**Project 2. Author: Matteo Pianella**

The following exercise is based on the second empirical project of the course "Using Big Data to Solve Economic and Social Problem", thought by Raj Chetty and Gregory Bruich at Harvard University. I found the material at opportunityinsight.org/course/.
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).

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.

If we regressed the result of test scores on class size, we would get a biased estimator due to the presence of variables that have an impact on test score and that are correlated with class size. For instance, the quality of the school (including school size) and demographic factors such as age, gender and race may be clustered by class size. Since these variables are usually associated with higher or lower test scores, if we fail to take those into account, the variable would be biased upward or downward depending on whether there is a positive or negative association between these factors and test score. Even if we can control for some of these characteristics, it is unlikely that we will be able to control for all of them. Therefore, we cannot show the causal relationship between test score and class size by a simple comparison of test scores, even by controlling for other variables.

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?

The Tennessee STAR experiment uses a randomization procedure to study causal relationship between class size and test score. It randomly assigned a group of 6,323 kindergarten students (K-3 grade) to small (target size 13-17 students) and regular-sized (target size 20-25 students) classes. The randomization gets rid of any possible clustering of other factors correlated with test score and it produces two statistically identical groups that only differ by their treatment status. In this way is possible to determine the impact of class size on test score all else equal.
The experiment shows a significant positive impact of the reduction in class size on students' test performance. However, the impact of class size on test score becomes statistically insignificant by 8th grade. Chetty et al. 2011 also studied the impact of class size on adult outcome by tracing 95% of the students to the United States tax record. They show that students assigned to small classes are significantly more likely to attend college and display other significant improvements on other outcomes in adulthood. It is likely that the long-term impact of class size is due to better non-cognitive skills.

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

The binned scatter plot is usually used with large datasets, when a simple scatter plot would produce an overcrowded table. A binned scatter plot shows the relationship between two variables by aggregating the data by groups that contain an equal number of observations. A binned scatter plot is produced in the following way: (1) the data on the x-axis are grouped into bins of an equal number of observations; (2) the means of the x-axis and of the y-axis variables are calculated for each bin; (3) these means constitute the data points that are displayed in a scatter plot; (4) the population regression line is superimposed to the scatter plot.
The binned scatter plot allows to compare a non-parametric estimate and the best linear estimate of the conditional expectation function. When the scatterpoints lie around the population regression line, then the slope is precisely estimated, otherwise it is not.

4) Graphical regression discontinuity analysis, focusing on the 40 student school enrollment threshold.

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 evolves 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.

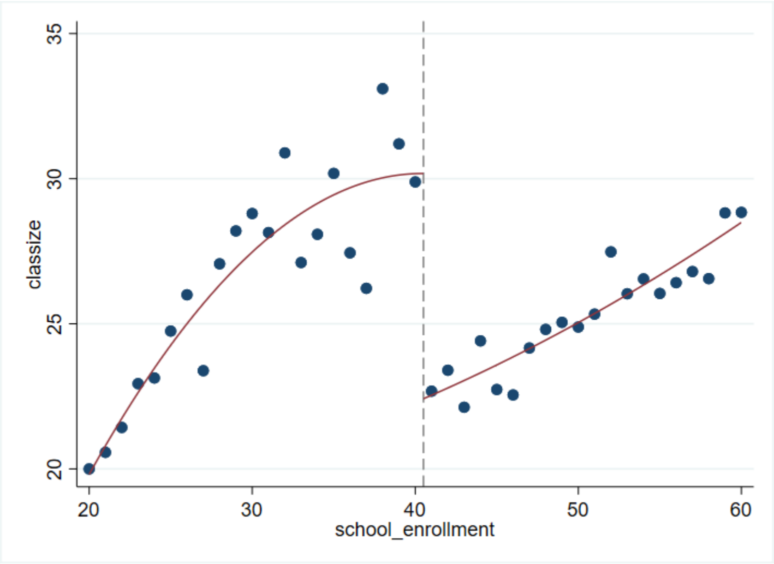

*Figure 1. The above scatterplot shows how the class size changes at the 40 students threshold.*

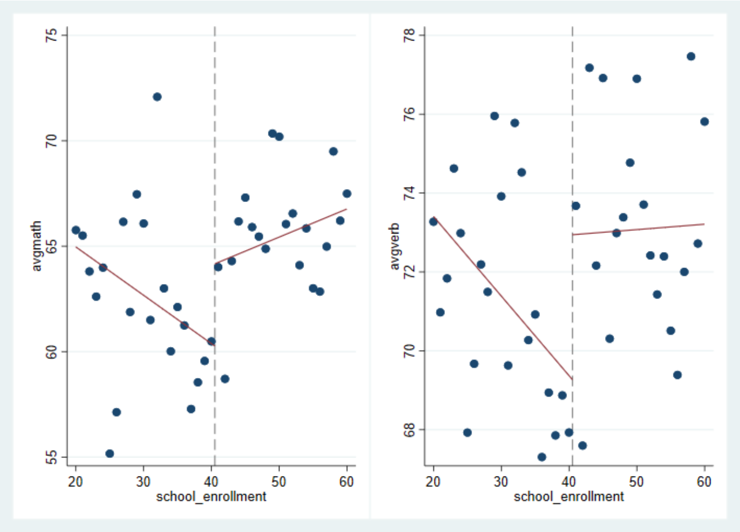

*Figure 2. The above scatterplots show how the average math and verbal score change at the 40 students threshold.*

