Causal Inference: Regression Discontinuity Design

POLI 803 Research Methods in PS

Howard Liu

Week 10, 2024

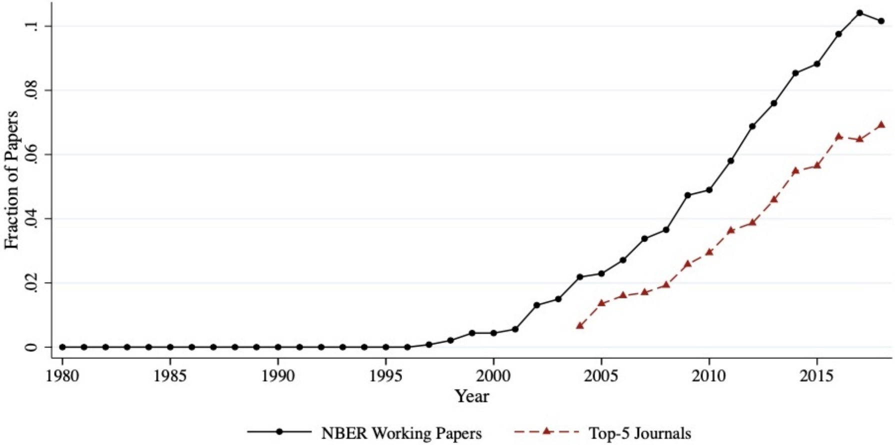
Roadmap

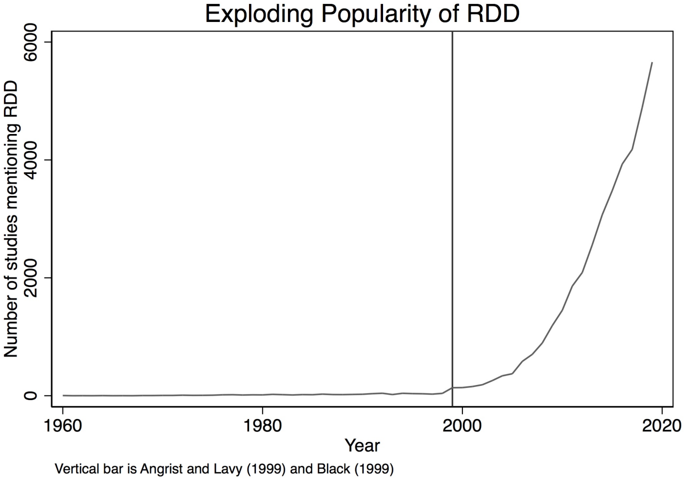
Mom’s knee issue

What is regression discontinuity?

- RDD is a popular particular type of research design.
- Often thought to be the most "credible" of the observational designs, even though it does not depend on randomization for identification
- A viz-heavy design, so let's see some figures. (tons of pictures in this lecture)

B: Regression Discontinuity





”Jumps are so unnatural that when we see them happen, they beg for explanation” (p.245)

Tell me what you think is happening

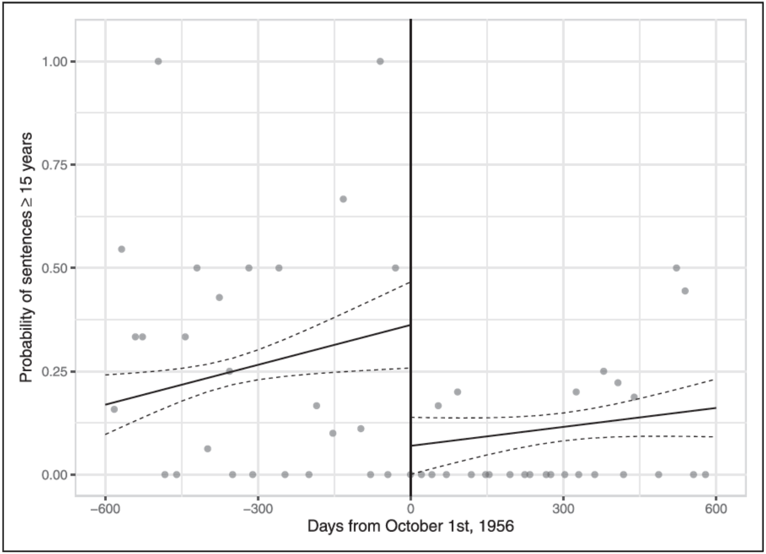


Figure 1. The 1956 reform and the severity of sentences.

RDD features

- We want to estimate some causal effect of a treatment on some outcome

RDD features

- We want to estimate some causal effect of a treatment on some outcome
- No comparison groups: But we can't compare two groups (treated and not treated) because of the self selection which creates selection bias

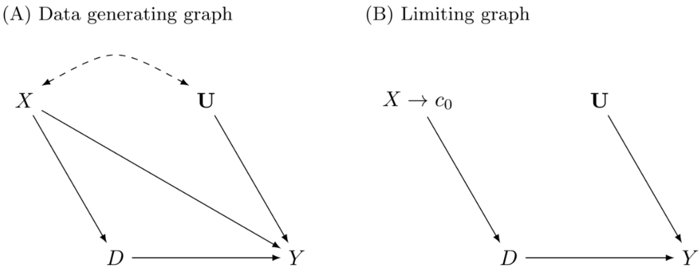
RDD features

- We want to estimate some causal effect of a treatment on some outcome
- No comparison groups: But we can't compare two groups (treated and not treated) because of the self selection which creates selection bias
- But what if treatment assignment was forced on units because the firm or agency uses a multi valued variable and splits the sample when units are above or below some threshold?
- RDD formalizes this setup and under some assumptions will identify causal effects

