

Return of the Phantom Menace

*Omitted Variable Bias in Political Research**

KEVIN A. CLARKE

University of Rochester, USA

Scholars often assume that the danger posed by omitted variable bias can be ameliorated by the inclusion of large numbers of relevant control variables. However, there is nothing in the mathematics of regression analysis that supports this conclusion. This paper goes beyond textbook treatments of omitted variable bias and shows, both for OLS and for generalized linear models, that the inclusion of additional control variables may increase or decrease the bias, and we cannot know for sure which is the case in any particular situation. The last section of the paper shows how formal sensitivity analysis can be used to determine whether omitted variables are a problem. A substantive example demonstrates the method.

KEYWORDS: control variables; model specification; omitted variable bias; research design

Introduction

Scholars often include extra variables in their specifications to address the fear that omitted relevant variables will bias the results. The belief is that the inclusion of every additional relevant variable serves to reduce this potential threat. That is, a researcher cannot know all of the variables that appear in a certain data generating process, but if she knows and includes 15 of them, she is better off than if she knows and includes only 10 of them. A corollary belief is that the inclusion of such variables is considered to be, for the most part, benign. Gelpi and Feather

*An earlier version of this paper was delivered at the annual meeting of the Society for Political Methodology in Tallahassee, FL (2005) and the annual meeting of the American Political Science Association in Washington, DC (2005); I thank the participants for their comments. I also thank Chris Achen, Jake Bowers, Rob Franzese, John Jackson, and three anonymous reviewers for helpful comments. Fabiana Machado provided excellent research assistance. Support from the National Science Foundation (Clarke: Grant #SES-0213771) is gratefully acknowledged. Correspondence to Kevin A. Clarke, Department of Political Science, University of Rochester, Rochester, NY 14627-0146, USA. E-mail: kevin.clarke@rochester.edu

(2002: 783) for instance, argue that the inclusion of “additional control variables cannot artificially inflate the estimated impact of our variable of interest.”¹

As Clarke (2005) demonstrates, however, the mathematics of econometric analysis do not support these conclusions. The standard omitted variable result addresses the omission of a single variable or a single set of variables from an ordinary least squares regression model. The result does not include the situation in which a *subset* of the set of omitted variables is included in a specification as controls, nor is the result easily extended to more complicated models such as generalized linear models. The reasons are not mysterious: the effect of including such variables depends, even in a simple case, on a host of factors.

The argument in this paper is a generalization and extension of the results of Clarke (2005). Going beyond OLS to include generalized linear models, we show that the only thing that can be said for certain about control variables is that unless we find ourselves in the precise situation described by textbooks, we cannot know the effect of including an additional relevant variable on the bias of a coefficient of interest. The addition may increase or decrease the bias, and we cannot know for sure which is the case in any particular situation.

An appreciation of these results does not mean, however, that political scientists are helpless in the face of the unknown. It is possible to make reliable inferences without controlling for large numbers of possible confounders, and in the penultimate section the paper, we discuss one such technique, formal sensitivity analysis. We demonstrate in a substantive example that robust inferences are possible with a modest number of covariates. The technique is largely unknown to political scientists at the present, but given its growing application in the economics literature (see Altonji, Elder, and Taber, 2005), it is likely to see greater use in the future.

Background Results

A brief review of some familiar, and less familiar, aspects of the omitted variable result is useful for following the discussion.² Consider an unrestricted model, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, and a restricted model, $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where $\boldsymbol{\beta}_2 = \mathbf{0}$. If we assume that the unrestricted model is the data generating process (DGP), and thus the correct specification, the restricted model is therefore misspecified and becomes $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*$, where $\boldsymbol{\epsilon}^* = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

The expected value of $\hat{\boldsymbol{\beta}}_1$ from the misspecified equation, under the usual assumptions, is well known and given by Greene (2003) to be $E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2}\boldsymbol{\beta}_2$, where $\mathbf{P}_{1.2} = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2$ is the matrix of regression coefficients from the auxiliary regressions of the excluded variables, \mathbf{X}_2 , on the included variables, \mathbf{X}_1 . Unless the omitted variables are orthogonal to the included variables, and thus $\mathbf{X}'_1\mathbf{X}_2$ is a zero matrix, or $\boldsymbol{\beta}_2$ is a zero vector, our estimate of $\hat{\boldsymbol{\beta}}_1$ is biased.³ The effect of omitting \mathbf{X}_2 depends upon the magnitude of the excluded coefficients, $\boldsymbol{\beta}_2$, the magnitudes of the coefficients from the auxiliary regressions,

¹Emphasis in the original.

²These results are developed at a more leisurely pace in any number of standard econometrics texts.

$\mathbf{P}_{1,2}$, the correlations between the included variables, \mathbf{X}_1 , and the variances of the independent variables (Hanushek and Jackson, 1977).

After demonstrating the above point, the standard textbook treatments of omitted variable bias (see for example, Gujarati, 2003; Johnston and DiNardo, 1997; Kmenta, 1986) turn to discussing the inclusion of irrelevant variables, those which have no effect on the dependent variable. The result is that $\hat{\beta}_1$ is unbiased (OLS correctly estimates β_2 as $\mathbf{0}$), but no longer minimum variance. While never explicitly endorsing the inclusion of irrelevant variables, most standard treatments note the “asymmetry” (Gujarati, 2003: 514) in these two results and imply that including an irrelevant variable has fewer deleterious effects than leaving out a relevant variable. Johnston & Dinardo (1997: 110), for example, write that the consequences of including irrelevant variables are “generally less serious than those pertaining to the exclusion of relevant variables.”

Finally, we note an additional result found in more extensive treatments of the subject. The OLS estimates of the coefficients of the restricted model are $\hat{\beta}_1 = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y}$, with variance $\sigma^2(\mathbf{X}'_1\mathbf{X}_1)^{-1}$. Making use of the Frisch-Waugh-Lovell theorem (see Davidson and MacKinnon, 1993: 21), the OLS estimates of the coefficients of the unrestricted model are $\hat{\beta}_1 = (\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{M}_2\mathbf{y}$ with variance $\sigma^2(\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1)^{-1}$, where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2$.

The difference in the covariance matrices of the unrestricted and restricted models, $\sigma^2(\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1)^{-1} - \sigma^2(\mathbf{X}'_1\mathbf{X}_1)^{-1}$, is shown by Greene (2003: 150) and Davidson and MacKinnon (1993: 95) to be positive semidefinite. The restricted estimator, therefore, is more efficient than the unrestricted estimator. That is, the variance of the restricted estimator is never larger than the variance of the unrestricted estimator. Adding a control variable can never decrease the variance of the coefficient of interest; the variance can only increase or remain the same.

The fact that the restricted estimator is both biased and more efficient than the unrestricted estimator when the unrestricted estimator is the DGP means that choosing between these estimators requires consideration of the mean square error criterion. The difference in the mean square error matrices for the unrestricted and restricted estimators is given by Judge et al. (1985) to be

$$\text{MSE}_U - \text{MSR}_R = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2\{\sigma^2[\mathbf{X}'_2\mathbf{X}_2 - \mathbf{X}'_2\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2]^{-1} - \beta_2\beta'_2\}\mathbf{X}'_2\mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1},$$

³Another condition under which $\mathbf{P}_{1,2}\beta_2$ is the null vector, one rarely mentioned in textbooks, is for each row to sum to zero. For example, if \mathbf{X}_1 and \mathbf{X}_2 are $n \times 2$, then $\mathbf{P}_{1,2}$ is a 2×2 matrix of auxiliary regression coefficients and β_2 is a 2×1 vector of the effects of the omitted variables. $\mathbf{P}_{1,2}\beta_2$ is then a vector of two sums,

$$\mathbf{P}_{1,2}\beta_2 = \begin{bmatrix} b_{13}\beta_3 + b_{14}\beta_4 \\ b_{23}\beta_3 + b_{24}\beta_4 \end{bmatrix}$$

where b_{ij} are the coefficients from the auxiliary regressions of the omitted variables on the included variables, and β_j are the effects of the omitted variables. If $b_{ij}\beta_j = -b_{i,j+1}\beta_{j+1}$, then $\mathbf{P}_{1,2}\beta_2 = \mathbf{0}$.

which is positive semidefinite if $\Sigma_{\hat{\beta}_2} - \beta_2\beta_2'$ is positive semidefinite, where $\Sigma_{\hat{\beta}_2}$ is the covariance matrix for $\hat{\beta}_2$.

