Regression Discontinuity Designs[1]

Prepared for the PPRN Methods Masterclass

Rocío Titiunik
Department of Politics
Princeton University
titiunik@princeton.edu

March 12, 2021
https://ucl.zoom.us

[1]NSF support through grants SES-1357561, SES-1459931, SES-1947805, SES-2019432.

Overview

Part 1:

Introduction to Regression Discontinuity Designs

Causal Inference and Program Evaluation

- Main goal: learn about treatment effect of policy or intervention
- If treatment randomization available $\rightarrow$ easy to estimate effects
- If treatment randomization not available $\rightarrow$ observational studies
  - Selection on observables
  - Instrumental variables, etc.
- **Regression discontinuity (RD) design**
  - Simple assignment, based on known external factors
  - Objective basis to evaluate assumptions
  - *Careful*: very local!

Regression Discontinuity Design

Defined by the triplet: score, cutoff, treatment.

- Units receive a **score**.
- A treatment is assigned based on the score and a *known* **cutoff**.
- The **treatment** is:
  - given to units whose score is greater than the cutoff.
  - withheld from units whose score is less than the cutoff.
- Under assumptions, the abrupt change in the probability of treatment assignment allows us to learn about the effect of the treatment.

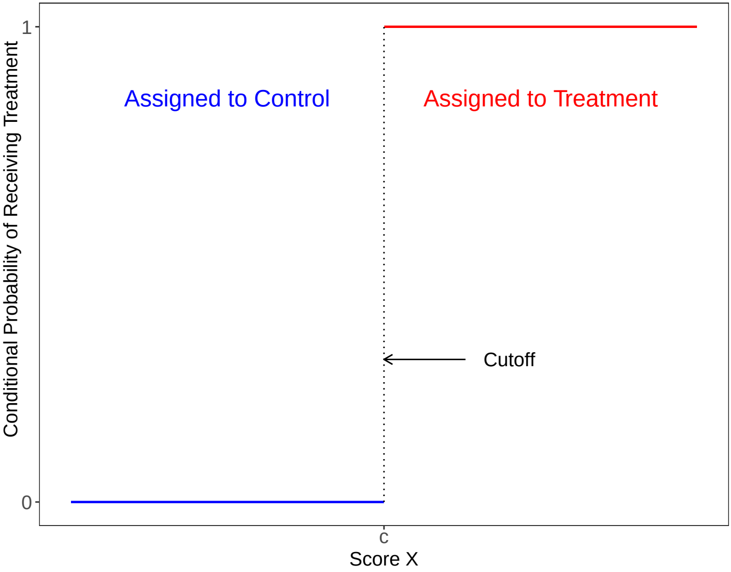
Regression Discontinuity Design

Defined by the triplet: score, cutoff, treatment.

- Units receive a **score**.
- A treatment is assigned based on the score and a *known* **cutoff**.
- The **treatment** is:
  - given to units whose score is greater than the cutoff.
  - withheld from units whose score is less than the cutoff.
- Under assumptions, the abrupt change in the probability of treatment assignment allows us to learn about the effect of the treatment.
- **Some examples**:

	X_i	Y_i
Education:	entry test score	test score, enrollment, performance, etc
Development:	pov index	educ, labor, health, etc
Health:	age / birthdate	insurance coverage, mortality, etc.

Treatment Assignment in (Sharp) RD Design



Sharp Regression Discontinuity Design

- n units, indexed by $i = 1, 2, \ldots, n$
- Unit's score is X_i, treatment is $T_i = \mathbf{1}(X_i \geq c)$
- Each unit has two potential outcomes:

  $Y_i(1)$: outcome that would be observed if i received treatment

  $Y_i(0)$: outcome that would be observed if i received control

- The *observed* outcome is

$$Y_i = \begin{cases} Y_i(0) & \text{if } X_i < c, \\ Y_i(1) & \text{if } X_i \geq c. \end{cases}$$

- Fundamental problem of causal inference: only observe $Y_i(0)$ for units below cutoff and only observe $Y_i(1)$ for units above cutoff

RD Designs: Taxonomy

- **Frameworks**.
  - Identification: Continuity/Extrapolation, Local Randomization.
  - Score: Continuous, Many Repeated, Few Repeated.
- **Settings**.
  - Sharp, Fuzzy, Kink, Kink Fuzzy.
  - Multiple Cutoff, Multiple Scores, Geographic RD.
  - Dynamic, Continuous Treatments, Time, etc.
- **Parameters of Interests**.
  - Average Effects, Quantile/Distributional Effects, Partial Effects.
  - Heterogeneity, Covariate-Adjustment, Differences, Time.
  - Extrapolation.

RCTs

- **Notation**: $(Y_i(0), Y_i(1), X_i)$, $i = 1, 2, \ldots, n$.

- **Treatment**: $T_i \in \{0, 1\}$, $\qquad T_i$ independent of $(Y_i(0), Y_i(1), X_i)$.

- **Data**: (Y_i, T_i, X_i), $i = 1, 2, \ldots, n$, with

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1 \end{cases}$$

- **Average Treatment Effect**:

$$\tau_{\texttt{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i | T = 1] - \mathbb{E}[Y_i | T = 0]$$

(Sharp) RD Designs

- **Notation**: $(Y_i(0), Y_i(1), X_i)$, $i = 1, 2, \ldots, n$, $\qquad X_i$ score.

- **Treatment**: $T_i \in \{0, 1\}$, $\qquad T_i = \mathbb{1}(X_i \geq c)$, $\qquad c$ cutoff.

