Conflict Management and Peace Science, 22:341–352, 2005
Copyright © Peace Science Society (International)
ISSN: 0738-8942 print / 1549-9219 online
DOI: 10.1080/07388940500339183

Routledge
Taylor & Francis Group

The Phantom Menace: Omitted Variable Bias in Econometric Research

KEVIN A. CLARKE

Department of Political Science
University of Rochester
Rochester, New York, USA

Quantitative political science is awash in control variables. The justification for these bloated specifications is usually the fear of omitted variable bias. A key underlying assumption is that the danger posed by omitted variable bias can be ameliorated by the inclusion of relevant control variables. Unfortunately, as this article demonstrates, there is nothing in the mathematics of regression analysis that supports this conclusion. 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. A brief discussion of alternative strategies for achieving experimental control follows the main result.

Keywords omitted variable bias, specification, control variables, research design

Quantitative political science is awash in control variables. It is not uncommon to see statistical models with 20 or more independent variables. An article in the August 2004 issue of the *American Political Science Review,* for example, reports a model with 22 independent variables (Duch & Palmer, 2004).[1] The situation is no different if we consider the *American Journal of Political Science*, the *Journal of Conflict Resolution*, the *Journal of Politics*, or *Political Analysis*.

The justification for these bloated specifications, when one is provided, is usually the fear that omitted relevant variables will bias the results. The unspoken, and often unwritten, belief seems to be that the inclusion of every additional relevant variable serves to reduce the potential threat from omitted variable bias. That is, a researcher cannot know all 20 variables that appear in the data-generating process, but if she knows and includes 12 of them, she is better off than if she knows and includes only 10 of them.

Unfortunately, there is nothing in the mathematics of regression analysis that supports this conclusion. The omitted variable result so familiar from the standard econometrics texts and graduate school addresses the omission of one variable or a set of variables from a regression. The result tells us nothing about the situation where a subset of the set of omitted variables is included in a specification as controls. The reason for this lacuna is no

Earlier versions of this article were presented at the Annual North American Meeting of the Peace Science Society, Houston, Texas, November 2004 and the Annual International Studies Association Convention, Honolulu, Hawaii, March 2005. I thank participants for their comments, and Sara McLaughlin Mitchell for inviting me to the Peace Science meeting. I also thank two anonymous reviewers for their comments and Fabiana Machado and Arthur Spirling for research assistance. Support from the National Science Foundation (Clarke: Grant #SES-0213771) is gratefully acknowledged.

Address correspondence to Kevin A. Clarke, Department of Political Science, University of Rochester, Rochester, NY, 14627-0146. E-mail: kevin.clarke@rochester.edu

[1] The median number of independent variables for the entire issue is 11 with an outlier at 3.

mystery: the effect of including such variables depends, even in a simple case, on a host of factors.

The only thing that can be said for certain 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. Omitted variable bias truly is a phantom menace. This result has important implications for our understanding of control variables and the way in which we approach the specification of our statistical models.

The Familiar Result

A brief review of the familiar omitted variable result is useful for following the discussion. Suppose that the correct specification of a regression model is

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{1}$$

but we estimate the misspecified model

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i^*, \tag{2}$$

where $\epsilon_i^* = \beta_4 X_{i4} + \epsilon_i$, and β_2 is the coefficient of interest.

Under the assumption that the expected value of ϵ_i is zero, the expected value for $\hat{\beta}_2$ is given by Hanushek and Jackson (1977) as

$$E[\hat{\beta}_2] = \beta_2 + \beta_4 b_{42}, \tag{3}$$

where

$$b_{42} = \frac{(r_{42} - r_{32} r_{43})}{1 - r_{32}^2} \sqrt{\frac{V_4}{V_2}}. \tag{4}$$

b_{42} is the regression coefficient on X_2 in the "auxiliary" regression of the excluded variable, X_4, on the included variables, X_2 and X_3. Thus, the effect of omitting X_4 depends on the magnitude of the excluded coefficient, β_4, the correlations between the included variables and the excluded variable, r_{42} and r_{43}, the correlation of the included variables, r_{32}, and the variances of X_2 and X_4 (denoted V_2 and V_4).[2]

