

empirical_proj_2

Lindsey Greenhill

4/9/2022

Question 1

The fundamental problem of causal inference is that you cannot see the outcome for a child who both was in a small class and not in a small class. In other words, it is impossible to measure the dual outcomes when a child both is in a small class and not in a small class, because the child can only either be in a small class or not in a small class.

Question 2

The following variables have missing values:

- towncode: 318 missing
- math: 58 missing
- verb: 466 missing
- ses_index: 318 missing
- boy: 1037 missing

```
##      student_id      towncode      schlcode      class_id
## Min.      :    1  Min.      : 166  Min.      :11005  Min.      :110051
## 1st Qu.:13600  1st Qu.:26104  1st Qu.:31049  1st Qu.:310491
## Median :27412  Median :62000  Median :41284  Median :412841
## Mean   :27395  Mean   :50434  Mean   :39776  Mean   :397762
## 3rd Qu.:41158  3rd Qu.:79004  3rd Qu.:51247  3rd Qu.:512471
## Max.   :54860  Max.   :98004  Max.   :61365  Max.   :613651
##                NA's      :318
## school_enrollment  class_size      math      verb
## Min.      : 9.00  Min.      : 7.00  Min.      : 1.0  Min.      : 1.00
## 1st Qu.: 57.00  1st Qu.:28.00  1st Qu.:23.0  1st Qu.:25.00
## Median : 75.00  Median :32.00  Median :27.0  Median :27.00
## Mean   : 80.41  Mean   :31.58  Mean   :25.3  Mean   :25.97
## 3rd Qu.:105.00  3rd Qu.:36.00  3rd Qu.:29.0  3rd Qu.:29.00
## Max.   :188.00  Max.   :44.00  Max.   :30.0  Max.   :30.00
##                NA's      :58  NA's      :466
##      ses_index      boy      born_isr      religious
## Min.      : 0.00  Min.      :0.0000  Min.      :0.0000  Min.      :0.0000
## 1st Qu.: 4.00  1st Qu.:0.0000  1st Qu.:1.0000  1st Qu.:0.0000
## Median : 9.00  Median :1.0000  Median :1.0000  Median :0.0000
## Mean   :13.17  Mean   :0.5078  Mean   :0.9608  Mean   :0.2242
## 3rd Qu.:18.00  3rd Qu.:1.0000  3rd Qu.:1.0000  3rd Qu.:0.0000
## Max.   :76.00  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000
## NA's      :318  NA's      :1037
```

