Regression Discontinuity Design

Prof. Tzu-Ting Yang
楊子霆

Institute of Economics, Academia Sinica
中央研究院經濟研究所

May 13, 2026

Introduction

Introduction

Selection Bias and RCT

- A major problem of estimating causal effect of treatment is the threat of **selection bias**
- In many situations, individuals can **select into treatment** so those who get treatment could be very different from those who are untreated
- The best to deal with this problem is conducting a randomized controlled trial (RCT)

Main Idea of Regression Discontinuity Design

- In an RCT, researchers can eliminate selection bias by **controlling treatment assignment process**
  - An RCT randomizes who receives a treatment –the treatment group - and who does not –the control group
  - Since we randomly assign treatment, the probability of getting treatment is unrelated to other confounding factors
- But conducting an RCT is very expensive and may have ethical issue

Main Idea of Regression Discontinuity Design

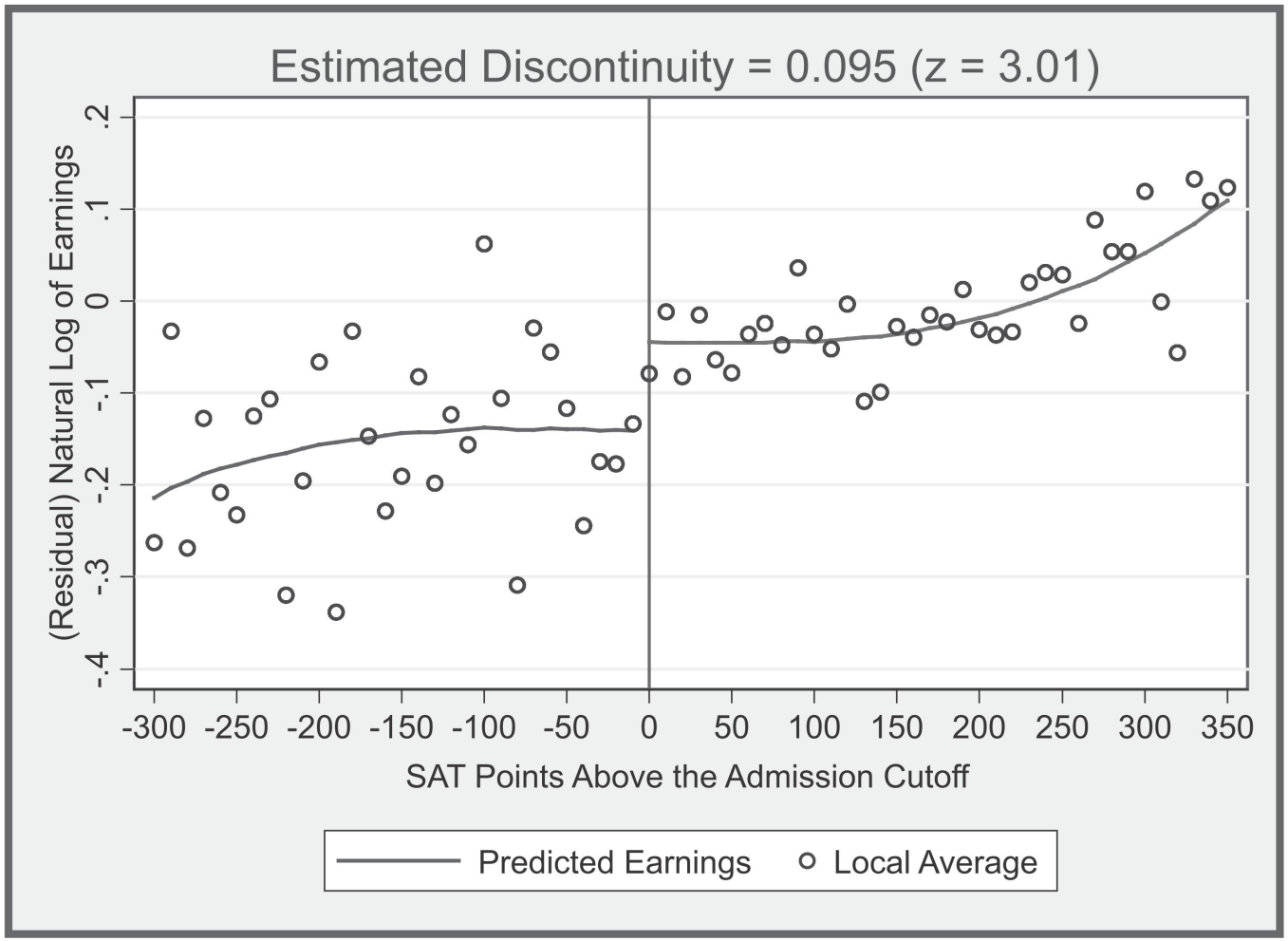
- Instead of controlling treatment assignment process, if researchers have **detailed institutional knowledge of treatment assignment process**
- Then we could use this information to create an "experiment"

Main Idea of Regression Discontinuity Design

- Regression Discontinuity Design (RDD) exploits the facts that:
  - Some rules can generate a discontinuity in treatment assignment
    - The treatment assignment is determined based on whether a unit exceeds some threshold on a variable.
    - Such variable is called **assignment variable**, **running variable** or **forcing variable**
  - Assume other factors do NOT change abruptly at threshold
  - Then any change in outcome of interest can be attributed to the assigned treatment

SAT Score and Earnings

FIGURE 2.—NATURAL LOG OF ANNUAL EARNINGS FOR WHITE MEN TEN TO FIFTEEN YEARS AFTER HIGH SCHOOL GRADUATION (FIT WITH A CUBIC POLYNOMIAL OF ADJUSTED SAT SCORE)



More Examples

- **Where there is a cutoff there is a RD**

Public Health

The Effect of Health Intervention

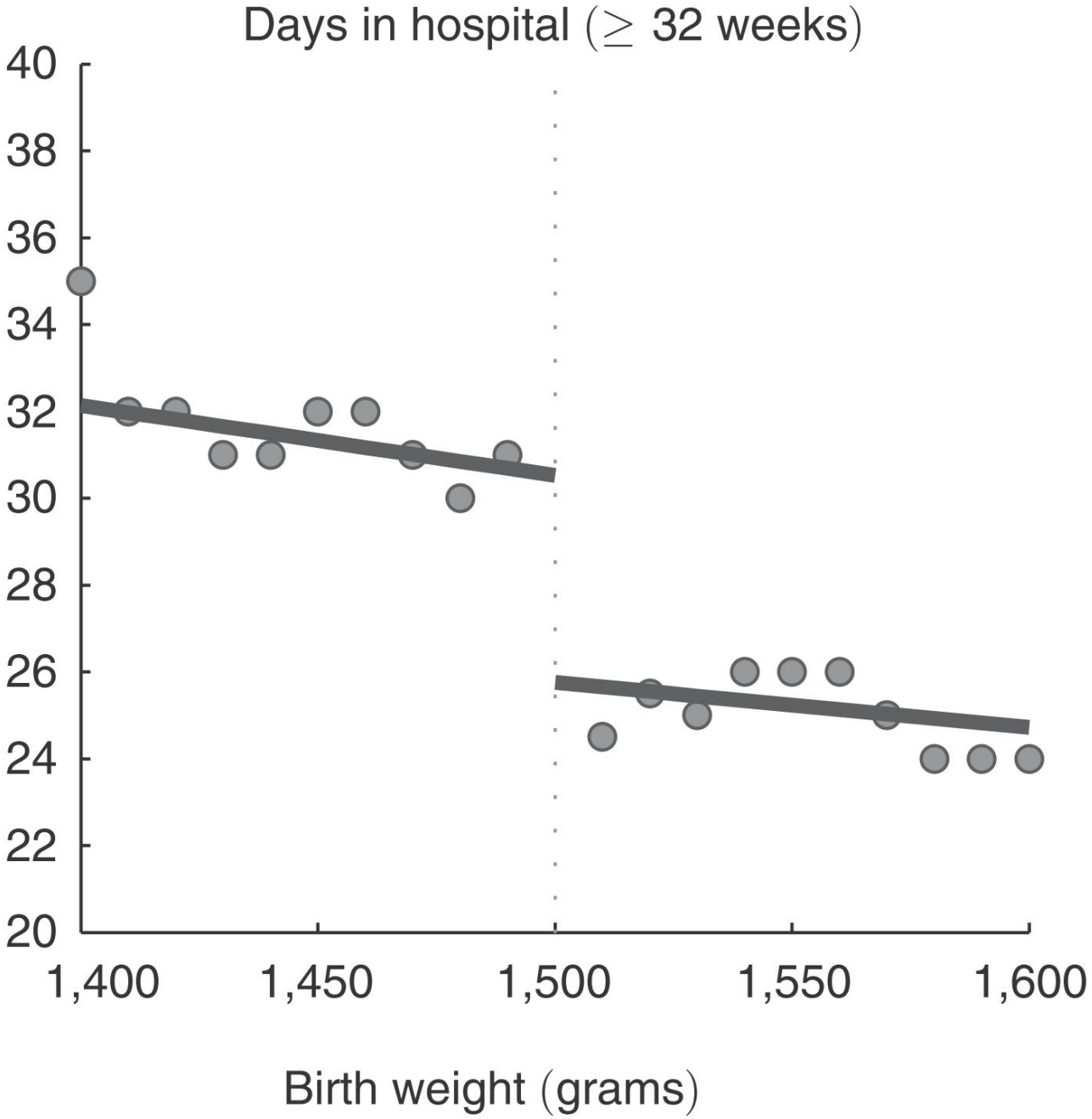
Prashant Bharadwaj, Katrine Vellesen Løken, and Christopher Neilson (2013) "**Early Life Health Interventions and Academic Achievement**" AER

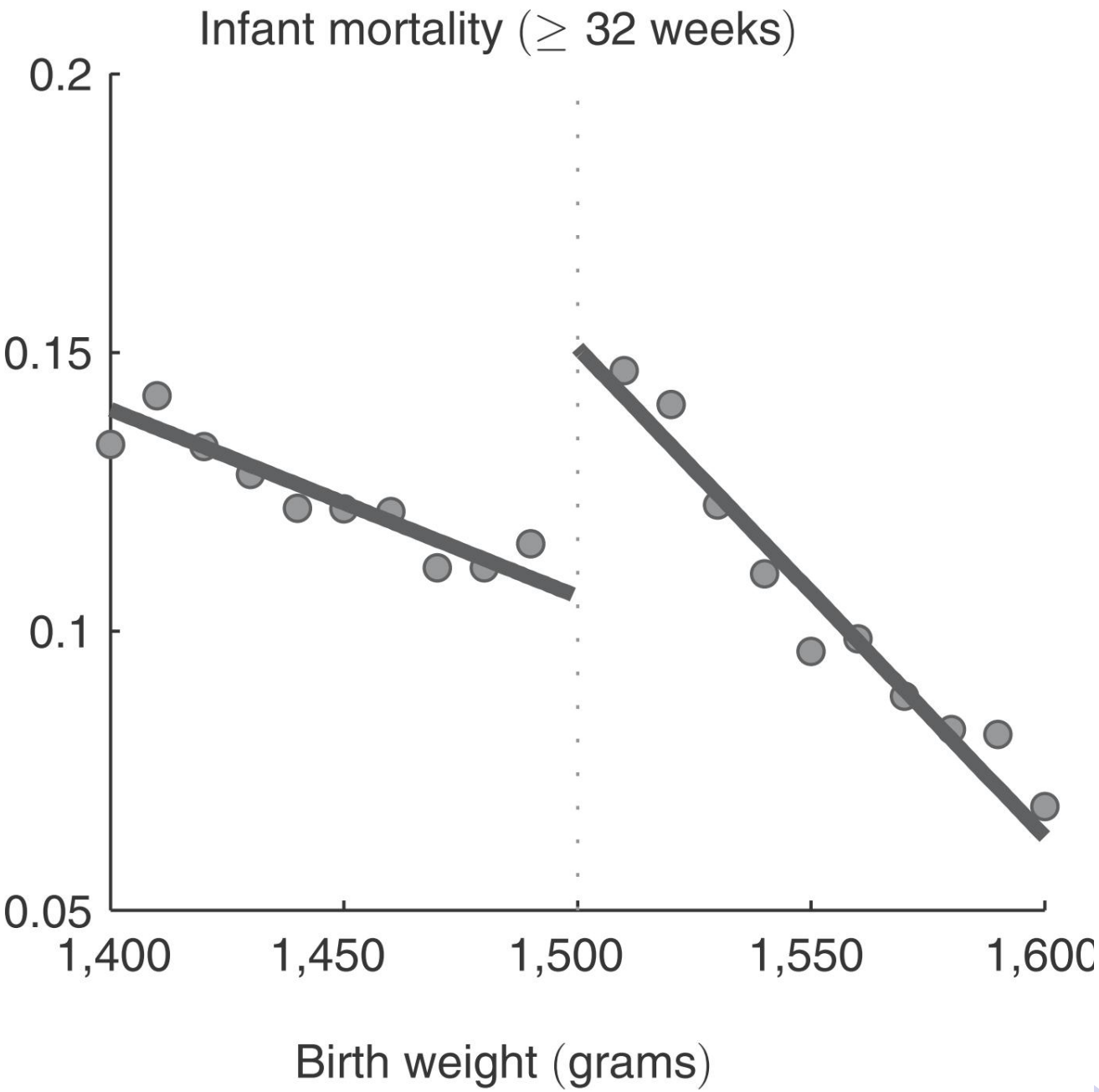
- The effect of **health intervention** in early childhood on **later life outcomes**
- **Selection bias**: children who need health intervention in early childhood could be very sick and might have bad later life outcome (e.g. low educational attachment)

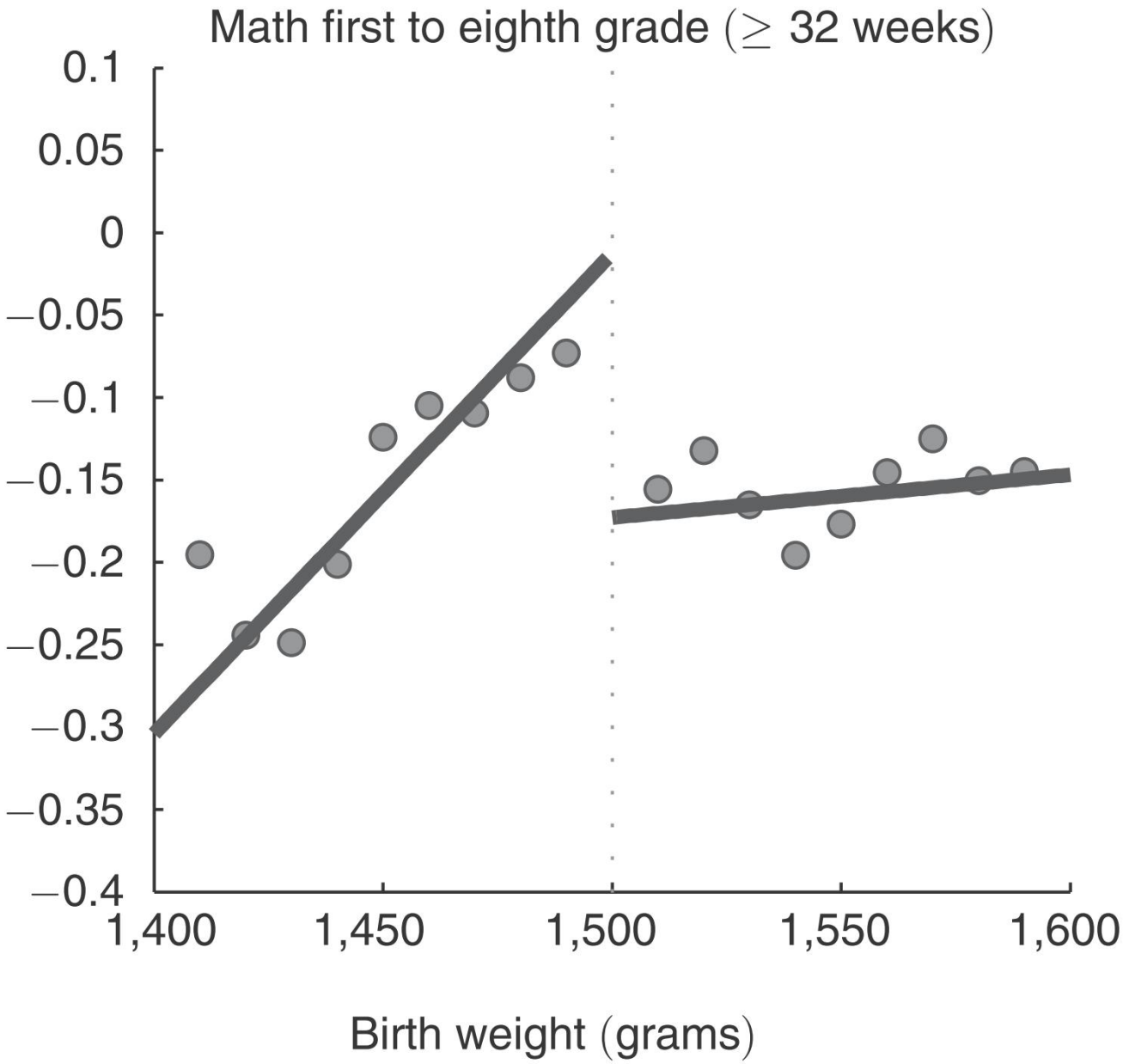
Public Health

The Effect of Health Intervention

- **RDD solution**:
  - Infants with a birth weight below **1500 grams** were eligible for additional healthcare while those with a birth weight just above the cutoff were not eligible
- Compares mortality rates and academic achievement between those infants **just below and above the cutoff of 1500 grams**


