When Can History be Our Guide? The Pitfalls of Counterfactual Inference[1]

Gary King[2] Langche Zeng[3]

October 9, 2005

[1]Easy-to-use software to implement the methods introduced here, called "WhatIf: Software for Evaluating Counterfactuals," is available at http://GKing.Harvard.Edu/whatif. At the suggestion of the editors, we minimized proofs and detailed mathematical arguments in this paper and wrote a separate technical companion piece that overlaps this one: it includes complete mathematical proofs, more general notation, and other methodological results not discussed here, but fewer examples and less pedagogical material. See "The Dangers of Extreme Counterfactuals," *Political Analysis*, Vol.14, No.2, 2006, forthcoming, copy at http://GKing.Harvard.edu/files/abs/counterft-abs.shtml.

Thanks to Jim Alt, Scott Ashworth, Neal Beck, Jack Goldstone, Sander Greenland, Orit Kedar, Walter Mebane, Maurizio Pisati, Kevin Quinn, Jas Sekhon, Simon Jackman for helpful discussions; Michael Doyle and Nicholas Sambanis for their data and replication information; and the National Institutes of Aging (P01 AG17625-01), the National Science Foundation (SES-0318275, IIS-9874747), and the Weatherhead Initiative for research support.

[2]David Florence Professor of Government, Harvard University (Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; `http://GKing.Harvard.Edu`, `King@Harvard.Edu`, (617) 495-2027).

[3]Professor of Political Science, University of California-San Diego (Department of Political Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0521, `zeng@ucsd.edu`).

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. Unfortunately, standard statistical approaches assume the veracity of the model rather than revealing 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. If an analysis fails the tests we offer, then we know that substantive results are sensitive to at least some modeling choices that are not based on empirical evidence. 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 separate analyses of the effects of UN peacebuilding efforts. We find evidence that many scholars are inadvertently drawing conclusions based more on modeling hypotheses than on evidence in the data. For some research questions, history contains insufficient information to be our guide. Free software that accompanies this paper implements all our suggestions.

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. In early 2003, Saddam Hussein was obviously either alive or dead, but the world did not know which it was until he was found. 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 whether the Americans and British would have invaded Iraq if the 9/11/2001 attack on the World Trade Center had not occurred. Counterfactuals are also crucial for making forecasts, such as whether there will be peace in the Mideast in the next two years, since the quantity of interest is not knowable at the time of the forecast, although it will eventually become known. Counterfactuals are essential as well in making causal inferences, since causal effects are differences between factual and counterfactual inferences: for example, how much more international trade would Syria have engaged in during 2003 if the Iraqi war had been averted.

Counterfactual inference has been a central topic of methodological discussion in political science (Fearon, 1991, Thorson and Sylvan, 1982; Tetlock and Belkin, 1996; Tetlock and Lebow, 2001), 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. "Counterfactuals are an essential ingredient of scholarship. They help determine the research questions we deem important and the answers we find to them" (Lebow, 2000: 558). 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.

The problem is easy to see in the simple example in Figure 1. Here, we fit linear and quadratic models to a simple set of simulated data (with the one explanatory variable on the horizontal axis and the dependent variable and its expected value on the vertical axis). The fit of the two models to the observed data is almost indistinguishable and we have little statistical reason to choose one over the other. This is not a problem if we are interested in a prediction of Y for any X between 1 and 2 where the data can be found; in this region, the choice of model is unimportant since either model (or most any other model with a reasonably smooth functional form) would yield similar predictions. However, predictions of Y for values of X outside the range of the data would be exquisitely sensitive to the choice of the model. In other words, inferences in the range of the data are far less *model-dependent* than inferences outside the data. The risk with model-dependent inferences is that substantive conclusions are based more on apparently minor modeling choices than on the empirical evidence.

[Figure 1 about here.]

But how can we tell how model-dependent inferences are when the counterfactual inference is not so obviously extreme, or when it involves more than one explanatory variable? The answer to this question cannot 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 ratios, predicted values, test statistics, first differences, p-values, etc. (For example, although not shown in the figure, the confidence intervals for the extrapolations in Figure 1 do not contain the predictions from the other model for much of the range of the extrapolation.) 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 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 the 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. The identified class of models normally excludes at least some that have a reasonable probability of returning different substantive conclusions. Most often, this approach is skipped entirely.

What the approach offered here provides is several easy-to-apply methods that reveal the degree of model dependency, without having to run all the models. As a consequence, it applies for the class of nearly all models, whether or not they are formalized, enumerated, and run, and for the class of all possible dependent variables, conditional only on the choice of a set of explanatory variables. If an analysis fails our tests then we know it will fail a sensitivity test too, but without being in the impossible position of having to run all possible models to find out.

Our field includes many discussions of the problem of strained counterfactuals in *qualitative* research. For example, Lebow (2000) and Fearon (1991) 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). Qualitative analysts

seem to understand this issue well. Scholars frequently ask questions like whether the conflict in Iraq is sufficiently like Vietnam so that we can infer the outcome from this prior historical experience. Unfortunately, although extreme counterfactuals is one of the most serious problems confronting comparative politics and international relations, *quantitative* empirical scholarship rarely addresses the issue. Yet, it is hard to think of many quantitative analysts in comparative politics and international relations in recent years who do not stop to interpret their results by asking what happens, for example, to the probability of conflict if all control variables are set to their means and the key causal variable is changed from its 25th to its 75th percentile value (King, Tomz, and Wittenberg, 2000). Every one of these analyses is making a counterfactual prediction, and every one needs to be evaluated by the same ideas well known in qualitative research. In this paper, we provide quantitative measures of these and related criteria that are meant to complement the ideas for qualitative research discussed by many authors.

We offer two empirical examples. The first evaluates inferences 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 most studied questions in comparative politics and international relations. Our results show that many analyses about democracy include at least some counterfactuals with little empirical support, so that scholars in these literatures are asking some counterfactual questions that are far from their data, and are therefore inadvertently drawing conclusions about the effects of democracy in some cases based on indefensible model assumptions rather than empirical evidence.

Whereas our example about democracy applies approximately to a large array of prior work, we also introduce an example that applies exactly to one groundbreaking study on designing appropriate peacebuilding strategies (Doyle and Sambanis, 2000). We replicate this work, apply our methods to these data, and find that the central causal inference in the

study involves counterfactuals that are too far from the data to draw reliable inferences, regardless of the methods employed. We illustrate by showing how inferences about the effect of UN intervention drawn from these data are highly sensitive to model specification.

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 and UN peacekeeping. 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. 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), which is merely a general expression for the linear or nonlinear regression line, i.e., how the expected value (or mean) of Y depends on X. In linear regression, the CEF is $\mathrm{E}(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$, whereas in logistic regression the CEF is $\mathrm{E}(Y|X) = 1/(1 + e^{-X\beta})$. These CEFs and others are illustrated in Figure 2 with one statistical model in each of four graphs, and with three CEFs displayed in each based on different choices of parameter values from the chosen functional form. For example the top right graph displays only the linear functional form, with three lines that differ based on their parameter values (the intercept and slope). The task of the analyst is to chose the statistical model (the graph), whereas the task of the parametric statistical analysis optimization routine is to find the parameter values that select one member of the assumed family of curves that best fits the data. The optimization routines usually work exceptionally well, but they can only choose within the given family. If the data are generated by one family of CEFs (one graph) but another is assumed by the investigator, we'll still get an approximation (such as the best linear approximation to the logit curve), but the estimated predictions can then be far off the mark, as Figure 1 illustrates.

[Figure 2 about here.]

Interestingly, no matter how good the fit to the data, each of these CEFs can be computed for *any* (real) values of the counterfactual point x. The model never complains, and exactly the same calculation can be applied for any 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, 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 is obviously capable of producing better forecasts (and "what if" evaluations) for some values of x than others. Predictive confidence intervals for forecasts farther from the data are larger, but confidence intervals computed in the usual way still assume the veracity of the model, and thus the uncertainty it represents does not include model dependence, no matter how far the counterfactual is from the data.

Worrying about model choice may be good in general, but it will not help here. Other models will not do verifiably better with the same data; one 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 reliably 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 are effectively seeking 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.

Indeed, this point is crucial since *the greater the distance from the counterfactual to the closest reasonably sized portion of available data, the more model dependent inferences can be about the counterfactual.* In our technical companion paper, we define this claim more precisely and, apparently for the first time, prove it mathematically. That is, no matter what the counterfactual, no matter what class of models one identifies as plausible, no matter how well the models tested fit the observed data, the farther the counterfactual from the data the higher the degree of model dependence becomes possible. Counterfactual questions sufficiently far from the data produce inferences with little or no empirical

content. Moreover, our proof is highly general. It does not assume knowledge of the model, its functional form, the estimator, or the dependent variable, and only assumes that the CEF (conditional on X) satisfies a general continuity condition, which fits almost all statistical models used and theoretical processes hypothesized in the discipline.

We now offer two procedures for measuring the distance from a counterfactual to the data that can be used to assess whether a question posed can be reliably answered from any statistical model. Neither requires any information about the model, estimator, or even the dependent variable.

