Problem Set 6

ECON 480 - Fall 2020

Theory and Concepts

Question 1

In your own words, describe what *fixed effects* are, when we can use them, and how they remove endogeneity.

Question 2

In your own words, describe the logic of a *difference-in-difference* model: what is it comparing against what, and how does it estimate the effect of treatment? What assumption must be made about the treatment and control group for the model to be valid?

R Questions

Answer the following questions using `R`. When necessary, please write answers in the same document (knitted `Rmd` to `html` or `pdf`, typed `.doc(x)`, or handwritten) as your answers to the above questions. Be sure to include (email or print an `.R` file, or show in your knitted `markdown`) your code and the outputs of your code with the rest of your answers.

Question 3

- `PeaceCorps.csv`

How do people respond to changes in economic conditions? Are they more likely to pursue public service when private sector jobs are scarce? This dataset contains variables at the U.S. State (& D.C.) level:

Variable	Description
`state`	U.S. State
`year`	Year
`appspc`	Applications to the Peace Corps (per capita) in State
`unemployrate`	State unemployment rate

Do more people apply to the Peace Corps when unemployment increases (and reduces other opportunities)?

a. Before looking at the data, what does your economic intuition tell you? Explain your hypothesis.
b. To get the hang of the data we're working with, `count` (separately) the number of `state`s, and the number of `year`s. Get the number of `n_distinct()` `state`s and `year`s[1], as well as the `distinct()` values of each[2].
c. Continuing our pre-analysis inspection, (install, and) load the `plm` package, and check the dimensions of the data with `pdim`.**[3]
d. Create a scatterplot of `appspc` (Y) on `unemployrate` (X). Which State is an outlier? How would this affect the pooled regression estimates? Create a *second* scatterplot that does not include this State.
e. Run two *pooled* regressions, one with the outliers, and one without them. Write out the estimated regression equation for each. Interpret the coefficient, and comment on how it changes between the two regressions.
f. Now run a regression with State fixed effects using the dummy variable method.[4] Interpret the marginal effect of `unemployrate` on `appspc`. How did it change?
g. Find the coefficient for Maryland and interpret it. How many applications per capita does Maryland have?
h. Now try using the `plm()` command, which de-means the data, and make sure you get the same results as Part F.[5] Do you get the same marginal effect of `unemployrate` on `appspc`?
i. Now include *year* fixed effects in your regression, using the dummy variable method. Interpret the marginal effect of `unemployrate` on `appspc`. How did it change?
j. What would be the predicted number of applications in Maryland in 2011 at an unemployment rate of 5%?k. Now try using the `plm()` command, which de-means the data, and make sure you get the same results as Part I.[6] Do you get the same marginal effect of `unemployrate` on `appspc`?
k. Can there still be endogeneity in this model? Give some examples.

[1] Do this inside the `summarize()` command
[2] Don't use the `summarize()` command for this part
[3] Set `index=c("state","year")` to indicate the group and time dimensions.
[4] Ensure that `state` is a factor variable, and insert in the regression. You can either `mutate()` it into a `factor` beforehand, or just do `as.factor(state)` in the `lm` command.
[5] Inside `plm()`, set `index = "state"` to indicate variable, and `model = "within"` to indicate a fixed effects model.
[6] Inside `plm()`, set `index = c("state", "year")` to indicate both variables, and `effect = "twoways"` to indicate a 2-way fixed effects model.

l. Create a nice regression table (using `huxtable`) for comparison of the regressions in E, G, and I.**

Question 4

- `TexasSchools.csv`

Are teachers paid more when school board members are elected "off cycle" when there are not major national political elections (e.g. odd years) than "on cycle?" The argument is that during "off" years, without attention on state or national elections, voters will pay less attention to the election, and teachers can more effectively mobilize for higher pay, versus "on" years where voters are paying more attention. This data comes from Anzia, Sarah (2012), "The Election Timing Effect: Evidence from a Policy Intervention in Texas." *Quarterly Journal of Political Science* 7(3): 277-297, and follows 1,020 Texas school board districts from 2003–2009.

From 2003–2006, all districts elected their school board members off-cycle. A change in Texas policy in 2006 led some, but not all, districts to elect their school board members on-cycle from 2007 onwards.

Variable	Description
`LnAvgSalary`	logged average salary of teachers in district
`Year`	Year
`OnCycle`	`=1` if school boards elected on-cycle (e.g. same year as national and state elections), `=0` if elected off-cycle
`pol_freedom`	Political freedom index score (2018) from 1 (least) top 10 (most free)
`CycleSwitch`	`=1` if district switched from off- to on-cycle elections
`AfterSwitch`	`=1` if year is after 2006

a. Run a pooled regression model of `LnAvgSalary` on `OnCycle`. Write the estimated regression equation, and interpret the coefficient on `OnCycle.` Are there any sources of bias (consider in particular the argument in the question prompt)?
b. Some schools decided to switch to an on-cycle election after 2006. Consider this, `CycleSwitch` the "treatment." Create a variable to indicate post-treatment years (i.e. years after 2006). Call it `After`. Create a second, *interaction* variable to capture the interaction effect between those districts that *switched*, and *after* the treatment.
c. Now estimate a difference-in-difference model with your variables in Part B: `CycleSwitch` is the treatment variable, `After` is your post-treatment indicator, and add an *interaction* variable to capture the interaction effect between those districts that *switched*, and *after* the treatment. Write down the estimated regression equation (to four decimal places).
d. Interpret what each coefficient means from Part C.
e. Using your regression equation in Part C, calculate the expected logged average salary (Y) for districts in Texas:

- i. *Before* the switch that did *not* switch
- ii. *After* the switch that did *not* switch
- iii. *Before* the switch that *did* switch
- iv. *After* the switch that *did* switch

f. Confirm your estimates in Part E by finding the mean logged average salary for each of those four groups in the data.[7]
g. Write out the difference-in-difference equation, and calculate the difference-in-difference. Make sure it matches your estimate from the regression.
h. Can we say anything about the types of districts that switched? Can we say anything about all salaries in the districts in the years after the switch?

[7] Hint, `filter()` properly then `summarize()`.

i. Now let's generalize the diff-in-diff model. Instead of the treatment and post-treatment dummies, use district-and year-fixed effects and the interaction term. Interpret the coefficient on the interaction term[8]

j. Create a nice regression table (using `huxtable`) for comparison of the regressions in (a), (c), and (i).

[8]This is doable with the dummy variable method, but there will be a *lot* of dummies! I suggest using `plm()`.