Days in hospital (≥ 32 weeks)
40
38
36
34
32
30
28
26
24
22
20
1,400
1,450
1,500
1,550
1,600
Birth weight (grams)


Infant mortality ($\geq$ 32 weeks)
0.2
0.15
0.1
0.05
1,400
1,450
1,500
1,550
1,600
Birth weight (grams)

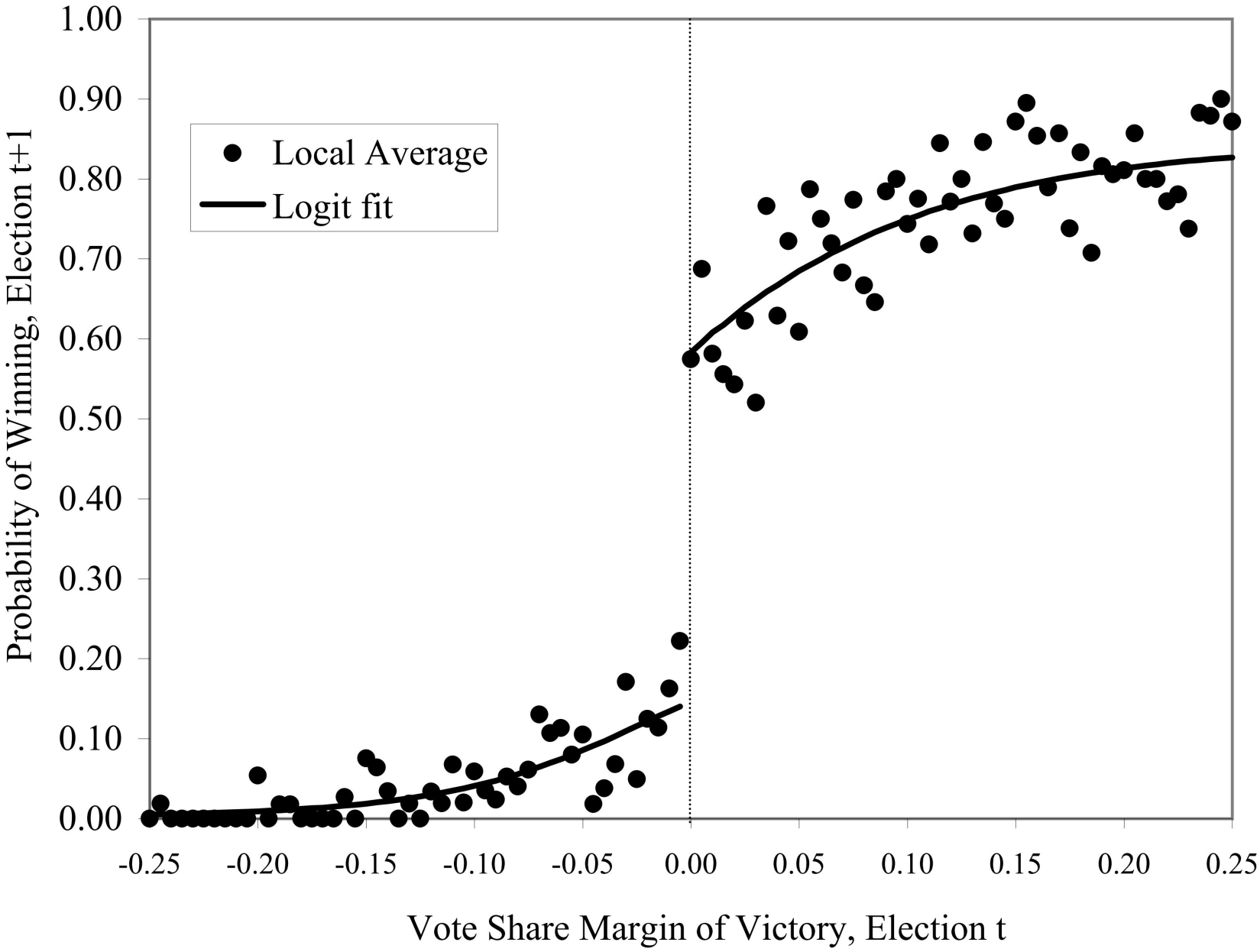

Math first to eighth grade (≥ 32 weeks)
0.1
0.05
0
−0.05
−0.1
−0.15
−0.2
−0.25
−0.3
−0.35
−0.4
1,400
1,450
1,500
1,550
1,600
Birth weight (grams)

David Lee (2007) "**Randomized Experiments from Non-random Selection in U.S. House Elections**" Journal of Econometrics

- Does **political incumbency** provide an **electoral advantage**
- **Selection bias**: people who win election (incumbency) should be more popular

- **RDD solution**:
  - Candidates who just **barely won** an election (barely became the incumbent) are likely to be ex ante comparable in all other ways to candidates who **barely lost**
  - So their differential electoral outcomes in the **next election** should represent a true **incumbency advantage**

Figure IIa: Candidate's Probability of Winning Election t+1, by Margin of Victory in Election t: local averages and parametric fit



Rafael Lalive (2008), "**How do extended benefits affect unemployment duration? A regression discontinuity approach**", Journal of Econometrics

- This paper studies a targeted program that extends the maximum duration of unemployment benefits from 30 weeks to 209 weeks in Austria

- Sharp discontinuities in treatment assignment at age 50 and at the border between eligible regions and control regions identify the effect of **extended benefits** on **unemployment duration**

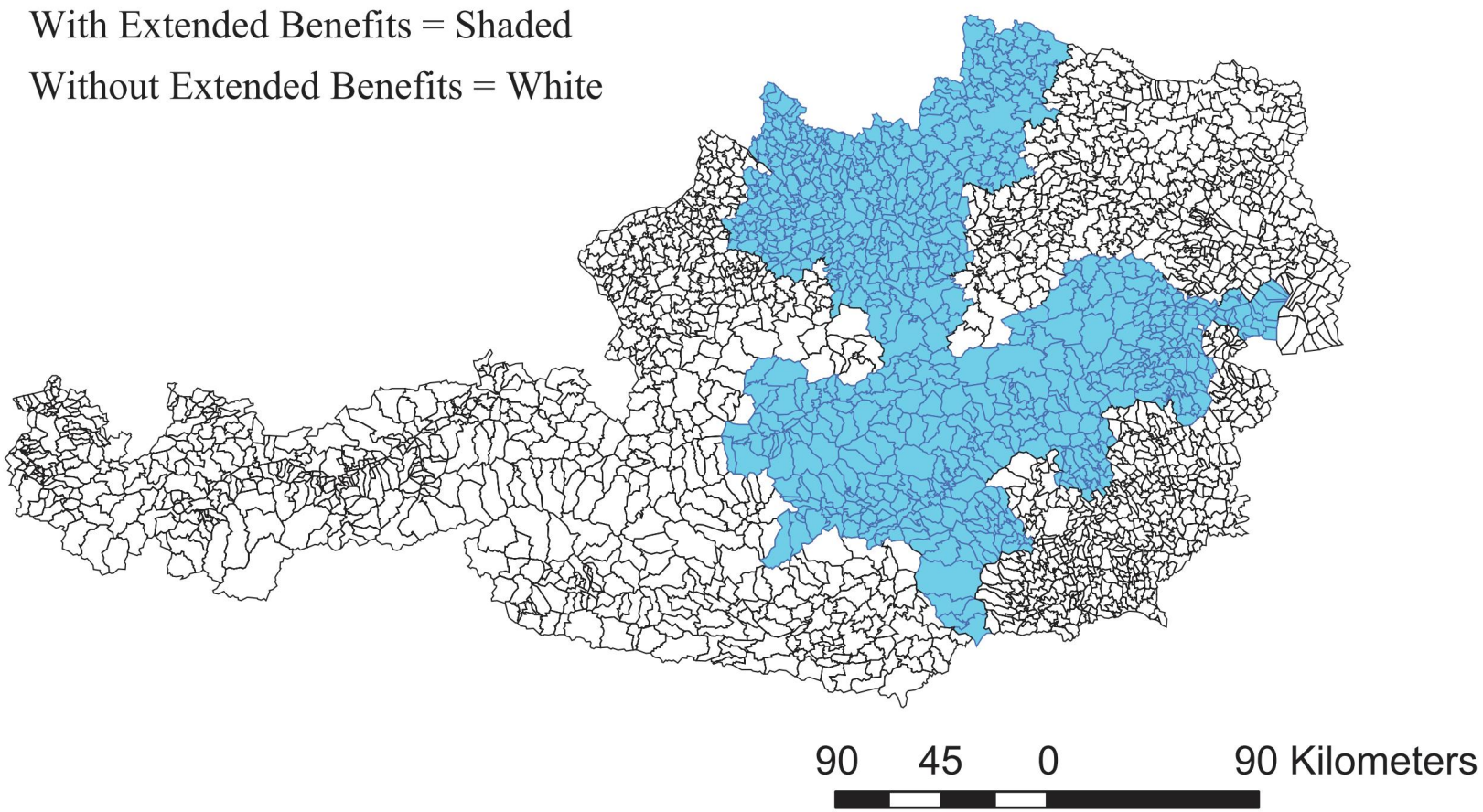
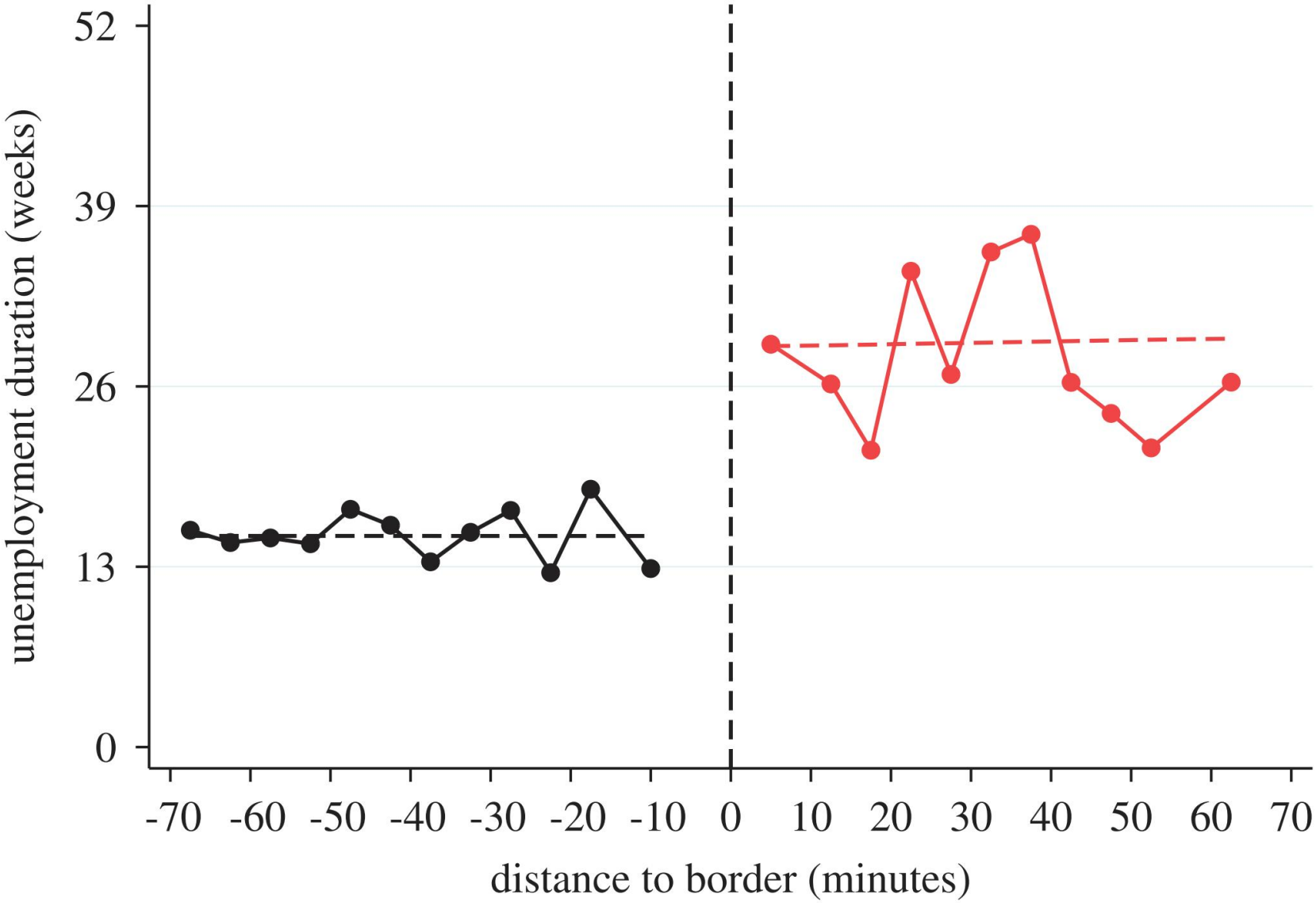


Fig. 1. Regional distribution of REBP.

