0)d\pi$. Comparing this result with $P(Z|D = 1) = \int P(Z|D = 1, \pi)P(\pi|D = 1)d\pi$, which is also true by definition, proves that (17) holds if (19) does.

¹⁴Of course, the propensity score itself is often unknown and must be estimated from data. The standard practice is to run a logit model of D on the variables Z , adding enough higher order terms such as interactions or squared terms into the model so that the estimated propensity score does behave as a propensity score: that is, Equation 18 has testable implications such as the means of Z within each given small interval of π (such as $[0,0.1]$, $[0.1,0.2]$, etc.) should be approximately the same for $D = 1$ and $D = 0$ countries. Formal tests of the condition can be performed using, for example, Hotelling's T^2 test. In our empirical example below we use a neural network model to estimate the propensity score that passes the Hotelling's T^2 test.

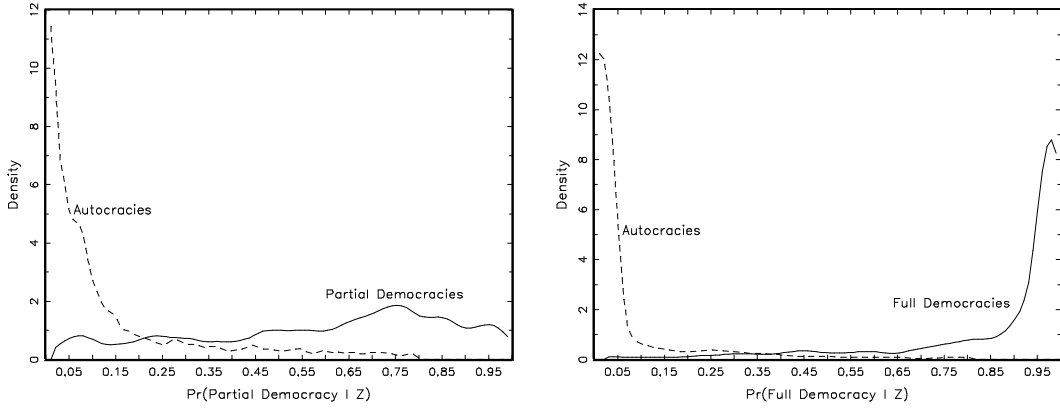


Figure 3: *Propensity Score Density Estimates for partial democracies and autocracies in the left graph and full democracies vs. autocracies in the right graph. Note the differences in the densities (which contribute to Δ_d), and especially the areas of zero density (nonoverlap) for each curve (which produce Δ_n).*

of histograms, computed with kernels that have bounded support) for partial democracies and autocracies (in the left graph) and for full democracies and autocracies (in the right graph) in Figure 3.

Figure 3 clearly shows that the densities within each pair of comparisons are not the same, which is no surprise. The ex ante probability of being a democracy should be higher for actual democracies than actual autocracies, and this is clearly the case: most of the mass of democracies is on the right of each graph (indicating high probability values) and even more of the density for autocracies is at the left (indicating low probability values). Also as expected, the differences are more dramatic for causal effect of full democracy vs. autocracy (on the right) than for partial democracy vs. autocracy (on the left). The extent to which the two densities are not identical in each graph is a vivid description of the bias of the difference in means estimator (11) due to the Δ_n and Δ_d components. If, in contrast, the values of D were assigned randomly, then the ex ante probabilities would not be related to D , and so the two densities would be the same. It is the observational, nonrandom aspects of this assignment that produce these components of bias. In the figure the *nonoverlap bias*, Δ_n , refers to portions of each graph for which one curve has positive density and the other has zero density. In each plot in this figure, areas of nonoverlap are at the ends. For example, no actual autocracies have an ex ante probability of being a partial democracy or full democracy of more than about 0.8. Yet, a large fraction of democracies have this high an ex ante probability. Similarly, a large number of autocracies'

ex ante probabilities of being democracies are less than about 0.02, but no democracies can be found with such low probabilities.