2.1 Interpolation vs. Extrapolation

A simple but powerful distinction in measuring the distance of a counterfactual from the data, and thus assessing the counterfactual question x, is whether answering it by computing the CEF $\mathrm{E}(Y|x)$ would involve interpolation or extrapolation (e.g., Kuo, 2001; Hastie et al., 2001). Except for some unusual situations we offer diagnostics for below, datasets contain more information about counterfactuals that require interpolation than those that require extrapolation. Hence answering a question involving extrapolation normally requires far more model-dependent inferences than one involving interpolation.

For intuition, imagine we have data on foreign aid received by countries with two natural disasters in a year and and we wished to estimate how much foreign aid countries receive when they have two natural disasters in a year. (Suppose for simplicity that each of the natural disasters is approximately the same size and of roughly the same consequence.) If we have enough such data, no modeling assumptions are necessary. That is, we can make a model-free inference by merely averaging the amount of money spent on foreign aid in these countries.

However, suppose we were still interested in foreign aid received by countries with two

natural disasters, but we only observed countries with one or three disasters in a year. This is a simple (counterfactual) "what if" question because we have no data on countries with two natural disasters. The interpolation task, then, is to draw some curve from expected foreign aid received in countries with a single natural disaster to the expected aid received in countries with three natural disasters; where it crosses the two-natural-disaster point is our inference. Without any assumptions, this curve could go anywhere, and the inferred amount of foreign aid received for countries with two disasters 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) is quite reasonable for this example, as it is for most political science problems; it is also intuitive, but it is stronger than necessary to prove our point. The consequence of this smoothness assumption is to narrow greatly the range of foreign aid into which the interpolated value can fall, especially compared to an extrapolation. Even if the aid received by countries with two disasters is higher than the aid received for countries with three disasters or lower than nations with only one, it probably won't be too much outside this range.

However, now suppose we observe the same data but need to extrapolate to foreign aid received for countries with four natural disasters. We could impose some smoothness again, but even allowing 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. This is also an example of

our general proof since the counterfactual requiring interpolation in this example is closer to more data than the counterfactual requiring extrapolation, and so the interpolation is less model dependent.

If we learn that a counterfactual question 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 determine whether a question involves extrapolation with one variable should now be obvious. Ascertaining whether a counterfactual requires extrapolation with more than one explanatory variable requires only one additional generalizing concept: Questions that involve interpolation are values of the vector x which fall in the *convex hull* of X.

Formally, the convex hull of a set of points is the smallest convex set that contains them. This is easiest to understand graphically, such as via the example in Figure 3 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). In other words, if the right graph in Figure 3 were a cork board, and the dots were nails, the convex hull would be a rubber band stretched around all the points. With this definition of a convex hull, a counterfactual question x that appears outside the polygon requires extrapolation. Anything inside involves interpolation.

[Figure 3 about here.]

Although Figure 3 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 from a mathematical perspective, the task of deciding whether a point lies within the hull generalizes directly.

The concept of a convex hull is well known in statistics and has been used regularly to convey the idea of extrapolation and interpolation. However, it has almost never been used in practice for problems with more than a couple of explanatory variables. The problem is not conceptual, but rather computational. Identifying the hull with even a few explanatory variables can take an extraordinary amount of computational power. Doing it with more than about ten variables appears nearly impossible. Moreover, the problem of locating whether a counterfactual point lies within or outside the hull is itself a difficult computational problem that also has no solution known in the statistical literature.

In our technical companion paper, we solve this problem with a new algorithm capable of ascertaining whether a point lies within a convex hull quickly even for large numbers of variables and data points. We have also developed easy-to-use software, "WhatIf: Software for Evaluating Counterfactuals," that automates this convex hull memberhship check as well as implementing the other methods discussed in this paper (see Stoll, King and Zeng, 2005). The result is that the convex hull can now easily be used in any applied statistical analysis to sort counterfactual questions that may be close enough to the data to be answered by the empirical evidence from those that are farther away and may require more highly model dependent inferences.

2.2 How Far Is The Counterfactual from the Data?

The interpolation vs. extrapolation distinction introduced in Section 2.1 is a simple dichotomous assessment of the distance from a counterfactual to the data. In our experience, this distinction is sufficient in most instances to ascertain whether the data can support a counterfactual inference without excessive model dependence. In some instances, however, a finer distinction is warranted. For example, points just outside the convex hull are arguably less of a problem than those farther outside, and they are clearly closer to the data and, by our proof, less model dependent. Another related issue is that it is theoretically possible (although probably empirically infrequent) for a point just outside the interpolation region defined by the convex hull of X to be closer to a large amount of data than one inside the hull that occupies a large empty region away from most of the data. Thus, in addition to assessing whether a counterfactual question requires interpolation or extrapolation, we also more explicitly measure the distance from the counterfactual to the data.

Our goal here is some measure of the number or proportion of observations "nearby" the counterfactual. To construct this quantity, we begin with a measure of the distance between two points (or rows) x_i and x_j based on Gower's (1971) metric (which we call G^2). It is defined simply as the average absolute distance between the elements of the two points, divided by the range of the data:

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

where the range is $r_k = \max(X_{.k}) - \min(X_{.k})$ and the min and max functions return the smallest and largest elements respectively in the set including the values of the kth explanatory variable. 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 continuous and discrete data.[1] Since the counterfactual x may be outside the convex hull of X, our version of G^2 may range anywhere from zero on up. Thus, $G^2 = 0$ indicates that x and the row in question of X are identical, and the larger G_{ij}^2, the more different the two rows are. (If G^2 is greater than 1 for any row of X, then the counterfactual x lies outside the convex hull of X, but the reverse does not necessarily hold.) We interpret G^2 as *the distance between the two points as a proportion of the distance across the data*, X. So a distance between two points of $G^2 = 0.3$ means that to get from one point to the other, we need to go the equivalent of 30% of the way across the range of the data set.

With G^2 applied to our problem, we need to summarize n numbers, the distances between the counterfactual and *each* row in the data X. If space permits, we suggest presenting a cumulative frequency plot portraying vertically 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 typically happen if many counterfactuals need to be evaluated, any fixed point on this graph could be used as a one-number summary. Our recommendation for a rule of thumb in defining observations that are sufficiently close to the counterfactual to make for relatively reliable inferences is to use the fraction (or number) of observations in the data with distances (values of G^2) less than the "geometric variability" (GV) of X — which is roughly the average distance among all pairs of observations in the data. Then we could report the fraction of rows in the data with G^2 values less than one GV. We interpret the resulting measure as the fraction of the observed data *nearby* the counterfactual. We have found this rule of thumb to be useful in practice for determining the effective number of observations available to make inferences without high levels of model dependence

Observations farther than one GV away from the counterfactual normally have little empirical content for inference about the counterfactual, and can produce considerable

model dependence. Researchers should consider downweighting or even discarding these observations from the data, unless they are in the unusual situation of being certain that their model specification is correct. Of course, this is only a rule of thumb and so more data conserving rules could be applied (such as discarding data only 1.5 or 2 GVs away from the counterfactual) or even rules that result in less model dependence, depending on how much confidence one puts in the chosen model.

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., 1998). See King and Zeng (2002) for an independent evaluation.

These data are among the most extensively used in this area, in part because the authors had considerable resources from 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 main limitation on types and especially combinations of data the task force could collect 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, no matter what the dependent variable. Thus, to the extent that we find that certain counterfactual questions of interest are unanswerable, our point regarding problems in the literatures on the effects of democracy are all that much firmer.

After elaborate searches Esty et al. (1998) used as explanatory variables 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) improved their forecasts by, among other things, adding to these the fraction of the population in the military, population density, and legislative effectiveness.

The task force's dependent variable is the onset of state failure, but since we do not require specifying the dependent variable, our analyses apply to all dependent variables one might ever want to use. "What would happen if more of the world were democratic" is a question that 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.

Of course, just because our analysis applies to all possible dependent variables, the subject of any one article will normally be one or a small number of these. To see how widely our analyses apply, we began collecting other articles in the field that use a set of explanatory variables with a fair degree of overlap with the set used here, and stopped at twenty after searching only the last few years. The methods presented in this section would need to be repeated to draw more precise conclusions from each of these other articles, but the overlap in the explanatory variables was sufficient to infer that the results presented here will likely apply at least roughly to a large number of articles in the field.

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. Before turning to empirically reasonable counterfactuals, we begin with examples that are deliberately extreme. Extreme examples are of course useful for ensuring expository clarity, but they are also useful here since, although almost no serious researcher would expect the data to provide information about such counterfactuals if intentionally asked, almost all empirical analysts estimating the effects of democracy have implicitly asked precisely these questions.

This is always the case when all observations are used in the estimation and causal effect evaluation, as is typical in the literature. So although the two examples we now introduce are obviously extreme, we show that many actually asked in the literature are in fact also quite extreme.

Our first extreme counterfactual is to suppose that Canada in 1996 had become an autocracy, but its values on other variables remained at their actual values? We find, as we would expect, that this extreme counterfactual is outside the convex hull of the observed data and therefore requires extrapolation. In other words, we can ask what would have happened if Canada had become autocratic in 1996, but we cannot use history as our guide, since the world (and therefore our data) includes no examples of autocracies that are similar enough to Canada on other measured characteristics. 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.

We now ask two counterfactual questions that are as obviously reasonable as the last two were unreasonable. Thus, we ask what would have happened 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 probably not much model dependence. In other words, the world has examples of autocracies that are like Poland in all other measured respects, and so history can be our guide. 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 therefore also require only interpolation and little model dependence to draw inferences.