- **Data**: (Y_i, T_i, X_i), $i = 1, 2, \ldots, n$, with

$$Y_i = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1 \end{cases}$$

- **Average Treatment Effect at the cutoff** (Continuity-based):

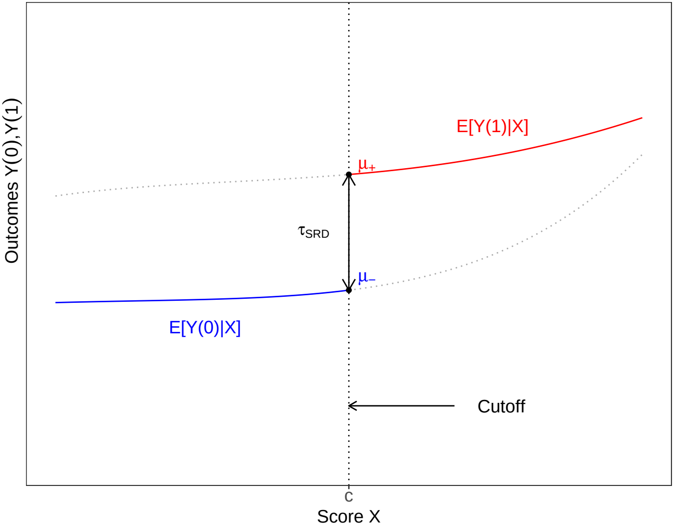
$$\tau_{\texttt{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]$$

- **Average Treatment Effect in a neighborhood** (LR-based):

$$\tau_{\texttt{LR}} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathcal{W}] = \frac{1}{N_1} \sum_{X_i \in \mathcal{W}, T_i = 1} Y_i - \frac{1}{N_0} \sum_{X_i \in \mathcal{W}, T_i = 0} Y_i$$

Continuity-based RD

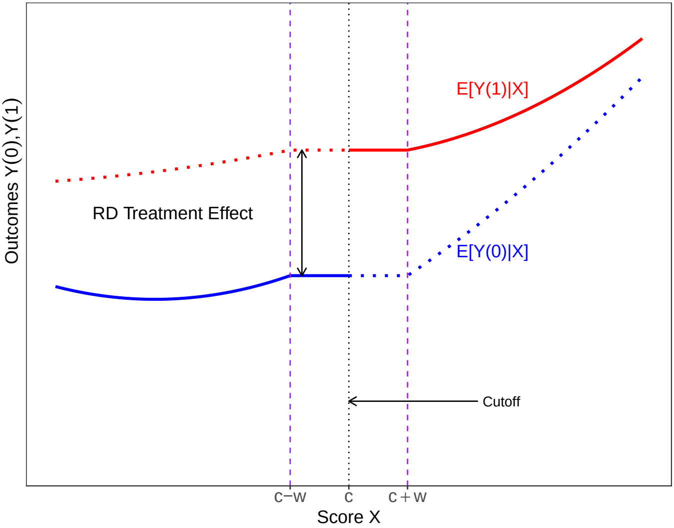
$$\tau_{\texttt{SRD}} = \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]}_{\text{Unobservable}} = \underbrace{\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x]}_{\text{Estimable}} - \underbrace{\lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]}_{\text{Estimable}}$$



Local Randomization RD

T_i independent of $(Y_i(0), Y_i(1))$ for all $X_i \in \mathcal{W} = [c - w, c + w]$

+ exclusion restriction



Part 3:

Graphical illustration of RD models

RD plots

- Main ingredients:
  - Global fit: smooth approximation to the unknown regression functions

    4th or 5th order polynomials, separately above and below the cutoff.
  - Local sample means:

    disjoint intervals (bins) of score, calculating mean of outcome within each bin.
- Main goals:
  - Graphical (heuristic) representation.
  - Detention of discontinuities.
  - Representation of variability.
- Tuning parameters:
  - Global polynomial degree.
  - Location (ES or QS) and number of bins.
- **Great to convey ideas but horrible to draw conclusions**.

Example: Incumbency Advantage in U.S. Senate

- **Problem**: incumbency advantage (U.S. senate).

- **Data**:

  Y_i = Democratic election outcome at $t+1$.

  T_i = whether Democratic party wins election at t .

  X_i = margin of victory at $t \quad (c = 0)$.

  Z_i = covariates (*demvoteshlag1*, *demvoteshlag2*, *dopen*, etc.).

- **Potential outcomes**:

  $Y_i(0)$ = election outcome at $t+1$ if **had not been** incumbent.

  $Y_i(1)$ = election outcome at $t+1$ if **had been** incumbent.

- **Causal Inference**:

  $$Y_i(0) \neq Y_i | T_i = 0 \qquad \text{and} \qquad Y_i(1) \neq Y_i | T_i = 1$$

- Cattaneo, Frandsen & Titiunik (2015, JCI).

RD Packages

https://rdpackages.github.io/

- `rdrobust`: estimation, inference and graphical presentation using local polynomials, partitioning, and spacings estimators; bandwidth selection.
- `rdlocrand` package: covariate balance, binomial tests, randomization inference methods (window selection & inference).
- `rddensity`: discontinuity in density test at cutoff (a.k.a. manipulation testing) using novel local polynomial density estimator.
- `rdmulti`: RD plots, estimation, inference, and extrapolation with multiple cutoffs and multiple scores.
- `rdpower` : power calculation and sample selection for local polynomial methods.

Part 3:

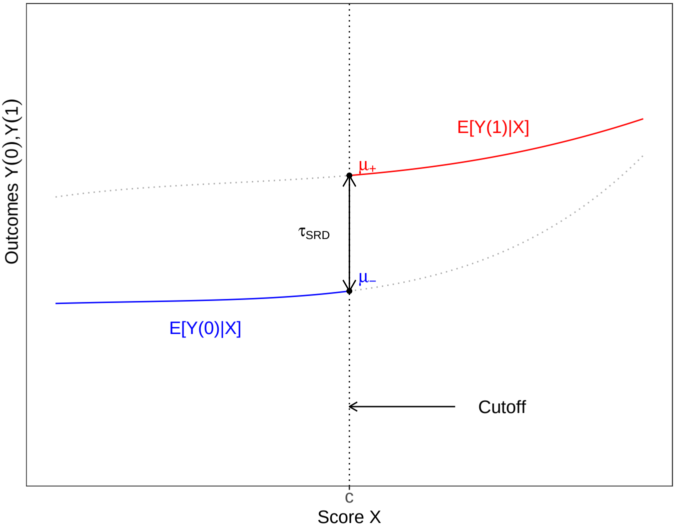
RD Estimation and Inference

Estimation and Inference Methods

- **Continuity/Extrapolation**: Local polynomial approach.
  - Localization: bandwidth selection (trade-off bias and variance).
  - Point estimation: "flexible" (nonparametric).
  - Inference: robust bias-corrected methods.
- **Local Randomization**: finite-sample and large-sample inference.
  - Localization: window selection (via local independence implications).
  - Point estimation: parametric, finite-sample (Fisher) or large-sample (Neyman/SP).
  - Inference: randomization inference (Fisher) or large-sample (Neyman/SP).
- Many refinements and other methods exist (EL, Bayesian, Uniformity, etc.).
  - Do not offer much improvements in applications.
  - Can be overly complicated (lack of transparency).
  - Can depend on user-chosen tuning parameters (lack of replicability).

