Omitted Variable Bias

Paul A. Jargowsky
University of Texas at Dallas

I. The Danger of Omitted Variable Bias
II. Omitted Variable Bias in the Bivariate Case
III. Omitted Variable Bias in Multivariate Analyses
IV. Coping with Omitted Variable Bias
V. Conclusion

GLOSSARY

bias the difference between the expected value of an estimator and the parameter being estimated.

coefficient in regression, the estimated effect of an independent variable on the dependent variables, controlling for the other independent variables in the model.

correlation a measure of the linear association between two quantitative variables

estimator a statistic based on a sample for the purpose of providing information about a population parameter.

expected value the mean of the probability distribution of a random variable

partial correlation a measure of the linear association between two quantitative variables after controlling for the effect of one or more additional variables

unbiased the property of an estimator that has an expected value equal to the parameter being estimated.

Omitted variable bias, also know as left out variable bias, is the difference between the expected value of an estimator and the true value of the underlying parameter due to failure to control for a relevant explanatory variable or variables.

I. The Danger of Omitted Variable Bias

The usual goal of a quantitative social science study is to estimate the direction, magnitude, and statistical significance of the effect of an independent variable on a dependent variable. When one or more variables that ought to be included in a model is left out, our estimate of the effect of the variables we do include in the model is likely to be in error. Even increasing the sample size or repeating the study multiple times will not help to solve the problem.

Social science phenomena are, by their very nature, multivariate. Yet human reasoning tends to operate in a bivariate manner. When we see two variables correlated in time or space, we tend to leap to the conclusion that one variable must be causing the other. Moreover, data tables and graphics are printed on a two-dimensional page and are most effective at showing bivariate relationships. Unfortunately, except in very favorable circumstances, conclusions based on bivariate reasoning are likely to be incorrect to one degree or another because of omitted variable bias. In fact, even analyses which take dozens of variables into account will still be biased if even one important explanatory variable is overlooked. In all likelihood, omitted variable bias is the most serious and pervasive threat to the validity of social science research.

This article describes why left out variable bias occurs, why researchers need to be concerned about it, and how to deal with it. The article works within a linear regression framework, but the concerns are generalizable to many different types of analysis, both more and less sophisticated than linear regression. The next section describes how an omitted variable biases the coefficient in a bivariate regression, and sets forth guidelines for determining the direction of the bias. Section III extends the analysis to multiple regression, and describes a more refined rule for predicting the direction of the bias which is less well know than the procedure for the bivariate case. Section IV identifies strategies for coping with omitted variable bias.

II. Omitted Variable Bias in the Bivariate Case

This section addresses the bias in the context of a bivariate regression. The next section extends the analysis to multiple regression.

A. The Algebra of the Bias.

For the purpose of understanding how the influence of an omitted variable can be picked up by other variables in an analysis, we begin with a simple case in which a dependent variable *Y* is determined by two variables X_2 and X_3, plus a stochastic disturbance term. Thus, the true model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad [1]$$

The betas are the parameters to be estimated. The disturbance term, u_i, is assumed to be uncorrelated with the X variables.

Let us assume, however, that the researcher lacks data on the variable X_3 and proposes to regress *Y* on X_2 alone. To clear X_3 from the equation, and thus recast the true model solely as a function of X_2, we first need to express X_3 as a function of X_2 as follows:

$$X_{3i} = \gamma_1 + \gamma_2 X_{2i} + \varepsilon_i \, . \qquad [2]$$

We need not concern ourselves about whether equation 2 expresses a causal relationship or merely a spurious correlation; the gammas are simply the parameters of the ideal line that captures the linear correlation between the two X variables. The epsilon term is the difference between the actual value of X_3 and the conditional mean of X_3 given X_2. In other words, it is the part of X_3 that is uncorrelated with X_2.

