Omitted-variable bias

From Wikipedia, the free encyclopedia

In statistics, **omitted-variable bias** (**OVB**) occurs when a model created incorrectly leaves out one or more important factors. The "bias" is created when the model compensates for the missing factor by over- or underestimating the effect of one of the other factors.

More specifically, OVB is the bias that appears in the estimates of parameters in a regression analysis, when the assumed specification is incorrect in that it omits an independent variable that is correlated with both the dependent variable and one or more included independent variables.

Contents

In linear regression

Intuition

Two conditions must hold true for omitted-variable bias to exist in linear regression:

- the omitted variable must be a determinant of the dependent variable (i.e., its true regression coefficient is not zero); and
- the omitted variable must be correlated with an independent variable specified in the regression (i.e., cov(z,x), is not equal to zero).

Suppose the true cause-and-effect relationship is given by

$$y = a + bx + cz + u$$

with parameters *a, b, c*, dependent variable *y*, independent variables *x* and *z*, and error term *u*. We wish to know the effect of *x* itself upon *y* (that is, we wish to obtain an estimate of *b*). But suppose that we omit *z* from the regression, and suppose the relation between *x* and *z* is given by

$$z = d + fx + e$$

with parameters *d, f* and error term *e*. Substituting the second equation into the first gives

$$y = (a + cd) + (b + cf)x + (u + ce).$$

If a regression of *y* is conducted upon *x* only, this last equation is what is estimated, and the regression coefficient on *x* is actually an estimate of (*b*+*cf*), giving not simply an estimate of the desired direct effect of *x* upon *y* (which is *b*), but rather of its sum with the indirect effect (the effect *f* of *x* on *z* times the effect *c* of *z* on *y*). Thus by omitting the variable *z* from the regression, we have estimated the total derivative of *y* with respect to *x* rather than its partial derivative with respect to *x*. These differ if both *c* and *f* are non-zero.

Detailed analysis

As an example, consider a linear model of the form

$$y_i = x_i\beta + z_i\delta + u_i, \qquad i = 1, \ldots, n$$

where

- x_i is a $1 \times p$ row vector of values of p independent variables observed at time i or for the i^{th} study participant;
- β is a $p \times 1$ column vector of unobservable parameters (the response coefficients of the dependent variable to each of the p independent variables in x_i) to be estimated;
- z_i is a scalar and is the value of another independent variable that is observed at time i or for the i^{th} study participant;
- δ is a scalar and is an unobservable parameter (the response coefficient of the dependent variable to z_i) to be estimated;
- u_i is the unobservable error term occurring at time i or for the i^{th} study participant; it is an unobserved realization of a random variable having expected value 0 (conditionally on x_i and z_i);
- y_i is the observation of the dependent variable at time i or for the i^{th} study participant.

We collect the observations of all variables subscripted $i = 1, \ldots, n$, and stack them one below another, to obtain the matrix X and the vectors Y, Z, and U:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times p},$$

and

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}, \quad U = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \in \mathbb{R}^{n \times 1}.$$

If the independent variable z is omitted from the regression, then the estimated values of the response parameters of the other independent variables will be given by, by the usual least squares calculation,

$$\hat{\beta} = (X'X)^{-1}X'Y$$

(where the "prime" notation means the transpose of a matrix and the -1 superscript is matrix inversion).

Substituting for Y based on the assumed linear model,

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + Z\delta + U) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U \\ &= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'U. \end{aligned}$$

On taking expectations, the contribution of the final term is zero; this follows from the assumption that U is uncorrelated with the regressors X. On simplifying the remaining terms:

$$\begin{aligned} E[\hat{\beta}|X] &= \beta + (X'X)^{-1}X'Z\delta \\ &= \beta + \textbf{bias}. \end{aligned}$$

The second term after the equal sign is the omitted-variable bias in this case, which is non-zero if the omitted variable z is correlated with any of the included variables in the matrix X (that is, if $X'Z$ does not equal a vector of zeroes). Note that the bias is equal to the weighted portion of z_i which is "explained" by x_i.

Effects on ordinary least squares

The Gauss–Markov theorem states that regression models which fulfill the classical linear regression model assumptions provide the best, linear and unbiased estimators. With respect to ordinary least squares, the relevant assumption of the classical linear regression model is that the error term is uncorrelated with the regressors.

The presence of omitted-variable bias violates this particular assumption. The violation causes the OLS estimator to be biased and inconsistent. The direction of the bias depends on the estimators as well as the covariance between the regressors and the omitted variables. A positive covariance of the omitted variable with both a regressor and the dependent variable will lead the OLS estimate of the included regressor's coefficient to be greater than the true value of that coefficient.

This effect can be seen by taking the expectation of the parameter, as shown in the previous section.

See also

- Confounding variable

References

- Barreto; Howland (2006). "Omitted Variable Bias". *Introductory Econometrics: Using Monte Carlo Simulation with Microsoft Excel*. Cambridge University Press.
- Clarke, Kevin A. (2005). "The Phantom Menace: Omitted Variable Bias in Econometric Research". *Conflict Management and Peace Science*. **22**: 341–352. doi:10.1080/07388940500339183.
- Greene, W. H. (1993). *Econometric Analysis* (2nd ed.). Macmillan. pp. 245–246.
- Wooldridge, Jeffrey M. (2009). "Omitted Variable Bias: The Simple Case". *Introductory Econometrics: A Modern Approach*. Mason, OH: Cengage Learning. pp. 89–93. ISBN 9780324660548.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Omitted-variable_bias&oldid=752379073"

Categories: Regression analysis | Bias

- This page was last modified on 30 November 2016, at 23:01.
- Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.