Continuity-based RD Estimation and Inference

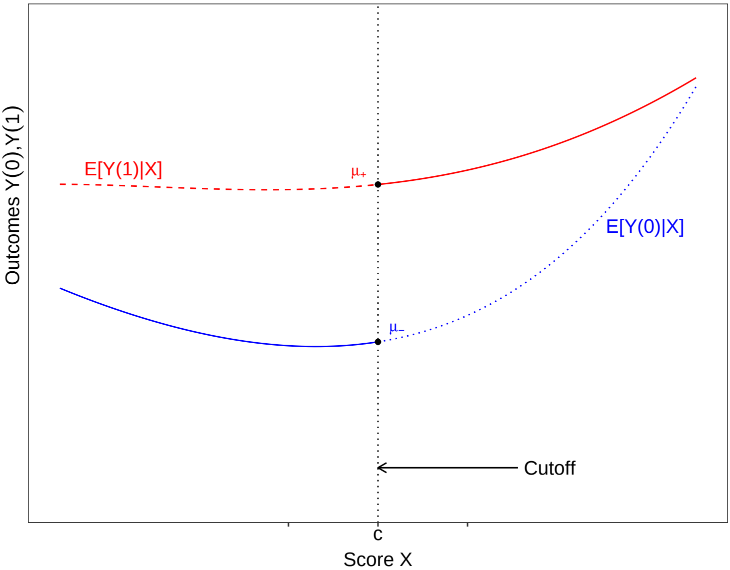
$$\tau = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|X_i = x_0]}_{\text{Unobservable}} = \underbrace{\lim_{x\downarrow x_0}\mathbb{E}[Y_i|X_i = x]}_{\text{Observable}} - \underbrace{\lim_{x\uparrow x_0}\mathbb{E}[Y_i|X_i = x]}_{\text{Observable}}$$



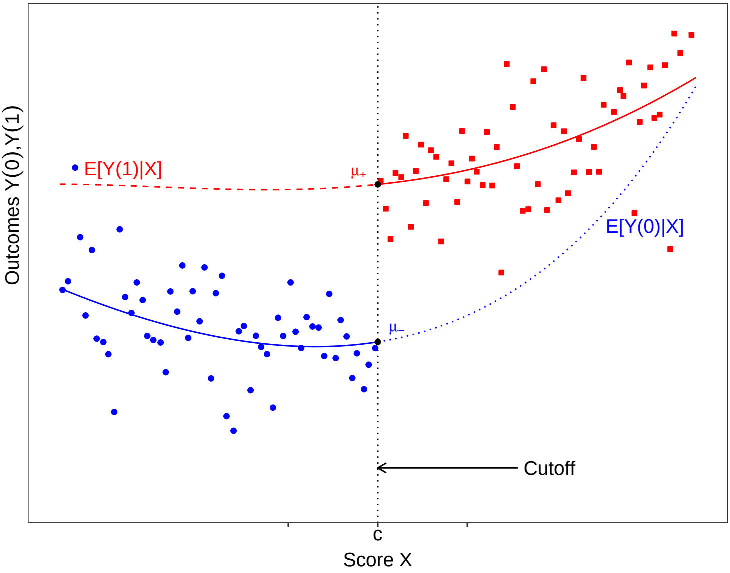
Local Polynomial Treatment Effect Estimation

- $\mathbb{E}[Y_i|X_i = x]$ approximated in neighborhood of x_0 by polynomial function
- Local polynomial estimation:
  - Choose order of polynomial p
  - Choose bandwidth h to keep observations in $[x_0 - h, x_0 + h]$
  - Choose kernel function to weigh observations, $w_i = K(\frac{x_i - x_0}{h})$