The standard omitted variable bias lesson often concludes with results that show that the inclusion of irrelevant variables produces inefficient coefficient estimates. Textbook

[2] The general result addresses the situation where a set of relevant variables is omitted from the estimated equation. Suppose that the correct specification of the regression model is

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

but we estimate

$$\mathbf{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*,$$

where $\boldsymbol{\epsilon}^* = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$. The expected value of $\boldsymbol{\beta}_1$, under the usual assumptions, is given by Greene (2003) to be

$$E[\hat{\boldsymbol{\beta}}_1] = \boldsymbol{\beta}_1 + \mathbf{P}_{1.2} \boldsymbol{\beta}_2,$$

where $\mathbf{P}_{1.2} = (\boldsymbol{X}_1^{'} \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^{'} \boldsymbol{X}_2$ is the matrix of regression coefficients from the auxiliary regression of the excluded variables, $\boldsymbol{X}_2$, on the included variables, $\mathbf{X}_1$.

authors then note that there exists a trade-off between bias and inefficiency when adding variables to a regression specification.[3]

Unfortunately, the standard lesson rarely applies in practice because, as data analysts, we are never in the situation described in the textbooks. That is, we are never in the position of choosing between including the final relevant variable (or set of variables) and including an irrelevant variable. Rather, we are faced with choosing to include an additional relevant variable (or set of variables) out of a larger set of relevant omitted variables.[4] The important question is the effect of adding to a regression some, but not all, of these relevant omitted variables.

The Logic of Control Variables

Although we are rarely in the situation described in econometrics texts, the logic of control variables flows directly, albeit mistakenly, from the standard omitted variable bias result. The argument seems to be that we decrease the aggregate bias for every additional relevant variable that we include. The inefficiency part of the equation is rarely mentioned, as control variables often do have real effects. 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.[5]

The basic logic is seductive and can be 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, & Verba, 1994, 173, 175). Kadera and Mitchell (2005, 13), in their contribution to this volume, consider not including a set of control variables and, being "schooled in the King, Keohane, and Verba (1994) approach to research design," they wonder " 'what about omitted variable bias?' "

The logic of control variables comprises some version of the following four points:

- control variables have real effects so their inclusion does not cause inefficiency,
- control variables have real effects so their absence, when correlated with included variables, may well cause bias and inconsistency,
- the bias caused by omitted variables is an aggregation of the bias caused by each individual omitted variable,
- as the inclusion of all relevant, correlated variables is impossible, we should include as many as possible in order to reduce the bias.

To put the argument in mathematical terms, consider a true model

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{5}$$

[3] The inefficiency discussion often adds fuel to the control variable fire. Johnston and DiNardo (1997, 110) write that the consequences of including irrelevant variables are "generally less serious than those pertaining to the exclusion of relevant variables."

[4] Even those in favor of including multiple control variables, such as Oneal and Russett (2000, 5), do not claim that their analyses include every variable in the data-generating process.

[5] There 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.

and two misspecified models

$$\text{Model 1: } Y_i = \beta_{11} + \beta_{21} X_{i2} + \epsilon_{i1}, \tag{6}$$

$$\text{Model 2: } Y_i = \beta_{12} + \beta_{22} X_{i2} + \beta_{32} X_{i3} + \epsilon_{i2}. \tag{7}$$

