31E00910 Applied Microeconometrics I, Fall 2021

Problem Set 5

Model answers

Due: Friday October 22 at 22:00

Note: All papers cited in the problem set are available in the webpage of the course (mycourses), in the folder "references". The datasets can be found in the folder "assignments".

1. **College cost and time to complete a degree (Garibaldi et al, 2012)**

For many students enrolled in academic programs around the world, it takes longer than the normal completion time to obtain a degree. Garibaldi et al. (2012) use a regression discontinuity design on data from *Bocconi University* in Italy in order to study how an increase in tuition would affect the probability of late graduation. Upon enrollment in each academic year, *Bocconi* students are assigned to one of twelve tuition levels on the basis of their family income, assessed by the university administration through the income tax declaration of the student's family and through further inquiries. A regression discontinuity design (RDD) can then be used to compare students who, in terms of family income, are immediately above or below each discontinuity threshold. The authors focus on students in the last regular year of the program, exploiting the fact that their current tuition is a good predictor of the tuition they would pay if they stayed in the program one more year. Thus, students on the two sides of a discontinuity threshold in the last regular year, have paid different tuitions and should expect to keep on paying different tuitions in the following year if they do not graduate on time. Using this source of identification, the authors show that if the official tuition assigned to a student in the last regular year were to increase by 1,000 euros, the probability of late graduation would decrease by 5.2 percentage points (with respect to an observed probability of 80%).

1.1 Explain intuitively what are the crucial assumptions that allow giving the regression discontinuity results a causal interpretation. Explain also whether you expect them to hold in this particular case. *[max. 100 words*

Ansewr: The crucial assumption for the validity of the regression discontinuity design is that there are no discrete changes in any relevant variable at the threshold, other than the treatment (i.e. the amount of tuitions payed). The biggest threat to validity here is that, if families can anticipate the threshold, they might try to manipulate their income. Given that these thresholds were determined ex-post (after the income tax was paid), this is unlikely to be the case.

1.2. How can you verify empirically whether there is manipulation? Explain explicitly what would you check in this particular case. *[max. 100 words]*

Answer: First, we should verify whether the density function is continuous at the discontinuity threshold (we can use the McCrary test or any other similar test). Second, we should verify that all relevant factors evolve "smoothly" with respect to the running variable at the threshold. For instance, we should expect students just above and below the threshold to have similar educational performance in the past, similar socio-economic background, and so on. (For these tests we can use the same rdd machinery that we use for the main analysis.)

2. **Academic probation and student achievement**

In many countries, universities use academic probation as a tool to ensure that enrolled students achieve minimum academic standards. Typically, academic probations work in the following way: If a student's GPA is below a certain standard, she is placed in probation that implies that she must reach a university-specific GPA in the following year or she will be suspended from the university for a year. The hope is that this policy will incentivize the students to perform better next year. However, it is unclear whether this policy is effective in improving student performance because some students who are placed in probation may be discouraged drop out of university altogether.

Lindo et al (2010) is an example of a paper that tries to study this question with data from an unnamed Canadian university where students who fail to meet an explicit GPA threshold in the first year of university studies are placed on academic probation in year 2. In this exercise we try to replicate some of their analysis with a sample from their data.

2.1 Can we simply compare the year 2 GPA and the drop-out probabilities of students who are in probation and who aren't to draw conclusions about the causal effect of probation policy on student outcomes? [Max 50 words]

Answer: Students who are placed on probation will probably differ from students who are not probation in many ways that are not observable for us and are likely to affect GPA. Therefore, it is not feasible to draw conclusions on the causal relationship between these variables just based on this comparison of means.