Labor Economics

Spatial Regression Discontinuity



RDD Became Popular since late 1990s

- The first RDD paper is Thistlethwaite and Campbell (1960), "RD Analysis: An Alternative to Ex Post Fact Experiments," Journal of Education Psychology
- RDD was not used much in economics until the late 1990s
- But hundreds of studies since then, starting with Van der Klaauw (2002)
- Two possible explanations:
  - Cutoff rules are very wide spread$\cdots$
  - Much more data available now, especially administrative data sets

RDD Became Popular since late 1990s

- An important advantage of RD designs is that they are well suited to large administrative data sets with
  - Few covariates
  - Lots of observations and all the relevant information about cutoffs and assignment variables
  - Since those have to be used in the administration of programs

Sharp RDD and Fuzzy RDD

- In general, depending on enforcement of treatment assignment, RDD can be categorized into two types:
  1. Fuzzy RDD: the probability of getting the treatment jumps discontinuously at the cutoff
  2. Sharp RDD: the probability of getting the treatment jumps from 0 to 1 at the cutoff

Fuzzy RDD: Setup

Fuzzy RDD

Overview

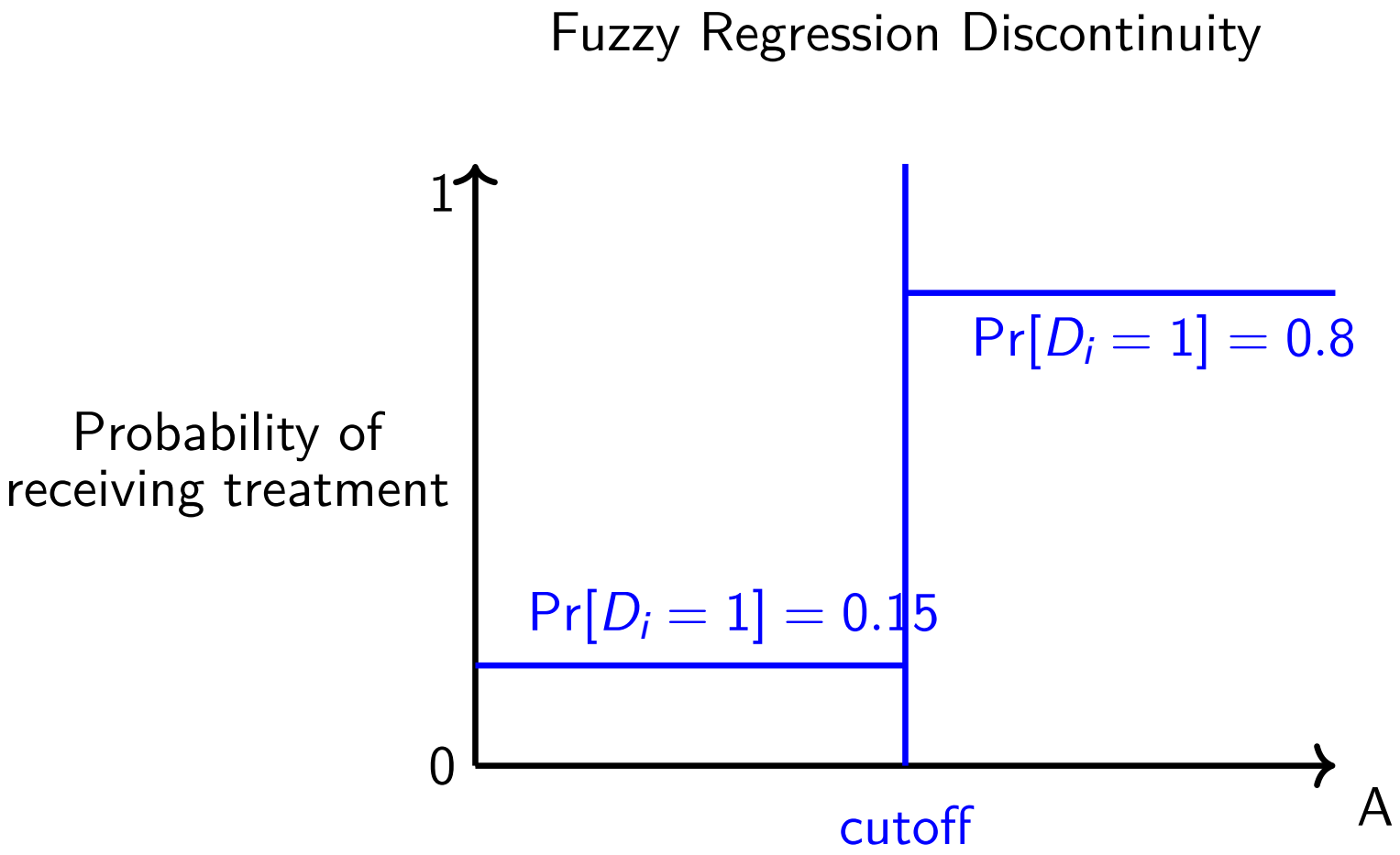
Fuzzy RDD

$$\lim_{\varepsilon \to 0} \Pr[D_i = 1 | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \Pr[D_i = 1 | A_i = c - \varepsilon] \neq 0$$

- The probability of getting the treatment jumps discontinuously at the cutoff
  - Some individuals above cutoff do NOT get treatment and some individuals below cutoff do receive treatment

Treatment Probability and assignment variable

Fuzzy RDD



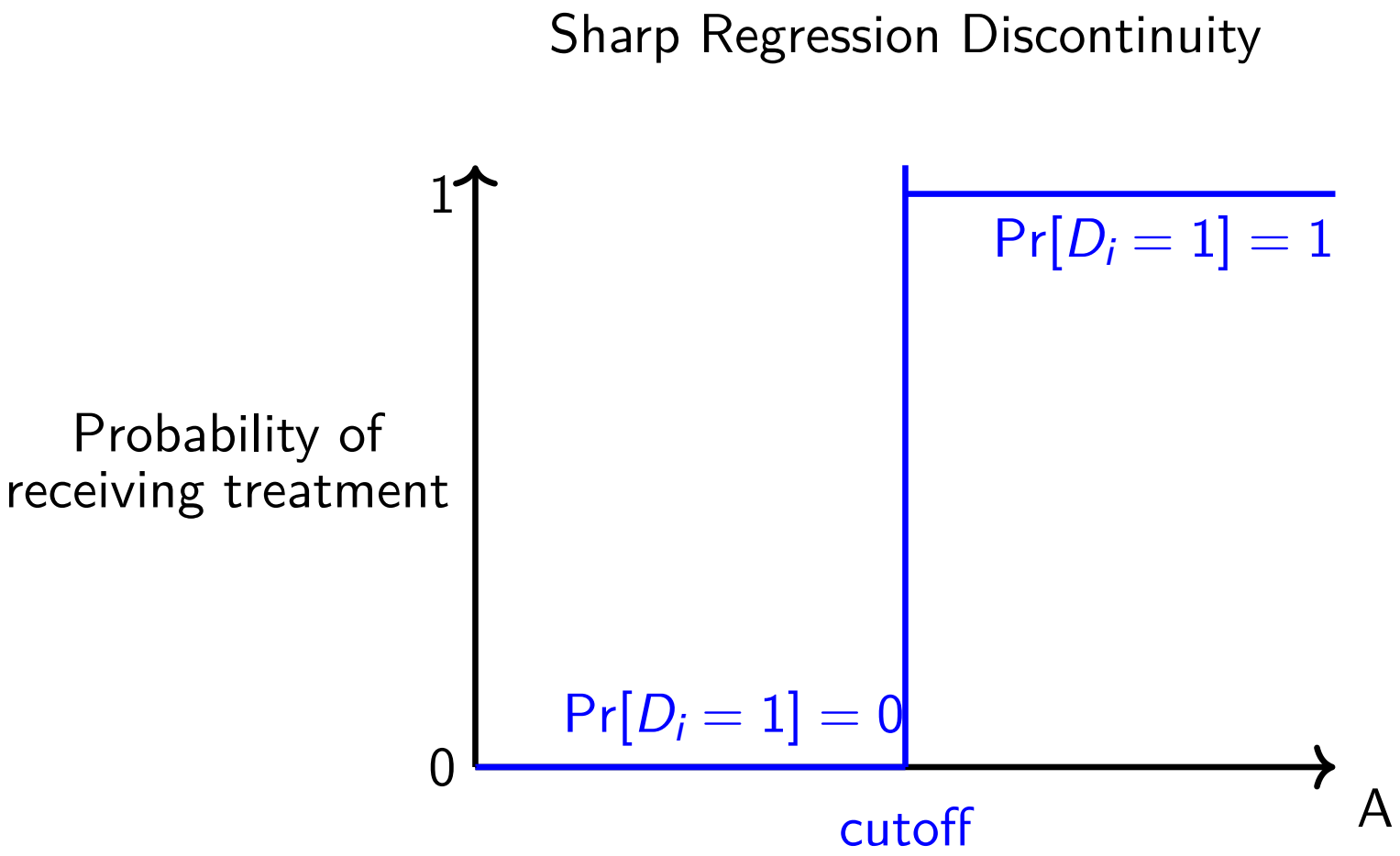
Sharp RDD

$$\lim_{\varepsilon \to 0} \Pr[D_i = 1 | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \Pr[D_i = 1 | A_i = c - \varepsilon] = 1$$

- Sharp RDD is a special case of fuzzy RDD
- It further required the jump in probability to be from 0 to 1
  - Nobody below the cutoff gets the "treatment", everybody above the cutoff gets it

Treatment Probability and assignment variable

Sharp RDD



Potential Outcomes Framework

- Treatment Eligibility

$$Z_i = \begin{cases} 1 & \text{if } A_i \geq c, \text{ eligible for a treatment} \\ 0 & \text{if } A_i < c, \text{ not eligible for a treatment} \end{cases}$$

Potential Outcomes Framework

- Potential Treatments
  - D_i^z: treatment status given the value of Z
  - D_i^1: treatment status if eligible for a treatment (above cutoff c)

$$\mathrm{D}_i^1 = \begin{cases} 1 & \text{if getting a treatment} \\ 0 & \text{if not getting a treatment} \end{cases}$$

  - D_i^0: treatment status if not eligible for a treatment (below cutoff c)

$$\mathrm{D}_i^0 = \begin{cases} 1 & \text{if getting a treatment} \\ 0 & \text{if not getting a treatment} \end{cases}$$

Potential Outcomes Framework

- Observed Treatment

$$D_i = \begin{cases} \mathrm{D}_i^1 & \text{if } Z_i = 1, A_i \geq c \\ \mathrm{D}_i^0 & \text{if } Z_i = 0, A_i < c \end{cases}$$

- or, in a more compact notation: $\mathrm{D}_i = Z_i \mathrm{D}_i^1 + (1 - Z_i)\mathrm{D}_i^0$

Potential Outcomes Framework

- In sharp RDD, the **eligible for a treatment** Z_i is the same as **getting a treatment** D_i
  - $Z_i = D_i$
- In fuzzy RDD, the **eligible for a treatment** Z_i does NOT represent the **getting a treatment** D_i
  - $Z_i \neq D_i$

Identification

Use Passing Cutoff as an IV

- The discontinuity in outcome is actually the **average causal effect** of **treatment eligibility** $Z_i = 1(A_i \geq c)$ at cutoff c

$$\lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c - \varepsilon] \tag{1}$$

- To recover the causal effect of **receiving treatment** D_i
- Divide (2) by the jump in the treatment probability at cutoff

$$\lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c - \varepsilon]$$

Fuzzy RDD is IV

- The the average causal effect of **receiving treatment** defined in fuzzy RDD :

$$\alpha_{FRD} = \frac{\lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c - \varepsilon]}{\lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c - \varepsilon]}$$

- This is a **Wald estimate** at cutoff c
- So we can consider fuzzy RDD as an IV estimate
- Use treatment eligibility $Z_i = 1(A_i \geq c)$ as an instrument for treatment received D_i
    - Use an indicator for the test score above threshold as an instrument for attending NTU

Fuzzy RDD is IV

Assumptions

- **First-Stage Relationship**: $Z_i = 1(A_i \geq c)$ affects treatment probability
- **Local Independent Assumption**: In a neighborhood of cutoff c

$$(\mathrm{Y}_i^1, \mathrm{Y}_i^0, \mathrm{D}_i^1, \mathrm{D}_i^0) \perp\!\!\!\perp Z_i$$

  - In a neighborhood of cutoff c, the assignment to treatment is random

Fuzzy RDD is IV

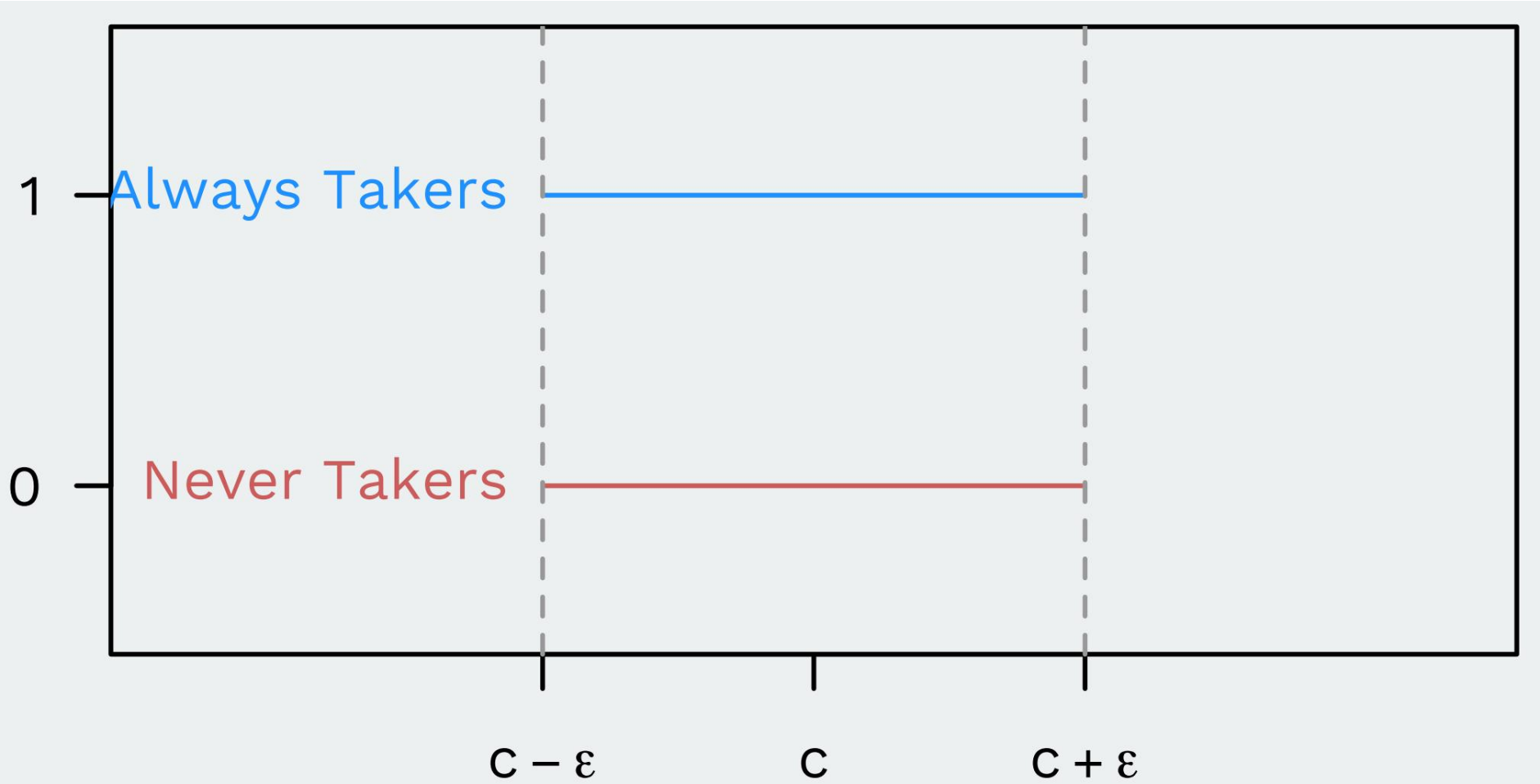
Assumptions

- **Exclusion Restriction**: $Z_i = 1(A_i \geq c)$ affects outcome Y_i only through changing treatment status D_i
- **Monotonicity Assumption**: $\mathrm{D}_i^1 \geq \mathrm{D}_i^0$
  - No one is discouraged from taking the treatment by crossing the threshold