The claim is that the bias on $\hat{\beta}_{21}$, the estimated coefficient on X_2 in model 1, is greater than the bias on $\hat{\beta}_{22}$, the estimated coefficient on X_2 in model 2. Letting the bias on $\hat{\beta}_{21}$, $E[\hat{\beta}_{21}] - \beta_2$, be denoted as $b(\hat{\beta}_{21}, \beta_2)$, and the bias on $\hat{\beta}_{22}$, $E[\hat{\beta}_{22}] - \beta_2$, be denoted as $b(\hat{\beta}_{22}, \beta_2)$, the mathematical argument is that

$$b(\hat{\beta}_{21}, \beta_2) \geq b(\hat{\beta}_{22}, \beta_2). \tag{8}$$

The mathematics of regression analysis, however, do not support the conclusion that the bias on $\hat{\beta}_{21}$ is necessarily greater than or equal to the bias on $\hat{\beta}_{22}$. The inclusion of additional relevant variables *may* reduce the bias on the X_2 coefficient, but it may also have the opposite effect and increase the bias on the X_2 coefficient. Short of knowing all omitted relevant variables, the researcher cannot know which is the case.

A Simple Example

As a simple example, consider the situation given above where the true model is

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{9}$$

If we estimate a misspecified model that omits X_3 and X_4,

$$Y_i = \beta_{11} + \beta_{21} X_{i2} + \epsilon_{i1}, \tag{10}$$

the expected value of $\hat{\beta}_{21}$, given that the expectation of the true stochastic term ϵ_i is 0, is

$$E[\hat{\beta}_{21}] = \beta_2 + \beta_3 \left(r_{23} \sqrt{\frac{V_4}{V_2}} \right) + \beta_4 \left(r_{24} \sqrt{\frac{V_4}{V_3}} \right).^{6} \tag{11}$$

If we estimate a misspecified model that omits only X_4,

$$Y_i = \beta_{12} + \beta_{22} X_{i2} + \beta_{32} X_{i3} + \epsilon_{i2}, \tag{12}$$

the expectation of $\hat{\beta}_{22}$, given that the expectation of the true stochastic term ϵ_i is 0, is

$$E[\hat{\beta}_{22}] = \beta_2 + \beta_4 \left(\frac{(r_{24} - r_{23} r_{34})}{1 - r_{23}^2} \sqrt{\frac{V_4}{V_2}} \right). \tag{13}$$

[6] As before, V_k is the variance of the kth independent variable, and r_{kj} is the correlation between the kth and jth independent variables.

In order to investigate the logic of control variables, we want to know under what conditions the bias on $\hat{\beta}_{21}$ is greater than or equal to the bias on $\hat{\beta}_{22}$,

$$b(\hat{\beta}_{21}, \beta_2) \geq b(\hat{\beta}_{22}, \beta_2). \tag{14}$$

Solving equations (11) and (13) for the conditions under which the above inequality (14) holds yields messy and unilluminating results due to the large number of moving parts. We can, however, investigate the effects of a few of these moving parts with relative ease through simulation. Two particular factors invite closer scrutiny, r_{34}, the correlation between the newly included variable, X_3, and the remaining omitted variable, X_4, is one of the major differences between $E[\hat{\beta}_{21}]$ and $E[\hat{\beta}_{22}]$ and thus will be allowed to vary in the simulation. The sign of β_4, the coefficient on X_4, plays a significant role in all discussions of omitted variable bias, and thus it will also be allowed to vary.

β_4 will be allowed to vary between -5.0 and 5.0 in order to investigate scenarios where the effect of the remaining omitted variable is either positive or negative. By the same token, the correlation between X_3 and X_4 will be allowed to vary over its full range. That is,

- $\beta_4 \in \{-5, \ldots, 5\}$,
- $r_{34} \in \{-1, \ldots, 1\}$.

To help isolate the effects of these factors, the first two regression coefficients, β_2 and β_3, are set at 4, all variances are set at 1, and the correlations between X_2 and X_3 and X_2 and X_4 are set to 0.5. That is,

- $\beta_2 = \beta_3 = 4$,
- $V_2 = V_3 = V_4 = 1$,
- $r_{23} = r_{24} = 0.5$.

