

ARE213

Econometrics

Spring 2006 UC Berkeley Department of Agricultural and Resource Economics

ORDINARY LEAST SQUARES III:

OMITTED VARIABLE BIAS AND PROXY VARIABLES(W 4.3)

A. OMITTED VARIABLE BIAS

Often we estimate a linear regression function but we are not completely sure that we have included all the relevant regressors. Here we investigate how omitting a variable affects the coefficients on the regressors we are most interested in. Suppose the true regression function is

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \dots + \beta_K \cdot X_{iK} + \beta_Z \cdot Z_i + \varepsilon_i,$$

with $\varepsilon_i \perp (X_i, Z_i)$, and $\mathbb{E}[\varepsilon_i] = 0$. We refer to this as the “long regression.” Now suppose we estimate the regression function without Z_i :

$$Y_i = \gamma_0 + \gamma_1 \cdot X_{i1} + \dots + \gamma_K \cdot X_{iK} + \eta_i,$$

referred to as the “short regression.” This regression is largely definitional: the coefficients are defined to be $\gamma = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1}(\mathbb{E}[\mathbf{X}\mathbf{Y}])$, so that $\mathbb{E}[\eta_i \cdot X_i] = 0$, but not necessarily $\eta_i \perp X_i$. In addition it is useful to consider the “artificial regression” of the omitted Z_i on a constant and the X_{ik} :

$$Z_i = \delta_0 + \delta_1 \cdot X_{i1} + \dots + \delta_K \cdot X_{iK} + \nu.$$

Again this is definitional, choose $\delta = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1}(\mathbb{E}[\mathbf{X}\mathbf{Z}])$ so that $\nu = \mathbf{Z} - \mathbf{X}\delta$ is by definition uncorrelated with \mathbf{X} .

If we estimate the short regression, and focus on the k th regressor, we will estimate γ_k . We are interested in the relation between γ_k and β_k . To see what this will look like, consider

the long regression, and substitute in for the omitted Z_i :

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \cdot X_{i1} + \dots + \beta_K \cdot X_{iK} + \beta_Z \cdot Z_i + \varepsilon_i \\ &= \beta_0 + \beta_1 \cdot X_{i1} + \dots + \beta_K \cdot X_{iK} + \beta_Z \cdot (\delta_0 + \delta_1 \cdot X_{i1} + \dots + \delta_K \cdot X_{iK} + \nu) + \varepsilon_i \\ &= (\beta_0 + \beta_Z \cdot \delta_0) + (\beta_1 + \beta_Z \cdot \delta_1) \cdot X_{i1} + \dots + (\beta_K + \beta_Z \cdot \delta_K) \cdot X_{iK} + (\beta_Z \cdot \nu + \varepsilon_i). \end{aligned}$$

Since the composite error term $\beta_Z \cdot \nu + \varepsilon$ is uncorrelated with the X 's by definition, the regression coefficients in this representation are what you get from the short regression. So,

$$\gamma_k = \beta_k + \beta_Z \cdot \delta_k,$$

or the omitted variable bias (the difference between the coefficient in the short regression, γ_k , and the coefficient in the long regression, β_k), is equal to the product of the coefficient on the omitted variable, β_Z , and the coefficient on the included regressor X_{ik} in a regression of the omitted variable on all included regressors, δ_k .

These equalities also hold exactly for the estimated regression coefficients, or

$$\hat{\gamma}_k = \hat{\beta}_k + \hat{\beta}_Z \cdot \hat{\delta}_k.$$

The practical relevance of this may seem small. In practice we do not observe the omitted variable, and so we cannot estimate these regression coefficients. If we could, we would not have the bias! Nevertheless, this result is extremely useful. Let us see how this is used in practice. Consider a wage regression of log earnings on education of the type we looked at before:

$$\begin{aligned} \widehat{\log(\text{earnings})}_i &= 5.0455 + 0.0667 \cdot \text{educ}_i \\ &\quad (0.0849) \quad (0.0062) \end{aligned}$$

We may be concerned that we did not control for differences in ability between different individuals. We can think of that as having erroneously omitted ability from this regression. The long regression would thus have been

$$\log(\text{earnings})_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{ability}_i + \varepsilon_i.$$

Now what can we say about this? It is likely that the coefficient on ability is positive. Also, it seems plausible that the correlation between ability and education is positive. Hence the bias is positive: we over-estimate the returns to education because high-ability people who would already have relatively high earnings also have high levels of education.

