*"Truth" Is Stranger than Prediction, More Questionable than Causal Inference**

Gary King, *Department of Government, Harvard University*

Robert Luskin's article in this issue provides a useful service by appropriately qualifying several points I made in my 1986 *American Journal of Political Science* article. Whereas I focused on how to avoid common mistakes in quantitative political science, Luskin clarifies ways to extract some useful information from usually problematic statistics: correlation coefficients, standardized coefficients, and especially R^2. Since these three statistics are very closely related (and indeed deterministic functions of one another in some cases), I focus in this discussion primarily on R^2, the most widely used and abused. Luskin also widens the discussion to various kinds of specification tests, a general issue I also address. In fact, as Beck (1991) reports, a large number of formal specification tests are just functions of R^2, with differences among them primarily due to how much each statistic penalizes one for including extra parameters and fewer observations.[1]

Reasons for Concern about Model Specification

Quantitative political scientists often worry about model selection and specification, asking questions about parameter identification, autocorrelated or heteroscedastic disturbances, parameter constancy, variable choice, measurement error, endogeneity, functional forms, stochastic assumptions, and selection bias, among numerous others. These model specification questions are all important, but we may have forgotten why we pose them. Political scientists commonly give three reasons: (1) finding the "true" model, or the "full" explanation; (2) prediction; and (3) estimating specific causal effects. I argue here that (1) is used the most but useful the least; (2) is very useful but not usually in political science where forecasting is not often a central concern; and (3) correctly represents the goals of political scientists and should form the basis of most of our quantitative empirical work.

*My thanks go to Jim Alt and Neal Beck for many helpful discussions and the National Science Foundation for grant SES-89-09201.

[1] In King (1991) I responded to a related critique of my 1986 *AJPS* article by Lewis-Beck and Skalaban (1991). The arguments here complement, but usually do not supplement, those by focusing on more general methodological issues.

American Journal of Political Science, Vol. 35, No. 4, November 1991, Pp. 1047–53
© 1991 by the University of Texas Press, P.O. Box 7819, Austin, TX 78713

Finding the "True" Model

Although Luskin speaks of all three reasons at times, he focuses principally on the first one. For example, at one point he writes, "But now suppose, more realistically, that the model may be incorrect. Perhaps a major influence has been omitted." He qualifies this at another point by seeking a model relatively "more true": "We shall never discover the 'true' regression function, known only to God, but the smaller the R^2, the likelier it is, other things being equal, that we could come up with a truer one." Finally, Luskin adds, "A further qualification is that the truest model may not be the best," an argument that is unclear at best (after all, if we introduce the criterion of "truth," how can one justify a less true model?).

The search for the true model is the most frequently cited justification for model specification tests in the political science literature, but I believe it to be largely vacuous. First, although I fully accept the premise that knowable truths exist in the *world,* the idea of a "true *model*" makes no sense. My point is not the usual cynical caveat that we can never attain perfection. Instead, a model is necessarily (and preferably) an abstraction and thus a drastic simplification, one that if successful will enable one to study only the essential elements of reality. Models may be good or bad for some purpose or another, but labeling models as true or false is not fruitful. Can one distinguish between true and false models of an airplane? Presumably either all models are false or the only true (sufficiently realistic) model is the airplane itself (although even actual airplanes do differ from one another). In either case, the goal of finding a "true model" is neither worthy nor useful. The concept of a "true model" also dredges up misleading goals of "perfect predictions," something we know to be impossible in a probabilistic world (or even in a deterministic world where we know we shall never have knowledge and measures of every possible influence).

Second, even if the meaning of a "true model" were perfectly clear, this goal does not represent the pursuit of a substantive political science question and thus generates particularly bad research designs. In order to reduce uncertainty, one should always seek to *maximize leverage* over research questions—by increasing the number of observations and finding evidence for as many observable implications of a theory as possible, as well as limiting the number of inferences we must make with the same set of data. If we start with a dependent variable and try to search for all possible (or all "big" or all "important") explanatory variables, we shall continually lose leverage over the problem. For each additional causal inference we wish to explore (or explanatory variable we add to an equation), the precision with which we can know all the other causal inferences in our study diminishes. By pursuing this goal, our "success" at finding more causes will *automatically* produce failure in learning about any one. Whenever possible, one should design research so as to focus on a small number of particularly important research questions (or, as I shall argue below, causal inferences).

Prediction

Prediction is by far the most straightforward of the three justifications for asking model specification questions, since it provides an extremely clear goal and an unambiguous and uncontroversial standard for judging success. One simply compares one's forecasts to future realizations of the process being explained. Thus, in this case, R^2 and the whole battery of specification tests mentioned by Luskin are irrelevant. In fact, there are numerous circumstances in which large values of R^2 (and the other specification tests) produce especially bad out-of-sample predictions (see Beck 1991). Inasmuch as political scientists do not often find themselves forecasting the processes they model, and since the solution to model specification here is so much easier and more obvious, I put this aside for now.

