Lecture 7

Multiple Regression II

Outline

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

Takeaways about R-square and adjusted R-square

- Adjusted R-square is a standardized measure that can be used to compare models. It is useful if our objective is to making prediction
- In micro-econometrics, R-square is not that useful. Because our objective is NOT fitting the data, but to find whether one specific X has causal impact on Y.
  - e.g. How many factors can explain variation in one's health status?
  - Perhaps many more than we can include in the regression
  - It's natural to get a small R-square if we only include education as X
  - Our interest lies in whether education has caused variation in health, not in predicting health status based only on one variable

Mathematical formula for β (difficult)

- We run a multiple regression
  $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_N X_{Ni} + \varepsilon_i$
- We could run a regression of X_{1i} on all of the other X variables, and then compute the residuals- call it $\widetilde{X_{1i}}$(partial out the effect of all the other X variables)
- Then $\hat{\beta}_1 = \frac{Cov(\widetilde{X_{1i}}, Y_i)}{Var(\widetilde{X_{1i}})}$
- The relationship between $X1$ and Y holding all other X variables fixed

Test of joint hypothesis

- Let *STR* = student teacher rat*io, Expn* = expenditures per pupil, and *PctEL =* percent of English learners
- Consider the population regression model:
- $TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$
- The null hypothesis is that "school resources don't matter," and the alternative that they do, corresponds to:
- H_0: $\beta_1 = 0$ ***and*** $\beta_2 = 0$
- vs. H_1: ***either*** $\beta_1 \neq 0$ ***or*** $\beta_2 \neq 0$ ***or both***
- $TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$

Test of joint hypothesis

- H_0: $\beta_1 = 0$ ***and*** $\beta_2 = 0$
- vs. H_1: ***either*** $\beta_1 \neq 0$ ***or*** $\beta_2 \neq 0$ ***or both***
- A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.

The "restricted" and "unrestricted" regressions

Example: are the coefficients on *STR* and *Expn* zero?

<u>Unrestricted population regression (under H_1):</u>

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

<u>Restricted population regression (that is, under H_0):</u>

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i \qquad (why?)$$

- The number of restrictions under H_0 is $q = 2$ (*why*?).
- The fit will be better (R^2 will be higher) in the unrestricted regression (*why*?)

By how much must the R^2 increase for the coefficients on *Expn* and *PctEL* to be judged statistically significant?

F-statistic

- The *F*-statistic tests all parts of a joint hypothesis at once.

F-test

- With a multiple regression $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_N X_{Ni} + \varepsilon_i$, we can still perform statistical inference using p-value or confidence interval to determine if a specific β is **statistically different** from 0 (or other number)
- Now if we want to test whether a group of variables jointly have any effect on Y, we will use F-test.
- H_0: $\beta_1 = \beta_2 = .. = \beta_N = 0$
- H_a: at least one of the βs is not zero

Simple formula for the F-statistic:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

where:

$R^2_{restricted}$ = the R^2 for the restricted regression
$R^2_{unrestricted}$ = the R^2 for the unrestricted regression
q = the number of restrictions under the null
$k_{unrestricted}$ = the number of regressors in the unrestricted regression.

- The bigger the difference between the restricted and unrestricted R^2's – the greater the improvement in fit by adding the variables in question – the larger is the F.

Example:

Restricted regression:

Test score = 644.7 – 0.671*PctEL*, $R^2_{restricted} = 0.4149$

(1.0) (0.032)

Unrestricted regression:

Test score = 649.6 – 0.29*STR* + 3.87*Expn* – 0.656*PctEL*

(15.5) (0.48) (1.59) (0.032)

$$R^2_{unrestricted} = 0.4366,\ k_{unrestricted} = 3,\ \ q = 2$$

so
$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1-R^2_{unrestricted})/(n-k_{unrestricted}-1)}$$

$$= \frac{(.4366-.4149)/2}{(1-.4366)/(420-3-1)} = \mathbf{8.01}$$

F-statistic – summary

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

- The F-statistic rejects when adding the two variables increased the R^2 by "enough" – that is, when adding the two variables improves the fit of the regression by "enough"

