Regression Discontinuity (Optional)

(Jing Li, Miami University)

1. This note introduces Regression Discontinuity (RD), a method for estimating treatment effect when the treatment is assigned according to a rule with a cutoff value. For instance, we may use RD to evaluate the causal effect of switching to Daylight Saving Time on car accidents, where the cutoff is the date when Daylight Saving Time begins. As another example, a rule may be that a person cannot drink legally until 18 year old. Then we can use RD to examine the effect of drinking on driving behavior.

2. I use academic probation (ap) data as an illustration[1]. The rule is that a student will be placed on academic probation if first-year GPA is below the cutoff of 1.5 or 1.6[2]. My goal is to quantify the treatment effect of academic probation on several outcome variables including second-year GPA and probability of dropping school.

3. The continuous variable that determines who gets treatment is called running variable. For ap data, the running variable is gpanorm computed as the difference between first-year GPA and cutoff value

$$gpanorm = GPA\_year1 - gpacutoff \qquad (running\ variable)$$

   A student is in treatment group (being subject to academic probation) if $gpanorm < 0$. More explicitly, we can generate a treatment dummy variable d indicating treatment/no treatment

$$d = 1, \quad if \quad gpanorm < 0 \quad (treatment)$$

$$d = 0, \quad if \quad gpanorm \geq 0 \quad (no\ treatment)$$

4. R codes below read data, generate $gpanorm$ (rounded to two decimal spaces), the treatment dummy d, and interaction term of d and $gpanorm$. To highlight the neighborhood near $gpanorm = 0$, I focus on a subsample of students with $-0.6 \leq gpanorm \leq 0.6$. Formally, this interval of $(-0.6,\ 0.6)$ is called bandwidth.

[1] Data is download from `https://www.aeaweb.org/articles?id=10.1257/app.2.2.95`

[2] Regression discontinuity becomes "fuzzy" when the rule is not strictly enforced. Then instrumental variable technique is needed to estimate the treatment effect. In this note I assume the rule is strictly enforced, so regression discontinuity is "sharp".

```
> data = read.csv("311r_apdata.csv")
> data$gpanorm = data$GPA_year1 - data$gpacutoff
> data$gpanorm = round(data$gpanorm,2)
> data$d = as.numeric(data$gpanorm<0)
> data$i = data$gpanorm*data$d
> da = data[data$gpanorm>=-0.6&data$gpanorm<=0.6,]
> attach(da)
> da[1:5,c("GPA_year1","gpacutoff","gpanorm", "d", "i")]
   GPA_year1 gpacutoff gpanorm d     i
3       1.95       1.5    0.45 0  0.00
22      1.78       1.6    0.18 0  0.00
46      1.94       1.5    0.44 0  0.00
50      2.10       1.6    0.50 0  0.00
51      1.18       1.6   -0.42 1 -0.42
```

For instance, the 51-th student is placed on academic probation ($d = 1$) because the first-year gpa 1.18 is less than cutoff 1.6, their gap is $gpanorm = -0.42$.

5. Regression discontinuity is motivated by comparing students whose gpa are just below the cutoff (treatment group) to students just above the cutoff (control group). Intuitively, those two groups would be otherwise similar (a critical identification assumption for RD). Thus, the control group would provide the counterfactual or potential outcome of what would have happened to the treatment group in the absence of treatment. The causal or treatment effect can be revealed by contrasting the actual outcome with potential outcome.

6. For instance, let the outcome variable be nextGPA computed as the difference between second-year GPA and cutoff:

$$nextGPA = GPA\_year2 - gpacutoff \qquad (outcome\ variable)$$

The average nextGPA of the two groups and their difference are

```
> mean(nextGPA[gpanorm<0&gpanorm>-0.1],na.rm=T)
[1] 0.4895549
> mean(nextGPA[gpanorm>0&gpanorm<0.1],na.rm=T)
[1] 0.3776385
```

```
> t.test(nextGPA[gpanorm<0&gpanorm>-0.1],nextGPA[gpanorm>0&gpanorm<0.1])
Two Sample t-test
t = 2.7492, df = 1524.4, p-value = 0.006044
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.03206619 0.19176657
```

(a) The average nextGPA of treatment group (students just below cutoff) is 0.4895549, greater than average nextGPA 0.3776385 of control group (students just above cutoff)

(b) The gap $0.1119164 = 0.4895549 - 0.3776385$ suggests positive treatment effect—being in academic probation is associated with the second year GPA of those affected students rising by 0.1119164. In other words, in the absence of academic probation, the average nextGPA of those students would have been 0.3776385.