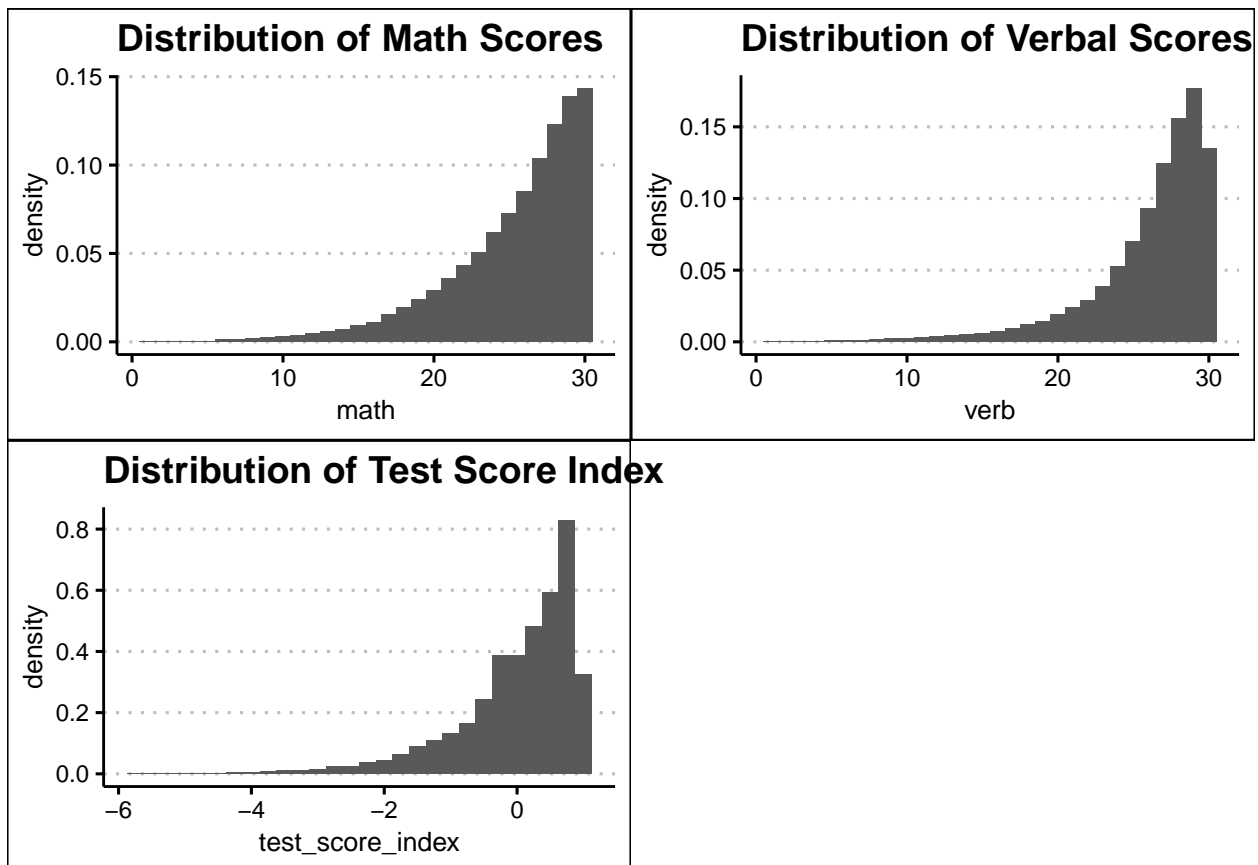
Questions 3 and 4

Table 1: Summary Statistics

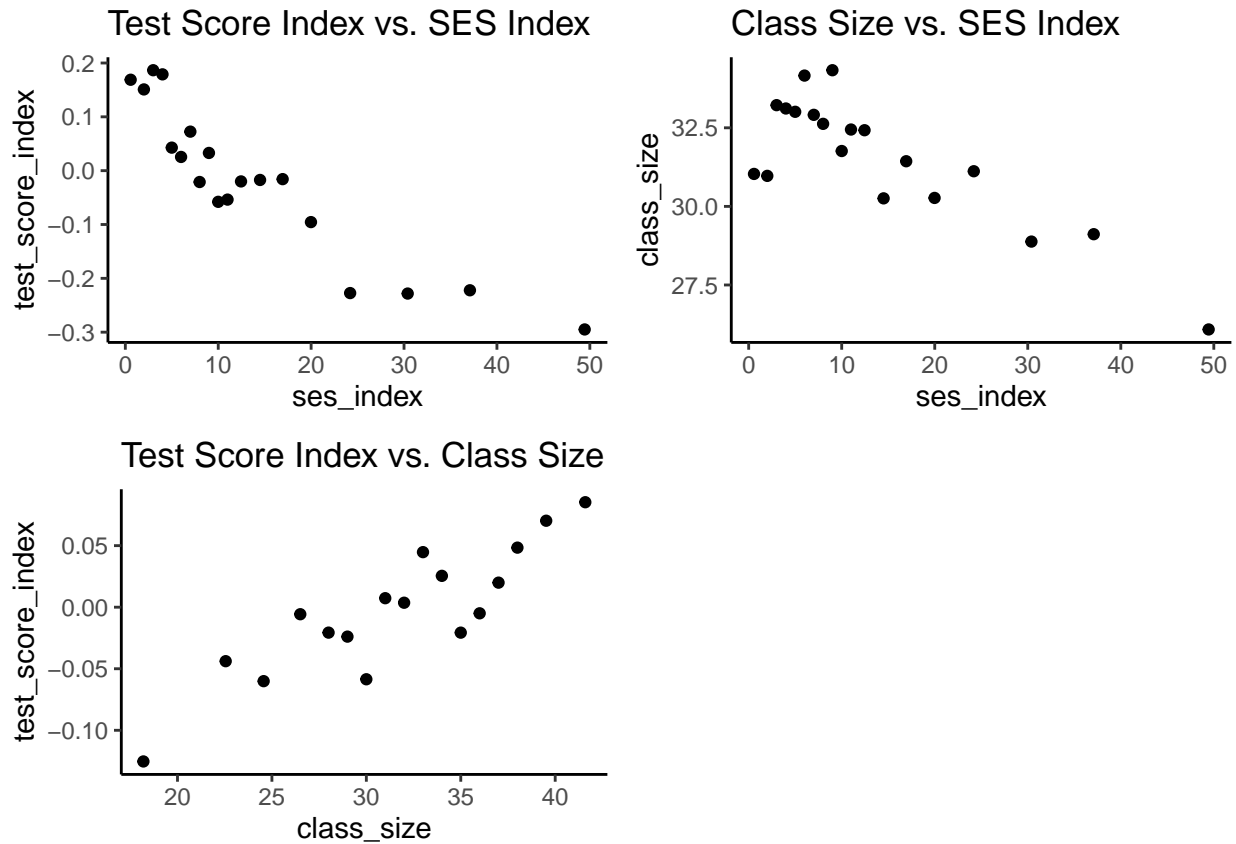
Variable	Mean	Sd	Min	Max
math	25.3	4.527	1	30
verb	26	4.279	1	30
test_score_index	0	0.903	-5.835	1.038

