Why OLS is not always Blue*

Carl-Johan Dalgaard
Institute of Economics
University of Copenhagen

February 9, 2005

Abstract

The OLS estimator is not always the Best Linear Unbiased Estimator. This note, aimed at readers without prior knowledge of the issues at hand, explains why the OLS estimate becomes "biased" when relevant regressors are either (i) measured with error, (ii) endogenous or, (iii) omitted from the equation being estimated.

1 A Few Definitions

Before we begin we need to have a few definitions in place. First, the mean of a variable, x, is denoted $\bar{x}$:

$$\bar{x} = \frac{1}{N}\sum_{i}^{N} x_i$$

where N is the number of observations and i denotes the individual observation (each country if you like). Second, the estimated variance of the same variable is calculated as

$$var\left(x_i\right) = \frac{1}{N}\sum \left(x_i - \bar{x}\right)^2.$$

Third the estimated covariance between x and y is calculated as

$$cov\left(x_i, y_i\right) = \frac{1}{N}\sum_{i}^{N} \left(y - \bar{y}\right)\left(x - \bar{x}\right).$$

*Lecture notes for Economic Growth, Spring 2005.

2 The OLS estimator

Consider the following regression model, relating growth in country i, to human capital in country i, h_i

$$g_i = \alpha_0 + \alpha_1 h_i + u_i, \tag{1}$$

where α_0 and α_1 are the structural parameters we wish to estimate. u_i is an error term, which is assumed to fulfill: $E(u_i) = 0$ and $var(u_i) = \sigma^2$; where $E(.)$ is the expectation operator.

When we proceed to estimate (1) we will a priori be strong believers in it being the "right" model. Hence, equation (1) will be referred to as "the true" model. Now, "the true model" may, upon closer reflection, change as we proceed. But for now, that's what the world looks like.

OLS estimation involves choosing α_0 and α_1 such that the squared residuals are minimized. So the problem is to choose

$$\{\alpha_0, \alpha_1\} = \arg\min \sum^{N} (g_i - \alpha_0 - \alpha_1 h_i)^2 .$$

Solving the two first order conditions for α_0 and α_1 yields the OLS estimators for α_0 and α_1, i.e. $\hat{\alpha}_0$ and $\hat{\alpha}_1$, respectively

$$\hat{\alpha}_0 = \bar{g} - \hat{\alpha}_1 \bar{h} \tag{2}$$

$$\hat{\alpha}_1 = \frac{cov(g_i, h_i)}{var(h_i)} = \frac{\sum^{N} (g_i - \bar{g})\left(h_i - \bar{h}\right)}{\sum^{N} \left(h_i - \bar{h}\right)^2}. \tag{3}$$

In what follows we'll focus attention on $\hat{\alpha}_1$, since its typically the case that we're preoccupied with assessing the possible causal effect of control variables, like human capital, on growth. In the present case, α_1 reflect the "true" causal effect of human capital on growth. Does OLS allow us to identify this important parameter?

To answer this question, its useful to note that the numerator of equation

(3) can be rewritten, using equation (1), to yield

$$\begin{aligned} &= \sum^{N} \left(\alpha_0 + \alpha_1 h_i + u_i - \left(\alpha_0 + \alpha_1 \bar{h}\right)\right)\left(h_i - \bar{h}\right) \\ &= \sum^{N} \left(\alpha_1 \left(h_i - \bar{h}\right) + u_i\right)\left(h_i - \bar{h}\right) \\ &= \alpha_1 \sum^{N} \left(h_i - \bar{h}\right)^2 + \sum^{N} u_i \left(h_i - \bar{h}\right). \end{aligned}$$

