XII. Regression Discontinuity Design

Alexandra Avdeenko

Evaluation

2022

Announcement: Exam

Preview of exam structure:

The exam will have three parts

A Multiple choice

B Theory questions and exercises; Questions on unknown research design example

C Questions on known research paper

This week: Send your questions for the review session next week.

Sum up: Careers in Evaluation

Academic	**Policy**
Long deadlines	Short deadlines
Self motivate	Good in teams
Being novel	Being right
Direction of effect	Magnitude of effect
Convince economists	Convince noneconomists
Find a question you can answer well	Answer the question as well as you can
Become an expert on one issue	Apply your tools to many issues
Find the optimal	Optimize within constraints

Source: Glennester, JPAL, Link

Randomization vs Natural Experiments

Last weeks

- we got to know natural experiments (NEs)
- reliance on 'natural' sources of randomness
- in contrast to controlled experiments, in NEs randomization and treatment design are not controlled by researcher
- makes NE studies observational rather than experimental
- NEs 'discovered' rather than designed

Difference-in-differences (diff-in-diffs)

- Estimates the effect of an intervention by comparing how the outcome variable changes among treated subjects versus untreated subjects
  - Diff-in-diffs uses observational panel data and has a better chance to estimate the effect of an intervention than using cross-sectional data, because it conditions on pre-intervention outcomes
  - Diff-in-diffs gives a good estimate of the effect if the parallel trends assumption holds
  - We can't test the parallel trends assumption directly, but examining pre-intervention trends can give indirect evidence

Matching (on the Propensity Score)

Matching is the process of closing back doors between a treat- ment and an outcome by constructing comparison groups that are similar according to a set of matching variables. Usually this is applied to binary treated/untreated treatment variables, so you are picking a "control" group that is very similar to the group that happened to get "treated."

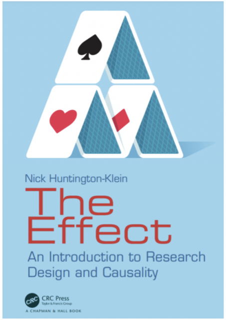
- Matching and then difference taking is repeated for all $x = 1$ observations.
- The estimated effect of x on y is then the average of those differences.

- If all confounders are included, the propensity score incorporates all endogenous sources of variation in the causal variable.

- In practice, many possible decisions...

Regression Discontinuity Design (RDD)

- Bekes book: Chapter 21.10;
- Further book: “The Effect: An Introduction to Research Design and Causality”
- by: Nick Huntington-Klein



Link

RDD: requirements and intuition

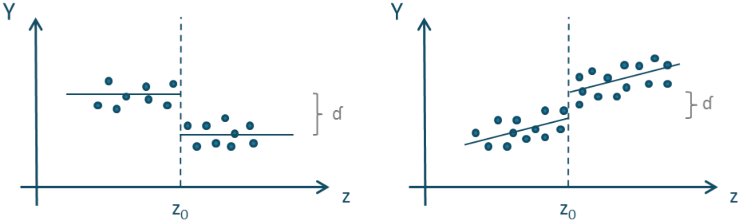
- Today we focus on a particular type of natural experiment: 'regression discontinuity designs'
- *idea:* exploit particularities in laws and institutions, usually thresholds
  - treatment occurs just above (below) threshold
  - control group are observations just below (above) threshold
- *examples:* date of birth cutoffs, class size, marginal admits/rejects
- *identification requirement:*
  discontinuity ('jump') in the mapping $Z \rightarrow D$ at some threshold z_0
- *identifying assumption:*
  continuity of $Z|D \rightarrow Y$ (in potential outcomes) and of other variables ($Z \rightarrow X$) around threshold z_0

$\Rightarrow$ 'identifying requirement' and 'identifying assumption' are necessary and sufficient, respectively, for identification of causality

RDD: requirements and intuition

1. Running variable
   - The running variable, also known as a forcing variable., is the variable that determines whether you're treated or not. For example, if a doctor takes your blood pressure and will assign you a blood-pressure reducing medicine if your systolic blood pressure is above 135, then your blood pressure is the running variable.
2. Cutoff
   - The cutoff is the value of the running variable that determines whether you get treatment. Using the same blood pressure example, the cutoff in this case is a systolic blood pressure of 135. If you're above 135, you get the medicine. If you're below 135, you don't
3. Bandwidth
   - It's reasonable to think that people just barely to either side of the cutoff are basically the same other than for the cutoff. But people farther away (say, downtown San Diego vs. further inside of Mexico) might be different for reasons other than the cutoff. The bandwidth is how much area around the cutoff you're willing to consider comparable. Ten feet to either side of the US/Mexico border? 1000 feet? 80 miles?