To clear X_3 from the model, we simply substitute equation 2 into equation 1:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 \left(\gamma_1 + \gamma_2 X_{2i} + \varepsilon_i \right) + u_i \\ &= \beta_1 + \beta_2 X_{2i} + \beta_3 \gamma_1 + \beta_3 \gamma_2 X_{2i} + \beta_3 \varepsilon_i + u_i \\ &= \underbrace{\left(\beta_1 + \beta_3 \gamma_1 \right)}_{\alpha_1} + \underbrace{\left(\beta_2 + \beta_3 \gamma_2 \right)}_{\alpha_2} X_{2i} + \underbrace{\left(\beta_3 \varepsilon_i + u_i \right)}_{\tilde{u}_i} \end{aligned} \qquad [3]$$

After regrouping, the terms in the first parenthesis are all constants, so for convenience we can label the sum of these constants "${}_1$. The terms in the second parenthesis are slopes; that is, changes in Y associated with changes in X_2. We label the sum of these slopes "${}_2$. The terms in the final parenthesis are stochastic (random error) terms that are uncorrelated with X_2.

The point is that when we mistakenly estimate the bivariate regression of Y on X_2, omitting X_3 and assuming that the disturbance term is uncorrelated with X_2, we actually obtain an unbiased estimate of "${}_2$ rather than an unbiased estimate of the true causal effect, $\$_2$. Specifically,

$$E\left[\hat{\alpha}_2 \right] = \alpha_2 = \beta_2 + \beta_3 \gamma_2 \qquad [4]$$

The first term ($\$_2$) is the true causal effect, the quantity we wish to estimate. The second term is the omitted variable bias. (The formal derivation of omitted variable bias involves substituting the true model for Y_i into the formula for the OLS slope coefficient, and then taking the expectation. See Appendix A.) Note that the bias has two components. The first ($\$_3$) is the true effect of the omitted variable. The second ($(_2$) is the parameter on X_2 in Equation 2 above. Because Equation 2 is in the form of a bivariate regression, we can express this parameter as follows:

$$\gamma_2 = \rho_{23}\left(\frac{\sigma_3}{\sigma_2}\right) \qquad [5]$$

in which D_{23} is the correlation between X_2 and X_3, and the sigmas are the standard deviations of the corresponding X variables.

B. The Direction of the Bias.

The two components of the bias make it possible to estimate the direction of the bias. Since standard deviations are always positive, the sign of the bias will be determined by the product of the signs of $\$_3$, the true effect of the omitted variable, and the correlation between the included and the excluded variables. If both are positive, or both are negative, then the bias will be positive, meaning that the coefficient estimate in the bivariate estimation is likely to be to high. If one is positive and the other negative, then the bias term will be negative, and the coefficient estimate in the bivariate case is likely to be too low. If either of these two terms is zero, then we need not worry about an omitted variable bias.

We don’t know these signs for certain; after all, the variable is omitted from our analysis. But often it is possible to have a good idea about probable sign of these quantities based on theoretical considerations, prior empirical research, or common sense.

C. Examples.

Suppose we regress wage in dollars on years of education, but omit a variable for experience. We might obtain a result like the following bivariate regression:

wage = 1.00 + 0.75*educ

If our model were correctly specified, we would be justified in inferring that an additional year of education leads to an increase of 75 cents in the wage rate. However, we have omitted experience, which is not directly available in most cross-sectional data sets, such as the Current Population Survey.

Even without a variable for experience, we can make an educated guess about the direction of the bias. Based on human capital considerations, we expect the sign of experience to be positive. The correlation between education and experience is likely to be negative, both because people stay out of the labor market to pursue education and because older cohorts of workers generally obtained less education but have accumulated much work experience. A positive times a negative is negative, so the bias is negative. The bivariate estimate of the effect of education on wage is, therefore, likely to be too low relative to the unbiased estimate that controls for experience.

While a biased estimate is a bad thing, at least in this case we can say that our estimate of the effect of education on wage is conservative; were we able to include experience, the coefficient would likely be even higher. That is, the effect of education on wage is probably greater than 75 cents per year of education.

On the other hand, we have also omitted ability. Ability is likely to have a positive effect on wage and a positive correlation with education. Therefore the bias will be positive as well.

