Problem Set 1

Francisco Brady

4 Feb 2025

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

```
. clear all

. set graphics off

. * load data
. use solis_dataset.dta, clear

. * create threshold crossing variable --
. * based on getting a score greater than 475 on PSE test
. gen m475t1 = (psut1 >= 475)

. label variable m475t1 "Indicator for scored above cutoff in year 1"

. *tab m475t1
. * pre-selected -- in 1-4 income quintiles
. gen pre_sel1 = (qqt1 <= 5 & qqt1 ~= .)

. label variable pre_sel1 "Pre-selected indicator for year 1"

. *tab qqt1 pre_selt1, missing
. * create centered psu in t1 score
. gen centered = psut1 - 475

. *hist centered
. gen psu_taker1 = ~missing(psut1)

. * tab psu_taker1, missing
```

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

```
. tabstat psut1 centered enrolt1 everretakepsu2, statistics(mean sd)
columns(statistics)
```

Variable	Mean	SD
psut1	489.5556	105.0928
centered	14.5556	105.0928
enrolt1	.3432197	.4747846
everretake~2	.2527354	.4345809

```
. tab qqt1, missing
```

Income quintile for year 1	Freq.	Percent	Cum.
1	108,442	22.82	22.82
2	49,779	10.48	33.30
3	36,982	7.78	41.08
4	35,450	7.46	48.54
5	29,147	6.13	54.68
.	215,365	45.32	100.00
Total	475,165	100.00	

For the entire sample, the mean PSU score is 489.55. By income quintile, over 50% of the sample are in the first and second quintiles, with over 45% missing a quintile assignment in year 1. In the overall sample, 34% enrolled in college in time period 1.

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?

```
. tab pre_sel1 m475t1
```

Pre-selected indicator for year 1	Indicator for scored above cutoff in year 1		Total
	0	1	
0	126,179	89,186	215,365
1	94,964	164,836	259,800
Total	221,143	254,022	475,165

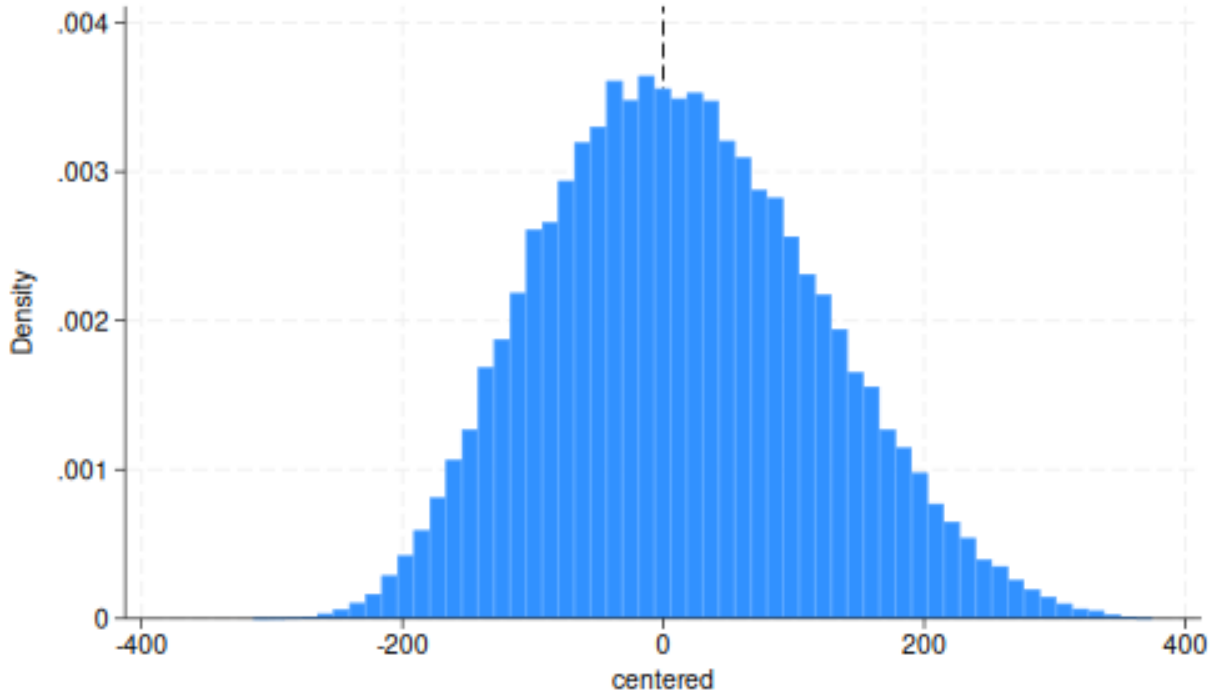
259,800 students are pre-selected. 164,836 are preselected and over the cutoff.