Illustration of RDD as a graph



- simple examples of RDD with single linear equation models
- left: even diff-in-means calculationon either side of z_0 sufficient

Two types of RDD: characteristics

Two types of RDD

1. 'sharp' or 'clean' RDD
   - D is *deterministically* related by Z
   - i.e. *all* subjects above (below) threshold z_0 are treated ...
   - ... while *all* subjects below (above) threshold z_0 are untreated

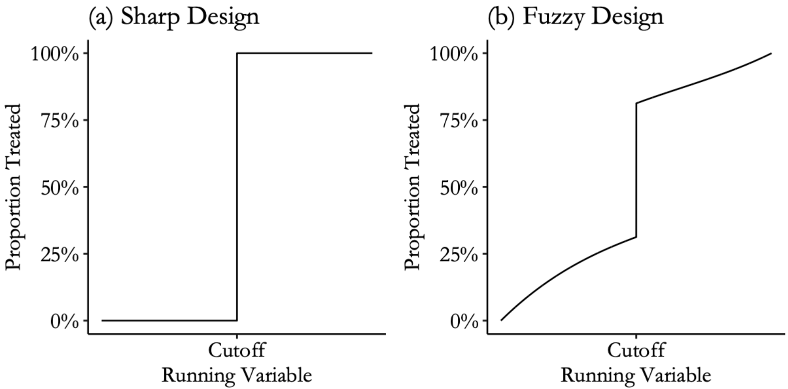
Two types of RDD: characteristics

Two types of RDD

1. 'sharp' or 'clean' RDD
   - D is *deterministically* related by Z
   - i.e. *all* subjects above (below) threshold z_0 are treated ...
   - ... while *all* subjects below (above) threshold z_0 are untreated

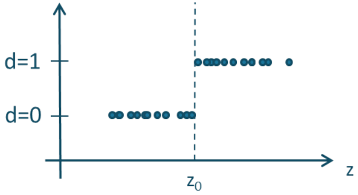
2. 'fuzzy' RDD
   - D is *probabilistically* related by Z
   - i.e. treatment probability is higher for subjects above (below) threshold z_0 than for those just below (above)