Nonoverlap is a serious problem. Using nonoverlap data requires highly model-dependent extrapolation to areas where no data exist. The problem is closely related to “what if” counterfactuals that lie outside the convex hull we discuss earlier (at the limit, being in the overlap region is a sufficient condition for being in the convex hull). In the causal inference literature where scholars have begun to pay attention to the problem, the “solution” mostly takes the form of changing the question to be asked, that is, eliminating the portions of the population to which the areas of nonoverlap correspond and hence changing the target of inference. The counterfactuals required for inference on the original population are considered too strained for that purpose. Just as we recommend in Section 2, we must conclude that these causal effects cannot be reliably estimated in existing data. In our empirical example here, the area of nonoverlap in both graphs in Figure 3 is substantial. In the comparison between full democracy vs. autocracy in particular, the nonoverlap area is so large that the counterfactuals for causal inferences would be too far from the data to allow empirical justification for the vast majority of such inferences.

Before ending this section we comment briefly on the component of bias due to density differences in the overlap region, Δ_d . This can be a substantial component of the total bias in many empirical problems. Unlike the nonoverlap problem which is just beginning to receive attention, however, adjusting density differences is a topic extensively researched and a practice routinely performed. A major function of standard regression type models used in political science, for example, is precisely for that purpose. Parametric models are subject to incorrect functional form assumptions, so alternatively nonparametric methods such as matching or inverse propensity score weighting can be used instead (e.g., Rosenbaum and Rubin, 1984; Heckman et al., 1998; Robins, 1999, 1999b; Winship and Sobel, 2000). The discovery of the propensity score has made these methods more feasible for most problems, as the task is reduced to one dimensional operations without the curse of dimensionality. These techniques are gaining popularity in many disciplines in recent years and have much to offer political scientists, but we save a detailed discussion of them for another paper.

We conclude this section by noting that although decompositions of bias are very

useful in understanding the problems, all four components of bias must ordinarily be eliminated simultaneously to produce reliable inferences. With one exception that we discuss momentarily, a partial “fix” may make the total bias worse. For example, if Δ_z and Δ_x are both known not to be zero, it is possible (although perhaps unlikely) that fixing one rather than both components may actually increase the total bias.

The exception is Δ_n . Although blind luck can always occur, eliminating the nonoverlap region, hence recognizing the impossibility of making certain causal inferences, should always be helpful. It may be disappointing of course to know that the desired questions have no good answers in available data, but it is better to know this than to ignore it.

4 Concluding Remarks

Even far-out questions with answers that are highly model-dependent may still be important enough to warrant further study. For a few examples, what would happen to the stability of the U.S. Government if inflation increased to 200% or had Gore refused to concede the 2000 election? What would the future of military conflict be in a world without nation states? How bad would the devastation be from a third world war? If a new virulent infectious disease that is ten times as bad as AIDS strikes the developed world and lasts longer than the AIDS crisis, would current international institutions survive? Scholars can and certainly still should ask questions like these, but we would be better served if we knew whether and to what degree our answers are based on empirical evidence rather than model assumptions. Sometimes, with the data at hand, no statistical model can give valid answers, and we must rely on theory or new data collection efforts. The techniques offered in this paper may be useful in ascertaining the degree to which this is the case. In this regard, it may be useful for empirical researchers to report these or other statistics in evaluating their counterfactuals.

We have used the methods discussed here to evaluate counterfactuals in the large area of research devoted to assessing the effects of democracy. We found that questions about democracy with empirical answers that are not highly model-dependent are a subset, sometimes a small subset, of those that have been asked. Usually scholars combine data on all available democracies to make predictions, ask “what if” questions, or estimate causal effects. Unfortunately, many of the explicit or implied questions have no available

control groups or otherwise cannot be estimated without making assumptions that even the authors would probably be unwilling to defend. We might like to know what would happen if Iraq became a full democracy, for example, but almost no evidence exists in our data with which to evaluate such a question. Having time series-cross-sectional data sets with thousands of observations does not change this basic fact and will not make inferences like these any more secure. In fact, these data sets must be analyzed with more care than has been common since, as it turns out, they do not include much evidence on many otherwise interesting counterfactuals. Asking questions about the effects of changes in democracy averaged over all countries — the predominant approach taken in the literature — almost always implies questions without useful empirical evidence. Statistical analyses in data sets like these should change: scholars could seek different types of evidence, develop better theory, or narrow their inferential target to a subset of countries and counterfactuals that have empirical support in their data.