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?

```
. sum centered
```

Variable	Obs	Mean	Std. dev.	Min	Max
centered	475,165	14.5556	105.0928	-314.5	375

```
. hist centered, xline(0)
(bin=56, start=-314.5, width=12.3125)

. graph export psu_hist.png, width(500) replace
file psu_hist.png saved as PNG format
```



The distribution of scores looks fairly normal.

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.

```
. tab enrolt1 if pre_sel1 == 0

Enrolled in |
```

college in t=1	Freq.	Percent	Cum.
0	167,038	77.56	77.56
1	48,327	22.44	100.00
Total	215,365	100.00	

```
. tab everenroll1 if pre_sel1 == 0
```

Ever enrolled flag	Freq.	Percent	Cum.
0	143,016	66.41	66.41
1	72,349	33.59	100.00
Total	215,365	100.00	

```
. tab enrolt1 m475t1 if pre_sel1 == 1 & enrolt1 == 1, row
```

Key
frequency row percentage

Enrolled in college in t=1	Indicator for scored above cutoff in year 1		Total
	0	1	
1	11,031	103,728	114,759
	9.61	90.39	100.00
Total	11,031	103,728	114,759
	9.61	90.39	100.00

```
. tab everenroll1 m475t1 if pre_sel1 == 1, row
```

Key
frequency row percentage

Ever	Indicator for scored above cutoff in year

enrolled flag	1		
	0	1	Total
0	75,283	40,874	116,157
	64.81	35.19	100.00
1	19,681	123,962	143,643
	13.70	86.30	100.00
Total	94,964	164,836	259,800
	36.55	63.45	100.00

Non-preselected enrolled in T1: 48,327 (22.4%)
Non-preselected ever enrolled in T1: 72,349 (33.59%)
Pre-selected enrolled in T1, below threshold score: 11,031 (9.61%)
Pre-selected enrolled in T1, above threshold score: 103,728 (90.39%)
Pre-selected ever enrolled, below threshold score: 19,681 (13.70%)
Pre-selected ever enrolled, above threshold score: 123,962 (86.30%)

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).

```
. tabstat enrolt1 everenroll1, by(qqt1) statistics(mean sd)
columns(variables)

Summary statistics: Mean, SD
Group variable: qqt1 (Income quintile for year 1)
```

