Problem Set 4- Regression Discontinuity & Difference-in-Differences

MaCCS 201 - Fall 2025

2025-11-10

Tentative Due Date: Before Thanksgiving.

Please submit markdown file named [last_name]_[first_name]_ps4.Rmd or a pdf with all code and answers.

#Part A: Difference in Differences.

Let's replicate the OG. Card and Krueger. We have been talking about them forever. Plus the consequences of raising the minimum wage will be forever policy relevant. The dataset is in Stata form, which is easy to import (I will show you below). It is called cardkrueger.dta.

The interviewed Fast food restaurants in two waves in New Jersey and Pennsylvania 3/1992 and 11/1992). In the middle the minimum wage in NJ went up from $4.25 to 5.05 per hour but did not change across the border. You will DiD to see whether you detect an effect of raising the minimum wage. If you see a 1 or a 2 at the end of a variable that indicates which survey wave it is from.

The variables you need are:

- state: `NJ=1`, `PA=0`
- `wage_st` / `wage_st2`: Starting wage at the restaurant
- `fte` / `fte2`: Full-time equiv. employment = #(Full time employees) + #(Part-time Employees)/2. Excludes managers.
- `chain`: which fast food chain you are dealing with (there are 4.)
- `co_owned`: = 1 if restaurant is company-owned, =0 if franchised
- `sample`: Dummy variable = 1 if wage and employment data are available for both survey waves

1. Dump all observation for which `sample=0`, so you have balance.
2. Create a variable `treated` which equals one if the state if NJ and zero otherwise.
3. Create a dummy called `after`, which equals 1 if the observation is from round 2.
4. Calculate the difference in average (across stores) starting wages before and after for each state and then calculate the difference in difference by differencing the two. What do you get?
5. Do the same, but with average full time employment. What do you get?
6. Now set your data up for proper DiD estimation (stacked or long format). This is one of the most pain in the neck (seemingly simple but in practice annoying feats.) Instead of having observations "next" to each other, you want the round 1 observations to be the top block of rows and round 2 observations be the bottom block of rows. You should have the indicator `after`, which is now 0 for the before and 1 for the after periods as a nice column. If you struggle, text me (925) 360-6473.
7. Run a difference in difference regression on these data. First use `wage` as the outcome. Then use full time eployment as the outcome. The unit of observation is the store here!
8. Do the same thing as in 7, but control for whether the store is a franchise or not. What do you see?
9. Cluster your standard errors by state. Then cluster by store. What do you see?

Part B: Regression Discontinuity (Fuzzy)

The Effect of Tutoring Eligibility on Student Performance. The data are on github for 1000 students. This is a new problem I wrote. It worked for me, but if you are encountering Gremlins, reach out!

A local school district offers free after-school tutoring to students who score less than **600 points** on a standardized placement test.
Students with scores above the cutoff are *eligible*, but participation is voluntary.
Some eligible students do not take up the offer, and a few students above the cutoff manage to enroll through appeals.

You have access to a sample of 10,000 students simulated in the file `fuzzy_rd_tutoring_cutoff600.csv`. For each student, you observe:

Variable	Description
`test_score`	Standardized placement test score (running variable)
`eligible`	Indicator = 1 if `test_score`$\leq$ 600
`tutoring`	Indicator = 1 if the student participated in tutoring
`final_score`	End-of-year standardized exam score (outcome)
`prior_gpa`	Prior year GPA (continuous)
`parent_college`	Indicator = 1 if at least one parent attended college
`female`	Indicator = 1 if student is female
`household_income`	Annual household income in dollars

---

1. Plot `tutoring` against `test_score` using a binned scatterplot (e.g., 100 bins) and add a vertical line at the cutoff = 600. Describe what you see. Is there a visible discontinuity? What does this represent in terms of program design?

2. Plot `final_score` against `test_score` in the same way. Does the pattern suggest a treatment effect?

3. Count observations in bins and look for evidence of bunching. Is there? [You could look up what a McCrary test does and run it, but this is not required.]

4. Estimate the discontinuity in the probability of receiving tutoring at the cutoff using `rdrobust(y = tutoring, x = test_score, c = 600)`. Report the estimated jump, its standard error, and the bandwidth used. Explain in words what this first-stage coefficient measures. Why is it less than 1? What would happen if the discontinuity were exactly 1? What design would that correspond to?

5. Estimate the discontinuity in `final_score` at the cutoff, ignoring treatment status. Report the coefficient and interpret it substantively. Why is this not yet the causal effect of tutoring participation?

6. Estimate the causal effect of tutoring on final scores using the fuzzy RD option:

   `rdrobust(y = final_score, x = test_score, c = 600, fuzzy = tutoring)` Report the estimated treatment effect and 95% confidence interval. Interpret the result as a **Local Average Treatment Effect (LATE)**. Who are the "compliers" in this setting? Compare the magnitude and precision of this estimate to the reduced-form effect.

7. Using `ivreg` from the **AER** package, estimate

$$\text{final\_score}_i = \beta_0 + \tau\,\text{tutoring}_i + f(\text{test\_score}_i) + \gamma' Z_i + \varepsilon_i$$

   instrumenting `tutoring` with `eligible`. Include a cubic polynomial in `test_score` and the controls

`prior_gpa`, `parent_college`, `female`, and `household_income`. Report the 2SLS estimate and compare it to the local fuzzy RD estimate.