Estimating Causal Effects

A final reason for asking model specification questions is to improve causal inference, a goal that I believe to be at the heart of most of what we are and should be doing in quantitative political science.[2] If we do not apply the proper specification tests, we can easily make incorrect inferences. For example, omitted variable bias can cause one incorrectly to attribute the explanatory power of an omitted variable to the key causal effect we are trying to estimate. Measurement error similarly can bias estimates of causal effects. Ignoring any of these or other model specification questions may have serious consequences for the estimation of causal effects. Statistics that help us to evaluate whether we have omitted variables, measurement error, selection bias, and the like, therefore are extremely important in pursuing this goal.

Once we drop the goal of finding the "true" model and focus on making causal inferences (or predictions), using R^2 to distinguish between "true" and "less true" models is pointless. Since I find that methodological disagreements frequently evaporate when real data or examples are discussed, and since neither this paper (until now) nor Luskin's includes an example, consider the following. Suppose one is interested in the causal effect of crude oil prices on the public's opinion as to whether there is an energy shortage. One can imagine gathering 50 opinion polls about this question over as many months and regressing opinion on prices (probably also controlling for lagged opinion in some form), a reasonable specification if one wishes to learn about this causal effect. Furthermore, one could calculate an R^2 value. Now suppose we also found a monthly measure of the amount of television news coverage of oil price rises. Should we include TV coverage in the original equation? This variable is probably a major influence on public opinion about the existence of an energy crisis, and it will be highly

[2] Estimating effect parameters from linear regression equations is one way of gauging these causal relationships, but numerous other statistical methods are also used.

correlated with oil prices. If we omit this variable, would we not bias our inference? If we include it, R^2 will certainly rise by a good deal. Indeed, most of the specification tests discussed by Luskin would unambiguously point to including a TV coverage variable in the equation. However, if we are interested in the causal effect of oil prices per se, it makes little sense to control for a variable such as TV coverage, since it is in part a *consequence* of our key causal variable. If we controlled for it, we would be underestimating the causal effect of oil prices on public opinion.[3] (Similarly, if we were interested in the effect of party identification on voting behavior, we would not control for the voter's intention five minutes before walking into the voting booth!) Thus, the best specification in this typical example includes just the oil price variable.[4]

Now suppose instead that we are primarily interested in the causal effect of TV coverage of oil prices on public opinion about an energy shortage. In this case, we clearly *should* control for oil prices. Oil prices are both prior to and correlated with TV coverage; controlling for oil prices will largely eliminate a plausible confounding influence. As noted above, R^2 in this modified equation will be a good deal higher than in the original.

Thus, the "right model" depends entirely on the use to which it is put—the precise causal inference (or prediction) one wishes to make.[5] In this example, and in most others, a "true model" does not exist even in theory, but there is no "universally best" model either. The *usefulness* of a particular model specification depends entirely on what causal or forecasting goals one pursues. The second model fits the data better than the first, but the fit of the model to our data, as measured by R^2 or any other statistic, is largely irrelevant to the specific goal of our analysis.[6] Indeed, the first model has a substantially lower R^2, but for the purposes of that model it is considerably better. Thus, our theoretical reason for a model is our best guide to specification. Large R^2 values are not even empirically associated with better models in general.[7]

[3]Of course, we might wish to include it for a different purpose, such as I describe below or for separating this total effect into a direct and indirect effect of oil prices through television coverage. However, these are still different research purposes, for which we would generally want different models.

[4]Note that we arrived at this specification without any reference to the concept of parsimony.

[5]This is what Luskin is getting at when he speaks of a true model existing only "at a given causal distance."

[6]This is not to say that we should ignore residual plots and the like, since they can often suggest other hypotheses or different plausible confounding effects. One should always attempt to extract as much information from our data as possible. (This is indeed another reason why Luskin's article is valuable: if we read an article that presents only what one would consider the wrong statistics, it still makes sense to see if we can extract all available information from it.)

[7]To make the point in a somewhat different way, suppose we continue with the example about the influence of TV coverage of oil price rises on public opinion about an energy shortage. However, on this occasion suppose we construct an almost perfectly controlled, randomized, and very large-n experiment to estimate the causal effect. In this event, we can omit oil prices and indeed all other

Although he may not have been thinking about a "true model" and model specification statistics, Mark Twain said aptly, "Why *shouldn't* truth be stranger than fiction? Fiction, after all, has to make sense."

Uses for Relatively Unfit Fit Statistics

Finally, I turn to the most important contribution of Luskin's article. That is, when these statistics are used, how should we interpret them? I entirely agree with Luskin that one can glean relevant information from them, but we disagree somewhat about the interpretation and most useful form for this information. Anyone having looked at a lot of regression output will be led to the impression that R^2 does not seem crazy: it tends to be high when coefficients are large and standard errors small.