The point of the above equation is that we might prefer the restricted model *even when the unrestricted model is the DGP*. As Davidson and MacKinnon (1993: 96) put it, “[t]hus it may be desirable to use the restricted estimator. . . when the restrictions are false, provided they are not too false.” Including a previously omitted variable is therefore not an unambiguously positive choice even when the conditions of the standard omitted variable bias discussion are met.⁴ Thus, whether we are concerned about bias, efficiency, or mean squared error, the inclusion of even relevant control variables may make the situation worse. Adding a relevant control variable to a regression does not necessarily “minimize” or “reduce” the threat from omitted variable bias.

Most of the lesson described above does not apply in practice. Political scientists are faced with including additional relevant variables out of a larger set of relevant omitted variables that we either do not know or cannot measure. What we need to understand is the effect of including some, but not all, of these relevant omitted variables.

Why We Use Control Variables

As argued by Clarke (2005: 343), the use of control variables comes directly from the omitted variable bias result. The reasoning is that we decrease the aggregate bias on the coefficient of interest for every additional relevant control variable that we include. The inefficiency part of the equation is rarely mentioned, as control variables often do have real effects.⁵ Included on the basis of previous empirical work, control variables do not engender efficiency concerns and are thus supposed to affect only the issue of bias.⁶

This reasoning is found throughout the discipline. The leading research design text in political science advises quantitative researchers to “systematically look for omitted control variables” and notes that if “relevant variables are omitted, our ability to estimate causal inferences correctly is limited” (King, Keohane, and Verba, 1994: 173, 175). Ansolabehere, Gerber, and Snyder (2002: 770), in their study of court-ordered redistricting and public expenditures, include variables such as poverty, unemployment rates, median income, percentages of the population that are school-aged, black, and elderly, and population change “[t]o minimize the

⁴The estimated σ^2 when the restrictions are false is biased upward. While this bias affects neither our efficiency nor our mean-square-error calculations, it does affect the estimated standard errors. Simulations show, however, that the magnitude of the bias on $\hat{\sigma}^2$ is rarely large enough to affect our findings.

⁵We define a control variable as a potential confounder that has been measured and used for covariance adjustment through regression or matching. Thus, a control variable is not of direct theoretical interest and generally serves to address charges of omitted variable bias.

⁶There also exists a separate logic of “robustness,” and some might argue that the use of control variables follows from it. Such a logic, however, would include the incremental addition of variables, multiple operational definitions for important variables, alternative functional forms, and alternative error structures. Rarely is this logic seen at work in political science.

danger of omitted variables bias....” Similarly, Bailey, Kamoie, and Maltzman (2005: 78) in studying the role of the solicitor general in Supreme Court decision making write that “[t]o minimize the possibility that omitted variable bias affects our results, we control for a number of factors that may affect court voting.” In the same vein, Hegre et al. (2001: 37), in a study of democracy and civil war, identify a number of control variables “whose omission might bias the results for the regime change variable.” Finally, Rudolph and Evans (2005: 64), looking at the relationship between public trust and government spending, control for a number of individual-level factors “[t]o reduce the risk of omitted variable bias,” and Krause (2003: 184), while studying uncertainty and budgeting, argues that leaving some variables out in an alternative specification “generated an omitted variable bias problem.”

What is obvious in the above quotes, and in hundreds of others throughout political science, is that control variables are added to a specification to “minimize” or “reduce” the likelihood of omitted variable bias. At the same time, there is little doubt that responding to or warding off reviewer attacks contributes to the growth of specifications significantly. King et al. (2001: 51) note, “[a]s critics we use omitted variables as the first line of attack, and as authors we know that controlling for more variables helps protect us from potential criticism; from this perspective, the more variables in X_2 the better.”

Mathematically, we can express the argument for the inclusion of control variables by considering a data generating process in scalar notation,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

and two misspecified models,

$$\text{Model 1: } Y_i = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1},$$

$$\text{Model 2: } Y_i = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2}.$$

The claim, based on the reasoning above, is that the bias on $\hat{\beta}_{11}$, the estimated coefficient on X_1 in Model 1, is greater than the bias on $\hat{\beta}_{12}$, the estimated coefficient on X_1 in Model 2. Letting the bias on $\hat{\beta}_{11}$, $E[\hat{\beta}_{11}] - \beta_1$, be denoted as $b(\hat{\beta}_{11}, \beta_1)$, and the bias on $\hat{\beta}_{12}$, $E[\hat{\beta}_{12}] - \beta_1$ be denoted as $b(\hat{\beta}_{12}, \beta_1)$, the mathematical argument is that

$$|b(\hat{\beta}_{11}, \beta_1)| \geq |b(\hat{\beta}_{12}, \beta_1)|.$$

This conclusion cannot be supported mathematically. The inclusion of additional relevant variables can increase or decrease the bias on the X_1 coefficient, and short of knowing all omitted relevant variables, the researcher cannot know which is the case.

A Monte Carlo Demonstration

In this section, we go beyond known results to demonstrate what can occur when a subset of the set of omitted variables is included in a regression.⁷ As an example, consider the situation where the true data generating process is

⁷To our knowledge, similar demonstrations do not exist in the econometrics literature.

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

The two misspecified models can both be written as $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*$ where \mathbf{X}_1 comprises X_1 in Model 1 and X_1 and X_2 in Model 2.

For Model 1, let the auxiliary regression of the excluded variables on the included variable be $X_2 = \gamma_0 + \gamma_1 X_1 + \epsilon_2$ and $X_3 = \delta_0 + \delta_1 X_1 + \epsilon_3$. For Model 2, let this auxiliary regression be $X_3 = \delta_0^* + \delta_1^* X_1 + \delta_2^* X_2 + \epsilon_3$.⁸

In Model 1, therefore, the bias on $\boldsymbol{\beta}_1$ is

$$E[\hat{\boldsymbol{\beta}}_1] - \boldsymbol{\beta}_1 = \mathbf{P}_{1,2}\boldsymbol{\beta}_2 = \begin{bmatrix} \gamma_0 & \delta_0 \\ \gamma_1 & \delta_1 \end{bmatrix} \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \gamma_0\beta_2 + \delta_0\beta_3 \\ \gamma_1\beta_2 + \delta_1\beta_3 \end{bmatrix}.$$

In Model 2,

$$E[\hat{\boldsymbol{\beta}}_1] - \boldsymbol{\beta}_1 = \mathbf{P}_{1,2}\boldsymbol{\beta}_2 = \begin{bmatrix} \delta_0^* \\ \delta_1^* \\ \delta_2^* \end{bmatrix} \begin{bmatrix} \beta_3 \end{bmatrix} = \begin{bmatrix} \delta_0^*\beta_3 \\ \delta_1^*\beta_3 \\ \delta_2^*\beta_3 \end{bmatrix}.$$