In addition to sensitivity testing, some other related approaches to the problem we study might include Bayesian model averaging and other methods that improve model generalization, such as committee methods. These approaches are useful in numerous circumstances, but they require the analyst to specify a particular class of models for exploration. In all cases, for practical purposes, the class of models must be narrower than the set of all possible models. In contrast, the questions about counterfactuals we ask here are broader and are not feasible to attack by these other approaches as they apply to the class of all possible models given a data set. One productive procedure might be to use the methods we offer here and then to use substantive assumptions or theories to narrow the class of models. Then we might be able to assess the degree to which model dependence is reduced by adding theoretical information, rather than data, to the estimation problem.

Suppose we read about a model that fits the data exceedingly well, has big likelihood ratio or F statistics, narrow confidence intervals, significance on all coefficients, large causal effect estimates, predictions with path breaking policy implications, and fascinating answers to a range of “what if” questions. With statistical reporting standards now commonly used in political science, essentially all such models would be published and taken seriously by readers. A subset of these, however, would involve counterfactuals that are so model-dependent as to be nearly unrelated to the data at hand, and so are

based more on the authors’ hypotheses and convenient model assumptions than their data. The main message of this paper is that assessing model dependence of counterfactual questions needs to be a routine and expected part of statistical reporting for anyone making predictions, asking “what if” questions, and estimating causal effects — which together encompasses the goals of a large fraction of empirical work in the discipline. The goal of this paper is to keep this empirical work based on empirical evidence.

A Checking Membership in a Convex Hull as a Linear Programming Problem

The goal is to check whether a given point $x_{1 \times k}$ is in the convex hull of $X_{n \times k}$, where n is the number of data points in X and k the number of variables. Let S be the set of vertices of the convex hull of X , so that S contains all the “boundary” or “extreme” points of X . By definition, x being in the convex hull of X means that x can be expressed as a convex combination of points in S . Since all points of X are of course also convex combinations of points in S , the condition is equivalent to x being a convex combination of all points in X . Identification of S can be computationally very expensive, but we show that the second form of the condition can be checked easily using standard linear programming software.

To do so, we formulate the problem as one of checking the existence of a feasible solution for a standard linear programming problem. If x can be expressed as a convex combination of points in X , then there exists a vector of coefficients $\eta_{n \times 1}$ constrained to the simplex so that $X'\eta = x'$. This last equation contains k linear constraints, each stating that an element (variable) of x is a convex combination of the corresponding elements in X . Combining this with the constraint that the elements of η sum to one, we have a total of $k + 1$ linear constraints in the form $A'\eta = B'$, where A' and B' are X' and x' with a row of ones added respectively.

To check whether x is in the convex hull of X therefore is equivalent to checking the existence of a feasible solution to the following standard form linear programming problem:

$$\begin{aligned} \min \quad & C'\eta \\ \text{s.t.} \quad & A'\eta = B' \\ & \eta \geq 0 \end{aligned} \tag{20}$$

where C is a vector of zeros (so there is no objective function to minimize). Checking

whether there is a feasible solution to problem (20) is what all standard LP software does in Phase I, and it can be done very efficiently for very large n and k .

B Decomposition of Causal Effect Estimation Bias

We now derive the bias of the estimator d in Equation 11. Note that d is the simplest estimator for all three causal effect parameters (the average causal effect in democracies, θ , its counterpart for nondemocracies for which we have assigned no symbol, and the average causal effect overall, γ , from Equation 10). In this appendix we prove that the bias of d as an estimator of any of these three parameters has the same four types of components as given in Equation 13 and discussed in section 3.1.