2.2 Open the data set *lindo_graphs.*dta. These data are aggregated by bins of first year GPA of students measured as a distance from the probation GPA cutoff so that the cutoff takes value zero. In the dataset you have the following variables:

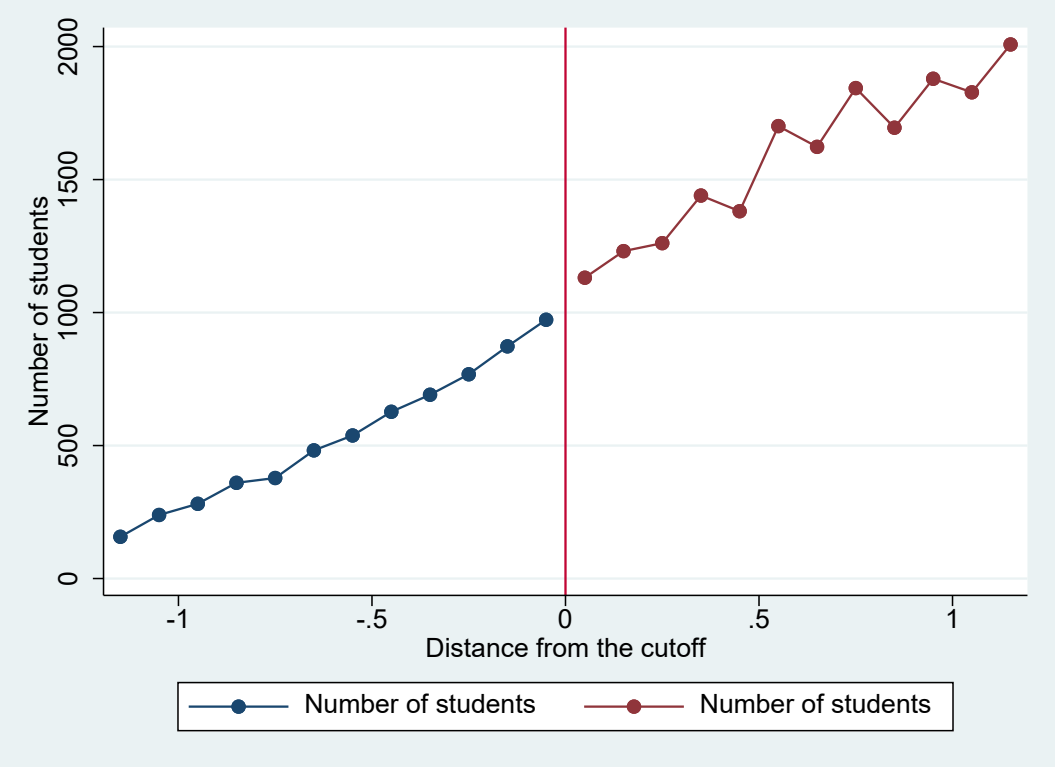
dist_from_cut_round	Distance from the GPA cutoff
freq	Number of students
hsgrade_pct	High school GPA percentile
age_at_entry	Age at entry to the university
male	Dummy for male
probation_year1	Dummy for being placed in probation in year 1
left_school	Dummy for leaving school after year 1
nextGPA	GPA in year 2
gradin	Dummy for graduating in year 4

Now, plot the number of students as a function of the distance from the GPA cutoff. In STATA you would write something like this:

twoway (scatter freq dist_from_cut_round if dist_from_cut_round<0, connect(l)) (scatter freq dist_from_cut_round if dist_from_cut_round>0, connect(l)), xline(0)

What does this graph tell you about the suitability of this setting for studying the effect of probation on student outcomes with a regression discontinuity strategy? [Max 50 words]

