AMERICAN UNIVERSITY - CENTRAL ASIA 21.11.2022; 28.11.2022
Econometrics I (ECO 608)
Nurgul Tilenbaeva

Seminar 7

Endogeneity (Part III)

1 Omitted Variables

1. Suppose that, at the elementary school level, the average score for students on a standardized exam is determined by

   $avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u$

   where

   `expend`=expenditure per student;
   `povrate`=poverty rate of the children in the school.

   Using school district data, we only have observations on the percentage of students with a passing grade and per student expenditures; we do not have information on poverty rates. Thus, we estimate β_1 from the simple regression of `avgscore` on `expend`.

   What is the direction of bias in $\tilde{\beta}_1$?

2. Suppose that average worker productivity at manufacturing firms `avgprod` depends on two factors, average hours of training `avgtrain` and average worker ability `avgabil`:

   $avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u$

   Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that `avgtrain` and `avgabil` are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of `avgprod` on `avgtrain`?

3. The following equation describes the median housing price in a community in terms of amount of pollution `nox` (for nitrous oxide) and the average number of rooms in houses in the community `rooms`:

   $log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 rooms + u$

What are the probable signs of β_1 and β_2? Why might `nox` [or, more precisely, $log(nox)$] and `rooms` be negatively correlated? If this is the case, does the simple regression of $log(price)$ on $log(nox)$ produce an upward or a downward biased estimator of β_1?

4. Suppose that you are interested in estimating the ceteris paribus relationship between y and x_1. For this purpose, you can collect data on two control variables, x_2 and x_3. (For concreteness, you might think of y as final exam score, x_1 as class attendance, x_2 as GPA up though the previous semester, and x_3 as SAT or ACT score). Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3.

   - If x_1 is highly correlated with x_2 and x_3 in the sample, and x_2 and x_3 have large partial effects on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain.
   - If x_1 is almost uncorrelated with x_2 and x_3, but x_2 and x_3 are highly correlated, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.

5. Import the Stata data file `"htv"` from the e-course platform. This data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

   - Estimate the regression model

     $educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$

     by OLS. How much sample variation in `educ` is explained by parents' education?
   - Add the variable `abil` (a measure of cognitive ability) to the regression. Does `abil` help to explain variations in education, even after controlling for parents' education? Explain.

6. Let `math10` denote the percentage of students at a Michigan high school receiving a passing score on a standardized math test. We are interested in estimating the effect of per student spending on math performance. A simple model is

   $math10 = \beta_0 + \beta_1 log(expend) + \beta_2 log(enroll) + \beta_3 poverty + u,$

where

`expend`=per student spending;

`enroll`=student enrollment;

`poverty`=percentage of students living in poverty.

- The variable `lnchprg` is the percentage of students eligible for the federally funded school lunch program. Why is this a sensible proxy variable for `poverty`?
- The table that follows contains OLS estimates, with and without `lnchprg` as an explanatory variable.

Dependent Variable: *math10*		
Independent Variables	**(1)**	**(2)**
log(*expend*)	11.13 (3.30)	7.75 (3.04)
log(*enroll*)	.022 (.615)	–1.26 (.58)
lnchprg	—	–.324 (.036)
intercept	−69.24 (26.72)	−23.14 (24.99)
Observations *R*-squared	428 .0297	428 .1893

  Explain why the effect of expenditures on `math10` is lower in column (2) than in column (1). Is the effect in column (2) still statistically greater than zero?
- Interpret the coefficient on `lnchprg` in column (2).
- What do you make of the substantial increase in R^2 from column (1) to column (2)?

7. Import the Stata data file `"wage2"` from the e-course platform. This data set contains information on monthly earnings, education, several demographic variables, and IQ scores for 935 men in 1980.
   - Estimate the regression model

     $log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 south + \beta_6 urban + \beta_7 black + u$

by OLS.

- Our primary interest is in the estimated return to education. If we suspect the omitted variable bias arising from not including cognitive ability in the regression, what is the expected direction of the bias on the coefficient for education?
- Add the variable `IQ` to the regression to account for omitted ability bias. What happens to the estimated return to education? Is the coefficient on `IQ` statistically significant?

8. What are the rules of thumb for including variables in the model?

2 Reversed Causality

1. A study using depression as the dependent variable and smoking as the independent variable of interest finds a statistically and economically significant effect of smoking on depression. What are the potential problems with the conclusions of this study? Explain.

3 Sample Selection

1. Suppose we are interested in the effects of campaign expenditures by incumbents on voter support. Some incumbents choose not to run for reelection. If we can only collect voting and spending outcomes on incumbents that actually do run, is there likely to be endogenous sample selection?

2. We estimated a model relating number of campus crimes to student enrollment for a sample of colleges. The sample we used was not a random sample of colleges in the United States, because many schools in 1992 did not report campus crimes. Do you think that college failure to report crimes can be viewed as exogenous sample selection? Explain.