Now let us see how this works out when we actually have a measure for ability. Here we take one such measure from the NLSY, namely an IQ measure. This is obviously a flawed measure of ability even if there is such thing, but it will do for our purpose. First look at the long regression:

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.7050 + 0.0443 \cdot \text{educ}_i + 0.0063 \cdot \text{iq}_i \\ &\quad (0.1003) \quad (0.0071) \quad (0.0010) \end{aligned}$$

The coefficient on education is indeed much smaller than when we did not control for ability. The difference in the coefficients is $0.0667 - 0.0443 = 0.0224$. This should be equal to the product of the coefficient on the omitted variable (0.0063) and the coefficient in a regression of the omitted variable on the included ones. That regression leads to

$$\begin{aligned} \widehat{\text{iq}}_i &= 53.6872 + 3.5388 \cdot \text{educ}_i, \\ &\quad (2.6229) \quad (0.1922) \end{aligned}$$

with slope coefficient of 3.5388, so that the product is indeed 0.0024.

B. PROXY VARIABLES

So far we have largely noted the omitted variable bias, and developed methods for at least assessing the sign of the bias. What else can we do? An additional approach is to use

a proxy variable. (For this I reasonably closely follow Wooldridge's discussion.) The idea is to include another covariate into the regression and thus eliminate at least part of the bias that stems from omitting the earlier variable. So suppose we would like to run the long regression:

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \dots + \beta_K \cdot X_{iK} + \beta_Z \cdot Z_i + \varepsilon_i.$$

We do not observe Z_i , but instead we observe a proxy W_i . This proxy variable does not enter into the original regression (but it is uncorrelated with ε_i). It is correlated with the omitted variable Z_i , and so we run the regression

$$Y_i = \gamma_0 + \gamma_1 \cdot X_{i1} + \dots + \gamma_K \cdot X_{iK} + \gamma_W \cdot W_i + \nu_i.$$

Wooldridge discusses two assumptions that make W_i a perfect proxy variable. The first is that W_i is uncorrelated with ε_i . Hence, if we were to run the regression

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \dots + \beta_K \cdot X_{iK} + \beta_Z \cdot Z_i + \beta_W \cdot W_i + \varepsilon_i,$$

the estimator for β_w would converge to zero. Second, partialling out the proxy variable W_i , the omitted variable Z_i is uncorrelated with the included variables X_{i1}, \dots, X_{iK} . Formally, if we run the regression

$$Z_i = \delta_0 + \delta_1 \cdot X_{i1} + \dots + \delta_K \cdot X_{iK} + \delta_W \cdot W_i + \nu,$$

the estimators for $\delta_1, \dots, \delta_K$ converge to zero. Under those assumptions we can directly use the earlier results on omitted variable bias. Think of the regression function including all $X_{i1}, \dots, X_{iK}, Z_i, W_i$ as the long regression, the regression functions omitting Z_i (but including $X_{i1}, \dots, X_{iK}, W_i$) as the short regression, and the regression of Z_i on $X_{i1}, \dots, X_{iK}, W_i$ as the artificial regression. The coefficient on X_{ik} in the short regression is by the omitted variable bias result equal to $\gamma_k = \beta_k + \beta_Z \cdot \delta_k = \beta_k$ because δ_k is zero, and so there is no bias

for this coefficient. The coefficient on W_i in the short regression is $\beta_W + \beta_Z \cdot \delta_W = \beta_Z \cdot \delta_W$, which is biased, but we typically do not care about the coefficient on the proxy variable.

Now let us look at this without assuming that W_i is a perfect proxy variable. There is now a large number of regressions floating around. For expositional reasons we look at the case with a single covariate X_i . We are interested in β_1 in the regression:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_Z \cdot Z_i + \varepsilon_{1i}.$$

We wish to compare the bias resulting from omitting Z from this regression, and estimating the regression

$$Y_i = \lambda_0 + \lambda_1 \cdot X_i + \varepsilon_{2i}, \tag{1}$$

with the bias of the regression where we replace Z_i with W_i and estimate

$$Y_i = \gamma_0 + \gamma_1 \cdot X_i + \gamma_W \cdot W_i + \varepsilon_{3i}. \tag{2}$$