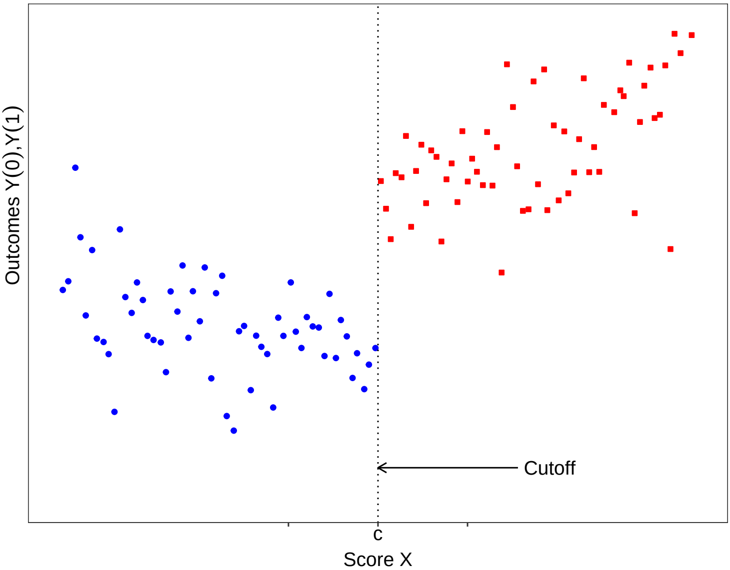
Local Polynomial Estimation



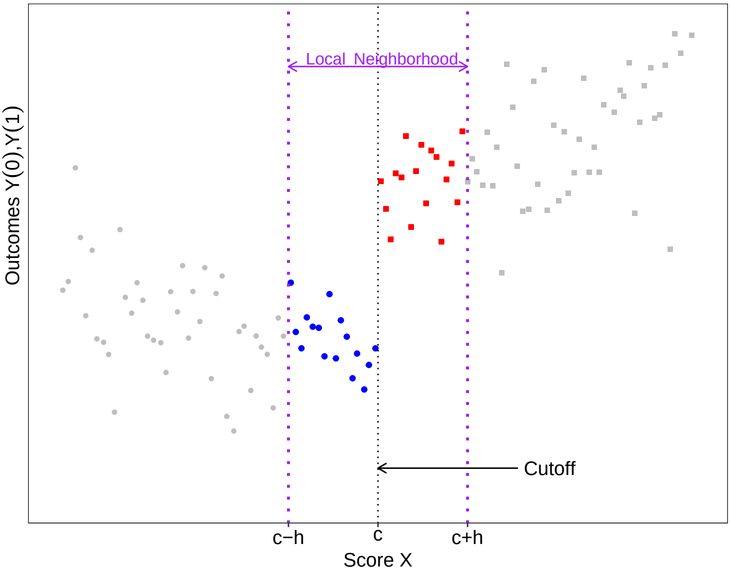
Local Polynomial Estimation



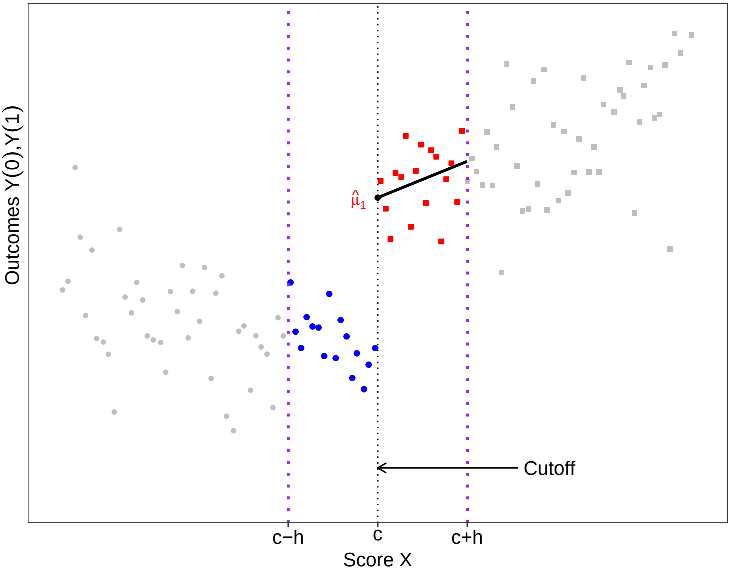
Local Polynomial Estimation



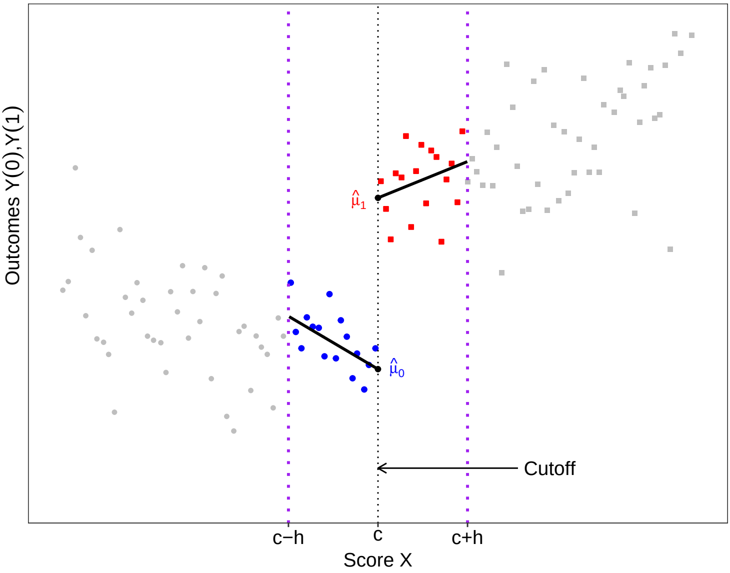
Local Polynomial Estimation



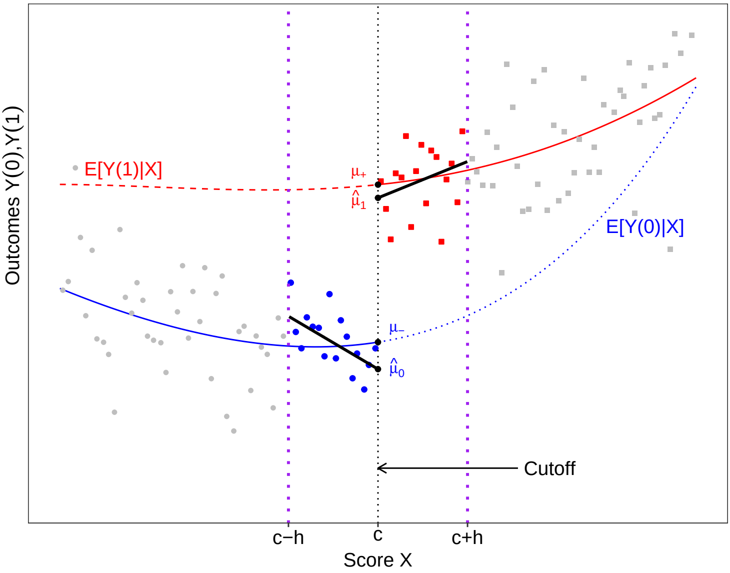
Local Polynomial Estimation



Local Polynomial Estimation



Local Polynomial Estimation



RD Local Polynomial Estimation and Inference

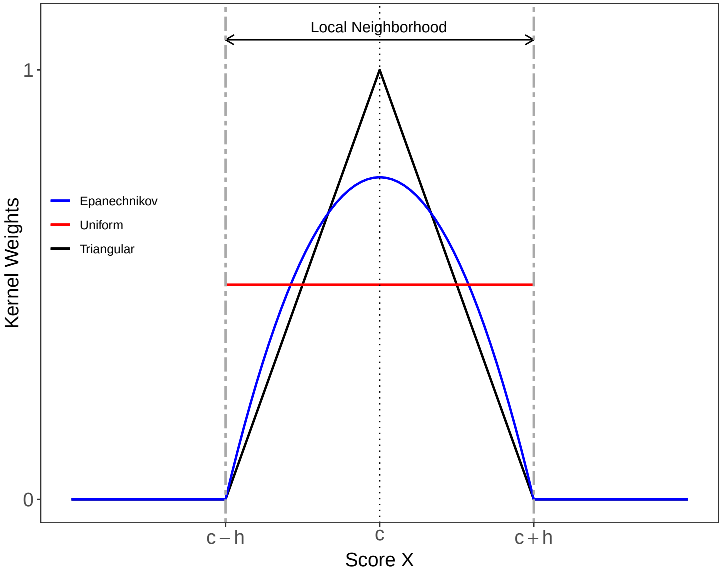
Choose low p and a kernel function $K(\cdot)$

Choose bandwidth h: MSE-optimal or CER-optimal

Construct point estimator $\hat{\tau}_n$ (optimal)

Given above steps, how do we make inferences about τ?

