Basu, Deepankar

**Working Paper**

# When can we determine the direction of omitted variable bias of OLS estimators?

Working Paper, No. 2018-16

**Provided in Cooperation with:**
Department of Economics, University of Massachusetts

*Suggested Citation:* Basu, Deepankar (2018) : When can we determine the direction of omitted variable bias of OLS estimators?, Working Paper, No. 2018-16, University of Massachusetts, Department of Economics, Amherst, MA

This Version is available at:
https://hdl.handle.net/10419/202949

# DEPARTMENT OF ECONOMICS

## Working Paper

# When Can We Determine the Direction of Omitted Variable Bias of OLS Estimators?

Deepankar Basu

Working Paper 2018-16


## UNIVERSITY OF MASSACHUSETTS AMHERST

# When Can We Determine the Direction of Omitted Variable Bias of OLS Estimators?

Deepankar Basu*

November 1, 2018

**Abstract**

Omitted variable bias (OVB) of OLS estimators is a serious and ubiquitous problem in social science research. Often researchers use the direction of the bias in substantive arguments or to motivate estimation methods to deal with the bias. This paper offers a geometric interpretation of OVB that highlights the difficulty in ascertaining its sign in any realistic setting and cautions against the use of direction-of-bias arguments. This analysis has implications for comparison of OLS and IV estimators too.

JEL Codes: C20

Keywords: omitted variable bias; ordinary least squares.

*Department of Economics, University of Massachusetts, 310 Crotty Hall, Amherst, MA 01003, email: dbasu@econs.umass.edu.

# 1 Introduction

It is common for researchers in the social sciences to be confronted with situations where unobservability of variables or unavailability of data force them to omit such variables from regression models. Omitting relevant variables from the econometric model leads to asymptotic omitted variable bias (OVB) in the ordinary least squares (OLS) estimators of parameters appearing in the population regression function. This is a serious and ubiquitous problem and has been discussed widely in the applied econometrics literature.

In discussing the problem of OVB, and of strategies to deal with it, researchers have frequently relied on arguments about the *direction* of the bias. Let us look at some examples of the use of direction-of-bias arguments in papers published over the last few decades.[1]

- "One of the longest-running debates in empirical labor economics regards bias in OLS estimates of the economic return to schooling. The overriding concern pertains to individual-specific productivity components not reflected in the usual human-capital measures, as these ability components may be positively correlated with both wages and schooling. If the return to schooling is estimated with no account taken of the role of ability, the estimate is *generally expected to be biased upward.* (Blackburn and Neumark, 1995, pp. 217, emphasis added).

- "Equation (7) generalizes the conventional analysis of ability bias in the relationship between schooling and earnings. Suppose that there is no heterogeneity in the marginal benefits of schooling (i.e., $b_i = \bar{b}$) and that log earnings are linear in schooling (i.e. $k_1 = 0$). Then (7) implies that

$$\operatorname{plim} b_{ols} - \bar{b} = \lambda_0$$

[1] This list of examples is purely for the purpose of illustration and does not pretend to completeness.

which is the standard expression for the asymptotic bias in the estimated return to schooling that arises by applying the omitted variables formula to an earnings model with a constant schooling coefficient $\bar{b}$. According to the model presented here, this bias arises through the correlation between the ability component $a_i$ and the marginal cost of schooling $r_i$. If marginal costs are lower for people who would tend to earn more at any level of schooling, then $\sigma_{ra} < 0$, implying that $\lambda_0 > 0$." (Card, 2001, pp. 1134).

- "Ordinary least-squares (OLS) estimates of the proportionate increase in wages due to an extra year of education in the United States (the Mincerian rate of return) are believed to be reasonably consistent. It appears that *upward bias* due to omitted variables is roughly offset by attenuation bias due to errors in the measurement of schooling. Orley Ashenfelter and Cecilia Rouse (1998) find a net *upward bias* on the order of just 10 percent of the magnitude of the OLS estimate. David Card's (2001) survey of instrumental variables-based estimates reaches a similar conclusion, as do Ashenfeiter et al. (1999)." (Hertz, 2003, pp. 1354, emphasis added).

- "Our IV results, together with the results on neighboring districts and the historical data, lead us to conclude that our OLS results are *not biased upward* due to omitted district characteristics." (Banerjee and Iyer, 2005, pp. 1206, emphasis added).

- "There are several possible threats to our strategy. One is that product demand shocks may be correlated across high-income countries. In this event, both our OLS and IV estimates may be contaminated by correlation between import growth and unobserved components of product demand, making the impact of trade exposure on labor-market outcomes *appear smaller than it truly is*." (Autor et al., 2013, pp. 2129, emphasis added).