In this case the bias is more troubling. Depending on the size of the bias, the true effect of education could be zero or even negative.

As the two examples show, an educated guess about the direction of the bias can be very helpful, particularly if the bias has the opposite sign of the coefficient. A positive coefficient with a negative bias is likely to be an even larger positive number; similarly, a negative coefficient with a positive bias is likely to indicate a larger – that is, more negative – coefficient.

It is worth noting at this point that bias is a property of the expected value of the coefficient, not any specific estimate. In any given sample, the value of a coefficient can be quite different than its expectation, so there is no guarantee that a negatively biased coefficient will be smaller than the true parameter value or vice versa.

III. Omitted Variable Bias in Multivariate Analyses

The analysis of omitted variable bias in the bivariate case is useful to develop an understanding about the mechanics of the bias. Most of the time, however, we are concerned about the effect of an omitted variable in a multiple regression analysis. The analysis of the previous section has to be modified as described below.

A. The Algebra of the Bias.

Anticipating the direction of the bias is somewhat more complex when we move beyond bivariate regression models. For example, suppose the true causal model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad [6]$$

Further, suppose that X_4 is related to X_2 and X_3 as follows:

$$X_{4i} = \gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \varepsilon_i \qquad [7]$$

Substituting and rearranging terms:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 \left(\gamma_1 + \gamma_2 X_{2i} + \gamma_3 X_{3i} + \varepsilon_i \right) + u_i \\ &= \underbrace{\left(\beta_1 + \beta_4 \gamma_1 \right)}_{\alpha_1} + \underbrace{\left(\beta_2 + \beta_4 \gamma_2 \right)}_{\alpha_2} X_{2i} + \underbrace{\left(\beta_3 + \beta_4 \gamma_3 \right)}_{\alpha_3} X_{3i} + \underbrace{\left(u_i + \beta_4 \varepsilon_i \right)}_{\tilde{u}_i} \end{aligned} \qquad [8]$$

Thus, when estimating the regression of Y on X_2 and X_3, omitting X_4, the slope parameters for both of the included independent variables are potentially biased. The expected values for the slope parameters are:

$$\begin{aligned} E\left[\hat{\alpha}_2\right] &= \beta_2 + \beta_4 \gamma_2 \\ E\left[\hat{\alpha}_3\right] &= \beta_3 + \beta_4 \gamma_3 \end{aligned} \qquad [9]$$

As in bivariate case, there is no bias if the true effect of the omitted variable – β_4 – is zero. Unlike the bivariate case, the gammas are now multiple regression coefficients. Thus, they cannot be expressed as a function of the simple bivariate correlation coefficients between the omitted variable and the included variables. Instead, γ_2 is the slope coefficient for X_2 that one would obtain from a regression of X_4 – the missing variable – on X_2 and X_3. If that coefficient would be zero, there is no bias. The sign of the that coefficient is the sign of the *partial* correlation between X_2 and X_4 controlling for X_3. As in the bivariate case, the direction of the bias is determined by the product of the signs of the two terms in the bias equation.

Extension to situations involving more variables is straightforward. (For a formal derivation of the bias in the multivariate case, see Greene (2000: 334).) If there are two excluded variables, X_4 and X_5, the expected value of the coefficient on X_2 is:

$$E\left[\hat{\alpha}_2\right] = \beta_2 + \beta_4\gamma_{42} + \beta_5\gamma_{52} \qquad [10]$$

where γ_{42} is the coefficient expressing the partial effect of X_2 on X_4 and γ_{52} is the coefficient expressing the partial effect of X_2 on X_5, with X_3 controlled in both cases.

In general, if X_j are the *J* variables included in the regression, and Z_k are the *K* variables omitted from the regression, the expected value of the coefficient any one of the included variables is:

$$E\left[\hat{\alpha}_j\right] = \beta_j + \sum_k \beta_k\gamma_{kj} \qquad [11]$$

Thus, each left out variable potentially contributes to the bias. For a given left out variable *k*, there is no bias if the omitted variable has no effect on the dependent variable after controlling for the included variables, i.e. $\beta_k = 0$ for that variable. There is also no bias for a given included variable *j* if the partial correlation between the omitted and the included variables is zero after controlling for the other included variables, i.e. $\gamma_{kj} = 0$.

