DS-UA 201: Causal Inference: Regression Discontinuity Part 2

Parijat Dube

New York University Center for Data Science

August 16, 2022

Acknowledgement: Slides including material from DS-US 201 Fall 2021 offered by Marco Morucci.

Overview of the Sharp RD Estimator

Our goal is to identify the **local** effect of assignment to treatment D_i knowing that assignment is being driven by a **running** variable X_i and a cut-point c:

- Units with X_i above the cut-point receive treatment
- Units below the cut-point receive control

In this setting we can identify:

$$E[Y_i(1) - Y_i(0) | X_i = c].$$

with

$$\lim_{x \to c^+} E[Y_i | X_i = x] - \lim_{x \to c^-} E[Y_i | X_i = x].$$

Overview of the Sharp RD Estimator

RDD Estimation strategy:

1. Subset the data to only observations that are within h of the treshold
2. Fit one regression model of Y_i on X_i above the cut-point and another of Y_i on X_i below the cut-point.
3. Use the models to predict the value of Y_i at the cut-point.
4. The difference in predictions is the estimated treatment effect.

RDD Challenges:

- How do we choose a model?
- How do we choose h?
- Can we test the RDD assumptions?
- Imperfect treatment assignment at the threshold.

Illustration: Lee (2008) Election RDD

We'll use the Lee (2008) election dataset to illustrate our results

Research question: Does being an incumbent give you an advantage at the next election?

- Currently in power politicians have an easier time winning than runner ups for many different reasons
- identification problem: do they win because the people that previously elected them still like them or because they have an advantage?

Variables:

- X: Democratic margin of victory in time t
- Y: Democratic vote share in time $t+1$
- D: Victory in time t (margin > 0).

Design: Compare democrats that won by a small margin to democrats that lost by a small margin: similar voters but different incumbency outcomes.

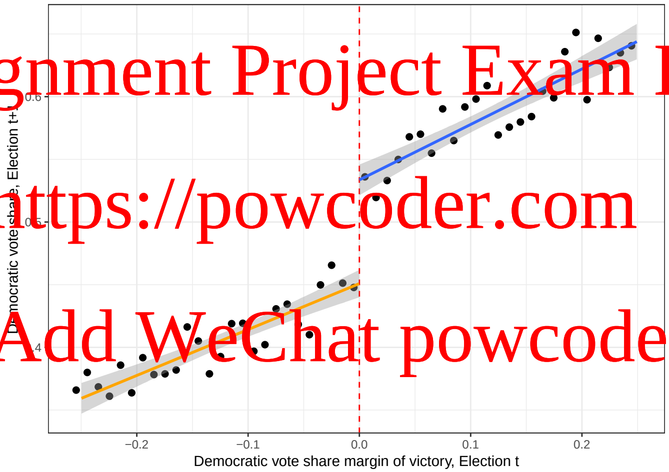
Illustration: Lee (2008) Election RDD

```
# generate a treatment indicator
house$d <- as.integer(house$x > 0)

# Subset to the close observations
house_close <- subset(house, x>-.25&x<.25)

# Fit the regression model w/ interaction
rd_reg <- lm_robust(y ~ d + x + d*x, data=house_close)
> rd_reg
              Estimate Std. Error t value Pr(>|t|) CI
                 Lower CI Upper
(Intercept)     0.4509    0.00558   80.82 0.00e+00
     0.4399    0.4618
d               0.0827    0.00838    9.87 1.31e-22
     0.0663    0.0991
x               0.3665    0.04135    8.86 1.37e-18
     0.2854    0.4476
d:x             0.0760    0.06288    1.21 2.27e-01
     -0.0473    0.1993
```

Illustration: Lee (2008) Election RDD


```
# Add it to the plot
# Scatterplot w/ regression
bin_scatter_close_reg <- ggplot(aes(x=x, y=y), data=
    house_close) +
  stat_summary_bin(fun.y='mean', bins=50,
                   size=2, geom='point') +
  geom_vline(xintercept=0, col="red", lty=2) +
  geom_smooth(data=subset(house_close,d==1), formula= y
      ~ x, method="lm_robust") +
  geom_smooth(data=subset(house_close,d==0), formula= y
      ~ x, method="lm_robust", col="orange") +
  xlab("Democratic vote share margin of victory
     Election t") +
  ylab("Democratic vote share, Election t+1") +
  theme_bw()
```