In order to assess the logic of control variables, we want to know under what conditions the bias on our estimate of β_1 from Model 1 is greater than or equal to the bias on our estimate of β_1 from Model 2,

$$\begin{aligned} |b(\hat{\beta}_{11}, \beta_1)| &\geq |b(\hat{\beta}_{12}, \beta_1)| \\ |\gamma_1\beta_2 + \delta_1\beta_3| &\geq |\delta_1^*\beta_3|. \end{aligned}$$

We can investigate with relative ease the precise effects of varying these components through simulation. With the exception of γ_1 , which we set at 2, all other values will be allowed to vary between -5.0 and 5.0 in order to explore scenarios where the omitted variables have both positive and negative effects on the dependent variable and where the omitted variables have both positive and negative effects on the included variables.⁹

The Monte Carlo Results

The results are in Figure 1, which contains graphs of the difference in the absolute values of the two biases for various combinations of β_2 and β_3 ,

$$|b(\hat{\beta}_{11}, \beta_1)| - |b(\hat{\beta}_{12}, \beta_1)|.$$

Negative values, therefore, indicate that the inclusion of the additional relevant variable, X_2 , increases the bias on the estimated coefficient of X_1 compared to the case where both X_2 and X_3 are omitted. Positive values indicate that the

⁸ $\delta_1 \neq \delta_1^*$ unless X_1 and X_2 are orthogonal.

⁹ γ_1, δ_1 , and δ_1^* can all be written in terms of correlations and variances (see Hanushek and Jackson, 1977), and the choices we have made imply certain restrictions. For example, r_{12} , the correlation between X_1 and X_2 , must be positive because γ_1 is set to 2. By the same token, r_{13} is positive when δ_1 is positive and negative otherwise. It is possible to show that by paying attention to these restrictions and setting the variance of X_1 to 1, all the points on the graphs are possible.

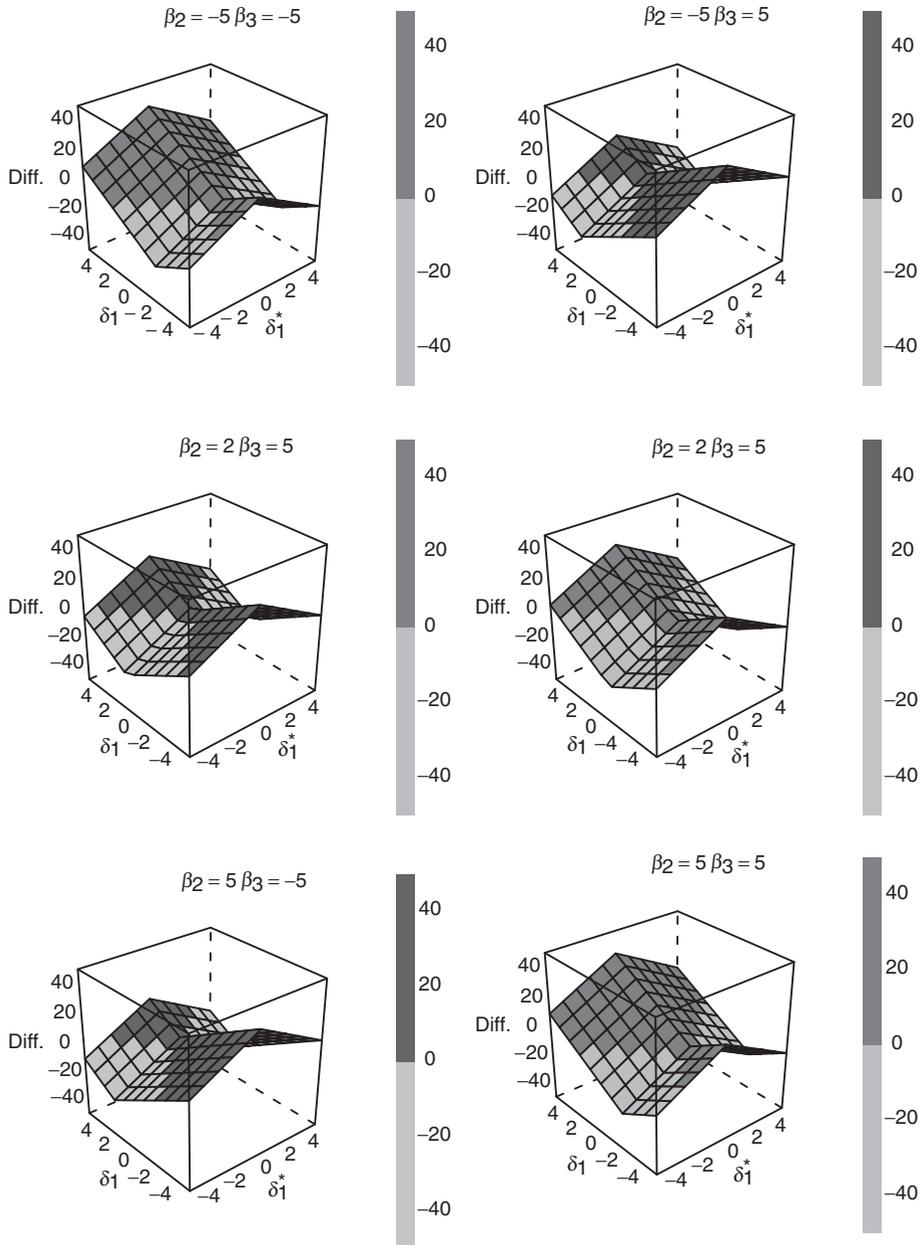


Figure 1. Difference in the Absolute Values of the Biases for Various Values of β_2 and β_3

inclusion of the additional relevant variable, X_2 , decreases the bias on the estimated coefficient of X_1 compared to the case where both X_2 and X_3 are omitted.

Figure 1 shows that including X_2 in the regression is just as likely to increase the bias on $\hat{\beta}_1$ as it is to decrease it. The lighter shaded areas on Figure 1 indicate

values of δ_1 and δ_1^* for which the absolute value of the bias on $\hat{\beta}_{12}$ is greater than the bias on $\hat{\beta}_{11}$. How this occurs is not mysterious. Consider the upper left-hand panel of Figure 1, where $\beta_2 = \beta_3 = -5$. Let $\delta_1 = -4$ and $\delta_1^* = -5$. Then,

$$|\gamma_1\beta_2 + \delta_1\beta_3| - |\delta_1^*\beta_3| = |2(-5) + -4(-5)| - |-5(-5)| = -15.$$

The only condition under which a researcher can be assured that the inclusion of the addition control variable will not make matters worse is when $\delta_1^* = 0$.¹⁰ This condition occurs when X_2 is added to the auxiliary regression of X_3 on X_1 and reduces the effect of X_1 on X_3 , δ_1^* , to zero. Otherwise, for every combination of δ_1 and δ_1^* , there are values of β_2 and β_3 that will make the addition of X_2 increase the bias on the coefficient of interest.

The *Monte Carlo* experiment demonstrates that unless a researcher knows the remaining omitted variable and, furthermore, knows the relationship of that variable with the newly included variable, she cannot know the effect that the newly included variable will have on the bias of a coefficient of interest. The newly included variable *may* decrease or increase the bias. We need to know, even if we cannot measure all of it, the complete and true specification in order to know which is the case.