(c) The difference 0.1119164 is statistically significant as the t value 2.7492 exceeds 1.96.

(d) However, that difference tends to underestimate the treatment effect because it ignores the fact that students in the treatment group have lower first-year GPA than control group. Recall the formula for omitted variable bias (OVB)

$$OVB = \frac{\beta_2 cov(d, o)}{var(d)}$$

where the omitted variable o is gpanorm. We can show $\beta_2 > 0, cov(d, gpanorm) < 0$, and as a result $OVB < 0$

7. The OVB of two-sample t test suggests a better approach of running a multiple regression that controls for gpanorm and its interaction term with the treatment dummy (denoted by i). The RD regression is given as

$$nextGPA = \beta_0 + \beta_1 d + \beta_2 gpanorm + \beta_3 i + \beta_4 w + u \quad (RD\ regression) \qquad (1)$$

where w denotes control variables or covariates (such as high school GPA). Ignore w for a moment, we can show

$$\beta_1 = E(nextGPA|d = 1, gpanorm = 0) - E(nextGPA|d = 0, gpanorm = 0)$$

or equivalently,

$$\beta_1 = E(nextGPA|d = 1, GPA\_year1 = gpacutoff) - E(nextGPA|d = 0, GPA\_year1 = gpacutoff)$$

Thus, β_1 measures the gap in average nextGPA across two groups at the cutoff.

8. The results of regression (1) without w are

```
> summary(lm(nextGPA~d+gpanorm+i),data=da)$coef
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 0.30673577 0.01877586 16.3367085 2.508890e-59
d           0.23990185 0.03049014  7.8681777 3.919906e-15
gpanorm     0.72198857 0.05101269 14.1531163 4.263153e-45
i           0.03268557 0.09074901  0.3601755 7.187225e-01
```

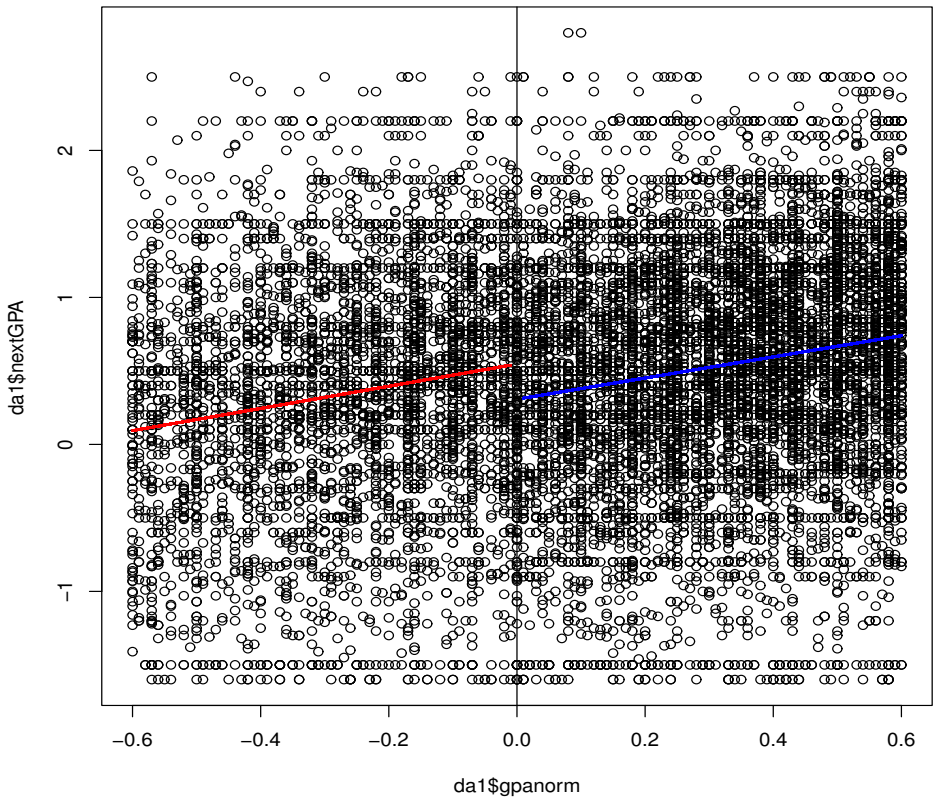
   (a) We find that the gap in average nextGPA across the two groups at $gpanorm = 0$ is 0.23990185. Its being greater than the two-sample-t-test gap 0.1119164 confirms that OVB is indeed negative.