Question 5

All of the histograms appear to not be symmetrical and all of them have a long left tail.



Question 6



Question 7

As we can see from the binned scatter plots above, the socio economic index variable is correlated with both class size and test score index. As such, it is a confounding variable and mean we cannot interpret the relationship between class size and test scores causally because it is likely that a student's socioeconomic status influences both outcomes, meaning that we don't know how greatly class size affects test scores versus ses affects test scores.

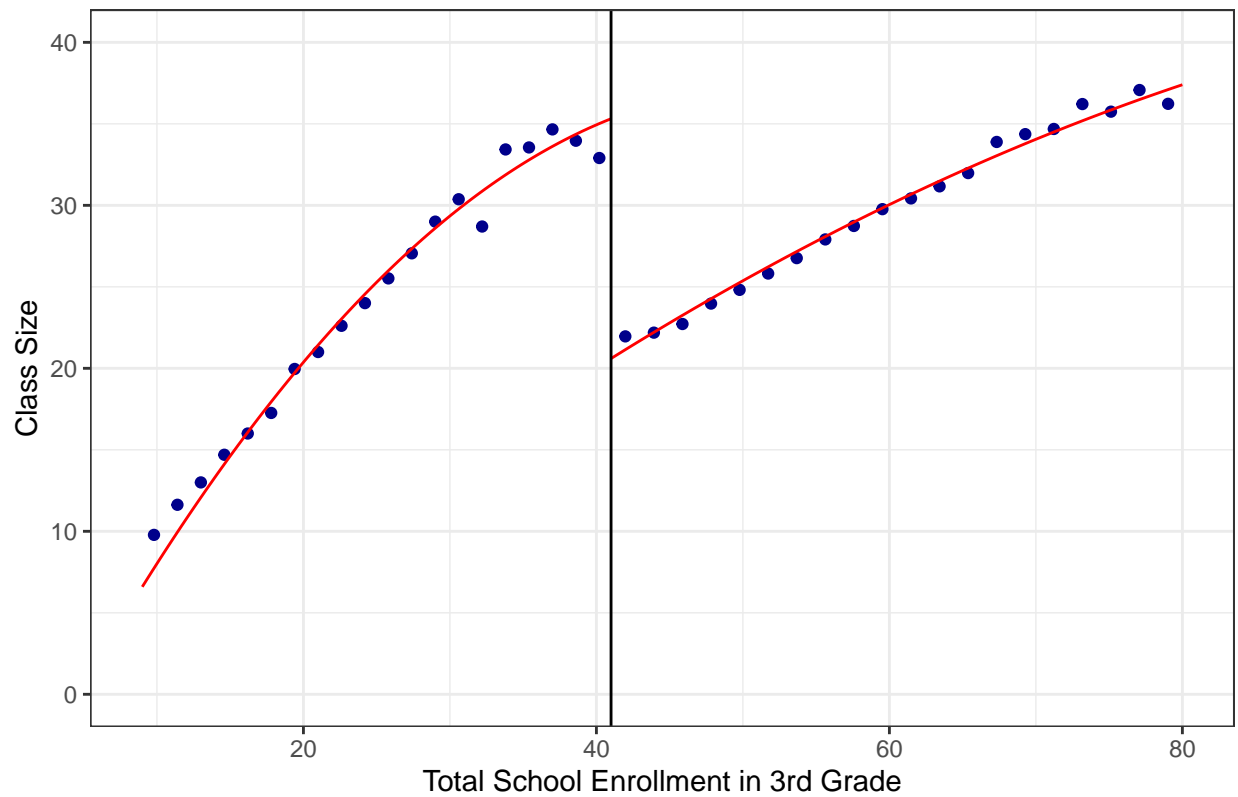
Question 8

Part a

