Professor Raj Chetty and Dr. Gregory Bruich
Department of Economics, Harvard University

Empirical Project 2
Do Smaller Classes Improve Test Scores? Evidence from a Regression Discontinuity Design
Posted on Thursday, February 21, 2019
Due at midnight on Thursday, March 7, 2019

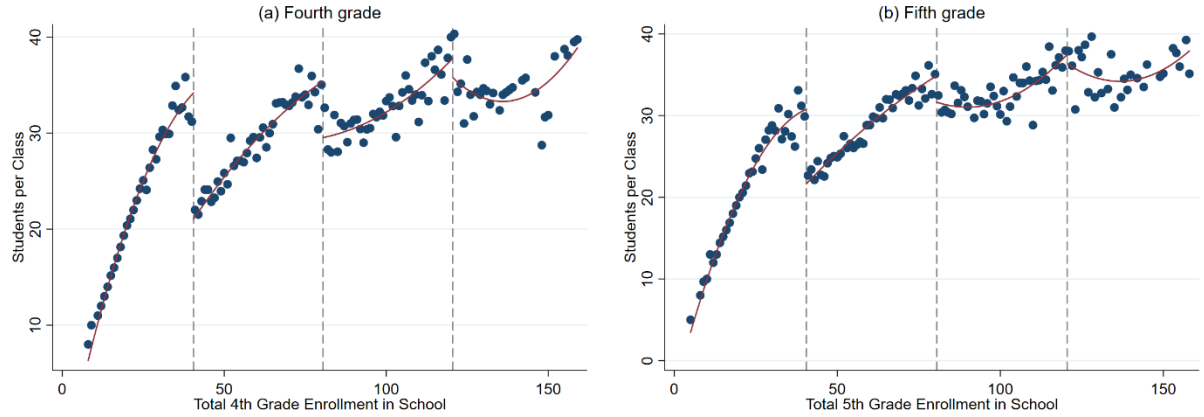
In this empirical project, you will use a regression discontinuity design to estimate the causal effect of class size on test scores. To answer some of the questions, you will need to refer to the following papers:

1. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," *Quarterly Journal of Economics* 126(4): 1593–1660.

2. Angrist, Joshua D., and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 114(2): 533–575.

The Stata data file grade5.dta consists of test scores in fifth grade classes at public elementary schools in Israel. These data were originally used in Angrist and Lavy (1999). The graphs below were drawn using the same data.

Figure 1
Class Size as a Function of Total School Enrollment in Public Schools in Israel



Note: These figures plot class size as a function of total school enrollment for fourth grade and fifth grade classes in pubic schools in Israel in 1991.

Instructions

Please submit your Empirical Project on Canvas. Your submission should include three files:
1. A 4-6 page replication as a word or pdf document (double spaced and including references, graphs, and tables)
2. A do-file with your STATA code or an .R script file with your R code
3. A log file of your STATA or R output

Specific questions to address in your replication

1. Explain why a simple comparison of test scores in small classes versus large classes would not measure the causal effect of class size. Would this simple comparison likely be biased upwards or biased downwards relative to that true causal effect? Explain.

2. (To answer this and the next question, read Chetty et al. 2011). How did the Tennessee STAR experiment overcome this problem? What did it find?

3. What is a binned scatter plot? Explain how it is constructed.

4. Graphical regression discontinuity analysis, focusing on the 40 student school enrollment threshold. See Table 2a and 2b for more guidance.

    a. Draw a binned scatter plot to visualize how class size changes at the 40 student school enrollment threshold. Display a linear or quadratic regression line based on what you see in the data.

    b. Draw binned scatter plots to visualize how math and verbal test scores change at the 40 student school enrollment threshold. Display a linear or quadratic regression line based on what you see in the data.

    c. Draw binned scatter plots to test whether (i) the percent of disadvantaged students, (ii) the fraction of religious schools, and (iii) the fraction of female students evolve smoothly across the 40 student school enrollment threshold. Display a linear or quadratic regression line based on what you see in the data.

    d. Produce a histogram of the number of schools by total school enrollment. Note that you must collapse the data by *school* to produce this graph.

5. Regression analysis. Run the regressions that correspond to your three graphs in 4a and 4b to quantify the discontinuities that you see in the data. In estimating these regressions, use all the observations with school enrollment less than 80. Report a 95% confidence interval for each of these estimates. See Table 2a and 2b for more guidance.