Next, substituting this equation back into equation (3)

$$\begin{aligned} \hat{\alpha}_1 &= \frac{cov\left(g_i, h_i\right)}{var\left(h_i\right)} = \frac{\alpha_1 \sum^{N} \left(h_i - \bar{h}\right)^2 + \sum^{N} u_i \left(h_i - \bar{h}\right)}{\sum^{N} \left(h_i - \bar{h}\right)^2} \\ \hat{\alpha}_1 &= \frac{cov\left(g_i, h_i\right)}{var\left(h_i\right)} = \alpha_1 + \frac{\sum^{N} u_i \left(h_i - \bar{h}\right)}{\sum^{N} \left(h_i - \bar{h}\right)^2}, \end{aligned}$$

which could also be stated, using the definitions from Section 1

$$\hat{\alpha}_1 = \frac{cov\left(g_i, h_i\right)}{var\left(h_i\right)} = \alpha_1 + \frac{cov\left(u_i, h_i\right)}{var\left(h_i\right)}. \tag{4}$$

Accordingly, OLS is dead on target if $cov\left(u_i, h_i\right) = 0$. But when is this assumption violated?

2.1 Measurement Error

We assumed above – based on theory – that human capital affects growth. The problem is that we cannot measure "human capital" directly. Thus, in order to proceed we need some other variable which can proxy for what we really have in mind. A candidate is (average) years of schooling in the population, e, since - conceivably - formal training builds up skills and thereby human capital. But e may not capture the theoretical concept of human capital fully. Accordingly, when we use e as a measure of human capital we are probably making an error in measuring human capital. To formalize this, let

$$e_i = h_i + v_i, \tag{5}$$

where v_i is white noise with variance σ_v^2. Observe that we maintain the assumption that e is a reasonable proxy for h, in the sense that it "gets it right" on average: $\bar{e} = \bar{h}$.

Now suppose we substitute equation (5) into our true model, equation (1) :

$$g_i = \alpha_0 + \alpha_1 (e_i - v_i) + u_i \equiv \alpha_0 + \alpha_1 e_i + \eta_i \tag{6}$$

where the new error term is

$$\eta_i \equiv u_i - \alpha_1 v_i.$$

The key thing to notice is that η_i depends on the measurement error v_i. As a consequence, if we estimate equation (6) we get (invoking formula (4)):

$$\begin{aligned} \hat{\alpha}_1 &= \alpha_1 + \frac{cov(\eta_i, e_i)}{var(e_i)} \\ &= \alpha_1 + \frac{cov(u_i - \alpha_1 v_i, e_i)}{var(h_i) + \sigma_v^2}, \end{aligned}$$

where we have calculated $var(e_i)$ using equation (5).[1] To push this issue a little further we could proceed to calculate $cov(u_i - \alpha_1 v_i, e_i)$. But at this point it is easy to make out the sign of the covariance. First, suppose human capital spur growth, so that $\alpha_1 > 0$. Now, if v_i is positive, $u_i - \alpha_1 v_i$ is "small", while e_i becomes "large" (Cf. equation (5)). So the covariance $cov(u_i - \alpha_1 v_i, e_i)$ is negative. As a result: The OLS estimate becomes biased towards zero, since $\hat{\alpha}_1 < \alpha_1$.[2]

Is it always the case that measurement error imply that the OLS estimate is biased towards zero? This is a question of some practical importance. Suppose you are going over an empirical study. You quickly realize that measurement error might be an issue in the case at hand. But the (OLS) estimation results reveal that the parameters of interest are significantly different from zero. If indeed measurement error always leads to a downward

[1] In general, if we have the equation $y_i = a_i + bx_i$ it holds that $var(y_i) = var(a_i) + b^2 var(x_i)$. The present case is the simpler one where $b = 1$.

[2] Notice that our assumption of $\alpha_1 > 0$ is completely unimportant for this conclusion. If $\alpha_1 < 0$, the covariance will be positive, and the estimate is still biased towards zero.

bias, you might feel inclined to conclude that the (true) causal effect is likely to be even stronger than what the estimate reveal and certainly significant. This can, however, be the wrong conclusion.