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 with our rule of thumb by asking what fraction of observations in our data are within one GV of G^2, which is approximately 0.1 (i.e., an average distance that is equivalent to 10% of the distance from the minimum to the maximum values on each variable in X). Essentially no real country-years are within 0.1 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. Recall that we do not need all the data in our collection to be near a counterfactual, only as much as needed to base our inferences on.

We now examine a larger sets of counterfactuals all at once. We start with all variables set at their actual values and then 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 full democracies and 4,039 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 would require extrapolation and would thus be highly model-dependent regardless of the model applied or dependent variable analyzed. This is quite important, since the usual practice of stacking up all the data, running an analysis, and interpreting the coefficients in standard ways is equivalent to directly evaluating these counterfactuals.

[Table 1 about here.]

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, but they are crucial for drawing valid inferences without high degrees of model dependence.

The first few columns of Table 1 break down these average results for counterfactuals from three different regions. The rest of the table provides the fraction of countries within a modified Gower metric of about one GV, or 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 rest of the data do not add much empirical content and generate considerable model dependence.

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 approximately to almost every such analysis with democracy as an explanatory variable in evert field with the same or similar control variables, regardless of the choice of dependent variable. The results here appear to suggest that many inferences in these fields (or most countries within each analysis) have little information content for the questions being posed and are highly model-dependent. Consequently, many conclusions are based more on unverifiable assumptions about the model than on empirical data. The result varies by region and by counterfactuals, and it would of course vary more if we changed the set of explanatory variables. We can only really know for sure by applying the methods introduced here to these other data sets, but no matter how you look at it, the problem of reaching beyond one's necessarily limited

data comes through in Table 1 with clarity.

Numerous interesting case studies could 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 accurately answered 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.

2.4 Counterfactuals About UN Peacekeeping

In "the first quantitative analysis of the correlates of successful peacebuilding and of the contribution of UN operations to peacebuilding outcomes," Doyle and Sambanis (2000, p.782) build and analyze a data set of 124 post-World War II civil wars. They characterize their results as firm enough to go beyond merely academic conclusions and to provide "broad guidelines for designing the appropriate peacebuilding strategy" (p.779) in practice. This work opens up a new area of quantitative analysis about an important public policy question for our field. We follow their lead and study the authors' "main concern" — "how international capacities, UN peace operations in particular, influence the probability of peacebuilding success" (p.783). Applying our methods, we found that the empirical conclusions offered in the article on this issue depend mostly on statistical modeling assumptions rather than empirical evidence. We do not address the veracity of the article's conclusions, only the weight of the data used to support them, and of course the authors should not be faulted for being unaware of methods we introduce here, years after their article was published. We also do not address the nine other hypotheses they test or other methodological issues raised by their analysis.

Doyle and Sambanis were helpful in providing us their data. We begin our analysis by replicating their key logistic regression model, numbered A8 in their article (Doyle and Sambanis, 2000: Table 3, p.790). Other models (each with different measures of UN intervention or other variables) in the article showed no effect of any specific type of UN intervention considered. It was therefore only the final specification in their Model A8 that the authors offered to support the article's key conclusion that "multilateral United Nations peace operations make a positive difference . . . and are usually successful in ending the violence" (abstract, p.779).

Our replication appears in Table 2, marked "original model". Doyle and Sambanis report results in odds ratios, whereas we report the more traditional logit coefficients, but the replication is otherwise exact.

[Table 2 about here.]

The theoretical justification for Doyle and Sambanis' logit specification comes from what they call their "interactive model," which posits peacebuilding success as the result of interactions among the level of hostility, local capacities, and international capacities such as the UN involvement. Their main concern is the effect of multidimensional UN peacekeeping operations, which include "missions with extensive civilian functions, including economic reconstruction, institutional reform, and election oversight," and which they find "are extremely significant and positively associated with peacebuilding" (p.791). In their original model replicated in Table 2, this is their UNOP4 variable, which is dichotomous: coded 1 for the seven multidimensional UN peacekeeping operations in their data and 0 for all other observations. Clearly the result is considerably larger than its standard error, even given the small number of observations available. When translated into an odds ratio, which is Doyle and Sambanis' preferred form, the odds of peacebuilding

success with a multidimensional UN peacekeeping operation is 23 times larger than with no such operation, holding constant a list of potential confounding control variables. We return to this remarkable result in Section 3.7 when we discuss causal inferences.

In this section, we examine the counterfactuals of interest. Assessing the causal effects of multidimensional UN peacekeeping operations implicitly involves asking the question: In civil wars with multilateral UN involvement, how much peacebuilding success would we have witnessed if the UN had not gotten involved? Similarly, in civil wars without UN involvement, how much success would there have been if the UN had gotten involved? In other words, the goal is counterfactual predictions with the dichotomous UNOP4 variable set to 1−UNOP4, which is one counterfactual for each observation. To begin with, we check how many counterfactuals are in the convex hull of the observed data. We found *none*. That is, every single counterfactual in the data set is a risky extrapolation rather than what would have been a comparatively safer interpolation. We also computed the Gower distance of each counterfactual from the data and found that few of the counterfactuals were near much of the data. For example, for all counterfactuals, an average of only 1.3% of the observations were within one GV (which is 0.11 in these data). Thus, not only are the counterfactuals all extrapolations, but they do not lie just outside the convex hull and are instead mostly fairly extreme extrapolations well beyond the data. These results strongly indicate that the data used in the study contain little information to answer the key causal question asked, and hence the conclusions reached there are based more on theory and model specifications than empirical evidence.

We now proceed to give relatively simple examples of the consequences of this result in terms of model dependence. We begin by making only one change in the logit specification by including a simple interaction between UNOP4 and the duration of the civil war, leaving the rest of the specification as is. (This is of course only one illustrative example, and not

the only aspect of the specification sensitive to assumptions.) Including this interaction would seem highly consistent with the "interactive" theory put forward in the article, and so would not seem possible to exclude on theoretical grounds alone. Excluding this interaction, which the original specification does, is equivalent to assuming that the effect of UN peacekeeping operations is identical for wars of all durations (except for the trivial assumed nonlinearities due to the logit model). Unfortunately, nothing in the theory expounded in the article, or in other literature in the field, justifies the use of such an assumption without empirical testing.

The result of this new specification is given in the second set of columns in Table 2. The coefficient on the interaction is positive and clearly distinguishable from zero (the p-value is 0.001), with a slightly higher likelihood and pseudo R^2 values, and so represents clear evidence by the usual rules of thumb to indicate that this model might even be preferred to the original one. To be clear, however, we do not necessarily favor this model or the original; we present both as two of many plausible alternatives not ruled out on the grounds of theory or data fitting. Moreover, no appropriate theory of statistical inference or analysis suggests whether to use the original or modified model to draw inferences without empirical testing. Thus, we consider the decision about whether to estimate the coefficient on the interaction as compared to fixing it to zero (or, equivalently, excluding it) as an apparently minor specification decision and now show how remarkably sensitive inferences are to this choice.

We now offer the left graph in Figure 4, which plots predicted values from both models based on the actual values of UNOP4 and the other explanatory variables. This graph shows that almost all the predicted values from the two models fall on the 45 degree line, which marks identical predictions. Only six points fall markedly off the 45 degree line. Of these points, we marked with a circle the five points for which the model with the

interaction fits the in-sample data better than the original model, and with a square the one point for which the original model fits better than the modified model. However one looks at the results in this graph, we conclude that the two models give extremely similar in-sample "factual" predictions. This means that, except for these six points, the models are indistinguishable on the basis of the observed data.

[Figure 4 about here.]

We now turn to the counterfactuals of interest by setting UNOP4 to $1 - \text{UNOP4}$, leaving all other variables at their observed values, computing the same predicted values, and redoing the same plot. Thus, this is the central question implied by Doyle and Sambanis's (2000) analysis: What happens if all the civil wars with multidimensional UN intervention did not have such an UN intervention, and all the civil wars without such UN interventions did have them. The logit model is effectively used in the article to evaluate these counterfactual effects. The graph on the right in Figure 4 gives the results by again plotting the predictions for the original vs. modified models. The result could hardly be more dramatic, with very few of the points anywhere near the 45 degree line. That is, for any value of the probability from the original model on the vertical axis (say 0.5 for example), the probabilities from the modified model are spread horizontally over almost the entire range from 0 to 1. This dramatic result indicates that these two models — that differ only very slightly in their fit to the in sample data — are giving wildly different counterfactual predictions, with very little relationship between them. Of course, these are precisely the results we would expect when counterfactual predictions are well outside the convex hull and thus confirms the prediction suggested by the convex hull test: Although the two models we chose fit the data almost identically, their counterfactual predictions are completely different because the counterfactuals are far from the data.

The counterfactual questions asked in this analysis are undoubtedly very important, but this analysis demonstrates that they cannot be reliably addressed by the data used.

We return to the analysis of these data, as well as the massively different substantive implications of these results, in Section 3.7.

3 Causal Inference

We now turn to causal inference and the counterfactual evaluation necessary as part of causal inference. We start with a definition of causal effects and then our decomposition of the bias in estimation, and finally a discussion of the components of bias. We devote the most space to discussing the components of bias due to interpolation and extrapolation, during which we show how the techniques introduced in Section 2 can also help solve an existing problem in causal inference. We illustrate with analyses in the same data used in Section 2.4 on UN peacekeeping.