RDD Words and Pictures

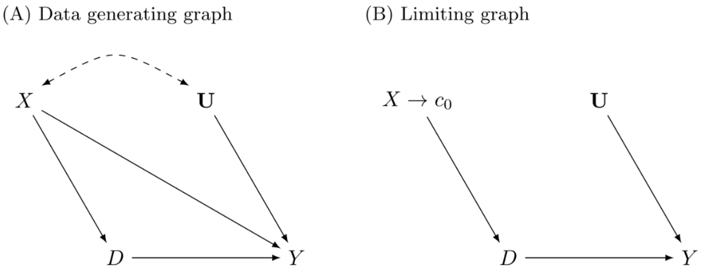
- Just keep in mind as do this – RDD is a method of mimicking the experimental design, as opposed to merely a regression model
- There's a lot of **new terminology** if you're new to RDD
- A picture is worth a thousand words: tons of pictures, but tons of new concepts too

DAG for RDD



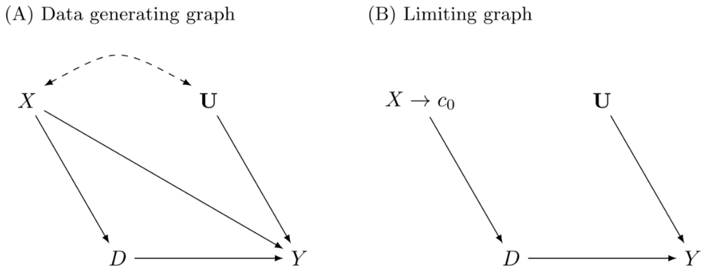
- In case you forget, DAG is Directed Acyclic Graphs
- X: running/assignment variable. A **continuous variable** assigning units to treatment D ($X \rightarrow D$), based on a "cutoff" score c_0

DAG for RDD



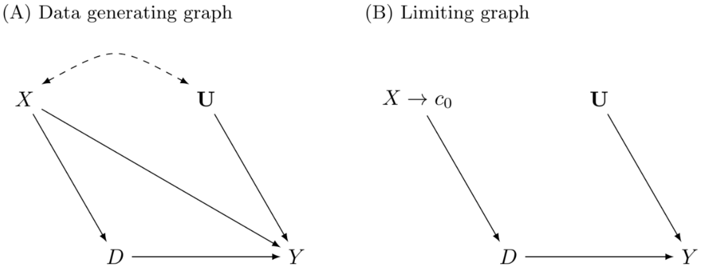
- In case you forget, DAG is Directed Acyclic Graphs
- X: running/assignment variable. A **continuous variable** assigning units to treatment D ($X \rightarrow D$), based on a "cutoff" score c_0
- Lack of common support: never have units that are both in the treatment and control groups for the same value of X $\rightarrow$ no overlaps..bad?

DAG for RDD



- In case you forget, DAG is Directed Acyclic Graphs
- X: running/assignment variable. A **continuous variable** assigning units to treatment D ($X \to D$), based on a "cutoff" score c_0
- Lack of common support: never have units that are both in the treatment and control groups for the same value of X $\to$ no overlaps..bad? we extrapolate in RDD

DAG for RDD



- Continuity assumption: right at c_0, the assigement variable X no longer has a direct efffect on Y
- In words, things (the expected potential outcomes) would have continue if there was no assigement

An RDD solution to non-overlaps (Lack of Common Support)

- Comparing units in a close neighborhood around some cutoff c_0

An RDD solution to non-overlaps (Lack of Common Support)

- Comparing units in a close neighborhood around some cutoff c_0
- ATE for a subpopulation can be identified as (X $\rightarrow c_0$)

An RDD solution to non-overlaps (Lack of Common Support)

- Comparing units in a close neighborhood around some cutoff c_0
- ATE for a subpopulation can be identified as (X $\rightarrow c_0$)
- Becasue we focus on a "subpopulation," we identify LATE (local average treatment effect), not ATE

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments
- **Cutoff,** c_0 or **threshold**: a particular value at a point on the running variable above which actors assign treatments to unit

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments
- **Cutoff,** c_0 or **threshold**: a particular value at a point on the running variable above which actors assign treatments to unit
- **Discontinuity** and/or **Jump**: Since we are *estimating* breaks in the outcome right at the cutoff, and when that happens we say that there is a "discontinuity"

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments
- **Cutoff,** c_0 or **threshold**: a particular value at a point on the running variable above which actors assign treatments to unit
- **Discontinuity** and/or **Jump**: Since we are *estimating* breaks in the outcome right at the cutoff, and when that happens we say that there is a "discontinuity" $\rightarrow$ also means everything else should be in **continuity**

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments
- **Cutoff,** c_0 or **threshold**: a particular value at a point on the running variable above which actors assign treatments to unit
- **Discontinuity** and/or **Jump**: Since we are *estimating* breaks in the outcome right at the cutoff, and when that happens we say that there is a "discontinuity" $\rightarrow$ also means everything else should be in **continuity**
- **Regression**: Many of the models are simple difference in means, "local" regressions from OLS or global regressions from OLS