I chose a quadratic model it seems to fit the data better than a linear model, especiaally for the data to the left of the break

```
## [1] "Mass points detected in the running variable."
```

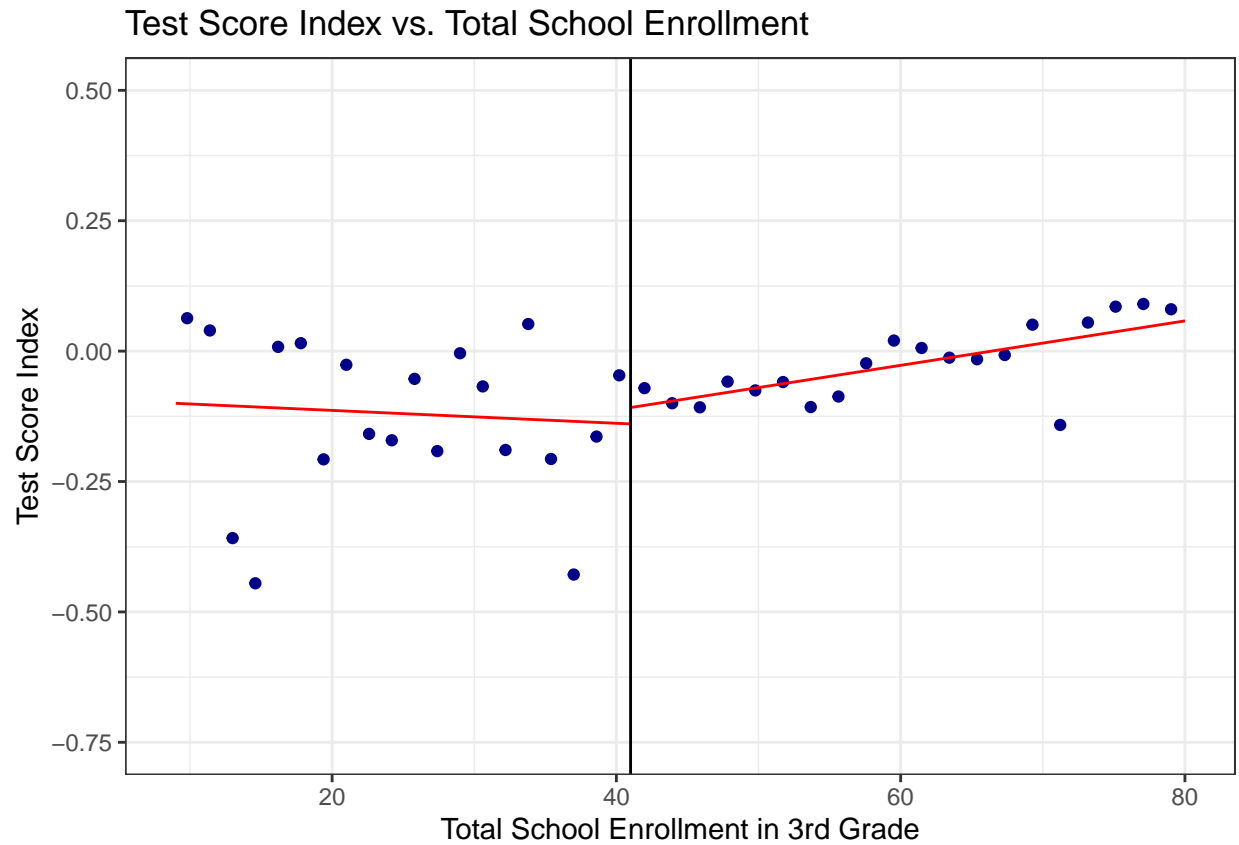
Class size vs. total school enrollment



Part b

I chose to use a linear model because it seems to fit the data better than a quadratic model. When I chose a quadratic model, it seemed like residuals on the right side of the break were much more in the positive than the negative, which would be bad. Looking at the graph, it doesn't seem like there is a wide gap in schools with 40 kids and schools with 41 kids enrolled.