3.1 Causal Effects Definition

To fix ideas, we use a version of the democratic peace hypothesis as a running example, which holds that democratic dyads are less conflictual than other dyads. Let D denote the "treatment" (or "key causal") variable where $D = 1$ denotes a democratic dyad and $D = 0$ denotes a nondemocratic dyad. The dependent variable is Y, the degree of conflict (but our discussion generalizes to all other dependent variables too).

To define the causal effect of democracy on conflict, denote Y_1 as the degree of conflict that would be observed if the dyad contained two democracies and Y_0 as the degree of conflict if this dyad were not both democracies. Obviously, only either Y_0 or Y_1 are observed for any one dyad at any given time, but not both, since (in our present simplified formulation) a dyad either is or is not democratic. This is known as the fundamental problem of causal inference (King, Keohane and Verba, 1994).

In principle, the democracy variable can have a different causal effect for every dyad in the sample. We can then define the causal effect of democracy by averaging over all dyads, or for the democratic and nondemocratic dyads separately (or for any other subset of dyads). For democratic dyads, this is known as the "average causal effect among the treated," which we define as follows:

$$\begin{aligned} \theta &= \mathrm{E}(Y_1|D=1) - \mathrm{E}(Y_0|D=1) \\ &= \text{“Factual”} - \text{“Counterfactual”} \end{aligned} \tag{2}$$

We call the first term — the average level of conflict among democratic dyads — factual since Y_1 is observable when $D = 1$. We refer to the second as counterfactual because Y_0 — the degree of conflict that would exist in a dyad if it were not democratic — is not observed and indeed is unobservable in democratic dyads ($D = 1$). (The causal effect for nondemocratic dyads ($D = 0$) is directly analogous and also involves factual and counterfactual terms.)

Although medical researchers are almost always interested in the average causal effect among the treated θ, political scientists are also interested in the average causal effect for the entire set of observations,

$$\gamma = \mathrm{E}(Y_1) - \mathrm{E}(Y_0), \tag{3}$$

where both terms have a counterfactual element, since each expectation is taken over all dyads, but Y_1 is only observed for democratic dyads and Y_0 only for nondemocratic dyads. These definitions of causal effects are used in a wide 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, the Mexico-Spain dyad with all its attributes fixed but with the regime type in both changed to autocracy. Of course, we can easily evaluate how reasonable it is to ask about this counterfactual in one's data with the

methods already introduced in Section 2: By checking whether x falls in the convex hull of the observed X and computing the distance from x to X. In addition, since x has only one counterfactual element we show that we can easily consult another criterion, whether x falls on the *support* of X, although we discuss some problems with this alternative in Section 3.6. The support of X is the range of values of X that are possible (i.e., have positive density) whether or not they occur in our data.

In real applications, the true causal effect, θ or γ, is unknown and needs be estimated. In Section 3.2, we discuss the sources of potential problems in using observational data to estimate these causal effects. We focus on θ there for expository purposes as is usual in the statistical and econometric literature. However, unlike prior literature, our companion paper includes proofs that are generalized to accommodate these quantities of interest to political science to show that our results also hold for the effect on nondemocracies and for the overall average treatment effect, γ, as well.

3.2 Bias Decomposition

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 a constant and D):

$$d = \text{mean}(Y|D=1) - \text{mean}(Y|D=0) \tag{4}$$

which is the average level of conflict in democratic dyads minus the average level of conflict in nondemocratic dyads. To identify the sources of potential problems using observational data in causal inference, we now present a new decomposition of the bias involved in using the simple difference in means estimator 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 clarify the threats to causal inference in our discipline, as well as to focus in on where

counterfactual inference is most at issue. In addition to identifying another key component of bias, we also present the decomposition for both quantities of interest, γ and θ, whereas Heckman et al. (1998b) only derived the result for the latter. Both results appear in our technical companion paper and require a fair amount of mathematical derivation (they are not merely analogies). Yet the results are simple. For θ, we show that,

$$\text{bias} \equiv \mathrm{E}(d) - \theta = \Delta_o + \Delta_p + \Delta_e + \Delta_i \tag{5}$$

We derive the equality and give the precise mathematical definition of the terms Δ_o, Δ_p, Δ_e, and Δ_i our technical companion paper. These four terms denote exactly the four sources of bias in using observational data, with the subscripts being mnemonics for the components. The bias components are due to, respectively, omitted variable bias (Δ_o), post-treatment bias (Δ_p), interpolation bias (Δ_i), and extrapolation bias (Δ_e). Briefly, Δ_o is bias due to omitting relevant variables such as common causes of both the treatment and the outcome variables; Δ_p is bias due to controlling for the consequences of the treatment; Δ_i is bias that can result if not properly adjusting for included controls within the region of the data; Δ_e is bias from extrapolating beyond the range of data in adjusting for included controls. We now explain and interpret each of these components in more detail with particular focus on extrapolation bias, including a discussion of how to use the methods we developed in Section 2 to help identify extreme counterfactuals in causal inference.

3.3 Omitted Variable Bias

The absence of all bias in estimating θ with d would be assured if we knew that it was 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 controlling for Z, treatment assignment is effectively random — that is, if we measure and control for the right set of control variables (those that are causally prior to and correlated with D and affect Y after controlling for D), then the first component of bias vanishes: $\Delta_o = 0$. Thus, this first component of bias, Δ_o, is due to pertinent control variables being omitted from X. This is the familiar omitted variable bias, which can plague any model.

Figure 5 illustrates omitted variable bias by plotting hypothetical data on a dependent variable vertically and a control variable horizontally. The treatment variable values are labeled in the graph. If we ignore the control variable, and thus project all the points to the left axis, we are left with two histograms. The histograms mostly overlap, but the control group (indicated by the dashed line) has a higher mean than the treated group (the solid line). However, if we adjust for the control variable Z, and thus look at the spread of the points in the body fo the graph, the causal effect estimate is revealed by the vertical distance between points for given values of Z. Where data are available, we see that the treated group data points are clearly above that for the control group data points, thus reversing the original conclusion of no effect or a negative treatment effect. (Ranges of Z where points do not exist for either the treatment or control group require extrpolation, about which more in Section 3.6.)

[Figure 5 about here.]

Since endogeneity bias and selection bias can be written as omitted variable bias, Δ_o encompasses these problems as well. In regression-type models, endogeneity bias, selection bias, and omitted variable bias each cause inferential problems by inducing a correlation between the explanatory variables and the error term. If we control for the correct variables, then it is sometimes possible to eliminate these problems. In the omitted

variable case, we can avoid the bias by including relevant variables such as common causes of D and Y. Similarly, we can avoid the biases due to nonrandom selection if we control for the probability that each unit is selected into the sample, and we can eliminate endogeneity bias by including in the controls covariates that eliminate the conditional relationship between X and the error term.

3.4 Post-Treatment Bias

Post-treatment bias is the second component of bias in our decomposition, Δ_p, and it deviates from zero when some of the control variables Z are at least in part consequences of the key causal variable D. If Z includes these post-treatment variables, then when the key causal variable D changes, the post-treatment variables may change too and the plan to interpret the model as revealing the effect of the treatment "holding other variables constant" becomes impossible.

As a simple example that illustrates the bias of controlling for post-treatment variables, suppose we are predicting the duration of a African dictatorship using the unemployment rate as the key explanatory variable. If we control for the existence of a well armed cabal inside the palace gates five minutes before a coup attempt is launched, our estimate of the effect of unemployment 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 the presence of natural resources in the country, since it cannot be a consequence of unemployment but may be a common cause of both the explanatory and dependent variables. 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; King, 1991). It is well known in the statistical literature, but is assumed away in most models and decompositions. 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 and especially in 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 surgical 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_p, we need to ensure that we control for no post-treatment variables, or that the distribution of our post-treatment variables do not vary with D. If this assumption holds, then $\Delta_p = 0$ and so this component of bias in (5) vanishes.

In our field, unfortunately, we almost always need to consider both Δ_o and Δ_p together, and in many situations we cannot fix one without making the other worse. The same is not true in some other fields (which is perhaps the reason the Δ_p component was ignored by Heckman et al., 1998), but it is rampant in ours. Unfortunately, the news gets worse,

since even the methodologist's last resort — try it both ways and, if it doesn't make a difference, ignore the problem — does not work here. Rosenbaum (1984) studies the situation where we run two analyses, one including and one excluding the variables that are partly consequences and partly causes of X. He shows that the true effect could be greater than these two or less than both. It is hard to emphasize sufficiently the seriousness of this problem and how prevalent it is in comparative politics and international relations.

Although we have no general solution to this problem, we can offer one useful way to avoid both Δ_p and Δ_o in many practical applications. Aside from choosing better research designs in the first place, of course, our suggestion is to study what we call *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 we have specified a parametric model with both variables, we can study this question by *simultaneously* moving both GDP and democracy while holding constant other variables at (say) their means. An alternative would be to recode the two variables into one on, as much as possible, a single dimension.

If this alternative formulation provides an interesting research question, then it can be studied without bias due to Δ_p 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_o, since both variables would be part of the treatment and could not be potential confounder's. 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. Either way, this is a serious problem that needs to move higher on the agenda of political methodology.

