Econometrics III: Applied Methods
NovaSBE, 2026

Problem Set 4: Regression Discontinuity Design

Exercise A: Assessing an RDD

This exercise asks you to critically evaluate an RDD identification strategy.

Mahzab (2023) studies the effect of electing "dishonest" politicians on public goods provision in Bangladesh using an RDD based on close elections. The setup is as follows:

- **Running variable**: Vote margin in local elections (winner's share minus runner-up's share)
- **Cutoff**: Zero (positive = honest politician won; negative = dishonest politician won)
- **Treatment**: Whether the elected politician is "dishonest"
- **Outcome**: Social safety net provision, public goods index, nighttime light growth

A politician is classified as "dishonest" if their reported income (from mandatory disclosure forms) exceeds the tax threshold but they did not pay income taxes. The paper finds that constituencies electing dishonest politicians have significantly lower public goods provision.

The paper reports the following descriptive statistics and balance tests:

Table 1: Constituency characteristics at the cutoff

Variable	Discontinuity at cutoff	Significant?
Total population	Small	No
Literacy rate	-2.9 pp	Yes ($p < 0.10$)
Electricity access	Small	No
School enrollment	Small	No
Pre-election nighttime light	Small	No

Table 2: Politician characteristics (full sample)

Characteristic	Honest	Dishonest	Significant?
Individual income (affidavit)	607,000 BDT	960,000 BDT	Yes
Total assets (affidavit)	4.0M BDT	5.5M BDT	No
Current criminal records	23.0%	33.8%	Yes
Years of schooling	13.78	13.57	No

1.1 Write down the continuity assumption required for identification in this RDD. What must be true about potential outcomes at the cutoff?

1.2 Consider the definition of "dishonest". What does this definition imply about the relationship between dishonesty status and politician income?

1.3 Based on the two tables above, assess the validity of this RDD. Explain. Suppose the true causal effect of dishonesty on public goods provision is zero. Could the paper's findings still arise? Under what conditions?

Exercise B: RDD Estimation in Practice

This exercise walks you through RDD estimation using simulated data from an educational intervention. Use R or Stata for the empirical questions. Both languages have the `rdrobust` *and* `rddensity` *packages.*

Setting: A university awards merit scholarships to students who score 70 or above on an entrance exam. The outcome of interest is first-year GPA. The dataset `rdd_scholarship.csv` contains:

- `student_id`: Unique identifier
- `score`: Entrance exam score (running variable)
- `scholarship`: Binary indicator for scholarship receipt
- `gpa`: First-year GPA (outcome)
- `family_income`: Family income (predetermined covariate)
- `high_school_gpa`: High school GPA (predetermined covariate)

B.1 Graphical Analysis

1.1 Create an RDD plot: Plot mean GPA in bins of the running variable (exam score), with separate local linear fitted lines on each side of the cutoff. Does there appear to be a discontinuity at the cutoff?

1.2 Create the same plot using three different bin widths: 1 point, 5 points, and 10 points. Comment on the trade-off you observe. Which bin width would you choose for presentation, and why?

B.2 Local Linear Estimation

2.1 Estimate the RDD effect "by hand" using local linear regression. Specifically, estimate:

$$Y_i = \alpha + \tau D_i + \beta_1(X_i - c) + \beta_2 D_i \cdot (X_i - c) + \varepsilon_i$$

where $D_i = \mathbb{1}(X_i \geq 70)$ and $c = 70$. Use only observations within a bandwidth of $h = 10$ points from the cutoff (i.e., scores between 60 and 80). Report $\hat{\tau}$.

2.2 Now use the `rdrobust` package (available in both R and Stata) to estimate the RDD effect with MSE-optimal bandwidth selection. Report:

- The point estimate and standard error
- The optimal bandwidth selected
- The effective number of observations used

2.3 Compare your "by hand" estimate (with $h = 10$) to the `rdrobust` estimate. Are they similar? If they differ, explain why.

B.3 Bandwidth Sensitivity

3.1 Estimate the RDD effect for bandwidths $h \in \{5, 7, 10, 12, 15, 20\}$ using local linear regression. Create a table showing:

Bandwidth	Estimate	Std. Error	N (left)	N (right)
5				
7				
$\vdots$				

3.2 Plot the point estimates with 95% confidence intervals as a function of bandwidth. Is the estimate stable across bandwidths? At what bandwidth do the confidence intervals start to widen substantially?