The concept of measurement error was illustrated above by the case where a proxy variable is adopted. That is, a scenario where we cannot directly measure the variable our theory tells us is relevant. In general, however, measurement error arises as soon as data is of poor quality. In the interest of clarity, imagine that e_i isn't a proxy variable at all. In stead, suppose our theory tells us that e_i is exactly the right hand side variable to use. Consequently, we'll entertain the idea that the "true model" is:

$$g_i = \alpha_0 + \alpha_1 e_i + u_i,$$

where u_i is white noise. Even so, we might still be in "hot water" since the data on years of schooling could be flawed, plain and simple. Perhaps because the statistical agencies do not have the necessary micro data (ideally, information on schooling for all citizens). Consequently, denote by e_i the true observation of years of schooling, while $\hat{e}_i$ is the data the statistician hand us. The difference between the two is a pure measurement error, v_i (again, white noise)

$$\hat{e}_i = e_i + v_i.$$

Now, suppose we go no further in making assumptions. If we go through the exact same steps as above, in assessing the bias of the OLS estimate, we would reach exactly the same conclusion: the OLS estimate will be biased towards zero. (If this is not obvious, you should take a few minutes to convince yourself that this statement is correct, before you proceed.)

So let's try to make matters a bit "worse" in order to show how a new conclusion may emerge. Specifically, suppose there is some underlying variable, z_i, with the following properties

(A1) Wherever z_i is "large", the measurement error, v_i, is "small": $v_i'(z_i) < 0$.

(A2) Wherever z_i is "large", the level of schooling, e_i, is "large": $e_i'(z_i) > 0$.

(A3) z_i is completely independent of the growth rate g_i. That is, z_i is neither a determinant of growth, g_i, nor is it affected by income growth.

If (A3) is violated we will be faced with other problems aside from the measurement error – omitted variables and endogenous regressors, respectively. We turn to such issues momentarily. But for now we would like to side-step such complications.

The final assumption we make is that

$$\frac{\partial \hat{e}}{\partial z_i} = e_i'(z) + v_i'(z) > 0, \tag{A4}$$

i.e. $e_i'(z)$ is larger in absolute value than $v_i'(z)$.

Our OLS estimate of α_1 is

$$\hat{\alpha}_1 = \alpha_1 + \frac{cov\left(u_i - \alpha_1 v_i(z_i), \hat{e}_i\right)}{var\left(e_i\right)}.$$

Under A1-A4; if we increase z_i (i.e. move from a country with a low value for z to another country with a high value) we will increase both $u_i - \alpha_1 v_i(z_i)$ and $\hat{e}_i$. Consequently the estimate will be biased *upward,* rather than towards zero.

Hence if measurement error is not entirely random, the OLS estimate may in fact be biased away from zero and thus overestimates the effect from – in this case – years of schooling on growth. The case where the measurement error is random is often referred to as "classical" measurement error.

2.2 Endogenous Regressors

Suppose we are able to wink the measurement error problem away. That is, suppose e is the correct variable to use (theoretically), and that we can measure it perfectly.

In stead, consider the possibility that people's desire to attend school is affected by how fast the economy is growing. Specifically, imagine that we have the following **true system**

$$g_i = \alpha_0 + \alpha_1 e_i + u_i, \tag{7}$$

$$e_i = \beta_0 + \beta_1 g_i + v_i. \tag{8}$$

Both u_i and v_i are white noise; in addition we'll assume (to focus on the main point) that they are completely independent of one another.

If we simply press the OLS bottom, estimating equation (7), while ignoring equation (8) entirely, we get the following

$$\hat{\alpha}_1 = \alpha_1 + \frac{cov\left(u_i, e_i\right)}{var\left(e_i\right)}, \tag{9}$$

our "standard result". But since e_i is in fact endogenous (depends on g_i), the last term is non-zero. To see this, solve the system (7)-(8) for e_i and you obtain

$$e_i = \frac{\beta_0 + \beta_1\alpha_0 + \beta_1 u_i + v_i}{1 - \alpha_1\beta_1}. \tag{10}$$

Clearly, e_i is not independent of u_i. This is because of the feed back loop from growth to education. If u_i rises (for some reason) it will imply an increase in g_i which will work to raise e_i through equation (8).