3.5 Interpolation Bias

Even if we can be sure that no omitted variable or post-treatment biases exist, we still have to control for the observed pre-treatment variables properly. The two remaining components of bias — interpolation bias and extrapolation bias — both have to do with correctly identifying the necessary control variables but failing to adjust for them properly. Interpolation bias or Δ_i results from adjusting incorrectly for the correct control variables in regions of interpolation, and extrapolation bias results from adjusting for the correct controls where data are needed but do not exist. Interpolation bias is normally the less serious of the two since it is more amenable to empirical testing.

Interpolation bias may exist in the simple difference in means estimator if the measured control variables Z are related in any way to the treatment variable, that is if the multivariate density of Z for the treatment group differs from that for the control group (within the region of interpolation). If in addition to these density differences Z also affects the outcome variable, then interpolation bias will exist if the density differences in Z are not properly adjusted.

When using a parametric model to adjust for control variables, this component of bias arises from controlling for Z with the wrong functional form. For example, in an application without post-treatment bias, with all control variables that could cause bias identified, and where extrapolation is unnecessary, our estimator could still generate bias by choosing a linear model to adjust for controls if the data were generated from a quadratic. Fortunately, standard regression diagnostics are quite useful for checking model fit within

the range of the data. Ultimately, whatever method of adjustment is used, the two multivariate histograms of Z for the control and treatment groups need to be the same for interpolation bias to be eliminated. We provide further insight into interpolation bias during our discussion of extrapolation bias, to which we now turn.

3.6 Extrapolation Bias

The last component of bias, and the one most related to the central theme of the paper, is extrapolation bias. This component is the second of the two that arise from not adjusting or improperly adjusting for identified control variables.

Extrapolation bias may arise when the support (or possible values) of the distribution of Z for the treatment group differs from that of the control group. That is, there may be certain values of Z that some members of one group take on but no members of the other group possess. For example, we might observe no full democracies with GDP as low as in some of the autocracies, but still somehow need to control for GDP. Intuitively, these autocracies have no comparables in the data, and so are not immediately useful for estimating causal effects. To make causal inferences in situations with nonoverlapping support, we must therefore either eliminate the region outside of common support — as is a standard practice in statistics and medicine — or attempt to extrapolate to the needed data (e.g., autocracies with high GDP), such as by using a parametric model — as is standard practice in political science and most of the other social sciences. As we demonstrate in Section 2, extrapolation in forecasting involves considerable model dependence. The same issue applies in causal inference, as we discuss here. Thus, unless we happen to be in the extraordinary situation where a known theory or prior evidence makes it possible to narrow down the possible models to one, or where we happen to guess the right model, we will be left with extrapolation bias, $\Delta_e \neq 0$.

3.6.1 Illustration with a Single Control Variable

Figure 6 illustrates some key issues involved in data that generate the need to extrapolation in causal inference. The figure also illustrates the connection between the problems of extrapolation in causal inference and extrapolation in forecasting and what if questions discussed earlier. Figure 6 plots hypothetical data on the dependent variable vertically and a single control variable Z horizontally. The treatment and control groups are labeled and the points are clearly separated in the figure. To estimate the causal effect in these data, we make comparisons between the treatment and control groups on the vertical axis (which corresponds to the outcome variable). The key extrapolation problem is that there exist no treated units for values of $Z > 2$ where some control data do exist, and so any comparison between the treated and control groups in this region would be based on extrapolating the treatment group data from where it is observed to where it is needed. In other words, a study seeking to estimate a causal inference from data where extrapolation is necessary is the same problem in that region as not having data for one of the two groups at all.

[Figure 6 about here.]

As the figure shows, the two models fit to the treated data, one linear and one quadratic, fit the treated data almost identically, but in the region to which the counterfactual extrapolations are needed (i.e., where control units exist but treated units do not), the difference between the models is vast. This illustrates model dependence of course, but it also illustrates extrapolation bias, since at least one of the models shown must be false in the extrapolation region, so if used would generate bias and make $\Delta_e \neq 0$. Since we have no data to test which model is appropriate, or whether both are wrong in the extrapolation region, we have no means to rule out extrapolation bias based on empirical evidence.

Interpolation bias could be seen in the figure if the different functional forms fit to the treated data differed in sample. If that were the case (and it is not as drawn), then bias would result if the estimation model were not close to the model that represented the data. In practice, because model dependence is much less of an issue in areas of interpolation (or on the common support) than in extrapolating, interpolation bias can often be detected and corrected in ways that extrapolation bias cannot.

Extrapolation bias is far more difficult than interpolation bias. If we use the data outside the region common support, we must extrapolate and will therefore have some degree of model dependence and risk some bias for almost any model one would choose. Alternatively, we can delete nonoverlap data, which eliminates the need to extrapolate. Of course, this procedure would fail to produce any estimates at all in applications where no data lie on the common support, a problem with some prevalence in our field. If some data do lie within the common support region, and the quantity of interest is the average treatment effect (γ in Equation 3), dropping observations outside of common support will produce bias by definition, as it changes the population of inference and the quantity of interest. Similarly, in the situation where we convince ourselves that we are interested only in the average treatment effect on the treated (θ in Equation 2), dropping treatment units not on common support will result in bias by changing the population of inference.

Although extrapolation bias is hard to correct without access to better data or willingness to change the population of inference (and thus the research questions), identifying the regions of extrapolation is important in all applications. 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.

3.6.2 Identifying Multivariate Extrapolation Regions with the Convex Hull

In the simple case when a model contains just one pretreatment variable Z, we can simply plot a histogram of this variable for the treated ($D_i = 1$) units on the same scale as a histogram of this variable for the control ($D_i = 0$) units, and compare them. Areas requiring extrapolation can easily be identified as the areas of the histograms that do not overlap. (Interpolation bias can arise where the histograms overlap but differ.)

In most real applications, of course, Z contains many control variables, and so identifying the extrapolation region requires comparing multidimensional histograms (as estimates of multivariate densities) for the treatment and control groups. For more than a few explanatory variables, this is a difficult or impossible task. In practice, scholars have checked common support by first collapsing their data to one variable via what is known as the "propensity score," but we show in our companion paper that this widespread use of the propensity score is invalid. Fortunately, as we now show, a simpler procedure is available based on the same convex hull concept and algorithm already introduced.

If we are interested in estimating the average treatment effect on the treated (θ in Equation 2), then we simply discard any control units for which Z is not within the convex hull of the treated units Z. (Even if some of the treated units are outside the convex hull of the control units and thus would require extrapolation, they would not be omitted so that the quantity of interest remains the same, although it would be worth identifying them so that the source of the remaining model dependence is identified.)

If instead we are willing to change the quantity being estimated to something different, but is reliably estimated without high levels of model dependence, then we would also want to drop treated units that fall outside the convex hull of the control units. If this alternative is desired, we can consolidate the two steps and estimate the common support by the convex hull of the subset of observed Z within which the counterfactual points

$\{1-D, Z\}$ fall. In other words, begin with all the counterfactuals (which are $\{1-D, Z\}$). Then, select only those that fall within the convex hull of the observed data. Our estimate of the common support is then the convex hull of Z of this subset of the counterfactuals. *Thus, the same procedures for identifying whether points fall within the convex hull (as described in Section 2.1) can be used to identify the region of common support.* To avoid the risk of voids within the region of common support, we can use the Gower distance to assess whether any of the counterfactual points within the hull are far from any observed data.

This strategy has not been used in the literature before, mostly because ascertaining whether counterfactual points fall in the convex hull have not previously been viewed as feasible. Given our new algorithm for finding whether points fall within the hull, this strategy is now feasible, and easy to apply. Indeed, a key advantage of the strategy suggested here is that at least a good first cut at finding the region of common support can now be automated and easily included in standard statistical software. It is already included in the software that accompanies this paper, (Stoll, King and Zeng, 2005), and has also been implemented as part of a general purpose matching software package called MatchIt (Ho et al., 2005).

Discussion of extrapolation bias in the quantitative empirical literature in most of the social sciences focusing on causal inference is rare and relatively few studies attempt to diagnose this issue formally, much less to do anything about it. Yet, using data without complete common support produces highly model-dependent extrapolations to areas where no data exist, and thus inferences become based partly on theoretical modeling rather than empirical data analyses.

3.7 The Causal Effect of UN Peacekeeping

We now apply the ideas introduced in Section 3.2 to the Doyle and Sambanis (2000) example we first studied in Section 2.4. We focus on extrapolation bias, the component of causal effect estimation bias most relevant to the theme of this paper. In the earlier section, we showed how all the counterfactuals in these data were extrapolations and far from the convex hull and thus inferences about them were highly model-dependent. We now demonstrate the same point by the common support criterion, and show the consequence of this model dependence on the causal inferences of interest.

Identifying common support by using our convex hull check is easy. We merely observe how many of the counterfactuals that result by switching all multinational UN interventions to no intervention, and switching all noninterventions to interventions, are inside the convex hull of the observed data. In Section 2.4, we found that none of these counterfactuals are within the hull, and so the common support is empty — the data include *no* information with which to reliably compare the two groups and estimate the causal effects of interest. In other words, there exist no civil wars in the data without UN intervention that are sufficiently like the civil wars with UN intervention to construct an adequately comparable control group. Going forward in this situation will generally produce high levels of model dependence. In fields where scholars have paid attention to these issues, the data would be judged to contain no information about the quantity of interest and no estimates would be attempted unless strict assumptions were warranted.