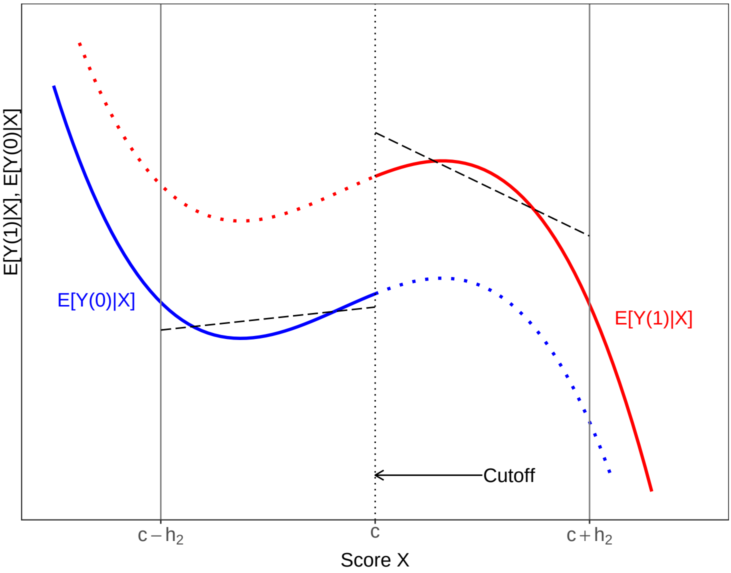
Choice of Kernel Weights



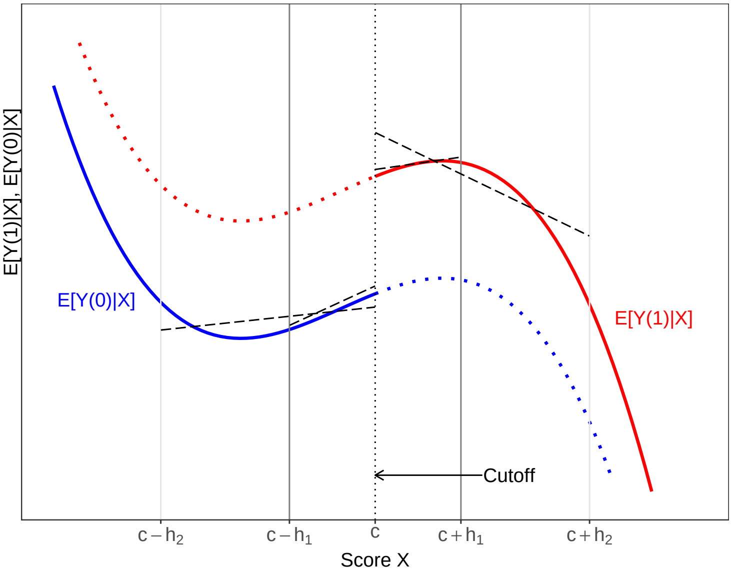
Choice of Polynomial Order (p)

- The higher p, the more flexible the approximation
- However, since approximation is local, p should be low to avoid overfitting
- Given p, approximation can be improved by focusing on a smaller neighborhood around the cutoff
- Standard practice is to choose $p = 1$ (“local linear”)

Approximation for fixed $p = 1$



Approximation for fixed $p = 1$



Choice of Bandwidth

- Given p, find h to ensure optimal properties of the point estimator $\hat{\tau_{RD}}$
- MSE-optimal

$$MSE(\hat{\tau_{RD}}) = Bias^2 + Variance \approx h^{2(p+1)}\mathcal{B}^2 + \frac{1}{nh}\mathcal{V}$$

$$h_{\texttt{MSE}} = C_{\texttt{MSE}}^{1/(2p+3)} \cdot n^{-1/(2p+3)} \qquad C_{\texttt{MSE}} = C(K) \cdot \frac{\mathsf{Var}(\hat{\tau}_{\texttt{SRD}})}{\mathsf{Bias}(\hat{\tau}_{\texttt{SRD}})^2}$$

  - **Key idea**: trade-off bias and variance of point estimator $\hat{\tau}$

$$\uparrow \mathsf{Bias}(\hat{\tau}) \Longrightarrow \downarrow \hat{h} \qquad \text{and} \qquad \uparrow \mathsf{Var}(\hat{\tau}) \Longrightarrow \uparrow \hat{h}$$

- Coverage Error Rate (CER) optimal

$$h_{\texttt{CER}} = n^{-\frac{p}{(3+p)(3+2p)}} \times h_{\texttt{MSE}}$$

  - **Key idea**: choose optimal bandwidth rate to minimize coverage error of the RBC confidence intervals.

Conventional Local Polynomial Point Estimation

- "Local-linear" estimator (w/ weights $K(\cdot)$):

$-h_n \leq X_i < c :$	$c \leq X_i \leq h_n :$
$Y_i = \alpha_0 + (X_i - c) \cdot \beta_0 + \varepsilon_{0,i}$	$Y_i = \alpha_1 + (X_i - c) \cdot \beta_1 + \varepsilon_{1,i}$

- RD effect: $\hat{\tau}_n = \hat{\alpha}_1 - \hat{\alpha}_0$

- When choosing MSE-optimal h, this point estimator $\hat{\tau}_n$ is optimal (also consistent)

Conventional Local Polynomial RD Inference

- RD effect: $\hat{\tau}_n = \hat{\alpha}_1 - \hat{\alpha}_0$
- Once $\hat{\tau}_n$ is estimated with optimal h, we might be tempted to use conventional (OLS) inference
- Construct usual t-statistic. For $\mathsf{H}_0 : \tau = 0$,

$$\mathsf{T} = \frac{\hat{\tau}_n}{\sqrt{\mathsf{V}_n}} = \frac{\hat{\alpha}_1 - \hat{\alpha}_0}{\sqrt{\mathsf{V}_{1,n} + \mathsf{V}_{0,n}}} \to_{\mathrm{d}} \mathcal{N}(0,1)$$

- 95% Confidence interval:

$$\mathsf{CI} = \left[\ \hat{\tau}_n \ \pm \ 1.96 \cdot \sqrt{\mathsf{V}_n}\ \right]$$

Conventional Local Polynomial RD Inference

- However, with conditions on $h_n \to 0$, the distributional approximation

$$\mathsf{T} = \frac{\hat{\tau}_n}{\sqrt{\mathsf{V}_n}} \to_{\mathrm{d}} \mathcal{N}(\mathsf{B}_n, 1) \quad \neq \mathcal{N}(0, 1)$$

  - Bias B_n in RD point estimator captures "curvature" of regression functions
  - In particular, the bias B_n occurs when the MSE-optimal bandwidth is used

- Conventional approach $\to$ assume bias negligible or undersmoothing

$$\mathsf{T} = \frac{\hat{\tau}_n}{\sqrt{\mathsf{V}_n}} \to_{\mathrm{d}} \mathcal{N}(0, 1) \qquad \Bigg| \qquad \mathsf{CI} = \left[\ \hat{\tau}_n \ \pm \ 1.96 \cdot \sqrt{\mathsf{V}_n}\ \right]$$

  $\Longrightarrow$ Not clear guidance & power loss!