d(z) in clean/sharp (left) and fuzzy (right) RDD

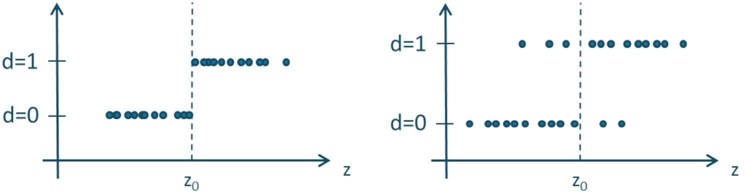


Link

d(z) in clean/sharp (left) and fuzzy (right) RDD

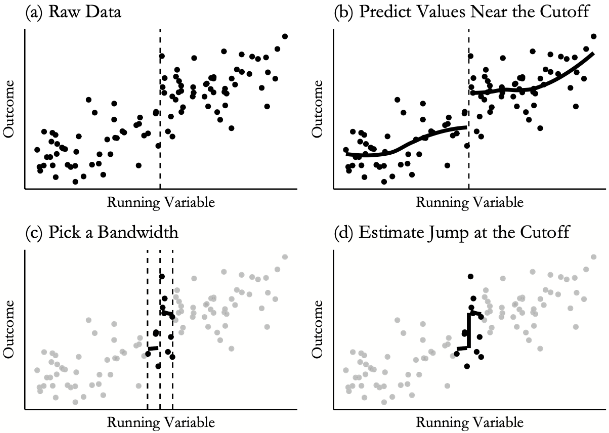


d(z) in clean/sharp (left) and fuzzy (right) RDD



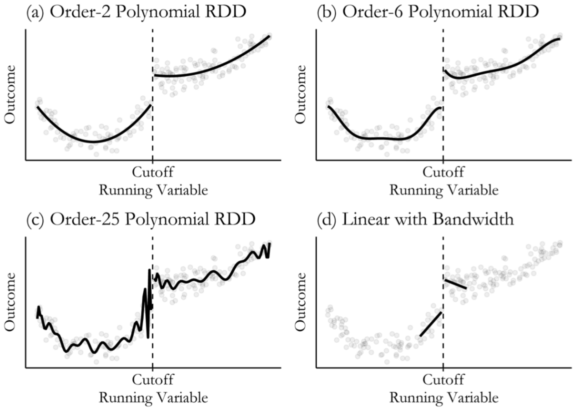
- left: $P(d=1)=0$ below z_0 and $=1$ above z_0
- right: $P(d=1)=\rho_{low}$ below z_0 and $=\rho_{high}$ above z_0, with $\rho_{high} \gg \rho_{low}$

RDD: requirements and intuition



Link

Fitting lines: Intuition



Link

Fitting lines: Intuition

- Adding all those polynomials, we get stranger and stranger results, with huge swings near the cutoff that don't even seem to track the data all that well.
- Because of this problem, it's not a great idea to go above the 2nd-order term when performing regression discontinuity with ordinary least squares. If there's a complex shape that needs fitting, instead try limiting the range of the data with a bandwidth and use a simpler shape.

Estimating a sharp RDD: equation

$$Y = \beta_0 + \beta_1(Running - Cutoff) + \beta_2 Treated + \\ \beta_3(Running - Cutoff) \times Treated + \varepsilon$$

$$Y = \beta_0 + \beta_1(Running - Cutoff) + \beta_2(Running - Cutoff)^2 + \\ \beta_3 Treated + \beta_4(Running - Cutoff) \times Treated + \\ \beta_5(Running - Cutoff)^2 \times Treated + \varepsilon \quad (20.3)$$

Treated is both an indicator for being treated and an indicator for being above the cutoff.
Link

Estimating a fuzzy RDD: intuition

- If we're only getting a jump in treatment rates of, say, 15 percentage points, then we should only expect to see 15% of the effect. We have to scale it back to the full size. Thankfully this can be done by applying instrumental variables, using basically the same regression discontinuity equations as for the sharp RDD, that it now happens in two stages.
  - These equations now become our second-stage equations.
  - Our first stage uses AboveCutoff as an instrument for Treated (and interactions with AboveCutoff as instruments for the interactions with Treated). This scales the estimate exactly as we want it to. In its basic form, instrumental variables divides the effect of the instrument on the outcome by the effect of the instrument on the endogenous/treatment variable.
- In other words, we're scaling the effect of being above the cutoff on the outcome (i.e., what we'd get from a typical sharp-regression-discontinuity model) but dividing to account for the fact that being above the cutoff only led to a partial increase in treatment rates.

Two types of RDD: estimation

Implications for estimation

1. clean RDD:
   - comparison of observations just above and just below threshold yields estimate of causal effect
   - several complications in model specification and estimation (see below)

Two types of RDD: estimation

Implications for estimation

1. clean RDD:
   - comparison of observations just above and just below threshold yields estimate of causal effect
   - several complications in model specification and estimation (see below)

2. fuzzy RDD:
   - being above ($z_i \geq z_0$) or below ($z_i < z_0$)threshold is used as instrument for actual treatment status d_i
   - estimation analogue to intention-to-treat design with Wald or 2SLS-IV estimation
   - 'double local' interpretation of estimated effect
     a. local in the LATE sense (applies only to compliers)
     b. local because only estimated at threshold

Placebo tests: Intuition

Since we're isolating variation in such a narrow window of the running variable, it's pretty plausible to claim that the only thing changing at the cutoff is treatment, and by extension anything that treatment affects (like the outcome).

Placebo tests: Intuition

Since we're isolating variation in such a narrow window of the running variable, it's pretty plausible to claim that the only thing changing at the cutoff is treatment, and by extension anything that treatment affects (like the outcome).