The frequent use of direction-of-bias arguments is problematic because in any realistic

situation, it is difficult to rule out more than one omitted variable, and in such a scenario the direction of OVB cannot be ascertained unambiguously (other than on the basis of rather restrictive assumptions). This latter fact is also well-known.

> Omitted-variable bias could be equally problematic, although *it is impossible to predict the direction of this bias* in a multivariate context. (Forbes, 2000, pp. 870, emphasis added).

How then do we square this - the impossibility of predicting the direction of omitted variable bias in a multivariate context - with the numerous examples of papers that explicitly use arguments about the direction of OVB (a small list of which I have referred to above)? The next line of the above quotation provides one possible answer.

> If there are strong univariate correlations between an omitted variable, inequality [included regressor], and growth [the dependent variable], however, these relationships could outweigh any multivariate effects and generate a significant, predictable bias. (Forbes, 2000, pp. 870).

This is not very persuasive. While intuitive arguments about relationships between omitted variables, included regressors and the dependent variable can often offer insights into the *signs* of partial effects involving omitted variables, it is difficult to see how such informal arguments can also give information about their *relative magnitudes* (especially when these might be unobserved). And, without knowledge about the relative magnitudes of the various partial effects involved, it is not possible to "generate a significant, predictable bias" in a multivariate context.[2]

The above examples suggest that there is less than full clarity in the applied economics literature about the nature of OVB in a multivariate context. In this paper, I offer a simple

[2]It should also be noted that, in this context, univariate correlations are not relevant; rather we need to deal with partial effects.

geometric interpretation of the OVB that helps us think rigorously about the issue. This analysis suggests that it is best to avoid using arguments about the direction of OVB. This is because the direction of OVB is ambiguous in most realistic research scenarios - when there are bound to be many omitted variables.

The rest of the paper is organized as follows: in section 2, I derive an expression for the OVB in the OLS estimator in a general setting which shows that, in general, it is not possible to ascertain the direction of OVB; in section 3, I develop a geometric argument to think about the direction of the OVB; in section 4, I discuss two special cases, and in the following section, I conclude the paper. The proof of Proposition 1 is given in the appendix.

## 2 Omitted Variable Bias of OLS Estimators

To fix ideas, suppose the population regression model is given by

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + u \tag{1}$$

where $\boldsymbol{y}$ is $N \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ and $\mathbf{Z}$ denote $(N \times K)$ and $(N \times M)$ matrices, respectively, of regressors, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote $(K \times 1)$ and $(M \times 1)$ vectors of population regression coefficients, and $u$ is the $N \times 1$ vector of errors. I assume that the error term is orthogonal to the regressors, i.e.

$$\mathbb{E}\left(\boldsymbol{X'u}\right) = \mathbb{E}\left(\boldsymbol{Z'u}\right) = \mathbf{0}, \tag{2}$$

to ensure that the method of ordinary least squares (OLS) estimation gives consistent estimates of the true parameters in the population regression function.

Suppose a researcher is unable to include the set of regressors, $\mathbf{Z}$, in the regression, either because those variables are unobservable or because data on them are not available. Hence

the researcher estimates the following model

$$\boldsymbol{y} = \mathbf{X}'\tilde{\boldsymbol{\beta}} + v \tag{3}$$

by OLS. Let us call the OLS estimator of $\tilde{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}$ and note that it is likely to be biased and inconsistent for the true parameter vector, $\boldsymbol{\beta}$. This follows from the fact that (3) is a misspecified model because the set of regressors, $\mathbf{Z}$, that appears in the true model (1), has been omitted from the estimated model (3).

**Proposition 1.** *The asymptotic omitted variable bias (OVB) in* $\hat{\boldsymbol{\beta}}$ *is given by*

$$\operatorname{plim} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{\delta}\boldsymbol{\gamma} \tag{4}$$

*where the* $m$*-th column of the* $K \times M$ *matrix* $\boldsymbol{\delta}$ *is the coefficient vector in the linear projection of the* $m$*-th omitted variable on the full set of included regressors,* $\boldsymbol{X}$*, and* $\boldsymbol{\gamma}$ *denotes the* $(M \times 1)$ *vector of coefficients associated with the omitted variables in the population regression function in (1).*

Using the result in Proposition 1, we can see that the OVB bias of the OLS estimator for the coefficient on the $k$-th included regressor in (3) is given by

$$OVB_k = \operatorname{plim} \hat{\beta}_k - \beta_k = \gamma_1\delta_{1k} + \gamma_2\delta_{2k} + \cdots + \gamma_M\delta_{Mk} = \sum_{m=1}^{M} \gamma_m\delta_{mk} \tag{5}$$

where $\delta_{mk}$ is the $k$-th element of the coefficient vector $\boldsymbol{\delta}_m$ in (11), with $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, M$.