Beyond Ordinary Least Squares

While these results are instructive, OLS regression is no longer the workhorse of quantitative political science. Generalized linear models (see McCullagh and Nelder, 1989) have long since replaced OLS in this capacity. In this section, we extend our simple demonstration to the case of generalized linear models.¹¹

As in the previous section, we need to review a few known results first. That the standard OLS omitted variable result does not directly extend to generalized linear models is well known among econometricians, although less so among political scientists. The basic result is that an omitted variable can bias a coefficient of interest even if the omitted variable is uncorrelated with the included variables (Gail, Wieand, and Piantadosi, 1984).¹² The conditions for bias caused by an independent omitted variable concern the shape of the link function, g . Only if the link function is linear or the log link, $g(u) = (1/b)\log(a + bu)$, do independent omitted variables not cause bias (Neuhaus, 1993: 810). Therefore, link functions such as the logistic, probit, and the complementary log-log do not prevent bias. In fact, any link function based on the normal, exponential, gamma (with shape parameter greater than one), or Weibull distributions (with shape parameter greater than one) is subject to this problem.

Cramer (2003) demonstrates this point in the logit case with a method that will prove useful. Consider a latent variable model,

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad \text{where } Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

¹⁰Note the “ridge line” in every graph where $\delta_1^* = 0$

¹¹As before, this demonstration does not exist in the literature.

¹²Wooldridge (2002: 470) refers to this situation as the “neglected heterogeneity” problem.

Suppose that we run a restricted model where β_2 is set to zero, $Y_i^* = \beta_0 + \beta_1 X_{i1} + \epsilon_i^*$, and the omitted variable is a function of the included variable and an error term that is independent of ϵ_i , $X_{i2} = \gamma_0 + \gamma_1 X_{i1} + v_i$. By substituting X_2 into equation (1) and collecting terms, Cramer gets an expression for the main effect of omitting X_{i2} ,

$$Y_i^* = (\beta_0 + \beta_2 \gamma_0) + (\beta_1 + \beta_2 \gamma_1) X_{i1} + (\epsilon_i + \beta_2 v_i).$$

The normalized coefficient on X_1 is therefore

$$\hat{\beta}_1 = \frac{c}{\sqrt{\sigma^2 + \beta_2^2 \sigma_v^2}} (\beta_1 + \gamma_1 \beta_2), \tag{2}$$

where $c = \pi/\sqrt{3} \approx 1.814$ in the logit case. Note that even if X_1 and X_2 are independent, $\gamma_1 = 0$, the denominator of equation (2) does not reduce to σ as it should. Rather, it reduces to $\sqrt{\sigma^2 + \beta_2^2 \text{var } X_2}$ (Cramer, 2003: 81). Thus, the change in the normalized coefficient depends upon the value of β_2 and the variance of X_2 .

The fact that an omitted variable can affect a coefficient of interest in a generalized linear model even if the omitted variable is uncorrelated with included regressors seems to make an even stronger argument for including all possible relevant variables. Once again, this intuition is mistaken, and we can make use of Cramer’s method to demonstrate the point.

Consider the following latent data generating process

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad \text{where } Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Also consider two misspecified models where X_2 and X_3 are left out of Model 1, and only X_3 is left out of Model 2,

$$\text{Model 1: } Y_i^* = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1},$$

$$\text{Model 2: } Y_i^* = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2}.$$

The normalized coefficient on X_1 (see the Appendix for a derivation) is

$$\hat{\beta}_{11} = \frac{c[\beta_1 - \beta_1 \gamma_2 \delta_2 + \beta_2 (\gamma_1 + \gamma_2 \delta_1) + \beta_3 (\delta_1 + \delta_2 \gamma_1)]}{\sqrt{\sigma^2 + \sigma^2 \gamma_2^2 \delta_2^2 + \sigma_\eta^2 (\beta_2^2 \gamma_2^2 + \beta_3^2) + \sigma_v^2 (\beta_2^2 + \beta_3^2 \delta_2^2)}}. \tag{3}$$

For Model 2, the normalized coefficient is

$$\hat{\beta}_{12} = \frac{c}{\sqrt{\sigma^2 + \beta_3^2 \sigma_\eta^2}} (\beta_1 + \beta_3 \delta_1). \tag{4}$$

As in our earlier demonstration, we want to know under what conditions the bias on $\hat{\beta}_{11}$ is greater than or equal to the bias on $\hat{\beta}_{12}$,

$$|b(\hat{\beta}_{11}, \beta_1)| \geq |b(\hat{\beta}_{12}, \beta_1)|.$$

We again use simulation to investigate the bias inequality due to the large number of moving parts. Two of the major differences between equations (3) and (4) are the presence of β_2 and γ_1 in equation (3). We therefore allow β_2 and γ_1 to vary between -5.0 and 5.0 in order to investigate scenarios where their effects are both positive and negative. By presenting a panel of graphs, we also let β_1 and β_3 vary between -5.0 and 5.0 . To isolate the effects of these factors, we set all variances to 1 and all other coefficients to 2.

The Monte Carlo Results, Part 2

The effects of β_1 , β_2 , β_3 , and γ_1 are shown in Figure 2, which contains graphs of the difference in the absolute values of the two biases,

$$|b(\hat{\beta}_{11}, \beta_1)| - |b(\hat{\beta}_{12}, \beta_1)|.$$

As before, negative values indicate that the inclusion of the additional relevant variable, X_2 , increases the bias on the estimated coefficient of X_1 compared to the case where both X_2 and X_3 are omitted. Positive values indicate that the inclusion of the additional relevant variable, X_3 , decreases the bias on the estimated coefficient of X_1 compared to the case where both X_2 and X_3 are omitted.

Figure 2 shows that there are many conditions under which the inclusion of X_2 is just as likely to increase the bias on $\hat{\beta}_1$ as it is to decrease it. Each panel of the figure shows the full range of β_2 and γ_1 . The left-hand column is for low values of β_3 (1), while the right-hand column is for high values of β_3 (5). Each successive row, moving from the top to the bottom, shows more negative values of β_1 ($-1, -3, -5$).¹³ As Figure 2 shows, the inclusion of X_2 in the logit specification makes the bias on $\hat{\beta}_1$ worse across the entire range of β_2 and γ_1 , although it is particularly the case when both β_2 and γ_1 are negative.¹⁴

Figure 2 also shows that the inclusion of X_2 in the logit specification can make the bias on $\hat{\beta}_1$ worse for almost all combinations of β_1 and β_3 . This is true when both β_1 and β_3 are small ($-1, 1$), when β_1 is small and β_3 is large ($-1, 3$), and when β_1 and β_3 are large ($-5, 5$). The only combination where the inclusion of X_2 never makes the bias worse is when β_1 is large and β_3 is small ($-5, 1$).¹⁵

As in the linear case, the bottom line here is that unless a researcher knows the remaining omitted variable and the relationship of that variable with the newly included variable, she cannot know the effect that the newly included variable

¹³Other sets of panels are available from the author. If β_1 is positive and β_3 is negative, the surfaces in Figure 2 simply rotate to the right a quarter turn. If both β_1 and β_3 are positive, the surface folds along the left-right axis with the lighter-shaded areas in the center of the fold.

¹⁴When β_1 is positive and β_3 is negative, the inclusion of X_2 makes the bias on $\hat{\beta}_1$ worse particularly when β_2 is positive and γ_1 is negative. When β_1 and β_3 are both positive, the inclusion of X_2 makes the bias on $\hat{\beta}_1$ worse particularly when β_2 is negative and γ_1 is positive, or when β_2 is positive and γ_1 is negative.

¹⁵This pattern also holds for the cases where β_1 is positive and β_3 is negative and when both β_1 and β_3 are positive.

