Advance Access publication December 14, 2008 *Political Analysis* (2009) 17:107–112
doi:10.1093/pan/mpn010

Empirical versus Theoretical Claims about Extreme Counterfactuals: A Response

Gary King
Department of Government and Institute for Quantitative Social Science, Harvard University, Cambridge MA 02138
e-mail: king@harvard.edu (corresponding author)

Langche Zeng
Department of Political Science, University of California, San Diego, La Jolla, CA 92093
e-mail: lazeng@ucsd.edu

In response to the data-based measures of model dependence proposed in King and Zeng (2006), Sambanis and Michaelides (2008) propose alternative measures that rely upon assumptions untestable in observational data. If these assumptions are correct, then their measures are appropriate and ours, based solely on the empirical data, may be too conservative. If instead, and as is usually the case, the researcher is not certain of the precise functional form of the data generating process, the distribution from which the data are drawn, and the applicability of these modeling assumptions to new counterfactuals, then the data-based measures proposed in King and Zeng (2006) are much preferred. After all, the point of model dependence checks is to verify empirically, rather than to stipulate by assumption, the effects of modeling assumptions on counterfactual inferences.

We are grateful for the attention devoted to our work (King and Zeng 2006, hereinafter KZ) in Sambanis and Michaelides (2008, hereinafter SM) and appreciate this opportunity to respond. KZ develops methods that reveal when analyses are model dependent without the laborious process of sensitivity testing; KZ does this by building on the general understanding that inferences on the empirical support of the data are less model dependent than interpolations which, in turn, are usually less model dependent than extrapolations. KZ put this position on firmer footing by proving the mathematical conditions under which it holds and solving an open problem in computational geometry to make it easy, in the kinds of data political scientists use, to determine whether an inference requires extrapolation. We (and King and Zeng 2007) also confirmed the advantage of these methods in checking model dependence in political science applications. SM does not question any of these points.

KZ uses the long-standing consensus definition of extrapolation in the statistical literature that is *data based* (Jevons 1874, 120; Hastie, Tibshirani, and Friedman 2001). SM

Authors' note: Easy-to-use software to implement the methods discussed here, called "WhatIf: Software for Evaluating Counterfactuals," is available at http://gking.harvard.edu/whatif. All information necessary to replicate the analyses herein can be found in King and Zeng (2008). *Conflict of interest statement*. None declared.

criticizes this choice as "too severe" (2008, 1) and "too conservative" (2008, 3, 6, 10, 12, 16, 17). SM uses instead a new definition that is *assumption based*, does not depend on the data, and defines extrapolation as an inference off the *population support* (the set of all possible values of X under an assumed population distribution) and causal inference as supposedly guaranteed free of model dependence when the treated and control groups are drawn from a common population distribution, regardless of the content or quantity of data or parameters to be estimated. Population support encompasses the empirical support and includes additional space in X only by assumption since the population distribution is of course almost never known in observational studies. Yet, the difference between SM's population support and our *empirical* support is the basis on which SM argue that our diagnostics, using the literature's data-based definition of extrapolation, are too conservative.

To facilitate discussion, we define terms standard in the literature and used in KZ. Without loss of generality, consider a simple example, with all observations on a single covariate X at 0, 1, or 3 and inferences about the expected value of the outcome variable Y, given some value x of X, $E(Y|X = x)$, such as estimated via the predicted value in a regression. Then, consider three types of inferences: (1) those *on the empirical support* of the data: $E(Y|X = 0)$, $E(Y|X = 1)$, and $E(Y|X = 3)$; (2) *interpolations*: $E(Y|X = x)$ for any x between 0 and 3 but excluding values at 0, 1, and 3; and (3) *extrapolations*: $E(Y|X = x)$ for any x less than 0 or greater than 3. Inferences that fall in categories (2) or (3) are called *counterfactuals*. The *convex hull* is the boundary marking off the extrapolation region, in this case points 0 and 3. These concepts are defined for the sample data, without an imaginary superpopulation, and require no hypothesis tests; they also apply when X contains many covariates.