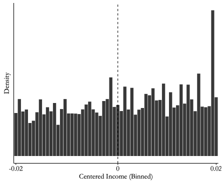
- Simply run your regression discontinuity model on something that the treatment shouldn't affect. Anything we might normally use as a control variable should serve for this purpose.
- If we do find an effect, our original regression discontinuity design might not have been quite right. Perhaps things weren't as random at the cutoff as we'd like, for some reason. There's not too much to say here.
- The equations, code, everything is the exact same for this test as for regular regression discontinuity. We just swap out the actual outcome variable for some other variable that the treatment shouldn't affect.

Summary: Inclusion of controls possible but ideally matters little (continuity assumption).

Caution: Check for manipulation

Check whether manipulation seems to be occurring at the cutoff.

- Conceptually it's very simple. We just look at the distribution of the running variable around the cutoff. It should be smooth, as we might expect if the running variable were being randomly distributed without regard for the cutoff.
- But if we see a distribution that seems to have a dip just to one side of the cutoff, with those observations sitting just on the other side, that looks a lot like manipulation.



Case Study

Thresholds in example paper: Gagliarducci & Nannicini (2013)
Do Better Paid Politicians Perform Better? Disentangling Incentives from Selection
Link to IZA Working Paper

Case Study: Gagliarducci & Nannicini (2013) - Summary

- *Background*: The wage paid to politicians affects both the choice of citizens to run for an elective office and the performance of those who are appointed.
  1. If skilled individuals shy away from politics because of higher opportunities in the private sector, an increase in politicians' pay may change their mind.
  2. If the reelection prospects of incumbents depend on their in-office deeds, a higher wage may foster performance.
  3. Goal: Authors seek to estimate effects of politicians' pay
- *Data*: On all Italian municipal governments from 1993 to 2001 and test these hypotheses in a quasi-experimental framework.
  - In Italy, the wage of the mayor depends on population size and sharply rises at different thresholds.
  - *Outcomes*: look at quality (and budget management performance) of mayoral candidates and elected mayors in small towns in Italy

Case Study: Gagliarducci & Nannicini (2013) - Summary

- Pay is increasing with town population
- *Identification of effects*: Apply a regression discontinuity design to the only threshold that uniquely identifies a wage increase – 5,000 inhabitants – to control for unobservable town characteristics.
  - *RDD identification:* mayor pay is increasing discretely at certain population thresholds
  - Implement a sharp (instead of a fuzzy) RDD, because in Italy it is the statutory wage that varies with population size.
  - Focus on the mayor as the chief executive of the municipality, and then look at budget indicators as performance outcomes.
- *Results* show that a higher wage attracts more educated candidates, and that better paid politicians size down the government machinery by improving internal efficiency.

Population thresholds for Italian municipalities

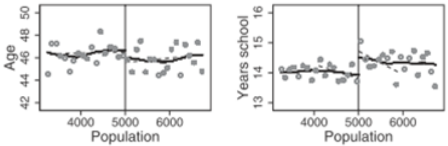
Table 1: Legislative thresholds for Italian municipalities

Population	Wage Mayor	Wage Ex. Com.	Fee Council	Ex. Com. Size	Council Size	Electoral Rule	Neighbor. Councils	Hospital/ Health
Below 1,000	1,291	15%	18	4	12	single	no	no/no
1,000-3,000	1,446	20%	18	4	12	single	no	no/no
3,000-5,000	2,169	20%	18	4	16	single	no	no/no
5,000-10,000	2,789	50%	18	4	16	single	no	no/no
10,000-15,000	3,099	55%	22	6	20	single	no	no/no
15,000-20,000	3,099	55%	22	6	20	runoff	no	no/no
20,000-30,000	3,099	55%	22	6	20	runoff	no	yes/no
30,000-50,000	3,460	55%	36	6	30	runoff	allowed	yes/no
50,000-60,000	4,132	75%	36	6	30	runoff	allowed	yes/no
60,000-100,000	4,132	75%	36	6	30	runoff	allowed	yes/yes
100,000-250,000	5,010	75%	36	10	40	runoff	yes	yes/yes
250,000-500,000	5,784	75%	36	12	46	runoff	yes	yes/yes
Above 500,000	7,798	75%	36	14-16	50-60	runoff	yes	yes/yes

At the 5,000 threshold, the gross salary of the mayor increases by 28.6% (from 2,169 to 2,789 euros), which is 33.3% before 2000 (see Table 1). Since other things change at all but one of the thresholds where mayor salary increases, only observations around the 5,000-people threshold can be used.