However, in my 1986 article, I claimed that the regression model contains no parameter for which R^2 is an estimate. The veracity of this argument is proved again by Luskin's equation (1), which is a fine statement of a regression model and which includes no parameter for R^2 to estimate. Although Luskin includes an interesting argument that proves that as the number of observations increase, R^2 converges in probability to a fixed point, which he labels ρ^2 (and defines as $1 - \sigma_u^2/\sigma_y^2$), he must assume that the "correct" explanatory variables are included in the model in order to show that R^2 is a reasonable estimate of ρ^2. That is, his probability limit calculations are conditional on prior knowledge of the correct specification. Of course, the "correct" specification depends entirely on the causal or predictive purpose of the model, as well as one's current priors about the specification and even preliminary experience with the data. Hence, the proof presupposes that one already knows the answer to his most important question. In other words, *R^2 measures no parameter in the regression model, and the parameter it does estimate is not at all helpful in choosing a particular model specification.*[8]

I show in King (1991) that R^2 and the standard error of the regression, $\hat{\sigma}$, contain precisely the same information; however, R^2 expresses it in a form much more likely to mislead. I demonstrate this in a different way here in the following

possible control variables (since they would be uncorrelated with the treatment variable, TV coverage). This research design will provide a very reliable estimate of the causal effect. Suppose, however, that R^2 is only 0.03. This R^2 value *might* indicate that numerous unmeasured variables affect public opinion as to the existence of an energy shortage, but that is entirely irrelevant to our purpose, which was to estimate a causal effect, not the unattainable (and unclear) goal of determining every possible cause of public opinion.

[8]We all agree that maximizing R^2 will not get us anywhere. However, suppose we could observe ρ^2. If we maximized it with respect to the specification, we would still not produce the right model, since in this formulation the correct specification is a condition, not something to be estimated. Trying to do so is essentially the same mistake as using inverse probability to estimate parameters when one should be using likelihood or a Bayesian posterior distribution (see King 1989).

three examples. First, suppose one is estimating the effect of unemployment on presidential approval, measured by the monthly Gallup opinion polls. The inherent sampling error of one of these polls is about 2%. If we obtain a value of $\hat{\sigma}$ smaller than this figure, we would *know* something is seriously wrong with the model specification: no statistical model can generate predictions of numbers more accurately than they can be measured. Although with $\hat{\sigma}$ we can sometimes see when we are overfitting a data set, doing the same with R^2 would be difficult.

More obvious, perhaps, is the relative frequency with which the two statistics are misused. Although statistical practice has been improving substantially in recent years, finding four or five consecutive issues of a major journal without some misuse of R^2—even a misuse that Luskin and I agree about—would be fairly difficult. However, finding even a single misuse of $\hat{\sigma}$, other than omission, is quite challenging (though not impossible!).

Finally, consider this thought experiment. Suppose one group of people were given a regression equation, R^2, and the variance of the dependent variable, whereas another group were given the same equation, the variance of the dependent variable, and the corresponding value of $\hat{\sigma}$. Which group would re-create the scatter plot more accurately without the benefit of seeing the original data? Each group would have essentially the same information. However, since the first group would need to recall how to compute $\hat{\sigma}$ from R^2 and the variance of the dependent variable, I think most would expect the second group to outperform the first. Thus, even as a measure of fit, something that is not terribly interesting, $\hat{\sigma}$ is more useful than R^2.[9]

The thrust of Luskin's article is well taken and very useful. In his words, "Weathervanes and anemometers in the cities do not diminish the value of a wet finger aloft in the wilderness." True though this may be, I would prefer a meteorologist in a well-equipped weather station.

[9]Luskin also warns about standardized coefficients but indicates that they contain useful information because they represent causal effects *and* the "variables capacity for change." I agree that we should always look at both pieces of information. However, why confound the two by combining them in one number? More important is that standardized coefficients do not properly measure the "variables capacity for change." For example, consider the variable age. For any one year, this variable changes by exactly one unit (year); however, if we judged by the standard deviation in a cross-section, we might incorrectly think that the age of a person could change by 30 points in a single year! The situation is even more extreme for race or gender, or even income. Standardized coefficients tend to be used primarily in cross-sectional data, but their use in time series data is not much better. For example, most economic variables change quite gradually over time, but many vary over large ranges when looking across many years. Inflation, for example, will increase or decrease by a point or two over a year, but the standard deviation over a multiyear data set would indicate that the variables short-term capacity for change is enormous—perhaps 10 points or more. Luskin is correct that standardized coefficients contain some useful information, but as with the other statistics, there are better forms.

REFERENCES

Beck, Nathaniel. 1991. "Model Selection: Are Time Series Techniques Useful in Cross-Sectional Problems?" Presented at the annual meeting of the Midwest Political Science Association, Chicago.

King, Gary. 1986. "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science." *American Journal of Political Science* 30:666–87.

———. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press.

———. 1991. "Stochastic Variation: A Comment on Lewis-Beck and Skalaban's 'The *R*-Square.'" *Political Analysis* 2:185–200.

Lewis-Beck, Michael S., and Andrew Skalaban. 1991. "The *R*-Squared: Some Straight Talk." *Political Analysis* 2:153–72.

Luskin, Robert. 1991. "Abusus Non Tollit Usum: Standardized Coefficients, Correlations, and R^2s." *American Journal of Political Science* 35:1030–44.