Answer: Here is the plot

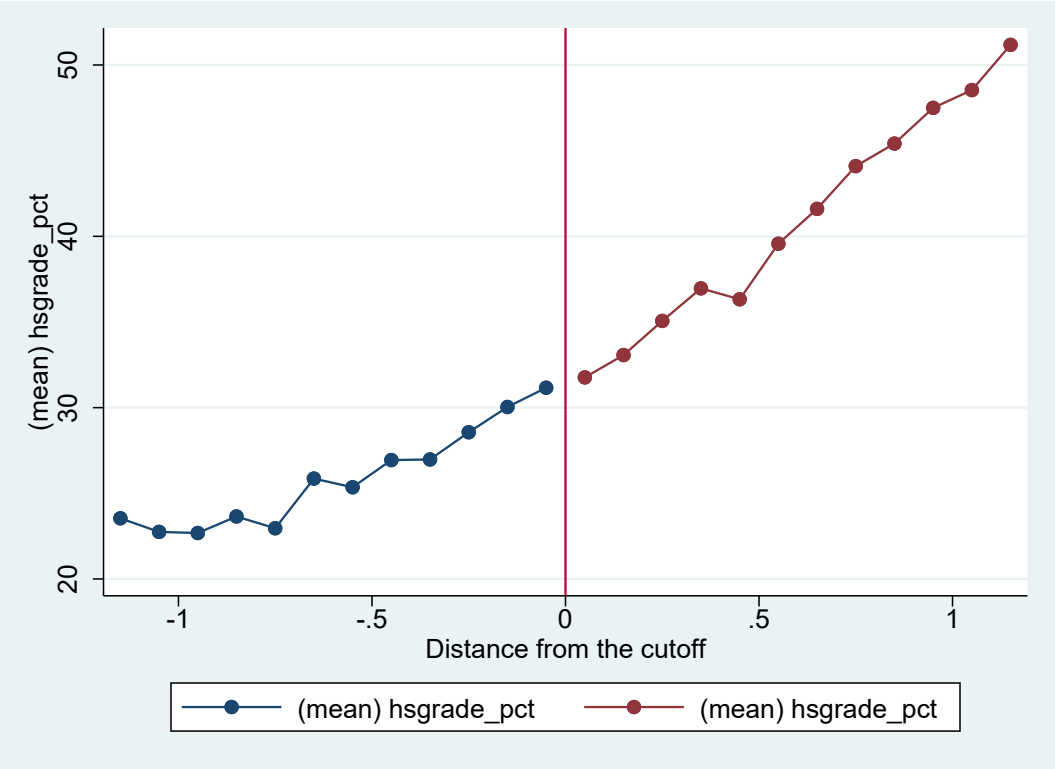


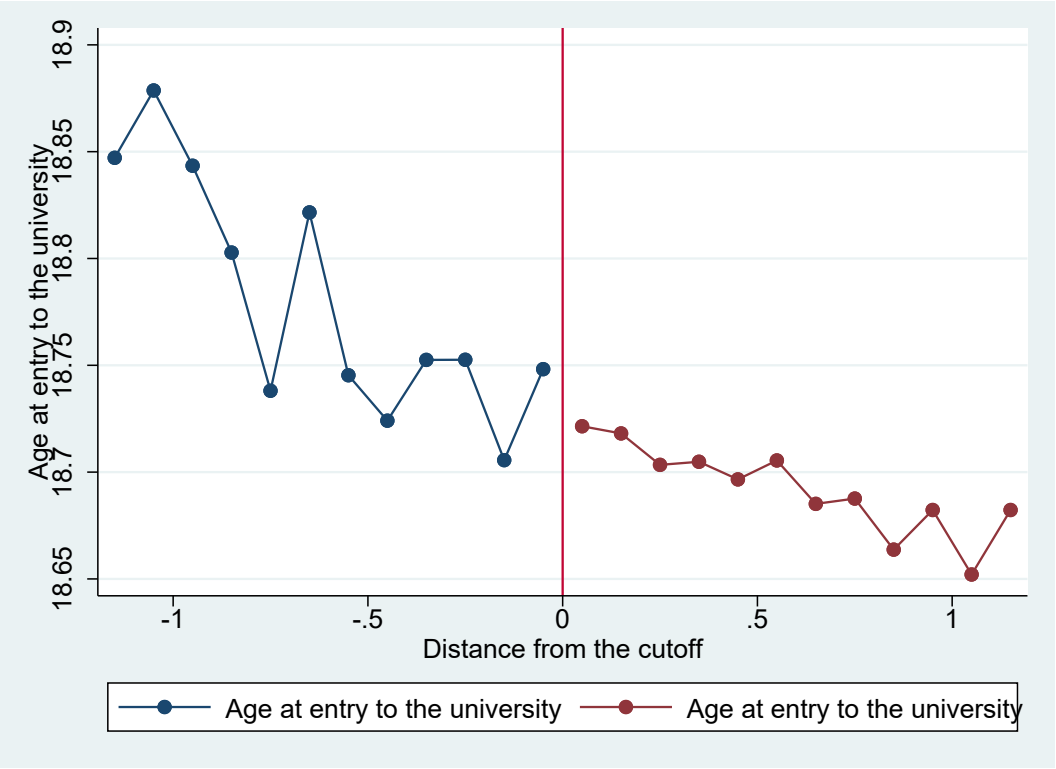
The relationship between the running variable and the number of students seems to be smooth around the cutoff which suggests that the students are not able to manipulate their GPA precisely enough to end up on either side of the threshold. Therefore, this looks like a suitable setting for RDD.