We start by showing that the bias of d in estimating the total effect γ is a convex combination of its bias in estimating the two group-specific causal effect parameters. We have

$$E(d - \gamma) = [E(Y_1|D = 1) - E(Y_0|D = 0)] - [E(Y_1) - E(Y_0)].$$

Let $\tau = \Pr(D = 1)$ be the size of the treatment group and then rewrite the terms in the definition of γ above as $E(Y_1) = \tau E(Y_1|D = 1) + (1 - \tau)E(Y_1|D = 0)$ and $E(Y_0) = \tau E(Y_0|D = 1) + (1 - \tau)E(Y_0|D = 0)$. Thus,

$$\begin{aligned} E(d - \gamma) &= E(Y_1|D = 1) - E(Y_0|D = 0) \\ &\quad - \tau E(Y_1|D = 1) - (1 - \tau)E(Y_1|D = 0) + \tau E(Y_0|D = 1) + (1 - \tau)E(Y_0|D = 0) \\ &= (1 - \tau) [E(Y_1|D = 1) - E(Y_1|D = 0)] + \tau [E(Y_0|D = 1) - E(Y_0|D = 0)] \\ &= (1 - \tau)B_0 + \tau B_1 \end{aligned} \tag{21}$$

where $B_1 = E(Y_0|D = 1) - E(Y_0|D = 0) = E(d - \theta)$ is the bias of using d to estimate θ , the causal effect on the treated (democracies), as derived in Equation 12. In a directly analogous way, $B_0 = E(Y_1|D = 1) - E(Y_1|D = 0)$ is the bias of d as an estimator of the causal effect in the control group (nondemocracies). (Note that, quite intuitively, B_1 is a function of unobservables among the treated, and B_0 is a function of unobservables among the untreated.)

We now derive the four components of bias for B_1 , and then in an identical fashion for

B_0 , before we combine them as per Equation 21. We have:

$$\begin{aligned}
B_1 &= \mathbb{E}(Y_0|D = 1) - \mathbb{E}(Y_0|D = 0) \\
&= \mathbb{E}_{z_0}[\mathbb{E}(Y_0|D = 1, Z_0) - \mathbb{E}(Y_0|D = 0, Z_0)] \\
&\quad - \mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)] + \mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)] \\
&= [\mathbb{E}_{z_0}\mathbb{E}(Y_0|D = 1, Z_0) - \mathbb{E}_z\mathbb{E}(Y_0|D = 1, Z)] + \mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)] \\
&= \Delta_z + \mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)]
\end{aligned}$$

where

$$\Delta_z = \mathbb{E}_{z_0}\mathbb{E}(Y_0|D = 1, Z_0) - \mathbb{E}_z\mathbb{E}(Y_0|D = 1, Z) \quad (22)$$

is the bias due to controlling for post-treatment variables, and $\mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)]$ can be further decomposed, following and generalizing the approach in Heckman et al. (1998b). Let S_j denote the support of $F(Z|D = j)$ for $j = 0, 1$ and S the common support. Then

$$\begin{aligned}
&\mathbb{E}_z[\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)] \\
&= \int_{S_1} \mathbb{E}(Y_0|D = 1, Z)dF(Z|D = 1) - \int_{S_0} \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 0) \\
&= \left\{ \int_{S_1 \setminus S} \mathbb{E}(Y_0|D = 1, Z)dF(Z|D = 1) + \int_S \mathbb{E}(Y_0|D = 1, Z)dF(Z|D = 1) \right\} \\
&\quad - \left\{ \int_{S_0 \setminus S} \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 0) + \int_S \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 0) \right\} \\
&\quad + \left\{ \int_S \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 1) - \int_S \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 1) \right\} \\
&= \Delta_n + \Delta_d + \Delta_x
\end{aligned}$$

where, through regrouping the terms,

$$\Delta_n = \int_{S_1 \setminus S} \mathbb{E}(Y_0|D = 1, Z)dF(Z|D = 1) - \int_{S_0 \setminus S} \mathbb{E}(Y_0|D = 0, Z)dF(Z|D = 0) \quad (23)$$

$$\Delta_d = \int_S \mathbb{E}(Y_0|D = 0, Z)\{dF(Z|D = 1) - dF(Z|D = 0)\} \quad (24)$$

$$\Delta_x = \int_S \{\mathbb{E}(Y_0|D = 1, Z) - \mathbb{E}(Y_0|D = 0, Z)\}dF(Z|D = 1), \quad (25)$$

Combining results above gives:

$$B_1 = \Delta_z + \Delta_n + \Delta_d + \Delta_x \quad (26)$$

(which proves Equation 13).