By the standard definition of extrapolation, in our running example, $E(Y|X = 200)$ is a massive extrapolation from 0, 1, and 3 and, thus, highly model dependent, but it is on the population support of the Poisson distribution (the set of nonnegative integers) and so SM would deem it not model dependent if they chose to make that assumption for the population distribution. Yet, in almost no real data would we know the population density. And even if it were known, we would also need to assume that the functional form of the model for $E(Y|X)$ is sufficiently smooth and stable to extrapolate so far from the data, which would be a heroic assumption indeed. The model dependence here would be easy to confirm by proper sensitivity testing. The Poisson model assumption seems extreme, but it is not as extreme as SM's standard normal densities for which the empirical support is almost always in [–5, 5], but the population support is $(-\infty, \infty)$; thus, SM's definition implies that no counterfactuals are ever model dependent, which is clearly false.

The same logic applies to causal inference where one needs treated and control groups that are identical in-sample in your data, rather than on average across hypothetical repeated experiments. This explains why matched-pair (MP) randomized experiments, which guarantee exactly matched treatment-control groups in-sample, are superior to complete randomization (CR) designs, which only guarantee equivalence on average across experiments (Box, Hunger, and Hunter 1978; Imai, King, and Stuart 2008).[1] SM's standard for their simulations is the inefficient CR design, which explains how they can incorrectly conclude that the convex hull is too conservative. In

[1] For example, imagine one medical experiment using CR (using a coin flip to determine treatment for each subject, ignoring X) that happened to draw only healthy people in the treated group and only sick people for controls. Experimentalists know that MP experiments (match pairs on observed pretreatment variables like health and flip a coin within each pair for treatment) can massively increase efficiency, power, and robustness (King et al. 2007; Imai, King, and Nall 2008).

https://doi.org/10.1093/pan/mpn010 Published online by Cambridge University Press

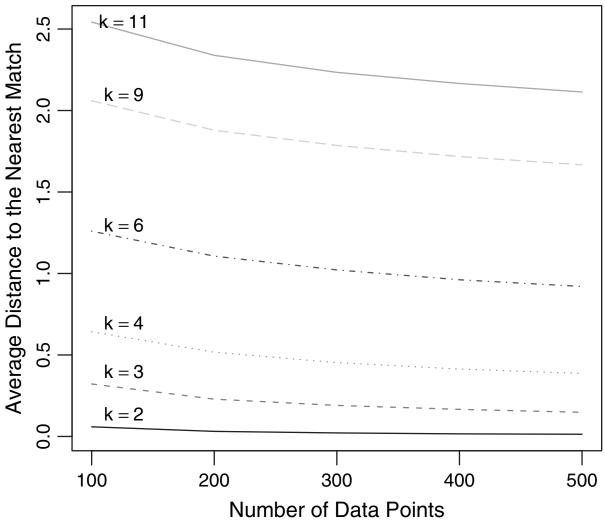

Fig. 1 Average distance to nearest match by n and k.

fact, the convex hull identifies the inefficiencies in CR, as it confirms that no counterfactuals are extrapolations when using an MP design, since inferences are all on the empirical support. Randomness is neither a panacea nor a proper standard of comparison when used inefficiently.

This difference in designs is also apparent when SM describe their "*n*/*k* problem" which confuses the nature of highly sparse high-dimensional space filled randomly in SM by CR rather than controlled exactly by MP. To appreciate the sparsity, take SM's case of $k = 10$. To properly fill 10 dimensional space with data, so counterfactuals randomly drawn from it would be nearby, n would need to be enormous. If we divide the possible values of each variable into just 10 groups and define "closeness" as being in the same group, one would need $n = 10^{10}$ to fill the space, so each point is "close" to at least one other. This is the curse of dimensionality. To claim that 10 dimensional space of continuous variables can be well packed with 1000 observations massively mischaracterizes the nature of high-dimensional space. The convex hull appropriately captures the available data regardless of how sparse and how much smaller the empirical support is relative to the population support: It keeps empirical tests empirical.