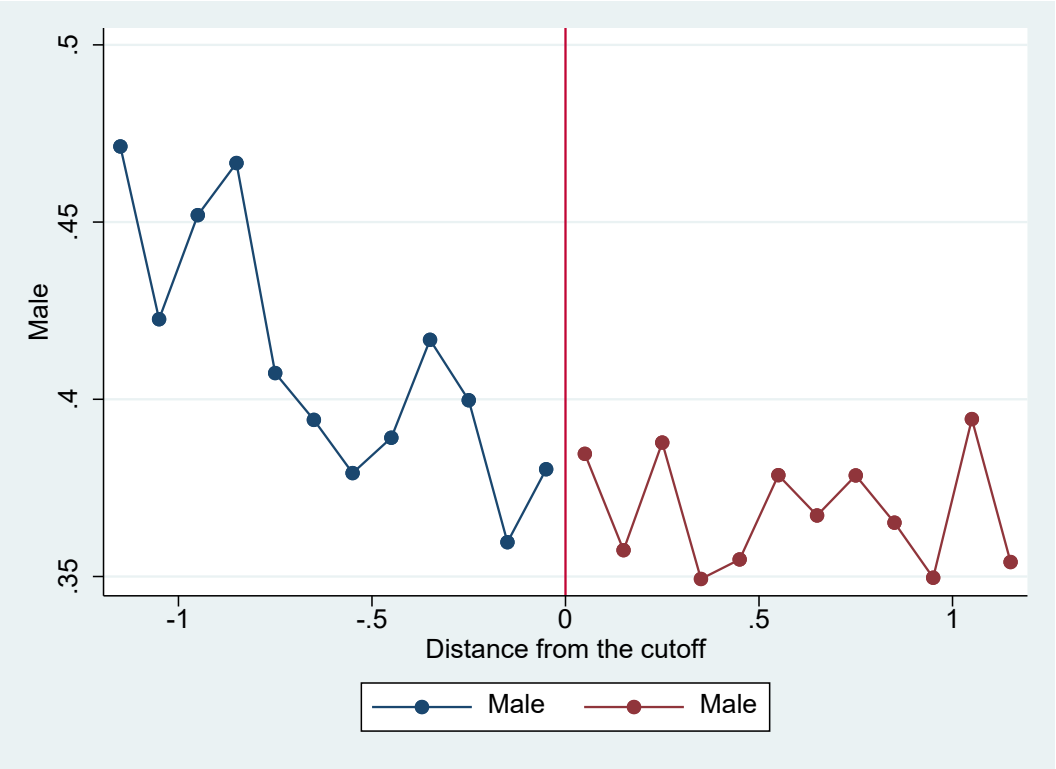
2.3 Variables *hsgrade_pct*, *age_at_entry*, and *male* are pre-determined background characteristics of students. Now, plot these variables, one-by-one, as a function of the distance from the GPA cutoff in the same way as you did in exercise 2.2. What do these graphs tell you about the suitability of this setting for studying the effect of probation on student outcomes with a regression discontinuity strategy? [Max 50 words]

Answer:

Here are the figures


(mean) hsgrade_pct
20
30
40
50
-1
-.5
0
.5
1
Distance from the cutoff
(mean) hsgrade_pct
(mean) hsgrade_pct