   (b) The new t value 7.8681777 indicates significance

   (c) We see $\beta_2 = 0.72198857$ is indeed positive

   (d) The small t value of the interaction term 0.3601755 implies no across-group difference in the marginal effect of gpanorm on nextGPA.

9. You may recognize that the RD regression (1) is just an application of dummy variable regression (DVR). Therefore, its results can be alternatively obtained by running two group-wise or separate regressions—one for treatment group, and the other for the control group, and each regression regresses nextGPA onto gpanorm. That explains a common practice to plot the outcome variable against running variable, with two regression lines superimposed, one for the treatment group and the other for the control group, see the graph below



(a) The blue line, which represents control group, is upward sloping because $\beta_2 = 0.72198857 > 0$. It shows as gpanorm of control group rises, the nextGPA rises as well.

(b) The red line represents treatment group. The red and blue lines are almost parallel because the interaction term in RD regression (1), which measures the change in slope, is insignificant

(c) Most importantly, we see a jump or discontinuity at gpanorm=0, or equivalently, at $GPA\_year1 = gpacutoff$. The punchline of a RD study is revealing that jump at the right location, and, for the right variable.

(d) That jump implies that, on average, second-year gpa of students just below the cutoff (treatment group) is greater than students just above the cutoff (control group) by 0.23990185

(e) If we cannot think of other reasons for that jump, then it must result from the

treatment of academic probation.

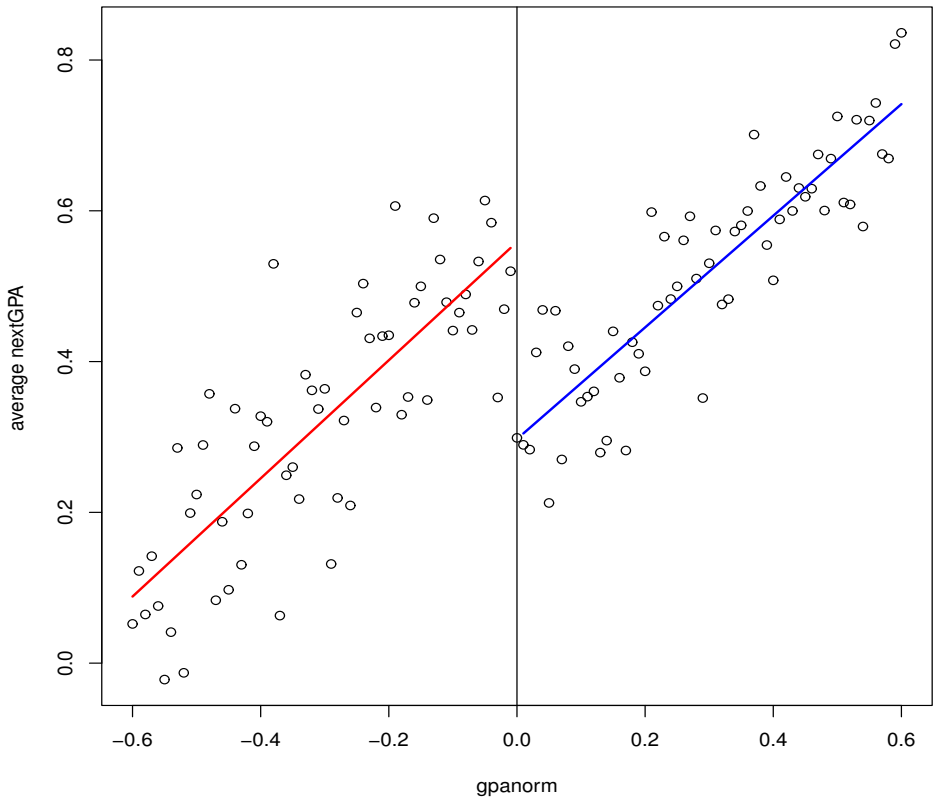
The codes to generate the plot are below

```
sum(is.na(da$nextGPA))
da1 = da[!is.na(da$nextGPA),]


plot(da1$gpanorm,da1$nextGPA)
x = da1$gpanorm; y = da1$nextGPA
x1 = x[x<0]; y1 = y[x<0]; m1 = lm(y1~x1)
y1h = predict(m1)
lines(x1,y1h,col = "red", lwd = 2)
x2 = x[x>0]; y2 = y[x>0]; m2 = lm(y2~x2)
y2h = predict(m2)
lines(x2,y2h,col = "blue", lwd = 2)
abline(v=0)
```