The expression in (4) and in (5) are both well-known (Angrist and Pischke, 2009, pp. 60–61). It shows that the OVB is the product of two types of effects summed over all the omitted variables: (a) the first is the marginal effect of the $m$-th omitted variable on the dependent

variable, $\boldsymbol{y}$, in the correctly specified model (1) in the population: $\gamma_m$; and (b) the second is the marginal effect of the $k$-th included regressor on the $m$-th omitted variable in a linear projection of the latter, $\boldsymbol{z}^m$, on the whole set of included regressors, $\boldsymbol{X}$ in the sample: $\delta_{mk}$. What is not emphasized is the following fact: since $OVB_k$ is the sum of $M$ terms, each of which can be of any sign, it is not possible in general to unambiguously ascertain its sign.

# 3 Direction of Bias: A Geometric Interpretation

Define the $1 \times M$ vector,

$$\boldsymbol{\delta^k} = [\delta_{1k} \quad \delta_{2k} \quad \cdots \quad \delta_{Mk}], \tag{6}$$

and note that this is the $k$-th row of the $\boldsymbol{\delta}$ is the $K \times M$ matrix. Hence, the $m$-th element of the $M \times 1$ vector $\boldsymbol{\delta^k}$ gives the coefficient on the $k$-th included regressor in a linear projection of the $m$-th omitted variable, $\boldsymbol{z}^m$, on the whole set of included regressors, $\boldsymbol{X}$. Hence, the vector $\boldsymbol{\delta^k}$ collects together the coefficient on the $k$-th included regressor in linear projections, successively, of the 1-st, 2-nd, $\cdots$, $M$-th omitted variable on the whole set of included regressors.

Since $\boldsymbol{\gamma}$ is a $M \times 1$ vector, the expression for the omitted variable bias in (5) is the inner product of the two vectors, $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$. Hence,

$$OVB_k = \boldsymbol{\delta^k}.\boldsymbol{\gamma} = \left\|\boldsymbol{\delta^k}\right\| \left\|\boldsymbol{\gamma}\right\| \cos(\theta) \tag{7}$$

where $\left\|\boldsymbol{x}\right\|$ denotes the $L_2$-norm of the vector, $\boldsymbol{x}$, $\theta$ is the angle - measured in radians - between $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, each considered as an element in $\mathbb{R}^M$, and $0 \leq \theta \leq \pi$.

**Definition 1.** *Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^M$ with $\theta$ denoting the angle between the two vectors defined by (7).*

1. *We will say that $\boldsymbol{a}$ and $\boldsymbol{b}$ are similar in orientation if the angle between them is acute,*

*i.e.*, $0 < \theta < \pi/2$.

2. *We will say that* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *are dissimilar in orientation if the angle between them is obtuse, i.e.* $\pi/2 < \theta < \pi$.

This definition is inspired by the notion of "cosine similarity" in the machine learning literature and can help us ascertain the direction of OVB.

**Proposition 2.** *The direction of omitted variable bias of the OLS estimator of the k-th included regressor in a misspecified model with many omitted variables is positive (negative) if the vectors* $\boldsymbol{\delta^k}$ *and* $\boldsymbol{\gamma}$ *are (dis)similar in orientation.*

*Proof.* The proof follows immediately from an inspection of the expression in (7). □

Figure 3 depicts the various possibilities related to the two vectors $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, and the direction of bias in a 2-dimensional setting. In this figure, we denote $\boldsymbol{\delta^k}$ by the solid (black) line and $\boldsymbol{\gamma}$ with the broken (red) lines. We start with a given value of $\boldsymbol{\delta^k}$, and then show the various configurations of $\boldsymbol{\gamma}$ that will lead to positive or negative bias.

