Applied Microeconometrics I

Lecture 10: Regression discontinuity design

Stefano Lombardi

Aalto University

October 5, 2023

Lecture Slides

What did we do last time?

- **Difference-in-differences (DID)**:
  - RCT is seldom feasible, CIA is unlikely to hold, and good instruments are hard to find
  - Often there is a *before* and *after* a treatment is assigned (e.g. policy change) and some units are treated while others are not.
  - We need *panel data*

- **Idea**: we use the evolution of Y in the control group to build the counterfactual of "what would have happened" in the treatment group in the absence of the treatment.

- **Identifying assumption**: in the absence of the treatment, the outcomes of both groups would follow *parallel trends*.

- We allow for time-invariant (level) differences between the two groups, but rule out time-variant differences (trends).

What did we do last time?

- Regression form (FE notation):

$$Y_{i,t} = \alpha_i + \lambda_t + \rho D_{i,t} + \varepsilon_{i,t}$$

  - $D_{i,t} = 1$ if death penalty is abolished and zero otherwise
  - $E(\varepsilon_{i,t}|\alpha_i, \lambda_t, D_{i,t}) = 0$

- Simplest case: 2 time periods, 2 cross-sectional units (States)
  - Comparison of the two States across time periods gives ρ

- An equivalent way to write the 2x2 DID equation:
$$Y_{ist} = \alpha + \gamma TREAT_s + \lambda AFTER_t + \rho(AFTER_t * TREAT_s) + \varepsilon_{ist}$$

- Threats to validity:
  - Lack of parallel trends
  - Anticipatory behavior (e.g. leading to GE effects)

- ρ interpretation: shocks/policies implemented at same time?

What did we do last time?

- Several model extensions:
  - Add controls X (to achieve identification)
  - Multiple time periods (before and after)
    - inspect pre-trends
    - placebo policy changes
    - dynamic effects
  - Multiple treated groups, treated at different times
  - Different types of controls (never-treated; treated later)
  - i and t can be *anything*!
    - Jimeno and Boeri (2005): “workers” and “firms”

Today: Regression Discontinuity Design

- Exploit institutional rules that create RCT's
  - a known (to us) threshold rule assigns people to treatments
  - close to the threshold, the treatment assignment is randomized
  - if nothing else discontinuously jumps at the threshold (other than the treatment status), we can causally relate outcome change to the treatment change at the threshold
- Randomization in *observational data*
- Identification: Imperfect manipulation, continuity assumption
- Types of RDD: sharp vs. fuzzy
- Sharp RDD:
  - model specification, estimation, and testing validity
  - sharp RDD as a local RCT

Regression discontinuity design (RDD)

Institutional rules create experiments

- Institutional rules often assign individuals to "treatments"; this can be exploited for estimating causal effects
- Typical case: **threshold rule** that is based on ex-ante variable
  - *Score* in entry exams for college admission
  - *Income* for subsidy eligibility
  - *Project quality score* for public R&D subsidies eligibility
  - *Age limit* for legal alcohol consumption
- The ex-ante variable X is called the **running variable** (W refers to pre-determined observable characteristics)
- A pre-determined **threshold** c of X assigns individuals to the binary **treatment status** $D \in \{0, 1\}$
  - $D = 1$ if treated and $D = 0$ if not treated
  - In principle, D can be continuous

Regression discontinuity design

Institutional rules create experiments

- RDD exploits the **randomness** of the treatment assignment **around the threshold**
- Provided that D is as good as randomly assigned at c, we can compare the Y below (control) and above (treated) c.
- **Main assumption**: *people cannot perfectly manipulate* X
  - X manipulation means being able the set own D_i value
  - some manipulation is allowed
  - but no one can perfectly decide to be just below/above c
  - imperfect self-selection around c cannot be fully tested

Regression discontinuity design

Imperfect manipulation and identification

- *Imperfect manipulation*:
  - It **implies** that X is continuous around c
    - lack of continuity is consistent with precise sorting around c
  - The continuity of X implies *local randomization*
  - Randomization is key to identify the causal estimand of interest

- **Why does RDD work?**
  1. Aim: estimate effect of D on Y for people at the threshold
  2. Institutional rules assign D based on X (above/below c)
  3. At the threshold, we want to causally relate the change in Y with the discontinuity in D induced by the running variable
  4. For this to work, *no other variable can jump at* c
  5. In other words, any other discontinuities at c (in either X or observable/unobservable characteristics) will bias the estimates
  6. Continuity implies randomization: people close to c are similar