Decomposition of B_0 proceeds identically and we omit the intermediate steps. The results are:

$$B_0 = \Delta_z^0 + \Delta_n^0 + \Delta_d^0 + \Delta_x^0 \quad (27)$$

where

$$\Delta_z^0 = E_z E(Y_1|D=0, Z) - E_{z_1} E(Y_1|D=0, Z_1) \quad (28)$$

$$\Delta_n^0 = \int_{S_1 \setminus S} E(Y_1|D=1, Z) dF(Z|D=1) - \int_{S_0 \setminus S} E(Y_1|D=0, Z) dF(Z|D=0) \quad (29)$$

$$\Delta_d^0 = \int_S E(Y_1|D=0, Z) \{dF(Z|D=1) - dF(Z|D=0)\} \quad (30)$$

$$\Delta_x^0 = \int_S \{E(Y_1|D=1, Z) - E(Y_1|D=0, Z)\} dF(Z|D=1) \quad (31)$$

Now, to arrive at the decomposition of bias in estimating the total effect γ , we only need to combine Equations (26) and (27) as per Equation (21). Omitting tedious but straightforward intermediate steps, the results are:

$$E(d - \gamma) = \Delta_z^t + \Delta_n^t + \Delta_d^t + \Delta_x^t \quad (32)$$

where

$$\begin{aligned} \Delta_z^t &= (1 - \tau) \{E_z E(Y_1|D=0, Z) - E_{z_1} E(Y_1|D=0, Z_1)\} \\ &\quad + \tau \{E_{z_0} E(Y_0|D=1, Z_0) - E_z E(Y_0|D=1, Z)\} \\ \Delta_n^t &= \int_{S_1 \setminus S} \{(1 - \tau) E(Y_1|D=1, Z) + \tau E(Y_0|D=1, Z)\} dF(Z|D=1) \\ &\quad - \int_{S_0 \setminus S} \{(1 - \tau) E(Y_1|D=0, Z) + \tau E(Y_0|D=0, Z)\} dF(Z|D=0) \\ \Delta_d^t &= \int_S \{(1 - \tau) E(Y_1|D=0, Z) + \tau E(Y_0|D=0, Z)\} \{dF(Z|D=1) - dF(Z|D=0)\} \\ \Delta_x^t &= \int_S (1 - \tau) [E(Y_1|D=1, Z) - E(Y_1|D=0, Z)] \\ &\quad + \tau [E(Y_0|D=1, Z) - E(Y_0|D=0, Z)] dF(Z|D=1) \end{aligned}$$

Clearly, the four components of bias have the same qualitative interpretations across Equations (26), (27), and (32).

References

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

- Cuadras, C.M. and J. Fortiana. 1995. "A Continuous Metric Scaling Solution for A Random Variable." *Journal of Multivariate Analysis* 52:1–14.
- Cuadras, C.M., J. Fortiana and F. Oliva. 1997. "The Proximity of an Individual to a Population with Applications to Discriminant Analysis." *Journal of Classification* 14:117–136.
- Dozois, Gardner and Stanley Schmidt, eds. 1998. *Roads not Taken: Tales of Alternative History*. New York: Del Rey.
- Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger and Robert S. Chen. 1998. *The State Failure Task Force Report: Phase II Findings*. McLean, Virginia: Science Applications International Corporation.
- Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger and Robert S. Chen. 1998b. "The State Failure Project: Early Warning Research for U.S. Foreign Policy Planning." In *Preventive Measures: Building Risk Assessment and Crisis Early Warning System*, ed. John L. Davies and Ted Robert Gurr. Lanham, Maryland: Rowman and Littlefield.
- Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Marc Levy, Geoffrey D. Dabelko and Alan N. Unge. 1999. "The State Failure Report: Phase II Findings." *Environmental Change and Security* 5.
- Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Pamela T. Surko and Alan N. Unger. 1995. *State Failure Task Force Report*. McLean, Virginia: Science Applications International Corporation.
- Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43:169–195.
- Frangakis, Constantine E. and Donald Rubin. N.d. "The defining role of principal effects in comparing treatments using general post-treatments using general post-treatment variables: from surrogate endpoints to censoring by death." <http://biosun01.biostat.jhsph.edu/~cfrangak/papers/>.