Although the convex hull check is easy and fast, and the resulting meaning is clear (the absence of civil wars in the data set without UN intervention that are otherwise the same as the civil wars where the UN did intervene), it is not always easy to understand a high dimensional calculation like this. Thus, we also illustrate the problem with a couple of univariate checks in Figure 7. For example, the left graph in this figure shows a

simple histogram for whether the parties signed a treaty to end the civil war. The darker histogram shows that all UN interventions in these data were in civil wars with signed treaties, but the lighter histogram indicates that only about 20% of the other civil wars had signed treaties. Thus, any civil war without a signed treaty (the left bar, at 0, in the figure) is outside the common support and cannot be used as a control group to evaluate the effect of UN intervention. This is quite intuitive: The goal of the causal inference is to isolate the effect of UN intervention, and so we want a treatment group that differs from the control group only by intervention status. The problem is that the only way to use these data would be to take countries without a signed treaty and without UN intervention and to somehow guess using some model what their peacebuilding success would be if they had no UN intervention but had signed a treaty. Extrapolations like this are the source of model dependence.

[Figure 7 about here.]

The right graph in Figure 7 illustrates the same point for a continuous variable, pre-war per capita electricity consumption. The density of electricity consumption in civil wars with UN intervention are all clustered near the low end of the continuum. Anything above that is outside the area of common support.

A key point is that, in general, one-dimensional graphs like these can identify some areas *not* on the multivariate common support (here for example any civil war where the parties have not signed a peace treaty or where pre-war electricity consumption is high) — since what is not on the common support for any single control variable is certainly not on the multivariate common support — but is not sufficient to identify the multivariate common support. To do that, we must consider all the explanatory variables together. That is, it is not enough for some observations to meet the test in the left graph and a

different set of observations meet the test in the right graph. Observations on common support must meet the common support test (have overlapping multivariate histograms) for all variables simultaneously. Fortunately the convex hull provides a simple way to do this for as many dimensions as are needed.

Finally, we illustrate the model dependence that results from the fact that none of the data are on the common support for both the intervention and nonintervention groups. To do this, we now compare estimates of the causal effects in the two logit analyses in Table 2, which we showed fit the data almost identically and which differ only by one interaction term. From these logit models, we compute the marginal effects of UN peacekeeping operations as a function of the duration of the civil war, holding constant all other variables at their means. Figure 8 plots these results.

[Figure 8 about here.]

The vertical axis in Figure 8 is the marginal effect, which, conditional on the veracity of the logit model, is a causal effect. The horizontal axis is the duration of the civil war. The dotted line is the causal effect of UN peacekeeping estimated by the the model originally presented in Doyle and Sambanis (2000). Without a formal interaction term, the modest nonlinearities of the logit model allow the effect of UNOP4 to vary with civil war duration almost linearly, and the consequence is clear: The effect of UN peacekeeping operations is smaller for civil wars that have gone on for a longer time before the UN stepped in. This is quite a plausible result, since we might expect that long conflicts would be more difficult to resolve.

However, as it turns out, the empirical support for this result depends almost entirely on what are treated in the article as minor modeling assumptions not worth discussion much less detailed justification. This can be seen by examining the solid line in the

figure, which portrays the causal effect for the modified model. As can be seen, the effects for the two models are massively different. In the modified model, the probability of success is hardly affected by UN operations for relatively short wars, it increases fast, and then declines in parallel to the original model. The huge differences for shorter wars and the opposite slopes of the two lines in that region suggest diametrically opposed policy implications for UN missions: According to one model, wars with the shortest duration should receive the most attention from the UN, since that is where the model indicates that the UN can have its biggest effect; according to the other model, the same civil wars should receive the least attention from the UN. (Confidence intervals confirm that inferences from the two models using more than point estimates also differ with dramatically different policy implications.) Although the two models gave nearly identical fit to the factual data, the counterfactuals are far enough from the observed data to make conclusions highly sensitive to modeling assumptions. To be clear, no theory offered in Doyle and Sambanis (2000) or the literature would rule out one of the two models, and the data do not enable us to choose one either. We also do not prefer the modified model over the original and introduce the modification only as an illustration of the extreme model dependence that can result from using data outside of common support that requires extrapolation.

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 the future of military conflict if globalization led to 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 to them 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 or to at least address the problem in some way when they evaluate 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 and autocracies 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 history cannot be our guide since almost no evidence exists in our data with which to evaluate such a question. Although having small numbers of cases will often make finding a proper control group harder, in this example, having access to 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 — al-

most always implies questions without adequate empirical evidence to answer. 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.

Suppose we read about a model that fits the data exceedingly well, has a 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 inferences 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.

Notes

[1]Following standard practice in data analyses, ordinal explanatory variables are typically assumed interval or coded as a set of dichotomous variables. Nominal variables are usually coded as a set of dichotomies. With these changes, Equation 1 applies directly.

References

Doyle, Michael W. and Nicholas Sambanis. 2000. "International Peacebuilding: A Theoretical and Quantitative Analysis." *American Political Science Review* 94(4, December):779–801.

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.

Fearon, James D. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43(2, June):169–195.

Gelman, Andrew and Gary King. 1994. Party Competition and Media Messages in U.S. Presidential Election Campaigns. In *The Parties Respond: Changes in the American Party System*, ed. L. Sandy Maisel. Boulder, Colorado: Westview Press pp. 255–295. http://gking.harvard.edu/files/abs/partycomp-abs.shtml.

Gower, J.C. 1971. "A General Coefficient of Similarity and Some of its Properties." *Biometrics* 27:857–872.

Hastie, Trevor, R. Tibshirani and J. Friedman. 2001. *The Elements of Statistical Learning*. New York: Springer Verlag.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998. "Characterizing selection bias using experimental data." *Econometrika* 66(5):1017–1098.

Ho, Daniel, Kosuke Imai, Gary King and Elizabeth Stuart. 2005. "Matching as Nonparametric Preprocessing for Parametric Causal Inference." http://gking.harvard.edu/files/matchp.pdf.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

King, Gary. 1991. "Truth' is Stranger than Prediction, More Questionable Than Causal Inference." *American Journal of Political Science* 35(4, November):1047–1053. http://gking.harvard.edu/files/abs/truth-abs.shtml.

King, Gary and Langche Zeng. 2002. "Improving Forecasts of State Failure." *World Politics* 53(4, July):623–658. http://gking.harvard.edu/files/abs/civil-abs.shtml.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2, April):341–355. http://gking.harvard.edu/files/abs/making-abs.shtml.

King, Gary, Robert O. Keohane and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Kuo, Yen-Hong. 2001. "Extrapolation of Association Between Two Variables in Four General Medical Journals." (September). Forth International Congress on Peer Review in Biomedical Publication.

Kvart, Igal. 1986. *A Theory of Counterfactuals*. Indianapolis: Hackett Publishing Company.

Lebow, Richard Ned. 2000. "What's so Different About a Counterfactual?" *World Politics* 52(July):550–85.

Lewis, David K. 1973. *Counterfactuals*. Cambridge: 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(5):656–666.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.

Stoll, Heather, Gary King and Langche Zeng. 2005. "WhatIf: Software for Evaluating Counterfactuals.". http://gking.harvard.edu/whatif/.

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(2, April):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* 95(4, December).

Thorson, Stuart J. and Donald A. Sylvan. 1982. "Counterfactuals and the Cuban Missle

Crisis." *International Studies Quarterly* 26(4):539–571.

List of Figures

1 *Linear and quadratic models with equal fit to simulated data but massively different out-of-sample implications.* 50

2 *Each graph displays three members from one family of statistical models. The families are, clockwise from the upper left, logit, linear, exponential, and quadratic. Typically one family of models is chosen by the investigator and the statistical analysis chooses the member of the family that best fits the data. Whether a different family fits better instead is not considered by the statistical analysis program.* 51

3 *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 interval between the minimum and the maximum values of the observed data (as portrayed as the points farthest to left and right on the left graph). With two explanatory variables, the convex hull is a polygon with vertices at the extreme points of the data (as in the right graph). Neither graph portrays the dependent variable, since it is not needed to ascertain whether the counterfactual is an interpolation or extrapolation.* 52

4 *Sensitivity of Predictions to Small Changes in the Model. The vertical axes are predicted probabilities from the model in (Doyle and Sambanis, 2000) and the horizontal axes are from the same model with the addition of an interaction term (see Table 2). The left graph shows how both models produce almost identical probabilities; dots in this graph that are farther from the line are marked with a square if the original model fits better and a circle if the modified model fits better. The right graph makes predictions for all observations by switching the 0/1 value of the UN peacekeeping indicator variable while holding others constant. Note how the in-sample predictions from both models in the left graph are approximately as good, but the out-of-sample predictions differ wildly.* 53

5 An Illustration of Omitted Variable Bias. The vertical axis is Y, a dependent variable. The histograms are projections of the points in the graph to the left by ignoring the control variable Z on the horizontal axis. The dashed line histogram is for the control group, and solid line is for the treatment group. 54