Effect of a 33% pay rise on politicians' age and schooling

Candidate characteristics below and above threshold



As for selection, the empirical results show that the 33% wage increase at 5,000 attracts more educated candidates: from 0.9 to 1.2 years of schooling more, depending on the specification, which means an increase in education from 6.4% to 8.6% (with respect to an average of 14 years of schooling in municipalities between 3,000 and 5,000 inhabitants).

Formalizing the RDD idea

Suppose the treatment variable of interest, d relates to z as follows:

$$d_i = \begin{cases} 1 & \text{if} \quad z_i \geq z_0 \\ 0 & \text{if} \quad z_i < z_0, \end{cases}$$

and the observed and potential outcomes can be written by a simple, linear model

$$\begin{aligned} E[Y_i^0|Z] &= \alpha + \beta z_i, && (1) \\ E[Y_i^1|Z] &= E[Y_i^0|Z] + \delta, \quad \text{which simplifies to} && (2) \\ Y_i^1 &= Y_i^0 + \delta, && (3) \end{aligned}$$

where δ is the causal effect of interest, assumed to be locally constant.

Formalizing the RDD idea

Suppose the treatment variable of interest, d relates to z as follows:

$$d_i = \begin{cases} 1 & \text{if} \quad z_i \geq z_0 \\ 0 & \text{if} \quad z_i < z_0, \end{cases}$$

and the observed and potential outcomes can be written by a simple, linear model

$$\begin{aligned} E[Y_i^0|Z] &= \alpha + \beta z_i, & (1)\\ E[Y_i^1|Z] &= E[Y_i^0|Z] + \delta, \quad \text{which simplifies to} & (2)\\ Y_i^1 &= Y_i^0 + \delta, & (3) \end{aligned}$$

where δ is the causal effect of interest, assumed to be locally constant.

Because for every i either Y^1 or Y^0 is observed, one can write the regression

$$Y_i = \alpha + \beta z_i + \delta d_i + \varepsilon_i,$$

where d_i is an 'above threshold' dummy variable.

RDD estimation in practice

In practice ...

- 'single equation' estimation of RDD (with a simple above-threshold dummy) rare

Three main problems

P1 observations from treatment group influence prediction of control group outcomes, and vice versa

P2.a observations far from z_0 change $\hat{y}$ close to z_0 which can lead to bias

P2.b complex functional forms necessary to adequately model the data, which often creates bias near threshold

P3 there may be very few observations close to the threshold

RDD estimation in practice

In practice ...

- 'single equation' estimation of RDD (with a simple above-threshold dummy) rare

Three main problems

P1 observations from treatment group influence prediction of control group outcomes, and vice versa

P2.a observations far from z_0 change $\hat{y}$ close to z_0 which can lead to bias

P2.b complex functional forms necessary to adequately model the data, which often creates bias near threshold

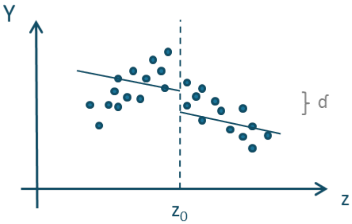
P3 there may be very few observations close to the threshold

RDD estimation in practice

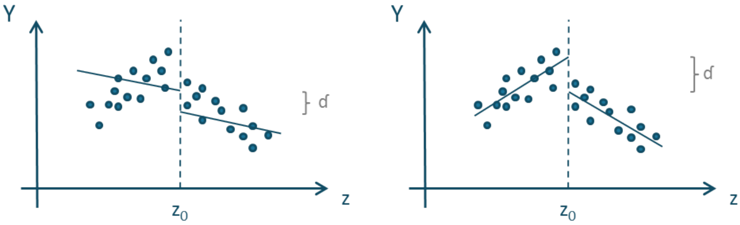
Solutions for ...