RDD key termologies

- **Running variable,** X: a usually continuous score that some actors use to assign treatments
- **Cutoff,** c_0 or **threshold**: a particular value at a point on the running variable above which actors assign treatments to unit
- **Discontinuity** and/or **Jump**: Since we are *estimating* breaks in the outcome right at the cutoff, and when that happens we say that there is a "discontinuity" $\rightarrow$ also means everything else should be in **continuity**
- **Regression**: Many of the models are simple difference in means, "local" regressions from OLS or global regressions from OLS

Common types of RDD: Close elections

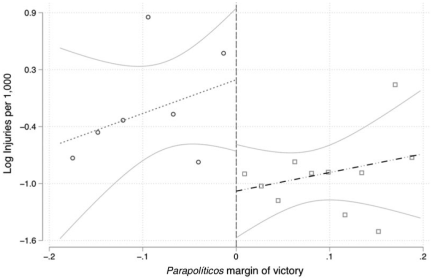
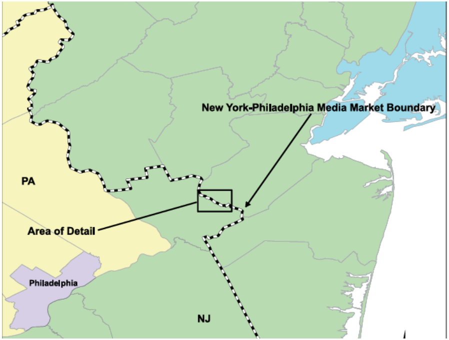


Fig. 1. RD effect of paramilitary-mayor winning on bodily harm.
Note: The horizontal axis displays the winning margin of the paramilitary-mayors (winners and runners-up). The dashed lines are the linear fit. The solid lines are the 95 per cent confidence intervals at both sides of the threshold.

- A cross-sectional discontinuity
- RQ: effects of winning elections on something (e.g., violence)
- Compare parties that win or lose at the margin, assuming that parties are quite similar in everything else (except winning)

Common types of RDD: Geographic RDD



- A cross-sectional discontinuity
- RQ: effects of arbitrary/unnatural borders on something (e.g., violence)
- Compare behavior of populations right at the border, assuming that population are quite similar at both sides of the border

Common types of RDD: RDD in Time

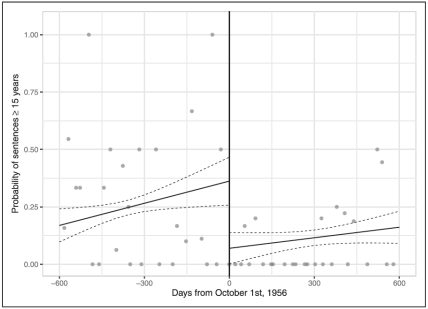


Figure 1. The 1956 reform and the severity of sentences.

- A time-series discontinuity
- RQ: effects of an arbitrary intervention in time on something (e.g., violence)
- Compare behavior of populations right at the time cutoff, assuming that population are quite similar at both sides of the time cutoff

Data requirements

Large sample sizes are characteristic features of the RDD

- If there are strong trends in the running variable, one typically needs a lot more data than if there weren't

Data requirements

Large sample sizes are characteristic features of the RDD

- If there are strong trends in the running variable, one typically needs a lot more data than if there weren't
- If the observations tend to be noisy, we need more data than if it was less noisy

Data requirements

Large sample sizes are characteristic features of the RDD

- If there are strong trends in the running variable, one typically needs a lot more data than if there weren't
- If the observations tend to be noisy, we need more data than if it was less noisy
- We need a lot of data bc we need significant mass at the running variable to reject the null

Data requirements

Large sample sizes are characteristic features of the RDD

- If there are strong trends in the running variable, one typically needs a lot more data than if there weren't
- If the observations tend to be noisy, we need more data than if it was less noisy
- We need a lot of data bc we need significant mass at the running variable to reject the null
- Rewards people with access to firm level data since it can be large

Sharp vs. Fuzzy RDD

- There's two classes of RD designs:
    1. Sharp RDD: Treatment is a **deterministic function (Yes or No)** of running variable, X. Example: Medicare benefits.

Sharp vs. Fuzzy RDD

- There's two classes of RD designs:
  1. Sharp RDD: Treatment is a **deterministic function (Yes or No)** of running variable, X. Example: Medicare benefits.
  2. Fuzzy RDD: Discontinuous "jump" in the **probability [0,1]** of treatment when $X > c_0$. Cutoff is used as an instrumental variable for treatment. Example: attending state flagship

Sharp vs. Fuzzy RDD

- There's two classes of RD designs:
  1. Sharp RDD: Treatment is a **deterministic function (Yes or No)** of running variable, X. Example: Medicare benefits.
  2. Fuzzy RDD: Discontinuous "jump" in the **probability [0,1]** of treatment when $X > c_0$. Cutoff is used as an instrumental variable for treatment. Example: attending state flagship
- Fuzzy is a type of IV strategy and requires explicit IV estimators like 2SLS; sharp is reduced form IV and doesn't require IV-like estimators – we study it later with IV therefore