Illustration: Lee (2008) Election RDD



Model Choice in RDD

It is a popular idea to sometimes fit **polynomial regressions** to the data to estimate RDDs:

$$Y_i = \alpha + D_i\tau + X_i\beta_1 + X_i^2\beta_2 + X_i^3\beta_3 + \cdots + D_i(X_i\lambda_1 + X_i^2\lambda_2 + X_i^3\lambda_3 + \dots) + \epsilon_i,$$

This is usually advocated because:

- We need to predict the expected outcome well for the RDD to be valid
- The relationship between X and Y could be nonlinear

Illustration: Lee (2008) Election RDD

```
# Suppose we used a polynomial fit to the entire data!
bin_scatter_close_poly_full <- ggplot(aes(x=x, y=y),
    data=house) +
  stat_summary_bin(fun.y='mean', bins=100,
                   size=2, geom='point') +
  geom_vline(xintercept=0, col="red", lty=2) +
  geom_smooth(data=subset(house,d==1),
              formula= y ~ x + I(x^2) + I(x^3),
              method="lm_robust") +
  geom_smooth(data=subset(house,d==0),
              formula= y ~ x + I(x^2) + I(x^3),
              method="lm_robust", col="orange") +
  xlab("Democratic vote share margin of victory,
      Election t") +
  ylab("Democratic vote share, Election t+1") +
  theme_bw()
```

Illustration: Lee (2008) Election RDD

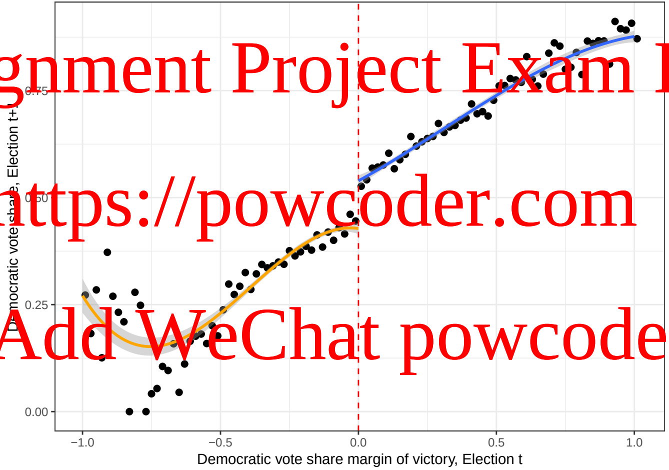


Illustration: Lee (2008) Election RDD

```
# Polynomial fit to the entire dataset gives the most
    extreme estimate
rd_reg_poly_full <- lm_robust(y ~ d*(x + I(x^2) + I(x
    ^3)), data=subset(house))
> rd_reg_poly_full
                Estimate Std. Error t value Pr(>|t|) CI
                   Lower CI Upper
(Intercept)      0.4278    0.00659   64.94 0.00e+00
    0.4149   0.4407
d                0.1115    0.00929   12.00 7.61e-33
    0.0933   0.1297
x               -0.0971    0.07859   -1.24 2.17e-01
   -0.2512   0.0569
I(x^2)          -1.7177    0.23197   -7.41 1.47e-13
   -2.1725  -1.2630
I(x^3)          -1.4636    0.17089   -8.56 1.33e-17
   -1.7986  -1.1286
d:x              0.4524    0.10366    4.36 1.29e-05
    0.2492   0.6556
d:I(x^2)         1.9109    0.28731    6.65 3.14e-11
    1.3477   2.4741
d:I(x^3)         1.2525    0.20437    6.13 9.37e-10
```

Caution with polynomials!

Even though polynomials are often used in RDD analysis, there are problems with them (Gelman and Imbens 2016)

- Noisy estimates
- Sensitivity to model specification and degree of polynomial
- Poor coverage of confidence intervals.