- Bias-correction approach

$$\mathsf{T}^{\mathrm{bc}} = \frac{\hat{\tau}_n - \mathsf{B}_n}{\sqrt{\mathsf{V}_n}} \to_{\mathrm{d}} \mathcal{N}(0, 1) \qquad \Bigg| \qquad \mathsf{CI}^{\mathrm{bc}} = \left[\ \left(\hat{\tau}_n - \hat{\mathsf{B}}_n\right) \ \pm \ 1.96 \cdot \sqrt{\mathsf{V}_n}\ \right]$$

  $\Longrightarrow$ Poor finite sample properties!

Robust Local Polynomial Inference

- Key observation: $\hat{\mathsf{B}}_n$ is constructed to estimate leading bias

$$\mathsf{T}^{\mathrm{bc}} = \frac{\hat{\tau}_n - \hat{\mathsf{B}}_n}{\sqrt{\mathsf{V}_n}} = \underbrace{\frac{\hat{\tau}_n - \mathsf{B}_n}{\sqrt{\mathsf{V}_n}}}_{\to_d \mathcal{N}(0,1)} + \underbrace{\frac{\mathsf{B}_n - \hat{\mathsf{B}}_n}{\sqrt{\mathsf{V}_n}}}_{\to_p 0}$$

- **Our robust approach** $\to$ *Non-standard Asymptotics*

$$\mathsf{T}^{\mathrm{bc}} = \frac{\hat{\tau}_n - \hat{\mathsf{B}}_n}{\sqrt{\mathsf{V}_n}} = \underbrace{\frac{\hat{\tau}_n - \mathsf{B}_n}{\sqrt{\mathsf{V}_n}}}_{\to_d \mathcal{N}(0,1)} + \underbrace{\frac{\mathsf{B}_n - \hat{\mathsf{B}}_n}{\sqrt{\mathsf{V}_n}}}_{\to_d \mathcal{N}(0,\gamma)}$$

- Robust Bias-Correction Approach:

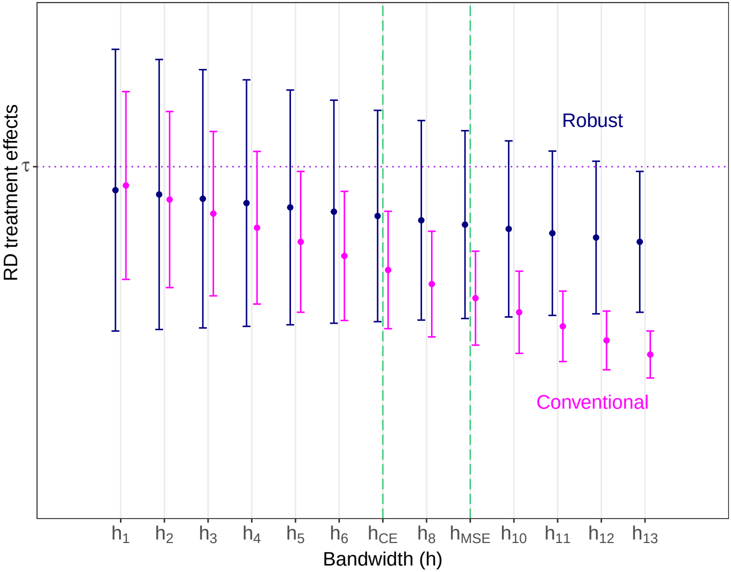
$$\mathsf{T}^{\mathrm{rbc}} = \frac{\hat{\tau}_n - \hat{\mathsf{B}}_n}{\sqrt{\mathsf{V}_n + \mathsf{W}_n}} \to_d \mathcal{N}(0, 1)$$

$$\mathsf{CI}^{\mathrm{rbc}} = \left[\left(\hat{\tau}_n - \hat{\mathsf{B}}_n \right) \pm 1.96 \cdot \sqrt{\mathsf{V}_n + \mathsf{W}_n} \right]$$

Table: Local Polynomial Confidence Intervals

	Centered at	Standard Error
Conventional: $\texttt{CI}_{\texttt{us}}$	$\hat{\tau}_{\texttt{SRD}}$	$\sqrt{\hat{\mathscr{V}}}$
Bias Corrected: $\texttt{CI}_{\texttt{bc}}$	$\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathscr{V}}}$
Robust Bias Corrected: $\texttt{CI}_{\texttt{rbc}}$	$\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}}$

Confidence Intervals for Different Bandwidths



Local Randomization RD Estimation and Inference

Local Randomization Approach to RD Design

- Gives an alternative that can be used as a robustness check.
- **Key assumption**: exists window $W = [-w, w]$ around cutoff $(-w < c < w)$ where (assuming random potential outcomes)

$$T_i \text{ independent of } (Y_i(0), Y_i(1)) \qquad \text{(for all } X_i \in W)$$

- Thus, inside W_0 subjects are as-if randomly assigned to either side of cutoff
  - The distribution of running variable same for all units inside W_0
  - Potential outcomes in W_0 depend on running variable only through threshold indicators within W_0
- Stronger than Continuity-Based Approach$\Rightarrow$ Relevant population functions are not only continuous at x_0, but also completely unaffected by the running variable in W_0

Local Randomization Framework

- **Key idea**: treatment assignment as-if randomly assigned "near" cutoff. There exists window $\mathcal{W} = [-w, w]$, with $-w < c < w$, such that

$$\text{for all } X_i \in \mathcal{W} \Longrightarrow T_i \text{ independent of } (Y_i(0), Y_i(1))$$

  and possibly other conditions hold (e.g., knowledge of assignment mechanism).

