Problem Set 4

Due Date: 11/12/21

Questions

Problem 1: Unweighted vs. Weighted Average

Using the Titanic dataset, answer the following parts. You can find information on the dataset variables here

a) Write down in expected value notation the estimator for a simple difference in means for this data set. Assume that we are interested in the effect of being in first class on survival.

b) Using R, estimate the estimator that you wrote down in part a

c) Assume that assignment might be confounded by gender and age. Write down the new estimator that we are interested in to test the effect of being in first class on survival.

d) Calculate the weighted average treatment effect by doing the following. i) Define a variable `s` that takes on four values: 1 for a male child, 2 for a male adult, 3 for a female child, and 4 for a female adult. Define a child as an individual under 15. Compare the estimator in (b) to your weighted ATE. Which one do you think is more likely to be correct and why?

Problem 2

Read Hyde (2007). Then answer the following questions:

a) Based on lectures and readings about natural experiments, evaluate the plausibility of this research design as a natural experiment.

For the next set of questions, use the `HydeData.csv` dataset.

b) Replicate Table 1

c) Replicate Table 3

d) Replicate Table 4

e) Replicate Table 5

For each table, explain in your own words what the table means and how a reader should evaluate it in context of Hyde's research design and claims.

Problem 3

Read Jakiela (2021) available here. The dataset for this problem is `JakielaData.csv.`

a) What are Jakiela's two diagnostics? In your own words, why are they helpful in assessing the potential bias of the two-way fixed effects estimator?

b) Estimate two TWFE models. In both models, use country and year fixed effects and cluster by country. In your first model, regress the number of enrollees in primary school on the treatment. In your second model, regress the number of enrollees in secondary school on the treatment.

c) Did these countries pass laws at different times? To assess this, create a new variable in your dataset called `lengthTreat` that is the number of years between the first year of implementation and the last year of the dataset for each country. Make a bar graph (`geom_col`) of this variable.

d) Subset your data into two data frames, one for primary schools and one for secondary schools. For each data frame:

- *i*) Get the residuals of the regression of treatment on the fixed effects.
- *ii*) Append the residuals to the data frame. Add a column for treatment weights calculated as in Equation (2) of Jakiela's paper.
- *iii*) Show that the coefficient of treatment is equal to the sum of the outcome multiplied by treatment weights
- *iv*) Make a histogram (`geom_histogram()`) of the weights. Are any weights negative?
- *v*) Run a regression of the outcome variable on the fixed effects. Get the residuals of this regression. Plot the residuals of the treatment against the residuals of the outcome. Use the color argument in `aes()` to differentiate the points based on whether they are treated or untreated.
- *vi*) Statistically test if the slopes of the two groups are the same by running a regression of the outcome residuals on the treatment residuals and the interaction of the treatment residuals and treatment. For either model, is the interaction term significant? If it is, what does that mean, according to Jakiela?

e) Conceptually, in your own words, what robustness checks does Jakiela recommend and why?