Intuition: Polynomial fit is closer to the datapoints, therefore:

- It might exaggerate the gap at the threshold due to noise
- Variance estimates will be smaller, therefore CIs will also be smaller

Suggestion: Either use linear, quadratic at most regressions, or other smooth functions.

Bandwith Choice in RDD

Another problem in RDD analysis is that of **choosing the bandwith** h.

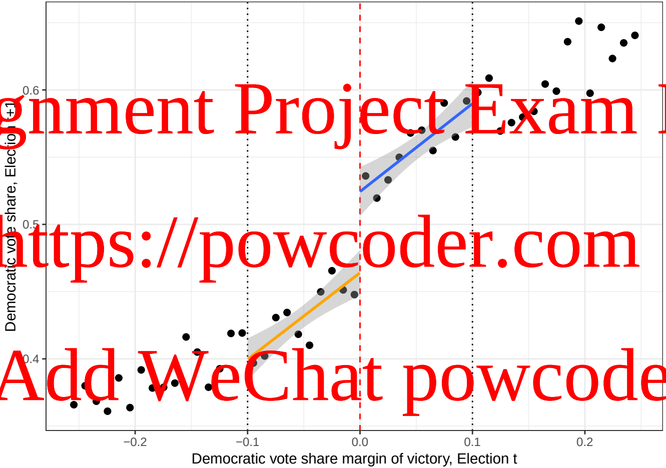
Choosing a bandwidth (h) for a local linear regression is a classic bias-variance trade-off.

- A metric for adjudicating between bias vs. variance is Mean Squared Error (MSE)

$$\text{MSE}(\hat{\tau}) = E[(\hat{\tau} - \tau)^2] = \text{Bias}(\hat{\tau})^2 + \text{Var}(\hat{\tau})$$

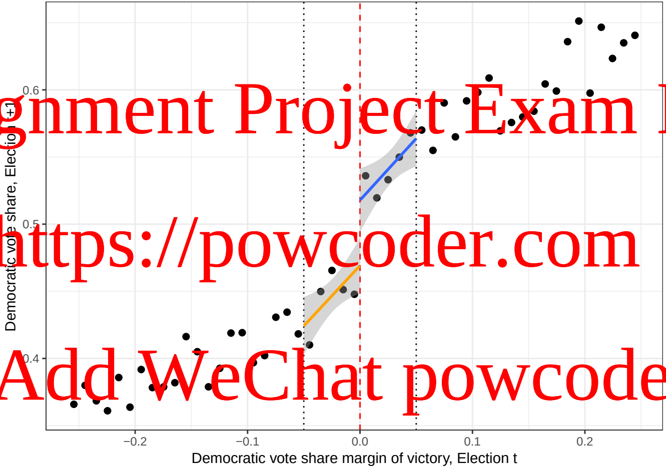
- Larger choices of h lead to smaller variance but larger bias
- Smaller choices of h lead to larger variance but smaller bias

Illustration: Lee (2008) Election RDD



- Bandwidth: 0.1
- Estimate: 0.06, 95% CI: [0.03, 0.08]

Illustration: Lee (2008) Election RDD



- Bandwidth: 0.05
- Estimate: 0.07, 95% CI: [0.03, 0.1]

Choosing a bandwidth

Strategies for choosing bandwidth:

- Plot estimates for a lot of bandwidths – assess "robustness" to bandwidth choice.
- Cross-validation: Randomly split the data into training and test sets – fit models of different bandwidths on training, compare predictive accuracy on the test for values of X_i close to the discontinuity (MSE)
- Optimal Bandwidth Selection: Imbens-Kalyanaraman (2008) develop a data-driven method for choosing h for local linear regression.

General intuition: Smaller samples $\rightsquigarrow$ larger bandwidth choices $\rightsquigarrow$ more dependence on the underlying model.

Overview of Fuzzy RDD

Last lecture, we looked at the case where treatment was perfectly determined by the running variable X_i

- Units with X_i above the cut-point c have $D_i = 1$
- Units with X_i below the cut-point c have $D_i = 0$

What if **not all units** above the cut-point receive treatment? What if not all units below the cut-point receive control?

Key idea: There is still a discontinuity at c in the probability of receiving treatment.

