Homework 4

STUDENT NAME

Due 7/25 11:59pm

NOTE: If you would like to answer the written portions using RMarkdown, simply type your answers below the questions. If you would like to print out the pdf to do the written portions, you should add `\vspace{50mm}` to add space in the knitted pdf for written work (adjust the 50mm as necessary). Alternatively, you can do the written portions on a separate paper.

Part 1: Theory

1

Use the following set of "data" describing the potential outcomes for a treatment to answer the following questions. D_i tells you whether that individual was actually treated. In an actual dataset, you would only observe Y_i^1 for treated individuals or Y_i^0 for untreated individuals; here you are provided with both so you can calculate exact values.

Individual	D_i	Y_i^1	Y_i^0
1	1	7	4
2	1	8	4
3	0	5	5
4	1	6	7
5	0	3	4
6	0	4	3

1.1

What is the difference in means of the *observed* outcomes (Y_i^1 for treated individuals, Y_i^0 for untreated) between the treated and untreated groups?

1.2

What is the ATE for this treatment?

1.3

What is the ATT for this treatment?

1.4

What is the ATU for this treatment?

1.5

In class I claimed that

$$\frac{1}{3}\sum_{i|D_i=1} Y_i - \frac{1}{3}\sum_{i|D_i=0} Y_i = ATE + [\mathbb{E}(Y_i^0|D_i=1) - \mathbb{E}(Y_i^0|D_i=0)] + (1-P(D_i=1))(ATT-ATU)$$

The middle term (in square brackets) is what we called the selection bias. What is the selection bias in this example?

2

Consider a randomized controlled trial intended to investigate whether the free provision of condoms reduces the transmission of sexually-transmitted diseases. The trial is conducted on a set of university dorms, where each resident of the dorm is assigned to either treatment (receives free condoms) or control (receives nothing). The residents are then asked six months later whether they have contracted any sexually-transmitted diseases in the past six months.

2.1

Name a potential violation of the Stable Unit Treatment Value Assumption (SUTVA) that could affect this RCT.

2.2

The surveyor conducting this randomized controlled trial collected lots of demographic information on the participants: their age, nationality, gender, blood type, religious affiliation, historical number of sexual partners, whether or not they smoke or drink, what their major is, and what clubs they are a part of. They conduct a balance test on all these variables and find that the treatment group is statistically significantly more likely to smoke than the control group. Does this indicate something went wrong during treatment randomization? Why or why not?

2.3

In the context of this RCT, what would it mean to be an always-taker? A never-taker? A complier?

2.4

The most basic regression you could use to analyze this RCT would be $STD_i = \beta_0 + \beta_1 D_i + u_i$, where STD_i is an indicator for whether individual i contracted a sexually-transmitted disease and D_i is an indicator for treatment. If you were interested in whether the treatment is more or less effective for students that drink (as opposed to students that do not drink), what would you add to this regression?

Part 2: Application

In this section, you will be using a subset of the data used in the paper "Randomized experiments from non-random selection in U.S. House elections" by David S. Lee.

It is well known that incumbent politicians (politicians currently holding an office) are more likely to win their elections. One obvious reason for this is *selection*: politicians who have already demonstrated an ability

to win an election in the past likely won because they have attributes that help with winning elections, like charisma or popular policy stances. These factors should help them again in future elections.

This paper seeks to find out whether this is the whole story, or if maybe there is something *purely* about incumbency, totally separate from the individual candidate's quality, that helps to win re-election.

The paper uses a regression discontinuity strategy. The idea is that politicians who just *barely* win their elections are not too different from politicians who just *barely* lose, but that only those that barely win enjoy incumbency in the next election. The running variable `difdemshare` is the share of votes going to the democratic party in election t minus 50 (so the running variable is $= 0$ at the cutoff, since getting $> 50\%$ of the votes is necessary to win). The dependent variable is `demsharenext`, the share of the vote that the democratic party gets in the next election.

Load the `tidyverse package` and the data, which is called `lee_data.csv`.

1

Create a histogram showing that the density of elections is not noticeably different in the immediate vicinity of the cutoff. Since we are most interested in observations around the cutoff, show only the density of elections with `difdemshare` within 0.05 of the cutoff (0). Does there appear to be manipulation to you?

2

Again using only observations where `difdemshare` is within 0.05 of the cutoff, plot the scatterplot with `difdemshare` on the x axis and `demsharenext` on the y axis. Does it look to you like there is a discontinuity at $difdemshare = 0$? Just eyeballing it, what does it look like the size of the treatment effect is?

3

Create an indicator for treatment and use it to estimate a linear regression discontinuity model. Use the `fixest` package so you can get heteroskedasticity-robust standard errors. Report the results.

4

According to the model you estimated in the last problem, what is the effect of being the incumbent party on your vote share in the next election?

5

Now estimate a quadratic model. How does the estimated treatment effect compare to the linear model?

6

The dataset also includes a few controls. `demelectexp` and `othelectexp` count the number of previous elections the Democratic candidate and other candidate (usually, but not always Republican) have participated in previously. `demofficeexp` and `othofficeexp` is defined similarly, but only counting elections that were won. Do we need to control for these to have unbiased estimates? Does controlling for them provide any other benefits?

7

Re-estimate the regression from problem 5 using all the controls. How does it affect your parameter estimate of interest?

8

Using the model estimated in problem 7, give the 95% confidence interval for the estimated treatment effect.