

Lecture 8

Multiple Regression III

Outline

- The notion of control
- Good, bad and useless controls
- Dummy Variable Trap

Unbiasedness of OLS

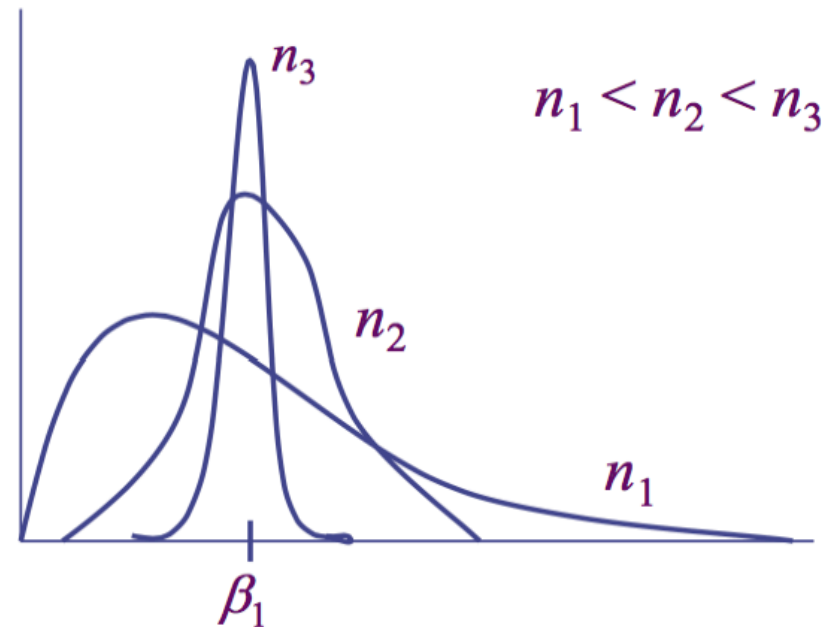
- In the econometric model we have population parameter β , and what we get from data is a sample estimate $\hat{\beta}$. If we draw many samples, we will get multiple $\hat{\beta}$. Those $\hat{\beta}$ form a sampling distribution with a certain mean and standard deviation
- Unbiased estimator:
 - If $E(\hat{\beta}) = \beta$, then $\hat{\beta}$ is unbiased
 - If the mean of all those $\hat{\beta}$ s equal to the true parameter β , then it is unbiased
- If the estimate is unbiased, it is likely we get the true causal effect.
 - It doesn't mean the specific $\hat{\beta}$ is the true β , it means if the sample is "typical", we are near the true effect

When is OLS estimate unbiased?

- There are four conditions
- The most important one is “zero conditional mean”, $E(\varepsilon | X)=0$
 - It is an assumption made in the population, it's not about a specific sample
 - The expected value of error term is zero for every given X .
 - In other words, the error is randomly distributed for a given X .
- It is an untestable assumption
- Likely to be true or not? Can you think of an example when OLS is biased?

Consistency of OLS

- The conditions for unbiased estimator are hard to meet, so we settle for estimators that are consistent.
- Consistency: the distribution of the estimate becomes more tightly distributed around β as the sample size grows
- As $n \rightarrow \infty$, the distribution of the estimate $\hat{\beta}$ collapse to the parameter value β



Unbiasedness and consistency

- For unbiasedness, we have “zero conditional mean”:

$$E(\varepsilon | X) = 0 \quad (1)$$

- For consistency, we have a weaker assumption-“zero mean and zero correlation”:

$$E(\varepsilon) = 0 \text{ \& Cov}(X, \varepsilon) = 0 \quad (2)$$

- $E(\varepsilon) = 0$ is always true for OLS estimator, so $\text{Cov}(X, \varepsilon) = 0$ is the key condition
- Why is (2) a weaker condition of (1)?
- Assumption (1) requires the error term to be completely random given each X
 - e.g. ε cannot be related to X^2
- Assumption (2) says that error is not correlate with X , but can be correlated with other stuff
 - e.g. ε can be correlated with X^2

Unbiasedness and consistency

- Unbiasedness is a **finite** sample property
- Consistency is a **large** sample property
- Unbiased means on average you will get the true β
- Consistent means on average you may not get the true β , but as sample size increase, you will be more likely to get true β
- Consistency requires a weaker assumption, therefore the condition to get correct causal estimate is relaxed to **$\text{Cov}(X, \varepsilon)=0$** .

Why do we have biased $\hat{\beta}$?