Regression discontinuity design

Continuity assumption

- Why do we require continuity (and not randomization)?
- In RDD:
  - All pre-determined observable and unobservable characteristics (W and U) are **allowed to be correlated with** D! (and Y)
    - we only require that they evolve smoothly around c.
  - X is also **allowed to be correlated with** W and U
    - we require that, conditional on (W, U), X passes smoothly across c
- Assuming that D is as good as randomly assigned conditional on (W, X) would mean relying on CIA (and we don't want to!)
- Remember, randomization *follows* from X continuity at c.
  - X continuity ensures that at c the effect of the discontinuous jump of D on Y is not confounded by a jump in X
  - X continuity implies that, conditional on X, on average the potential outcomes evolve smoothly across c

Regression discontinuity design

Consequences of the local random assignment

- Local random assignment means that treated and comparison units are not systematically different in any dimension at c
  - It is "pure luck" whether i is just above vs. below c
  - This follows from the inability of precisely manipulating X to be on one side vs. the other
- Three implications local random assignment:
  1. **Identification**: It is a necessary condition (and sufficient, in sharp RDD) to identify causal effect of D on Y at c
  2. **Validity testing**: test whether pre-determined W jump at c (similar to randomization test in RCTs)
  3. **Irrelevance of W**: their inclusion/exclusion should not matter for consistency (only for precision)

Types of Regression discontinuity designs

- Overall, RDD is compelling design (high internal validity):
  - local RCT
  - if X and c are known to the researcher, one can exploit local randomization *in observational data*
  - local randomization does not need to be assumed, it *follows* from imperfect manipulation of X.

- **Sharp RDD**:
  1. Treatment status D is a deterministic function of X
  2. D is a discontinuous function of X

  That is, D jumps from 0 to 1 at c (*everyone* with $X_i \geq c$ receives $D_i = 1$).

- **Fuzzy RDD** allows for a smaller jump at c.

Types of Regression discontinuity designs

- We can often redefine D to have a sharp vs. fuzzy design:

X	D	Y	RDD
Test score	college admission	Income	sharp
Test score	college enrollment	Income	fuzzy
Income	welfare benefits eligibility	Employment	sharp
Income	welfare benefits take-up	Employment	fuzzy
Age	legal alcohol consumption	Death	sharp
Age	actual alcohol consumption	Death	fuzzy

TABLE 1 – Example of RDD applications

- Typically:
  - Sharp RDD if X fully/deterministically captures the eligibility
  - Fuzzy RDD if D is participation/take-up
  - Fuzzy RDD if D assignment follows more rules than just X

Effect of Minimum Legal Drinking Age on death rates

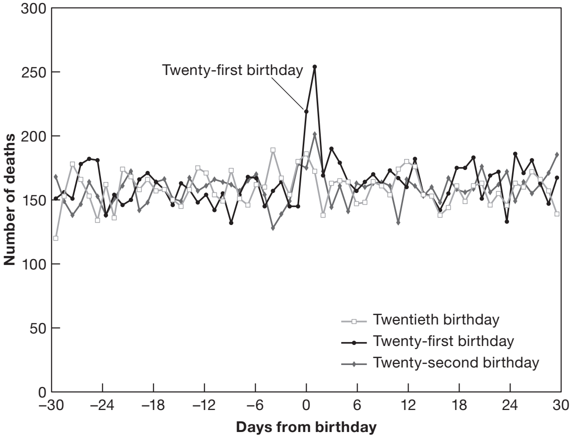
Carpenter and Dobkin (2009)

- We are interested in the effect of the minimum legal drinking age (in the US it's 21) on mortality.
- What type of RDD is this?
- Setting:
  - outcome variable Y_i: death rate
  - treatment status D_i: legal drinking status
  - running variable X_i: age
  - cutoff $c = 21$: MLDA rule transforms 21-year-olds from underage minors to legal alcohol consumers.

Effect of Minimum Legal Drinking Age on death rates

Carpenter and Dobkin (2009)

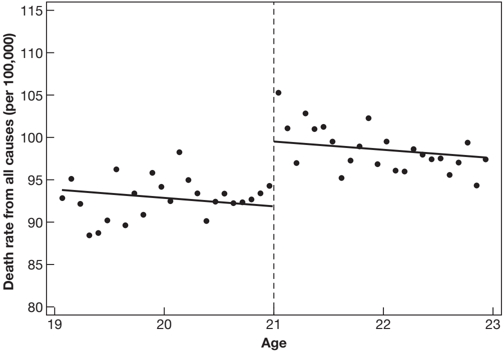
FIGURE 4.1
Birthdays and funerals



Effect of Minimum Legal Drinking Age on death rates

Carpenter and Dobkin (2009)

FIGURE 4.2
A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

Sharp Regression Discontinuity Design

- Suppose that D_i is a deterministic and discontinuous function of the running variable X_i
- D jumps from 0 to 1 at c. For each i:

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

  Or, more compactly, $D_i = 1[X_i \geq c]$.