We need to be aware of missing values denoted by NA in nextGPA since some students placed on academic probation choose to drop school in second year. The R code !is.na excludes those missing students. A key feature of the codes is running two regressions separately—one using data below cutoff and the other above cutoff.

10. Because ap data is very large, the previous graph that displays individual student is cluttered. A better way is plotting the average nextGPA conditional on each value of gpanorm, see new graph below



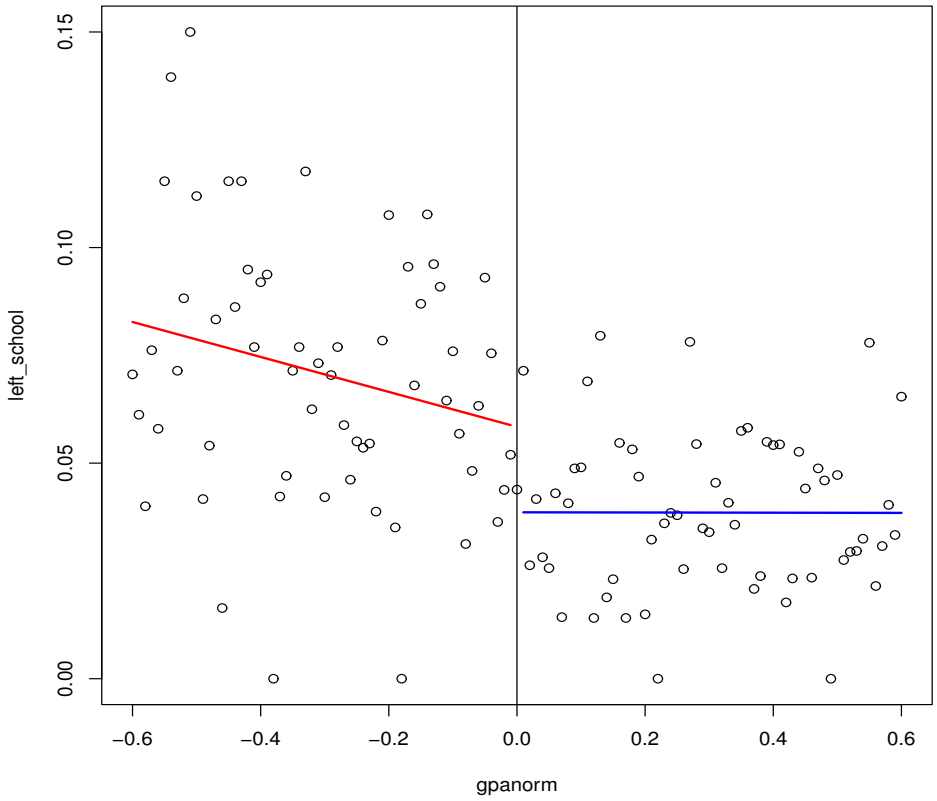
The codes are

```
count = NULL; x = sort(unique(da$gpanorm))
for (j in 1:length(unique(da$gpanorm))) {
        count[j] = mean(nextGPA[(da$gpanorm==sort(unique(da$gpanorm))[j])],na.rm=T)
}
plot(x,count,xlab="gpanorm",ylab="averge nextGPA")
x1 = x[x<0]; y1 = count[x<0]; m1 = lm(y1~x1)
y1h = predict(m1)
lines(x1,y1h,col = "red", lwd = 2)
x2 = x[x>0]; y2 = count[x>0]; m2 = lm(y2~x2)
y2h = predict(m2)
lines(x2,y2h,col = "blue", lwd = 2)
abline(v=0)
```

Notice that we use a for loop to compute $E(nextGPA|gpanorm)$ for each value of

gpanorm, and save those conditional means in count. Then two separate regressions below and above the cutoff that use count as the dependent variable are fitted. Finally, two regression lines based on those two regressions are drawn. In the new graph each point refers to a conditional mean as opposed to an individual student. We may call the graph as RD diagram.