6 Illustrating Nonoverlap and Density Differences: Hypothetical Data 55

7 *Univariate Checks for Common Support: Whether the parties signed a peace treaty (left graph) and energy consumption (right graph). Observations with UN interventions are the darker histogram on the left and a solid line on the right.* 56

8 *Causal Effect of Multidimensional UN Peacekeeping Operations* 57

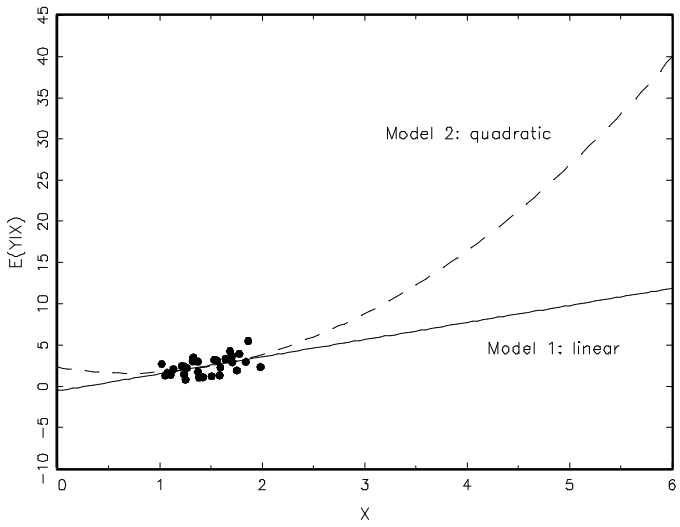

Figure 1: *Linear and quadratic models with equal fit to simulated data but massively different out-of-sample implications.*

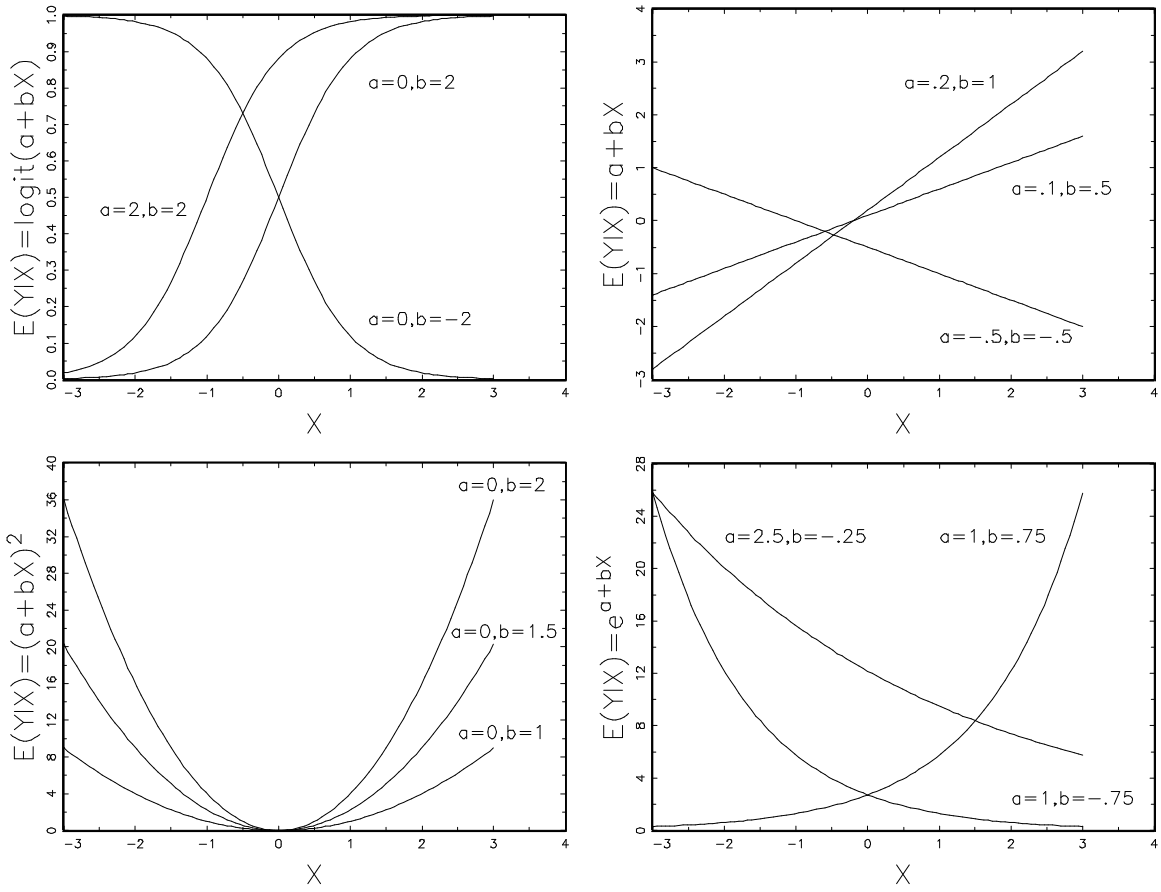

Figure 2: *Each graph displays three members from one family of statistical models. The families are, clockwise from the upper left, logit, linear, exponential, and quadratic. Typically one family of models is chosen by the investigator and the statistical analysis chooses the member of the family that best fits the data. Whether a different family fits better instead is not considered by the statistical analysis program.*

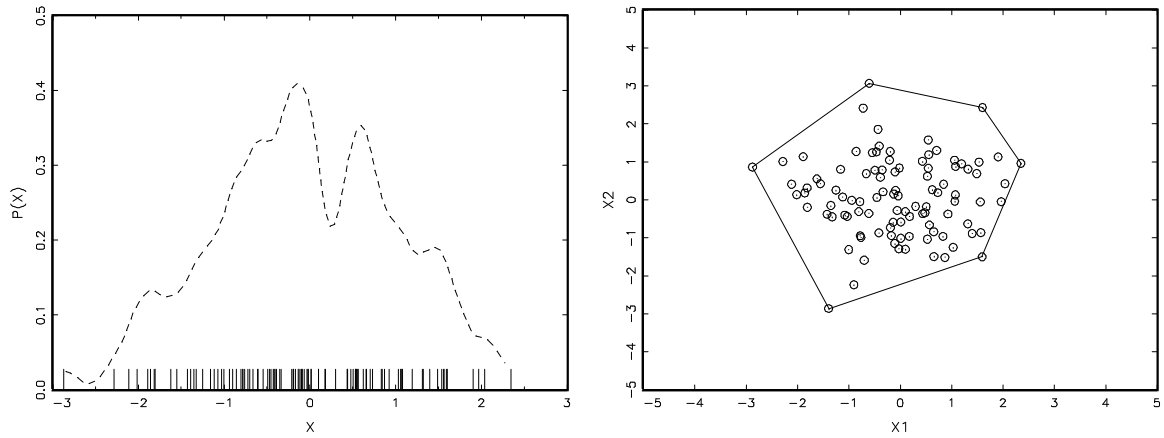

Figure 3: *Interpolation vs. Extrapolation: The convex hull of X is the smallest convex set that contains the data. Inference on points inside the convex hull requires interpolation, outside it requires extrapolation. With one explanatory variable, the convex hull is the interval between the minimum and the maximum values of the observed data (as portrayed as the points farthest to left and right on the left graph). With two explanatory variables, the convex hull is a polygon with vertices at the extreme points of the data (as in the right graph). Neither graph portrays the dependent variable, since it is not needed to ascertain whether the counterfactual is an interpolation or extrapolation.*

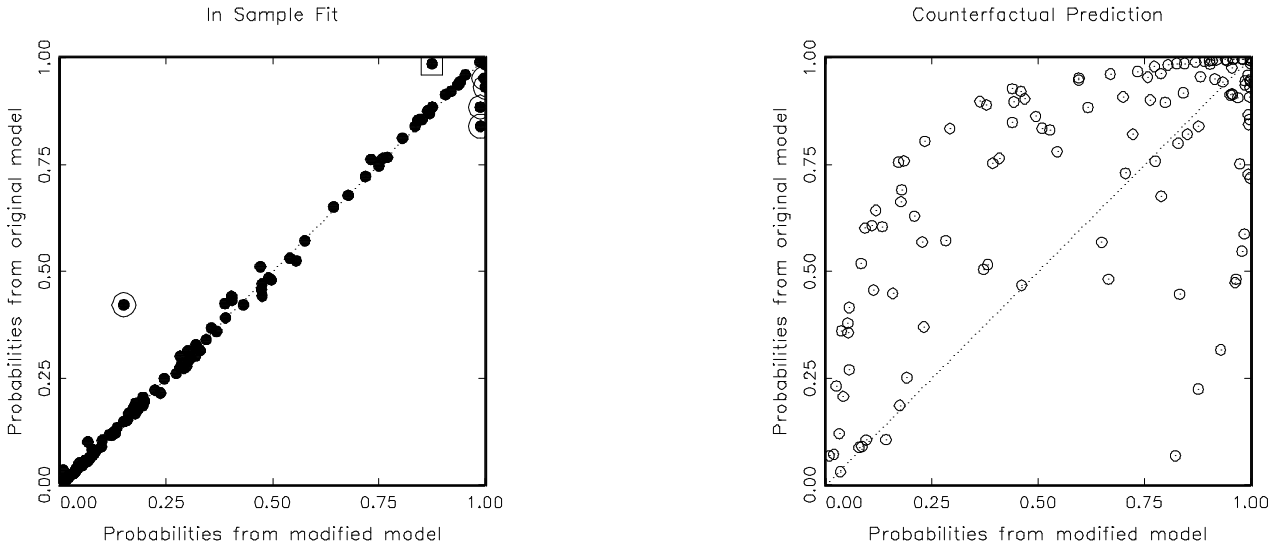

Figure 4: *Sensitivity of Predictions to Small Changes in the Model. The vertical axes are predicted probabilities from the model in (Doyle and Sambanis, 2000) and the horizontal axes are from the same model with the addition of an interaction term (see Table 2). The left graph shows how both models produce almost identical probabilities; dots in this graph that are farther from the line are marked with a square if the original model fits better and a circle if the modified model fits better. The right graph makes predictions for all observations by switching the 0/1 value of the UN peacekeeping indicator variable while holding others constant. Note how the in-sample predictions from both models in the left graph are approximately as good, but the out-of-sample predictions differ wildly.*