We also need to consider the long regression

$$Y_i = \alpha_0 + \alpha_1 \cdot X_i + \alpha_Z \cdot Z_i + \alpha_W \cdot W_i + \varepsilon_{4i}, \tag{3}$$

and the artificial regressions

$$W_i = \kappa_0 + \kappa_X \cdot X_i + \kappa_Z \cdot Z_i + \varepsilon_{5i},$$

$$Z_i = \delta_0 + \delta_X \cdot X_i + \delta_W \cdot W_i + \varepsilon_{6i},$$

and

$$Z_i = \theta_0 + \theta_X \cdot X_i + \varepsilon_{7i}.$$

First consider the relation between the coefficients of interest and the coefficients from the long regression. Using the omitted variable bias formula we have

$$\beta_Z = \alpha_Z + \alpha_W \cdot \kappa_Z,$$

and

$$\beta_1 = \alpha_1 + \alpha_W \cdot \kappa_X.$$

Next consider the bias from running (1). Using the omitted variable bias formula the bias is equal to

$$\beta_Z \cdot \theta_X = \alpha_Z \cdot \theta_X + \alpha_W \cdot \kappa_Z \cdot \theta_X. \quad (4)$$

Now consider the bias from replacing Z_i by W_i . Using the omitted variable bias formula for the third time we have

$$\gamma_1 = \alpha_1 + \alpha_Z \cdot \delta_X = \beta_1 - \alpha_W \cdot \kappa_X + \alpha_Z \cdot \delta_X,$$

so that the bias is

$$-\alpha_W \cdot \kappa_X + \alpha_Z \cdot \delta_X.$$

Now suppose the own effect of the proxy variable is fairly small ($\alpha_W \approx 0$). In that case the bias for omitting Z_i is $\alpha_Z \cdot \theta_Z$, and the bias from replacing it with W_i is $\alpha_Z \cdot \delta_X$. The latter is smaller if $|\delta_X| < |\theta_Z|$, that is, if controlling for W_i lowers the correlation between Z_i and X_i .

Finally, let us see how this plays out with real data. Suppose we are interested in regression log earnings on education controlling for iq. The estimated regression would be

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.7050 + 0.0443 \cdot \text{educ}_i + 0.0063 \cdot \text{iq}_i \\ &\quad (0.1003) \quad (0.0071) \quad (0.0010) \end{aligned}$$

If we did not observe IQ we could estimate the regression without it:

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 5.0455 + 0.0667 \cdot \text{educ}_i \\ &\quad (0.0849) \quad (0.0062) \end{aligned}$$

Alternatively we could estimate a regression using a test score KWW, knowledge of the world of work) as a proxy variable. This proxy variable regression is

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.8004 + 0.0479 \cdot \text{educ}_i + 0.0140 \cdot \text{KWW}_i \\ &\quad (0.0890) \quad (0.0066) \quad (0.0019) \end{aligned}$$

Here the proxy variable method seems to work quite well. To understand that better, let us look at the various components of the bias. First, the long regression:

$$\begin{aligned} \log(\widehat{\text{earnings}})_i &= 4.5925 + 0.0347 \cdot \text{educ}_i + 0.0046 \cdot \text{iq}_i + 0.0117 \cdot \text{KWW}_i \\ &\quad (0.1002) \quad (0.0072) \quad (0.0011) \quad (0.0019) \end{aligned}$$

The three artificial regressions are:

$$\begin{aligned} \widehat{\text{KWW}}_i &= 9.6469 + 0.8285 \cdot \text{educ}_i + 0.1475 \cdot \text{iq}_i, \\ &\quad (1.6603) \quad (0.1180) \quad (0.0172) \end{aligned}$$

$$\begin{aligned} \widehat{\text{iq}}_i &= 44.9921 + 2.8657 \cdot \text{educ}_i + 0.4950 \cdot \text{KWW}_i, \\ &\quad (2.7229) \quad (0.2009) \quad (0.0578) \end{aligned}$$

and finally

$$\begin{aligned} \widehat{\text{iq}}_i &= 53.6872 + 3.5388 \cdot \text{educ}_i. \\ &\quad (2.6229) \quad (0.1922) \end{aligned}$$

So first decomposing the bias from the omitted variable regression is

$$\alpha_Z \cdot \theta_X + \alpha_W \cdot \kappa_Z \cdot \theta_X = 0.0046 \times 3.5388 + 0.0117 \times 0.1475 \times 3.5388 = 0.0224$$

and the bias from the proxy variable regression is

$$-\alpha_W \cdot \kappa_X + \alpha_Z \cdot \delta_X = -0.0117 \times 0.8285 + 0.0046 \times 2.8657 = 0.0035.$$

The bias from the proxy regression is one sixth of that of the omitted variable regression. This is despite α_W being fairly large (0.0117), but due to the offsetting biases in the proxy variable regression.