- Assume the true model for whether insurance affect health is
 - $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3 Income + \varepsilon$
- We are now running this model:
 - $Health = \alpha' + \beta' Insured + \gamma_1' Age + \gamma_2' Educ + \varepsilon'$
- Is β' the true causal impact of insured on health (β)?
- Recall the condition for β' to be causal is $Cov(\varepsilon', Insured) = 0$
- Which variable is now left in ε' ?
 - Income
- Is $Cov(income, Insured)$ likely to be 0?
 - No. Therefore $\hat{\beta}'$ is biased
- Failure to include proper controls result in biased estimates

The notion of control

- Meaning of control #1:
 - When you control for variables you hold that variable constant or fixed
 - Similar notion to a lab experiment—keep other factors held fixed (*ceteris paribus*)
- Meaning of control #2:
 - e.g. $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3 Income + \varepsilon$
 - Within group comparison: compare mean health status of the insured with the uninsured for those with same age, education and income. It's more proper than across group comparison
- Meaning of control #3:
 - “Partialling Out”
 - β tells us the causal effect of whether insured on health after the confounding effect of age, education and income has been partialled (or netted) out

Omitted variable bias

- The bias in the OLS estimator that occurs as a result of an omitted factor is called **omitted variable** bias. For omitted variable bias to occur, the omitted factor “Z” must be:
 1. A determinant of Y (i.e. Z is part of ε); **and**
 2. Correlated with the regressor X (i.e. $\text{corr}(Z, X) \neq 0$)
- **Both** conditions must hold for the omission of Z to result in omitted variable bias.

Omitted variable bias direction

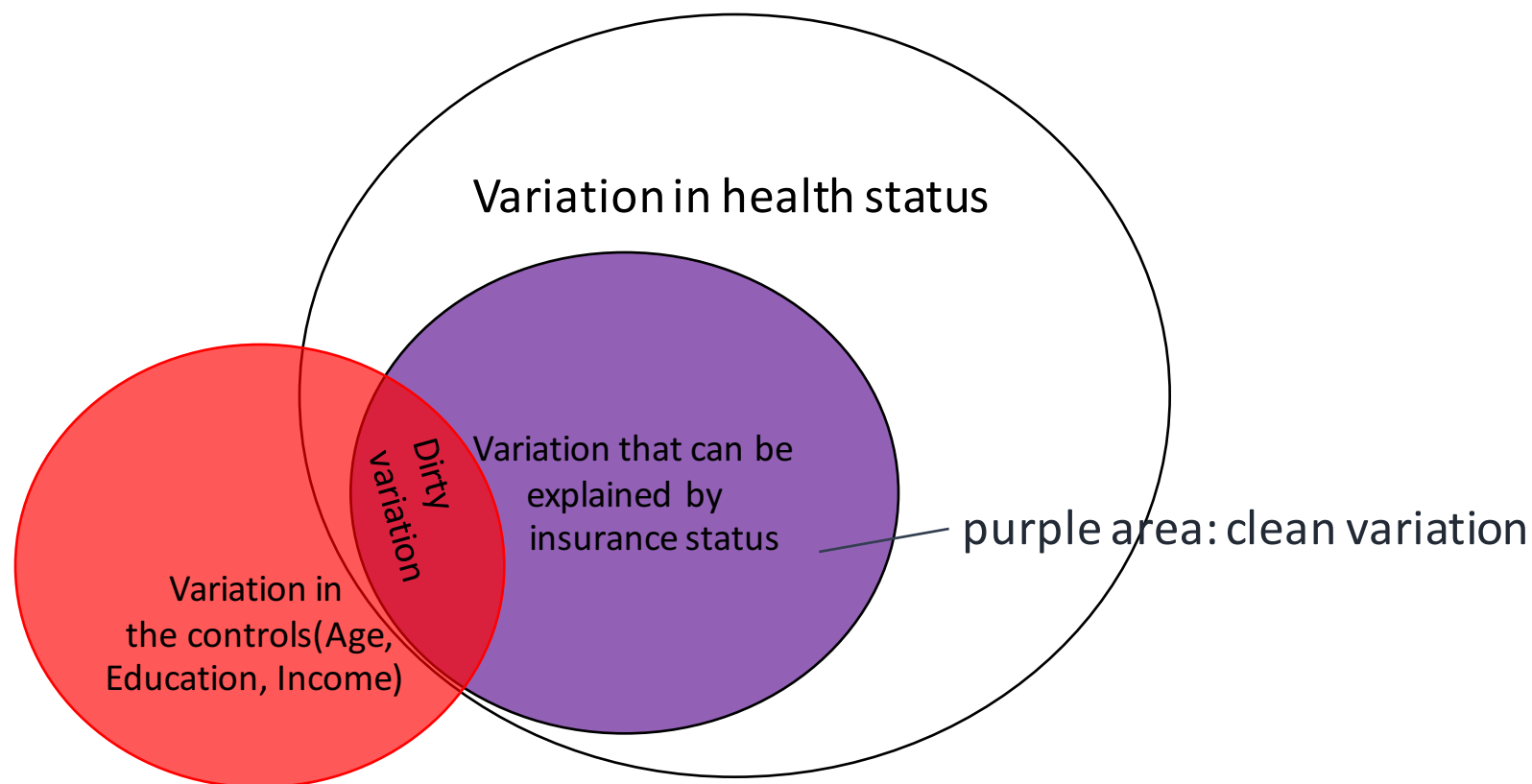
		Corr(omitted variable,x)	
		positive	negative
Corr(omitted variable,y)	positive	upward bias	downward bias
	negative	downward bias	upward bias

The notion of control

- $Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3 Income + \varepsilon$
- More on “partialling out”:
 - If insurance status is randomly assigned, we won’t have to control for other factors. In that case, all the variation in insured is “clean”.
 - Since insurance status is not randomly assigned, it contains both “clean” variation and “dirty” variation
 - The “dirty” variation is the part that correlates with age, educ or income.
 - By controlling for age, educ and income, we are partialling out the “dirty” variation, leaving the “clean” variation for identification. What does it exactly mean?