In the right panel in Figure 3, we start with a given $\boldsymbol{\delta^k}$ (shown in solid black), and then draw a plane that is perpendicular to $\boldsymbol{\delta^k}$ (labeled AB). If the vector $\boldsymbol{\gamma}$ lies anywhere to the right (or on top) of the plane, the direction of bias will be positive (because the angle between the two vectors will be between 0 and $\pi/2$. For instance, two possible values of the $\boldsymbol{\gamma}$ vector are shown in broken (red) lines. If we move to the left panel in Figure 3, we see configurations when the bias will be negative. For a given value of $\boldsymbol{\delta^k}$ (shown in solid), the perpendicular plane is AB. Any value of $\boldsymbol{\gamma}$ which leads to the vector falling below the plane AB will give rise to a negative OVB (because the angle between the two vectors will be larger than $\pi/2$ but less than $\pi$).

This intuition carries over to $\mathbb{R}^M$. In this general case, AB will be the subspace of $\mathbb{R}^M$ that is perpendicular to the $M$-vector $\boldsymbol{\delta^k}$. When the $M$-vector $\boldsymbol{\gamma}$ lies on top of the subspace

AB, the OVB will be positive; when lies below AB, the OVB will be negative. While this general characterisation allows us to see the conditions that lead to positive or negative OVB, in the next subsection, I will convert this discussion into sign requirements on the elements of the two vectors. That will allow us to interpret the general geometric requirement into signs of coefficients that capture partial effects.

## 3.1 Unambiguous Sign of OVB

The result in Proposition 2 shows that in general we will not be able to ascertain the sign of the OVB. Nonetheless, there are special configurations, as noted in Proposition 2, where we *will be* able to make unambiguous sign statements.

### 3.1.1 Unambiguously Zero Bias

We will be able to assert that there is no bias if the vectors $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$ are orthogonal or if one of them is a null vector. The two vectors are orthogonal when the all omitted variables are orthogonal to all the included regressors, and hence leaving out the omitted variables does not induce any correlation between the error term and the included regressors. That is why OLS is able to consistently estimate all the parameters. On the other hand, if either of the vectors is a null vector, it means that either the omitted variables are irrelevant or that the included regressors have no partial effect on the omitted variables (in the relevant linear projection). That is why OLS is able to, once again, estimate the parameters consistently.

### 3.1.2 Unambiguously Positive Bias

We will be able to unambiguously determine the sign of the OVB to be positive if both the vectors $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$ lie in the same orthant. This is because, in this case, the two vectors will be similar in orientation according to Definition 1. If the two vectors lie in the same orthant, they will have the same sign for each corresponding element. In this case, we will be able to