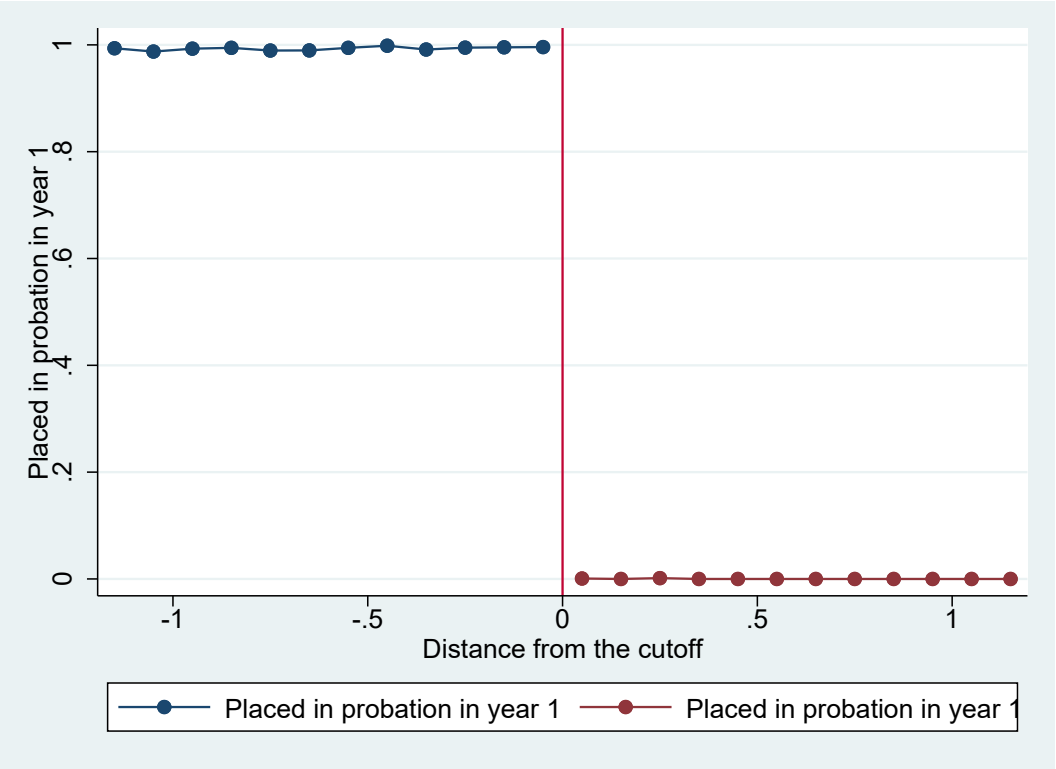

Age at entry to the university
18.65
18.7
18.75
18.8
18.85
18.9
-1
-.5
0
.5
1
Distance from the cutoff
Age at entry to the university
Age at entry to the university



Especially high school grades are very smooth around the cutoff. There is more volatility in age at entry and gender but the jumps at the cutoff are not larger than jumps elsewhere. Again, this looks lime a promising setting for RDD.

2.4 *probation_year1* is obviously the causal variable of interest here. Now, plot this variable as a function of the distance from the GPA cutoff. Does this look like a sharp or a fuzzy regression discontinuity setting? [Max 50 words]