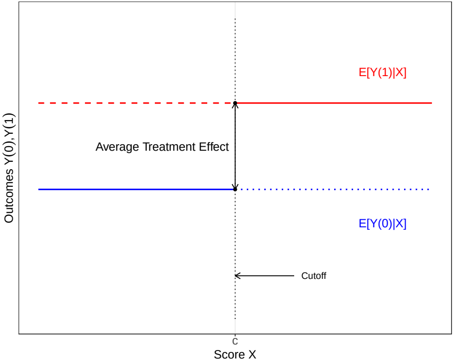
  - Conceptually different from continuity/extrapolation based methods.
  - Challenge: window (neighborhood) selection.
  - Challenge: small sample (estimation and) inference.

- **Two Steps** (analogous to local polynomial methods):

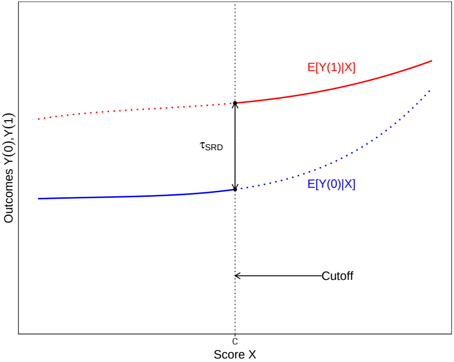
  1. Select window $\mathcal{W}$ based on "covariate balance" idea.
  2. Given window $\mathcal{W}$, (estimation and) inference is "standard": superpopulation, large-samples designed-based methods, randomization inference methods.

- **Catch**: as-if random assumption good approximation *only very near cutoff*!

Experiment versus RD Design

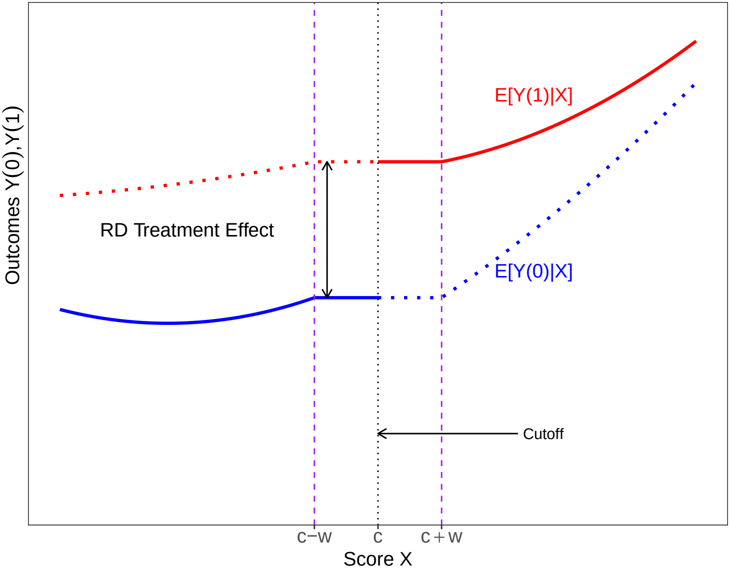


(a) Randomized Experiment



(b) RD Design

If as-if random interpretation is true: Local Randomization RD

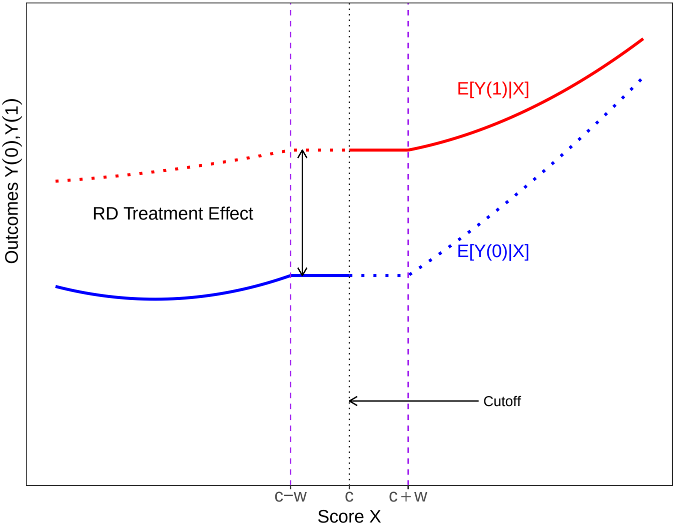


Step 1: Choose the window W_0

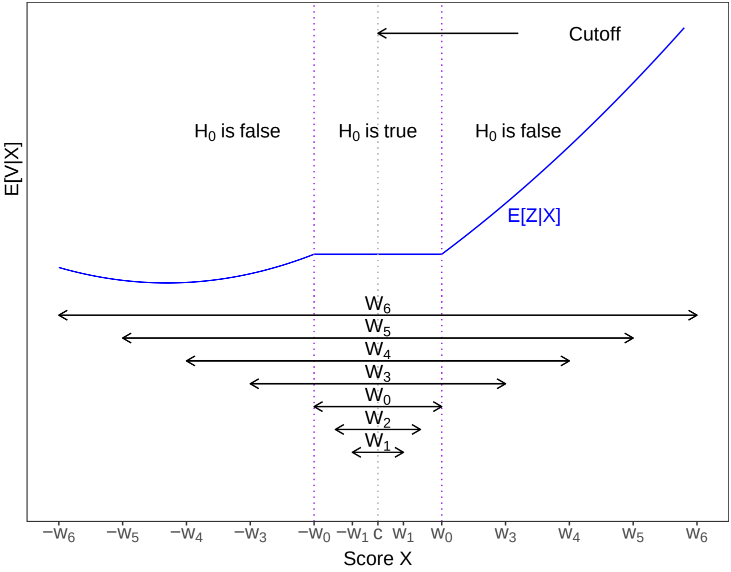
- How to choose $\mathcal{W}$?
  - Find neighborhood where (pre-intervention) covariate-balance holds.
  - Find neighborhood where outcome and score independent.
  - Domain-specific or application-specific choice.