Example of F-test

- $Health_i = \beta_0 + \beta_1 Educ_i + \beta_2 Age_i + \beta_3 male + \varepsilon_i$
- I want to know whether education, age and gender are jointly affecting health
- $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
- What is H_a?
- Where is F-test in STATA output?
  - Top right panel
  - Like t-stat, F-stat follows a specific distribution, a bigger F-stat means higher power to reject the H_0
  - Prob>F is like the p-value in t-test. We can look at "Prob>F" and compare it with α to decide whether rejecting H_0

```
Number of obs =     46950
F(  3, 46946) =   1272.88
Prob > F      =    0.0000
R-squared     =    0.0752
Adj R-squared =    0.0752
Root MSE      =    1.0959
```

Unbiasedness of OLS

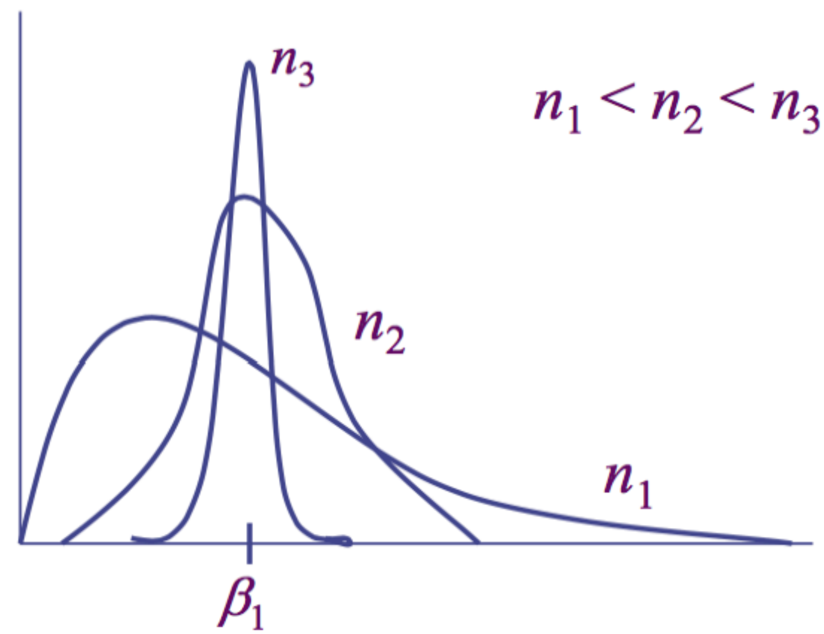
- In the econometric model we have population parameter β, and what we get from data is a sample estimate $\hat{\beta}$. If we draw many samples, we will get multiple $\hat{\beta}$. Those $\hat{\beta}$ form a sampling distribution with a certain mean and standard deviation
- Unbiased estimator:
  - If $E(\hat{\beta}) = \beta$, 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":

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

- $E(\varepsilon)=0$ is always true for OLS estimator, so $Cov(X, \varepsilon)=0$ 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 **Cov(X, ε)=0**.

OLS with control

- Look at multiple regression from another perspective:
- $Y = \alpha + \beta D + \gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_n X_n + \varepsilon$
- We have treated all independent variables equally. Now we give special attention to our variable of interest, ***the treatment variable D***, and consider other variables $\boldsymbol{X_1}$-$\boldsymbol{X_n}$ as ***controls***
- We put $\mathrm{X_1}$-$\mathrm{X_n}$ in the regression, because we can observe them and the data allow us to measure them. Therefore they are also called ***the observables***
- There can be factors that we want to control but are not observable to us (or we cannot measure them in the data). These factors are called ***unobserved variables (unobservables)***. They are represented by the ***error term ε***

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(ε',Insured)=0
- Which variable is now left in ε'?
  - Income
- Is Cov (income, Insured) likely to be 0?
  - No. Therefore $\widehat{\beta'}$ 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

