## LPO 7870 Research Design and Data Analysis II, 2024

**Assignment 5**

**Submission Guidelines:**

* **Please submit an electronic copy of your group’s solutions as a PDF file by 8:00 pm, Wednesday, April 24.**
* **All submissions via Brightspace in the “Assignments” section —only one submission per group.**
* Please include the Stata output (you can copy and paste the main Stata outcomes).
* Pro-tip: if you use Courier New font, Stata output will line up as it does in the output window.

**Question 1. Regression Discontinuity Design.**

Suppose you are asked to design and conduct a study aimed at estimating the impact of a multi-year high school career and technical education (CTE) program. For this program, students are recruited in Grade 8 to participate in CTE-focused courses during Grades 9–12. Students also participate in work-based learning activities and receive other support services while in high school. Because of limited capacity, students must apply for admission. Several factors are combined into a single score for each student that applies. Suppose that all students who score above a threshold score are offered spots in the program. You are interested in whether the program increases the probability of graduating from high school on time for students who participate.

The Stata dataset *data\_rdd\_s1.dta* (posted on Brightspace and Github) includes observations for 4,993 students and contains the following variables:

**grad1**, the outcome variable. Equal to 1 for students who graduate on time and zero otherwise.

**appscore1**, the rating variable (or running variable). This is the application score for students who applied to the program, rescaled so that the score is equal to 0 at the threshold of admission.

**enroll1**, an indicator for enrolling in the program. Equal to 1 for students who enrolled in the program and zero otherwise.

**pretest1**, the student’s score on a Grade 8 standardized test (before the intervention)

1. **Plot the raw data.** Create a scatterplot with *appscore1* on the x-axis and *grad1* on the y-axis. Is this figure helpful in evaluating the feasibility of a regression discontinuity design? Why or why not? (4 points)
2. **Plot the data using bins.** Install the *binscatter* command by typing *ssc install binscatter*. This command will create a scatterplot using “bins”—that is, it will divide the X variable into equal-width bins and plot the average of the Y value for each bin. For example the command: *binscatter Y X, nq(20) linestyle(none)* will create a scatterplot with 20 bins. Try this with at least three different bin counts (more bins=narrower bins). Do any of these plots show a visual discontinuity at the test score threshold? (4 points)
3. **Check for manipulation of the rating variable (the running variable)**. An RD design would be invalid if there were manipulation of the running variable (e.g., if students or administrators had influence over scores). Perform some visual checks for manipulation (9 points):
   1. Create a histogram of appscore1.
   2. Install the *rddensity* command using *ssc install rddensity*. This command estimates the density of the running variable around the threshold and tests for manipulation. Use this command with *appscore1*: *rddensity X, c(0) plot*. Note c(0) is telling Stata the threshold value.
   3. Based on your results above, is there any evidence that appscore1 has been manipulated? Explain your reasoning.

1. **Create an “eligible” variable.** Create a variable called *eligible* that equals 1 if the student’s score makes them eligible for the CTE program and 0 otherwise. Based on this variable (and the *enroll1* variable) determine whether this is a sharp or fuzzy RD. Explain your answer. (4 points)
2. **Check for balance before the intervention.** (5 points)
   1. Perform a balance test on the *pretest1* variable.
   2. How do you read the results? Please discuss in light of the main assumption for RD.
3. **Estimate the treatment effect**. Estimate the effect of participating in the high-school CTE program on the likelihood of graduating on time. (24 points)

* 1. Begin by using the full range of data (i.e., all students). Estimate the effect of the program on *grad1* using a linear regression that includes an intercept shift and a change in slope at the threshold. Interpret your results. Is the effect statistically significant?
  2. Repeat part (i) but narrow the bandwidth so that you only include students within +/-10 points from the threshold. Does this change your conclusion?
  3. Now repeat part (i) but use a quadratic function of *appscore1* (i.e., *appscore1* and *appscore1* squared). Be sure your regression still includes an intercept shift and a change in slopes at the threshold. Interpret your results. Is the effect statistically significant?

1. **Run a falsification test**. Check that the program has no effect on a variable it should not affect. (5 points)
   1. Repeat the linear regression from part f(i), but use *pretest1* as the outcome instead of *grad1*. Is there evidence of a jump at the cutoff?