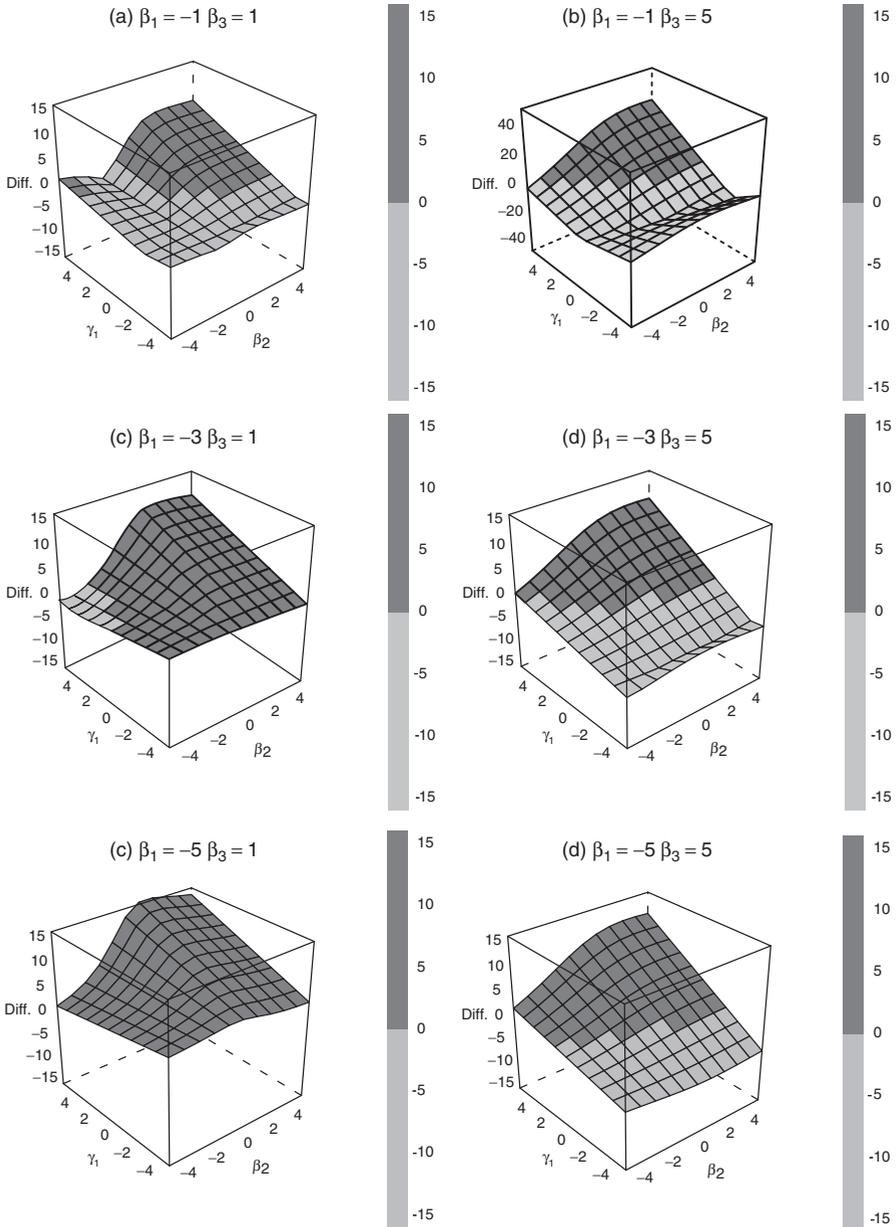


Figure 2. The Effects of β_1 , β_2 , β_3 , and γ_1 on the Difference in the Absolute Values of the Two Biases

will have on the bias of the coefficient of interest. The newly included variable may decrease the bias, or it may not. We simply cannot know the effect on the bias of including an additional control variable unless we also know the complete and true specification.

Engaging the Phantom Menace

Once the connection between omitted variable bias and control variables is gone, the main justification for using control variables is gone, and therein lies the importance of these results. Including more variables in a regression, even relevant ones, does not necessarily make the regression results more accurate.

Others have also argued against control variables, but on very different grounds. These include theory (Achen, 1992), the difficulty of data analysis with large numbers of variables (Achen, 2002), the threat of measurement error (Griliches, 1997), and the threat from unrecognized nonlinearity (Welch, 1975; Maddala, 1997; Achen, 2005).

Of course, variable selection is an enormously complex problem with a long history. A criminally short review of the literature would include estimated risk criteria such as Mallows C_p (Mallows, 1973), information theoretic model selection criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), Stein-like shrinkage estimators (Stein, 1955), various nonnested tests (Cox, 1961; Vuong, 1989; Clarke, 2001; Clarke, 2003; Clarke, 2007), and econometric methodologies such as the London School of Economics approach associated with David Hendry and colleagues (Hendry and Richard, 1982) and Christopher Sims's vector autoregression approach (Sims, 1980). Notably, none of these approaches is based on the assumption that larger specifications are desirable, and most are designed explicitly to guard against such specifications. Nowhere in the literature on variable selection does bigger equal better, and some statisticians go so far as to quite forthrightly argue that regression equations based on a few variables are simply "more accurate" than regression equations based on many variables (Breiman, 1992: 738).

Appreciation of the results in this paper should not give rise to either despair or nihilism on the part of quantitative political scientists. The use of large numbers of control variables is a flawed strategy for making reliable inferences, but other more successful strategies are available. While no technique produces reliable inferences in a deterministic fashion, the use of these other strategies serves to produce more compelling evidence than could be had without their use. Two popular strategies are natural experiments (for theory and an application, see Meyer, 1995; Brady and McNulty, 2004) and restricting the spatial or temporal domain of a study (see Hanushek and Jackson, 1977; Achen, 2002). A third strategy, one that is mostly unknown in political science, is formal sensitivity analysis.¹⁶

The goal of formal sensitivity analysis is to provide a sense of how large the effect of omitting a variable or variables would have to be in order to invalidate a finding. That is, sensitivity analysis provides a quantitative statement that in order to explain away a particular association, one would need a hidden bias of a certain size (Rosenbaum, 2002). The canonical example is smoking and lung cancer. Cornfield et al. (1959: 194) demonstrate that if cigarette smokers have 9 times the risk of nonsmokers for lung cancer but only because of some as yet unknown factor X (and thus smoking is not a causal factor), then the proportion of smokers with factor X must be at least 9 times greater than the proportion of nonsmokers with factor X . To

¹⁶Space considerations prevent a full-scale discussion of formal sensitivity analysis. The discussion that follows includes a brief introduction, an example, and some tips for implementation.

explain away an association as strong as that between smoking and lung cancer, then, it is necessary to hypothesize a hidden bias with a very large magnitude. Given that the existence of such a bias is unlikely, we gain confidence in the reliability of our finding, and controlling for every possible omitted variable is unnecessary.

Developed over the last few decades, sensitivity analysis is most closely associated with the work of Paul Rosenbaum (Rosenbaum and Rubin, 1983; Rosenbaum, 1986; Rosenbaum, 2002). The practical approach to sensitivity analysis that we take, however, comes from Imbens (2003). There are two reasons for this choice. First, sensitivity analysis is often discussed in the context of the literature on causal modeling (for an introduction and overview, see Holland, 1986). While Imbens’s discussion is also framed in the familiar “potential outcome” language of Rubin-style causal modeling, the method itself can be used and understood independently of the causal machinery. This independence is of particular importance as many of the variables political scientists would like to subject to sensitivity analysis would not be considered treatments in the Rubin/Holland sense because they are not experimentally manipulable. Second, the method relies on techniques—regression-type models and R^2 s—that are already familiar to quantitative political scientists. The steepness of the learning curve is therefore ameliorated significantly.