- **All people** to the right of the cut off are exposed to the treatment and **all those** to the left are denied the treatment

Sharp Regression Discontinuity Design

Linear case

- What is the effect of *college admission* (D) on *income* (Y)?
- We have information on the college admission rules:
  - admission is entirely based on applicant's *entry test score* (X).
  - students scoring at least c are admitted: $D = 1[X \geq c]$.
- We can write the model for Y_i as:

$$Y_i = \alpha + D_i\tau + X_i\beta + u_i$$

- Note that:
  - Y_i varies discontinuously with D_i (if $\tau \neq 0$)
  - D_i is discontinuous function of X_i generating treat. effect τ
  - here, *by assumption* the relationship between Y_i and X_i is linear (and the same above/below c)
- Side note: can we use X as an instrument for D?

Sharp Regression Discontinuity Design

Linear case

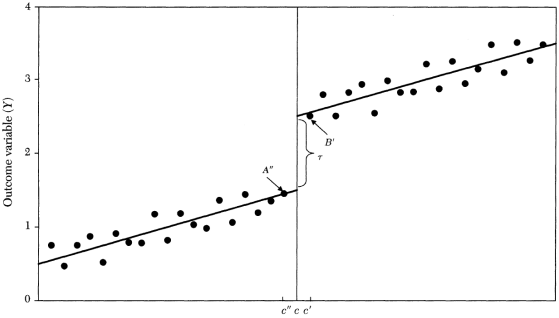


FIGURE 1 – Linear RDD

- The causal effect for a person whose $X_i = c$ requires us to *extrapolate* beyond c
- Suppose that all factors (other than college admission) affecting Y_i evolve smoothly with respect to X_i
- Then $B^{'}$ $(A^{''})$ is a reasonable guess for Y_i when $D_i = 1$ $(D_i = 0)$
- And $B^{'} - A^{''}$ would be the impact of treatment on Y_i

Sharp Regression Discontinuity Design

Specifying the functional form

- **Tradeoff**:
  - Estimates are more accurate, the closer we are to the thershold
  - The closer we are to the threshold, the less data we have
- We *always* need to use data away from the threshold
  - In the example, no individual had $X_i = c$!
  - extrapolation means that we must *always* control for X_i
- As a result we need to **assume a functional form** for the relationship between Y and X
  - In the example, a linear relationship seems appropriate
  - If the true population model is linear, OLS using all observations provides the best estimate for τ
- Since we need data away from threshold, the τ **estimate depends on chosen functional form**
  - In the example, if we restricted $\beta = 0$, OLS would biased

Sharp Regression Discontinuity Design

Specifying the functional form

- We saw the linear example
- But in general the relationship can be any $f(X_i)$:
$$Y_i = \alpha + \tau D_i + f(X_i) + u_i$$
- For example, ρ:th order polynomial:
$$f(X_i) = \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + ... + \beta_\rho X_i^\rho$$
- $f(X_i)$ can also be different at each side of the cutoff point
- Relies on the assumption that $f(X_i)$ is an adequate description of the relationship between Y and X
- The further away from c we are, the bolder this assumption is
- Therefore, it makes sense to use estimation methods for $f(X_i)$ well-suited *locally*

Sharp Regression Discontinuity Design

Nonlinear case

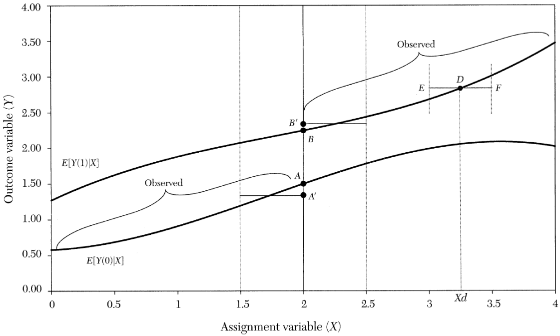


FIGURE 2 – Non-linear RDD

- $Y_i(0)$, $Y_i(1)$: potential outcomes. The i-level treatment effect is never observed.
- $E[Y_i(0)|X]$, $E[Y_i(1)|X]$: underlying relationships between the average outcomes and X (non-linear; different above/below c).
- We *never* observe $E[Y_i(1)|X]$ below $c = 2$ (the opposite holds for $E[Y_i(0)|X]$).