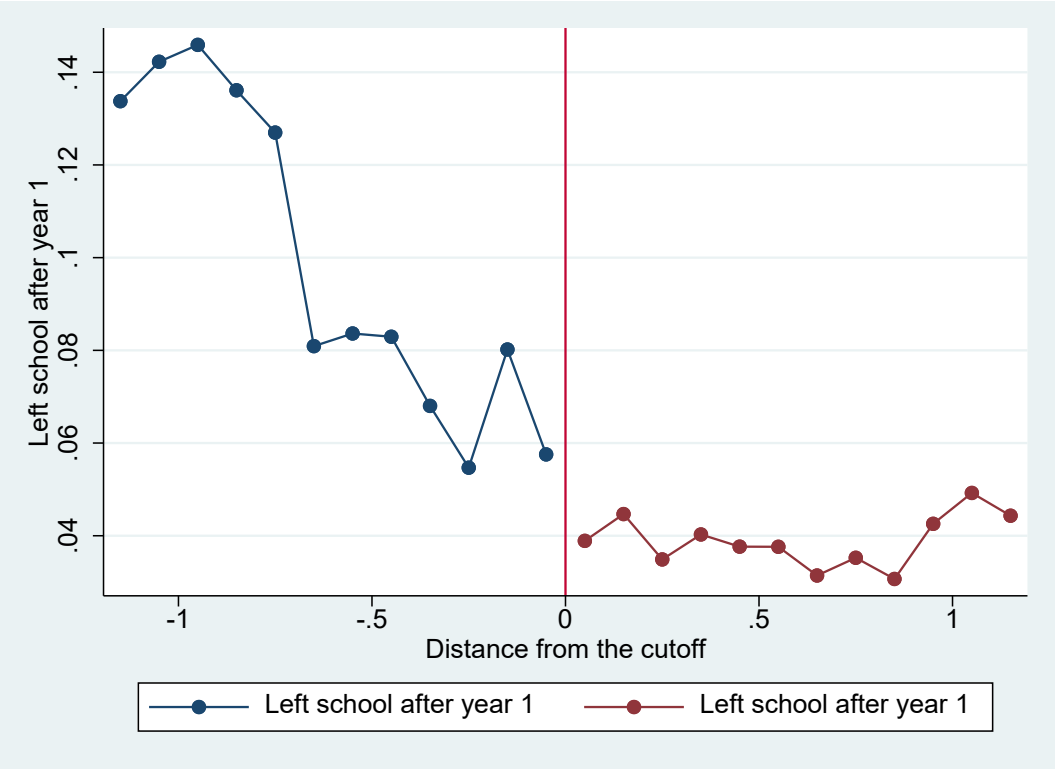
Answer: Here is the figure:

