Regression Discontinuity Design (RDD)

– Problem Set 3 –

In the problem set 3 we are going to practice RDD in Lee (2008) framework presented in the lecture 11. We employ the original simplified data set on the individual candidates for the US House of Representatives from 1946 to 1998. If a candidate obtains more votes than his or her competitors, he or she takes the office. Each elected candidate represents one of 435 congressional districts. The elections are held every two years. We seek the answer to the question whether winning the election has a causal influence on the probability that the candidate will win the next election.

The observations of the data set `individ_final.dta` are clustered by district and election year. It consists of the following variables:

1. *outcome* is a treatment variable; it is coded as 1 if a candidate won the election in the corresponding year and 0 – otherwise.

2. *outcomenext* is an outcome variable. It is coded as 1 if a candidate won the next election; as 0 if he or she did not win the next election; and as -1 if he or she did not participate in the next election.

3. *difshare* is an assignment variable; it is the winning candidate's vote share minus the vote share of the highest performing competitor. Therefore, 0 is the cutoff point: a candidate whose vote share is more than 0 is automatically assigned to treatment.

Task A. Theoretical foundation

(1) *What is the main assumption that makes RDD possible?* Define the local randomization condition in the simplified setup presented in the lecture.

Task B. Graphical presentation using local averages

A major advantage of the RD design over competing methods is its transparency, which can be illustrated using graphical methods. A standard way of graphing the data is to divide the assignment variable into a number of bins, making sure there are two separate bins on each side of the cutoff point. Then, the average value of the outcome variable can be computed for each bin and graphed against the mid-points of the bins.

(1) Create a new variable that groups the assignment variable values into 400 bins with a size of 0.005.

(2) Since we are interested in a causal influence on the probability that the candidate will win the next election based on winning the current election, drop the rows that do not have a comparable next election.

(3) Find the mean of the outcome variable for each bin or, in other words, local average. Draw this relationship on the scatterplot.

(4) For better visuality we also add to the graph the fitted values of logistic regression around the cutoff. For this apply logistic regression separately on either side of the threshold (we take the bins with the share values from -0.25 to 0.25 and use the package `LogisticRegression` from `sklearn.linear_model`). Extract probability estimates. Add them to the scatterplot in the proximity of cutoff. *Do you observe a discontinuity at the cutoff point?*

Task C. Local linear regression (LLR)

LLR as a method restricts the estimation to observations close to the cutoff. It is based on the assumption that regression lines within the bins around the cutoff point are close to linear. That helps to avoid some of the drawbacks of other parametric/non-parametrics approaches (Lee & Lemieux (2010))

(1) Run the LLR with a specification $Y = \alpha_r + \tau D + \beta X + \gamma X D + \epsilon$, where X is rectricted by a bandwidth: $h \geq X \geq -h$. Interpret the result. Experiment with few bandwidths on your choice.

Task D. Cross-validation

As you might find, the treatment effect result is sensitive to the bandwidth choice. In general, choosing a bandwidth in estimation involves finding an optimal balance between precision and bias. One the one hand, using a larger bandwidth yields more precise estimates as more observations are available to estimate the regression. On the other hand, the linear specification is less likely to be accurate (Lee & Lemieux (2010))

We are going to review one of the approaches for choosing a bandwidth – cross-validation "leave one out" procedure. The main idea is to take an observation i in the data, leave it out, run LLR, and use the estimates to predict the value of Y at $X = X_i$. Proceeding with each observation separately on each side of the cutoff, we obtain the predicted values of Y that can be compared to the actual values. The optimal bandwidth is then a value of h that minimizes the mean square of the difference between the predicted and actual values of Y. And overall mean square error is simply the average of the squares of the prediction errors on each side of the cutoff.

(1) If you want to practice your Python skills, we recommend to work with the packages `LeaveOneOut()` and `cross_val_score` from `sklearn.model_selection` and to write the code that finds the optimal bandwidth. Otherwise, we created our draft in `auxiliary.py`; you can use it to produce your solution. Draw the graph showing the relationship between the bandwidth and the mean square error. *What is the optimal bandwidth for LLR in our framework?*

References

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, *142(2)*, 675–697.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281-355.