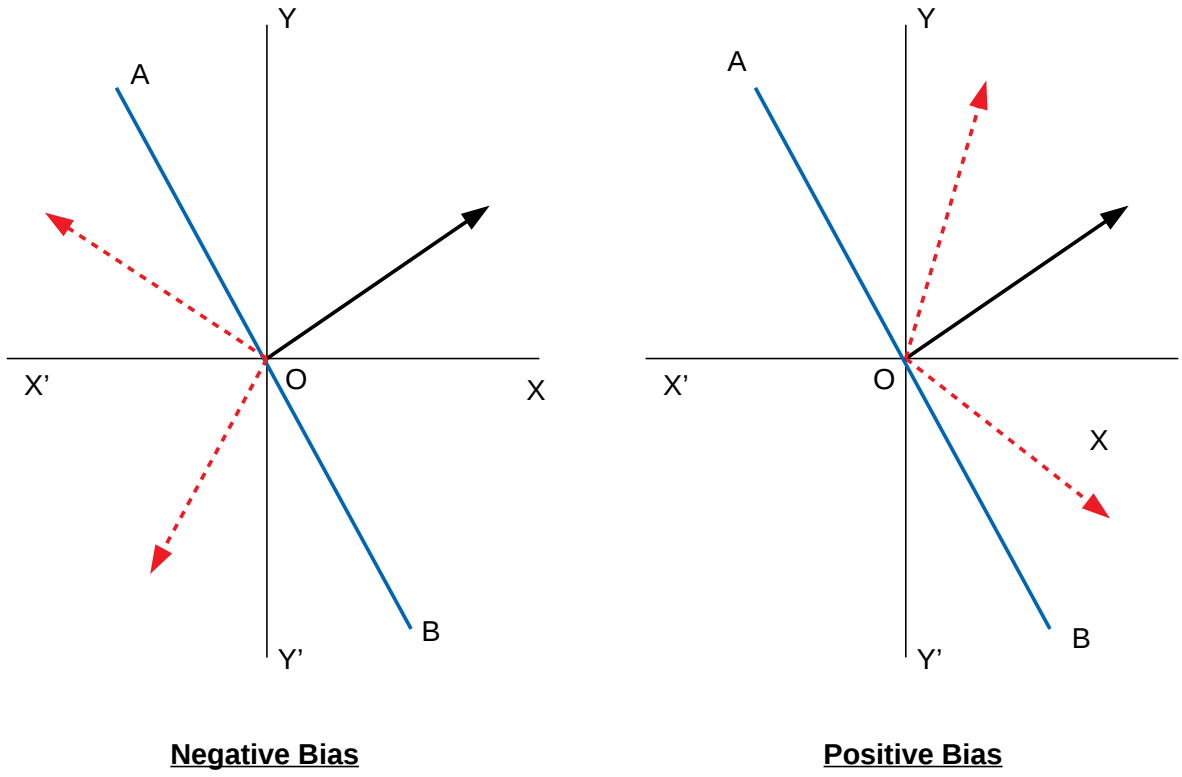


Figure 1: *Possible configuration of the vectors* $\boldsymbol{\delta^k}$ *(solid arrow) and* $\boldsymbol{\gamma}$ *(broken arrow), and the direction of bias. The vector* $\boldsymbol{\delta^k}$ *collects together the coefficients on the* $k$*-th included regressor in linear projections, successively, of the* 1*-st,* 2*-nd,* $\cdots$*,* $M$*-th omitted variable on the whole set of included regressors,* $\boldsymbol{X}$*. The vector* $\boldsymbol{\gamma}$ *is the vector of coefficients on the omitted variables in the population regression function. In the right panel, the two vectors are similar in orientation, which leads to positive asymptotic bias of the OLS estimators of the coefficients of* $\boldsymbol{X}$ *in the misspecified model (with omitted variables,* $\boldsymbol{Z}$*). In the left panel, the two vectors are dissimilar in orientation, so that the bias is negative.*

determine the sign of the OVB as positive irrespective of the magnitude of the elements of the two vectors. Translated into the meaning of the elements of the two vectors, $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, an unambiguously positive OVB will arise for the OLS estimate of $k$-th included regressor's coefficient in the misspecified model in (3) if the partial effect of each omitted variable on the dependent variable has the same sign as the partial effect of the $k$-th included regressor on that omitted variable (in a linear projection of the omitted variable on all the included regressors).

How likely is this scenario? To answer this question, let us abstract from the magnitudes of the elements of the two vectors, $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, and only consider their signs, which can be either positive or negative. Thus, let us consider two vectors of length $M$, whose elements belong to this two element set: $\{+, -\}$. The total number of possibilities of generating these two vectors is $2^M \times 2^M$. In geometric terms, generating these two vectors is exactly equivalent to choosing the orthant combination of two vectors in $M$ dimensional Euclidean space. Since there are $2^M$ orthants, when we choose *two* $M$-vectors, we can choose from $2^M \times 2^M$ orthant combinations.

This immediately gives us a way to identify cases when the two vectors will lie in the same orthant. That will happen only if the signs of *all* the elements in the two vectors are exactly the same. In this case, we can choose from one of the $2^M$ orthants - because both vectors must lie in the same orthant. Thus, if the elements of the two vectors were randomly assigned signs, the probability of having them lie in the same orthant - which generates an unambiguously positive OVB - would be $2^{-M} (= 2^M/2^{2M})$. Even for moderately large values of $M$, this probability is quite small. For instance, if $M = 5$, the probability is 0.03125, and if $M = 10$, this probability is 0.0001.

In most realistic research scenarios, we will have partial, rather than, zero information. Hence, we will often be able to convincingly argue about the sign of some of the omitted variables. This will reduce the severity of the problem. Suppose there are a total of $M$

omitted variables, and we are able to ascertain signs of $M' \leq M$ partial effects (of omitted variables on the dependent variable and of included regressors on the omitted variables), then the dimensionality of the problem reduces to $M - M'$. Now we need to choose the orthant combination of two vectors in $M - M'$ dimensional Euclidean space. Using the same argument as above, we can see that the probability of correctly guessing a positive OVB is, in this case, $2^{M'-M}$. For instance, if out of 10 omitted variables, we are able to sign the partial effects of 2 of them, then the probability of guessing the OVB correctly as being positive is 0.004.

### 3.1.3 Unambiguously Negative Bias

We will be able to unambiguously determine the sign of the OVB to be negative if the two vectors, $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, lie in "opposite" orthants, by which I mean that the sign of each element in the first vector is exactly opposite of the sign of the corresponding element in the second vector. This is because, in this case, the two vectors will be dissimilar in orientation, according to Definition 1. To see this, note that the inner product of the two vectors in this case will result in a negative scalar because each of the terms in the inner product is negative. Hence, the angle between the two vectors will be between $\pi/2$ and $\pi$.

