Program Analysis and Evaluation (ECON 672)
University of Maryland
Winter 2022
Prof. Laura Kawano

Problem Set 2
Due February 3, 2022, 6:45pm

Your problem set should be uploaded to ELMS by the start of class (6:45pm) on February 3. Late problem sets will not be accepted.

Your problem set should consist of two portions: (1) the answers to the questions to the questions; and (2) a Stata log file that shows how you got to your answers to the empirical questions. Your written answers should be typed. All answers should be concise. Your log file must be clear, and include comments that make it easy to find the commands leading to each answer.

Problem sets not turned in using the format described will receive no credit.

Data description

The data for this problem set come from an experimental program within the National Job Training Partnership Act (JTPA), the National JTPA Study (NJS). The JTPA was a large, national labor program that ran from 1982 to 2000. The experiment was conducted in 1989.

The file name is econ672_njs_data.dta.

The variables included in the dataset are:

bifid: Identification number
ra_stat: Random assignment status, 1 = treatment and 2 = control
enroll: Participation status, 1 = participated in JTPA and 0 =did not participate in JTPA
esum18i: Self-reported earnings 18 months after random assignment
sex: Gender, 1 = male and 0 = female
race: 1= white, 2= black, 3=Hispanic, 4 = other
age: Age in years at baseline interview
totch18: Number of children age 18 or under
bfeduca: Years of schooling at baseline interview

bfyrearn: Self-reported earnings in the year before random assignment
site_num: Categorical value coded for each experimental site, with values 1-16

1. Read in the data and summarize it.

2. Create a variable *treatment* that equals one for those who are assigned to the treatment group and zero for individuals assigned to the control group. Drop observations with missing values of earnings 18 months after random assignment. Drop treatment group members who did not enroll in JTPA from the data. The remaining data should consist only of participants and controls.

3. Create three new indicator variables: (1) *miss_child*, that equals one when the number of children 18 or under is missing and zero otherwise; (2) *miss_educ*, that equals one when education before the experiment is missing and zero otherwise; and (3) *miss_earn*, that equals one when earnings before the experiment is missing and zero otherwise. After creating these variables, replace *totch18, bfeduca*, and *bfyrearn* to take the value of 0 when the values are missing. (You should check to make sure that all observations now have values for each of these variables, or they will be dropped from the regressions.)

4. Using OLS, estimate the mean impact of the experimental assignment to the treatment group on self-reported earnings 18 months after random assignment. Do not include additional covariates in the model. Interpret your findings. What is the estimated impact? Is it statistically different from zero?

5. Repeat the regression analysis in (4), now including as covariates sex, race, age, total number of children age 18 or younger, years of schooling at baseline, self-reported earnings before random assignment, site number, and indicator variables for missing children aged 18 or younger, missing years of schooling before random assignment, and missing self-reported earnings before random assignment. (I will refer to this list of covariates will be the "baseline covariates".) Describe and interpret your findings. Be sure to relate these estimates to your previous answer. [Hint: think about how best to include race and site number into the regression. What would make sense in for interpretation in a regression analysis?]

Propensity Score Matching

The results below come from nearest neighbor matching estimation with replacement with the ten nearest propensity scores. The propensity scores were generated via a probit model using the baseline covariates.

```
. psmatch2 treatment, outcome(esum18i) pscore(pscore) neighbor(10)
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
esum18i	Unmatched	9836.70573	9274.31172	562.394015	183.609092	3.06
	ATT	9836.70573	9242.54394	594.161791	193.129997	3.08

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support On suppor	Total
Untreated	4,549	4,549
Treated	6,263	6,263
Total	10,812	10,812

6. How do the estimates from propensity score matching differ from the linear regression estimates? Account for any differences. Note that the standard errors reported by psmatch2 do not account for the estimation of propensity scores.

Difference-in-differences

7. Construct the difference-in-differences estimate of the impact of the JTPA participation on earnings using participants and the control group. Using Stata's regress command and use the difference between *esum18i* and *bfyrearn* as the dependent variable. Produce one set of difference-in-differences results including no covariates, and another set of results conditioning on all of the baseline covariates other than *bfyrearn.* Be mindful of how each covariate should enter the regression. Be sure to report both the estimates and their standard errors.

8. Interpret the estimates obtained in the preceding problem. Compare them to the estimates obtained by doing nearest-neighbor matching. Are they different from the matching estimates? Should they be?

Regression Discontinuity

9. Keep only observations from the control group that also have non-missing values of *bfyrearn.*

10. Suppose that a treatment was administered to all members of the control group with earnings in the year before random assignment less than or equal to $2650. Create and indicator variable called *rdtreat* equal to one for control group members who received the (hypothetical) treatment and equal to zero otherwise.

11. Construct a regression discontinuity impact estimate by taking the difference in means of *esum18i* for the control group members within $500 on either side of $2650. Be sure to provide an estimated standard error for your impact estimate.

12. State precisely what parameter the impact estimate in the preceding problem estimates.

13. Does the estimate that you found in (11) surprise you? Why or why not?

14. Construct two separate regression discontinuity estimates by estimating separate linear regressions of *esum18i* on *bfyrearn.* The first regression model should use all of the data, and the second should use only those within $500 of the cutoff, as in problem (11). For both estimates, you should let the parameters differ on each side of the discontinuity.

15. Are the estimates found in the preceding problem surprising? Explain why or why not. Which one do you prefer? Why?