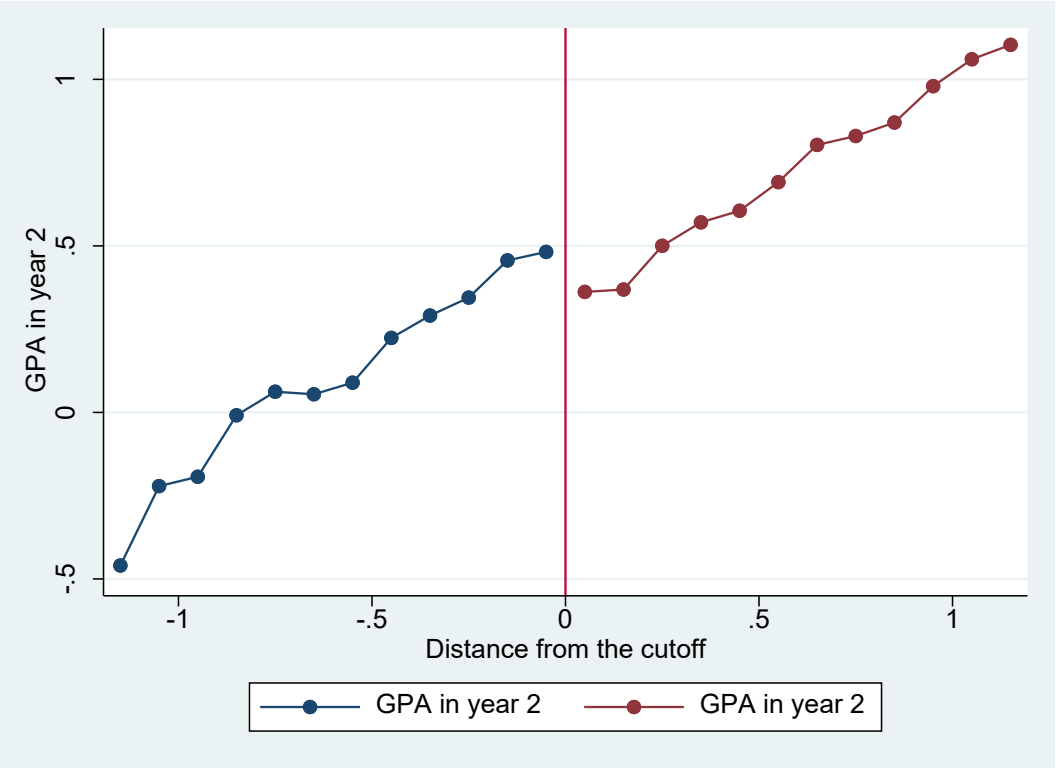


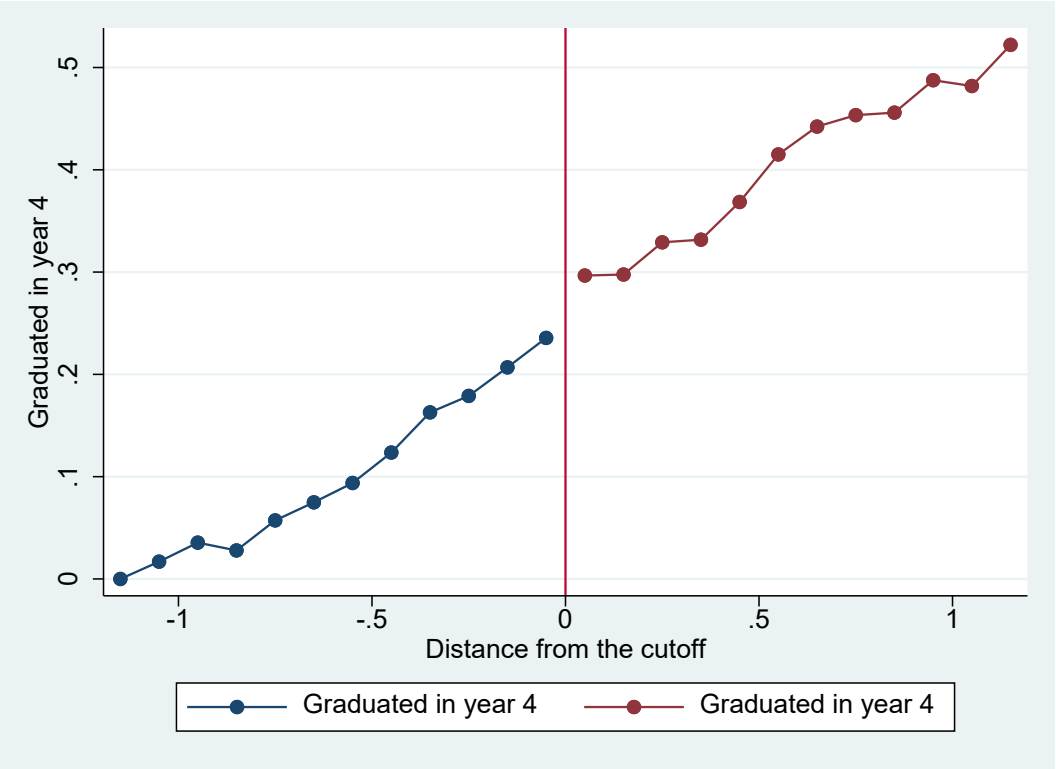
This looks like a sharp design.

2.5 *left_school*, *nextGPA*, *gradin4* are the outcome variables that we are interested in. Plot also these variables as a function of the distance from the GPA cutoff. Does it look like probation would have a causal effect on these variables? [Max 50 words]

Answer: Here are the figures:


Left school after year 1
.04
.06
.08
.1
.12
.14
-1
-.5
0
.5
1
Distance from the cutoff
Left school after year 1
Left school after year 1


GPA in year 2
-.5
0
.5
1
-1
-.5
0
.5
1
Distance from the cutoff
GPA in year 2
GPA in year 2



Probation seems to have a positive effect on leaving school the following year, a positive effect on year 2 GPA, and possibly a negative effect on graduating in four years.