Results

The effects of r_{34} and β_4 are shown in Figure 1, which contains a graph of the difference in the absolute values of the two biases,

$$|b(\hat{\beta}_{21}, \beta_2)| - |b(\hat{\beta}_{22}, \beta_2)|. \tag{15}$$

Negative values, therefore, indicate that the inclusion of the additional relevant variable, X_3, increases the bias on the estimated coefficient of X_2 compared to the case where both X_3 and X_4 are omitted. Positive values indicate that the inclusion of the additional relevant variable, X_3, decreases the bias on the estimated coefficient of X_2 compared to the case where both X_3 and X_4 are omitted.

What Figure 1 shows is that including X_3 in the regression is just as likely to increase the bias on $\hat{\beta}_2$ as it is to decrease it. The lighter shaded areas on Figure 1 indicate values of β_4 and r_{34} for which the absolute value of the bias on $\hat{\beta}_{22}$ is greater than the bias on $\hat{\beta}_{21}$. It is clear that the addition of X_3 in the regression is more likely to increase the bias on $\hat{\beta}_2$ when both β_4 and r_{34} are negative (the front lower corner of the cube). That said, there are additional combinations under which the bias on $\hat{\beta}_2$ may increase. For instance, the addition of X_3 may increase the bias when β_4 is positive (particularly between 3 and 5) and r_{34} is negative (the right front of the cube). The addition of X_3 may also increase the bias when β_4 is negative and r_{34} is positive (the back left corner of the cube).

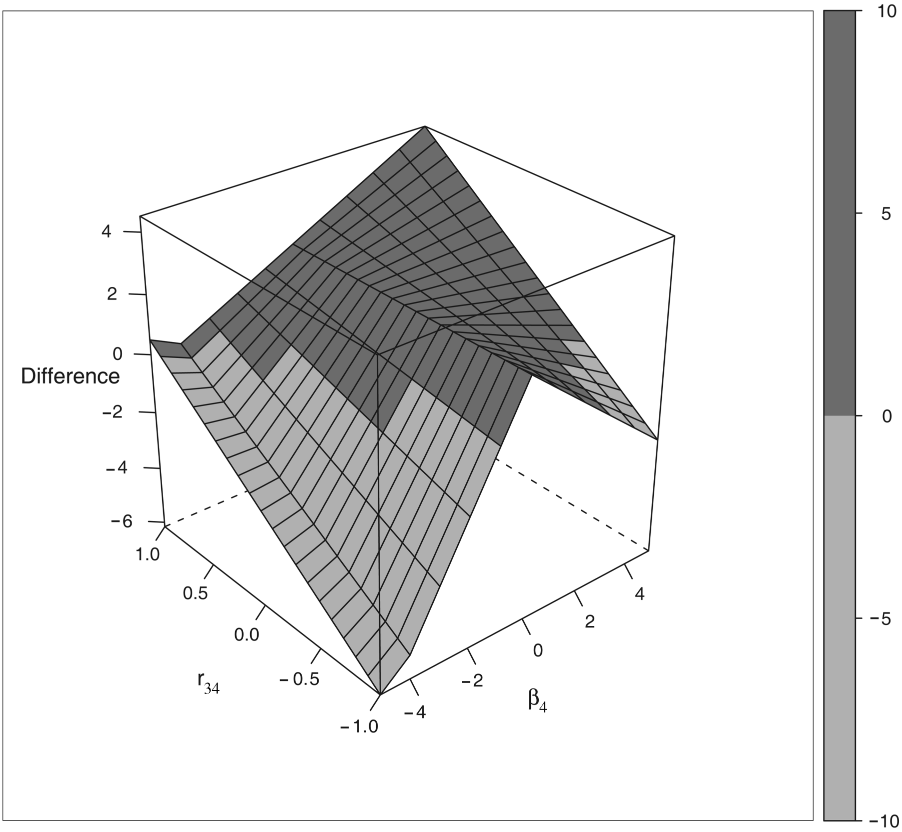

FIGURE 1 The effect of β_4 and r_{34} on the difference in the absolute values of the two biases.