6. Recall that any quasi experiment requires an identification assumption to make it as good as an experiment. What is the identification assumption for regression discontinuity design? Explain whether your graphs in 4c and 4d are consistent with that assumption.

7. (To answer this question, read Angrist and Lavy (1999)). If all schools followed the class size rule exactly as described in Angrist and Lavy (1999), how much would you expect class size to change at the 40 student enrollment threshold? Explain why the actual change in class size that you see in the data is less than this.

8. Suppose your school superintendent is considering a reform to reduce class sizes in your school from 40 to 35. Use your estimates above to predict the change in math and verbal test scores that would result from this reform.

   Hint: divide the RD estimate of the change in test scores by the change in number of students per class at the threshold.

9. Now suppose you are asked for advice by another school that is considering reducing class size from 20 to 15 students – a 5 unit reduction as above. Would you feel confident in making the same prediction as you did above about the impacts this change will have? Why or why not?

10. Compare your estimates in 8 with the estimates from (i) the Tennessee STAR experiment (Chetty et al. 2011) and (ii) data from Sweden (Fredriksson et al. 2013) discussed in lecture. Give <u>two reasons</u> that your estimates might differ from those of these other studies.

11. Chetty et al. (2011) show that being assigned to a smaller class in Kindergarten raises Kindergarten test scores, but has little impact in later grades. Does this "fade out" effect mean that class size doesn't really matter in the long run? Why or why not?

12. Given the evidence above, would you encourage your hometown school to reduce class size by hiring more teachers if the goal is to maximize students' long-term outcomes (e.g., college attendance rates, earnings)? Explain clearly what other data you would need to make a scientific recommendation and how you would use that data.

DATA DESCRIPTION, FILE: grade5.dta

The data consist of *n* = 2,019 fifth grade classes at 1,002 public schools in Israel in 1991. For more details on the construction of the variables included in this data set, please see Angrist and Lavy (1999).

Table 1
Definitions of Variables in grade5.dta

Variable	Label
(1)	(2)
schlcode	School id code
school_enrollment	Total school enrollment in fifth grade
grade	Class grade 5 = fifth grade for all observations in grade5.dta
classize	Number of students in the class
avgmath	Average composite year-end math score in the class, on a scale of 1 to 100, from a national elementary school test.
avgverb	Average composite year-end verbal score in the class, on a scale of 1 to 100, from a national elementary school test.
disadvantaged	Percent of class coming from a disadvantaged background, as defined by an index used by the Ministry of Education to allocate supplementary hours of instruction and other school resources. The index is based on fathers' education, fathers' continent of birth, and family size.
female	Fraction of students in the class that are female
religious	1 = School is a religious public school 0 = School is a secular public school

Note: This table describes the variables included in grade5.dta.

Table 2a
STATA Commands

STATA command	Description

`*Install binscatter`
`ssc install binscatter, replace`
 	
`*Draw graph (command all goes on one line)`
`binscatter yvar school_enrollment if inrange(school_enrollment,20,60), rd(40.5) discrete line(lfit)`
 	
`*Save graph`
`graph export figure1_linear.png, replace`
 	
`*Draw graph (command all goes on one line)`
`binscatter yvar school_enrollment if inrange(school_enrollment,20,60), rd(40.5) discrete line(qfit)`
 	
`*Save graph`
`graph export figure1_quadratic.png, replace`
	The first command installs binscatter, which only has to be done once. The second command produces a binned scatter plot of *yvar* against the total school enrollment with a linear best fit line, restricting the graph to observations with total school enrollment in [20,60]. The third line saves the graph. The fourth line shows how to change the best fit line to be quadratic by changing *line(lfit)* to *line(qfit)*.

`*Collapse data to school level`
`collapse (mean) school_enrollment, by(schlcode )`
 	
`*Graph counts (command all goes on 1 line)`
`twoway (histogram school_enrollment if inrange(school_enrollment,20,60), discrete frequency), xline(40.5)`
 	
`*Save graph`
`graph export school_counts.png, replace`
 	
`*Note after collapsing the data, you have to load in the original data in order to run your regressions.`
	These commands show how to create a graph showing the number of schools that have each value of *school_enrollment*. First, we collapse the data to convert from school-grade level data to school level data. Second, we draw a graph of the counts of schools, restricting the graph to schools with between 20 and 60 students enrolled. Finally, we save the graph.