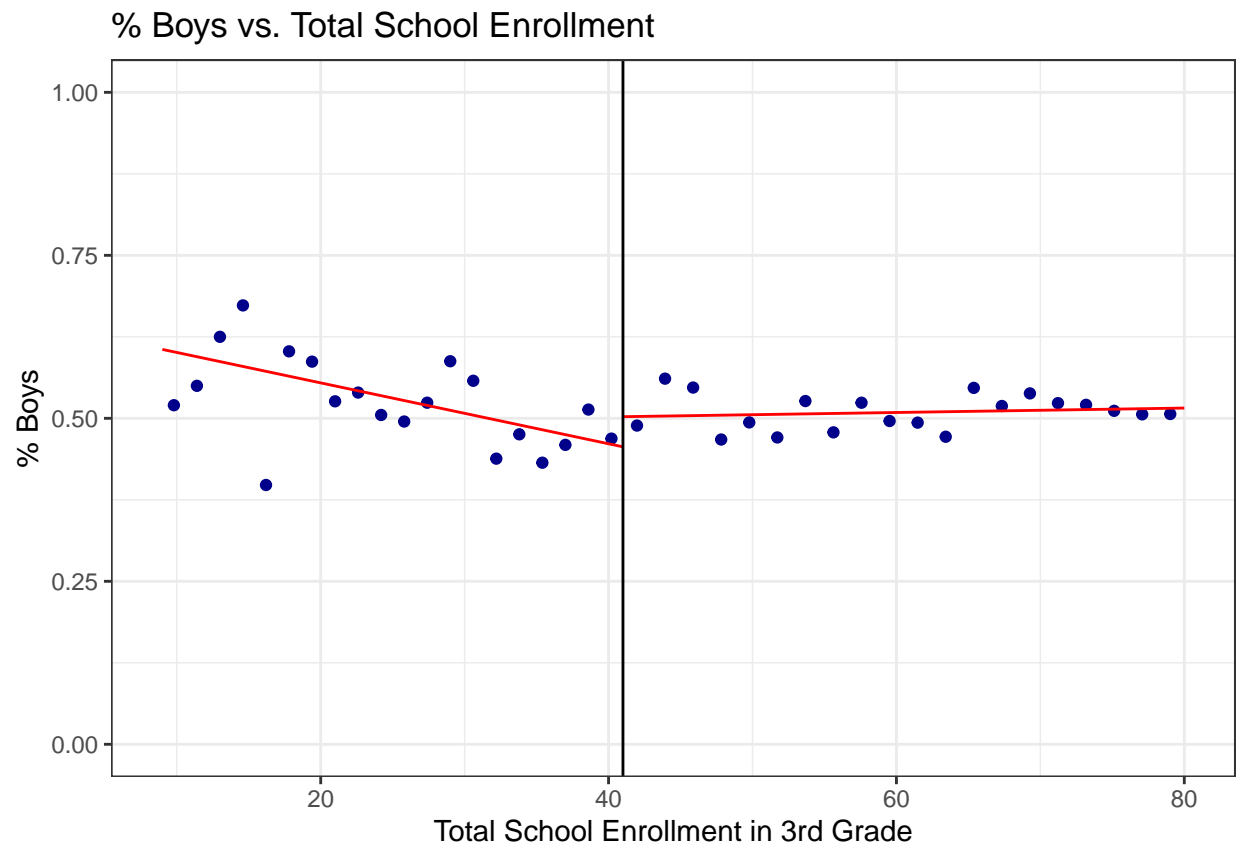
```
## [1] "Mass points detected in the running variable."
```