11. Next we try using the dummy variable $left\_school$ as the outcome variable, which equals one if a student drops school. The new RD diagram for $left\_school$ is below



Again each point represents a conditional mean of $left\_school$. Because it is a dummy variable, the mean is the same as proportion.

12. (Exercise) Can you reconcile the pattern in the above RD diagram with regression results below? Hint: pay attention to the results highlighted in red

```
> summary(lm(left_school~d+gpanorm+i),data=da)$coef
                 Estimate  Std. Error      t value     Pr(>|t|)
(Intercept)  3.957077e-02 0.004943450  8.004686715 1.300376e-15
d            1.958211e-02 0.007887538  2.482664197 1.305309e-02
gpanorm      8.470264e-05 0.013455498  0.006295021 9.949774e-01
i           -3.903109e-02 0.023338928 -1.672360174 9.447771e-02
```

13. To prove the existence of a significant treatment effect and defend that finding, we need to

    (a) show there is jump in outcome variable at cutoff. The direction and magnitude of the jump hopefully are consistent with prior belief.

    (b) show that the jump in outcome variable is insensitive to function form—we may add squared running variable and its interaction with treatment dummy in RD regression

    (c) use Placebo Study I to show there is no jump in covariates or control variables at cutoff. This exercise can shed light on the identification assumption that the treatment and control groups are otherwise similar. Put differently, we hope to see that ceteris paribus condition holds, or the comparison is apple-to-apple, not apple-to-orange.

    (d) use Placebo Study II to show there is no jump in outcome variable far away from cutoff. This exercise intends to rule out that a jump can occur just by chance.

    (e) run Robustness Check to show the results are insensitive to bandwidth—the width of interval of the running variable that is used to compute the conditional mean of the outcome variable

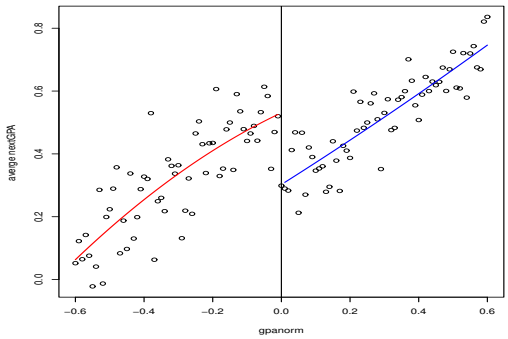
    (f) rule out Non-random Sorting—the possibility that people manipulate the system and choose treatment/no treatment endogenously. Non-random sorting leads to apple-to-orange comparison, and therefore compromises the finding of RD.

14. Let us check the robustness to function form by creating the squared term of running variable and its interaction term with the treatment dummy. Then re-run RD regression (1) with these two additional regressors. The results are below

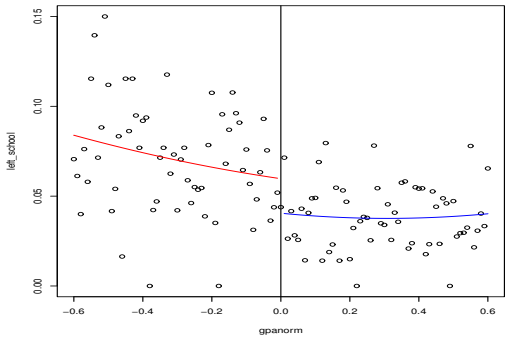
```
> da$rsq = da$gpanorm^2
> da$irsq = da$d*da$rsq
```

```
> summary(lm(nextGPA~d+gpanorm+i+rsq+irsq,data=da))$coef
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  0.3149037 0.02752748 11.4396127 3.855452e-30
d            0.2101283 0.04566868  4.6011470 4.246578e-06
gpanorm      0.6428428 0.20161094  3.1885312 1.433802e-03
i           -0.1102944 0.35443889 -0.3111803 7.556692e-01
rsq          0.1282358 0.31602887  0.4057723 6.849175e-01
irsq        -0.5050897 0.57294244 -0.8815715 3.780270e-01
```