- The discontinuity is due an imperfectly applied rule
- For example: Van der Klauuw (2002) looks at the effect of financial aid on college enrollment, knowing that the aid assignment function for the university being studied incorporated cut-offs based on a GPA/SAT score index.

Visualizing Fuzzy RDD

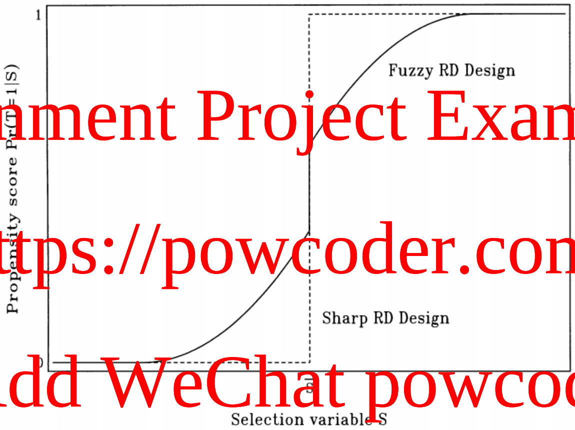


FIGURE 2

ASSIGNMENT IN THE SHARP (DASHED) AND FUZZY (SOLID) RD DESIGN

Figure taken from Van der Klauuw (2002) “Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach”

Fuzzy RDD Setup

- Running/forcing variable X_i with cut-off c.
- Z_i: Indicator for being above the cut-off.
- D_i: Actual receipt of treatment.
- Y_i: Outcome.

Fuzzy RDD Assumptions

Instead of assuming that the treatment assignment "jumps" from 0 to 1 at the cut-point c, the "fuzzy" RD design assumes that the **probability** of treatment is discontinuous at the threshold

Discontinuous Propensity of Treatment

$$\lim_{x \to c+} Pr(D_i = 1 | X_i = x) \neq \lim_{x \to c-} Pr(D_i = 1 | X_i = x)$$

In other words, there is a **subset** of units, who would take the treatment if they were above the discontinuity and control if below.

- Unobserved confounders could be affecting treatment take up!

Fuzzy RDD is IV

The Fuzzy RD set-up is **equivalent to an instrumental variables** design where the instrumental variable is the indicator for being above or below the cut-off.

Standard IV assumptions apply:

- Exogeneity of X_i (within the area around the discontinuity) – same as in the sharp RD design ("local randomization")

- Exclusion restriction (being slightly above the discontinuity only affects Y_i through its effect on D_i)
- Monotonicity (being above the discontinuity does not increase treatment propensity for some and decrease it for others).

Treating a fuzzy RD design as a "sharp" RD design gives us an **intent-to-treat effect** (what is the effect of being slightly above vs. below the discontinuity)

Fuzzy RDD is IV

Estimation is straightforward using classic **2SLS framework** w/ $(X_i - c)$ as a covariate in both regressions.

- Again, let $Z_i = \mathbb{1}(X_i \geq c)$ – the indicator for being above the threshold c.

First-stage:

$$D_i = \delta_0 + \delta_1 Z_i (X_i - c) + \delta_2 (1 - Z_i)(X_i - c) + \rho Z_i + \eta_i$$

Second stage:

$$Y_i = \beta_0 + \beta_1 Z_i (X_i - c) + \beta_2 (1 - Z_i)(X_i - c) + \tau D_i + \epsilon_i$$

Approach is equivalent to a Wald-type ratio estimator: the ratio of the sharp RD estimate over the estimated first-stage effect of the discontinuity on probability of treatment.

Fuzzy RDD is IV

As in IV, a Fuzzy RD effect is a **local** effect on **compliers**.

The Fuzzy RD identifies the Local Average Treatment Effect among compliers who would be induced to take the treatment if their X_i was slightly above the cut-off and take control if X_i were slightly below.

- We're adopting the "local randomization" interpretation of RDD – X_i is as-good-as-random within the vicinity of the cut-point c.

- Because treatment assignment is not deterministic, only some units would actually change D_i if they were above the cut-point vs. below. These are our "compliers"

The subset of units for which we estimate the ATE is even smaller!

