# 14.771 - Problem Set 4

MIT

In this problem set, we will focus on the use of Regression Discontinuity Design (RDD) to perform causal inference with observational data.

You would like to study the impact of secondary education on labour market outcomes, such as wages, in Ghana. We have provided you with an (artificial) dataset that you will need to use to answer the questions in this problem set. In this data, you observe the wages in 2015 of workers in Ghana, whether they obtained secondary education or not, and their grade in a national entrance exam for secondary education. There are two cohorts in the data set, so workers will either have taken the exam in 2000 or in 2001. Note however that even though having an above median exam grade predict having secondary education, there is not a deterministic relationship between the two, there is only a change in probability of getting secondary education.

**Exercise 1**

Imagine that you have data on adults with the following information: (i) schooling level; (ii) wages. Assume, with these data, you estimate the regression

$$y_i = \alpha + \beta S_i + \varepsilon_i \, (1)$$

by OLS, where $y_i$ is wage and $S_I$ is a dummy indicating whether $i$ completed secondary education.

(i) What identification hypothesis is necessary to estimate the return of secondary education with this equation? (max 5 lines)
(ii) Mention at least 2 reasons why the identification hypothesis might be violated in this context. For each of these reasons, explain the direction of the bias of $\hat{\beta}_{OLS}$, the OLS estimator of equation 1. (max 3 lines per reason)

**Exercise 2**

At some point, you discover that in Ghana, there is a national entrance exam for eighth grade (first secondary education grade in this country), where each student receives a *continuous* score ranging from 0 to 100. The students are admitted if they receive an *above median grade*.

Assume you have individual-level information of the grade in the national exam for years 2000 and 2001. How do you propose to estimate the return to secondary education in this country? What identification hypothesis are necessary? (tip: note that you have multiple years of data, and the median grade might not be the same for each year).

**Exercise 3** Using the data set you received with this problem set, estimate the return to education using a regression discontinuity design (RDD). Your *running variable* should be the student's percentile in the year they took the exam.

Report the following:

1. The estimating equation.
2. The population parameter you are estimating (i.e., it is a treatment effect for what population?) (tip: who are the "compliers" here?)
3. Report the typical RDD graph for the "first stage" (i.e., a scatter plot between the running variable and secondary education). To improve readability, you might want to do a binscatter (though you don't have to).
4. Report the typical RDD graph for the "reduced form" (i.e., a scatter plot between running variable and wages).
5. Report the point-estimate using a bandwidth $h = 0.2$ (i.e., 20 percentage points on either side of the median) and a triangular kernel, reporting also the confidence interval. Interpret this coefficient. (max 3 lines).
6. Estimate an OLS model using equation 1 above. How do the coefficients differ? What do you think explains the difference between these coefficients?

**Exercise 4**

1. What is the McCrary test and how is it useful in the context of an RDD?
2. Give a reason as to why the McCrary test might not be particularly informative in this setting.
3. Report the density plot of the test score and the result of the McCrary test.
4. How does the result of the McCrary test make you feel regarding the identification hypothesis of the RDD you just ran?

**Exercise 5**

Assume now that entry to eight grade in this country is determined by a lottery, rather than by an entry exam. Each student draws a unique lottery number from 0-100 *at random* and you have access to the lottery number for each of them.

For each of the following two scenarios, explain whether you think you should estimate the return to education by IV or using an RDD approach. (max 5 lines per item).

1. Each student receives a lottery number at year $t$. If their lottery number is larger than 50 they are enrolled to secondary education at $t$. Otherwise, they are not enrolled. (note: The 50 cut-off is common-knowledge for families and students).
2. Each student receives a lottery number at year $t-1$. At year $t$, the government announces the cut-off, which is expected to be close to 50, but with some variance because of uncertainty in the number of 8th grade classes the government will open. Students are then enrolled at $t$. Can you think of a reason why would you prefer to do an RDD *around the cutoff*, rather than using the full sample in this scenario?

14.771: Development Economics
Fall 2021