How likely is this scenario? To answer this question, we can use the same device we used to determine cases of unambiguously positive bias. Two vectors will lie in "opposite" orthants if the signs of *all* the elements are exactly opposite in the two vectors. We can choose one of the $2^M$ orthants for the first vector, and then flip the sign of each element of the vector to get the orthant for the second vector. The first can be done in $2^M$ ways, and the second in 1 way, giving us a total of $2^M$ combinations of such vectors. Thus, if the elements of the two vectors were randomly assigned signs, the probability of having a negative OVB would be $2^{-M} (= 2^M/2^{2M})$. Interestingly, this is the same magnitude as the probability of unambiguous positive OVB. Moreover, if we are able to correctly assign signs

for $M' \leq M$ partial effects, then the probability becomes $2^{M'-M}$, as in the previous case.

#### 3.1.4 Unambiguously Positive or Negative Bias

Bringing discussion of the two cases together, we can see that, if the elements of the two vectors, $\boldsymbol{\delta^k}$ and $\boldsymbol{\gamma}$, were assigned signs randomly, then the probability of being able to make an unambiguous assertion about positive or negative bias would be $2^{-M+1}(= 2^{-M} + 2^{-M})$. This is a rather small probability. Hence, when we do not have a firm basis for determining the sign of the partial effects of omitted variables on the dependent variable or of the partial effects of included regressors on omitted variables (in the relevant linear projections) or both, we would be able to make correct judgments about the direction of OVB with extremely small probabilities by making a random guess. For instance if there were 10 omitted variables, this probability would be 0.00195.

## 4 Special Cases Discussed in the Literature

There are two special cases of the general result in (5) that are often discussed in the literature.

### 4.1 One Omitted Variable, Many Included Variables

If the researcher is able to make a convincing argument that there is only one omitted variable, then the vector of omitted variables, $\boldsymbol{Z}$, reduces to a scalar, $Z$, in (1). Using (5), in this case, the OVB for the $k$-th included regressor becomes

$$OVB_k = \gamma_1 \delta_{1k} \tag{8}$$

where $\gamma_1$ is the marginal effect of the single omitted variable on the dependent variable in the true model, and $\delta_{1k}$ is the coefficient on the $k$-th element of $\boldsymbol{X}$ in the linear projection of the single omitted variable on the full set of included regressors, $\boldsymbol{X}$. This has often been used in the applied economics literature -for instance, in the applied labour economics literature - and has filtered down into textbook treatments of the OVB (Wooldridge, 2002, pp. 61-63.) In this case, we will be able to ascertain the direction of bias in an unambiguous manner just by knowing the signs of $\gamma_1$ and $\delta_{1k}$. We will not need to know the magnitude of the coefficients to make any statements about the direction of OVB.

The canonical case is a wage regression where the included variable under consideration is years of schooling and the only omitted variable is "ability". Since ability is likely to be positively correlated with log-wage (the dependent variable) and years of schooling, we might be able to make the case that the direction of the OVB is positive.

## 4.2 One Included Variable, Many Omitted Variables

In many textbook treatments, the OVB is motivated with examples where there is only one included variable but many omitted variables (Angrist and Pischke, 2009, pp. 60). Using (5), in this case, the OVB for the only included variable is given by

$$OVB_1 = (\gamma_1\delta_{11} + \gamma_2\delta_{21} + \cdots \gamma_M\delta_{M1})\,. \tag{9}$$

Consider the wage regression again, but now only with years of schooling as the included regressor. If ability, motivation, neighbourhood characteristics, family income, and other such variables are omitted from the model, then we are within the purview of this special case. Note that unless we make the strong assumption that we can replace all the omitted variables with a composite variable called "ability", we will be facing an expression for the OVB as given in (9). Thus, in this case too, we will not be able to ascertain the direction

of the OVB (because it is the sum of $M$ terms, each of which can be positive, negative or zero) other than by making claims about the *relative magnitudes* of the various parameters appearing in (9). This latter option is untenable because in most cases there is scant basis for making judgments about relative magnitudes of partial effects involving omitted (often unobservable) variables. Moreover, this case is qualitatively different from the previous special case where we could determine the direction of bias without any knowledge of the magnitude of the coefficients.

# 5 Comparison of OLS and IV Estimators

One common strategy to deal with the bias caused by omitted variables is to use instrumental variables estimators. In such a context, it is standard in the literature to make comparisons of the direction and magnitude of bias of OLS and IV estimators (Angrist and Krueger, 2001, pp. 79). The above analysis suggests that such comparisons can be difficult to pin down.