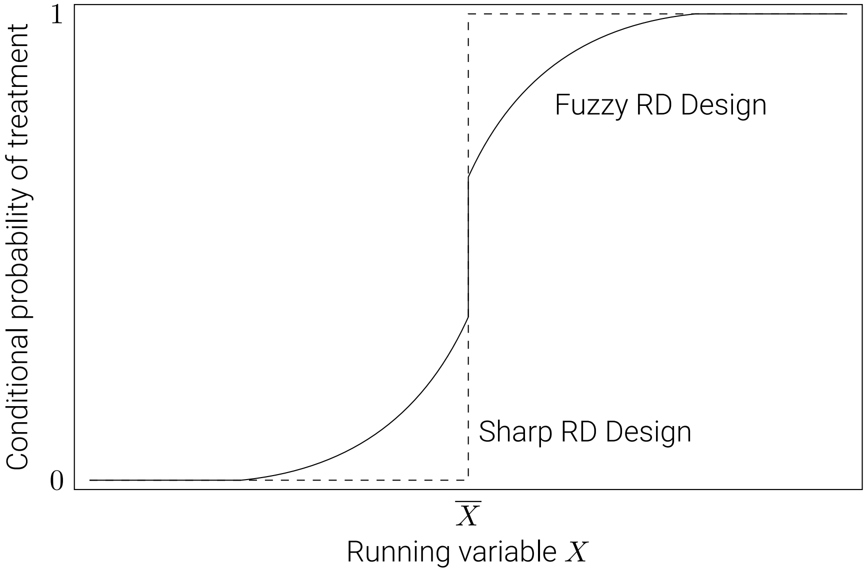


Figure: Sharp (dashed) vs. Fuzzy (solid) RDD

Sharp RDD example

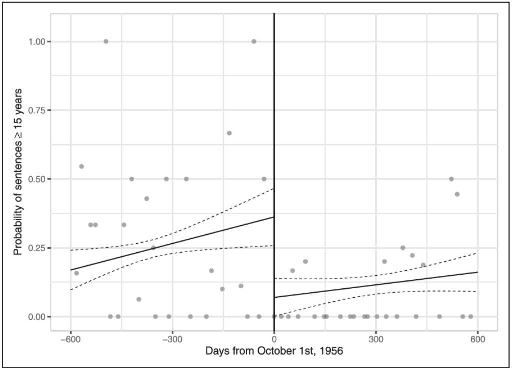
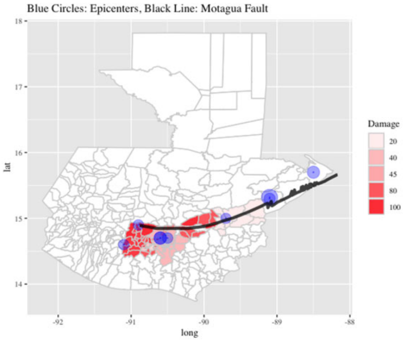


Figure 1. The 1956 reform and the severity of sentences.

- X (assigment variable): trial dates
- c (cutoff): deterministic time cutoff (1956.10.1)
- D (treatment status): 1, treated. 0, otherwise
- Y (potential outcome): individual sentencing levels for defendants

Fuzzy RDD example



$$y(t) = y_f + (y_0 - y_f)e^{-\alpha t}$$

Figure 1. Major earthquake shocks and damage by municipalities, 1976.

- X (assigment variable): Motagua fault zones
- c (cutoff): probabalistic geography cutoff, distance to the fault line
- D (treatment status): infrastructure damage
- Y (potential outcome): levels of repression in each municipalities

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we **extrapolate** using models like regression and nonparametric methods by comparing units **just below and above the cutoff to one another** but this is sensitive to data trends, bandwidths, and number of observations

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we **extrapolate** using models like regression and nonparametric methods by comparing units **just below and above the cutoff to one another** but this is sensitive to data trends, bandwidths, and number of observations $\rightarrow$ means a lot of robustness tests needed

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we **extrapolate** using models like regression and nonparametric methods by comparing units **just below and above the cutoff to one another** but this is sensitive to data trends, bandwidths, and number of observations $\rightarrow$ means a lot of robustness tests needed
- **Treatment effects**: We are estimating average treatment effects but only for people **at the cutoff** and that may not be informative of any other point on the running variable with extreme heterogeneity

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we **extrapolate** using models like regression and nonparametric methods by comparing units **just below and above the cutoff to one another** but this is sensitive to data trends, bandwidths, and number of observations $\rightarrow$ means a lot of robustness tests needed
- **Treatment effects**: We are estimating average treatment effects but only for people **at the cutoff** and that may not be informative of any other point on the running variable with extreme heterogeneity $\rightarrow$ means a lot of robustness tests needed

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we **extrapolate** using models like regression and nonparametric methods by comparing units **just below and above the cutoff to one another** but this is sensitive to data trends, bandwidths, and number of observations $\rightarrow$ means a lot of robustness tests needed
- **Treatment effects**: We are estimating average treatment effects but only for people **at the cutoff** and that may not be informative of any other point on the running variable with extreme heterogeneity $\rightarrow$ means a lot of robustness tests needed

Treatment assignment in the sharp RDD

Deterministic treatment assignment ("sharp RDD")

In Sharp RDD, treatment status is a deterministic and discontinuous function of a covariate, X_i:

$$D_i = \begin{cases} 1 \text{ if} & X_i \geq c_0 \\ 0 \text{ if} & X_i < c_0 \end{cases}$$

where c_0 is a known threshold or cutoff. In other words, if you know the value of X_i for a unit i, you know treatment assignment for unit i with certainty.