This simple example 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 the bias, but it is just as likely to increase the bias. In short, we cannot know the effect on the bias of including an additional control variable unless we know the complete and true specification.[7]

Efficiency and the Phantom Menace

What is the effect of the decision to include or exclude control variables on the efficiency of our coefficient(s) of interest? If the control variables are irrelevant, as previously noted, inclusion means that our estimator is no longer minimum variance. Is the opposite true when the control variables are relevant? That is, does the inclusion of relevant control variables decrease the variance of the coefficient of interest? The quick answer is no.

Consider again a true model

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \tag{16}$$

[7] Even if we could not measure all of the omitted variables, we would at least need to know what they are in order to ensure that their inclusion would cause the bias to decrease.

and two misspecified models:

$$\text{model 1: } Y_i = \beta_{11} + \beta_{21} X_{i2} + \epsilon_{i1}, \tag{17}$$

$$\text{model 2: } Y_i = \beta_{12} + \beta_{22} X_{i2} + \beta_{32} X_{i3} + \epsilon_{i2}. \tag{18}$$

The variance of $\hat{\beta}_{21}$ in model 1 is given by

$$\text{Var}(\hat{\beta}_{21}) = \frac{\sigma^2}{s_2}, \tag{19}$$

where $s_2 = \sum_{i=1}^{n} (X_{i2} - \bar{X}_2)^2$. The variance of $\hat{\beta}_{22}$ in model 2 is given by

$$\text{Var}(\hat{\beta}_{22}) = \frac{\sigma^2}{s_2(1 - r_{23}^2)}. \tag{20}$$

Thus, provided that the correlation between the two included variables, X_2 and X_3, is not 0, the variance of $\hat{\beta}_{21}$ is less than the variance of $\hat{\beta}_{22}$.[8] Adding a variable, therefore, can never decrease the variance of the coefficient of interest; the variance can only increase or stay the same.

The above result concerning the variance of our misspecified models and our previous result concerning the bias of our misspecified models can be combined and compared using the mean square error criterion. As is well known, the mean square error is the expected value of the squared difference between an estimator and its parameter,

$$\text{MSE}(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] \tag{21}$$

$$= \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2. \tag{22}$$

In our case, we want to compare the mean square error of $\hat{\beta}_{21}$, MSE($\hat{\beta}_{21}$), with the mean square error of $\hat{\beta}_{22}$, MSE ($\hat{\beta}_{22}$), and determine under what conditions the following inequality holds:

$$\text{MSE}(\hat{\beta}_{21}) \geq \text{MSE}(\hat{\beta}_{22}), \tag{23}$$

$$\text{Var}(\hat{\beta}_{21}) + \text{Bias}(\hat{\beta}_{21})^2 \geq \text{Var}(\hat{\beta}_{22}) + \text{Bias}(\hat{\beta}_{22})^2. \tag{24}$$

Equation (24) is particularly useful because we know that the variance of $\hat{\beta}_{21}$ will

[8] In general, suppose that the correct specification is

$$\mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}),$$

and we estimate two misspecified models:

$$\text{model 1: } \mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_{11} + \boldsymbol{\epsilon}_1^*,$$

$$\text{model 2: } \mathbf{y} = \boldsymbol{X}_1\boldsymbol{\beta}_{12} + \boldsymbol{X}_2\boldsymbol{\beta}_{22} + \boldsymbol{\epsilon}_2^*.$$

The difference in the inverses of the covariance matrices for $\hat{\boldsymbol{\beta}}_{11}$ and $\hat{\boldsymbol{\beta}}_{12}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}_{11})^{-1} - \text{Var}(\hat{\boldsymbol{\beta}}_{12})^{-1} = \frac{1}{\sigma^2}\boldsymbol{X}_1'\boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1,$$

which is nonnegative definite. The variance of $\hat{\boldsymbol{\beta}}_{11}$ therefore cannot be larger than the variance of $\hat{\boldsymbol{\beta}}_{12}$.

