PubPol713/Educ714
Causal Inference for Education Policy: Postsecondary Winter 2025
Prof. Kevin Stange
University of Michigan
Assignment 1[1]
Workshop: February 4, 2025
Due: February 6, 2025

In this assignment, you will replicate some of the main findings from:

> Solis, A. (2017). Credit access and college enrollment. *Journal of Political Economy, 125*(2), 562- 622.

We will see this paper a little later in the course when we discuss student loans.

This problem set gives you the opportunity to practice regression discontinuity designs, and to improve Stata programming skills. You will consider some of the research design choices inherent in an RD analysis, appropriate robustness checks of findings, and the use of graphics to communicate findings. If you would like a refresher on RDD, (re)read:

Bloom, H.S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*, 43-82.

Murnane & Willett, Chapter 9
http://ebookcentral.proquest.com.proxy.lib.umich.edu/lib/umichigan/detail.action?docID=57 8791

You may work in groups of up to 3 students including yourself, but make sure to submit assignments individually in Canvas. Type your answers and produce nice-looking tables and charts. Convert your write-up to PDF, making sure to include your do file code at the back. Include the names of all group members on page 1, or preferably, in a header on all pages. Save a log file with your final model specifications and results that we can refer to if there are any questions later.

You should skim the paper to get an idea of the context and some of the details. A few important things to note: Eligibility for the loan program is determined by the test score (psut1) being greater than or equal to 475 and also the student not being in the highest (i.e. 5th) income quintile (qqt1).

NOTES: The tables and columns referenced in this problem set reflect the PUBLISHED version of the paper. This is uploaded to the assignment in Canvas.

[1] We are grateful to Fernando Furquim for assignment materials. Source data for this assignment can be found at the Solis (2017) article page https://www.journals.uchicago.edu/doi/abs/10.1086/690829.

Assignment questions

1. Generate three variables for time period 1:

   a. An indicator that flags students scoring above the cutoff of 475
   b. An indicator for "pre-selected" status. Pre-selected students are those for whom income quintile is less than 5
   c. A running or forcing variable centered at the cutoff score of 475

2. Calculate some descriptive statistics and describe what you find:

   a. For time period 1, how many individuals are "pre-selected?" What proportion of PSU takers are pre-selected? What proportion of those scoring below/above the cutoff are pre-selected?
   b. Summarize and plot the distribution of the forcing variable (or of the PSU score) for time period 1. Briefly describe what you find. Do you see any evidence of bunching or manipulation around the 475-point threshold?
   c. Based on time period 1, calculate the rates of immediate enrollment (*enrolt1*) and ever enrollment (*everenroll1*) for 3 groups: non-pre-selected students and, among pre-selected students, those above and below the 475-point PSU cutoff. Do this only for observations that have a non-missing value for PSU in time period 1.
   d. Calculate the rate of immediate and ever enrollment for all students by family income quintile (again, among those with a value for PSU in time period 1).

3. There are two key assumptions to a regression discontinuity analysis: (1) the likelihood of being assigned to the treatment varies discontinuously through the cutoff; and (2) characteristics that are associated with the outcome of interest change smoothly through the cutoff. Present evidence (figures and/or tables) of assumption (2) by analyzing the distribution of scores across family income quintiles. Briefly discuss your findings. See the bottom panel of Figure A1 in the paper for an example. Also discuss any remaining sources of bias that your RD analysis cannot rule out.

4. Replicate columns 1 and 2 of Table 3, where the outcome is immediate college enrollment. Put these findings in a nice, clear table with all necessary information including bandwidth used. Briefly explain the relevant coefficient(s) in each column.

5. The loan eligibility rule lends itself to a "sharp" RD specification in the short term. However, the fact that individuals may retake the test and become eligible in later years introduces some "fuzziness" to the treatment assignment. Use a 2SLS RD setup where exceeding the threshold in year 1 is an instrument for *evereligl* to replicate Table 4 columns 1 and 2. Report first-stage results as well. What do you infer from column 1-2 results? Is this consistent with your estimates from question 4?

6. One of the nice features of this paper is that the RDD findings so clearly show the main result. Create your own version of Figure 1. [Just to warn you, you almost certainly will NEVER get an RD graph that looks this clean!]

7. This final question asks you to estimate a series of placebo effects to gage the size of the enrollment discontinuity at a score of 475 relative to other discontinuities at irrelevant scores.

a. First, estimate the Table 3 column 1 specification for every value of τ between 431 and 519. That is, substitute placebo values of τ in the equation (1) term $\boldsymbol{1}(T_i \geq \tau)$. Use a 44-unit bandwidth as in the main results, but note that this bandwidth will cover different PSU values in each placebo estimate. Store coefficients for the $\boldsymbol{1}(T_i \geq \tau)$ indicator (β_1), and plot the distribution of placebo coefficients. Mark where the true effect of $\tau = 475$ lies in the distribution of placebo effects. What share of placebo effects are smaller in absolute value than the true β_1?

b. Repeat part 7a, but for the Table 3 column 2 specification.

Notes and hints

- Variables are labeled intuitively. Please contact Kevin with any questions.

- There are canned routines in Stata to generate RD estimates and plots (rdrobust being the most useful). I'd like you to do the analysis using basic regression tools (regress and ivregress) and graphic commands. You can also use rdrobust and compare against yours.

- There are several variables that you don't need for this problem set, but that you may want to explore if you want to try to replicate additional results.

- Getting question 1 right is crucial. There are particular variables for time periods 2 and 3 that you can use as reference. They are (m475t2, m475t3); (pre_sel1, pre_sel2); and (psu475t2, psu475t3). HINT: you should get 230,653 pre-selected observations and the mean of the running variable should be 14.55.

- The way to constrain your regression to your desired bandwidth is to use ifs in your regression command.

- Figure 1 might be time consuming. The process for creating these RD figures generally is:
  - Collapse the data into bins for the running variable (and any other variable you care about such as pre-select status)
  - Run a regression on the collapsed data
  - Create a chart with two components:
    - A scatterplot of the observed outcome and bin
    - Lines for the predicted values of the regression, including confidence interval
    - Note that the lines should be differentiated by the cutoff score of the variable.
    - If you just want a plain linear fit, you can skip the regression and tell Stata to plot a best linear fit through a scatterplot. Figure 1 uses a quartic polynomial. Based on your findings, decide whether that's worth the trouble!