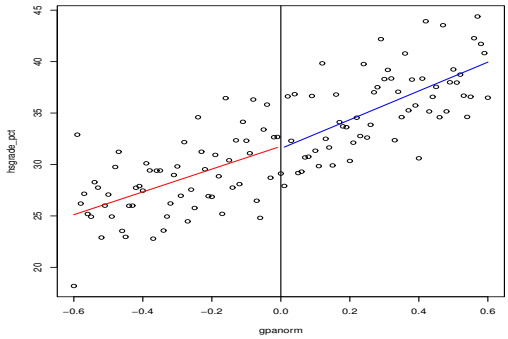
The newly-added regressors are insignificant, and the jump 0.2101283 is akin to the previous jump. With the squared term, now we draw two polynomial curves in the RD diagram shown below



The updated RD diagram with squared term for $left\_school$ is below



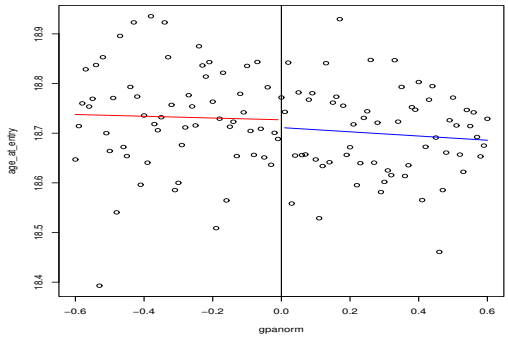
15. The RD diagram below conducts Placebo Study I to the covariate $hsgrade\_pct$, which measures the percentile of high school grade for a student.



Unlike nextGPA and *left_school*, this time we see no obvious jump at gpanorm=0. Thus, the jumps in the outcome variables *nextGPA* and *left_shool* cannot be attributed to *hsgrade_pct*. Results of RD regression (1) applied to *hsgrade_pct* are below. Notice the insignificant t value 0.6740133, which implies no jump

```
> summary(lm(hsgrade_pct~d+gpanorm+i),data=da)$coef
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 31.165173  0.5204449 59.8817969 0.000000e+00
d            0.559699  0.8303975  0.6740133 5.003150e-01
gpanorm     14.432870  1.4165906 10.1884549 2.762703e-24
i           -2.847013  2.4571150 -1.1586812 2.466077e-01
```

16. We also carry out Placebo Study I to another covariate *age_at_entry*, the age at which a student enters the college



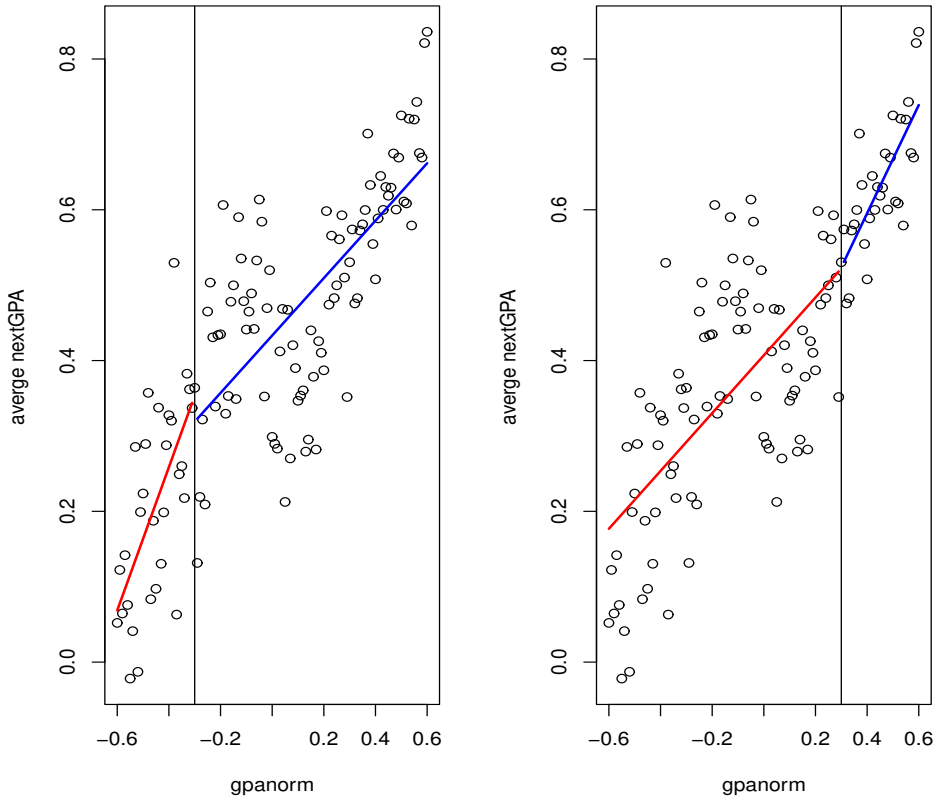
Results of RD regression (1) are

```
> summary(lm(age_at_entry~d+gpanorm+i),data=da)$coef
               Estimate Std. Error      t value  Pr(>|t|)
(Intercept) 18.71784312 0.01666694 1123.0520540 0.0000000
d            0.02044723 0.02659299    0.7688954 0.4419696
```