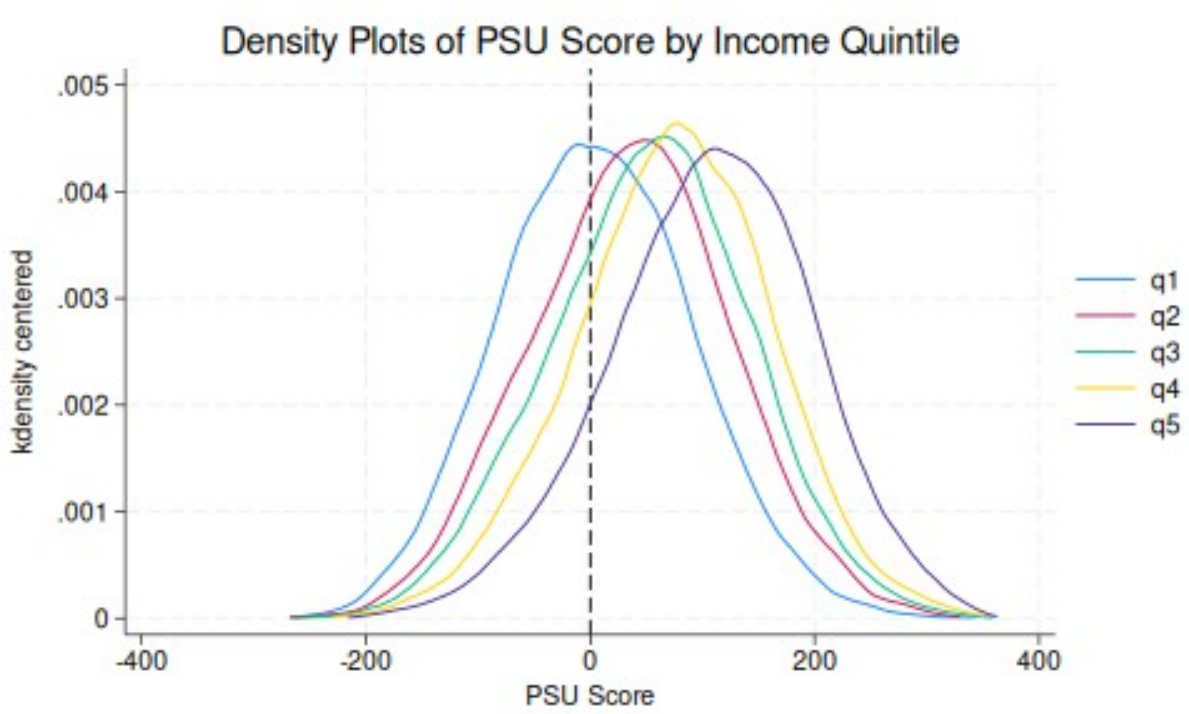
qqt1	enrolt1	everen~1
1	.3525663	.4363715
	.4777713	.4959372
2	.4570401	.5564997
	.498156	.4968025
3	.5227138	.6405008
	.4994906	.4798601
4	.5700987	.7029055
	.4950688	.456985
5	.4883521	.6866916
	.4998729	.4638466
Total	.4417206	.5528984
	.4965928	.4971948

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.

To show support for the assumption that treatment assigment varies through the cutoff, run a t-test across the two groups, above and below the score threshold, and whether the student is pre-selected.

```
. ttest pre_sel1, by(m475t1)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf.
interval]
---------+--------------------------------------------------------------------
       0 |
221,143    .4294235    .0010526     .494995    .4273604    .4314866
       1 |
254,022    .6489044     .000947    .4773137    .6470483    .6507606
---------+--------------------------------------------------------------------
Combined |
475,165    .5467574    .0007222    .4978095     .545342    .5481729
---------+--------------------------------------------------------------------
    diff |           -.2194809    .0014124               -.2222491
-.2167127
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -
1.6e+02
H0: diff = 0                                     Degrees of freedom =
475163

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff >
0
```

```
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) =
1.0000

. twoway  kdensity centered if qqt1==1 || ///
> kdensity centered if qqt1==2 || ///
> kdensity centered if qqt1==3 || ///
> kdensity centered if qqt1==4 || ///
> kdensity centered if qqt1==5 ||, ///
> legend(order(1 "q1" 2 "q2" 3 "q3" 4 "q4" 5 "q5")) ///
> xtitle("PSU Score") xline(0) ///
> title("Density Plots of PSU Score by Income Quintile")

. *graph export psu_income.png
```



To address assumption 2, the figure above shows PSU scores by family income quintile. Across the eligibility threshold (475, the density is smooth for all income quintiles. The continuity across the threshold gives us confidence that there is no bunching just above the cutoff by different income quintiles, which would be an indication of manipulation of the scores.

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.

```
. * col 1
. qui reg enrolt1 m475t1 centered if pre_sel1 == 1 & abs(centered) < 44, r

. eststo presel_linear

. qui reg enrolt1 m475t1 centered if pre_sel1 == 0 & abs(centered) < 44, r

. eststo nonpresel_linear

.
. esttab
```

	(1) enrolt1	(2) enrolt1
m475t1	0.160***	0.000402
	(26.56)	(0.07)
centered	0.00273***	0.00187***
	(22.88)	(16.54)
_cons	0.214***	0.154***
	(66.08)	(47.99)
N	84196	61994

```
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

.
```

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 `everelig1` 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 gauge 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 $1(T_i \geq \tau)$. 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 $1(T_i \geq \tau)$ 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.