While it is clearly more difficult in the multivariate case, one can still anticipate the direction of omitted variable bias by making educated guesses about the sign of the effect of the excluded variable and the *partial* correlation of the included variables in questions and the excluded variable. Most of the time, but certainly not always, the sign of the partial correlation and the simple bivariate correlation are the same. Nevertheless, one must guard against uncritically applying the bivariate analysis in the multivariate case, and the distinction is not as well appreciated as it ought to be. Table 1 summarizes the issues involved in evaluating the existence and direction of the bias. However, if there are several omitted variables and some of the expected biases are positive and some are negative, it may be difficult if not impossible to anticipate direction of the net expected bias.

[Table 1 about here.]

C. Broader Implications.

An often neglected point is that omitting relevant variables affects the standard errors of the included variables as well. There are several potentially offsetting effects, so that the standard errors may be larger or smaller relative to the correctly specified model. However, in view of the bias in the coefficients when relevant variables are left out, the effects on the standard errors seem like a second-order concern, since the wrong quantity is being estimated. There is no bias if the omitted variable is not correlated with the included variable, but the standard errors of the coefficients on the included variables will still be affected. On the one hand, including an omitted variable uses up one degree of freedom, potentially increasing the standard error. On the other hand, including an omitted variable reduces the residual variance, which will tend to reduce the standard errors of the coefficients of the included variables. In practice, the latter effect will usually predominate, especially in larger data sets where degrees of freedom are plentiful.

The foregoing discussion has been framed in the context of OLS regression, but the ideas developed are applicable to a broad array of models which include a linear function of several variables as part of their specification. Omitted variable bias has roughly the same properties in all models in which the dependent variable is modeled as some function of a linear combination of dependent variables and parameters. For example, a logit model is specified as follows:

$$P_i = \frac{1}{1+e^{-(\beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}} \qquad [12]$$

Solving for the linear function of the independent variables, the model can be written as:

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \qquad [13]$$

Although the relationship between the *Xs* and P_i is clearly non-linear, the relationship is monotonic; therefore, a positive $ implies that the probability increases as *X* increases and a negative coefficient implies that the probability declines as *X* increases and a larger coefficient implies a larger effect, other things equal. Omitted variable bias is a function of the relationship between the *X* variables, not the functional form of the relationship with the dependent variable. In other words, one can still make the same generalizations about the *direction* of the bias as described above when considering logit, probit, tobit, Poisson, and many other models.

On a conceptual level, we can also extend these ideas to tables, graphs, and maps. For example, given the rise of Geographic Information Systems (GIS), it is common for researchers to show maps and note that two problems, say percentage black and the high school drop out rate, are geographically concentrated in the same neighborhoods. This is nothing more than a visual bivariate correlation, without even the benefit of a test of significance. Any conclusion one would wish to draw from such maps is tainted by the omission of other variables, such as the poverty rate, which may well be the partly or fully determine the outcome variable.

Omitted variable bias is very closely related to path analysis (Alwyn and Hauser 1975). In fact, the mathematics of the two are identical; they differ only in the assumptions about the relations among the variables. What has been described as a bias in this article is called an indirect effect in path analysis. In order for the indirect effect interpretation to be valid, the researcher must be able to correctly specify the causal ordering of the variables in the model.

IV. Coping with Omitted Variable Bias

The definitive solution to omitted variable bias is to conduct a classical experiment, in which individuals are randomly assigned to treatment and control groups. Because both groups are random draws from the same population, they have the same expected value for any variable other than the treatment. Because no other variable, measured or unmeasured, known or unknown, is likely to be correlated with the treatment, there is little reason to be concerned with omitted variable bias in the analysis of experimental data. In many issues studied by social scientists, however, it is not possible to conduct experiments or, in cases where it might be possible, it is often not ethical. We can't randomly assign workers to a gender, and we wouldn't randomly assign children to abusive parents. And even where experiments are possible and ethical, they are expensive and difficult to conduct. Thus, non-experimental research is an unavoidable necessity, despite the possibility and probability of omitted variable bias.