Extrapolation, common support and functional form

- Sharp designs create common support problems because there will literally *never* be a unit in treatment and control across the running variable

Extrapolation, common support and functional form

- Sharp designs create common support problems because there will literally *never* be a unit in treatment and control across the running variable
- This requires "extrapolation"; prediction beyond the support of the data (i.e., where treatment switches at cutoff)

Extrapolation, common support and functional form

- Sharp designs create common support problems because there will literally *never* be a unit in treatment and control across the running variable
- This requires "extrapolation"; prediction beyond the support of the data (i.e., where treatment switches at cutoff)
- But since you're predicting, modeling choices like **functional form** are key and that's a structural assumption

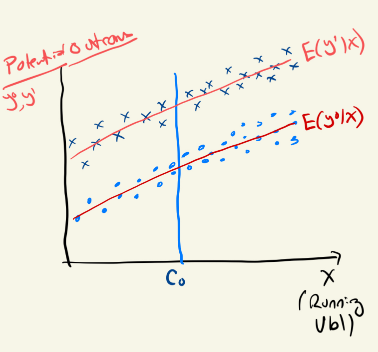
Smoothness/continuity as the identifying assumption

Smoothness of conditional expected potential outcome functions through the cutoff

$E[Y_i^0|X=c_0]$ and $E[Y_i^1|X=c_0]$ are continuous (smooth) in X at c_0.

- If population average *potential outcomes*, $E[Y^1]$ and $E[Y^0]$, are smooth functions of X across the cutoff, c_0, then expected potential average outcomes *won't* jump at c_0.
- Implies that the **confounders** should evolve smoothly across the cutoff

Smoothness vs Treatment Effect



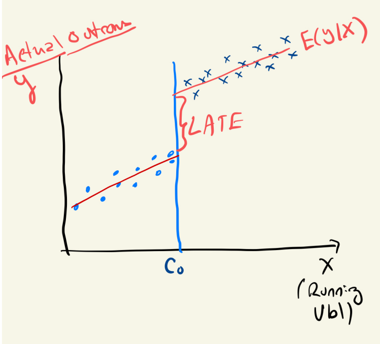


Figure: Smoothness of potential outcomes (left) vs estimation of LATE (right)

Discussion: Why is the left picture different from the right picture? Where did the two lines go?

Potential and observable outcomes

- Smoothness is about potential outcomes:
  - → Potential outcomes are on average smoothly changing across the threshold
- Discontinuity is about realized outcomes:
  - → The cutoff *is the assignment mechanism*
  - → The cutoff switches between potential outcomes
  - → Therefore if there is a treatment effect, we can observe the *realized* outcomes jump/drop at the cutoff
  - → If there is a treatment effect, it would be visible but it requires some extrapolation to see

Smoothness permits extrapolation

- Smoothness justifies the use of regression models to extrapolate missing potential outcomes from one side of the cutoff to the other (has a matching feel)

Smoothness permits extrapolation

- Smoothness justifies the use of regression models to extrapolate missing potential outcomes from one side of the cutoff to the other (has a matching feel)
- Average causal effect is defined **at the cutoff**, but estimation uses data left and right **around the cutoff**

Smoothness permits extrapolation

- Smoothness justifies the use of regression models to extrapolate missing potential outcomes from one side of the cutoff to the other (has a matching feel)
- Average causal effect is defined **at the cutoff**, but estimation uses data left and right **around the cutoff**
- Once we have the identification strategy justified (smoothness), we now can run regression (estimation of the causal effect)

Roadmap

Approximate the functional form

Two ways to estimate the treatment effect at $X = c_0$

1. Curve: Use global and local regressions with $f(X_i)$ equalling a p^{th} **order polynomial** (results highly sensitive to functional form)

Approximate the functional form

Two ways to estimate the treatment effect at $X = c_0$

1. Curve: Use global and local regressions with $f(X_i)$ equalling a p^{th} **order polynomial** (results highly sensitive to functional form)
   - $\rightarrow$ Is our finding of discontinuity due to us mis-specifying the curves?

Approximate the functional form

Two ways to estimate the treatment effect at $X = c_0$

1. Curve: Use global and local regressions with $f(X_i)$ equalling a p^{th} **order polynomial** (results highly sensitive to functional form)
   - $\rightarrow$ Is our finding of discontinuity due to us mis-specifying the curves?
   - $\rightarrow$ Allowing different slopes of regression lines at the both sides
   - $\rightarrow$ Allowing lines to become curves (higher order polynomials)
2. Weights: Nonparametric kernel methods and local linear regressions (less sensitive)

Estimation with extrapolation

- We use *extrapolation* to estimate average treatment effects with the sharp RDD which is unbiased under *smoothness*
- Our statistical models predict expected conditional *counterfactuals* using data on *the other side of the cutoff*
- Keep in mind though: the actual aggregate causal effect is $Y_i^1 - Y_i^0$ at any point on X_i – not across $X = c_0$

Re-centering the running variable

- Assume a linear function

$$Y_i = \alpha + \beta(X_i) + \delta D_i + \varepsilon_i$$

- People will often "re-center" by subtracting c_0 from X_i:

$$Y_i = \alpha + \beta(X_i - c_0) + \delta D_i + \varepsilon_i$$

- This doesn't change the interpretation of the treatment effect; **just the intercept**.