always be less than or equal to the variance of $\hat{\beta}_{22}$. Therefore, whenever the bias on $\hat{\beta}_{21}$ is less than the bias on $\hat{\beta}_{22}$, the MSE of $\hat{\beta}_{21}$ must be less than the MSE on $\hat{\beta}_{22}$. That is,

$$\text{MSE}(\hat{\beta}_{21}) < \text{MSE}(\hat{\beta}_{22}) \tag{25}$$

in all of the situations given in Figure 1 where the bias on $\hat{\beta}_{21}$ is less than the bias on $\hat{\beta}_{22}$ (the lighter shaded regions). Thus, whether we are concerned about bias, efficiency, or mean squared error, the inclusion of even relevant control variables may make the situation worse.

Discussion

The importance of these results lies in the fact that once the connection between omitted variable bias and control variables is gone, the main justification for using control variables is gone. The mathematics of regression analysis simply do not support the logic of control variables previously laid out. Including more variables in a regression, even relevant ones, does not necessarily make the regression results more accurate.

The result should not surprise us. Others have made similar arguments on different grounds. In his contribution to this volume, Achen (2005) shows that small amounts of nonlinearity in control variables can have deleterious effects, and Griliches (1977, 12) argues that small amounts of measurement error in control variables "are magnified as more variables are added to the equation in an attempt to control for other possible sources of bias." He notes that we may "kill the patient in our attempts to cure what may have been a rather minor disease originally." Welch (1975) and Maddala (1977) make similar points.

Yatchew and Griliches (1985), branching out beyond the linear model, examine the effects of misspecification, including omitted variables, on the estimation of probit models. The bottom line is that the nonlinear specification complicates everything. In the omitted variable case, there are various asymptotic biases that exist depending on how the samples are drawn. Unlike the linear model, coefficients may be biased even if the omitted variable is uncorrelated with the included variables. Discerning the effect of a single omitted variable is therefore difficult, and discerning the effect of including a subset of relevant omitted variables is nearly impossible.

Finally, 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, 2003), and econometric methodologies such as the London School of Economics approach associated with David Henry and colleagues (Hendry & Richard, 1982) and Christopher Sims's (1980) vector autoregression approach. 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. 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).

The Logic of Research Design

If the logic of control variables is flawed, experimental control must be achieved in another way. How this can be done is no mystery. In their seminal econometrics text, Hanushek and Jackson (1977) discuss how to use research design in place of using control variables to address potential omitted variable bias. These include basing specification on theory, finding "natural" experiments, and "controlling" for unmeasured effects through careful sample stratification. To these we can add the explicit use of competing theories and the choice of research hypothesis (Rosenbaum, 1999; Freedman, 1991). These techniques go back to the earliest days of econometrics and are suitable for use in any observational study.

Thus, in place of the logic of control variables, we can speak of the logic of research design. This logic comprises some version of the following points:

- it is impossible to include all the relevant variables in a regression equation,
- omitted variable bias is therefore unavoidable,
- the inclusion of a subset of relevant control variables may not ameliorate, and may increase, the bias caused by omitted variables,
- the inclusion of a subset of relevant control variables may also cause additional biases through measurement error,
- experimental control can, however, be achieved through careful research design.

The easiest way to utilize this logic in actual practice is to test broad theories in narrow, focused, controlled circumstances. As Rosenbaum (1999) argues, broad theories are important because such theories make predictions across a variety of domains. There is no requirement, however, that broad theories must be tested across all their domains at the same time. A broad theory can be tested in a particular place and at a particular time, and while such a test is far from definitive, a series of these narrow controlled tests is far more convincing than the alternative. The limited circumstances provide a level of experimental control that control variables cannot.[9]

Good international relations scholarship demonstrates the point. IR scholars too often attempt to test their theories on the population of militarized disputes from 1816 to the present.[10] In back-to-back articles, Zeev Maoz and Bruce Russett make a better choice when analyzing alternative explanations of the democratic peace (Maoz & Russett, 1992, 1993). Instead of anchoring their temporal domain to 1816, Maoz and Russett consider only the Cold War period, 1946–1986. Limiting their temporal domain to this 40-year period has two significant benefits. First, the authors point out that the number of states in the international system increased dramatically prior to this period (Maoz & Russett, 1992, 248). By limiting their temporal domain, Maoz and Russett do not need to include a new variable to control for the number of states. Oneal and Russett (2005, 10–11), on the other hand, extend their temporal domain back 60 years to 1885 and do have to include this control variable.