T_i independent of $(Y_i(0), Y_i(1))$ for all $X_i \in \mathcal{W} = [c - w, c + w]$

+ exclusion restriction



Window Selector Based on Covariate Balance in Locally Random RD



Step 2: Finite-sample and Large-sample Methods in $\mathcal{W}$

- Given $\mathcal{W}$ where local randomization holds:
  - Randomization inference (Fisher): sharp null, finite-sample exact.
  - Design-based (Neyman): large-sample valid, conservative.
  - Large-sample standard: random potential outcomes, large-sample valid.
- All methods require window ($\mathcal{W}$) selection, and choice of statistic.
  First two also require choice/assumptions assignment mechanism.
  Covariate-adjustments (score or otherwise) possible.

Example: Incumbency Advantage in U.S. Senate

- **Problem**: incumbency advantage (U.S. senate).

- **Data**:

  $Y_i =$ election outcome at $t+1$.

  $T_i =$ whether party wins election at t .

  $X_i =$ margin of victory at $t \quad (c = 0)$.

  $Z_i =$ covariates (*demvoteshlag1*, *demvoteshlag2*, *dopen*, etc.).

- **Potential outcomes**:

  $Y_i(0) =$ election outcome at $t+1$ if **had not been** incumbent.

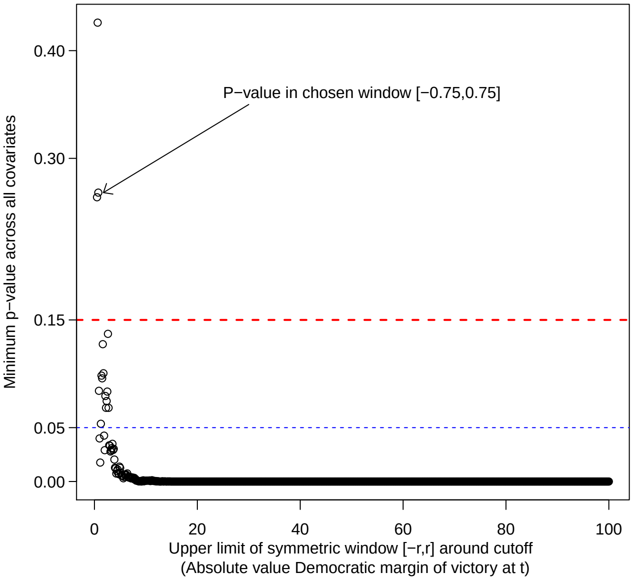
  $Y_i(1) =$ election outcome at $t+1$ if **had been** incumbent.

- **Causal Inference**:

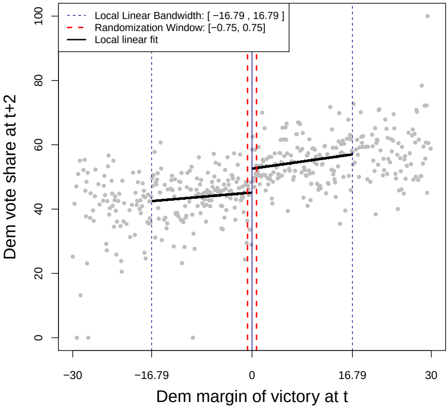
  $$Y_i(0) \neq Y_i | T_i = 0 \qquad \text{and} \qquad Y_i(1) \neq Y_i | T_i = 1$$

- Cattaneo, Frandsen & Titiunik (CFT, 2015, Journal of Causal Inference).

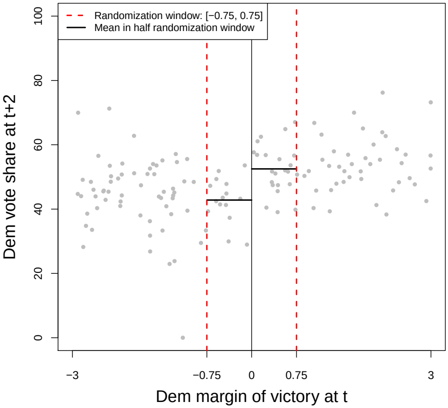
Window Selection Based on Covariates, CFT



Continuity-Based vs Local Randomization Analysis, CFT



(a) Continuity-Based Analysis



(b) Local Randomization Analysis

Part 4:

Falsification Analysis for RD designs

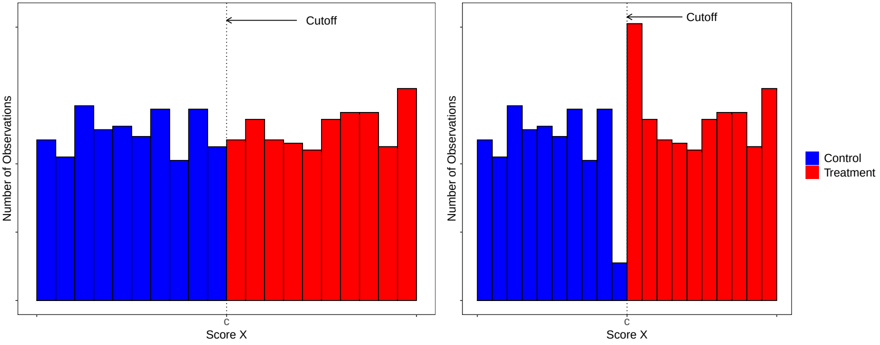
Falsification Methods

- RD rule of treatment assignment is not by itself enough to guarantee that continuity or local randomizations are met
- Qualitative information and quantitative falsification tests play crucial role
  - Qualitative information: were there mechanisms to appeal score? did people change their score?
  - Falsification: various statistical tests

Falsification Methods

- Density test of “sorting”: is number of observations below the cutoff surprisingly different from number of observations above it?
- Treatment effect on
  - Predetermined covariates
  - Placebo outcomes
- Also: effect at different cutoffs, effect at different bandwidths, dougnout hole

Falsification Methods: Density Test



(a) No sorting (b) Sorting

Falsification Methods: Tests on Predetermined Covariates and Placebo Outcomes

- Continuity-based falsification:
  - Test of continuity of density of the running variable
  - Local polynomial effects with optimal banwidth
  - Robust Inference
  - CRUCIAL: each covariate/placebo outcome must have its own optimal bandwidth
- Local randomization falsification:
  - Within chosen window, density test
  - Test that covariate and placebo outcome distributons are indistinguishable for treated and control
  - CRUCIAL: all tests are conducted within the same window for each covariate/placebo outcome

Thank you!

https://rdpackages.github.io/