**Week 7 Lab Handout- Self-Selection Bias and Working with Panel Data**

**PA 5033 – Multivariate Techniques**

April 29th, 2022

Wenchen Wang

**PART A: SELF-SELECTION BIAS ~20 min**

**PART B: WORKING WITH PANEL DATA ~30 min**

**Part A. Self-Selection Bias**

Retrieve **Union.dta** from CANVAS. Variables used in data set:

**Union** – presence of a union =1; no union =0

**Gr** – growth rate of a firm

**CR** – concentration rate

**Profit** – profit of the firm

**Age** – age of firm

**IPR** – import penetration rate

**Logsales** – amount of sales of a firm

We are interested in what effect unions have on a firm’s profits. We hypothesize that unions self-select into firms that are more profitable or unions may be associated with older firms (unions organized firms in the 1950-70s that are older). This data contains observations of large firms in manufacturing. If unions self-select into certain firms, our results (coefficient estimates) will be biased because the error term will be correlated with the union variable. Therefore, we must correct for this bias using a selection correction method similar to the “Instruments in the classroom” example.

1. Run a linear regression with **profits** as the dependent variable and **union**, **gr, ipr,** and **logsales** as the independent variables.





1. Next we **will** co**rrect for** the **bias. R**un a linear probability regression with **union** as the dependent variable and **Gr, Cr,** and **age** as the independent variables. We are running the union variables on Gr, Cr and age because we believe that these variables theoretically explain union. **BE SURE** to save the predicted values before running the regression.

*Z = f (gr, cr, age)*

*union = f (gr, cr, age)*





1. Create a new variable **lambda** that will be the selection correction coefficient. It will be gained by subtracting **union\_1** (the unstandardized predicted value of union) from **union**.







1. Finally, run a linear regression with **profit** as the dependent variable and **union**, **lambda, gr, ipr**, and **logsales** as the independent variables. **Union** should now be cleansed of selection bias. Note that it is now significant.







* *What are your results? How do the differ from initial OLS model?*

1. Order condition of Heckman Procedure: When selecting for independent variables to explain the variation in union, there has to be at least one independent variable that is not in the final regression

**Part B. Working with Panel Data: Problem 16.4 in Studenmund**

Retrieve the “**Problem-16.4.dta**” file from CANVAS. Variables in the data set are as below:

**State** – U.S. state

**Tax** – Whether or not a cigarette tax was imposed on a given state in 2003; 1 = tax imposed, 0 = no tax imposed

**Yr2000** – Cigarette consumption in the year 2000

**Yr2006** – Cigarette consumption in the year 2006

Problem 16.4 from Studenmund:

In 2003, ten states increased the taxes they placed on cigarettes. Because taxes increase the price of cigarettes, we’d expect that a tax increase would reduce the consumption of cigarettes. In Table 16.2, we present cross-sections of state level data on cigarette consumption for the years 2000 and 2006. Forty-four states plus the District of Columbia are listed here with those states that did not have a tax increase in 2003 listed first.

**a. Would you consider this to be a random assignment experiment data set, a natural experiment data set, or a panel data set? Explain.**

**b. Depending on your answer to part a, use the appropriate estimation technique to determine the impact of the cigarette tax increase on the consumption of cigarettes.**

Since we’re trying to see how a tax changed consumption, the tax can be thought of as a “treatment” to selected states. That is, some states did NOT impose the tax and thus serve as the CONTROL group, whereas ten states DID impose the tax and thus serve as the TREATMENT group.

To see whether or not the tax affected cigarette consumption, we need to measure the CHANGE IN CONSUMPTION of cigarettes over time.

Thus, our model becomes a difference-in-difference model:

ΔConsumption = β0 + β1Tax + ε

To create the proper dependent variable, we simply calculate the difference of “Yr2006” an “Yr2000”, as this is the change in cigarette consumption over time.



Now we simply regress “ConsumpChange” against the treatment variable, “Tax”:





**c. Do these results conform with your expectations? If they don’t, what problems do you see with this research design?**

Retrieve the “**16.4-Panel.dta**” file from CANVAS. Variables in the data set are as below:

**State** – U.S. state

**Tax** – Whether or not a cigarette tax was imposed on a given state in 2003; 1 = tax imposed, 0 = no tax imposed

**Consumption** – Cigarette consumption

**Year** – Year indicator

**ID**: Unique identifier for each state

Set the panel data environment and fun Fixed-effect and Random-effect regression to see how the results compare with before (**Optional content**)