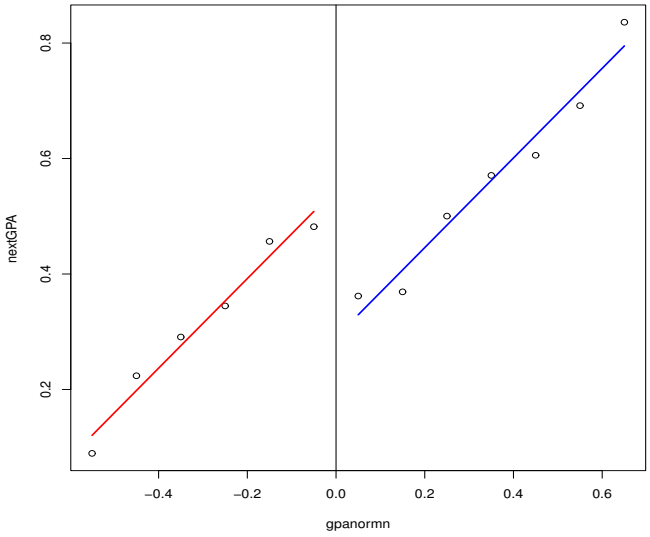
```
gpanorm      -0.02989352 0.04536548   -0.6589485 0.5099406
i             0.03247802 0.07868767    0.4127460 0.6797996
```

No jump is found for that covariate either. Good news.

17. Next, Placebo Study II is implemented to the outcome variable nextGPA at $gpanorm = -0.3$ and $gpanorm = 0.3$. We see no jumps at those two locations, implying that the jump at gpanorm=0 is unlikely to be there just by chance. Good news again!



18. To show robustness to bandwidth, we round the running variable gpanorm to one decimal space. The new RD diagram for nextGPA is below. Because of greater bandwidth, we obtain fewer conditional means of outcome variables, and the new diagram displays fewer points. Other than that, we see no obvious change from the previous diagram—the jump is still there at the right location, with right direction and similar magnitude. Another good news.