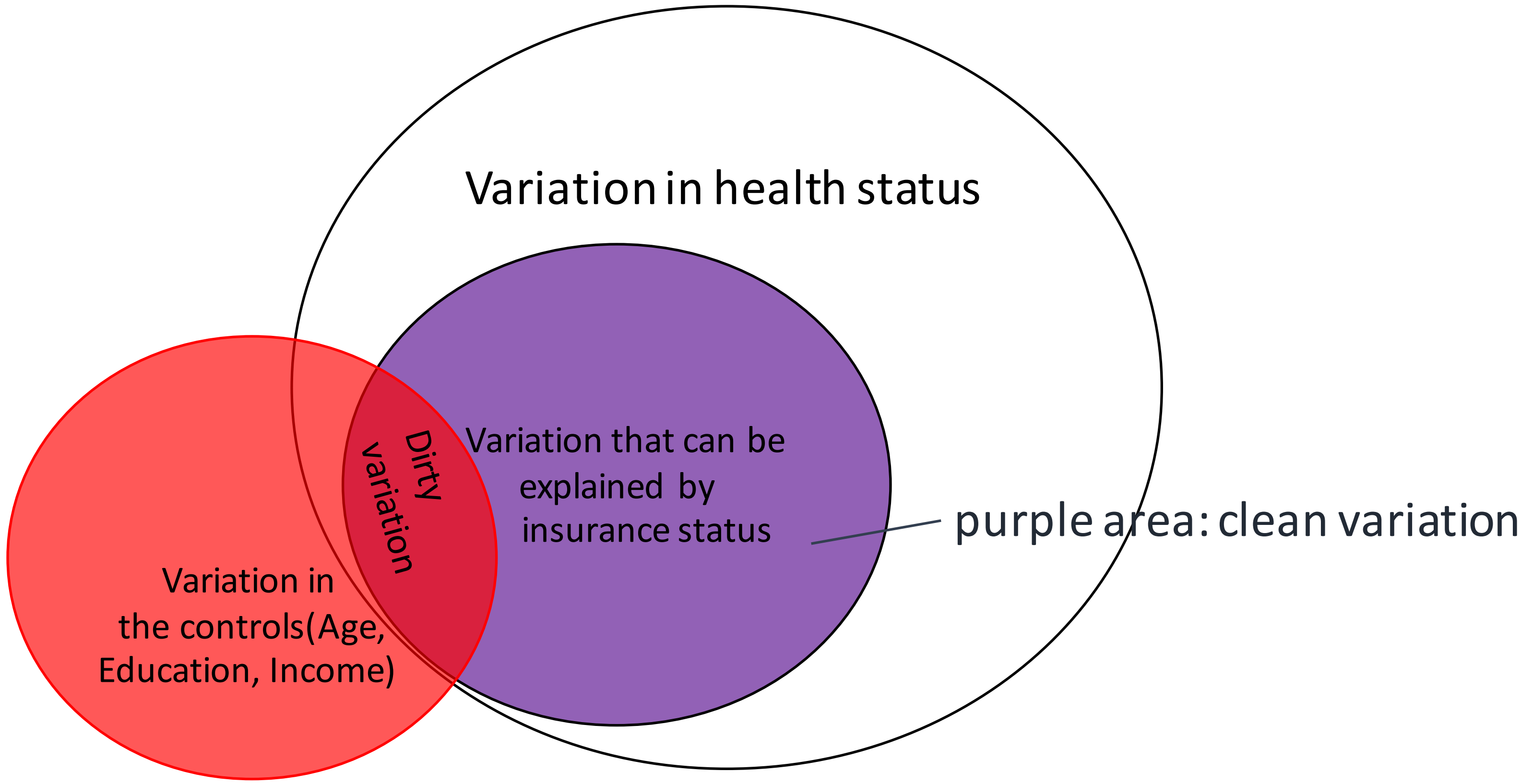
		Corr(omitted variable,x)	
		positive	negative
Corr(omitted variable,y)	positive	upward bias	downward bias
	negative	downward bias	upward bias

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: Immigrantion

- Suppose immigration positively impacts crime rate
- If I regress crime on immigration, I may find a positive effect
- $Crime = \alpha + \beta\ Number\ of\ Immigrants + \varepsilon, \hat{\beta}$>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 Health=$\beta_0 + \beta_1$Less_HS+β_2HS_grad+β_3College_or_above+ε
  - 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:
- Health=$\beta_0 + \beta_1$HS_grad+β_2College_or_above+ε
- 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 $\beta_{0,} \beta_2$

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

- Way#2: omit the constant term
- Health=$\beta_1 \mathrm{Less}HS + \beta_2HS_grad+\beta_3$College_or_above+ε
    - 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

- Know the meaning, null and alternative hypothesis of F-test
- Understand the conditions for unbiasedness and consistency and the relationship between them
- Understand why we need controls-Because cov(D,ε)=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.