- Gelman, Andrew and Gary King. 1994. "A unified method of evaluating electoral systems and redistricting plans." *American Journal of Political Science* 38:514–554.
- Gower, J.C. 1966. "Some Distance properties of latent root and vector methods used in multivariate analysis." *Biometrika* 53:325–388.
- Gower, J.C. 1971. "A general coefficient of similarity and some of its properties." *Biometrics* 27:857–872.
- Heckman, James, H. Ichimura, J. Smith and P. Todd. 1998b. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66:1017–1098.
- Heckman, James, H. Ichimura and P. Todd. 1998. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64:605–654.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- Imbens, Guido W. n.d. "The Role of the Propensity Score in Estimating Dose-Response Functions." *Biometrika*.
- King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95:49–69.
- King, Gary and Langche Zeng. 2002. "Improving Forecasts of State Failure." *World Politics* Winter.
- King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton University Press.
- Kuo, Yen-Hong. 2001. "Extrapolation of Association Between Two Variables in Four General Medical Journals." Forth International Congress on Peer Review in Biomedical Publication.
- Kvart, Igal. 1986. *A Theory of Counterfactuals*. Indianapolis: Hackett Publishing Company.

- Lechner, Michael. 1999. "Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumptions." IZA Discussion Papers no.91, University St. Gallen.
- Lewis, David K. 1973. *Counterfactuals*. Cambridge: Harvard University Press.
- Madych, W.R. and S.A. Nelson. 1992. "Bounds on Multivariate Polynomials and Exponential Error Estimates for Multiquadric Interpolation." *Journal of Approximation Theory* 70:94–114.
- Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Harvard University Press.
- Murphy, George G. S. 1969. "On Counterfactual Propositions." *History and Theory* 9:14–38.
- Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Robins, James M. 1999. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, ed. M.E. Halloran and D. Berry. Vol. 116 New York: Springer-Verlag pp. 95–134.
- Robins, James M. 1999b. "Association, causation, and marginal structural models." *Synthese* 121:151–179.
- Rosenbaum, Paul. 1984. "The Consequences of Adjusting for a Concomitant Variable That Has been Affected by the Treatment." *Journal of the Royal Statistical Society, A* 147:656–666.
- Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central role of the propensity score in observational studies for causal effects." *Biometrika* 70:41–55.
- Rosenbaum, Paul R. and Donald B. Rubin. 1984. "Reducing bias in observational studies using subclassification on the propensity score." *Journal of the American Statistical Association* 79:515–524.

- Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 6:688–701.
- Schaback, R. 1996. "Approximation by Radial Basis Functions with Finitely Many Centers." *Constructive Approximation* 12:331–340.
- Tally, Steve. 2000. *Almost America: From the Colonists to Clinton, A "What If" History of the U.S.* New York: Quill.
- Tetlock, Philip E. 1999. "Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are we Prisoners of our Preconceptions?" *American Journal of Political Science* 43:335–366.
- Tetlock, Philip E. and A. Belkin, eds. 1996. *Counterfactual Thought Experiments in World Politics*. Princeton: Princeton University Press.
- Tetlock, Philip E., Ned R. Lebow and G. Parker, eds. 2000. *Unmaking the West: Counterfactual Explorations of Alternative Histories*. New York: Columbia University Press.
- Tetlock, Philip E. and Richard Ned Lebow. 2001. "Poking Counterfactual Holes in Covering Laws: Cognitive Styles and Historical Reasoning." *American Political Science Review*.
- Thorson, Stuart J. and Donald A. Sylvan. 1982. "Counterfactuals and the Cuban Missile Crisis." *International Studies Quarterly* 26:539–571.
- Winship, Christopher and Michael Sobel. 2000. "Causal Inference in Sociological Studies." Harvard University.
- Wu, Z. and R. Schaback. 1993. "Local Error Estimates for Radial Basis Function Interpolation of Scattered Data." *Journal of Numerical Analysis* 13:13–27.