SM err when they use the propensity score as the gold standard for evaluating balance in their simulations since the goal of the propensity score check is population, not sample, balance. To establish sample balance, one must check common empirical (not population) support in the covariates directly (Ho et al. 2007). However, the only empirical check SM do is a classic example of the balance test fallacy (Imai, King, and Stuart 2008) and so is invalid. For example, with only 20 observations randomly spread over a huge 20 dimensional space, their check would still certify the data as free from model dependence, which is obviously incorrect. Appropriate checks would indicate that their simulated data are far from exactly balanced empirically.[2]

[2] Details appear in our replication file (King and Zeng, 2008). Moreover, even for checking population imbalance, the propensity score is not guaranteed to help if it is incorrectly specified, which is unfortunately the case in SM since in their simulation the true propensity score is exactly 0.5 whereas they use a misspecified logit model to estimate it.

SM's Fig. 2 argues that, because more data fall outside the convex hull when n shrinks and k grows, that something is amiss. However, they neglected to compute what should be the case. When this is done, their result confirms the value of our approach: We rerun SM's simulation and directly measure the nature of the data by computing the average distance of a data point to the nearest match, as a function of k and n. As the results in Fig. 1 show, the average distance between a data point in the treatment or control group and the closest point in the other group also grows when n shrinks (which can be seen by each line dropping) and k grows (which can be seen by lines with larger values of k appearing higher on the graph). This indicates that the convex hull diagnostic is appropriately empirical, as it should be: when data are farther away from the counterfactual, the diagnostic ought to indicate that.

SM's Fig. 3 is a spectacular demonstration of the advantage of the convex hull over population support: as the observed data region changes, the convex hull responds to reflect the empirical evidence, but the population support remains the entire shaded rectangle, always ignoring the data. SM's argument that higher correlations between a treatment variable and control variables causes a smaller interpolation region is as it should be, for the same reason as with linear regression, where high collinearity between the control variables and your causal variable leads to more difficult estimation and interpretation problems: holding one constant and letting the other change is a more difficult counterfactual in the presence of higher correlations. Similarly, by definition, we can make inferences more difficult by including additional covariates even if unrelated to existing variables in X and requiring counterfactuals to hold them constant, as SM do in their real data example; but that is of course the point: counterfactuals more counter to more facts are harder to infer. (And of course, the correct specification of X must be settled prior to the application of our diagnostics.)

In these simulations, sparsity is cleanly determined by n and k, whereas in real data other factors intervene. An application with any sized data can have 100% of inferences inside the convex hull and in another of the same size 0%; one outcome is not more likely than the other since data can take any structure and researchers can choose any counterfactuals they wish. Thus, SM's claim that "the probability that a data point is inside the hull of a set of points is always extremely small" (12) is false.[3] Similarly, SM's claim about the Gower distance that "there is a mechanical 'upper bound' for the percent of data that can be nearby the factuals relative to the counterfactuals" is false for the same reason: the data can be anywhere and scholars' counterfactual choices can be whatever they wish. No matter "the skewness of the treatment," the Gower distance and convex hull can take any value, depending on the data and inferential targets.

Finally, Section 3.1.3 of SM shows balancing on covariates unrelated to treatment, using the convex hull, is inefficient but not biased. This is unsurprising as it is a form of overfitting (Greene, 2008, 136). If, however, X is related to treatment, removing extreme counterfactuals can improve efficiency (Ho et al. 2007, 214). Of course, if one can rule out all models but the true one, as SM does by assumption, then model dependence diagnostics are

[3] SM's use of Elekes (1986) to support this claim is also incorrect: Elekes' is only relevant to evaluating all points that could in principle be drawn from a hypothetical population, not to SM's n countable counterfactuals constructed from switching the treatment variable value in their n observed data points. In fact, Elekes' proof does not even apply to SM's population since his proof requires bounded variables. That is why Eleke's probability bound applied to SM's data can be so wrong: for example, for $n = 500$ and $k = 2$, his bound on the [0,1] probability is 125!

https://doi.org/10.1093/pan/mpn010 Published online by Cambridge University Press

irrelevant. When the model is unknown, the convex hull diagnostic unambiguously indicates its worth, as our Fig. 1 and SM's Fig. 2 demonstrate.