Gaining some geometrical intuition for this result might be illuminating. Consider Figure 1, where the system (7)-(8) is illustrated in a (g, e) diagram. Accordingly, $g\left(e\right)$ represents equation (7), while $e\left(g\right)$ represents equation (8). For this illustration we assume that the slope of $g\left(e\right)$ is greater than the slope of $e\left(g\right)$; $\alpha_1 > 1/\beta_1$. Now, we can think of the picture as representing the "equilibrium" outcome in a country in the sample, thus reflecting a specific draw of the noise terms; $u_i = v_i = 0$, say. This gives us 1 data point: $\left(g^*, e^*\right)$. Next imagine you were to draw another picture, for another country. The difference would be that the values for u, v are different. As a result, the equilibrium would lie somewhere else in the diagram. For example, if $u_i > 0$ while $v_i < 0$ then it will be situated south-east of $\left(g^*, e^*\right)$. More generally we know that $g\left(e\right)$ will be shifting up and down with different realizations of u, bounded by a maximum and a minimum value for u (and therefore, the variance of u).[3] Similarily the exact location of $e\left(g\right)$ differ across countries depending on the realization of v. The outer boundaries for

[3]The variance may be as large as you want, as long as it is finite.

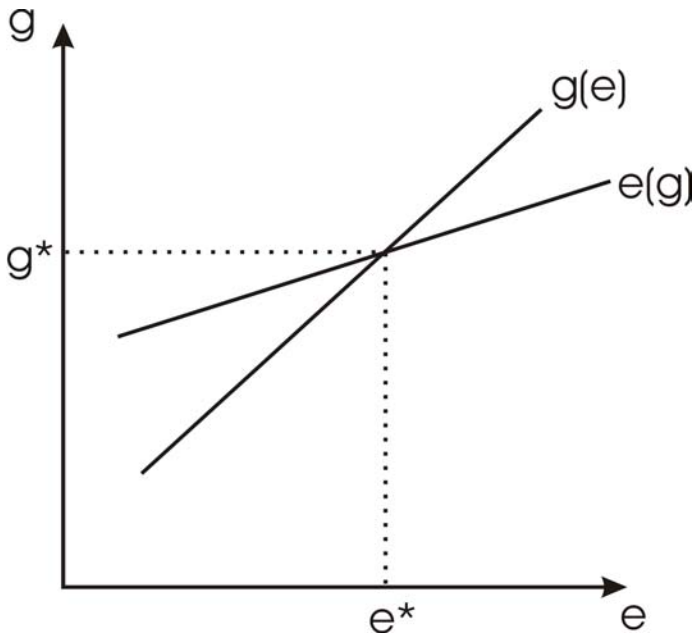

Figure 1: Schooling and growth when causality runs both ways.

the two functions (thus corresponding to the smallest and largest observation of u and v, respectively) are illustrated by the dotted curves in Figure 2. Now, suppose every observed data point correspond to an intersection point between the two curves. That is, assume "reality" is an equilibrium outcome. Then every data point will fall in the diamond shaped set: ABCD. The OLS estimator then selects an interscept and slope such that the sum of deviations from the associated line as small as possible. In practice this means a lineare function similar to the "OLS- line" depicted in Figure 2. As you can see, the slope will essentially be a convex combination of the slope of $g(e)$ and $e(g)$; α_1 and $1/\beta_1$, repectively. The weight on α_1 and $1/\beta_1$ will in practice depend on which of the two error terms exhibits the greater variance.[4] Since we don't know the variances, there is no way we can recover α_1 or β_1.

Can we at least say something about the *direction* of the bias of α_1?

[4]Figure 2 is essentially the case where the variance of u and v are identical. If the variance of, say u, is greater than that of v the "diamond" changes shape. You can convince yourself that the OLS slope estimate will be numerically smaller in this case.

