Problem Set 08
Regression Discontinuity Design: Concepts, Estimation, and Diagnostics

TAs: Aida Hatami, Sam Fathinejad, Fatemeh Salehi, Reza Sahour

May 23, 2025

In this problem set, we will study the impact of class size on student performance using a Regression Discontinuity Design (RDD). The dataset `maimonides.dta` includes information on class size, school enrollment, and average test scores from 1003 public schools in Israel. These data replicate those used in Angrist and Lavy (1999).

1. State the key identifying assumption required for causal inference in an RDD. Discuss the role of smoothness in potential outcomes at the cutoff. Describe one graphical and one formal test that can be used to evaluate this assumption.

2. In an RDD setting, how does the relationship between the running variable and treatment status affect the choice of design? Based on the dataset you are using, assess whether a sharp or fuzzy RDD is more appropriate, and justify your reasoning.

3. What is the Local Average Treatment Effect (LATE)? Explain how it differs from the Average Treatment Effect (ATE) in terms of both interpretation and applicability. How does the nature of the RDD design affect which population is captured in the LATE?

4. Describe how manipulation in the running variable can threaten identification in an RDD. What signs in the data would suggest such manipulation? Mention a diagnostic method commonly used to detect this issue.

5. Design an example of an RDD from a real-world policy or institutional rule. Clearly identify the running variable, the cutoff, the treatment, and the outcome. Explain how you would assess whether the assumptions of RDD are satisfied.

6. Why are visual diagnostic tools central in RDD analysis? Name at least three visualizations commonly used in practice. For each, describe its purpose and a potential risk or misinterpretation associated with it.

7. RDD relies on the assumption that nothing else changes at the cutoff besides the treatment. Explain why this is important for internal validity. Provide an example where this assumption may be violated and describe the potential bias introduced.

8. Suppose in a regression discontinuity setting, treatment assignment does not change deterministically at the cutoff but instead changes probabilistically. Discuss how such a fuzzy design can be interpreted using the instrumental variables (IV) framework. Be precise in explaining the key conditions required for identifying a causal effect and the type of effect that is identified.

9. Discuss the importance of bandwidth selection in local polynomial estimation. What trade-off exists when choosing narrower versus wider bandwidths? How do kernel functions affect which observations influence the estimate?

10. Suppose you observe a clear discontinuity at your cutoff, but also at a placebo cutoff far from the threshold. What might this imply about your research design? Propose two robustness strategies to investigate whether the original result is credible.

11. Generate a table of descriptive statistics for the following variables: class size, math scores, verbal scores, school enrollment, and the percentage of disadvantaged students. Compare your results to Table I in Angrist and Lavy (1999). Note any small differences you observe due to variation in the provided dataset.

12. Plot average class size as a function of school enrollment. Try to replicate Figure Ia from the paper. Comment on whether the pattern reflects clear thresholds in class assignment rules.

13. Focus on math scores. Estimate the conditional correlation between class size and math performance using OLS regression. Include the proportion of disadvantaged students and total enrollment as control variables. Try to reproduce columns 4, 5, and 6 from Table II in the paper.

14. Estimate the effect of class size on math scores using a sharp regression discontinuity design (RDD).

    (a) First, use only the first threshold at 41 students. Construct a dummy variable for "large class" and estimate its effect on math scores. Restrict your sample to schools with enrollment between 20 and 60 students. Control for the share of disadvantaged students and include a linear trend in enrollment.

    (b) Next, use all class size thresholds. Define the predicted class size as:

$$\text{predicted\_size} = \frac{\text{enrollment}}{\text{int}\left(\frac{\text{enrollment}-1}{40}\right) + 1}$$

    Use this predicted value to replicate column 6 in Table III.

15. Estimate the effect of class size on math performance using a fuzzy RDD approach.

    (a) First, use the first discontinuity as an instrument for class size and estimate the treatment effect.

    (b) Then, repeat the fuzzy RDD using all thresholds (predicted class size). Try to reproduce column 8 of Table IV.

16. If the RDD is valid, then the coefficient of interest should not change significantly when covariates are included or excluded. Re-estimate your models from Questions 4 and 5 with and without the control for disadvantaged students, and examine whether this assumption holds in your case.

17. Using the same dataset, assess the credibility and validity of your RDD implementation following the checklist of Lee and Lemieux (2010):

    (a) Plot the distribution of the forcing variable (school enrollment). Is there any evidence of manipulation around the cutoff?

    (b) Present the main RD graph using binned local averages (binscatter) of math scores against school enrollment. Can you detect a clear discontinuity around the cutoff?

(c) Overlay linear and quadratic trend lines on the graph. Do these specifications capture the non-linearities in the data near the cutoff?

(d) Explore the sensitivity of your estimated treatment effect to the choice of bandwidth and the polynomial order used in the local regression.

(e) Run a placebo RD using a baseline covariate (e.g., the proportion of disadvantaged students) as the outcome. Should this variable be affected by the cutoff? Interpret your results.

Important instructions:

- Submit your code (Stata/R/Python) and all output files and graphs in a clearly labeled and organized format.
- All written explanations and interpretations must be handwritten. Scan or photograph your handwritten work and save it as a clearly labeled, single PDF file. Similarly, all outputs (code snippets, tables, and graphs) must be compiled into one organized PDF file. You may submit the two PDFs either separately or merged into a single comprehensive file.