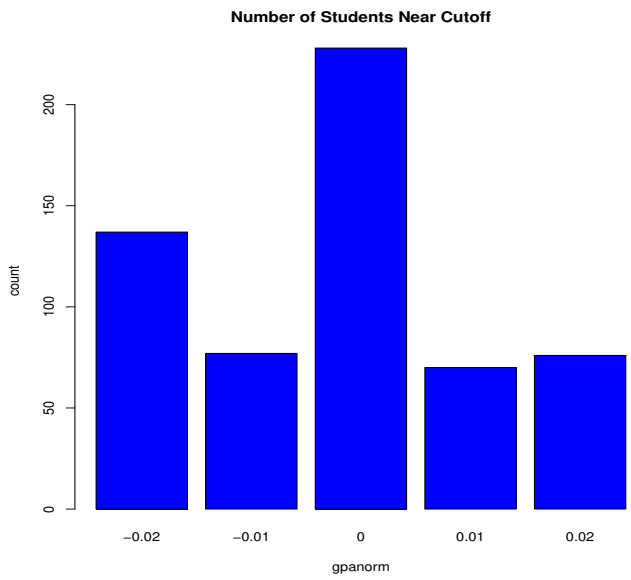


19. Here comes bad news! We do find evidence of non-random sorting or self-selection. In this case, the number of students with gpanorm=0 is remarkably greater than students with gpanorm=-0.01. One plausible explanation is that some students near the cutoff are trying to avoid academic probation by, say, asking professors to give a favorable grade or cheating in exams. In the absence of non-random sorting, the distribution of the running variable should be smooth around the cutoff[3], and we should see no big jump in number of students from gpanorm=-0.01 to gpanorm=0. Instead, below are our findings

```
> sum(da$gpanorm==-0.01)
[1] 77
> sum(da$gpanorm==0)
[1] 228
> sum(da$gpanorm==0.01)
[1] 70
```

Or we can use a bar graph to highlight the jump of student number at gpanorm=0

[3]That is the null hypothesis of McCrary Test, see `https://www.sciencedirect.com/science/article/abs/pii/S0304407607001133` for more detail.



The presence of non-random sorting implies that more students could have been in the treatment group than it actually was. Or in other words, because of those "self-select smart" guys, the average nextGPA of the control group is downward biased. We can no long assume that the treatment is assigned randomly—instead, we see some students are self-selected to avoid the treatment. In short, self-selection results in bias or endogeneity, and the true treatment effect may not be as big as indicated by the jump in the RD diagram of nextGPA or indicated by the following result of RD regression (1) applied to nextGPA with covariates

```
> summary(lm(nextGPA~d+gpanorm+i+hsgrade_pct+age_at_entry+totcredits_year1),data=da
                     Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)       0.733158731 0.2046293260  3.5828625 3.412525e-04
d                 0.238132639 0.0303101776  7.8565240 4.300175e-15
gpanorm           0.669759423 0.0509385925 13.1483692 3.330819e-39
i                 0.041420696 0.0902078466  0.4591695 6.461211e-01
hsgrade_pct       0.003558940 0.0003252232 10.9430688 9.823011e-28
age_at_entry     -0.029979363 0.0101160491 -2.9635446 3.047404e-03
totcredits_year1  0.005202526 0.0142327549  0.3655319 7.147211e-01
```

20. Finally, I will use potential outcome notations[4] to derive the RD regression (1). Let y_{1i} and y_{0i} denote the potential outcomes for the i-th student if the student is in treatment

[4]See https://www.jstor.org/stable/2692190 for more discussion

group $d_i = 1$ or control group $d_i = 0$. The actual outcome is y_i, which is either y_{1i} or y_{0i} depending on d_i :

$$y_i = \begin{cases} y_{1i} & \text{if } d_i = 1 \\ y_{0i} & \text{if } d_i = 0 \end{cases} \tag{2}$$

Let x_i be observed confounders or covariates, and u_i be unobserved one. We assume y_{0i} and x_i are related via an unknown function f :

$$y_{0i} = f(x_i) + u_i \tag{3}$$

We also assume the difference between y_{1i} and y_{0i} is the treatment effect β_1

$$y_{1i} - y_{0i} = \beta_1 \tag{4}$$

From (2), (3) and (4), it follows that

$$y_i = y_{0i} + (y_{1i} - y_{0i})d_i = f(x_i) + u_i + \beta_1 d_i = \beta_1 d_i + f(x_i) + u_i \tag{5}$$

We obtain the RD regression by (i) assuming f is linear function[5]; (ii) letting x_i contain the running variable r_i and control variable w_i; (iii) subtracting cutoff from r_i, and (iv) adding the interaction term of r_i and d_i

$$y_i = \beta_0 + \beta_1 d_i + \beta_2(r_i - c) + \beta_3 d_i \times (r_i - c) + \beta_4 w_i + u_i \quad (RD\ Regression) \tag{6}$$

where c is the cutoff. The key assumption to identify β_1 is conditional mean independence (CMI)

$$E(u_i|x_i, d_i) = E(u_i|x_i) \qquad (conditional\ mean\ independence) \tag{7}$$

That is, once x_i has been controlled for, treatment assignment d_i is as good as random. The CMI assumption is likely to be true if we can show x_i is balanced, that is, x_i of treatment and control groups are similar (no jump in covariates). A counter-example is non-random sorting, which indicates the distribution of running variable is unbalanced or not smooth across the cutoff.

[5]Alternatively, non-parametric method such as Lowess can be used to estimate $f(x_i)$