We hope readers agree with us that this exchange has been productive in clarifying the nature of high-dimensional space, and the differences between empirical and population support, for assessing model dependence. In summary, for those in the unusual position of being sure about the population from which the data were drawn, the functional form of the model that generates the dependent variable and the stability of the model when used to infer the counterfactual of interest, SM's population-based definitions of extrapolation can sometimes be reasonable. If not, as is almost always the case in practice, then the venerable empirical definition of extrapolation used in the literature, and our resulting measures, are far more likely to give an accurate assessment of the degree of model dependence for any given counterfactual. From a Bayesian perspective, our model dependence checks give the objective data, which in different circumstances may or may not overwhelm one's prior beliefs.

After observing that the Mississippi River had been shortening 1.3 miles/year recently, Twain (1883, 176) extrapolated: "Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million year ago next November, the Lower Mississippi River was upward of 1,300,000 miles long and stuck out over the Gulf of Mexico like a fishing rod. And by the same token any person can see that 742 years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo and New Orleans will have joined their streets together and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact." As Twain obviously understood, the problem with this inference was his assumed population model, the absurdity of which can only be seen by recognizing how far his counterfactual had veered outside the convex hull of the observed data.

References

Box, George E. P., William G. Hunger, and J. Stuart Hunter. 1978. *Statistics for experimenters*. New York: Wiley-Interscience.

Elekes, G. 1986. A geometric inequality and the complexity of computing volume. *Discrete & Computational Geometry* 1:289–92.

Greene, William H. 2008. *Econometric analysis*. 6th ed. New York: Prentice Hall.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2001. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236. http://gking.harvard.edu/files/abs/matchp-abs.shtml.

Imai, Kosuke, Gary King, and Clayton Nall. 2008. "The essential role of pair matching in cluster-randomized experiments, with application to the Mexican Universal Health Insurance Evaluation." Unpublished manuscript, submitted to Statistical Science. http://gking.harvard.edu/files/abs/cluster-abs.shtml.

Imai, Kosuke, Gary King, and Elizabeth Stuart. 2008. Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A* 171:481–502. http://gking.-harvard.edu/files/abs/matchse-abs.shtml.

Jevons, W. Stanley. 1874. *The principles of science: a treatise on logic and the scientific method*. New York: MacMillen and Co.

King, Gary, Emmanuela Gakidou, Nirmala Ravishankar, Ryan T. Moore, Jason Lakin, Manett Vargas, Martha María Téllez-Rojo, Juan Eugenio Hernández Ávila, Mauricio Hernández Ávila, and Héctor Hernández Llamas. 2007. A 'politically robust' experimental design for public policy evaluation, with application to the Mexican universal health insurance program. *Journal of Policy Analysis and Management* 26:479–506. http://gking.harvard.edu/files/abs/spd-abs.shtml.

https://doi.org/10.1093/pan/mpn010 Published online by Cambridge University Press

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14:131–59. http://gking.harvard.edu/files/abs/counterft-abs.shtml.

King, Gary, and Langche Zeng. 2007. When can history be our guide? The pitfalls of counterfactual inference. *International Studies Quarterly* 51:183–210. http://gking.harvard.edu/files/abs/counterf-abs.shtml.

King, Gary, and Langche Zeng. 2008. Replication data for: empirical vs. theoretical claims about extreme counterfactuals: a response. hdl:1902.1/11903, Murray Research Archive [Distributor].

Sambanis, Nicholas and Alexander Michaelides. 2008. A Comment on Diagnostic Tools for Counterfactual Inference. Political Analysis. Advance Access published February 12, 2008, doi: 10.1093/pan/mpm032.

Twain, Mark. 1883. *Life on the Mississippi*. London: Chatto and Windus.

https://doi.org/10.1093/pan/mpn010 Published online by Cambridge University Press