Model Specification

EC 320: Introduction to Econometrics

David Hall
University of Oregon

Review and Today's Goal

Last time, we looked at:

- Dummy/Indicator Variables
- Difference-in-Differences Design

Dummy variables help us understand the impact of more qualitative/categorical variation.

DiD is a way we use dummy variables to get around having to assume full exogeneity (or one of the ways).

Today, we cover model specification, and what happens to our coefficients/SE when the right variable(s) are not included.

Warm Up i

Suppose you wanted to test the impact of a change in the Federal Funds Rate on the pricing of Mortgage Backed Security pricing firms. *You believe that the impact of the change in the FFR might have a different impact on the price charged by firms small, medium and large.*

Let P_{it} be the price charged by firm $i \in \{1, ..., n\}$ at time $t \in \{1, \dots T\}$. Let F_t be the current Fed Funds Rate charged at time t. Denote the size of firms however you wish.

1. Write a regression equation corresponding to the RQ: What is the impact of the FFR on mortgage pricing, and how does it vary by firm size?

Warm Up ii

2. Suppose the Fed decides at time t_0 to charge different rates for small-cap and keep large-cap firms the same (the Fed does not think medium size firms exist). We have data before the rate change. Define your *treatment* and *control* variables, your *post* variable, and write the DiD Specification testing the impact of this change on MBS pricing.

Answers:

For #1:

$$P_{it} = \beta_0 + \beta_1(Small_i \times FFR_t) + \beta_2(Med_i \times FFR_t) + \Gamma X_{it} + \varepsilon_{it}$$

OR

$$P_{it} = \beta_1(Small_i \times FFR_t) + \beta_2(Med_i \times FFR_t) + \beta_3(Large_i \times FFR_t) + \Gamma X_{it} + \varepsilon_{it}$$

and we might also want to just include *Small*, *Med* and *Large* on their own to capture the base differences in pricing among different-size firms.

Answers for #2:

Define:

- Treatment = Small-cap (e.g. $Cap \leq \$N$)
- Control = Large-cap
- $Post_t = 1\{t > t_o\}$

Regression:

$$P_{it} = \beta_0 + \beta_1 Small_i + \beta_2 Post_t + \beta_3 (Small_i \times Post_t) + \Gamma X + \varepsilon_{it}$$

Q: How do we interpret this?

Transition Slide

Please do not actually forget everything, though.

Variable Specification

The following table shows the two ways we think about variable (mis)specification and their consequences.

Table 6.1 Consequences of variable specification

		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = b_1 + b_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Coefficients are unbiased (in general) but inefficient. Standard errors are valid (in general)	Correct specification, no problems

Two Issues:

1. Omission of a Variable – BIAS
2. Inclusion of a Redundant Variable – Inefficiency

Today, we will cover the first in order to set up our discussion of instrumental variables next week.

Omitted Variable Bias: Example

You have actually already seen an example of OVB when we discussed hypothesis testing with Hank.

- Hank thought the probability of a successful first date was solely dependent on the length of one's pickup line.
- Hank did *not* consider that attraction plays a significant role in dating.
- Without attractiveness as a variable, the length of a pickup line is biased.
    - In fact, it is picking up both the causal effect of the pickup line length on first-date-probability *as well as* how attractiveness impacts the success of your pickup line.

Other Examples of Omitted Variables:

1. Education and wage, what about natural ability?
   - Some even believe we should control for occupation (this is an ongoing debate).
2. Housing price and square footage, but what about age of the house?
3. Physical activity and bone density, what about weight?