- P1 estimate separate regressions on either side of the threshold
- P2 narrow z-range of observations considered (so called 'bandwidth') to reduce bias
- P3 avoid too small bandwidths
- $\Rightarrow$ there is an unavoidable trade-off between bias and precision

Problem 1: treatment and control observations influence each other

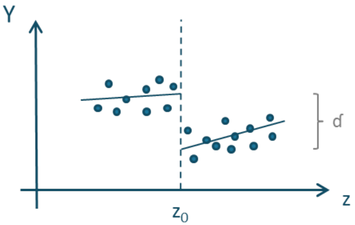


Problem 1: treatment and control observations influence each other

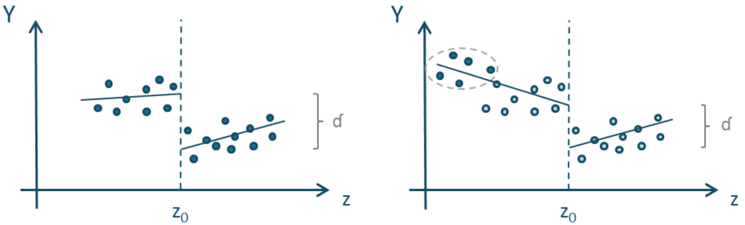


- with a single-equation linear model, slope constrained to be identical on both sides of threshold
- inappropriate modelling, leads to biased estimate of δ
- separate regressions (or one with an interaction of z with d) fit the data better

Problem 2a: bias from remote observations

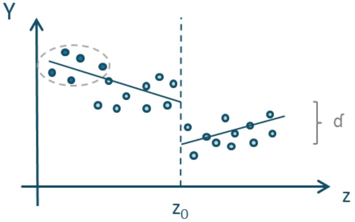


Problem 2a: bias from remote observations

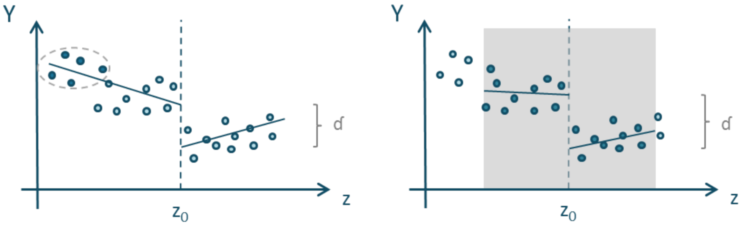


- a few additional observations far from threshold can bias estimate of effect δ at threshold

Problem 2a: bias from remote observations

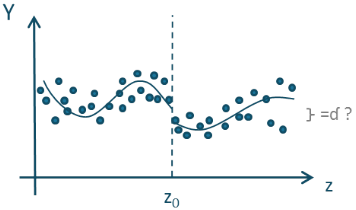


Problem 2a: bias from remote observations

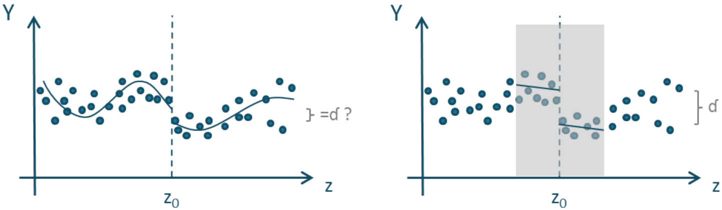


- choice of narrower bandwidth allows unbiased estimation of data
- only weaker assumption of local continuity necessary

Problem 2b: complex shapes of $y(z)$

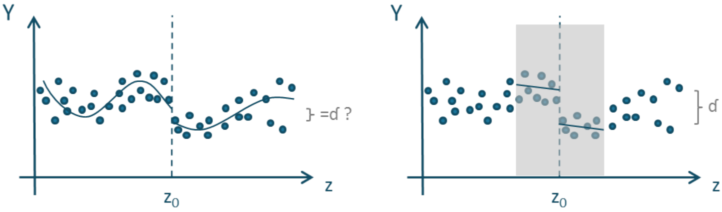


Problem 2b: complex shapes of $y(z)$



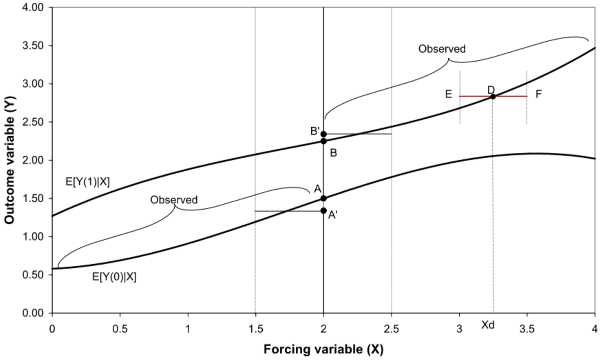
- on the left: third-order polynomial leads to downward slope of fitted model near threshold
- downward-biases estimate of δ
- narrower bandwidth choice allows unbiased estimation with local linear models