Linearity Problem 1: Smooth but **nonlinear** expected potential outcomes

- What if the trend relation $E[Y_i^0|X_i]$ does not jump at c_0 but rather is simply nonlinear? You could get spurious results

Linearity Problem 1: Smooth but **nonlinear** expected potential outcomes

- What if the trend relation $E[Y_i^0|X_i]$ does not jump at c_0 but rather is simply nonlinear? You could get spurious results
- You'll likely use higher order polynomial transformations of the running variable

Potential outcomes and nonlinear running variable

- But what if the potential outcomes aren't just nonlinear – the nonlinearities are different for $E[Y^1]$ than they are for $E[Y^0]$
- We can generalize the potential outcome expressions by allowing them to depend on the running variables, but in different ways depending on whether it is or is not treated

Potential outcomes and nonlinear running variable

- This will require saturated models in which you include them both individually and interacting them with D_i.

$$\begin{aligned} E[Y_i^0|X_i] &= \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \cdots + \beta_{0p}\tilde{X}_i^p \\ E[Y_i^1|X_i] &= \alpha + \delta + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \cdots + \beta_{1p}\tilde{X}_i^p \end{aligned}$$

  where $\tilde{X}_i$ is the centered running variable (i.e., $X_i - c_0$).
- Notice the treatment effect in the second line, and the intrinsic ATE when comparing the two equations, $E[Y_i^0 - Y_i^1|X_i]$

Linearity Problem 2: Interact running variable (X) with treatment (D)

- If you believe the effect of the running variable on the outcome differs above and below the threshold, adding an interaction term $D \times X$ can allow the slope of the relationship to change at the threshold. This helps model potential non-linearities.

Linearity Problem 2: Interact running variable (X) with treatment (D)

- If you believe the effect of the running variable on the outcome differs above and below the threshold, adding an interaction term $D \times X$ can allow the slope of the relationship to change at the threshold. This helps model potential non-linearities.

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 (X_i - c_i) + \beta_3 (X_i - c_i) * D_i + \epsilon_i \tag{1}$$

$$\beta_1 = E[Y_i | D_i = 1, X_i = c] - E[Y_i | D_i = 0, X_i = c] \tag{2}$$

- β_1 is the treatment effect at the cutoff.

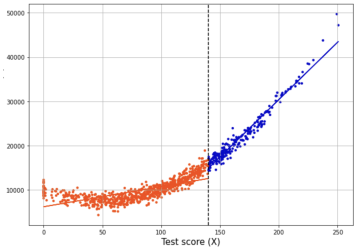
Regression equation: higher order polynomial and interaction

- Let $\tilde{x} = (X_i - c_i)$
- Regression model you estimate is:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \delta D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \varepsilon_i \end{aligned}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{21} - \beta_{21}$ and $\beta_p^* = \beta_{1p} - \beta_{0p}$

Estimation without and with specifying nonlinear running variable



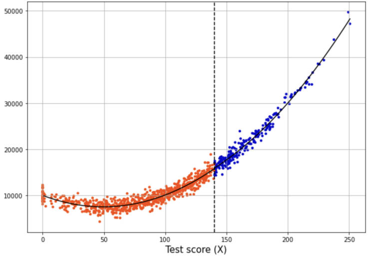


Figure: Spurious treatment effects with linear specification (left) versus 3rd order polynomial (right)

Look close: See how the lines don't touch on the left, but they do on the right?

The trade-off: Comment about higher order polynomials

- If you don't have a lot of data, you will likely have to have very large bandwidths just to get the sample size up
- With a lot of data far from the cutoff, you'll likely overfit with a higher order polynomial series
- But higher order polynomials can have overfitting problems leading to poor prediction beyond the cutoff
- Gelman and Imbens (2018) caution against overfitting on these global regressions (i.e., **quadratics**)

Some new terms: all about the windows

- **Kernels** make a window and give you the shape of the window (e.g., triangular kernels weight the observations differently within the window)
- **Bandwidth** is the “length” of the window (small ones are tiny windows, bigger ones, bigger windows – think of a histogram)
- **Bins** are about the interval itself (a partition)

Local linear nonparametric regressions

- Least squares approaches models the counterfactual using functional forms which is parametric, but it can have poor predictive properties on counterfactuals above/below the cutoff
- Another way of approximating the running variable flexibly $f(X_i)$ is to use a **nonparametric kernel**

Local linear nonparametric regressions

- Local linear nonparametric regression substantially reduces the bias
- Think of it as a weighted regression restricted to a window – kernel provides the weights to that regression.

Choices you have to make

1. Choose the bandwidth h: the window
2. Choose the kernel $K(\cdot)$: uniform vs. traingular
3. Choose the polynomial ordering p: linear vs. quadratic fit

We have a broad set of writings and suggestions around each of these things, and the issues around choices is always subjective researcher bias, uncertainty and various forms of bias $\rightarrow$ You do all!

Animation of a local linear regression

https://twitter.com/page_eco/status/958687180104245248

Types of kernels

- **Rectangular uniform** weights equivalent to $E[Y]$ at a given bin on X
- **Triangular** draws a straight line from the threshold to the edge of the bandwidth and weights along the line
- **Epanechnikov** is similar but is more like a parabola

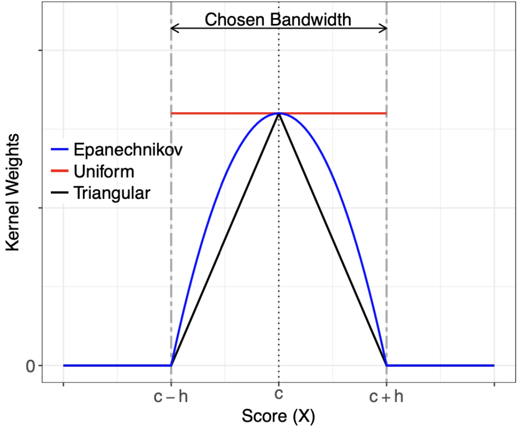


Figure: From Cattaneo, et al. (2019)

Estimation with kernels

- Cattaneo, et al. (2019) recommend using the triangular kernel because when you use it with a bandwidth that optimizes mean squared error, you can get a point estimate that is optimal.