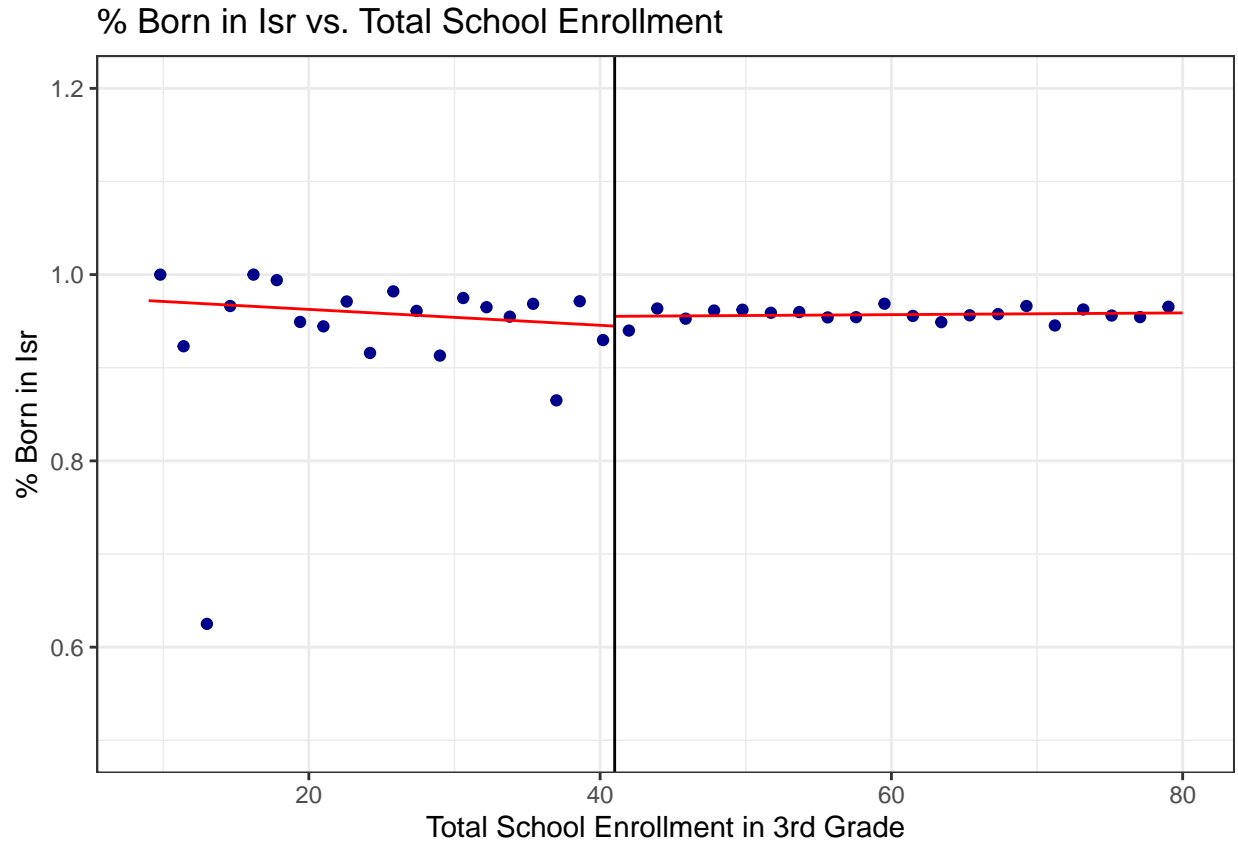
Question 9

Part a

```
## [1] "Mass points detected in the running variable."
```



```
## [1] "Mass points detected in the running variable."
```