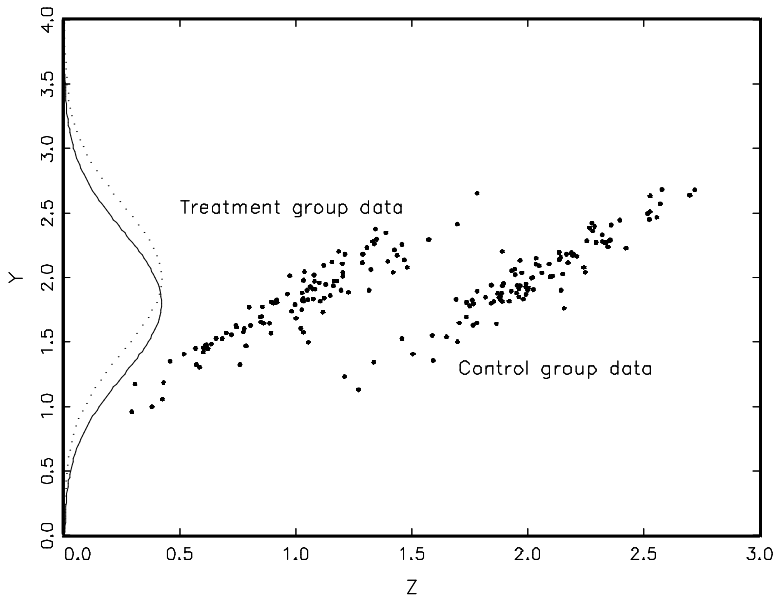

Figure 5: An Illustration of Omitted Variable Bias. The vertical axis is Y, a dependent variable. The histograms are projections of the points in the graph to the left by ignoring the control variable Z on the horizontal axis. The dashed line histogram is for the control group, and solid line is for the treatment group.

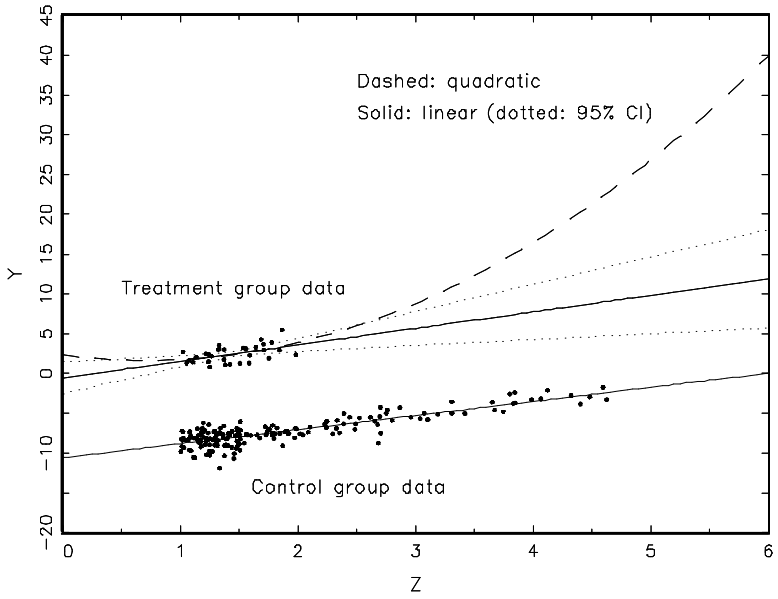

Figure 6: Illustrating Nonoverlap and Density Differences: Hypothetical Data

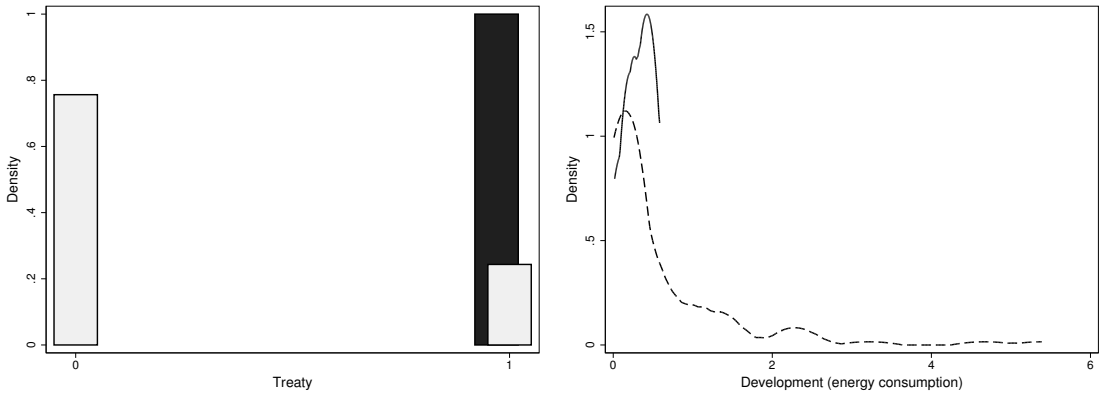

Figure 7: *Univariate Checks for Common Support: Whether the parties signed a peace treaty (left graph) and energy consumption (right graph). Observations with UN interventions are the darker histogram on the left and a solid line on the right.*

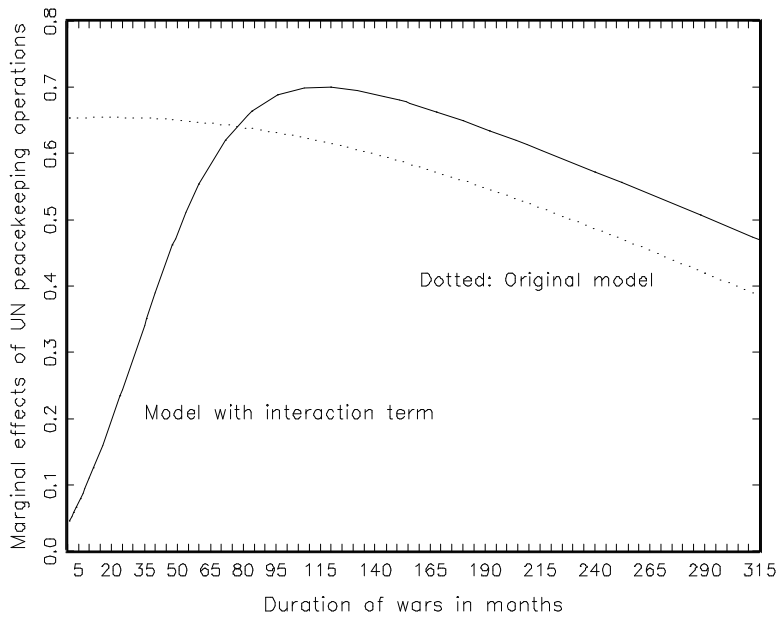

Figure 8: *Causal Effect of Multidimensional UN Peacekeeping Operations*

List of Tables

1 How Factual Are Counterfactuals About Democracy? 59
2 Peacebuilding Models With and Without Interaction Terms. The logit model on the left replicates (Doyle and Sambanis, 2000); the model on right is identical to the original except for the addition of an interaction term. 60

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

	Original Model			Modified Model		
Variables	Coefficient	Robust S.E.	P-value	Coefficient	Robust S.E.	P-value
Wartype	−1.742	.609	.004	−1.666	.606	.006
Logdead	−.445	.126	.000	−.437	.125	.000
Wardur	.006	.006	.258	.006	.006	.342
Factnum	−1.259	.703	.073	−1.045	.899	.245
Factnum2	.062	.065	.346	.032	.104	.756
Trnsfcap	.004	.002	.010	.004	.002	.017
Develop	.001	.000	.065	.001	.000	.068
Exp	−6.016	3.071	.050	−6.215	3.065	.043
Decade	−.299	.169	.077	−0.284	.169	.093
Treaty	2.124	.821	.010	2.126	.802	.008
UNOP4	3.135	1.091	.004	.262	1.392	.851
Wardur×UNOP4	—	—	—	.037	.011	.001
Constant	8.609	2.157	0.000	7.978	2.350	.000
N	122			122		
Log-likelihood	−45.649			−44.902		
Pseudo R^2	.423			.433		

Table 2: Peacebuilding Models With and Without Interaction Terms. The logit model on the left replicates (Doyle and Sambanis, 2000); the model on right is identical to the original except for the addition of an interaction term.