The analysis is conducted by making assumptions about the effect of an omitted variable on the dependent variable and on an independent variable of interest. Let the possibly omitted variable be U_i , the variable of interest be W_i , and the other covariates be \mathbf{X}_i . Both U and W are assumed to be 0,1 for simplicity. The distribution of the variable of interest, W_i , given the possibly omitted variable and the other covariates is assumed to be logistic,

$$\Pr(W = 1|\mathbf{X}, U) = \frac{\exp(\gamma'\mathbf{X} + \alpha U)}{1 + \exp(\gamma'\mathbf{X} + \alpha U)}.$$

Furthermore, we assume that the distribution of the dependent variable, Y , is normal, given U and \mathbf{X} ,

$$Y|\mathbf{X}, U \sim N(\tau w + \beta'\mathbf{X} + \delta U, \sigma^2).$$

The trick of sensitivity analysis is to choose values for α and δ , the effect of the possibly omitted variable on the variable of interest and the effect of the possible omitted variable on the dependent variable, respectively, and to calculate the maximum-likelihood estimate, $\hat{\tau}$, of the effect of the variable of interest. So, by varying α and δ , we can get a range of estimates for the effect of the variable of interest on the dependent variable.¹⁷

¹⁷The maximum likelihood estimator of τ comes from maximizing the following log-likelihood function,

$$\begin{aligned} L(\tau, \beta, \sigma^2, \gamma, \alpha, \delta) = & \sum_{i=1}^n \ln \left[\frac{1}{2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \times \exp \left(-\frac{1}{2\sigma^2} (Y_i - \tau W_i - \beta'\mathbf{X}_i)^2 \right) \right. \\ & \times \frac{(\exp(\gamma'\mathbf{X}_i))^{W_i}}{1 + \exp(\gamma'\mathbf{X}_i)} + \frac{1}{2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \times \exp \left(-\frac{1}{2\sigma^2} (Y_i - \tau W_i - \beta'\mathbf{X}_i - \delta)^2 \right) \\ & \left. \times \frac{(\exp(\gamma'\mathbf{X}_i + \alpha))^{W_i}}{1 + \exp(\gamma'\mathbf{X}_i + \alpha)} \right]. \end{aligned}$$

Imbens argues that the sensitivity parameters α and δ are not easily interpretable. In the logistic equation, for example, α 's effect, as opposed to its size, is determined in part by the values taken by \mathbf{X} . A workable solution is to translate the sensitivity parameters into partial R^2 s and compare “the amount of variation that is explained by the unobserved covariate relative to the amount not explained by the observed covariates” (Imbens, 2003: 128). The proportion of the previously unexplained variation in Y that is explained by the unobserved covariate is

$$R_{Y,par}^2 = \frac{R_Y^2(\alpha, \delta) - R_Y^2(0, 0)}{1 - R_Y^2(0, 0)} = \frac{\hat{\sigma}^2(0, 0) - \hat{\sigma}^2(\alpha, \delta)}{\hat{\sigma}^2(0, 0)},$$

where $R_Y^2(\alpha, \delta) = 1 - \hat{\sigma}^2(\alpha, \delta) / \Sigma_Y$ and $\Sigma_Y = \Sigma_i(Y_i - \bar{Y})^2 / N$. For the logistic regression, the partial R^2 is

$$R_{W,par}^2 = \frac{R_W^2(\alpha, \delta) - R_W^2(0, 0)}{1 - R_W^2(0, 0)}.^{18}$$

To perform the sensitivity analysis, we construct pairs of partial R^2 s, $R_{W,par}^2$ and $R_{Y,par}^2$, for pairs of α and δ so that the coefficient on the variable of interest, $\hat{\tau}$, changes by a preset amount. “If the set of all such values does not include reasonable values of the partial R^2 values,” then the sign of the estimated coefficient on the variable of interest is judged robust (Imbens, 2003: 128). Reasonableness is judged by comparing these partial R^2 s to pairs of partial R^2 values corresponding to the observed covariates. Using a criterion such as “reasonableness” may seem to introduce a note of subjectivity into the method; there is no test for reasonableness. On the other hand, the results should not be close. If the pairs of R^2 values obtained from the unobserved covariate are anywhere near those obtained from the observed covariates, then the robustness of the effect of interest must be called into question.

Although we have presented the method using particular parametric forms, logistic for the distribution of W and normal for the distribution of Y , other combinations are possible. Assuming the logistic distribution for W is a holdover from the causal modeling roots of sensitivity analysis, where treatments are generally thought of as present or not. It is certainly plausible, however, to assume a normal distribution for W or a logistic distribution for Y . All that would change is the log-likelihood function that is maximized and whether a natural or implicit R^2 is used.

To make these ideas concrete and transparent, we consider an article by Buhaug and Gates (2002), the objective of which is to “examine factors that determine

¹⁸As there is no natural R^2 in this case, Imbens uses the implicit R^2

$$R_W^2(\alpha, \delta) = \frac{\hat{\gamma}(\alpha, \delta)' \Sigma_{\mathbf{X}} \hat{\gamma}(\alpha, \delta) + \alpha^2 / 4}{\hat{\gamma}(\alpha, \delta)' \Sigma_{\mathbf{X}} \hat{\gamma}(\alpha, \delta) + \alpha^2 / 4 + \pi^2 / 3},$$

where $\Sigma_{\mathbf{X}}$ is the sample covariance matrix of the observed covariates \mathbf{X} with the constant term omitted.

location and scope of civil wars” (p. 420). Location is defined as the distance between the capital city and the conflict center point, and scope is defined as the geographic domain of the conflict zone, measured as the circular area centered around the conflict center and covering all significant battle zones (rounded to the nearest 50-km interval).¹⁹ The dataset includes 265 civil conflicts in the period 1946–2000.

The authors hypothesize that international borders are related to the size of a conflict zone because borders are valuable to the leaders of a rebellion. Rebels attempt to push conflicts to borders because neighboring states may provide safe refuge from government troops, and borders are natural places for rebels to gain access to the weapons and resources trade. 51% of the conflicts in the sample extend to or cross the border of the conflict-ridden state. The results from Buhaug and Gates’s (2002: 427) Model 5 are in Table 1.²⁰ Their results show that a conflict that abuts an international border is, on average, roughly 10 percentage points larger than if the conflict does not abut an international border.

The question we want to answer is how large the effect of an unobserved covariate or covariates would have to be to destroy the substantive significance of the estimated coefficient on border. We can imagine, for instance, any number of non-geographic variables that might be correlated with scope of conflict and border. Whether or not the rebel group constitutes an ethnic faction is just one such example.²¹ Such non-geographic variables might not be included in the specification for a host of reasons. The data on these variables might not be available, or perhaps they are measured with significant error, or the concepts might not be measurable, or including such variables is simply a distraction for a researcher interested solely in the effects of geographic variables. Formal sensitivity analysis can be used in any of these instances to help assess the robustness of the coefficient on border.

There are two practical issues that must be confronted before the sensitivity analysis can be performed. The first issue is by what preset amount the coefficient on the variable of interest, $\hat{\tau}$, should change. The second is how to choose values of α and δ that move $\hat{\tau}$ by that preset amount.

On the first issue, note that the standard error on border is roughly 4.5. If the observed coefficient were to decrease by two standard errors, it would be essentially zero, and border would not have any substantive significance. Thus, to know how large an effect a non-geographic variable or variables would have to have in order to render border spurious, we treat border as W and construct our pairs of partial R^2 s for pairs of our sensitivity parameters, α and δ , that decrease the estimated coefficient on border by two standard errors.

The second issue is choosing values for the sensitivity parameters, α and δ . A grid search is a reasonable, if not efficient, way to proceed. We can specify a range

¹⁹The *relative* scope of conflict is the conflict zone as a proportion of total land area.

²⁰The additional covariates include location (the distance from the conflict center to the capital center), land area (the size of the country), duration (the duration of the conflict), and resource (whether or not the conflict zone contains natural resources).