Sharp Regression Discontinuity Design

Nonlinear case

- With what is observable, we can estimate:

$$B - A = \lim_{\epsilon \to 0} E[Y_i | X_i = c + \epsilon] - \lim_{\epsilon \to 0} E[Y_i | X_i = c - \epsilon]$$

- Assume that X evolves continuously across c. This implies that also the underlying $E[Y_i(0)|X]$ and $E[Y_i(1)|X]$ do so.
- Then $B - A$ at the limit is equal to:

$$\tau = E[Y_i(D_i = 1) - Y_i(D_i = 0) | X_i = c]$$

- This is the treatment effect at the thershold c
- Continuity enables us to use the average outcomes of those right below c and valid counterfactuals for those right above
- We can *estimate* smooth $E[Y|X]$ relationships before/after c, but must *assume* relationship is smooth in the counterfactuals

Sharp Regression Discontinuity Design

Estimation within a bandwidth

- How to estimate $E[Y_i|X_i = c + \epsilon]$ and $E[Y_i|X_i = c - \epsilon]$?
    - $f(X_i)$ approximates well relationship between X and Y?
    - The further away from c we are, the bolder this assumption is
    - We need estimation methods for $f(X_i)$ well-suited *locally*
- **Non-parametric methods** flexibly estimate $E[Y|X]$ below and above c within a bandwidth (window) of width h.
- E.g., *local constant estimator* (Figure 2):
    - choose interval h around $c = 2$, e.g. $(-1.5, 1.5)$
    - use data in $(-1.5, 2)$ to compute A'; repeat above c to get B'
    - $B' - A'$ is the estimate for τ
- Alternatively, *local linear/polynomial regression*, and we can give more weight to observations closest to c
- How to choose h? **Precision vs. bias tradeoff**
    - Literature on optimal bandwidths

Model specification and estimation

Sharp RDD

- Standard approach is to choose a window of width h and use *local linear regression* with rectangular kernel.

- This is equivalent to estimating two linear regressions over window h on both sides of c:

$$\begin{aligned} Y_i &= \alpha_l + \beta_l \tilde{X}_i + u_i \quad \text{if } \tilde{X}_i \in [-h, 0) \\ Y_i &= \alpha_r + \beta_r \tilde{X}_i + u_i \quad \text{if } \tilde{X}_i \in [0, h] \end{aligned}$$

  - with $\tilde{X}_i = X - c$, hence $D_i = 1[\tilde{X}_i \geq 0]$
  - this way, α_l α_r are the value of the regression functions at c.
  - then, $\tau = \alpha_r - \alpha_l$

- Equivalently, **do OLS** once to estimate *pooled regression*:

$$Y_i = \alpha_l + \tau D_i + \beta_l \tilde{X}_i + (\beta_r - \beta_l) D_i \tilde{X}_i + u_i, \quad \tilde{X}_i \in [-h, h]$$

  - again, $\tau = \alpha_r - \alpha_l$, but now we directly get $\hat{\sigma}(\hat{\tau})$

Model specification and estimation

Sharp RDD

- Note that in general we can rewrite the pooled regression as:

$$Y_i = \alpha_l + \tau D_i + f(\tilde{X}_i) + u_i, \quad \tilde{X}_i \in [-h, h]$$

with $f(\tilde{X}_i) = f_l(\tilde{X}_i) + D[f_r(\tilde{X}_i) - f_l(\tilde{X}_i)]$

- for local linear regression: $f_l(\tilde{X}_i) = \beta_l \tilde{X}_i$ and $f_r(\tilde{X}_i) = \beta_r \tilde{X}_i$
- we can alternatively use local polynomial regression for $f(\tilde{X}_i)$
- alternative non-parametric methods exist, but local linear regression with rectangular kernel (same weight to all observations in h) is a good start
- optimal bandwidth algorithms and manual inspection help deciding h value (and robustness of results to it)

Testing for sharp/fuzzy RDD validity

- We cannot test whether individuals influence the running variable precisely. But we can check:
  1. **institutional rules defining** X_i:
     - is it likely/feasible to precisely sort on one side vs. the other?
     - do people even know about c?
  2. **local random assignment**:
     - do pre-determined W jump at c? (similar to balance test in RCTs)
     - does W inclusion change estimates? (should only affect precision)
  3. **continuity** (McRary test):
     - Is the density of X continuous across c?
  4. **relationship between** Y **and** X
     - are results robust to the use of different h? (bias-variance trade-off)
     - what is the optimal bandwidth?
     - are estimates sensitive to different functional forms for $f(X_i)$?
- **Transparency**: plot Y vs. X to *see* variation that identifies τ