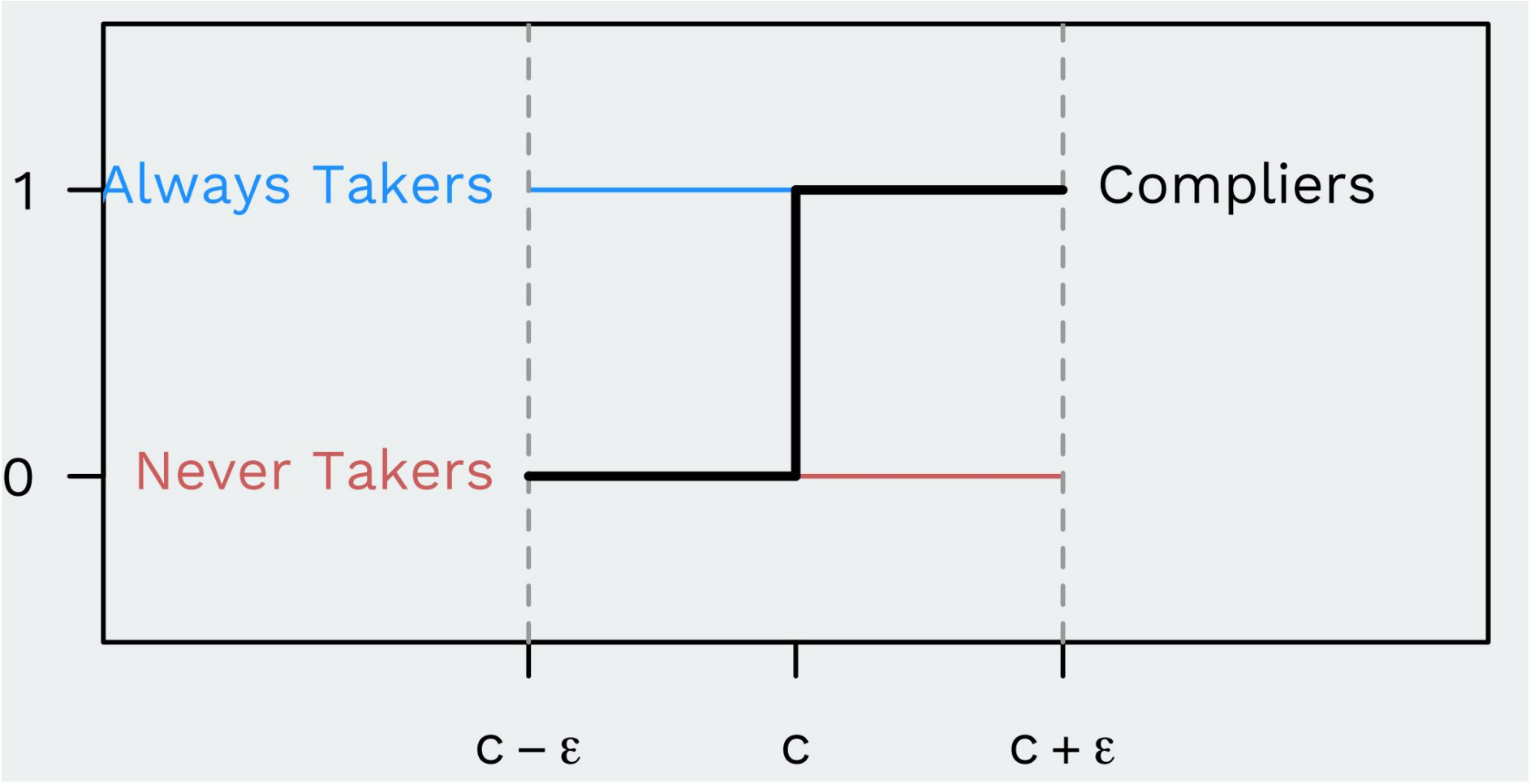
Fuzzy RDD and Compliers

- We can define four types of individuals based on whether they fellow the treatment assignment:
  - **Compliers**: $\mathrm{D}_i^1 > \mathrm{D}_i^0$ ($\mathrm{D}_i^0 = 0$ and $\mathrm{D}_i^1 = 1$)
    - David's test score is above NTU cutoff and enrolled in NTU
    - David's test score is below NTU cutoff and did not enroll in NTU
  - **Always-takers**: $\mathrm{D}_i^1 = \mathrm{D}_i^0 = 1$
    - John always can enroll in NTU (whether or not his test score is above NTU cutoff)
  - **Never-takers**: $\mathrm{D}_i^1 = \mathrm{D}_i^0 = 0$
    - Hank never enroll in NTU (whether or not his test score is above NTU cutoff)
  - **Defiers**: $\mathrm{D}_i^1 < \mathrm{D}_i^0$ ($\mathrm{D}_i^0 = 1$ and $\mathrm{D}_i^1 = 0$)
    - Jimmy's test score is above NTU cutoff and did NOT enroll in NTU
    - Jimmy's test score is below NTU cutoff but enrolled in NTU

Fuzzy RDD and Compliers



Fuzzy RDD and Compliers



Identification Results for Fuzzy RDD

Fuzzy RDD Identify LATE at cutoff c

$$\begin{aligned}\alpha_{FRD} &= \frac{\lim_{\varepsilon\to 0}\mathrm{E}[Y_i|A_i = x+\varepsilon] - \lim_{\varepsilon\to 0}\mathrm{E}[Y_i|A_i = x-\varepsilon]}{\lim_{\varepsilon\to 0}\mathrm{E}[D_i|A_i = x+\varepsilon] - \lim_{\varepsilon\to 0}\mathrm{E}[D_i|A_i = x-\varepsilon]} \\ &= \mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0|\mathrm{D}_i^1 > \mathrm{D}_i^0, A_i = c]\end{aligned}$$

- The estimate in fuzzy RDD represents the **causal effect for compliers** (local average treatment effect, LATE) at cutoff c
- **Compliers** are those who receive the treatment when they follow treatment eligibility rule ($A_i \geq c$), but would not otherwise receive it ($A_i < c$)

Examine Validity of Identification Assumptions

Test Internal Validity of RDD

Examine "Discontinuity" in Nonoutcome Variables

1. Examine "Discontinuity" in Nonoutcome Variables

- Construct a similar graph to the one before but using a covariate as the "outcome"
- There should be **NO jump in other covariates**
- If the covariates would jump at the cutoff one would doubt the identifying assumption

Test Internal Validity of RDD

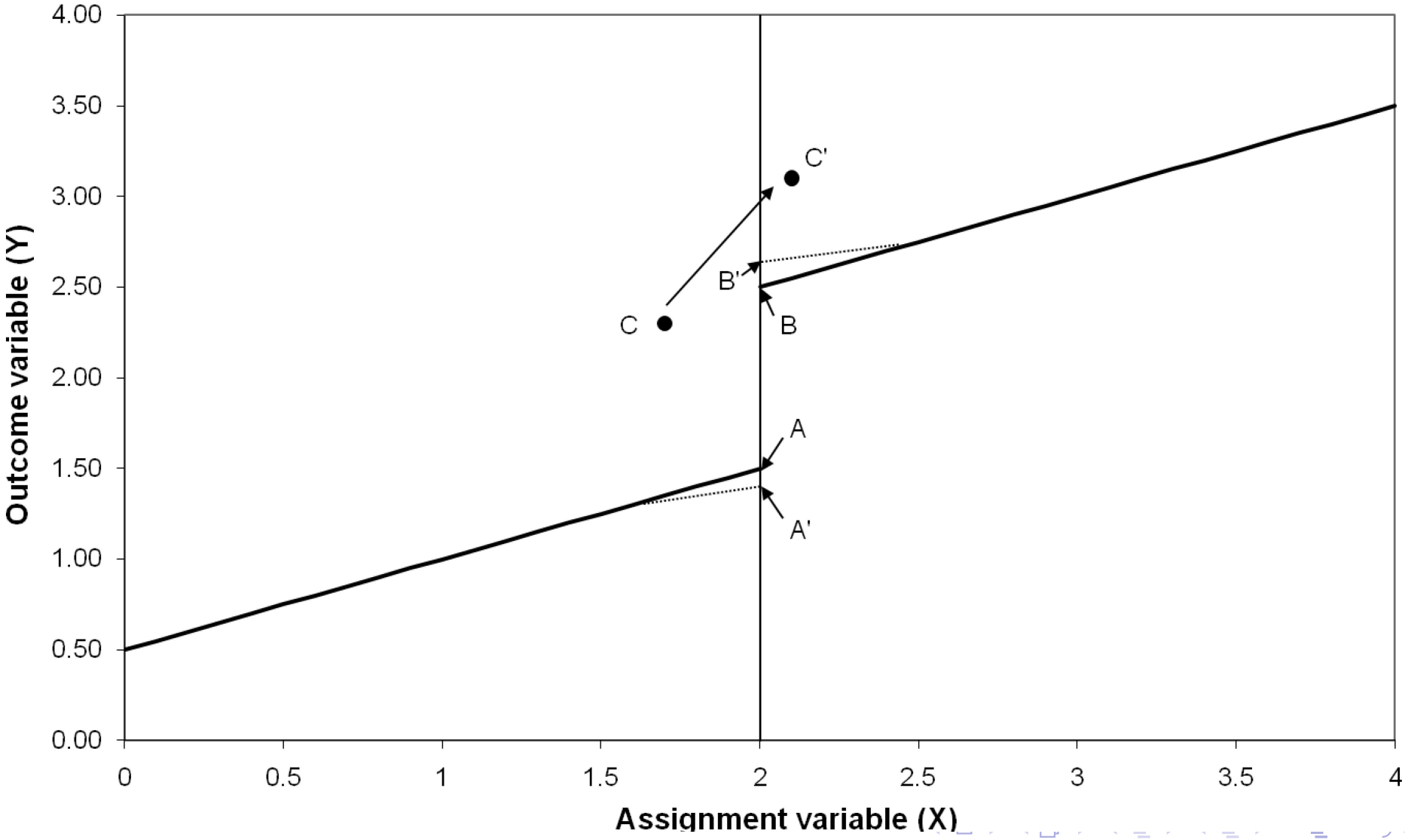
Sorting Behavior

2. Examine "Discontinuity" in Density of the assignment variable

- Individuals may invalidate the **continuity assumption** if they strategically manipulate assignment variable A to be just above or below the cutoff
- That is, people just above and just below the cutoff are no longer comparable

Consequence of Sorting Behavior

Example



Sorting Behavior

Example

- This is a concern especially if the exact value of the cutoff is known to the individuals in advance
  - Such sorting behavior may create a **discontinuity in the distribution of A at the cutoff**
  - That is, "bunching" to the right or to the left of the cutoff

Sorting Behavior

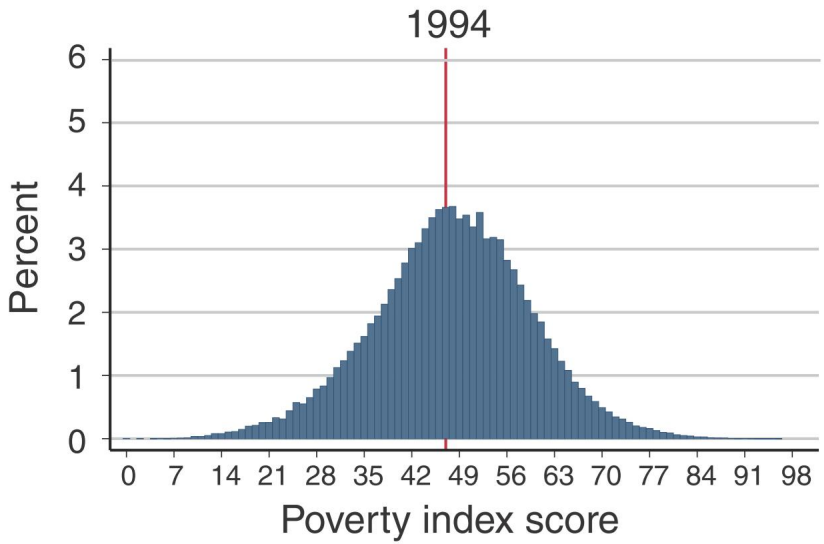
Example

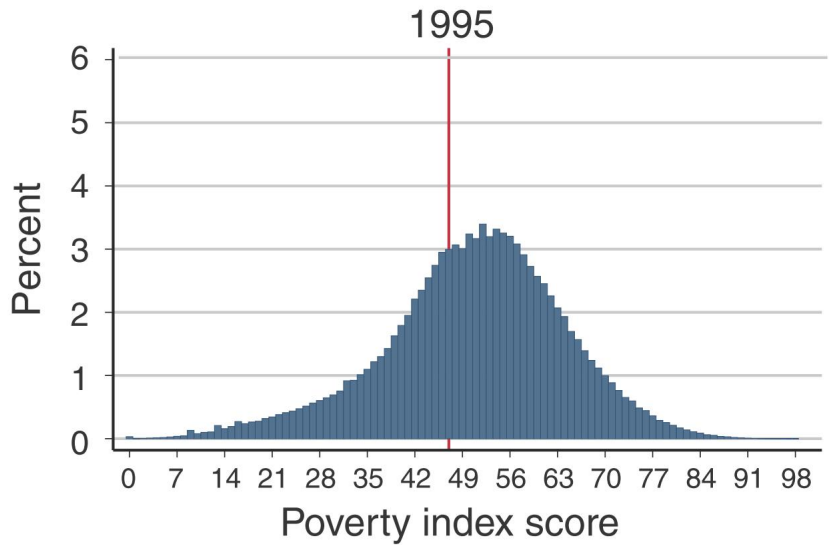
Adriana Camacho and Emily Conover (2011) "**Manipulation of Social Program Eligibility**" AEJ: Economic Policy

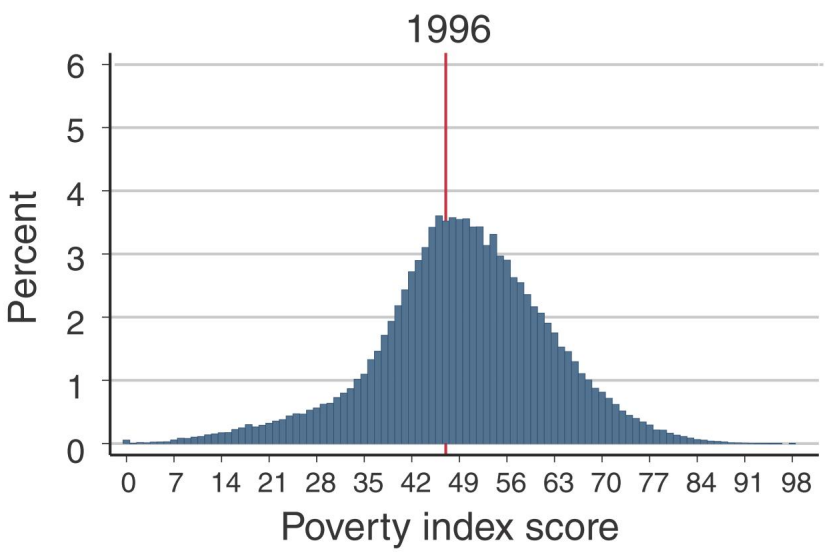
- Manipulation of a poverty index in Colombia
  - A poverty index is used to decide eligibility for social programs
- The algorithm to create the poverty index **becomes public** during the second half of 1997