The second best way to correct omitted variable bias is to collect new data which either include the variables omitted in previous analyses. One should then include the formerly omitted variables and perform the relevant t or F tests; depending on the results of these tests, either include or exclude the variable from the final model. However, this is not always practical or possible. For example, some variables are intrinsically hard to measure, such as a person's level of motivation. And in some cases a researcher may not even be aware of the variable which is being omitted because of a lack of understanding of the phenomenon under study.

For all of these reasons, researchers will often have no choice but to conduct analyses on data which may be missing important explanatory variables. This section addresses strategies and techniques for recognizing and addressing omitted variable bias when conducting an experiment or collecting new data are not possible.

A. Testing for Omitted Variable Bias.

Any time the R^2 from a regression is less than 1, some aspect of the process that generates the dependent variable is unknown. Thus, the first diagnostic for omitted variables is a low R^2 statistic. However, R^2 less than 1 is a necessary but not a sufficient condition for omitted variable bias. The unexplained variation could be random noise, or it could be the systematic impact of omitted variables which, taken together, are uncorrelated with any of the included variables.

A number of tests have been developed which can aid the researcher in diagnosing the likelihood of left out variable bias. The most commonly used is the Regression Specification Error Test (RESET) proposed by Ramsey (1969). The intuition behind the test is that the residuals from a regression with a left out variable will include a systematic component, reflecting the net influence of the omitted variables. However, by construction, the residuals are orthogonal to both the predicted values and all the included independent variables. Thus, the residuals from the suspect regression are themselves regressed against squares and higher powers of either the predicted values or the independent variables. Under the null hypothesis of no specification error, the sum of the squared residuals from the auxiliary will have an F distribution. Unfortunately, the test is a general mis-specification test, and cannot distinguish between omitted variables and incorrect functional form.

B. Correcting Omitted Variable Bias

When the data do not contain one or more variables of interest, there are a number of ways to control for them implicitly.

In panel data, fixed effects models may be estimated. The fixed effect model implicitly controls for all variables that are unchanging over the period in question. For example, fixed effects at the individual level can be used to control for innate ability in a wage regression; fixed effects for multiple-sibling families can be used to control for unmeasured characteristics of the parents. Similarly, paired difference tests and "difference in difference" estimators allow individual observations to serve as their own control.

An instrumental variable approach (IV) may be attempted, if variables can be found which are related to the included variables but not the omitted variables. By purging the included variable of its correlation with the omitted variable, consistent estimates of the included variable's influence may be obtained. For example, quarter of birth is correlated with schooling because of compulsory education laws, but presumably uncorrelated with ability (Angrist and Krueger 1991). Thus, an IV based on quarter of birth will allow a consistent estimate of the effect of schooling on wage that is not biased by the omission of a measure of ability. Obviously, this technique depends on the availability of suitable instruments.

If there is no way to add the omitted variable or to control for it implicitly, the best one can do is to try to evaluate the likely impact of the bias following the logic of Sections II and III above.

V. Conclusion

Social phenomena are the results of complex multivariate processes in which the causal variables are often correlated. Ultimately, the best way to recognize a left out variable problem is to use theory, prior empirical research, and common sense as a guide to what variables ought to be included in the model. Given the near impossibility of measuring all the relevant variables, virtually all social science analyses are plagued by omitted variable bias to one degree or another. However, understanding the mechanics of omitted variable bias can suggest the probable size and direction of the bias. Perhaps more importantly, it can help the researcher to decide which left out variables cause the most serious threat to the validity of research findings, and therefore indicate where further time, money, and effort can be most usefully deployed.