Consider a scenario that fits with the special case discussed above: one included regressor and many omitted variables. Let the dependent variable and the included (endogenous) regressor be denoted as $y$ and $x$, respectively, and suppose we have an instrumental variable, $z$ for $x$. Thus, the model is

$$y = \beta x + u$$

where $\mathbb{E}(xu) \neq 0$ because the error term contains many omitted variables. In this case, the instrumental variables estimator of the coefficient on $x$ is given by

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}$$

so that

$$\operatorname{plim} \hat{\beta}_{IV} - \beta = \frac{\operatorname{Cov}(z, u)}{\operatorname{Cov}(z, x)}. \tag{10}$$

On the other hand, the asymptotic bias of the OLS estimator is given by the expression in (9):

$$\text{plim}\,\hat{\beta}_{OLS} - \beta \equiv OVB_1 = (\gamma_1\delta_{11} + \gamma_2\delta_{21} + \cdots \gamma_M\delta_{M1})\,.$$

If $\text{Cov}\,(z,u) = 0$, the instrument is exogenous and the IV estimator will be consistent. In such a case, IV estimation is clearly superior to OLS because the latter will give asymptotically biased estimates. But if $\text{Cov}\,(z,u) \neq 0$, the instrument is not exogenous. Hence, the IV estimator will be biased, as can be seen from the expression in (10). In such situations, the question that is often of interest is a possible comparison of the OLS bias and the IV bias - both its direction and its magnitude. Is it possible to do such comparisons? The answer seems to be in the negative because, as argued in this paper, it is difficult to ascertain - other than in special cases - the magnitude and sign of the OVB of the OLS estimator.

Even if $\text{Cov}\,(z,u)$ is small, so that violation of instrument exogeneity is not very serious, the bias in the IV estimator can be large if $\text{Cov}\,(z,x)$ is small. This is the *weak instrument* problem and has been discussed extensively in the past several decades (Andrews et al., Working Paper). But the same problem of comparison of the bias of the IV and the OLS estimator remains. If neither the magnitude nor the sign of OLS bias can be determined, then it is not clear how one would compare it with the possibly large bias of the IV estimator with weak instruments? It is undeniable that the use of weak instruments can lead to large asymptotic bias. What is less clear is whether we can compare the sign and magnitude of that bias with the bias of the OLS estimator in the presence of many omitted variables - which presumably led to the use of instrumental variables in the first place.[3]

[3]The analysis in this section can be easily extended to the case of many instrumental variables and the 2SLS estimator.

# 6 Conclusion

In the social sciences, researchers are often confronted with bias and inconsistency in OLS estimators of parameters of interest due to omitted variables. In such situations, if the use of methods to deal with the omitted variable problem is not feasible, researchers often choose to deal with the situation with a direction-of-bias argument. The direction of bias argument is used in other cases too - possibly to motivate the use of instrumental variables estimation or related methods. In either case, researchers have to confront the problem highlighted in this paper: when there are many omitted regressors, it is not possible, in general, to ascertain the sign of the OVB of OLS estimators.

In this paper I have identified some special cases where we *will be* able to unambiguously determine the sign of the OVB using knowledge the signs of relevant partial effects only (and being ignorant about their magnitudes). These cases, discussed in section 3.1, are multivariate generalizations of the one-dimensional case that is frequently discussed in the literature: one omitted variable and one included regressor. In this latter case we are able to determine the direction of bias as soon as we know the signs of the two partial effects. In a similar way, for a case with $M$ omitted variables, we will be able to unambiguously determine the sign of the OVB as positive if the partial effects of omitted variables on the dependent variable are of the same sign as the partial effect of the regressors on the omitted variables, and as negative if the partial effects of omitted variables on the dependent variable are of exactly the opposite sign as the partial effect of the regressors on the omitted variables. In all other cases, we will not be able to unambiguously determine the sign of the OVB.

Much of the extant literature in applied economics, including standard textbook treatments, seems to have ignored the possible ambiguity of the sign of the OVB by considering the case of a single omitted variable. As soon we move beyond this simplified setup and allow for more than one omitted variables, it is no longer possible to unambiguously determine the

sign of the OVB of OLS estimators (other than in the two special cases discussed above). Attempts to club together multiple omitted variables into a composite category, for instance, as is often done in discussions of wage regressions, where "ability" stands for many omitted factors (like ability, motivation, family background), are bound to be misleading. The direction of bias conclusions used by researchers in substantive arguments using such composite omitted variables cannot be sustained in most realistic settings if we allow for many omitted variables. The upshot of the analysis presented in this paper is that researchers should not take recourse to direction of OVB arguments - even to motivate the use of methods to deal with OVB, like instrumental variables. Moreover, comparisons of the magnitude and direction of bias of OLS versus IV estimators in situations with many omitted variables might also be misleading.

