ECON3500 Lab 6: Internal validity and LPM

Materials

- `acs2024_4pct.dta`
- Do-file template `econ3500_lab_template.do`

Objectives

Today we're going to work with `acs2024_4pct.dta`, which contains information from the 2024 American Community 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
- Reason about omitted variable bias and measurement error

Data context

Each row in `acs2024_4pct.dta` is an individual from the 2024 ACS microdata. The file includes demographics, education, labor-force status, work hours, and earnings variables. We will restrict our sample to **married adults** and explore the gender wage gap, labor force participation, and how sample selection affects our estimates.

Tip: use `describe`, `codebook`, and `tab ... , nolabel` to check labels and coding for any variables you plan to use.

Variables we'll use

variable	meaning	notes
`incwage`	wage and salary income	check for topcodes (999999 = N/A)
`sex`	sex	1 = Male, 2 = Female
`age`	age in years	
`marst`	marital status	1 = married spouse present, 2 = married spouse absent
`labforce`	labor force status	0 = N/A, 1 = not in LF, 2 = in LF
`uhrswork`	usual hours worked per week	0 = N/A (did not work last year)
`wkswork1`	weeks worked last year	0 = did not work

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 var1, per(99)`	Calculate the 99th percentile of `var1`, 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 probability model**. There is plenty of debate about whether (and when) this is an okay idea, as it can lead to predictions 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 interpret coefficients in **percentage points** (not percents!)

Consider the following:

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

β_1 means that a 1-year increase in age is associated with a 100*β_1 **percentage-point change** in the probability of being married. So if β_1 is 0.05, that means that being one year older is associated with a 5 percentage point increase in the likelihood of being married.

LPM in Stata

A linear probability model looks exactly like a typical OLS regression — but your dependent variable is **binary (0/1)**:

```
regress lf female, robust
```

The coefficient on `female` tells you the change in the **probability** (in decimal form) of being in the labor force associated with being female. Multiply by 100 to express in percentage points.

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

Lab Video

** Note that this video is from an earlier version of the lab that used 2016 data from the Current Population Survey. Details may vary, but the implelmentation is the same!**

YouTube video

Workflow overview

1. Load the dataset and start your log file.
2. Restrict the sample (married adults only).
3. Inspect and clean variables (`codebook`, `tab`, replace N/A codes).
4. Generate binary indicators (`female`, `lf`).
5. Run regressions, adding controls sequentially and interpreting results.
6. Construct new variables (hourly wages, log wages) and analyze outliers.
7. Answer the worksheet questions about internal validity throughout.

Lab 6 Worksheet

What do I submit?

- Your written up answers to exercise questions (1) - (18). 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.

Use robust standard errors in all regressions.

Example:

```
regress incwage female, robust
```

Questions

1. Open Stata, start a new do-file (or use the template). Make sure you add code to start (and end) a log.
2. Open `acs2024_4pct.dta` and restrict the sample to adults (age 18+) who are married (spouse present or absent). Use `tab marst, nolabel` to identify the correct codes. Confirm that you have **59,039** observations.
3. Check work hours (`uhrswork`), weeks of work (`wkswork1`), and wage income (`incwage`) for any N/A codes. In this dataset, `uhrswork == 0` means "did not work last year" (N/A) — replace these with missing. Also check whether `incwage` has any topcode values (999999). Use the `codebook` command to help (e.g. `codebook uhrswork`). Ensure you have the correct means and number of observations:

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    wkswork1 |     59,039    30.70257    24.59739          0         52
    uhrswork |     37,796    39.18065    12.56177          1         99
-------------+---------------------------------------------------------
     incwage |     59,039    50505.14    84753.23          0     907000
```

4. Generate a binary variable `female` equal to one if `sex == 2`. Estimate the impact of `female` on wage income (`incwage`) among your sample of married individuals. What is the interpretation of the coefficient?
5. If our objective is to measure the impact of gender on wage income among married people, is sample selection bias likely to be important? Why or why not? 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 (`labforce == 2`), and 0 otherwise. Estimate the impact of gender on labor force status. What is the interpretation of the coefficient?

Reminder: This is a linear probability model! Your dependent variable (`lf`) is binary, so interpret the coefficient in percentage points.

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 $incwage = \beta_0 + \beta_1 female + u$ without controlling for labor force participation status?
8. Re-estimate the previous regression, including a control for `lf`: $incwage = \beta_0 + \beta_1 female + \beta_2 lf + u$. Was your prediction in part (7) correct?
9. Now, add your cleaned variable for usual hours worked to estimate $incwage = \beta_0 + \beta_1 female + \beta_2 lf + \beta_3 uhrswork + u$. What is the interpretation of each coefficient?
10. Why does your regression not include all 59,039 people? What type of bias might this introduce?
11. Is measurement error likely to be important in the previous regression, 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, which retains variable labels. Re-estimate the impact of gender, labor force status and `uhrsNZ` on wage income (`incwage`). That is, you're replacing `uhrswork` with `uhrsNZ`. What is the interpretation on *each* coefficient? Why did it change?
13. Now, re-estimate but exclude `lf`: $incwage = \beta_0 + \beta_1 female + \beta_3 uhrsNZ + u$. How do your results change? Conditional on including `female` and `uhrsNZ`, does it make sense to include `lf`?

14. Create a new variable that estimates log wages: `gen l_incwage = log(incwage)`. Estimate the impact of gender on logged wage income, including a control for `uhrswork`. How does the sample size change, and why? What is the interpretation of each coefficient?
15. Using the cleaned variables, calculate hourly wages: `gen hourwage = incwage / (uhrswork * 50)`. We assume that people work 50 weeks in one year. What are mean hourly wages for men and women?
16. Estimate the impact of gender on hourly wages for those with positive hourly wages, controlling for usual hours worked (`uhrswork`). Then, replace missing hourly wages with 0 for those who worked but earned no wages, and re-estimate. How does the impact of gender on earnings compare between the two regressions? Why does the sample size change?
17. Do outliers affect your results? Exclude observations that exceed the 99th percentile in wages based on `incwage`, and re-estimate both equations from the previous question. How do your results change?

Hint:

```
_pctile incwage, per(99)
ret list
```

This stores the 99th percentile value, which you can use to filter observations.

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

Submission checklist - Answers to questions (1)-(18) - Do-file with comments for each question - Log file that matches your do-file commands - Make sure your do-file includes `log close` at the end