3.3 Explain the bias-variance trade-off you observe. Why might the estimate change as you widen the bandwidth?

B.4 High-Order Polynomials

4.1 Estimate the RDD effect using a global polynomial of order 4:

$$Y_i = \alpha + \tau D_i + \sum_{p=1}^{4} \beta_p (X_i - c)^p + \sum_{p=1}^{4} \gamma_p D_i \cdot (X_i - c)^p + \varepsilon_i$$

using all observations (no bandwidth restriction). Report $\hat{\tau}$.

4.2 Now estimate using polynomials of order 1, 2, 3, and 4. Create a table comparing the estimates. Do the estimates vary substantially with polynomial order?

4.3 Plot the fitted values from your order-4 polynomial specification. Comment.

4.4 Explain in your own words why Gelman and Imbens (2019) recommend against high-order polynomials. What are the two main concerns?

Exercise C: Diagnostics and Robustness

Continue using `rdd_scholarship.csv`*.*

C.1 Testing for Manipulation

1.1 Create a histogram of the running variable (exam scores) with bins of width 1 point. Does there appear to be any “bunching” just above or below the cutoff?

1.2 Use the `rddensity` package (available in both R and Stata) to formally test for a discontinuity in the density of the running variable at the cutoff. Report the test statistic and p-value. What do you conclude about manipulation?

1.3 Suppose you found evidence of bunching just above the cutoff. What would this suggest about the validity of the RDD? Who might be manipulating their scores, and in what direction would this bias the estimate?

C.2 Covariate Balance

2.1 Test for discontinuities in *predetermined* covariates at the cutoff. Estimate the RDD specification with `family_income` and `high_school_gpa` as outcomes (instead of GPA). Report the estimates and standard errors.

2.2 Why is covariate balance important for RDD validity? If you found a significant discontinuity in a predetermined covariate, what would you conclude?

2.3 Create a single figure showing both covariate balance tests (similar to your main RDD plot, but with predetermined covariates as outcomes).

C.3 Donut Hole Specification

A “donut hole” RDD excludes observations very close to the cutoff.

3.1 Re-estimate the RDD effect, but exclude all observations with scores in $[68, 72]$ (i.e., within 2 points of the cutoff). Compare to your baseline estimate.

3.2 Give two reasons why a researcher might use a donut hole specification, focusing on the context described in Exercise B.

3.3 What is/are the downside(s) of using a donut hole?

C.4 Asymmetric Bandwidths

Standard RDD uses the same bandwidth on both sides of the cutoff. Sometimes asymmetric bandwidths are appropriate.

The test scores in `rdd_scholarship.csv` are drawn from a normal distribution, and the cutoff of 70 is relatively high.

4.1 Create a histogram of the running variable. Comment on the distribution of observations above versus below the cutoff.

4.2 Given this distribution, should the MSE-optimal bandwidth be wider on the left or the right of the cutoff? Explain the intuition.

4.3 Estimate the RDD effect using `rdrobust` with the option for MSE-optimal *different* bandwidths on each side (`bwselect = "msetwo"`). Report the two bandwidths. Does the result match your prediction?

Exercise D: Fuzzy RDD

The dataset `rdd_fuzzy.csv` contains data from a setting where scholarship eligibility is determined by the same test score cutoff, but compliance is imperfect: some students above the cutoff decline the scholarship, and some below receive it through appeals. Variables are the same as before, but now `scholarship` does not perfectly correspond to `score >= 70`.

1 **First stage.** Plot the probability of receiving a scholarship against the running variable (binned scatter). Estimate the first stage discontinuity. What is the jump in treatment probability at the cutoff?

2 **Reduced form.** Estimate the reduced form discontinuity—the effect of *crossing the eligibility threshold* on GPA (not necessarily receiving the scholarship).

3 **Fuzzy RDD estimate.** Calculate the Wald estimate $\hat{\tau}_{Fuzzy} = \hat{\rho}/\hat{\pi}_1$ from your answers above. Verify using 2SLS with $\mathbb{1}\left(X_i \geq 70\right)$ as instrument for `scholarship`.

4 **LATE interpretation.** The fuzzy RDD identifies a Local Average Treatment Effect.

   (a) Who are the compliers, always-takers, and never-takers in this context?

   (b) Why might the fuzzy RDD estimate differ from the sharp RDD estimate in Exercise B?