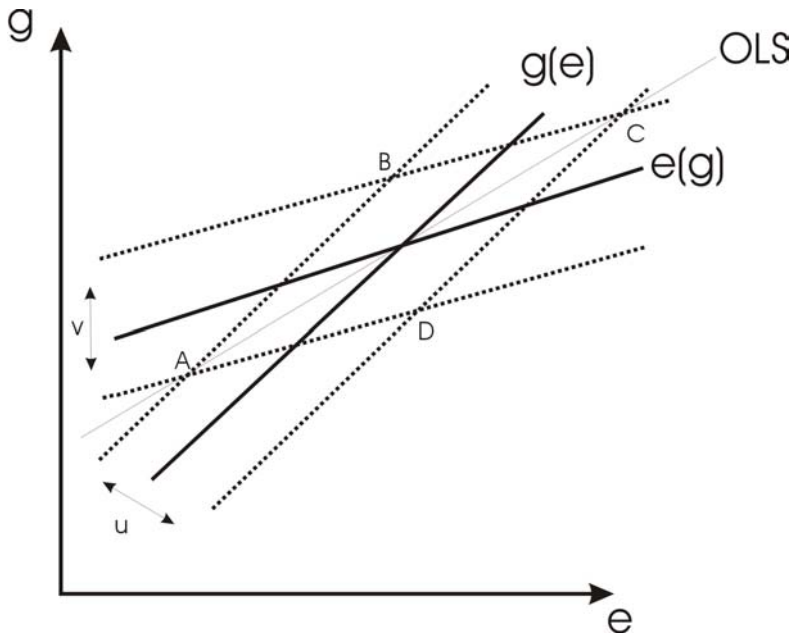

Figure 2: Getting nothing right using OLS.

Unfortunately, in general the answer is in the negative. The intuition can be obtained geometrically. The case illustrated in Figure 2 is where the slope of $g(e)$ is larger than the slope of $e(g) - \alpha_1 > 1/\beta_1 > 0$. Since the OLS estimate is a convex combination of the two, our OLS estimate $\hat{\alpha}_1 \in \left(\frac{1}{\beta_1}, \alpha_1\right)$. In this case we are clearly underestimating the "true" impact from e on growth. But if $0 < \alpha_1 < 1/\beta_1$ the opposite is the case! The OLS estimate $\hat{\alpha}_1$ will be smaller than its "true value". Hence, without strong priors as to the *relative magnitude* of education's effect on growth and vice versa, we are unable to assess whether the effect from education on growth is over- or underestimated by OLS.[5]

In sum, if two variables are jointly endogenous OLS is no longer BLUE. The direction of the bias, however, is often difficult to assess.

[5] Alternatively we can recover the same insight by looking at equation (10). It is immediately clear that the sign of the covariance between e and u depends on the sign of $(1 - \alpha_1\beta_1)^{-1}$. And the latter is positive (negative) iff $\alpha_1 < 1/\beta_1$ ($\alpha_1 > 1/\beta_1$), which (cf equation (9)) implies that $\hat{\alpha}_1 > \alpha_1$ ($\hat{\alpha}_1 < \alpha_1$).

2.3 Omitted Variables

Suppose education is in no way affected by growth, there is no measurement error issues either. Are we home free? Not necessarily, unfortunately.

Another sort of bias, which however boils down to exactly the same thing (i.e. the covariance between the right hand side variable and the disturbances is non-zero), arises when we forget a variable which should have been in the model.

To illustrate, while staying with out education example, suppose quality, q, matters for growth as well. Hence, consider the possibility that the true model is

$$g_i = \alpha_0 + \alpha_1 e_i + \alpha_2 q_i + u_i.$$

Next, suppose we ignore q_i when we estimate the equation. Then, in effect, the model we are estimating is

$$g_i = \alpha_0 + \alpha_1 e_i + \eta_i$$

where $\eta_i \equiv \alpha_2 q_i + u_i$ contains the influence from quality. Our OLS estimate for α_1 is

$$\begin{aligned} \hat{\alpha}_1 &= \alpha_1 + \frac{cov\left(\eta_i, e_i\right)}{var\left(e_i\right)} \\ &= \alpha_1 + \frac{cov\left(\alpha_2 q_i + u_i, e_i\right)}{var\left(e_i\right)}. \end{aligned}$$

Now if the quality of the educational system tends to be higher in places with more formal schooling then $cov\left(\alpha_2 q_i + u_i, e_i\right) > 0$, and we are overestimating the importance of e_i for growth. Of course, if the omitted variable is negatively correlated with variables entering the right hand side variables, then the opposite is the case – OLS will be biased downward.