²¹Ethnic groups often cross state lines, and a rebel ethnic minority funded by an ethnic majority elsewhere would serve to widen the scope of the conflict.

Table 1. Relative Scope

Variable	Coefficient	Standard Error
Location	5.64	1.235**
Land area	-14.88	1.226**
Duration	0.77	0.302**
Border	9.49	4.545**
Resource	17.51	5.533**
Constant	91.62	6.879**
N	246	
R ²	0.374	

** $p \leq 0.05$.

of values for α and the same set of values for δ . We then obtain values for $\hat{\tau}$ by maximizing the log-likelihood function for every pair of α and δ . For those pairs that change $\hat{\tau}$ by the preset amount, we calculate the partial R^2 s and report them. Running the analysis is computer intensive. If we let α range between -10 and 10 by distances of 0.2 , and at the same time, let δ range between -10 and 10 by distances of 0.2 , there are $10,201$ pairs of values to assess, which means maximizing the log-likelihood function $10,201$ times. Unfortunately, we cannot always get away with a smaller number because, depending on the preset amount chosen for the change in $\hat{\tau}$, we may need larger values of α and δ to produce the desired change. A relatively inexpensive method of checking the necessary size of the grid is to increase the distances between the values of α and δ .²²

The results of the sensitivity analysis on border are in Figure 3. The curve describes how strongly the unobserved covariate would have to be correlated with the scope of conflict and whether or not the conflict abuts a border in order to make the coefficient on border be zero. An unobserved covariate would have to explain, for example, 20% of the variation in border and 15% of the variation in scope not explained by the observed covariates to make the coefficient on border zero. Now, compare the effects of the observed covariates. All lie below the curve. Thus, in order for an unobserved covariate to wipe out the effect of border on the scope of conflict, it would have to explain more of the variation in border and scope than location of the conflict, the size of the state, the duration of the conflict, and whether or not the conflict zone includes natural resources. Given that the existence of a variable of that importance is unlikely, we judge the coefficient on border to be robust. It is possible, therefore, to assess the possible impact of omitted variables without resorting to including large numbers of control variables. The Buhaug and Gates (2002) regression includes a manageable number of covariates, and we can be reasonably certain that incorporating non-geographic variables into the specification would not change the results on border.

What if the coefficient on border had been judged non-robust? The sensitivity analysis tells us that another variable or set of variables remains unaccounted for, but not what those variables are or where to find them. Some might see this as a

²²R code for implementing the analysis in this article is available from the author.

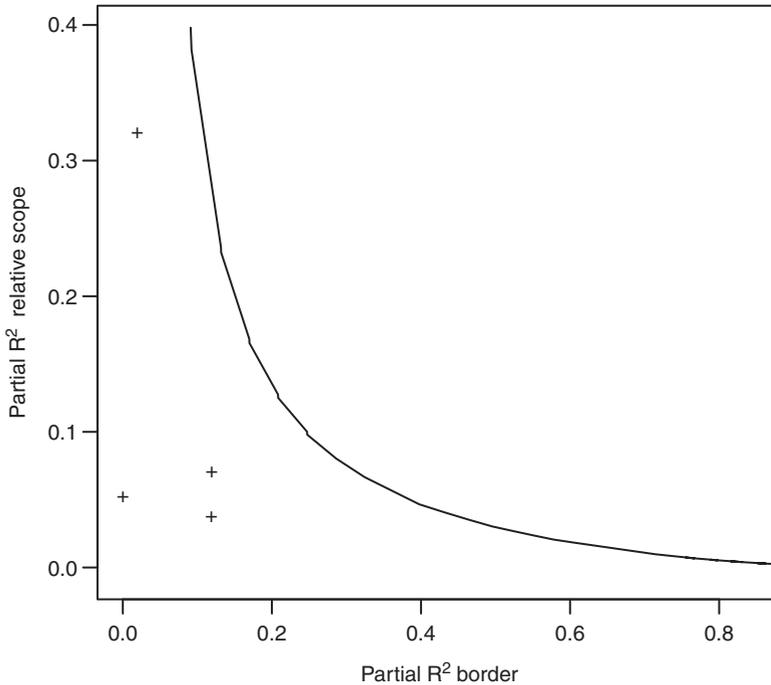


Figure 3. How Large a Hidden Bias Would Have To Be in Order To Make the Coefficient on Border Zero

drawback of the method. However, even if we cannot measure or discover the missing variable or variables, it is important to know that the effect of the coefficient in question cannot be trusted. The situation is analogous to our prejudice in favor of the null hypothesis in a significance test. We wish to guard against concluding that the effect is real when, in fact, it is not. Furthermore, the sensitivity analysis can serve as a spur toward additional theorizing in the hopes of discovering the unobserved variables.

No amount of theorizing, however, can tell us all the variables that should be included in a specification. The reason is that theory is general, and data are specific to a particular time and place. The effects of variables of interest depend in part on local variables about which good social theory is silent. Our results, then, are always vulnerable to omitted variable bias, and sensitivity analysis can always be useful.

Thus, sensitivity analysis can be used in two ways. First, the method can be used prophylactically to ward off reviewers intent on having their pet variable included in every regression in their area of expertise. Second, and more importantly, the method can be used to increase confidence in our statistical results. The more information we have about our results—good or bad—can only help the progress of our field. There is no reason why the phantom menace should continue to scare us into specifying large, unmanageable statistical specifications. By

making use of techniques such as formal sensitivity analysis, we can make reliable inferences with only a handful of carefully chosen variables.

Conclusion

Omitted variable bias is a serious problem, and it is the goal of textbook treatments of omitted variable bias to demonstrate that fact, in much the same way that textbooks demonstrate that the estimated coefficients of correctly specified models are minimum variance unbiased. These demonstrations, however, are equally far removed from the everyday practice of quantitative political science. Just as we are likely never in the position of working with a correctly specified model, we are likely never in the position of considering a single omitted variable or a single set of omitted variables. Rather, we are faced with models that are, at best, first-order approximations, and we are faced with decisions concerning the inclusion of a subset of the set of omitted variables.

The effect of including such a subset in a regression-type equation . . . depends. It depends on the effects of the included and excluded variables; it depends on the correlations between the included and excluded variables; it depends on the variances of all the variables. The phantom menace is elusive. By including additional control variables in our specifications, we could very easily be making the bias on the coefficient of interest worse. Knowing for sure requires knowing much more than we typically do in practice. In the absence of this kind of omniscience, we need approaches to achieving reliable inferences that have fewer debilitating side effects. Formal sensitivity analysis fulfils that role by providing evidence that is more convincing than a regression equation weighed down by half a dozen control variables, and convincing evidence is the foundation of a compelling science.

