Applied Microeconometrics I, Fall 2023

Problem Set 5

Due: 13th of October, 23:59 (Friday midnight)

Note that the deadline is strict and late submissions will not be accepted.
Submit your answers in one pdf file via MyCourses. Follow the word limit of the answers.

1. Credit access and college enrollment

Students from richer families are more likely to attend and graduate from university than students from poorer families. The gap may not only reflect differences in readiness for college, but also differences in access to credit. However, empirically studying whether credit constraints affect college enrollment is a challenging task for at least two reasons: 1) data on access to credit is usually not observed and 2) even when it is observed, there are many other unobserved factors that affect college enrollment and are correlated with access to credit, leading to biased estimates.

The paper by Solis (2017) overcomes these methodological challenges by exploiting eligibility rules for two college loan programs in Chile. These programs provide access to loans to students who come from the four poorest income quintiles and who score above 475 points on the national college admission test. The loans provide an amount that covers around 90% of the tuition costs.

1.1 How would you use the setting where only students scoring above 475 points are eligible for loans to study the effect of credit access on college enrollment?

We can do sharp RDD, where the running variable is students' score, the threshold is 475 (everyone above is eligible and everyone below is not). We can use the setting to instrument for enrolment at the threshold (which in general is endogenously chosen by people). Here one can also mention the assumptions needed to identify the effect.

1.2 In equation (1) from the paper, Y_i is college enrollment immediately after high-school graduation, and T_i is the score and τ is the 475-point eligibility threshold. How would you interpret β_0 and β_1? What is the running variable?

$$Y_i = \beta_0 + \beta_1 \cdot \mathbf{1}(T_i \geq \tau) + f(T_i - \tau) + \xi_i, \qquad (1)$$

beta_0 is the expected value of Y for students just below the threshold; beta_1 is the increase in the expected value for Y for students just above the threshold (what we called "tau" at slide 26 of lecture 10: the treatment effect at the threshold – the difference in the value of Y just above vs. just below the threshold).

1.3 Suppose you want to study the effect of credit access on college enrollment with an RD identification strategy using this set-up. How would you convince the reader that your design is valid in the sense that there is no selection at the 475-point cut-off? If you had the data on the following pre-determined student characteristics: high-school GPA, gender and mother's years of education, what kind of validity tests would you run?

We can discuss whether in principle people are likely to manipulate the threshold (e.g., do they know the actual threshold value?). With the characteristics mentioned, we can test whether they are balanced across the threshold: we run equation 1 three times with each of the characteristics mentioned as outcome variable. Ideally, we want to see no jump (beta_1 = 0 or close to 0). Alternatively, we can re-estimate (1) using the main outcome of interest (as in 1.2) by additionally controlling for all 3 pre-determined characteristics. The estimated beta_1 coefficient should not change when adding the characteristics.

1.4 The eligibility conditions are public knowledge. Does this pose a threat to the validity of the regression discontinuity design? How can you test for whether it is really a concern?

The fact that the eligibility conditions are public knowledge may pose a threat to the "no perfect manipulation" assumption (people can exactly put themselves just at the right vs. left of the threshold).

1.5 Open the data set *credit_access.dta* (adapted from Solis 2017). The data set contains information on the students' scores and whether they enrolled in college.

Create the following variables: a dummy variable for being above the threshold (*dummy_above)*, a variable that calculates the distance between an individual's given score and the threshold (*dif)* and a variable that gives the interaction between these two (*interact)*.

See attached do-file.

1.6 Run equation (1) above for a bandwidth of 44. Since $f(T_i-\tau)$ is linear, it takes the following form:

$$f(T_i - \tau) = \phi_0 \cdot (T_i - \tau) + \phi_1 \cdot (T_i - \tau) \cdot \mathbf{1}(T_i \geq \tau)$$

This means that the outcome is regressed on *dummy_above, dif* and *interact*. To restrict to observations within the given bandwidth, add an if statement: *if abs(dif) <= 44.* Finally, use

robust standard errors (, *robust* at the end of your regression command). Can you replicate the results in Table 3, column 1? How do you interpret your results?

See do-file for code. The coefficient on beta_1 is .175, which means that scoring above the threshold increases the probability of enrolling in college immediately after high-school by 17.5 percentage points.

1.7 How sensitive are your results to wider bandwidths of 22, 88, and 220 ? Interpret the patterns that you find.

See do-file for regressions.

Bandwidth	*Coefficient*
22	*.184 (.009)*
88	*.189 (.004)*
220	*.238 (.003)*

The closer we are to the threshold, the closer the estimates are to the baseline estimates, but they are slightly less precise. This is a great RDD setting where there are many observations around the threshold, so the loss in precision is not very large. The farther away from the threshold we go, the more biased the coefficients are but the more precision we have. This is what we called the precision-bias trade-off in the lecture notes (slide 23, lecture 10).

1.8 Solis (2017) also estimates the effect of access to credit on the probability of ever enrolling in college. However, as the author highlights, looking at a longer horizon (i.e. not at immediate enrollment as above) comes with the issue that students may self-select into treatment by retaking the test in subsequent years and scoring at or above the eligibility cutoff in those later attempts. Assuming you had data on all subsequent test attempts, how would you adjust the RD strategy to account for this behavior?

We could run a fuzzy RDD, with the first stage given by scoring above the threshold on any test attempt. The second stage regresses the outcome (ever enrolled) on eligibility as estimated in the first stage.

1.9 Open the data set *credit_access2.dta.* This data contains a new outcome variable, *everenroll1*, which takes the value 1 if those that took the test ever enrolled in college (regardless of whether they scored above 475 on their first attempt), and 0 otherwise. It also contains a variable *everelig1* which takes the value 1 if a student is eligible for loans in any admission process (i.e., if a student scores above the cutoff in any test attempt after being classified in one of the four poorest income quintiles), and 0 otherwise. Use these

two variables to implement the strategy you outlined in 1.8 (use the 44 bandwidth). Can you replicate the results in Table 4, column 1? How do you interpret your results? How do these results differ from those in 1.6?

See do-file for the regression. The coefficient on beta_1 is .154, which means that scoring above the threshold increases the probability of ever enrolling in college by 15.4 percentage points, which is slightly smaller than what we estimated in 1.6. See also the discussion in the paper by Solis.