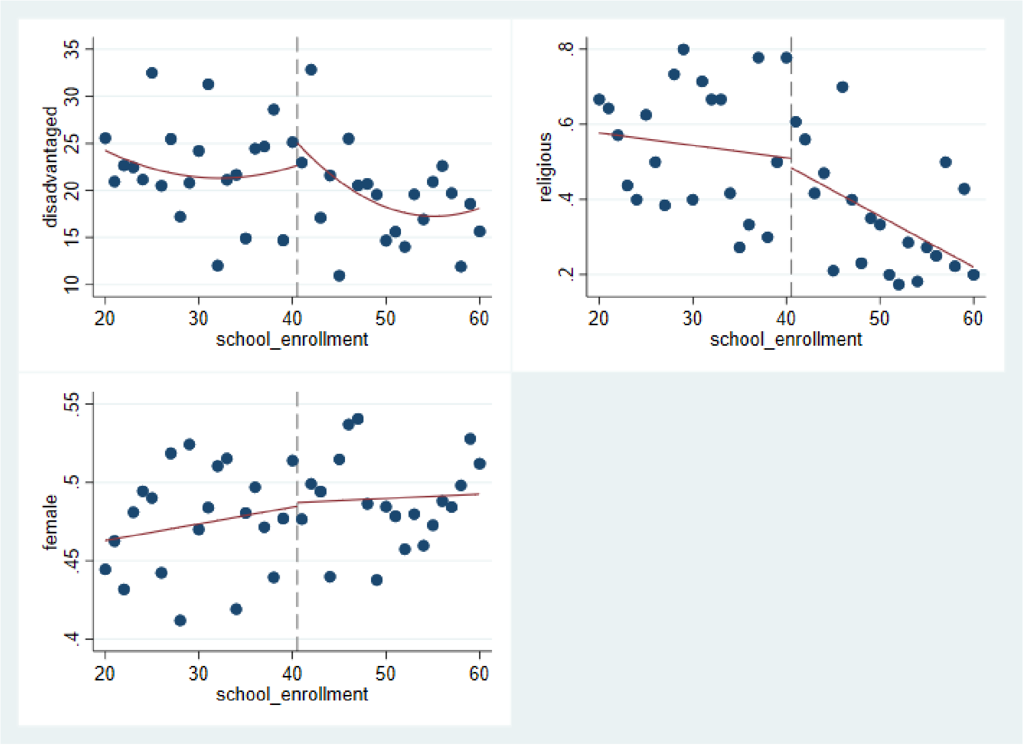

*Figure 3. The above scatterplots show how the percentage of disadvantage students, the fraction of religious schools, and the fraction of female students evolve across the 40 students school enrollment threshold.*

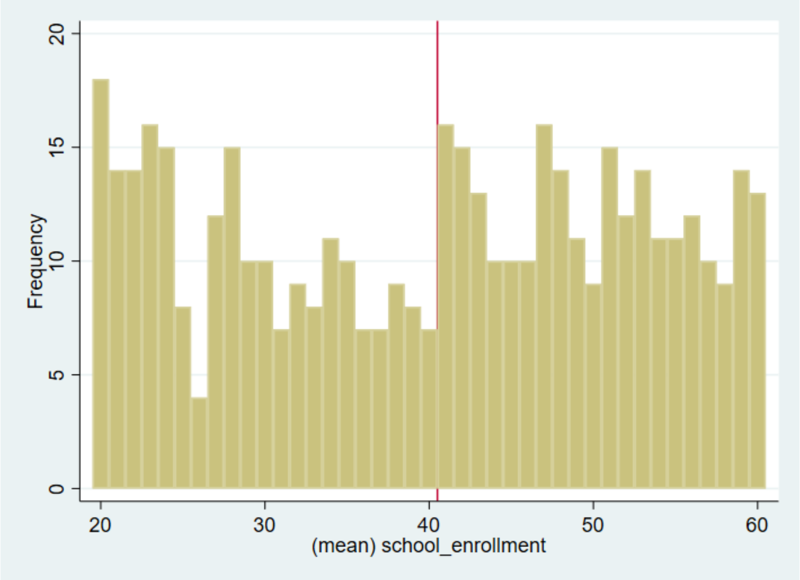

*Figure 4. The above histogram shows the frequency of school by total school enrollment.*

See the .do file and the .log file for the Stata program that generates the above figures.

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 the the .log file for point 5).

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.

A first identification assumption for the regression discontinuity design is that all explanatory variables except the treatment variable (in our case class size) should behave smoothly in the point where the discontinuity between the treatment and control occurs (in our case the 40 students threshold). This assumption is respected in our case. Indeed, figure 3 shows that the other 3 explanatory variables for which we have data behave smoothly across the threshold. If a discontinuity was present then this would have implied the presence of some correlation between class size and the explanatory variable.

The histogram in figure 4 shows a higher presence of schools with small classes compared to school with bigger classes and some discontinuity at the 40 students threshold. This may be due to an increased demand for these schools by high income parents who prefer to send their kinds to schools with small classes. If this was indeed the case, and small classes were attended by a higher percentage of students with high income parents, this would contradict the identification assumption. To test this idea we would need data on parental income.