Second, Maoz and Russett (1992, 248) argue that the meaning of democracy had changed "substantially," citing as reasons the enfranchisement of women and ethnic groups, mass participation, and the advent of mass media. In the 1993 article, the authors also argue that democratic norms are more deeply entrenched during the Cold War period than the pre-1945 period (Maoz & Russett, 1993, 627). By limiting their temporal domain, however, Maoz and Russett do not need to control for the changing meaning of democracy or the strength of democratic norms. (See Oren, 1995, for a convincing argument that the nature of democracy is not constant.)

One could, of course, argue that Maoz and Russett might have pursued this logic further by restricting their spatial domain to the Western Hemisphere, because including the rest of

[9] Starr (2005, 2) makes a similar point when discussing "domain specific" laws.

[10] A short list of influential articles in which the temporal domain begins at 1816 includes Lai and Reiter (2000); Rasler and Thompson (1999); Bueno de Mesquita et al. (1999); Werner (1999); Enterline (1998); Bennett (1997); Bennett and Stam (1996); Siverson and Starr (1994); and Huth, Gelpi, and Bennett (1993).

the world "introduces greater spatial and cultural variation in the data" (Maoz & Russett, 1992, 248). Demonstrating the effect of regime type, however, is not Maoz and Russett's only goal. The authors are also interested in demonstrating the *independent* effect of regime type, which means considering possible confounding explanations such as wealth, alliances, and political stability. Maoz and Russett need spatial variation beyond the Western Hemisphere in order to have variation on these other factors. They argue that spatial variation allows "a more complex test of the basic hypothesis," that is, "a test designed to display the power of competing hypotheses" (Maoz & Russett, 1993, 627). Thus, Maoz and Russett carefully chose their temporal and spatial domains to further their substantive goals without having to rely on a series of unnecessary control variables.

My point here is not to suggest that Maoz and Russett's work stands as the last word on the democratic peace, nor is it to comment on the current state of democratic peace research (see Huth & Allee, 2003, for an up-to-date review). Indeed, the fact that these two articles concern the democratic peace is quite beside the point. Rather, I am arguing that substituting research design for control variables can be as simple as testing a theory in a particular place and at a particular time. By limiting their temporal domain to the Cold War era, Maoz and Russett made research design choices that are consistent with the arguments of Rosenbaum (1999) and Freedman (1991). While not definitive, Maoz and Russett's results are more convincing than similar studies that use every bit of available data and attempt to achieve control through the inclusion of additional 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, much in 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 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 an approach to achieving convincing experimental control that has fewer debilitating side effects. Substituting design for control does exactly that. Narrow, focused, controlled tests of broad theories, while unlikely to be definitive, provide evidence that is far more convincing than a regression equation weighed down by half a dozen control variables, and convincing evidence is the foundation of a compelling science.

References

Achen, C. H. 2005. Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22: this issue.

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. Minnesota Studies in the Philosophy of Science. Budapest: Akademinai Kiado.

Bennett, D. S. 1997. Testing alternative models of alliance duration, 1816–1984. *American Journal of Political Science* 41: 846–878.

Bennett, D. S., and A. C. Stam. 1996. The duration of interstate wars, 1816–1985. *American Political Science Review* 90: 239–257.

Breiman, L. 1992. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* 87: 738–754.

Bueno de Mesquita, B., J. D. Morrow, R. M. Siverson, and A. Smith. 1999. Policy failure and political survival: The contribution of political institutions. *Journal of Conflict Resolution* 43: 147–161.

