Controls in Two-Step Procedures

Tom Glinnan

November 2021

The Question

One of the questions in class today (1a - PS 6) was all about using the OVB formula in situations where there are other controls in the regression. Suppose that we have short and long regressions as so:

$$\begin{aligned}
\text{wage}_i &= \alpha + \beta_1 \text{college}_i + \beta_2 \text{female}_i + \varepsilon_i \\
\text{wage}_i &= \alpha^l + \beta_1^l \text{college}_i + \beta_2^l \text{female}_i + \beta_3^l \text{age}_i + u_i
\end{aligned}$$

We're interested in how the coefficient on *college* changes when we add in the *age* regressor. Note that this is the exact same as the usual examples, apart from addition of the control *female*. Whether or not you use the 'chain rule' intuition that I've mentioned[1], you probably remember the OVB formula as something like:

$$\beta^s = \beta^l + \gamma\delta$$

where β^s is the coefficient of interest in the short regression, β^l is the coefficient of interest in the long regression, γ is the coefficient of the omitted variable in the long regression, and δ is the 'coefficient in the auxiliary regression'. For the purposes of question 1a (and in general!) we want to understand how to match up this abstract form of the OVB equation to the regressions above, so we can pick the numbers out of Stata tables

The first three are quite easy: β^s is β_1, β^l is β_1^l and γ is β_3^l. The question is, *where is the* δ?

We know we have to run some auxiliary regression of some kind. The question is, which one the following regressions is it?

$$\begin{aligned}
\text{college}_i &= \alpha^m + \beta_1^m \text{age}_i + v_i \\
\text{college}_i &= \alpha^m + \beta_2^m \text{female}_i + \beta_3^m \text{age}_i + v_i
\end{aligned}$$

In other words, do we include the control *female* in the auxiliary regression? Is δ equal to β_1^m or β_3^m? This is what I was[2] trying to ask in class

[1] If you want to remember what that is - I have another note on that on Github under one of the earlier weeks

[2] Very confusingly!

The Solution (plus notes on why we control)

If you look in your lecture slides, you'll see that we **should** include controls in the auxiliary regression - and so take δ to be β_3^m. For the purposes of the exam it's enough to just say that this is the right formula, or rely on the mathematical argument given in the lecture notes. That's fine, but it helps to have some intuition for why we do that

Remember what it means to add something as a control in a regression. Essentially, it moves us from a 'total derivative' interpretation of coefficients to a 'partial derivative' one - the coefficient β in $y_i = \alpha + \beta x_i + \gamma z_i + \varepsilon_i$ represents the expected change in y_i if x_i were one unit higher *holding* z_i *constant*; in other words, comparing people who have the same level of z_i. This interpretation holds for both discrete and continuous variables

Sidenote on why we control - going beyond EC221

This is true no matter whether controls are good or bad - what we're saying here is a completely correlation-based statement. The idea underlying whether a control is good or bad comes down to whether or not comparing people with the same level of z_i helps you or harms you in getting to causality. Namely: you should control for confounders because if you have a variable that causes both the treatment T and outcome O to go up, then you'll observe a positive correlation between T and O even if there's no relationship between them[3], because high values of the confounder will lead to high T and high O, and low values of the confounder lead to low T and O. By controlling for such a confounder, you'll only be looking at people who have the same value of the confounder, and so for all people who are being compared T and O will be pushed up by the same amount, so we would see no relationship between T and O (which is what we want). By contrast, controlling for a collider[4] is bad because if there's no relationship between T and O[5], then by comparing people with the same level of C they will be people with a high T and a low O or a low T and high O (assuming that the correlations along both arrows are positive) - and as such we will see a negative correlation even when there isn't one. Try to see if you can explain in your own words why we should control on confounders, but not colliders, and see me if you need help. Then, to see what we should or should not control for in a given causal graph (the thing with the arrows), we have to realise that controlling for any confounder-like substructures 'blocks' ways in which T affects O, and controlling for collidier-like structures 'opens' them. The key is to pick controls which block all paths from T to O except the one we care about. The 'bad control' structure given in EC221 is motivated by the idea that conditioning on the thing we say not to opens up one of these 'bad paths'. If interested: the pdf on controlling on Github is good at explaining this

The main takeaway, ignoring the details above, is that controlling for a variable z_i means you're comparing people at the same value of z_i. These people might not exist in your data but we're still effectively doing this

[3] ie when there is no $T \rightarrow O$ arrow

[4] The example I mentioned at the end of last week's class, where we have $T \rightarrow C$, $T \rightarrow O$ and $O \rightarrow C$. This isn't talked about so much in the course

[5] Such that we only have $T \rightarrow C$ and $O \rightarrow C$

Back to Solution

The big intuition is: in any two-step procedure, we need to have the same controls everywhere, because if we don't then **in one step we'll be comparing people with the same z_i, and in the other step we won't**

For example, in the OVB formula we're doing this 'two step' process - we're breaking down the total effect into partial effects. Both the γ and δ in the OVB formula must come from regressions with all of the controls included. This way, we're comparing the same groups of people throughout our whole procedure. If instead we took γ or δ from a shorter regression (ie without the *female* control variable here) then we would be holding female constant (and so comparing men and women separately) to get β^l, but not to get γ or δ (and so comparing men and women together). We're not being consistent here! Formally, we can show that this results in your procedure not being correct; here, that means that the OVB formula won't hold unless we have controls in all of our regressions. A similar situation happens when we do **2SLS with controls**. If we don't put the same controls in both our first and second stage, then we'll be comparing amongst different groups of people in both stages. This will mean we don't quite get what we want, which is the IV estimator

Please let me know if you have any questions!