Estimation with kernels

- Cattaneo, et al. (2019) recommend using the triangular kernel because when you use it with a bandwidth that optimizes mean squared error, you can get a point estimate that is optimal.
- Triangular kernels assign zero weight to all observations outside bandwidth h interval and positive weights within it
- Weights are maximized at the cutoff and decline symmetrically and linearly as the value of the running variable gets further away

Polynomial order

- Simple difference in means (i.e., p order of zero) is like a histogram with uniform weights
- Suffers from what is called the "boundary problem" – the estimation of the true expected potential outcomes at the cutoff is biased with trends in the running variable
- But even after choosing kernel weights, we aren't done as then there is the business of choosing polynomial order

Polynomial terms

- Two conceptual issues to keep in mind
    1. No polynomials has boundary problems, but
    2. **Higher order polynomials, though, suffer from severe overfitting problems**
- Local linear RD is the preferred method, but this is where we end up in the world of choosing the bandwidths, h, because that controls the width (and thus selects the units) of the neighborhood around the cutoff that will be used to fit the model

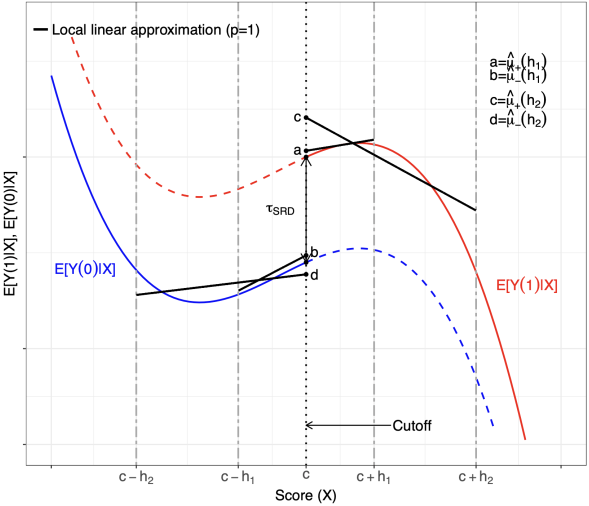


Figure: From Cattaneo, et al. (2019)

Bias and variance?

Bias term

- When we approximate the unknown functions with p, h and $K(\cdot)$, there's some approximation error because we do now actually know the true function
- Think about the earlier picture – when we used the larger bandwidth and p of zero, we came up short. Why? Because of the curvature of the functions we were approximating

Bias and variance?

Bias term

- When we approximate the unknown functions with p, h and $K(\cdot)$, there's some approximation error because we do now actually know the true function
- Think about the earlier picture – when we used the larger bandwidth and p of zero, we came up short. Why? Because of the curvature of the functions we were approximating
- $\rightarrow$ The smaller the window, the more precision we have in estimation

Bias and variance?

Variance term

- Variance depends on sample size and bandwidth h
- As number of observations near the cutoff falls, the contribution of the variance term to MSE grows and vice versa
- Variability of the the point estimator depends therefore on density at the cutoff (which gets back to why RD tends to be data intensive in the first place)

Bias and variance?

Variance term

- Variance depends on sample size and bandwidth h
- As number of observations near the cutoff falls, the contribution of the variance term to MSE grows and vice versa
- Variability of the the point estimator depends therefore on density at the cutoff (which gets back to why RD tends to be data intensive in the first place)
- $\rightarrow$ The more observation, the more variance we have in estimation

So what do we do? Optimal Bandwidths

- Most approaches have some balancing act between bias and variance that they're trying to address
- **Minimizing the MSE** of the local estimator, $\widehat{\delta}$, given a choice of p and $K(\cdot)$ has become the most popular since MSE is the sum of squared bias and variance

$$MSE(\widehat{\delta}) = Bias^2(\widehat{\delta}) + Variance(\widehat{\delta})$$

- If you choose to minimize MSE, you are choosing h – hence "optimal bandwidths"

$$min_{h>0}\left(h^{2(p+1)}B^2 + \frac{1}{nh}V\right)$$

Optimal Bandwidths

- Solution to that minimization problem is h_{MSE} and is the MSE-optimal bandwidth choice

$$h_{MSE} = \left(\frac{V}{2(p+1)} B^2 \right)^{\frac{1}{(2p+3)}} n^{-1/(2p+3)}$$

which directly addresses the bias-variance trade-off

Optimal Bandwidths

- Solution to that minimization problem is h_{MSE} and is the MSE-optimal bandwidth choice

$$h_{MSE} = \left(\frac{V}{2(p+1)} B^2 \right)^{\frac{1}{(2p+3)}} n^{-1/(2p+3)}$$

  which directly addresses the bias-variance trade-off
- Optimal bandwidths that minimize MSE are proportional to that last term and therefore MSE-optimal bandwidths increase with V (more observations) and decrease with B (less observations)