Appendix A: Expectation of the Bivariate Slope Coefficient with an Omitted Variable

Assume that the true model is:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad u_i \sim N\left(0, \sigma^2\right) \qquad [A.1]$$

However, the following model is estimated:

$$Y_i = \alpha_1 + \alpha_2 X_2 + \varepsilon_i \qquad [A.2]$$

For ease of presentation, let lower case letters represent deviations from the respective means. In other words:

$$y_i = Y_i - \overline{Y}$$
$$x_{2i} = X_{2i} - \overline{X}_2$$
$$x_{3i} = X_{3i} - \overline{X}_3$$

It is also useful to note the following relation:

$$\sum x_{2i} y_i = \sum x_{2i}\left(Y_i - \overline{Y}\right) = \sum x_{2i} Y_i - \overline{Y} \sum x_{2i} = \sum x_{2i} Y_i \qquad [A.3]$$

The second term disappears because the sum of deviations around a mean is always zero. The expectation of the slope coefficient from the incorrect bivariate regression is:

$$E\left[\hat{\alpha}_2\right] = E\left[\frac{\sum x_{2i} y_i}{\sum x_{2i}^2}\right] = E\left[\frac{\sum x_{2i} Y_i}{\sum x_{2i}^2}\right] \qquad [A.4]$$

The first step is the standard solution for the bivariate slope coefficient (see, for example, Gujarati 2003: 62). In the second step we make use of equation A.3. Now we substitute the true model (Equation A.1), simplify the expression, and take the expectation:

$$
\begin{aligned}
E\left[\hat{\alpha}_2\right] &= E\left[\frac{\sum x_{2i}\left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right)}{\sum x_{2i}^2}\right] \\
&= E\left[\frac{\beta_1 \sum x_{2i} + \beta_2 \sum x_{2i} X_{21} + \beta_3 \sum x_{2i} X_{3i} + \sum x_{2i} u_i}{\sum x_{2i}^2}\right] \\
&= E\left[0 + \beta_2 + \beta_3\left(\frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2}\right) + \frac{\sum x_{2i} u_i}{\sum x_{2i}^2}\right] \\
&= \beta_2 + \beta_3 \gamma_{32}
\end{aligned}
\qquad [A.5]
$$

This is the result noted in Equation 4 in the text. The first term goes to zero because the sum of deviations around a mean is always zero, and the final term has an expectation of zero because the model assumes no covariance between the *Xs* and the disturbance term. The raw values of X_2 and X_3 are replaced by their deviations (x_2 and x_3) in the third step by making use of the property demonstrated in Equation A.3.

Further Reading

Alwin, Duane F. and Robert M. Hauser (1975). "The Decomposition of Effects in Path Analysis." *American Sociological Review* 40:37-47.

Angrist, J. and A. Krueger (1991). "Does Compulsory Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106:979-1014.

Goldberger, Arthur S. (1974). “Unobservable Variables in Econometrics.” in *Frontiers in Econometrics*, ed. by P. Zarembka. New York: Academic Press.

Greene, William H. (2000). *Econometric Analysis*. 4th ed. Upper Saddle River, New Jersey: Prentice Hall.

Griliches, Zvi (1957). “Specification Bias in Estimates of Production Functions.” *Journal of Farm Economics* 39: 8-20.

----- (1977). “Estimating the Returns to Schooling: Some Econometric Problems.” *Econometrica* 45: 1-22.

Gujarati, Damodar N. (2003). *Basic Econometrics.* 4th ed. New York: McGraw-Hill.

Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statisitical Society. Series B (Methodological)* 31: 350-71.

Table 1: Existence and Direction of Omitted Variable Bias

Partial correlation between omitted variable and the included variable of interest, net of the other included variables:	**Marginal effect of the omitted variable on the dependent variables, controlling for all the included variables**		
	Negative	**Zero**	**Positive**
Negative	*Positive Bias*	None	*Negative Bias*
Zero	None	None	None
Positive	*Negative Bias*	None	*Positive Bias*

Note: if more than one variable is omitted, each can cause a bias. If the biases are in opposite directions, it is difficult if not impossible to tell the net expected direction of the bias.