A quick intro to a DAG

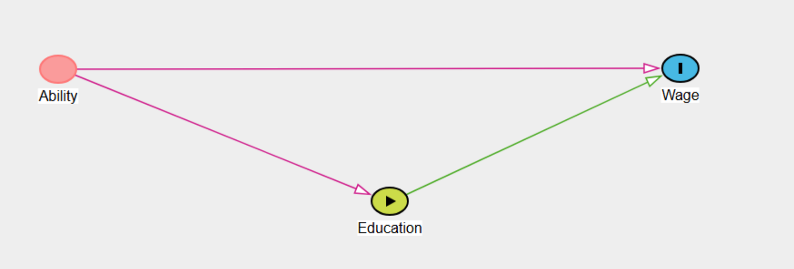

Figure 1: Classic OVB is thought of as a "confounder."

Let's Formalize This

Suppose the true model for Y, the data generating process, depends on both X and Z according to:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Suppose you thought Z did not matter really at all, so you modeled:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and after fitting you get our typical estimate of β_1 :

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

Recall: If $\hat{\beta}_1$ were unbiased... i

If $Y = \beta_0 + \beta_1 X + \varepsilon$ was the right model, we can show that our estimate of β_1 is unbiased if we assume that $E[\varepsilon_i | X] = 0$.

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\varepsilon}$
- $\implies Y_i - \bar{Y} = \beta_1 X_i - \bar{X}) + (\varepsilon_i - \bar{\varepsilon})$

Recall: If $\hat{\beta}_1$ were unbiased... ii

Thus we can write:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (\varepsilon_i - \bar{\epsilon})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

and since $E[\varepsilon_i | X] = 0$, $\hat{\beta}_1 = E[\beta_1] \checkmark$

BUT WAIT

Clearly, in the case of omitted variable bias, $E[\epsilon_i|X] \neq 0$. Work in groups and try to figure out what the bias is going to be for the problem given the following:

- True Model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon$
- Regression Basis: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- $\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$

OVB Formally

If the true model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon$ then:

$$Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + \beta_2(Z_i - \bar{Z}) + (\varepsilon_i - \bar{\varepsilon})$$

We can thus expand our OLS estimator of $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(\beta_1(X_i - \bar{X}) + \beta_2(Z_i - \bar{Z}) + (\varepsilon_i - \bar{\varepsilon}))}{\sum_i (X_i - \bar{X})^2}$$

Cont'd.

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(\beta_1(X_i - \bar{X}) + \beta_2(Z_i - \bar{Z}) + (\varepsilon_i - \bar{\varepsilon}))}{\sum_i (X_i - \bar{X})^2}$$

$$\implies \hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_i (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i (X_i - \bar{X})(\varepsilon_i - \bar{\epsilon})}{\sum_i (X_i - \bar{X})^2}$$

$$E[\epsilon_i | X = 0] \implies E\left[\frac{\sum_i (X_i - \bar{X})(\varepsilon_i - \bar{\epsilon})}{\sum_i (X_i - \bar{X})^2} | X\right] = 0$$

$$\implies \hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_i (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_i (X_i - \bar{X})^2}$$

The Bias Portion

There are two parts to the bias portion:

$$\hat{\beta}_1 = \beta_1 + \underbrace{\beta_2}_{\text{Effect of } Z \text{ on } Y} \underbrace{\left(\frac{\sum_i (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_i (X_i - \bar{X})^2} \right)}_{\text{OLS coefficient from regressing } Z \text{ on } X}$$

i.e. this is $\hat{\alpha}_1$ from a regression of $\hat{Z} = \hat{\alpha}_0 + \hat{\alpha}_1 X$. (This works for the simple regression only, really this is just the sample correlation between X and Z.)

What does this mean?

This means that, if you *must* omit a variable for whatever reason, we can use logic/theory to tell us what we might expect the sign of the bias to be! Let's think about some concrete examples.

1. Education on wage, omit natural ability.
    - What is the impact of ability on wage?
    - What is the correlation of ability and education?
2. Hank's problem, we omit attractiveness.
    - What is the impact of attractiveness on first-date probability?
    - What is the correlation of attractiveness and how well a pick up line works?

Questions?

Next Class

- Including a redundant variable.
- Instrumental Variables