Part b

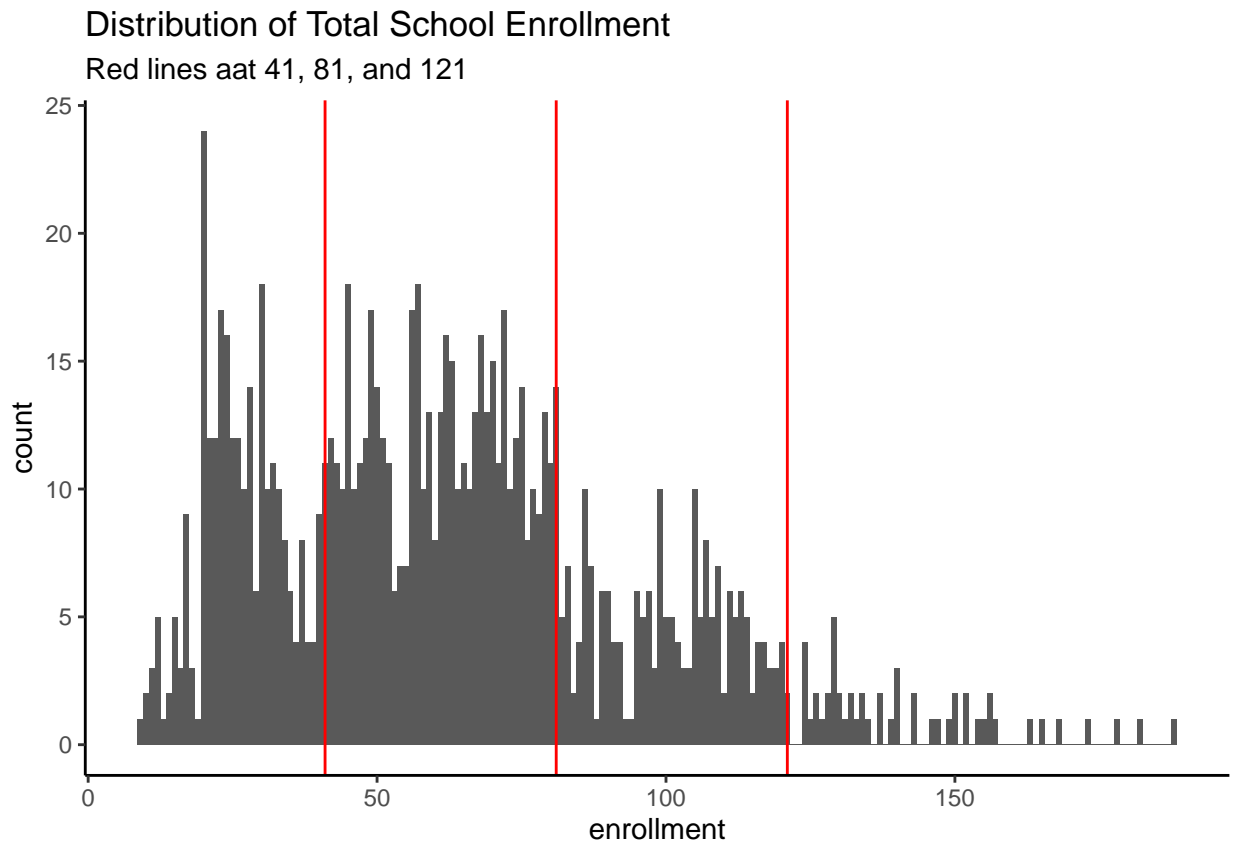
The identification assumption for the regression discontinuity design is that schools on either side of the cutoff are similar in all ways and as such the fact that they have school enrollment over or under the cutoff is as if random. The graphs from part a above are consistent with this identification assumption, as schools just at either side of the cutoff do not seem to have different percentages of boys or kids born in Israel enrolled.

Question 10

Part a

Manipulation of school enrollment could possibly invalidate the identification assumption because it means that schools could intentionally stay right below the discontinuity cutoff, meaning that the schools on either side of the cutoff would be different in this way.

Parts b, c, and d



Looking at the histogram above, it seems like there is a spike in enrollment number right after the break at 41. However, there seems to be a drastic drop off in enrollment right at the 81 threshold line. As such, it certainly seems possible that there was manipulation in enrollment, but there are conflicting behaviors at the two cutoffs. There doesn't appear to be any spike at the 121 line.

Question 11

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac@gmail.com % Date and time: Sat, Apr 09, 2022 - 16:07:14

- For the test score model, the estimate of the discontinuity at the 41 threshold = .026.
- For the class size model, the estimate of the discontinuity at the 41 threshold = -16.6

Table 2:

	<i>Dependent variable:</i>	
	avg_class_size	avg_test_score_index
	(1)	(2)
above41	-16.610*** (0.502)	0.026 (0.070)
dist_from_41	0.857*** (0.026)	-0.001 (0.004)
interaction41	-0.429*** (0.028)	0.005 (0.004)
Constant	38.175*** (0.424)	-0.135** (0.060)
Observations	744	744
R ²	0.790	0.021
Adjusted R ²	0.789	0.017
Residual Std. Error (df = 740)	2.971	0.417
F Statistic (df = 3; 740)	927.181***	5.193***

Note:

*p<0.1; **p<0.05; ***p<0.01