Conflict Management and Peace Science
© The Author(s). 2012. Reprints and permissions:
http://www.sagepub.co.uk/journalsPermissions.nav
[DOI: 10.1177/0738894211433170]
Vol 29(2): 239–241

More Phantom Than Menace[1]

KEVIN A. CLARKE
University of Rochester

In this article, I address whether the inclusion of control variables in the hopes of getting an unbiased estimate of the residual variance is a good reason for the inclusion of control variables. I conclude that the inclusion of an additional control variable is unlikely to decrease the estimated standard errors and that the tradeoffs involved with including a new control variable are simply too large.

KEYWORDS: control variables; model specification; omitted variable bias

In "The phantom menace" (Clarke, 2005) and "Return of the phantom menace" (Clarke, 2009), I tackle two common arguments for the inclusion of control variables in regression specifications. The first argument is that the omission of relevant control variables that are correlated with the variables already included in a specification biases the estimated coefficients on the included variables. The second argument is that including irrelevant control variables has relatively few harmful effects.[2] These two arguments lead to what I call the "logic of control variables", which states that if bias is the result of omitted variables and the inclusion of irrelevant variables causes little damage, then our regressions should include as many of the variables that we have access to as possible. Increasingly bloated specifications are the natural outcome of this thinking. My papers demonstrate that if a researcher is not in the exact situation described by the textbooks (the specification differs from the data generating process by a single variable or set of variables), then this logic does not hold.

There is, however, a third argument for the inclusion of control variables. I relegated it to a footnote in "Return of the phantom menace", and I am happy that Vance and Ritter (2012) have expanded upon it. The argument is that the omission of relevant control variables biases the residual variance, which in turn, affects the standard errors. Vance and Ritter show that it is possible that the estimated

[1] I thank Michael Peress for helpful comments and advice.

[2] In their comment, Vance and Ritter (2012: X) write that I make a "strong assumption that is rarely met in practice" (that the newly included variable adds no explanatory power to the regression). My discussion of the irrelevant control variable argument *requires* that I make this assumption.

standard error on a coefficient of interest may decrease from the inclusion of a relevant control variable, and their analysis is correct. What I want to address in this short note is a topic that neither Vance and Ritter nor I previously addressed: is the standard error argument a *good* reason for including control variables?

Answering the question I just raised requires assessing the frequency with which we expect the inclusion of a control variable to decrease the standard errors, as well as the tradeoffs involved. As to the former, the inclusion of a control variable decreases the standard error on the estimate of interest … sometimes. In Vance and Ritter's simulation (p. X–X), the inclusion of the control variable *fails* to decrease the standard error over 70% of the time. They also note that the necessary conditions for a decrease are sometimes met in natural experiments, but natural experiments account for a small percentage of the empirical work in political science. Finally, Vance and Ritter argue that it is unlikely that a newly included variable has no effect (and thus is irrelevant) on the dependent variable (p. X–X). By the same token, if the newly included variable has an effect on the dependent variable, it is likely to be correlated with the included variables at a significant level. Everyone who does empirical work has had the experience of including a control variable that changes the magnitude of the estimated coefficients on the already included variables because the variables are correlated. Just a moderate correlation is enough to prevent the newly included control variable from decreasing the standard errors.

Even if the control variable were to decrease the standard errors, the tradeoffs involved with including a new control variable could be large. First, as I demonstrated in my previous papers, including a control variable could increase the bias on the estimated coefficient of interest, and unless we know how that variable interacts with relevant variables that remain unobserved, we cannot be sure of its effect on the bias. Second, the control variable may introduce measurement error, which is a singularly destructive force in empirical analysis. Even dichotomous variables are often measured with error.[3] A similar argument can be made about the introduction of nonlinearities into a specification. Third, a control variable might introduce endogeneity. Finally, any number of other problems could be introduced that affect the standard errors. These include varying coefficients, multiplicative error terms, incorrect data transformations, influential observations, skewed regressors, aggregate data, cross-sectional data with units of different sizes, excessive smoothing, and data correlated over time and space. The bottom line is that including a control variable in the hopes of decreasing the standard error on the coefficient of interest is simply not worth the risk.

Vance and Ritter cite Angrist and Pischke (2009) in their comment, and it is instructive to turn to the "Last Words" section of Angrist and Pischke's text, where they write, "Your standard errors probably won't be quite right, but they rarely are" (2009: 327). Nearly 30 years ago, Chris Achen (1982) came to the same conclusion, writing that "the standard errors are likely to be wrong" (p. 39), and

[3] Use of a dichotomous variable is usually an admission that we do not know something, which is why they are called "dummy" variables. Examples include gender, race, and party identification.

in areas of political science where tens of thousands of regressions have been run, "wise investigators know far more about the true variability across observations and samples than any statistical calculation can tell them" (p. 40). A statistical model is a description, and descriptions are always partial and always context dependent. It follows that there is no such thing as a perfect description; there are only more or less useful descriptions (Clarke and Primo, 2012). A regression model describes the dependencies among the variables in a data set, and the description is likely to be more useful when it includes a manageable number of covariates (so that the relationships between the variables can be explored fully), makes defendable assumptions, and carries natural interpretations. An adequate description often requires multiple different regression models. Properly understood, omitted variables pose little threat to statistical analysis. They are more phantom than menace.

References

Achen, C. 1982. *Interpreting and Using Regression*. Thousand Oaks CA: Sage.

Angrist, J. D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Clarke, K. A. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22(4): 341–352.

Clarke, K. A. 2009. Return of the phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 26(1): 46–66.

Clarke, K. A., and D. M. Primo. 2012. *A Model Discipline: Political Science and the Logic of Representations*. New York: Oxford University Press.

Vance, C., and N. Ritter. 2012. The phantom menace of omitted variables: A comment. *Conflict Management and Peace Science* 29(2): XX–XX.

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