2.6 Now open the dataset *lindo_problem.dta*. This is an individual level dataset that contains the same variables as the previous dataset with the following additional variables:

gpalscutoff	Dummy for GPA in year 1 being less than cutoff
gpaXgpalscutoff	Distance from the GPA cutoff x GPA less than cutoff
gpaXgpagrcutoff	Distance from the GPA cutoff x GPA greater than cutoff

First, "individualize" the data set by removing the observation that corresponds to your birthday. For example, my birthday is 21st and in STATA I would write;

gen n=_n

drop if n==21

Now regress the predetermined background characteristics *hsgrade_pct*, *age_at_entry*, and *male* on *gpalscutoff, gpaXgpalscutoff,* and *gpaXgpagrcutoff.* In STATA you would write

reg hsgrade_pct gpalscutoff gpaXgpalscutoff gpaXgpagrcutoff, cluster(clustervar)

reg age_at_entry gpalscutoff gpaXgpalscutoff gpaXgpagrcutoff, cluster(clustervar)

reg malet gpalscutoff gpaXgpalscutoff gpaXgpagrcutoff, cluster(clustervar)

where the option "*, cluster(clustervar)*" asks the STATA to cluster the standard errors appropriately.

Then, regress the probability of being placed in probation, *probation_year1* on the same variables:

reg probation_year1 gpalscutoff gpaXgpalscutoff gpaXgpagrcutoff, cluster(clustervar)

What do these results tell you about the validity of the regression discontinuity design in this application? [Max 50 words]

Answer: Here are the regression tables:

	(1)	(2)	(3)	(4)
VARIABLES	hsgrade_pct	age_at_entry	male	probation_year1
gpalscutoff	0.437	0.0256	-0.000204	0.989
	(1.380)	(0.0380)	(0.0263)	(0.00319)
gpaXgpalscutoff	9.914	0.0526	-0.0226	-0.0164
	(3.001)	(0.0842)	(0.0592)	(0.00733)
gpaXgpagrcutoff	15.73	-0.0379	-0.0556	-0.00186
	(2.327)	(0.0668)	(0.0484)	(0.00187)
Constant	30.98	18.73	0.387	0.000835
	(0.846)	(0.0250)	(0.0157)	(0.000836)
Observations	6,264	6,264	6,264	6,264
R-squared	0.032	0.000	0.001	0.990
Robust standard errors in parentheses				

For the predetermined variables all the coefficients of gpalscutoff are close to zero and insignificant. For the treatment variable the coefficient of gpalscutoff is one and highly significant. This looks good.

2.7. Regress the outcome variables *left_school*, *nextGPA*, *gradin4* on the same variables as in 2.6 while clustering the standard errors. What do these results tell you about the causal effects of probation? [Max 50 words]

Answer: Here are the regression tables:

	(1)	(2)	(3)
VARIABLES	left_school	nextGPA	gradin4
gpalscutoff	0.0302	0.184	-0.0451
	(0.0103)	(0.0413)	(0.0243)
gpaXgpalscutoff	-0.0310	0.726	0.286
	(0.0291)	(0.0873)	(0.0561)
gpaXgpagrcutoff	0.000998	0.603	0.164
	(0.0124)	(0.0717)	(0.0507)
Constant	0.0376	0.340	0.296
	(0.00503)	(0.0273)	(0.0163)
Observations	6,264	5,640	4,396
R-squared	0.007	0.032	0.040
Robust standard errors in parentheses			

Probation has a positive effect on leaving the university and next year's GPA and a marginally significant negative effect on graduating in four years-

2.8 Students performance outcomes *nextGPA* and *gradin4* can only be measured for students who do not drop-out as a result of being placed on probation. Is this a problem for the inference about the causal effect of probation on these outcomes? [Max 50 words]

Answer: The problem is that we can only measure next year's GPA and graduating in four years for students who remain in the university. Therefore, these effects are estimated conditional on not leaving the school which is itself an outcome that is causally affected by probation. Therefore, there is a clear bad control problem here.

Please: make sure you attach the log file with your problem set (name the document "log_surname_name")