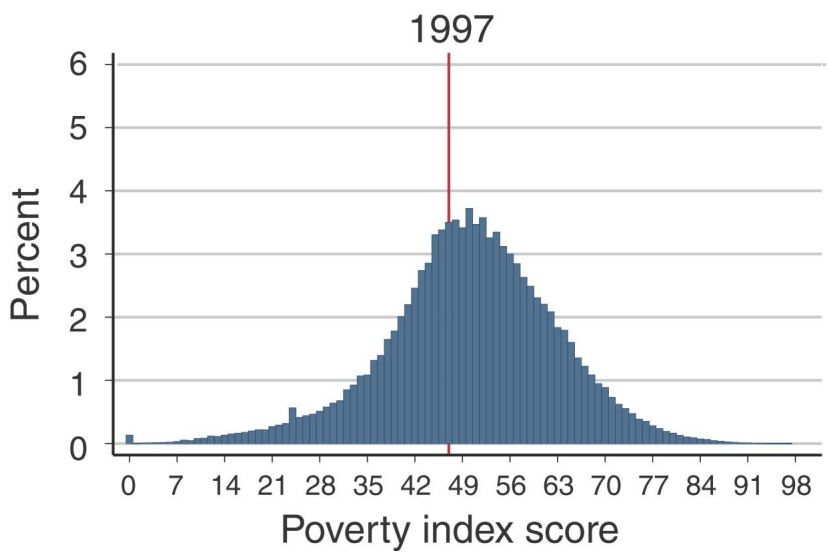
Sorting Behavior

Example



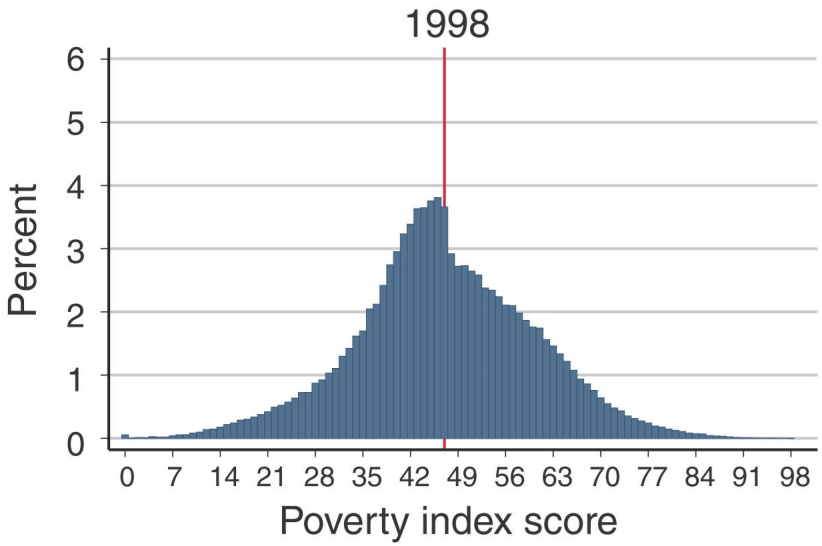


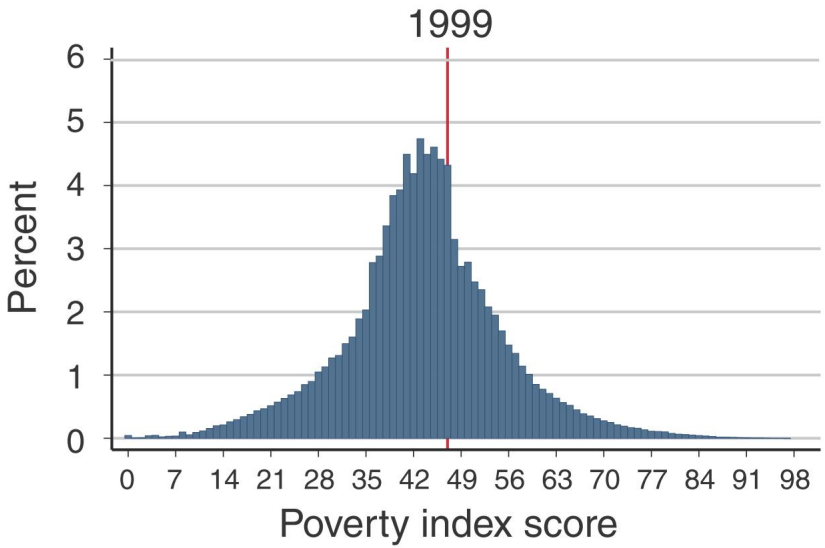


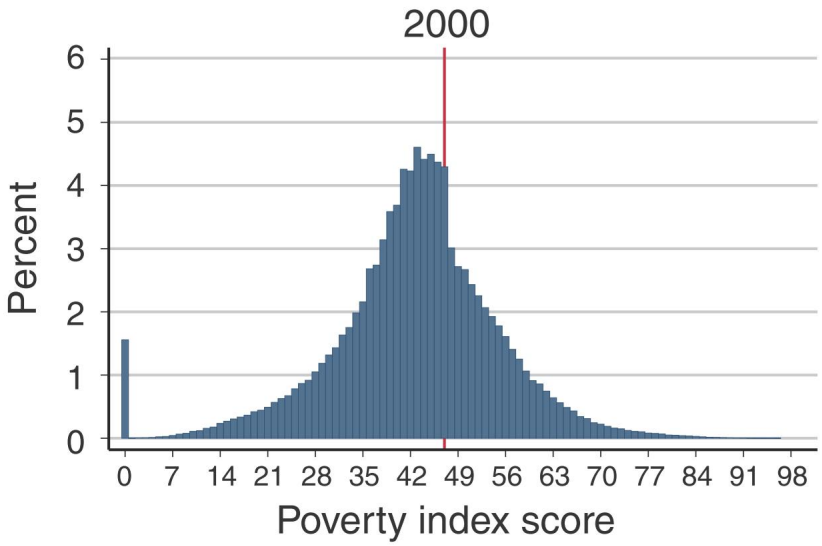


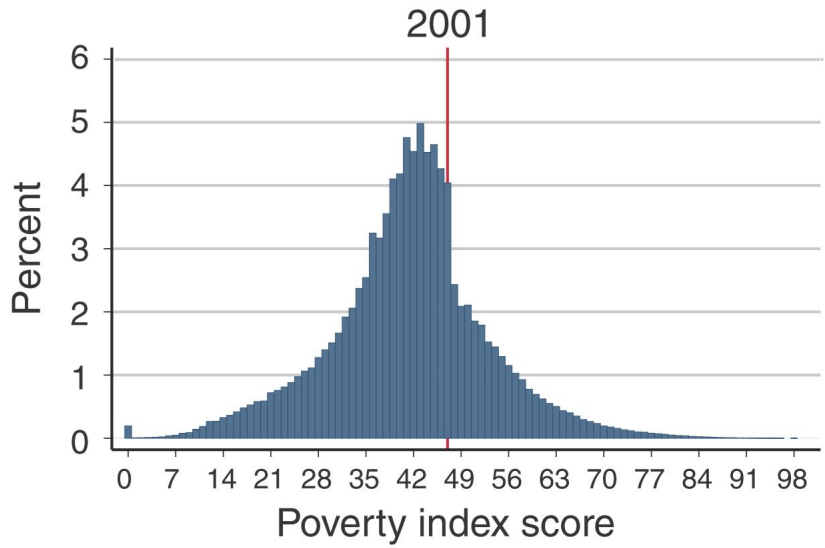
Sorting Behavior

Example









Test Internal Validity of RDD

Examine Discontinuity in Density of the Assignment Variable

How to Examine "Discontinuity" in Density of the assignment variable

- Plot the number of observations in each bin of assignment variable
- Investigate whether there is a **discontinuity in the distribution of the assignment variable** at the threshold
  - A discontinuity in the density suggests that people might manipulate the assignment variable around the threshold

Test Internal Validity of RDD

Examine Density of the assignment variable

- A formal test is provided by McCrary (2008)
  1. Partition the assignment variable into bins and calculate frequencies (number of observations) in each bins
  2. Calculate frequencies (number of observations) in each bin
  3. Ensure that no bin overlaps the cutoff
  4. **Run two local linear regressions**, one to the right and one to the left of the cutoff
  5. In these regressions, the bin midpoints are the regressors and **number of observations is outcome**
  6. Test whether log difference of the intercepts of the two regressions is statistically different from zero

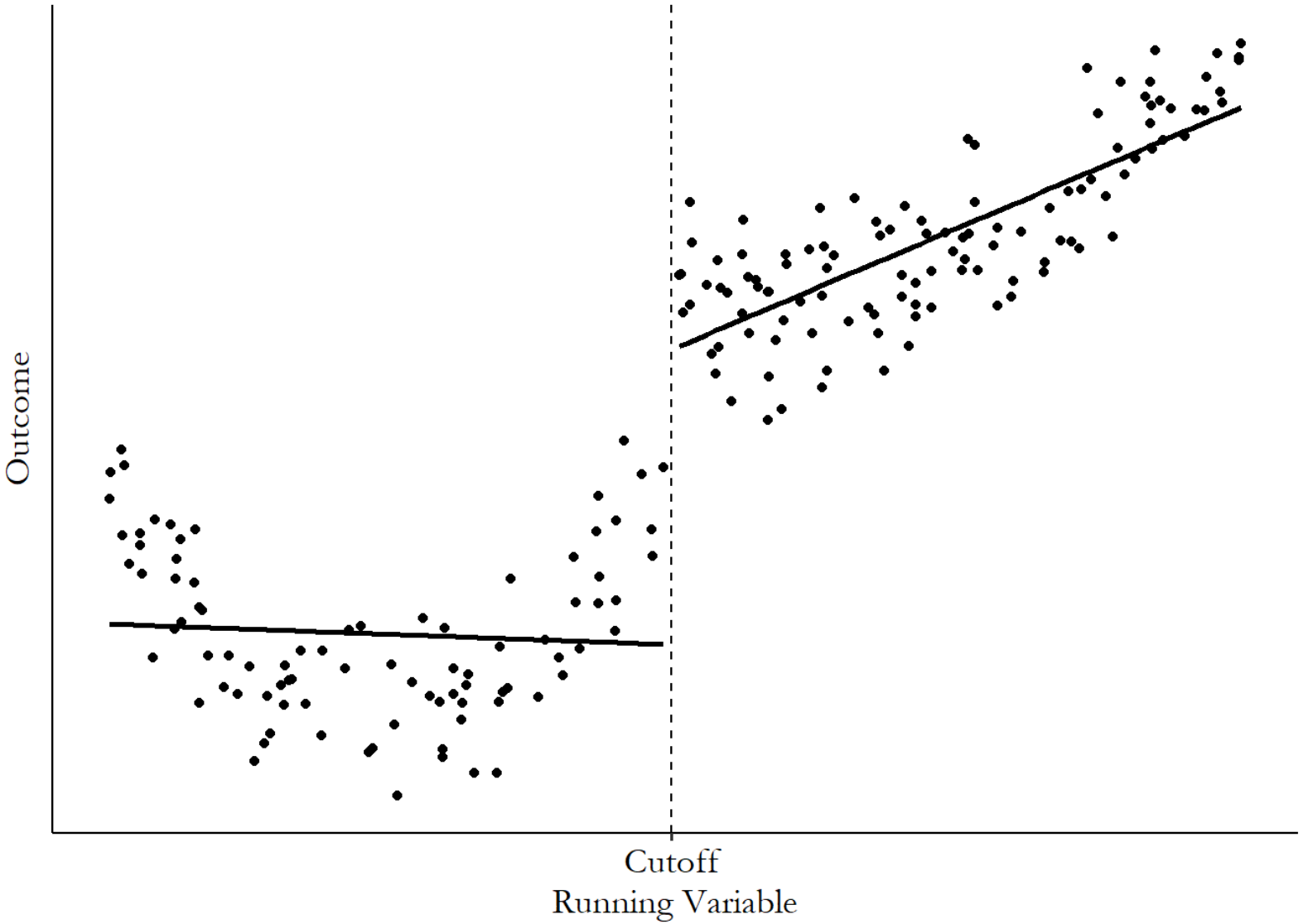
Estimation

RDD Estimation

Overview

- There are two strategies for getting RD estimates:
  1. Parametric/global method:
     - Use all available observations
     - Estimate treatment effects based on a **specific functional form** for the outcome and assignment variable relationship
  2. Nonparametric/local method:
     - Use the observations around cutoff
     - Compare the outcome of treated and untreated observations that lie **within specific bandwidth**

Estimate Discontinuity in Outcome



Source: Nick Huntington-Klein, The Effect: An Introduction to Research Design and Causality, Chapter 20

Parametric/Global Approach

- To **estimate the discontinuity at cutoff**, we need to model the relationship between assignment variable A and outcome Y
- Suppose that potential outcomes can be described by some reasonably smooth function $\mathrm{f}(A_i)$:

$$E[\mathrm{Y}_i^0|A_i] = \beta + \mathrm{f}(A_i)$$
$$\mathrm{Y}_i^1 = \mathrm{Y}_i^0 + \alpha$$

- We can get RD estimates by fitting:

$$Y_i = \beta + \alpha D_i + \mathrm{f}(A_i) + \eta_i$$

Parametric/Global Approach

- Assume $\mathrm{f}(A_i)$ is a linear function of A_i
- We usually do the following two settings:
  1. Allow the A_i terms to differ on both sides of the threshold
     - Include A_i both individually and interacting them with D_i
     - By doing this, we can estimate $\mathrm{f}(A_i)$ on each side
  2. Re-center A_i at c:
     - This step ensures that the treatment effect at $A_i = c$ is the coefficient on D_i in a regression model

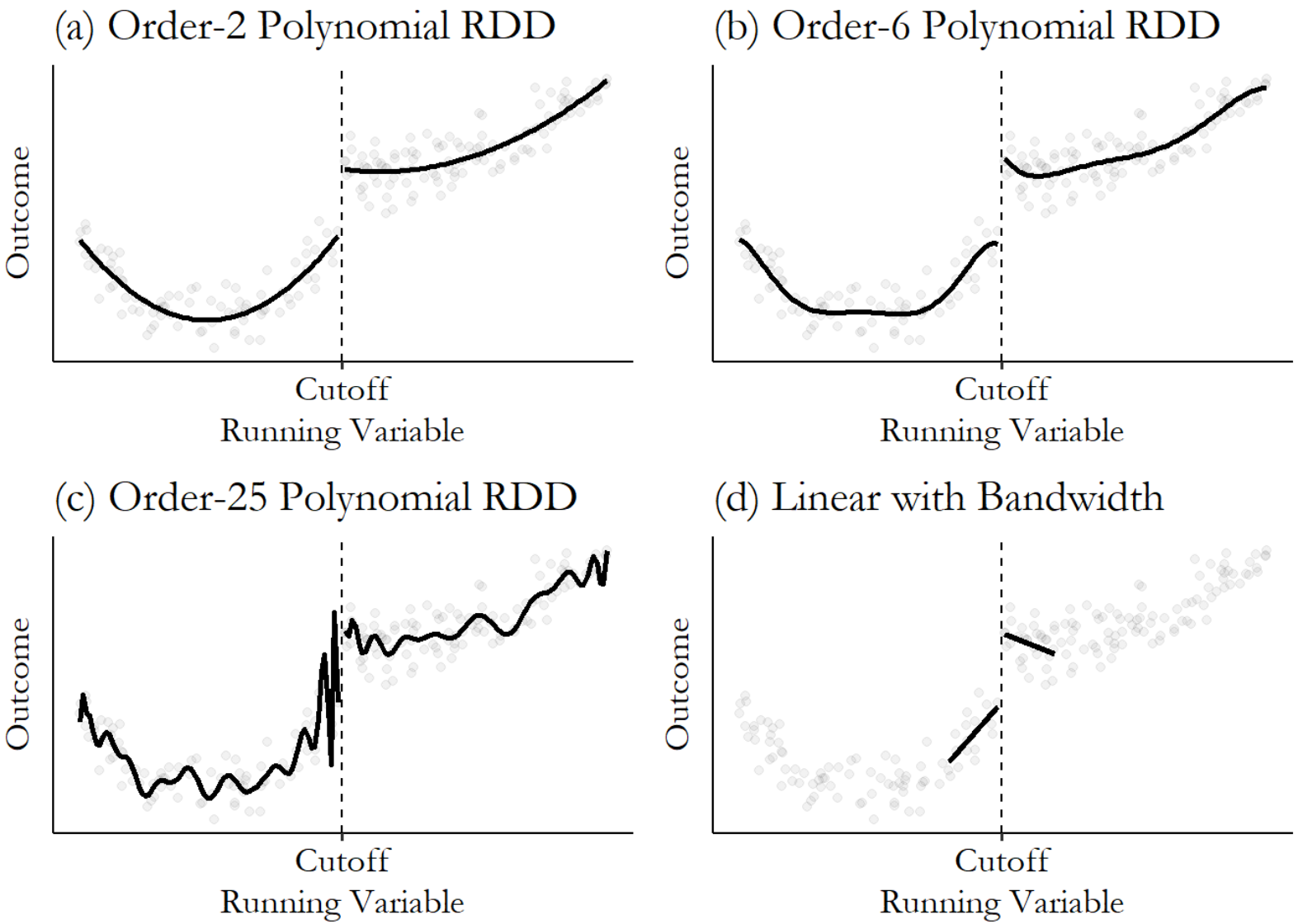
Parametric/Global Approach

- Therefore, we estimate the following regression model:

$$Y_i = \beta + \alpha D_i + \gamma_1(A_i - c) + \gamma_2(A_i - c) \times D_i + \eta_i$$