# References


Isaiah Andrews, James H. Stock, and Liyang Sun. Weak instruments in IV regression: Theory and practice. Working Paper. Accessed on November 1, 2018 from here.

Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, New York, 2009.

David H. Autor, David Dorn, , and Gordon H. Hanson. The china syndrome: Local labor market effects of import competition in the united states. *American Economic Review*, 103(6):2121–2168, 2013.

Abhijit Banerjee and Lakshmi Iyer. History, institutions, and economic performance: The

legacy of colonial land tenure systems in India. *American Economic Review*, 95(4):1190–1213, 2005.

McKinley L. Blackburn and David Neumark. Are ols estimates of the return to schooling biased downward? another look. *The Review of Economics and Statistics*, 77(2):217–230, 1995.

David Card. Estimating the return to schooling; progress on some persistent econometric problems. *Econometrica*, 69(5):1121–1160, 2001.

Kristin J. Forbes. A reassessment of the relationship between inequality and growth. *American Economic Review*, 90(4):170–192, 2000.

Tom Hertz. Upward bias in the estimated returns to education: Evidence from south africa. *American Economic Review*, 93(4):1121–1160, 2003.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2002.

# Appendix

This is a proof of Proposition 1.

*Proof.* Note that

$$\operatorname{plim}\hat{\boldsymbol{\beta}} = \operatorname{plim}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\right) = \boldsymbol{\beta} + \left[\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)\right]^{-1}\left[\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\right]\boldsymbol{\gamma},$$

where we have plugged in the expression for $\boldsymbol{y}$ from the true model in (1), and the last step follows from the orthogonality of the error term given in (2) and using a suitable law of large numbers (along with the Slutsky theorem) to replace $\operatorname{plim}\left(\boldsymbol{X}'\boldsymbol{X}\right)$ with $\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{X}\right)$, and to replace $\operatorname{plim}\left(\boldsymbol{X}'\boldsymbol{Z}\right)$ with $\left(\boldsymbol{X}'\boldsymbol{Z}\right)$.

Note, using the algebra of partitioned matrices, that

$$\boldsymbol{X}'\boldsymbol{Z} = \boldsymbol{X}'\begin{bmatrix}\boldsymbol{z}^1 & \boldsymbol{z}^2 & \cdots & \boldsymbol{z}^M\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'\boldsymbol{z}^1 & \boldsymbol{X}'\boldsymbol{z}^2 & \cdots & \boldsymbol{X}'\boldsymbol{z}^M\end{bmatrix}$$

where $\boldsymbol{z}^m$ refers to the $N\times 1$ vector representing the $m$-th column of $\boldsymbol{Z}$, with $m = 1, 2, \ldots, M$. Hence

$$\begin{aligned}
\operatorname{plim}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left[\mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{\gamma} \\
&= \left[\left[\mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{z}^1\right) \quad \cdots \quad \left[\mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{z}^M\right)\right]\boldsymbol{\gamma} \\
&= \left[\boldsymbol{\delta}_1 \quad \cdots \quad \boldsymbol{\delta}_M\right]\boldsymbol{\gamma} \\
&= \boldsymbol{\delta}\boldsymbol{\gamma}
\end{aligned}$$

where, for $m = 1, 2, \ldots, M$, $\boldsymbol{\delta}_m$ is the coefficient vector in the linear projection of the $m$-th omitted variable on the whole set of included regressors, i.e.

$$\boldsymbol{z}^m = \boldsymbol{X}\boldsymbol{\delta}_m + v_m, \tag{11}$$

with $\mathbb{E}\left(\boldsymbol{X}' v_m\right) = \mathbf{0}$, so that

$$\boldsymbol{\delta}_m = \left[\mathbb{E}\left(\mathbf{X}'\mathbf{X}\right)\right]^{-1} \mathbb{E}\left(\boldsymbol{X}'\boldsymbol{z}^m\right).$$

Columnwise stacking of $\boldsymbol{\delta}_m$, then gives the $K \times M$ matrix $\boldsymbol{\delta}$ and completes the proof. $\square$