Problem 2b: complex shapes of $y(z)$

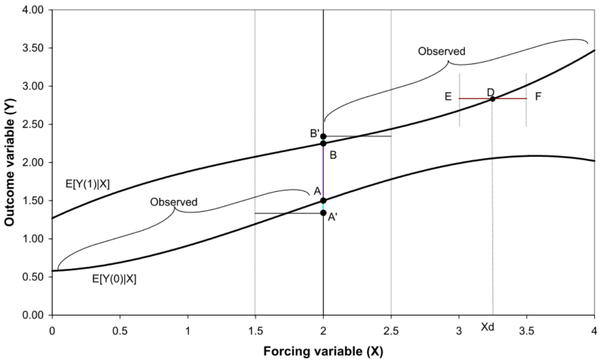


- on the left: third-order polynomial leads to downward slope of fitted model near threshold
- downward-biases estimate of δ
- narrower bandwidth choice allows unbiased estimation with local linear models
- but leads to problem 3: too few observations in bandwidth $\rightarrow$ less precision
- widening the bandwidth again leads to bias if $y(z)$ has a trend at threshold

Bandwidth-induced bias of local regressions



Bandwidth-induced bias of local regressions

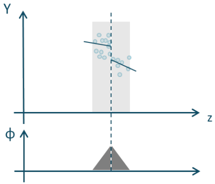


- with a (positive or negative) trend in $y(z)$, a non-zero bandwidt always leads to some bias
- a way to mitigate this is kernel-weighted local regression with e.g. a triangular

RDD with local kernel-weighted regressions

State-of-the-art approach

- estimate local regression on either side of threshold
- weight observations depending on distance from threshold (kernel-weighting)
- need to choose bandwidth and kernel shape (rectangular, triangular (in graph), Epanechnikov)
- local regression model can be linear or higher-order polynomial, but should not matter much if bandwidth narrow enough



RDD estimates based on local linear regressions

TABLE 3. Candidates and mayor selection, RDD estimates.

Population	Female	Age	Years school	Not employed	Entrepreneurs	White collar	Blue collar
				All candidates			
Effect	0.005	−0.903	0.905***	−0.025	−0.037	0.082**	−0.018
	(0.018)	(0.587)	(0.279)	(0.025)	(0.028)	(0.039)	(0.025)
Δ	1,300	1,700	900	900	1,700	1,300	1,400
Obs.	4,805	6,405	3,295	3,295	6,405	4,805	5,191
				Mayors			
Effect	−0.014	−0.847	0.879**	−0.007	−0.023	0.074	−0.035
	(0.022)	(0.822)	(0.346)	(0.033)	(0.046)	(0.046)	(0.035)
Δ	1,700	1,700	1,100	1,000	1,400	1,700	1,400
Obs.	2,971	2,971	1,905	1,738	2,396	2,971	2,396

Notes: Effect of the 33% wage increase at the 5,000 threshold on the characteristics of the three best candidates (top panel) and of the elected mayor (bottom panel). Terms from 1993 to 2001. Cities with population between 3,250 and 6,750 inhabitants. Local linear regression (LLR) with optimal symmetric bandwidth Δ. *Age* and *Years school* are measured in years; the other variables are dummies. *Not employed* includes unemployed, retired, and any other individual out of the labor force. *Entrepreneur* includes self-employed and entrepreneurs. *White collar* includes lawyers, professors, physicians, and managers. *Blue collar* includes blue collars, clerks, and technicians. Standard errors robust to clustering at the municipality level are in parentheses.
*Significant at 10%; **significant at 5%; ***significant at 1%.

