Causal Inference for Policy Evaluation – Spring Semester 2025

Lab Assignment 4

Due date: Friday, 16.05.2025 (at 23:59)

Instructions

Please send your solutions to mario.bernasconi@unibas.ch. The solutions should include two files. One file with your code saved in .R format. A second file with written answers, including tables and graphs, should be sent as a .pdf file and is limited to a maximum of 2 pages (12 points, single spacing). PhD student groups are allowed one more page as they have one more question. Any text or output beyond that will not be considered.

Regression Discontinuity Design (25 points for MSc, 30 for PhD)

In 1998, the Refah Party was banned by the Constitutional Court for engaging in activities against the secular nature of the state. After the Refah Party was banned, many of its members regrouped under a new party, the Fazilet Partisi. In the 1999 municipal elections, Fazilet Partisi did relatively well, becoming one of the major opposition parties. The goal of this assignment is to use a new data set to repeat some of the analyses covered during the lab session. The data set contains the 1999 election data from TurkStat as well as data from the 2006 population census. Download the image file `meyersson_RDD2.Rdata` and open it via `load(.)`. Make sure that the dataframe `meyersson.data` as well as the covariate matrix `COVS` are loaded into your global environment. You need these data for questions 1 and 2.

1. Meyersson (2014) presents larger estimates for women living in municipalities that are more religious and poorer (using literacy rate as a proxy). You now have data on a more direct measure of poverty, with a variable `anhinc99` capturing the average net household income per municipality in 1999. You want to use this new measure to check whether the treatment effect differs by level of income.

(a) Estimate the treatment effects using the RD method for municipalities above or below the median value of income. Use a linear control for the running variable and options kernel="triangular" and bwselect="mserd". Do not include any covariate. Report the point estimates and s.e. for each sub-sample and explain what the options kernel="triangular" and bwselect="mserd" do. **(2 points)**

(b) Repeat the estimation in point (a), but now control for all the covariates in `COVS`. Report the point estimates and s.e. for each sub-sample. Compare the results form point (a) and (b) and comment. Should we expect them to be similar? **(2 points)**

(b) Meyersson argues that the treatment effect could be larger in poorer municipalities. Provide some intuition as to why this might be the case. Are the RD estimates from points (a) and (b) in line with this reasoning? **(1 point)**

(c) What would happen to your estimator if you halved the bandwidth around the cutoff? Discuss the implications for the RD estimator. **(2 points)**

(d) Conduct a suitable check that there is no selection in terms of household income at the cutoff. Comment on the result, the validity of an RD approach, and what we can learn from the estimates in points (a) and (b). **(2 points)**

2. Further investigate the validity of the Regression Disconuity Design with the following checks:

(a) Perform the McCrary (2008) density discontinuity test for the running variable and plot the frequency of the running variable. Describe the results and explain whether they are indicative for bunching behavior. **(2 points)**

(b) Check whether municipalities differ in terms of the percentage of population above 60 (`ageshr60`) and below 19 (`ageshr19`) by running separate RD estimations. As in question 2, use a linear control for the running variable and options kernel="triangular" and bwselect="mserd". Report the point estimates and standard errors along with the corresponding p-values. What conclusions do you draw with regard to sorting in covariates? Briefly comment. **(2 points)**

(c) Given the results from (1d), (2a) and (2b), can we use an RD approach to estimate the desired treatment effect? Briefly make the case for or against the RD design by discussing the validity of each of the identifying assumptions. **(3 points)**

3. We now perform a simulation exercise.

   (a) Write a function that takes as input a scalar for the sample size of our simulated sample and generates the following data:

   - $e^{sim} \sim N(0, 50)$ (an error term that follows a normal distribution with s.d. of 50 and zero mean, where the subscript 'sim' stands for simulated)
   - $X^{sim} \sim U(-50, 50)$ (a variable X^{sim} that is uniformly distributed from -50 to 50, i.e. the running variable)
   - $D^{sim} = \mathbf{1}\{X^{sim} > 0\}$ (a dummy D^{sim} for being above 0, i.e. a treatment indicator)
   - $DX^{sim} = D^{sim} \times X^{sim}$ (an interaction term to allow different slopes on the two sides of the cut-off)
   - $Y^{sim} = 2 - 5D^{sim} + 0.5X^{sim} - DX^{sim} + e^{sim}$ (the outcome variable)

   The function should additionally (i) use the 'rdrobust' command to estimate the parameter of interest with options p=1, kernel="triangular" and bwselect="mserd", (ii) estimate the parameter of interest with OLS (use robust s.e.). Use 'set.seed(123)' within the function (before generating the data) such that we all get the same results. **(4 points)**

   (b) Since we generate the data ourselves, we know the value of the treatment effect that we will then try to recover. What is the value of the treatment effect that we are interested in? **(1 point)**

   (c) Present a table with the RD point estimates, the RD s.e., the RD bandwidth, the OLS estimates and the OLS robust s.e. for a sample size of 5,000, 10,000 and 20,000. What do the results suggest about the precision of the RD estimator compared to OLS? Explain why this is the case. **(2 points)**

   (d) Given the data generating process in point (a), is the OLS estimator biased or unbiased for the treatment effect of interest? Why? Would you use the RD or the OLS estimates from point (c) for inference? **(2 points)**

4. **This question is for PhD students only.** Imagine the Turkish Statistical Institute would deliver less precise data for your analysis due to a new data protection guideline. Instead of reporting vote shares and the Islamic win margin with several decimal places,

they would now only report the vote shares of the Islamic party and the largest secular party rounded to 2 percentage points.

(a) How would you define the running variable based on this data and how would this coarser measurement affect its quality? Could you still estimate a valid RDD effect? Explain why or why not. **(2 points)**

(b) Would your conclusion change if you did not need to construct the running variable yourself but got it delivered as a rounded variable by the statistical office (still in increments of 2 percentage points)? Would the quality of your estimates be affected by the coarser measurement and if yes, how? **(3 points)**