`*Load un-collapsed data`
`use grade5.dta, clear`
 	
`*Generate new variables`
`gen above40 = 0`
`replace above40 = 1 if school_enrollment > 40`
`gen x = school_enrollment - 40`
`gen x_above40 = x*above40`
 	
`*Run regression (all goes on one line)`
`reg yvar above40 x x_above40 if inrange(school_enrollment,0,80), cluster(schlcode)`
	These commands show how to run a regression to quantify the discontinuity in *yvar* at the 40 student threshold. We first generate an indicator variable for *school_enrollment* being above 40. We next generate a variable that equals *school_enrollment* minus 40 and the interaction term between this variable and the indicator for *school_enrollment* being above 40. Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *school_enrollment* between 0 and 80. The coefficient on *above40* is the estimate of the discontinuity in *yvar* at the threshold. We report standard errors that are clustered by school.

Table 2b
R Commands

R command	Description

`#Install and load rdrobust`
`install.packages('rdrobust')`
`library(rdrobust)`
 	
`#Subset data to observations in [20,60]`
`narrow <- subset(grade5, school_enrollment <= 60 &`
`school_enrollment >= 20)`
 	
`#draw binned scatter plot with linear fit`
`rdplot(narrow$yvar, narrow$school_enrollment, c = 40.5, p =`
`1, nbins = 20)`
`ggsave("figure1_linear.png")`
 	
`# draw binned scatter plot with quadratic fit`
`rdplot(narrow$yvar, narrow$school_enrollment, c = 40.5, p =`
`2, nbins = 20)`
`ggsave("figure1_quadradratic.png")`
	The first command installs rdrobust, which only has to be done once. The second command produces a binned scatter plot of *yvar* against the total school enrollment with a linear best fit line, restricting the graph to observations with total school enrollment in [20,60]. The last part shows how to change the best fit line to be quadratic by changing *p=1* to *p=2*.

`#Install and load dyplr`
`install.packages('dplyr')`
`library(dplyr)`
 	
`#Collapse data`
`by_school <- group_by(narrow, schlcode)`
`schools <- summarise(by_school, school_enrollment =`
`mean(school_enrollment, na.rm = TRUE))`
 	
`#Draw graph`
`ggplot(schools, aes(school_enrollment)) +`
`  geom_histogram(bins = 40) +`
`  geom_vline(xintercept=40.5,  color = "red")`
 	
`#Save graph`
`ggsave("school_counts.png")`
	These commands show how to create a graph showing the number of schools that have each value of *school_enrollment*. First, we collapse the data to convert from school-grade level data to school level data. Second, we draw a graph of the counts of schools, restricting the graph to schools with between 20 and 60 students enrolled. Finally, we save the graph.

`#For clustered standard errors`
`source("BM_StandardErrors.R")`
 	
`#Subset data and define indicator for above enrollment > 40`
`narrow <- subset(grade5, school_enrollment <= 80)`
`narrow$above40 <- 0`
`narrow$above40[which(narrow$school_enrollment > 40)] <- 1`
 	
`#Generate centered version of enrollment`
`narrow$x <- grade5_narrow$school_enrollment - 40`
 	
`#Generate interaction term`
`narrow$x_above <- narrow$above40*grade5_narrow$x`
 	
`#Run regression`
`mod1 <- lm(yvar~above40 + x + x_above, data = narrow)`
`summary(mod1)`
 	
`#Report clustered standard errors`
`clustervar <- as.factor(narrow$schlcode)`
`BMlmSE(mod1, clustervar, IK=F)`
	These commands show how to run a regression to quantify the discontinuity in *yvar* at the 40 student threshold. We first subset the data to observations with *school_enrollment* between 0 and 80. Next we generate an indicator variable for *school_enrollment* being above 40. We then generate a variable that equals *school_enrollment* minus 40 and the interaction term between this variable and the indicator for *school_enrollment* being above 40. Finally, we run a regression of *yvar* on these three variables, restricting the regression to observations with *school_enrollment* between 0 and 80. The coefficient on *above40* is the estimate of the discontinuity in *yvar* at the threshold. We report standard errors that are clustered by school, which will be reported as *$se.Stata* after running the last two lines.