- candidates and mayors in above-threshold towns have ca 0.9 additional years of schooling compared to those in below-threshold towns
- on average the population around threshold has 14 years schooling
- thus, increase of 0.9 reflects +6.4% for candidates and +6.2% for elected mayors
- few other significant results

The key RDD requirements/assumptions revisited

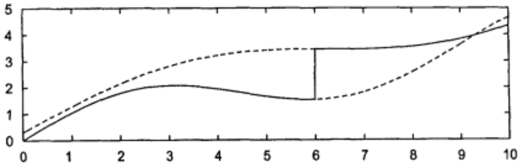
RDD requirements/assumptions:

1. *existence of discontinuity:* discrete 'jump' of cause of interest (d) at one or more values (thresholds) of z
2. *continuity assumption:* 'assignment' to above and below threshold(s) as-if random with respect to everything but d
   - continuity in observed and potential outcomes around the threshold if there were no threshold (see graph)

The key RDD requirements/assumptions revisited

RDD requirements/assumptions:

1. *existence of discontinuity:* discrete 'jump' of cause of interest (d) at one or more values (thresholds) of z
2. *continuity assumption:* 'assignment' to above and below threshold(s) as-if random with respect to everything but d
   - continuity in observed and potential outcomes around the threshold if there were no threshold (see graph)



Testing RDD assumptions

Can these requirements/assumptions be tested?

1. in-depth knowledge of laws and institutions necessary
   - check whether there is an actual jump in d; is it sharp or fuzzy?
   - check whether d and y do not display jumps at other points of z where there is no threshold
     - estimate 'effects' using preferred RD specification at many placebo thresholds
     - rule of thumb: should only be able to reject the null of 'no jump' in 5% of cases

Testing RDD assumptions

Can these requirements/assumptions be tested?

1. in-depth knowledge of laws and institutions necessary
   - check whether there is an actual jump in d; is it sharp or fuzzy?
   - check whether d and y do not display jumps at other points of z where there is no threshold
     - estimate 'effects' using preferred RD specification at many placebo thresholds
     - rule of thumb: should only be able to reject the null of 'no jump' in 5% of cases
2. continuity harder to test
   - assumption cannot be verified (because potential outcomes by definition unobserved)
   - but can be falsified because assumption is usually likely violated if
     (a) y or d are very 'jumpy' along z,
     (b) there is sorting around z_0 (e.g. due to manipulated population numbers),
     (c) or if there are other changes at the same threshold

   $\rightarrow$ check whether one of them applies

Trying to (fail to) falsify the continuity assumptions

Checks to falsify continuity assumption

(a) check if there are discrete jumps in y or d at values of z where there are no thresholds (placebo tests)

(b) check for smooth distribution of other variables x around z_0 (works for observables)

(b) check for smooth density around threshold, e.g. by visual inspection and McCrary (2007) test (also speaks to unobservables)

(c) in-depth knowledge of legal and institutional environment necessary

Tests of smooth density

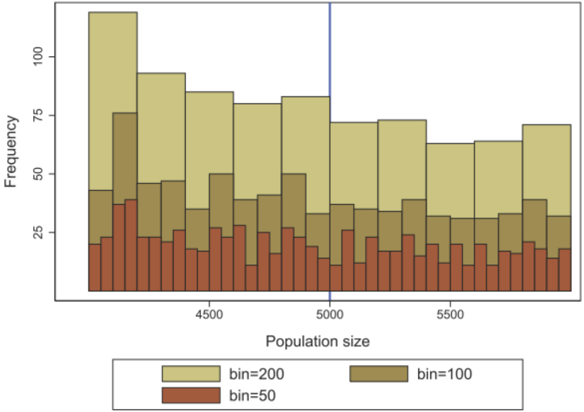


FIGURE 6. Population density around 5,000. Frequency of cities around the 5,000 threshold (vertical line), according to population size in the 2001 Census.

Assessing RDD

Advantages

- often only way for causal identification of certain effects
- particularly high internal validity if continuity holds
- assumptions fairly well and transparently testable (e.g. compare to IVs)
- graphical display of results often intuitive

Assessing RDD

Advantages

- often only way for causal identification of certain effects
- particularly high internal validity if continuity holds
- assumptions fairly well and transparently testable (e.g. compare to IVs)
- graphical display of results often intuitive

Disadvantages

- similar to any natural experiment
  - opportunities for RDD cannot be created, just discovered
  - method influences what is been studied
- local identification only $\rightarrow$ external validity
  (how far away from threshold is δ still constant?)

Non-parametric estimation

Basic idea of non-parametric estimation (‘running mean smoothing’ in G&N paper)

- data divided into many slices or ‘bins’
- many local means estimated
- kernel-weighting within slices possible