Diagram showing clean and dirty variation

$$Health = \alpha + \beta Insured + \gamma_1 Age + \gamma_2 Educ + \gamma_3 Income + \varepsilon$$



Good, bad and the useless

- Good controls are variables that you include in the regression in order to kill dirty identifying variation
- Useless controls are variables that you include in the regression that aren't correlated with treatment variable D.
 - Useless controls can still explain Y
 - Often it's worthwhile to include these controls to reduce your standard errors, but fundamentally, these controls don't do that much
- Bad controls are variables that are caused by D. Controlling for these variables is called "over controlling"

Example: Immigration

- Suppose immigration positively impacts crime rate
- If I regress crime on immigration, I may find a positive effect
- $Crime = \alpha + \beta \text{ Number of Immigrants} + \varepsilon, \hat{\beta} > 0$
- Is this relationship biased?
 - Possibly, because some other factors (e.g. big city indicator) is left in the error term. Once controlling for big city, the effect may change
- Note, when we include a control, we are more concerned whether it affects treatment **variable D**, not whether it affects Y.

Example: good, bad and useless controls

- Variable “big city” affects both immigration and crime rate, therefore it is a **good** control.
- Now you include another variable number of rainy days. You think number of rainy days will affect crime. But it has nothing to do with immigration. Therefore it is a **useless** control. (Although the adjusted R-square will increase, it doesn’t eliminate any “dirty” variation)
- Now consider the variable-local unemployment rate. This variable is likely to be caused by immigrants and therefore may be a **bad** control. Because it will take away the “clean” variation too. You will over-control by including it
 - Note: when you try to think of a bad control, think about the mechanisms why immigrants affect crime rate, controlling for these mechanisms are usually over-controlling

Criteria for good, bad and useless controls

- Good controls affect Y and D (variable of interest)
- Useless controls affect Y but not D
- Bad controls affects Y, and are caused by D (Bad controls are often mechanisms through which D affect Y)

Example: code categorical variables into dummies

- Sometimes, you want to turn a categorical variable into dummy variables to model the relationship more flexibly
- e.g. Education: 0 if HS dropout, 1 if HS grad, and 2 if college grad.

Recode it into three dummy variables:

- less_HS:1 if education=0, 0 otherwise
- HS_grad:1 if education=1, 0 otherwise
- college_or_above:1 if education=2, 0 otherwise

Dummy variable trap

- If we run $\text{Health} = \beta_0 + \beta_1 \text{Less_HS} + \beta_2 \text{HS_grad} + \beta_3 \text{College_or_above} + \varepsilon$
 - The regression will not work
 - This is called perfect multi-collinearity

Constant	less_HS	HS_grad	college	health status
1	1	0	0	3
1	0	1	0	4
1	0	1	0	4
1	0	0	1	3
1	0	0	1	2
1	1	0	0	1
1	0	1	0	4

**Less_HS+HS_grad+college_or_above
=constant
Cannot estimate the regression!**

How to get out of dummy variable trap(1)?

- There are two ways of getting out of the dummy variable trap
- Way #1: omit one category of the dummy variables
- e.g. if we run a regression:
- $\text{Health} = \beta_0 + \beta_1 \text{HS_grad} + \beta_2 \text{College_or_above} + \varepsilon$
- We omit HS dropouts dummy and treat it as the baseline. The categories we include are compared to the category we exclude
- How do we interpret β_1 ?
 - The average health status of HS grads relative to HS dropouts.
 - **It is the difference in average health between HS grad and HS dropouts**
- Exercise: Interpret β_0, β_2

How to get out of dummy variable trap(2)?

- Way#2: omit the constant term
- $\text{Health} = \beta_1 \text{LessHS} + \beta_2 \text{HS_grad} + \beta_3 \text{College_or_above} + \varepsilon$
 - How does this regression differ from the previous?
- How do we interpret β_1 ?
 - The average health status of HS dropouts.
- Exercise: Interpret β_2 and β_3

Review for quizzes

- Understand the conditions for unbiasedness and consistency and the relationship between them
- Understand why we need controls-Because $\text{cov}(D, \varepsilon) \neq 0$ is violated, and the estimate is biased
- Be able to come up with examples of three types of controls, and explain why they are good, bad and useless
- Know the dummy variable trap and how to get rid of it.