  - The intercept at untreated side around the cutoff: β
  - The intercept at treated side around the cutoff: $\beta + \alpha$
  - The discontinuity in outcome at cutoff: $(\beta + \alpha)$ - $\beta = \alpha$
    - $\alpha = \lim_{\varepsilon \to 0} \mathrm{E}[\mathrm{Y}_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[\mathrm{Y}_i | A_i = c - \varepsilon]$

More Flexible Function



Source: Nick Huntington-Klein, The Effect: An Introduction to Research Design and Causality, Chapter 20

Parametric/Global Approach

- We can use more flexible function $\mathrm{f}(A_i)$ to capture the relationship between assignment variable A and outcome Y
  - For example, use second-order polynomial of A
  - However, more flexible function $\mathrm{f}(A_i)$ might not give better estimate of discontinuity
- Gelman and Imbens (2019):
  - It's not a great idea to go above the second-order term when performing RDD
  - If there's a complex shape that needs fitting, instead try limiting the range of the data with a bandwidth and use a simpler function

Parametric/Global Approach

Fuzzy RDD Extension

- In **sharp RDD**: estimate one regression directly

$$Y_i = \beta + \alpha D_i + \gamma_1(A_i - c) + \gamma_2 D_i(A_i - c) + \eta_i$$

- In **fuzzy RDD**: $D_i \neq Z_i$, so we need two regressions
  - **First Stage** (effect of crossing cutoff on treatment):

$$D_i = \alpha_1 + \rho_1 Z_i + \gamma_1(A_i - c) + \delta_1 Z_i(A_i - c) + \upsilon_i$$

  - **Reduced Form** (intent-to-treat, effect on outcome):

$$Y_i = \alpha_2 + \rho_2 Z_i + \gamma_2(A_i - c) + \delta_2 Z_i(A_i - c) + \eta_i$$

  - **LATE** via 2SLS: instrument D_i with Z_i

$$\widehat{\mathrm{LATE}} = \frac{\hat{\rho}_2}{\hat{\rho}_1} = \mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 \mid \mathrm{D}_i^1 > \mathrm{D}_i^0, A_i = c]$$

Fuzzy RDD Estimation

2SLS — How It Works

- **Two-Stage Least Squares (2SLS)** removes the part of D_i correlated with unobservables
- **Stage 1** — regress treatment on the instrument and controls:

$$D_i = \alpha_1 + \rho_1 Z_i + \gamma_1(A_i - c) + \delta_1 Z_i(A_i - c) + \upsilon_i$$

  Obtain fitted values $\hat{D}_i = \hat{\alpha}_1 + \hat{\rho}_1 Z_i + \hat{\gamma}_1(A_i - c) + \hat{\delta}_1 Z_i(A_i - c)$
- **Stage 2** — regress outcome on *predicted* treatment:

$$Y_i = \alpha_2 + \rho_2 \hat{D}_i + \gamma_2(A_i - c) + \delta_2 Z_i(A_i - c) + \eta_i$$

  $\hat{\rho}_2$ is the **LATE** estimate
- **Intuition**: $\hat{D}_i$ retains only the variation in treatment driven by the cutoff Z_i (exogenous) and discards the part driven by unobservables (e.g. ability)
  - OLS uses all variation in D_i, including the confounded part $\Rightarrow$ biased
  - 2SLS uses only the "as-good-as-random" cutoff-driven variation $\Rightarrow$ LATE
- **Note**: in practice use `ivregress 2sls` / `ivreg()` — running two

Sharp RDD Estimation

Nonparametric/Local Approach

- The core idea of RDD is to compare outcomes just above and just below the cutoff c.
- Nonparametric/local method:
  - We do **not** assume a specific global functional form for the relationship between the outcome Y and the assignment variable A over their entire range.
    - This is what "nonparametric" means in this context.
  - Instead, we focus the analysis **locally**, using only observations close to the cutoff c.
  - We compare treated and untreated observations that fall within a specific **bandwidth (h)** around the cutoff.

Sharp RDD Estimation

Nonparametric/Local Approach

- Key insight: within a **sufficiently small local window (bandwidth h)** around the cutoff c, even if the true relationship is complex globally.
  - It can often be well **approximated by a simple linear function**.
- Therefore, we use **Local linear regression**:
  - This involves fitting separate linear regressions on each side of the cutoff, but *only* using data within the bandwidth h.
  - Optionally, kernel functions can be used to give more weight to observations closer to the cutoff.

Sharp RDD Estimation

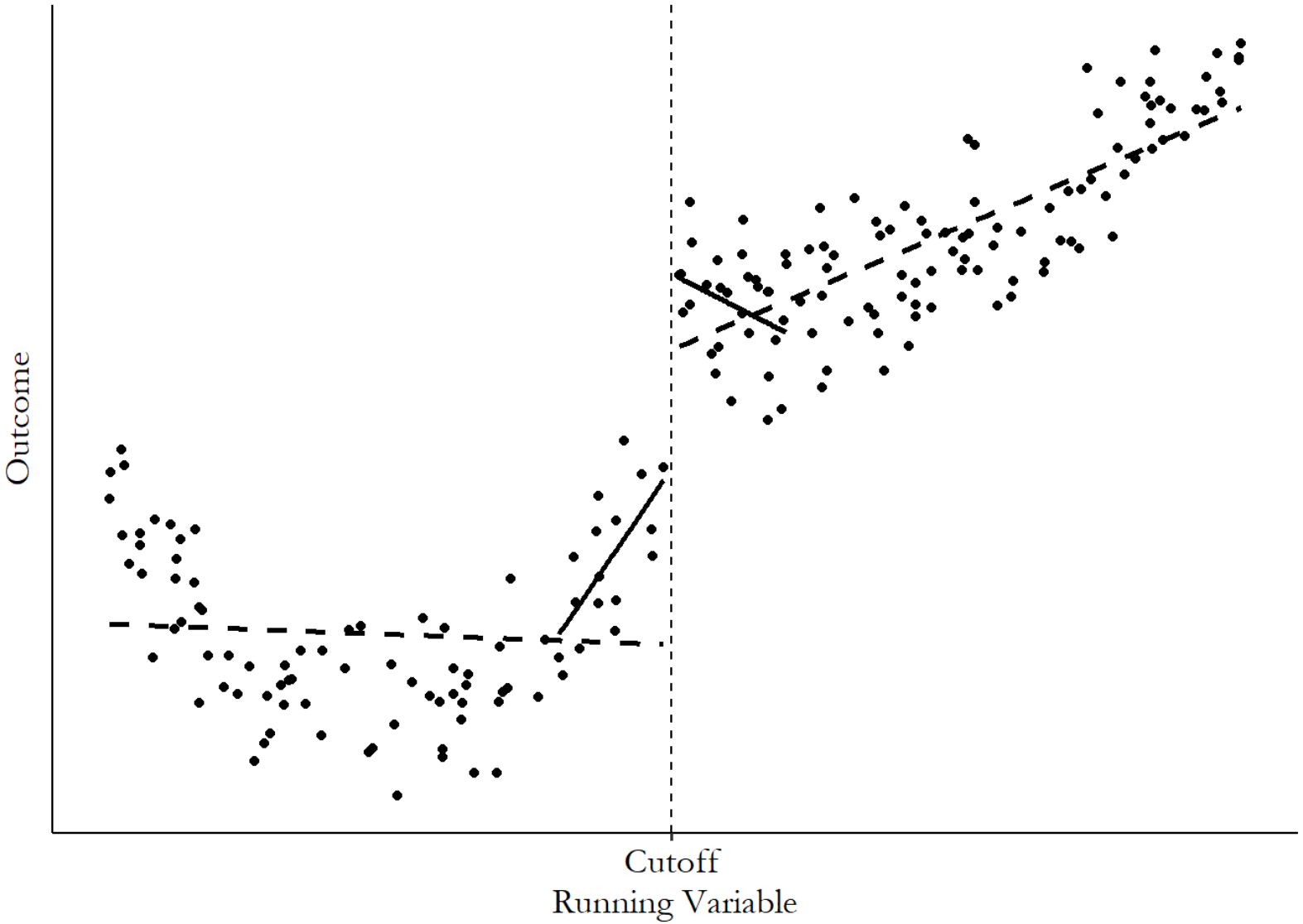
Implementing Local Linear Regression

- Specifically, we estimate the following regression model **using only observations within the chosen bandwidth** h:

$$Y_i = \beta + \alpha D_i + \gamma_1(A_i - c) + \gamma_2(A_i - c) \times D_i + \eta_i$$

- This linear model is just a **local approximation** of the relationship near the cutoff c, within the bandwidth h.
  - It is *not* a global assumption about the functional form.
- The estimated coefficient $\hat{\alpha}$ provides the RDD estimate of the treatment effect.

Bandwidth



Source: Nick Huntington-Klein, The Effect: An Introduction to Research Design and Causality, Chapter 20

Sharp RDD Estimation

Nonparametric/Local Approach

- The main challenge of nonparametric approach is to **choose a bandwidth**
- There is essentially a trade-off between **bias** and **precision** (efficiency)
- Use a larger bandwidth:
  - Since more data points are used in the regression
  - Get more **precise** treatment effect estimates
  - But use data points far from cutoff
  - The estimated treatment effect could be **biased**

Sharp RDD Estimation

How to Choose Bandwidth

1. **Cross-Validation (CV) Procedure**:
   - Aims to select h that minimizes prediction errors of the RDD model specifically in the neighborhood of the cutoff c.
2. **Plug-In Procedure**:
   - Derives a formula for the optimal bandwidth (h_{opt}) by analytically minimizing an (Asymptotic) Mean Squared Error for the RDD treatment effect estimator (e.g., $\hat{\alpha}$)
     - The h_{opt} formula depends on some unknown data characteristics at the cutoff c (like local trends and data variability).
     - These must first be estimated from your data, often using a "pilot" bandwidth.
   - See, e.g., Imbens and Kalyanaraman (2012), Calonico, Cattaneo, and Titiunik (2014).

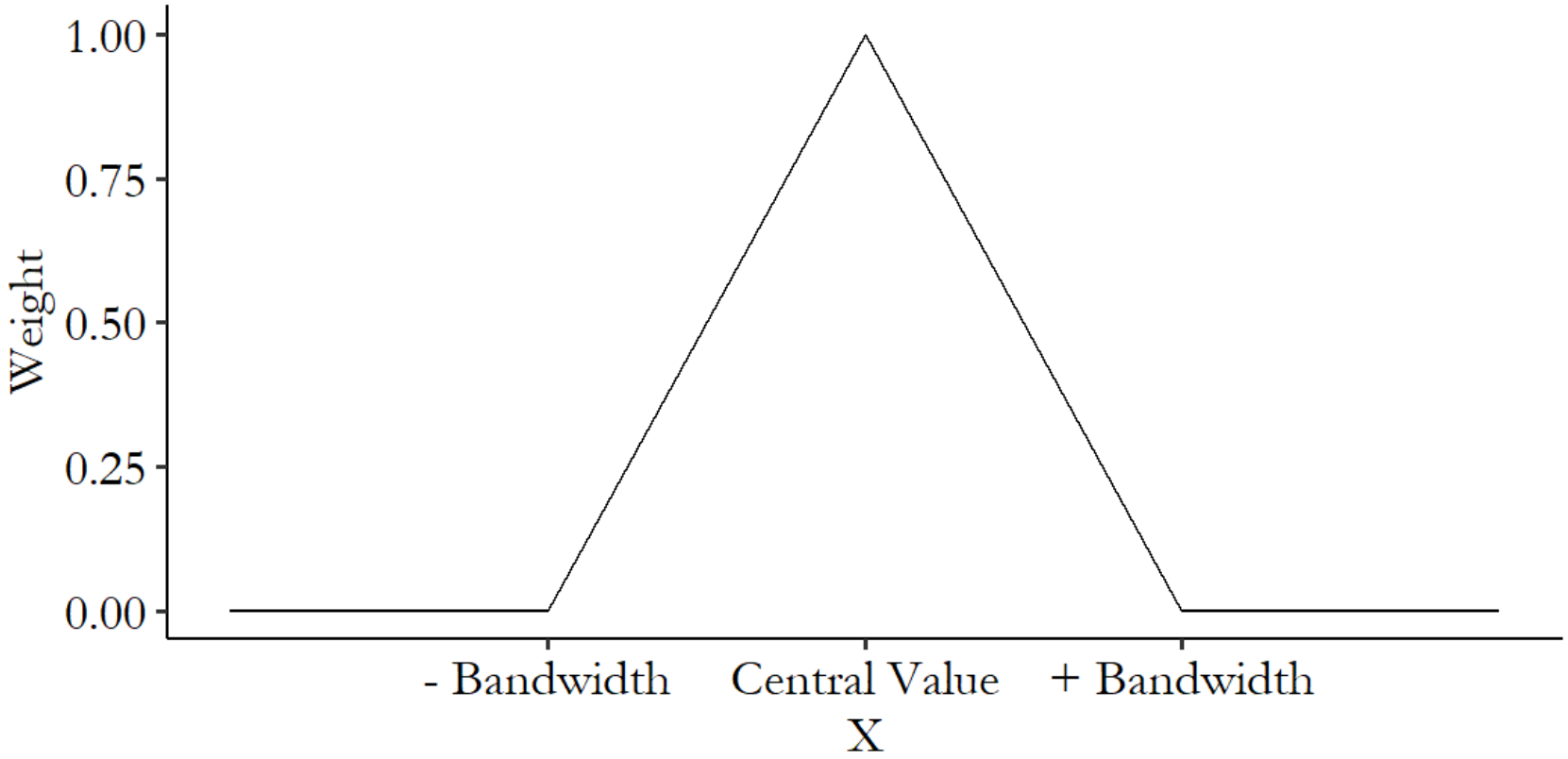
Sharp RDD Estimation

Nonparametric/Local Approach

- In practice, we also use a specific kernel function to weight observations more heavily around cutoff
- A very commonly-used kernel in regression discontinuity is the **triangular kernel**

$$K(A) = \begin{cases} 1 - \frac{|A_i - c|}{h} & \text{for } c - h < A_i < c + h, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Triangular Kernel Function



Source: Nick Huntington-Klein, The Effect: An Introduction to Research Design and Causality, Chapter 20

Fuzzy RDD Estimation

Nonparametric Approach — Fuzzy Extension

- In **sharp RDD**: one local linear regression within bandwidth h

$$Y_i = \beta + \alpha D_i + \gamma_1(A_i - c) + \gamma_2 D_i(A_i - c) + \eta_i, \quad |A_i - c| \leq h$$

  $\hat{\alpha}$ = RD treatment effect (ATE at cutoff)

- In **fuzzy RDD**: two local regressions within the *same* bandwidth h
  - **First Stage** (local): regress D_i on Z_i and $(A_i - c)$ within h
  - **Reduced Form** (local): regress Y_i on Z_i and $(A_i - c)$ within h
  - LATE $= \hat{\rho}_2^{RF} / \hat{\rho}_1^{FS}$ (Wald ratio at the cutoff)
- `rdrobust` with `fuzzy`(D_i) handles all this automatically:
  - Selects MSE-optimal bandwidth; uses triangular kernel
  - Provides bias-corrected, robust confidence intervals (Calonico et al. 2014)

Empirical Example: Stata/R

Bleemer & Mehta (2022)

Motivation

Zachary Bleemer and Aashish Mehta (2022) "**Will Studying Economics Make You Rich? A Regression Discontinuity Analysis of the Returns to College Major**", American Economic Journal: Applied Economics

- Examine the causal effect of majoring in economics on early-career earnings
- Key Research Questions:
  - What is the wage return to studying economics?
  - How does access to economics major affect students' career paths?
  - Do observational wage differences reflect causal effects?

