Omitted Variable Bias

Alan Liang

May 2018

1 Motivation

From the last note, we've learned that linear regression can help us control for potential covariates. But what is the effect of not considering and controlling for covariates?
In most real-life experiments, we will not have access to all the data of all possible covariates: so how can we get a feel for what direction the selection bias is going?
As its name suggests, **omitted variable bias (OVB)** accounts for this. For a particular omitted variable, OVB allows us to estimate the sign and magnitude of the bias associated with the variable. OVB calculuates this based on the omitted variable's relationship with other variables that we are taking into account of.

2 The Omitted Variable Bias

2.1 Intuition and Definition

Say you have measured explanatory variable X, covariate A, and outcome Y, but decided to naively not control for A. Your singular linear regression is known as the **'short' regression**, and follows the format:

$$Y = \alpha^s + \beta^s X$$

On the other hand, in reality the relationship involves us controlling for A. This is known as the **'long' regression**:

$$Y = \alpha^l + \beta^l X + \gamma A$$

Note that $\alpha^l \neq \alpha^s$ and $\beta^l \neq \beta^s$, as one takes into consideration of the confounding variable.
Intuitively, the difference in β, $\beta^s - \beta^l$, measures the difference in the slope if we did not account for our covaraite. Assuming that β^l is the true treatment effect slope (because we accounted for all possible covariates), then $\beta^s - \beta^l$ would be how wrong our short regression was. This is how we calculate the **omitted variable bias**:

$$\text{OVB} = \beta^s - \beta^l$$

2.2 Derivation from Partial Slopes

In the example above, we knew the correct results going in: we knew that A was the confounding covariate and had the data for A at every data point, allowing us to control for it in our linear regression. However, this does not happen in real life. In real life, we will not have recorded for A and hence not have been able to determine β^l.
In this case, what do we do? In essence, we are forced to look at the relationship of the omitted variable with variables we have data of. Following with the notation from above, define π:

$$\pi = \frac{\mathrm{d}A}{\mathrm{d}X}$$

This measures the relationship between the omitted variable to the explanatory variable: how much does our omitted variable change for every unit change in the explanatory variable? γ is the same from the

long regression equation:

$$\gamma = \frac{\delta Y}{\delta A}$$

This measures the relationship (partial slope) between the outcome and the omitted variable: how much does our outcome variable change for every unit change in the omitted variable, holding all other variables still. Note that unlike π, this is a partial slope, so it assumes that variables such as X are being held constant. Together, the omitted variable bias is:

$$\text{OVB} = \pi \times \gamma$$

Very very informally (to build intuition only), notice that if we were to multiply π and γ, we can cancel out the partial omitted variable:

$$\frac{\cancel{\mathrm{d}A}}{\mathrm{d}X} \times \frac{\delta Y}{\cancel{\delta A}} = \frac{\delta Y}{\delta X}$$

These units adheres to what we want, the units of β: the change in outcome due to change in explanatory variable.
Hence, the definition of the omitted variable bias is as follows:

$$\text{OVB} = \beta^s - \beta^l = \pi \times \gamma = \frac{\mathrm{d}A}{\mathrm{d}X} \times \frac{\delta Y}{\delta A}$$

2.3 An example

I'm going to shamelessly steal this example from lecture:
Say you are trying to determine the effect of neighborhood income (X from above) on coffee shop sales (Y from above). A confounding factor may be the number of competitors nearby (A from above).
However, let us assume that we do not have access to A, so we can only try to determine the sign of the OVB.
Let's first calculate γ, the change in number of coffee shops nearby times with respect to the change in local income. γ should be positive as with an increase in local income, we should expect to see more competitors.
π, the partial slope of sales with respect to number of competitors, should be negative: holding income the same, if there are more competitors, we should see a decrease in sales.
Hence, combined, our omitted variable bias should be a negative number, which means that $\beta^s < \beta^l$, so that controlling for competitors, the effect of income on coffee sales is larger than had we not considered it.