Optimal Bandwidths

- Solution to that minimization problem is h_{MSE} and is the MSE-optimal bandwidth choice

$$h_{MSE} = \left(\frac{V}{2(p+1)}B^2\right)^{\frac{1}{(2p+3)}} n^{-1/(2p+3)}$$

  which directly addresses the bias-variance trade-off
- Optimal bandwidths that minimize MSE are proportional to that last term and therefore MSE-optimal bandwidths increase with V (more observations) and decrease with B (less observations)
- Hence why optimal bandwidths are "data driven" and automated which takes away some of the subjective decisions researchers must make

Implementation with software

- You have choices for implementing this – manually (see Cattaneo, et al. (2019) section 4.2.4, or with packages like `rdrobust`
- Very flexible – choose kernels (e.g., triangular), choose polynomials, choose number of bandwidths h
- But remember choosing h is not advisable bc of what we just said, so there is a separate package called `rdbwselect` which selects the MSE-optimal bandwidth for the local estimator (but you still choose p and $K(\cdot)$)

Implementation with software

- Tons of options with `rdbwselect` – different kernels, even different bandwidths left and right of the cutoff
- Once you use it, you can pass it on to `rdrobust` in a second stage, or
- Just use `bwselect` within the syntax of `rdrobust` itself (we will review this with our Hansen exercise later)
- All of this can be incorporated into plotting too with `rdplot`

Main Challenges to RDD and **Robustness Tests**

Classify your concern regarding smoothness violations into two categories:

- **Manipulation** on the running variable $\rightarrow$ McCrary Density Test

Main Challenges to RDD and **Robustness Tests**

Classify your concern regarding smoothness violations into two categories:

- **Manipulation** on the running variable $\rightarrow$ McCrary Density Test
- Endogeneity of the cutoff $\rightarrow$ Donut hole RDD

Most robustness is aimed at building credibility around these

Manipulation of your running variable score

- Treatment is not as good as randomly assigned around the cutoff, c_0, when agents are able to manipulate their running variable scores. This happens when:
  1. the assignment rule is known in advance
  2. agents are interested in adjusting
  3. agents have time to adjust
  4. administrative quirks like nonrandom heaping along the running variable
- In other words, we are looking for evidence of people choosing their value of X_i so as to get just barely get into the treatment

McCrary Density Test

- Goal: show no apparant discontinuity in the number of observations around the cutoff

McCrary Density Test

- Goal: show no apparant discontinuity in the number of observations around the cutoff
- Assumes a null where the *density* is continuous at the cutoff point

McCrary Density Test

- Goal: show no apparant discontinuity in the number of observations around the cutoff
- Assumes a null where the *density* is continuous at the cutoff point
- Under the alternative hypothesis, the density increases at the cutoff as people sort onto the desirable side of the cutoff
- This is oftentimes visualized with confidence intervals illustrating the effect of the discontinuity on density - you need no jump to pass this test
- Not perfect, but pretty ingenious and is based on rational choice when you think about it

Steps for a density test in RDD

1. Count observations for a chosen bin (needs multiple units in other words per bin)
2. Estimate your nonlinear OLS model with quadratics in the running variable on the *counts*
3. Do you reject the null at the cutoff? No rejection is good. Rejection is bad.

There are updates to McCrary (2008) using other density tests but this is the basic idea

Simulations of density tests

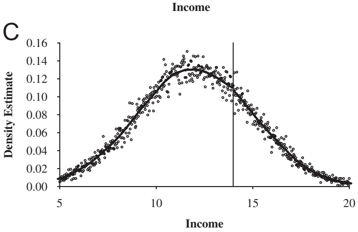


Figure: From McCrary (2008). Left shows failing to reject. Right shows rejection of the null.

Donut hole RDD

- Estimates should not logically be sensitive to the observations at the cutoff – if it is, then smoothness may be violated
- Drop units in the vicinity of the cutoff and re-estimate the model (called “donut hole”)
- Reanalyzing the birthweight mortality data, effects were 50% smaller than previously reported

Other common robustness checks

Table: Robustness checks used in the economics literature (Hausman and Rapson 2018)

Check	Proportion of publications
Data viz	0.79
bandwidth or polynomial order	0.79
Discontinuity test on controls	0.36
Placebo	0.29
Donut hole	0.14
Test for autoregression	0.14 (RDiT)

Discussion: RDD Pros

- Intuitive: mom's kneecap example
- RDD is viewed as very credible among observational designs; for some reason people feel the smoothness assumption is easier to defend

Discussion: RDD Pros

- Intuitive: mom's kneecap example
- RDD is viewed as very credible among observational designs; for some reason people feel the smoothness assumption is easier to defend
- Mild assumptions, easy to justify: It may be because you only have to defend the exogeneity of the treatment at c_0 since you're essentially arguing the potential outcomes wouldn't have jumped there in counterfactual

Discussion: RDD Pros

- Intuitive: mom's kneecap example
- RDD is viewed as very credible among observational designs; for some reason people feel the smoothness assumption is easier to defend
- Mild assumptions, easy to justify: It may be because you only have to defend the exogeneity of the treatment at c_0 since you're essentially arguing the potential outcomes wouldn't have jumped there in counterfactual
- Rewards people who have access to large datasets bc as N grows, the mass at the cutoff should as well, giving you shorter windows for estimation and therefore lower bias and lower variance

Discussion: RDD Caveats

- Not always easy to find a jump
- Obsession on counterfactuals: People want to see counterfactuals (on samples where there is no intervention) as your comparison sets