Empirical Example: Bleemer & Mehta (2022)

Motivation

- College graduates with economics degrees earn substantially higher wages
  - Median wage of \$90,000 for economics majors vs. \$65,000 for other social sciences
- However, estimating causal effects is challenging due to:
  - Students' nonrandom selection into majors
  - Universities' admissions and grade requirements
- Study exploits a GPA threshold policy at UC Santa Cruz that restricted access to the economics major
  - Students needed 2.8 GPA in Economics 1 & 2 to declare major

Empirical Example: Bleemer & Mehta (2022)

Identification Strategy

- **Fuzzy RDD**: Exploits GPA threshold (2.8) in Economics 1 and 2 for major access

$$\alpha_{FRD} = \frac{\lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[Y_i | A_i = c - \varepsilon]}{\lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c + \varepsilon] - \lim_{\varepsilon \to 0} \mathrm{E}[D_i | A_i = c - \varepsilon]}$$

- Where:
  - Y_i: Early-career wages (2017-2018)
  - A_i: Average GPA in Economics 1 & 2
  - D_i: Economics major indicator
  - c: GPA threshold (2.8)

Empirical Example: Bleemer & Mehta (2022)

Empirical Specification

- **First Stage**:

$$D_i = \alpha_1 + \rho_1 Z_i + \mathrm{f}_1(A_i) + \upsilon_i$$

- **Reduced Form**:

$$Y_i = \alpha_2 + \rho_2 Z_i + \mathrm{f}_2(A_i) + \eta_i$$

- Where:
  - D_i: Economics major indicator
  - $Z_i = 1(A_i \geq 2.8)$: Treatment eligibility
  - A_i: Average GPA in Economics 1 & 2
  - $\mathrm{f}_1(.), \mathrm{f}_2(.)$: Linear functions of GPA

- **Fuzzy RD Estimate**:

$$\alpha_{FRD} = \frac{\rho_2}{\rho_1} = \mathrm{E}[\mathrm{Y}_i^1 - \mathrm{Y}_i^0 | \mathrm{D}_i^1 > \mathrm{D}_i^0, A_i = 2.8]$$

Data and Program

STATA Implementation

- See `RDD_fuzzy.do`
- Use `RDD.dta` (calibrated to the paper)
- Install required packages:
  - `binscatter.ado`
  - `rdrobust.ado`
  - `rddensity.ado`
  - `DCdensity.ado`
- Key steps:
  1. Graphical analysis (first stage, reduced form, covariate balance)
  2. Sorting/manipulation test
  3. Preparation for estimation (centered running variable)
  4. Parametric estimation (first stage, reduced form, 2SLS)
  5. Nonparametric estimation (`rdrobust fuzzy()`)
  6. Robustness checks

Data and Program

R Implementation

- See `RDD_fuzzy.R`
- Use `RDD.dta` (calibrated to the paper)
- Install required packages:
  - `haven`, `dplyr` (data handling)
  - `rdrobust` (`rdrobust`, `rdplot`)
  - `rddensity` (manipulation test)
  - `AER` (`ivreg` for 2SLS)
- Key steps: same as STATA
  1. Graphical analysis (`rdplot`)
  2. Sorting test (`rddensity`)
  3. Preparation for estimation (`mutate`)
  4. Parametric estimation (`lm` + `ivreg`)
  5. Nonparametric estimation (`rdrobust` with `fuzzy=`)
  6. Robustness checks

Step 1: Graphical Analysis

First Stage

- In fuzzy RDD, **always** produce two RD plots before any regression
- **First stage**: plot the **treatment** (D_i = economics major) against the **running variable** (A_i = GPA in Econ 1 & 2)
  - Construct bins, average the treatment within bins on each side of the cutoff
  - Inspect whether P(economics major) **jumps** at GPA = 2.8
  - The size of the jump is the **first-stage estimate** ($\hat{\rho}_1$)
  - Bleemer & Mehta (2022): students just above 2.8 GPA were **36 percentage points** more likely to major in economics

STATA: Step 1

First Stage

- Example (first stage graph):

```
use "RDD.dta", clear
binscatter econ_major gpa_econ, n(50) rd(2.8) linetype(lfit) /*
*/ xtitle("GPA in Introductory Economics")                    /*
*/ ytitle("P(Economics Major)")                               /*
*/ title("First Stage: GPA Cutoff and Economics Enrollment")
graph export figure/econ_fs.png, replace
```

- **econ_major**: treatment D_i on the y-axis
- **gpa_econ**: running variable A_i on the x-axis
- **rd(2.8)**: split sample at the GPA cutoff
- **linetype(lfit)**: fit a linear regression line on each side

R: Step 1

First Stage

```
library(haven); library(rdrobust); library(dplyr)
data   <- read_dta("RDD.dta")
cutoff <- 2.8
# First stage: P(Economics Major) by GPA
rdplot(y = data$econ_major, x = data$gpa_econ, c = cutoff,
       title   = "First Stage: GPA Cutoff and Economics Enrollment"
           ,
       x.label = "GPA in Introductory Economics",
       y.label = "P(Economics Major)")
```

Empirical Example: Bleemer & Mehta (2022)

First Stage

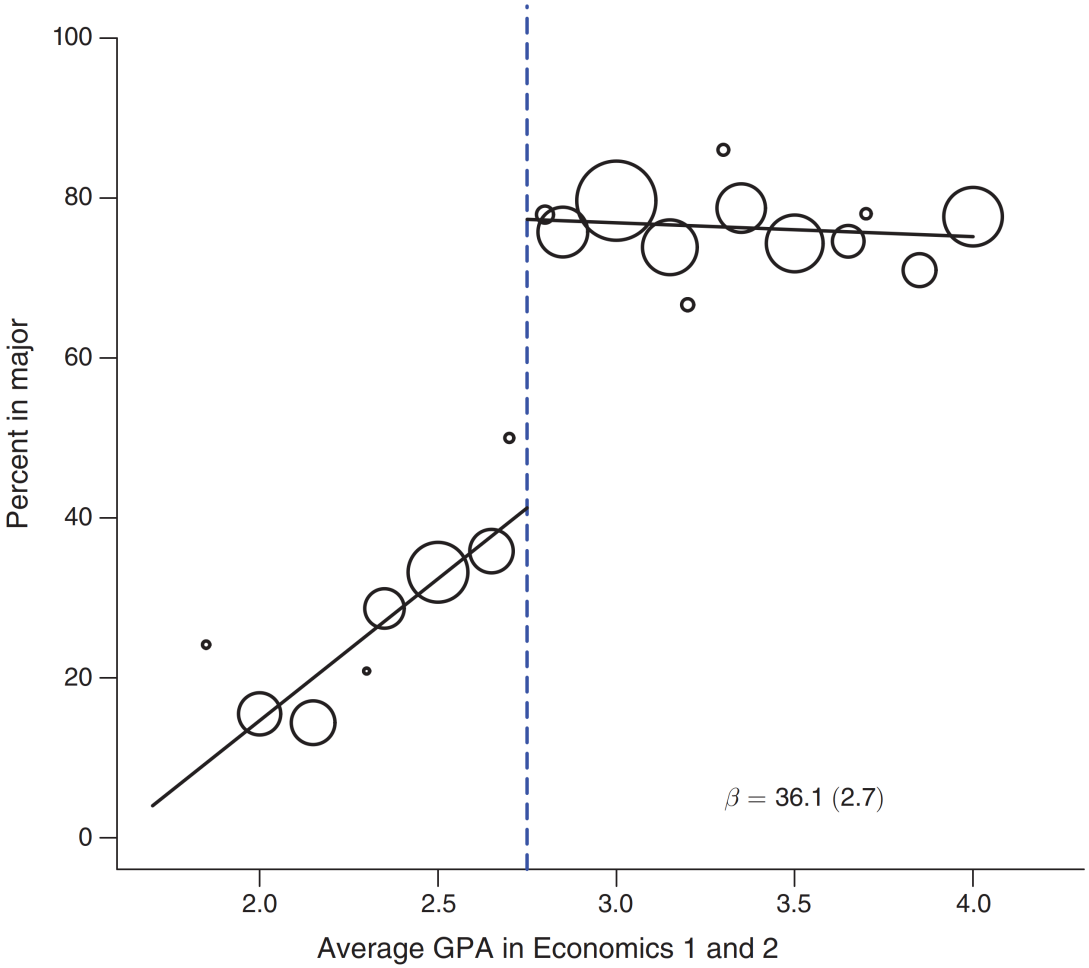


FIGURE 1. THE EFFECT OF THE UCSC ECONOMICS GPA THRESHOLD ON MAJORING IN ECONOMICS

Empirical Example: Bleemer & Mehta (2022)

First Stage

- Students just above 2.8 GPA threshold were:
  - 36 percentage points more likely to major in economics
  - Most would have otherwise majored in other social sciences

Step 1: Graphical Analysis

Reduced Form

- **Reduced form**: plot the **outcome** (Y_i = log wages) against the **running variable** (A_i = GPA)
  - Shows the **intent-to-treat (ITT)** effect: the total effect of *eligibility* (Z_i) on wages
  - The jump at GPA $= 2.8$ is the **reduced-form estimate** ($\hat{\rho}_2$)
  - LATE $= \hat{\rho}_2 / \hat{\rho}_1$ (Wald estimator)
  - Bleemer & Mehta (2022): wages rise $\approx 14\%$ at the cutoff (ITT)

STATA: Step 1

Reduced Form

- Example (reduced form graph):

```
binscatter logwage_1718 gpa_econ if employed_1718, /*
*/ n(50) rd(2.8) linetype(lfit)                    /*
*/ xtitle("GPA in Introductory Economics")         /*
*/ ytitle("Log Wages 2017-2018")                   /*
*/ title("Reduced Form: GPA Cutoff and Wages")
graph export figure/econ_rf.png, replace
```

- **if employed_1718**: restrict to employed workers (wages only observed for employed)
- A visible jump at GPA $= 2.8$ indicates that the GPA cutoff affects wages through the economics major pathway

R: Step 1

Reduced Form

```
data_emp <- data |> filter(employed_1718 == 1, !is.na(logwage_1718)
    )
# Reduced form: Log wages by GPA
rdplot(y = data_emp$logwage_1718, x = data_emp$gpa_econ, c = cutoff
    ,
       title   = "Reduced Form: GPA Cutoff and Log Wages",
       x.label = "GPA in Introductory Economics",
       y.label = "Log Wages (2017-2018)")
```

Empirical Example: Bleemer & Mehta (2022)

Second Stage

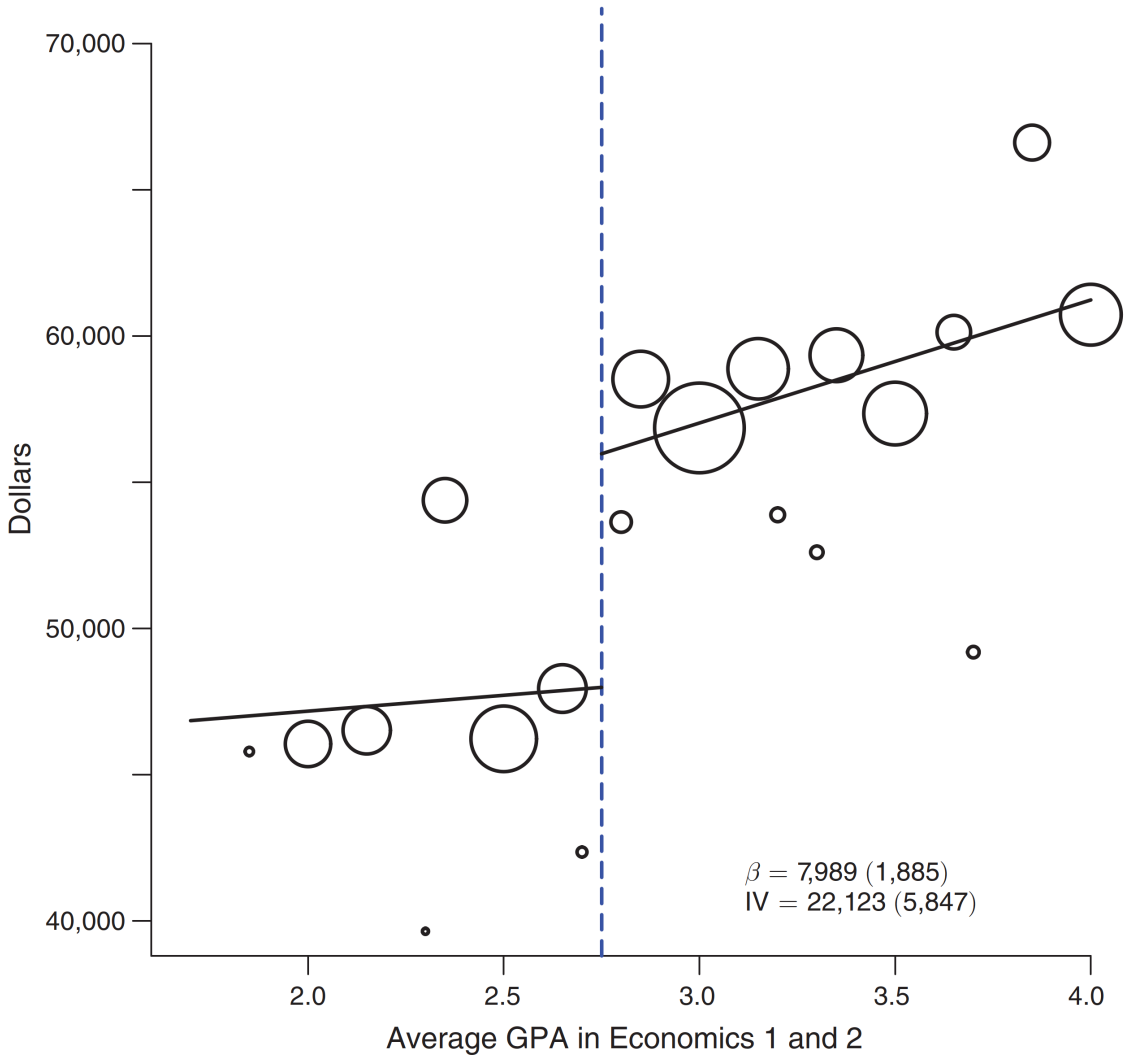


FIGURE 2. THE EFFECT OF THE UCSC ECONOMICS GPA THRESHOLD ON ANNUAL WAGES

Step 1: Graphical Analysis

Covariate Balance

- Also plot **pre-determined covariates** against the running variable
- These should show **no discontinuity** at GPA $= 2.8$
  - SAT score, gender, URM status, zip income were fixed *before* students knew their intro-econ GPA
  - A jump in any covariate would indicate that students above/below 2.8 differ systematically $\Rightarrow$ threat to continuity
- Bleemer & Mehta (2022): no significant jumps in any baseline characteristic

STATA: Step 1

Covariate Balance

- Example (SAT score and gender balance):

```
1 binscatter sat_score gpa_econ, n(50) rd(2.8) linetype(lfit) /*
2 */ title("Covariate Balance: SAT Score")
3 graph export figure/econ_sat.png, replace
4
5 binscatter female gpa_econ, n(50) rd(2.8) linetype(lfit)   /*
6 */ title("Covariate Balance: Gender")
7 graph export figure/econ_female.png, replace
```

