Version: October 22, 2021 Fall 2021
EC200: Econometrics and Applications

Lab 6: Internal Validity and LPM

Materials

- `cps_2016.dta`
- Do-file template `labtemplate.do`

Objectives

Today we're going to keep working with `cps_2016.dta`, which contains information from the 2016 Current Population Survey.

By the end of this lab, you should be able to complete the following tasks in Stata:

- Think about sample selection issues
- Estimate and interpret linear probability models

Key commands

command	description
`codebook var1`	Look at key details for `var1`
`clonevar var1 = var2`	Make a new variable, `var1` that duplicates `var2` (including labels!)
`_pctile hourwages,per(99)`	Calcualte the 99th percentile of hourly wages, and store as a local variable
`ret list`	Show locally stored variables (handy!)

Linear Probability Models

What happens when our dependent variable is binary? We can use it anyway! Using OLS with a binary dependent variable is called a **linear probablity model**. There is lots of debate about whether (and when) this is an okay idea, as it can lead to predcitions that are below zero or greater than 1, and it violates homoskedasticity assumptions. We can fix the latter by estimating heteroskedasticity-robust standard errors, and the general consensus *seems* to be that usually, we're okay using a LPM. (Though we can do better!)

What about interpretation? We intrepret coefficients are in **percentage points** (not percents!)

Consider the following to see:

$Married_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + u_i$

β_1 means that a 1-year increase in age is associated with a β_1 **percentage-point change** in the probability of being married.

For great slides on this (and a deeper dive), check out this resource!

Lab 6 Worksheet

What do I submit?

- Your written up answers to the exercise questions. This can be typed or written out then scanned (or photographed), in any reasonable format.
- The do-file you've created that runs this analysis
- A log file that contains the results from this exercise.

Exercises

1. Open Stata, start a new do-file (or bring in a template). Make sure you add code to start (and end) a log.
2. Open `cps_2016.dta` and restrict the sample to adults (age 18+) who are married (spouse present or absent). Drop anyone who reports "NIU" (not in universe) for labor force status. Confirm that you have **73,950** observations
3. Check work hours, weeks of work, and wage income for any weird recodes (that is, replace any 999999 values with missing values) generate the following variables, and ensure you have the correct means. You may want to use the `codebook` command to help (i.e. `codebook uhrsworkly`)

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    wkswork1 |     73,950     34.0054     23.5977         0         52
  uhrsworkly |     51,921    40.19379    11.33071         1         99
-------------+---------------------------------------------------------
     incwage |     73,950    38947.58    64901.47         0    1259999
```

4. Generate a binary variable `female` equal to one if `sex == 2`. Estimate the impact of `female` on wage income among married individuals. What is the interpretaion on the coefficient?
5. If our objective is to measure the impact of gender on wage income among married individuals, is sample selection bias likely to be important? Why? Is measurement error likely to be important, why or why not? If so, what is the likely impact of measurement error on your estimated coefficients?
6. Create a binary variable `lf` equal to 1 if an individual is in the labor force, and 0 otherwise. Estimate the impact of gedner on labor force status. What is the interpretation of the coefficient? Estimate the impact of
7. What is the impact of being in the labor force on wage income? Based on this and the previous question, what is the implication for the direction of omitted variable bias when you estimated $incwageNZ = \beta_0 + \beta_1 female + u$? without controlling for it?

8. Re-estimate, including a control for `lf`: $incwageNZ = \beta_0 + \beta_1 female + \beta_2 lf + u$. Was your estimate correct?

9. Now, add your cleaned variable for usual hours worked to estimate $incwageNZ = \beta_0 + \beta_1 female + \beta_2 lf + \beta_3 uhrsworkly + u$. What is the interpretation of each coefficient?

10. Why does your regression not include all 73,850 people? What type of bias might this introduce?

11. Is measurement error likely to be important, and if so, for which variables? What is the likely impact of measurement error on your estimated coefficients?

12. Generate a new variable `uhrsNZ` that recodes all missing work hours values as zeros. You can expedite this with the `clonevar` command. Re-estimate the impact of gender, labor force status and `uhrsNZ` on wage income. What is the interpretation on *each* coefficient? Why did it change?

13. Now, re-estimate but exclude `lf`: $incwageNZ = \beta_0 + \beta_1 female + \beta_3 uhrsworkly + u$. How do your results change? Conditional on including `female` and `uhrsworkly`, does it make sense to include `lf`?

14. Calculate log wages based on `incwageNZ`. Estimate the impact of gender on wage income, including a control for `uhrsworkly`. How does the sample size change, and why? What is the interpretation on each coefficient?

15. Calculate hourly wages, based on the cleaned variables. What is mean hourly wages for men and women?

16. Estimate the impact of gender on hourly wages for those with non-zero hourly wages, controlling for weekly work hours. Repeat then repeat to include all adults (Replace hourly wages with 0 for non-earners) How does the impact of gender on earnings compare?

17. Are there outlier wages? Exclude observations that exceed the 99th percentile in wages and re-estimate both equations. How does this affect your results?

18. Is measurement error likely to affect your dependent variable? Why or why not? If so, what are the implications?