Clarke, K. A. 2001. Testing nonnested models of international relations: Reevaluating realism. *American Journal of Political Science* 45: 724–744.

Clarke, K. A. 2003. Nonparametric model discrimination in international relations. *Journal of Conflict Resolution* 47: 72–93.

Cox, D. R. 1961. Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium* I: 105–123.

Duch, R. M., and H. D. Palmer. 2004. It's not whether you win or lose, but how you play the game: Self-interest, social justice, and mass attitudes toward market transition. *American Political Science Review* 98: 437–452.

Enterline, A. J. 1998. Regime changes, neighborhoods, and interstate conflict, 1816–1992. *Journal of Conflict Resolution* 42: 804–829.

Freedman, D. A. 1991. Statistical models and shoe leather. *Sociological Methodology* 21: 291–313.

Greene, W. H. 2003. *Econometric aanalysis*. 5th ed. Englewood Cliffs, NJ: Prentice Hall.

Griliches, Z. 1977. Estimating the results to schooling: Some econometric problems. *Econometric* 45: 1–22.

Hanushek, E. A., and J. E. Jackson. 1977. *Statistical methods for social scientists*. New York: Academic Press.

Hendry, D. F., and J.-F. Richard. 1982. On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20: 3–33.

Huth, P., C. Gelpi, and D. S. Bennett. 1993. The escalation of great power militarized disputes: Testing rational deterrence theory and structural realism. *American Political Science Review* 87: 609–623.

Huth, P. K., and T. Allee. 2003. *The democratic peace and territorial conflict in the twentieth century*. Cambridge, UK. Cambridge University Press.

Johnston, J., and J. DiNardo. 1997. *Econometric methods*. 4th ed. New York: McGraw-Hill.

Kadera, K. M., and S. M. Mitchell. 2005. Heeding Ray's advice: An exegesis on control variables in systemic democratic peace research. *Conflict Management and Peace Science* 22: this issue.

King, G., R. O. Keohane, and S. Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Lai, B., and D. Reiter. 2000. Democracy, political similarity, and international alliances, 1816–1992. *Journal of Conflict Resolution* 44: 203–227.

Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics* 15: 671–676.

Maoz, Z., and B. Russett. 1992. Alliance, contiguity, wealth, and political stability: Is the lack of conflict among democracies a statistical artifact? *International Interactions* 17: 245–267.

Maoz, Z., and B. Russett. 1993. Normative and structural causes of democratic peace, 1946-1986. *American Political Science Review* 87: 624–638.

Oneal, J. R., and B. Russett. 2005. Rule of three, let it be? When more really is better. *Conflict Management and Peace Science.* 22: 293–310.

Oren, I. 1995. The subjectivity of the democratic peace: Changing U.S. perceptions of imperial Germany. *International Security* 20: 174–184.

Rasler, K., and W. R. Thompson. 1999. Predatory initiators and changing landscapes for warfare. *Journal of Conflict Resolution* 43: 411–433.

Rosenbaum, P. R. 1999. Choice as an alternative to control in observational studies. *Statistical Science* 14: 259–304.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464.

Sims, C. A. 1980. Macroeconomics and reality. *Econometrica* 48: 1–48.

Siverson, R. M., and H. Starr. 1994. Regime change and the restructuring of alliances. *American Journal of Political Science* 38: 145–161.

Starr, H. 2005. Cumulation from proper specification: Theory, logic, research design, and "nice" laws. *Conflict Management and Peace Science* 22: 353–363.

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, Q. 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57: 307–333.

Welch, F. 1975. Human capital theory: Education, discrimination, and life-cycles. *American Economic Review* 65: 63–73.

Werner, S. 1999. The precarious nature of peace: Resolving the issues, enforcing the settlement, and renegotiating the terms. *American Journal of Political Science* 43: 912–934.

Yatchew, A., and Z. Griliches. 1985. Specification error in probit models. *The Review of Economics and Statistics* 67: 134–139.