- What can we say about populations of interest?ß

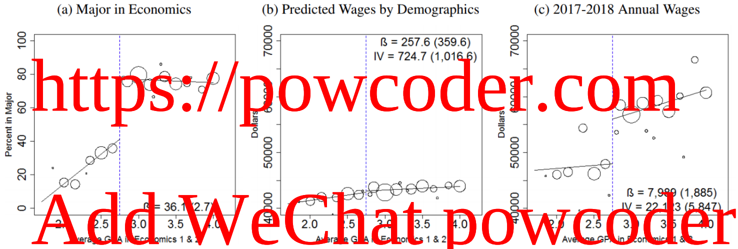
Example: Bleemer and Mehta (2020)

Question: Does majoring in Economics boost graduates' wages?

- Major choice is endogenous!
- Bleemer and Mehta (2020) leverage the fact that UC Santa Cruz's Economics department implemented a minimum GPA requirement in 2008 for students to be permitted to declare.
  - Students had to earn a 2.8 GPA in Econ 1 and 2 to declare. But policy was imperfectly implemented. Grades are relatively noisy, hard for students to manipulate to get exactly above the threshold.
- Students just above the threshold were about 36pp more likely to declare
- Being above the threshold raised post-graduation wages by about \$8,000 annually
- This generated an IV estimate of about a \$22,000 effect on annual early-career wages!

Example: Bleemer and Mehta (2020)

Figure 1: First-Stage and Reduced Form Effect of 2008-2012 UCSC Economics Major Restriction Policy



Assessing RD assumptions

How can we assess whether our "local randomization" assumption is plausible?

Problems:

- Units slightly above the discontinuity differ on pre-treatment covariates from those slightly below the discontinuity – imbalance.

- Units are able to selectively manipulate their score to land just above or just below the cut-point (essentially a kind selection effect on an unobservable)

Solutions:

- Balance tests (is there a covariate that also exhibits a discontinuity around c?)
- Placebo tests (does the discontinuity have an "effect" around some fake discontinuity)
- Density tests (is the density of X_i discontinuous in the area of the threshold)?

Density tests

If units are not able to manipulate their X_i, then the density of X_i around the discontinuity should be continuous. (McCrary 2008)

- We shouldn't expect surprisingly more or less units above versus below the cut-off

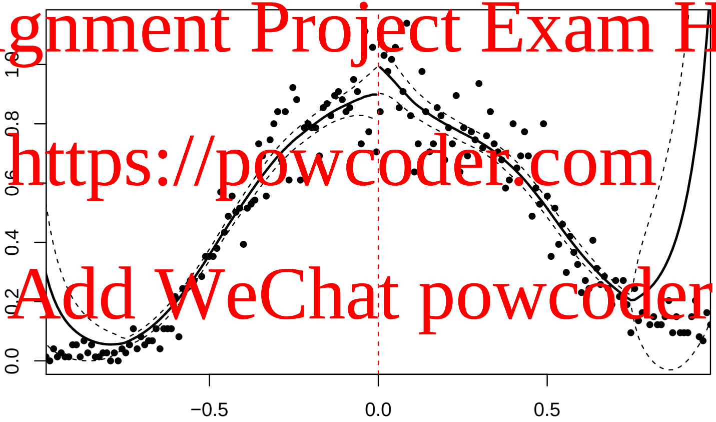
Intuition:

- Construct a histogram of the running variable (with bins selected to not overlap at the discontinuity)
- Smooth the histogram by fitting a local linear regression of the histogram heights on the bin mid-points
- Test for the difference in the smoothed histogram near the discontinuity

Implemented in the `rdd` package.

```
rdd::DCdensity(house$x, cutpoint=0)
```

Density tests



Summary

Today we looked at four issues in RDD designs:

1. Model and polynomial choices: try to use simple models.
2. Bandwidth selection: show estimates for many bandwidths.
3. Imperfect treatment assignment at the threshold: fuzzy RDD as IV.
4. Assumption checking for RDD: Placebo and density tests.

Last 4 lectures: Special topics.

- Sensitivity analysis
- Causal inference and ML
- Causal inference case studies
- Course review