Appendix: Finding the Normalized Coefficient

Consider the following latent data generating process

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad \text{where } Y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

Also consider two misspecified models where X_2 and X_3 are left out of Model 1, and only X_3 is left out of Model 2,

$$\text{Model 1: } Y_i^* = \beta_{01} + \beta_{11} X_{i1} + \epsilon_{i1},$$

$$\text{Model 2: } Y_i^* = \beta_{02} + \beta_{12} X_{i1} + \beta_{22} X_{i2} + \epsilon_{i2}.$$

To find the normalized coefficient on X_1 in Model 1, let X_2 be a function of both X_1 and X_3 and an independent error term, and let X_3 be a function of both X_1 and X_2 and an independent error term,

$$X_{i2} = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i3} + v_i$$

$$X_{i3} = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \eta_i.$$

Solving the above equations in terms of X_1 , we can substitute them into equation (5),

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 \left[\frac{\gamma_0 + \gamma_1 X_{i1} + \gamma_2 \delta_0 + \gamma_2 \delta_1 X_{i1} + \gamma_2 \eta_i + v_i}{1 - \gamma_2 \delta_2} \right] + \beta_3 \left[\frac{\delta_0 + \delta_1 X_{i1} + \delta_2 \gamma_0 + \delta_2 \gamma_1 X_{i1} + \delta_2 v_i + \eta_i}{1 - \delta_2 \gamma_2} \right] + \epsilon_i$$

and collect terms,

$$(1 - \gamma_2 \delta_2) Y_i^* = (\beta_0 - \beta_0 \gamma_2 \delta_2 + \beta_2 \gamma_0 + \beta_2 \gamma_2 \delta_0 + \beta_3 \delta_0 + \beta_3 \delta_2 \gamma_0) + (\beta_1 - \beta_1 \gamma_2 \delta_2 + \beta_2 \gamma_1 + \beta_2 \gamma_2 \delta_1 + \beta_3 \delta_1 + \beta_3 \delta_2 \gamma_1) X_{i1} + (\epsilon_i - \epsilon_i \gamma_2 \delta_2 \eta_i + \beta_2 v_i + \beta_3 \delta_2 v_i + \beta_3 \eta_i).$$

The normalized coefficient on X_1 is therefore

$$\hat{\beta}_{11} = \frac{c[\beta_1 - \beta_1 \gamma_2 \delta_2 + \beta_2(\gamma_1 + \gamma_2 \delta_1) + \beta_3(\delta_1 + \delta_2 \gamma_1)]}{\sqrt{\sigma^2 + \sigma^2 \gamma_2^2 \delta_2^2 + \sigma_\eta^2(\beta_2^2 \gamma_2^2 + \beta_3^2) + \sigma_v^2(\beta_2^2 + \beta_3^2 \delta_2^2)}}. \quad (6)$$

To perform the same analysis on Model 2, where only X_3 is omitted, assume that X_3 is a function of X_1 , X_2 , and an independent error term,

$$X_{i3} = \delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \eta_i.$$

Substituting the above into equation (5) and collecting terms, we get,

$$Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3(\delta_0 + \delta_1 X_{i1} + \delta_2 X_{i2} + \eta_i) + \epsilon_i = (\beta_0 + \beta_3 \delta_0) + (\beta_1 + \beta_3 \delta_1) X_{i1} + (\beta_2 + \beta_3 \delta_2) X_{i2} + (\epsilon_i + \beta_3 \eta_i),$$

which makes the normalized coefficient

$$\hat{\beta}_{12} = \frac{c}{\sqrt{\sigma^2 + \beta_3^2 \sigma_\eta^2}} (\beta_1 + \beta_3 \delta_1). \quad (7)$$

References

- Achen, Christopher H. 1992. Social psychology, demographic variables, and linear regression: Breaking the iron triangle in voting research. *Political Behavior* 14 (September): 195–211.
- Achen, Christopher H. 2002. Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423–450.
- Achen, Christopher H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22(4): 327–339.
- Akaike, H. 1973. Information theory and an extension of the likelihood ratio principle. In *Second International Symposium of Information Theory*, eds B. N. Petrov and F. Csaki, pp. 267–281. Budapest: Minnesota Studies in the Philosophy of Science.
- Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113 (February): 151–184.
- Ansolabehere, Stephen, Alan Gerber, and Jim Snyder. 2002. Equal votes, equal money: Court-ordered redistricting and public expenditures in the American states. *American Political Science Review* 96 (December): 767–777.

- Bailey, Michael A., Brian Kamoie, and Forrest Maltzman. 2005. Signals from the Tenth Justice: The political role of the Solicitor General in Supreme Court decision making. *American Journal of Political Science* 49 (January): 72–85.
- Brady, Henry E., and John E. McNulty. 2004. The costs of voting: Evidence from a natural experiment. Paper presented at the annual meeting of the Society for Political Methodology, Stanford University.
- Breiman, Leo. 1992. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* 87 (September): 738–754.
- Buhaug, Halvard, and Scott Gates. 2002. The geography of civil war. *Journal of Peace Research* 39 (July): 417–433.
- Clarke, Kevin A. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45 (July): 724–744.
- Clarke, Kevin A. 2003. Nonparametric model discrimination in international relations. *Journal of Conflict Resolution* 47 (February): 72–93.
- Clarke, Kevin A. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22 (Winter): 341–352.
- Clarke, Kevin A. 2007. A simple distribution-free test for nonnested hypotheses. *Political Analysis* 15 (Summer): 347–363.
- Cornfield, J., W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. 1959. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* 22(1): 173–203.
- Cox, David R. 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley Symposium I: 105–123.
- Cramer, J. S. 2003. *Logit models*. Cambridge: Cambridge University Press.
- Davidson, Russell, and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. Oxford: Oxford University Press.
- Gail, M.H., S. Wieand, and S. Piantadosi. 1984. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 71 (December): 431–444.
- Gelpi, Christopher, and Peter D. Feaver. 2002. Speak softly and carry a big stick? Veterans in the political elite and the American use of force. *American Political Science Review* 96 (December): 779–793.
- Greene, William H. 2003. *Econometric analysis*, 5th edn. New Jersey: Prentice Hall.
- Griliches, Zvi. 1977. Estimating the results to schooling: Some econometric problems. *Econometrica* 45(January): 1–22.
- Gujarati, Damodar N. 2003. *Basic econometrics*, 4th edn. New York: McGraw-Hill.
- Hanushek, Eric A., and John E. Jackson. 1977. *Statistical methods for social scientists*. New York: Academic Press.
- Hegre, Hävard, Tanja Ellingsen, Scott Gates, and Nils Petter Gleditsch. 2001. Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. *American Political Science Review* 95(March): 33–48.
- Hendry, David F., and Jean-Francois Richard. 1982. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20(October): 3–33.
- Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(December): 945–960.
- Imbens, Guido W. 2003. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* 93(May): 126–132.
- Johnston, Jack, and John DiNardo. 1997. *Econometric methods*, 4th edn. New York: McGraw-Hill.

- Judge, George G., W. E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. 1985. *The theory and practice of econometrics*, 2nd edn. New York: John Wiley and Sons.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.
- King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review* 95(March): 49–69.
- Kmenta, Jan. 1986. *Elements of econometrics*, 2nd edn. New York: Macmillan.
- Krause, George A. 2003. Coping with uncertainty: analyzing risk propensities of SEC budgetary decisions, 1949–97. *American Political Science Review* 97(February): 171–188.
- Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.
- Mallows, C.L. 1973. Some comments on Cp. *Technometrics* 15(November): 671–676.
- McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*, 2nd edn. New York: Chapman and Hall.
- Meyer, Bruce D. 1995. Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 13(April): 151–161.
- Neuhaus, John M. 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* 80(December): 807–815.
- Rosenbaum, Paul R. 1986. Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics* 11(Autumn): 207–224.
- Rosenbaum, Paul R. 2002. *Observational Studies*, 2nd edn. New York: Springer-Verlag.
- Rosenbaum, P. R., and D. B. Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B* 45(2): 212–218.
- Rudolph, Thomas J., and Jillian Evans. 2005. Political trust, ideology, and public support for government spending. *American Journal of Political Science* 49(July): 660–671.
- Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.
- Sims, Christopher A. 1980. Macroeconomics and reality. *Econometrica* 48(January): 1–48.
- Stein, C. 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, pp. 197–206.
- Vuong, Quang. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57(March): 307–333.
- Welch, Finis. 1975. Human capital theory: Education, discrimination, and life-cycles. *American Economic Review* 65(May): 63–73.
- Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

KEVIN A. CLARKE is an associate professor in the Political Science Department at the University of Rochester. He is a political methodologist with interests in quantitative theory comparison, philosophy of science, and international relations.