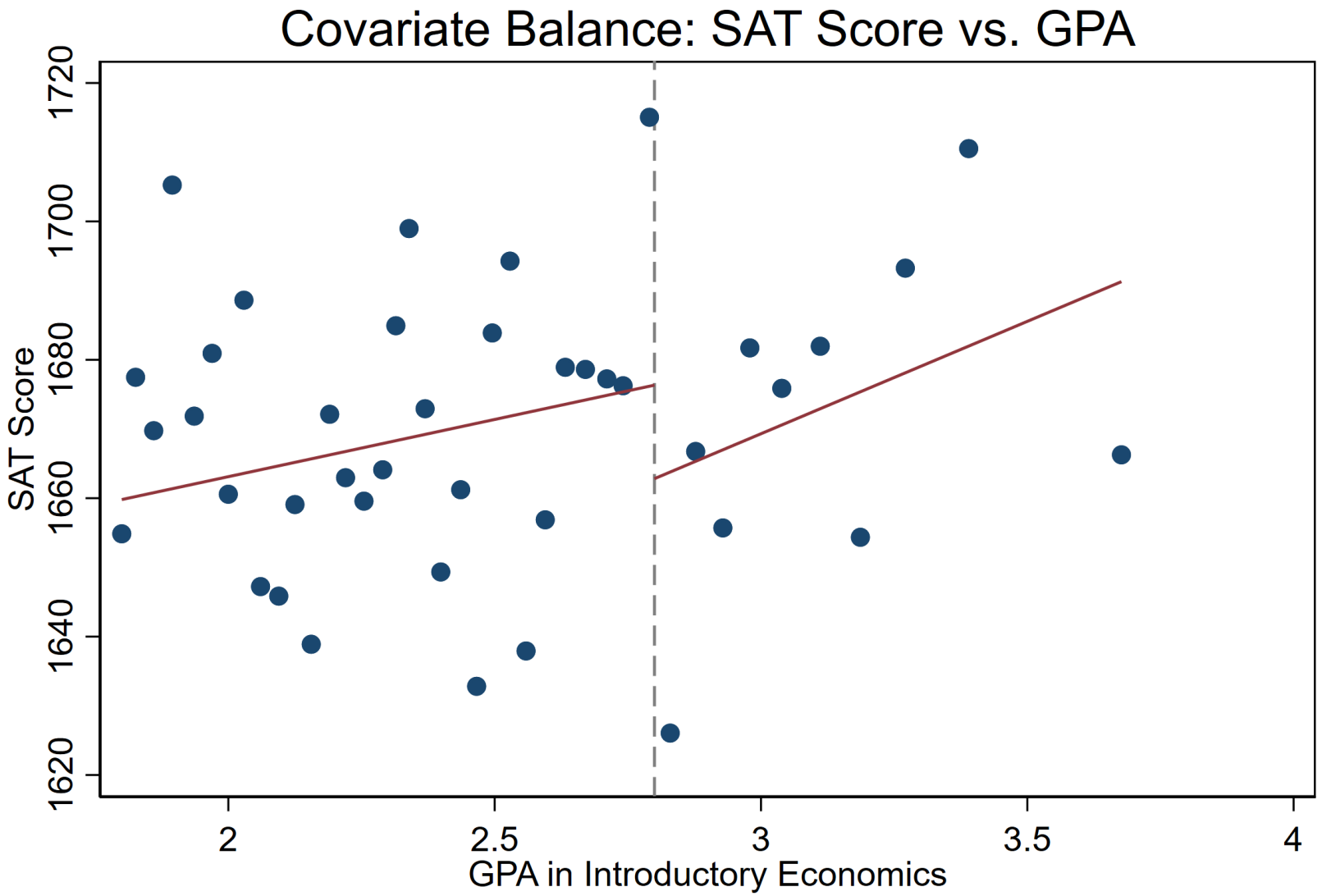
R: Step 1

Covariate Balance

```
rdplot(y = data$sat_score, x = data$gpa_econ, c = cutoff,
       title = "Covariate Balance: SAT Score")
rdplot(y = data$female, x = data$gpa_econ, c = cutoff,
       title = "Covariate Balance: Gender")
```

Covariate Balance

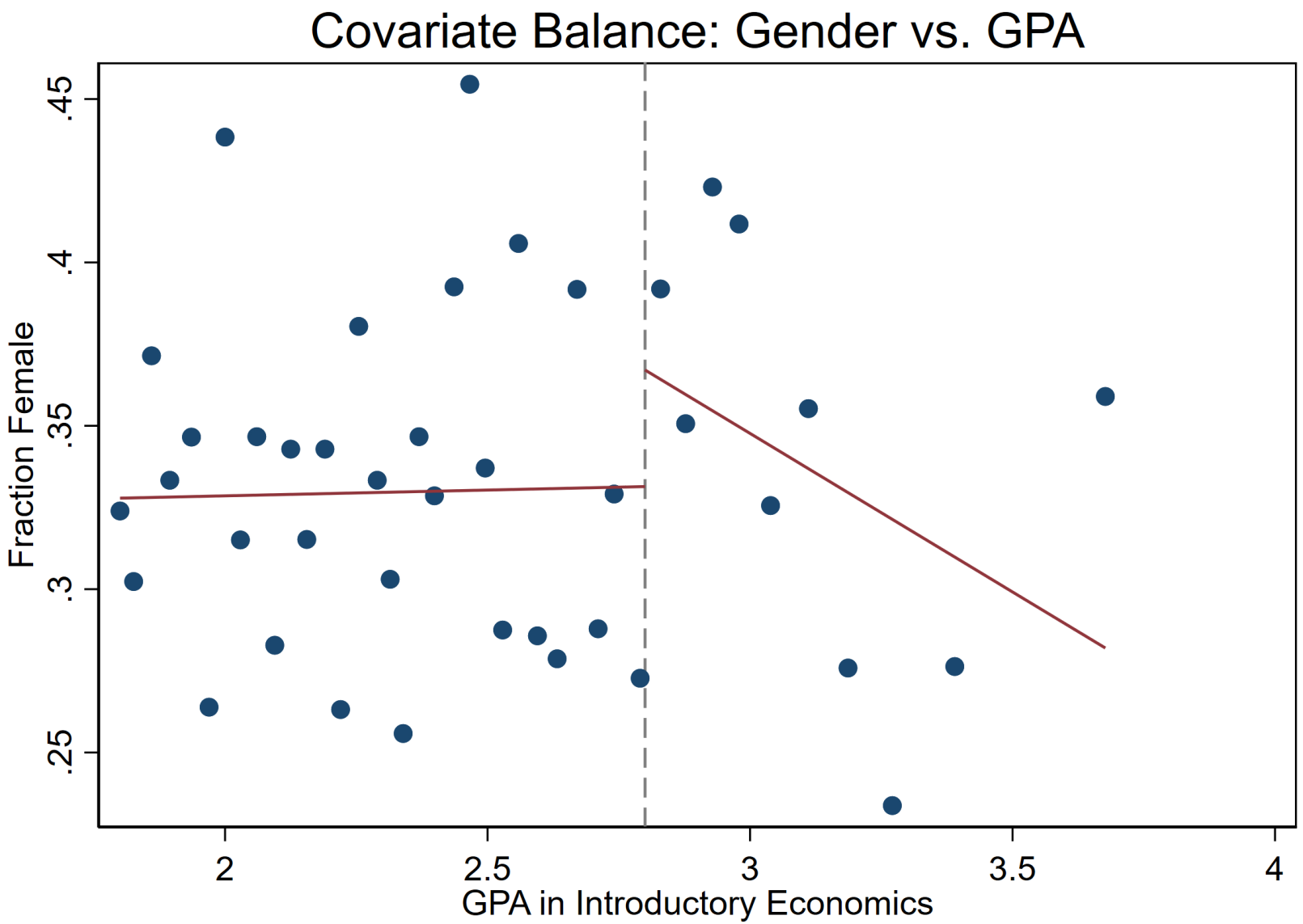
SAT Score by GPA in Introductory Economics



No discontinuity expected if RDD design is valid.

Covariate Balance

Female by GPA in Introductory Economics



Step 2: Test Sorting Behavior

- Plot the **density of the running variable** (GPA) around the cutoff
- If students can **manipulate** their GPA to just exceed 2.8 (e.g. by disputing grades or retaking the exam), we would see:
  - A spike in the density *just above* GPA $= 2.8$
  - A dip *just below*
- H_0: no density discontinuity at the cutoff
- Non-rejection ($p > 0.05$) supports the RDD validity assumption
- Bleemer & Mehta (2022): no evidence of manipulation at GPA $= 2.8$

STATA: Step 2

Sorting Test

- Syntax:

```
rddensity runvar [if] [in] [, options]
```

- Example:

```
* Option A: rddensity (Cattaneo, Jansson, Ma 2020)
rddensity gpa_econ, c(2.8) all plot
graph export figure/econ_density.png, replace

* Option B: McCrary (2008) DCdensity
DCdensity gpa_econ, breakpoint(2.8) generate(Xj Yj r0 fhat
    se_fhat)
drop Xj Yj r0 fhat se_fhat
```

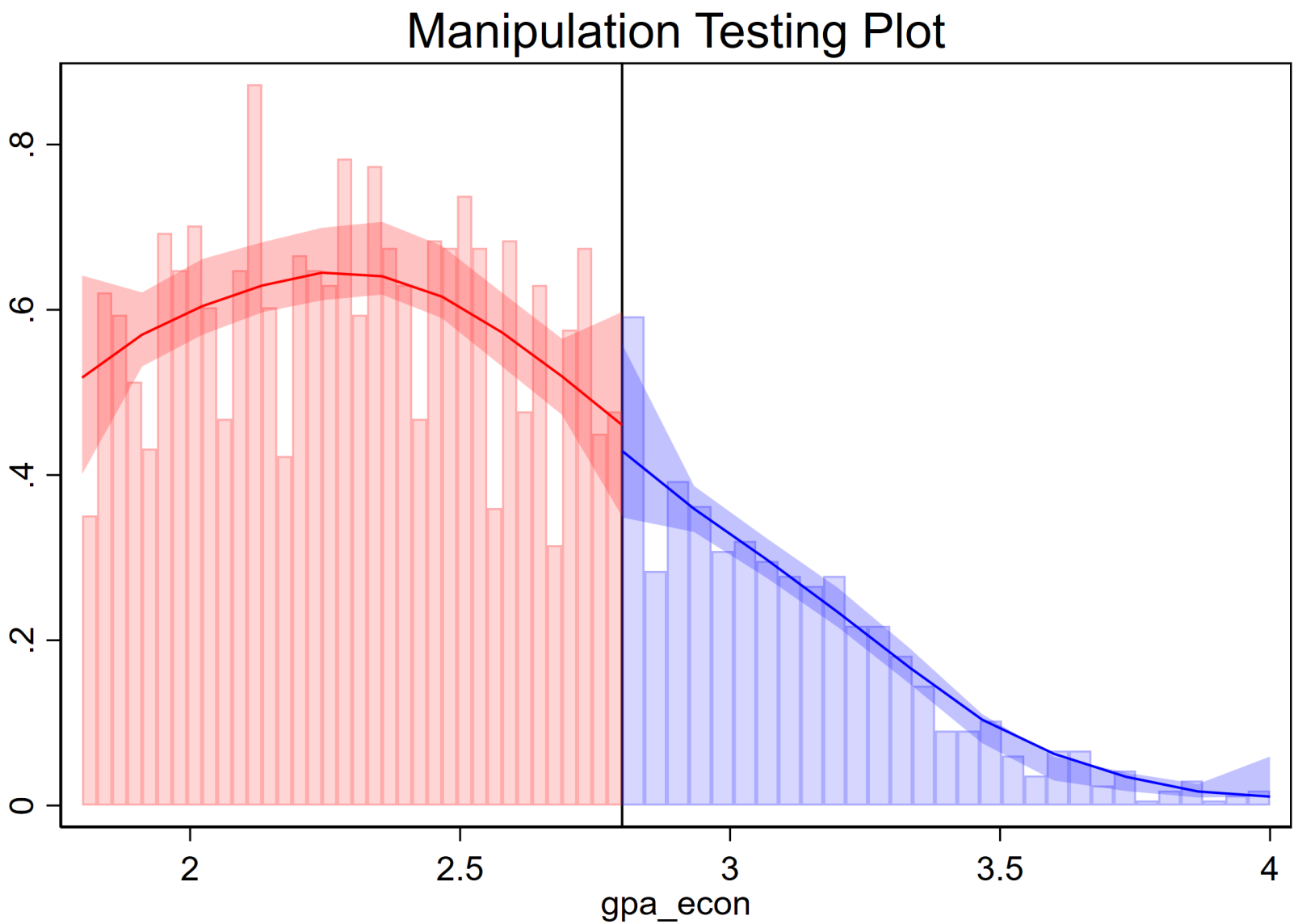
- **gpa_econ**: running variable whose density we test
- **c(2.8)**: cutoff value

R: Step 2

Sorting Test

```r
library(rddensity)
# H0: no density discontinuity at GPA = 2.8
# p-value > 0.05 -> no evidence of manipulation -> RDD valid
dens_test <- rddensity(X = data$gpa_econ, c = cutoff)
summary(dens_test)
rdplotdensity(dens_test, data$gpa_econ,
              title = "McCrary Test: GPA Density at Cutoff")
```

Test Sorting Behavior



Step 3: Preparation for Estimation

- Generate centered running variable and interaction terms:
  - **gpa_c**: $\tilde{A}_i = A_i - 2.8$, centered running variable ($= 0$ at cutoff)
  - **gpa_c2**: $\tilde{A}_i^2$, for quadratic robustness check
  - **above**: $Z_i = \mathbf{1}[A_i \geq 2.8]$ (already in dataset)
  - **above_gpa_c**: $Z_i \cdot \tilde{A}_i$, allows slope to differ on each side
  - **above_gpa_c2**: $Z_i \cdot \tilde{A}_i^2$, for quadratic specification
- Also define **data_emp**: subsample of employed workers (for wage regressions)

STATA: Step 3

Preparation for Estimation

```
1 gen gpa_c         = gpa_econ - 2.8
2 gen gpa_c2        = gpa_c^2
3 gen above_gpa_c   = above * gpa_c
4 gen above_gpa_c2 = above * gpa_c2
```

R: Step 3

Preparation for Estimation

```
data <- data |>
  mutate(gpa_c = gpa_econ - cutoff, gpa_c2 = gpa_c^2,
         above_gpa_c = above * gpa_c, above_gpa_c2 = above * gpa_c
             ^2)
data_emp <- data |> filter(employed_1718 == 1, !is.na(logwage_1718)
    )
```

Step 4: Parametric Estimation

First Stage, Reduced Form, and 2SLS

- Estimate three regressions; LATE = Reduced Form / First Stage
  - **First Stage**: effect of crossing GPA 2.8 on economics enrollment

$$D_i = \alpha_1 + \rho_1 Z_i + \beta_1 \tilde{A}_i + \delta_1 Z_i \tilde{A}_i + \upsilon_i$$

  - **Reduced Form** (Intent-to-Treat): effect of crossing cutoff on wages

$$Y_i = \alpha_2 + \rho_2 Z_i + \beta_2 \tilde{A}_i + \delta_2 Z_i \tilde{A}_i + \eta_i$$

  - **Fuzzy IV (2SLS)**: instrument D_i with Z_i to recover LATE $= \rho_2/\rho_1$
- Bleemer & Mehta (2022) simulated data: $\hat{\rho}_1 \approx 0.36$, $\hat{\rho}_2 \approx 0.14$, LATE ≈ 0.38 (46% wage premium)

STATA: Step 4

First Stage and Reduced Form

```
* First stage: Z_i -> D_i
reg econ_major above gpa_c above_gpa_c, cl(gpa_c)
display "First stage jump: " _b[above]

* Reduced form: Z_i -> Y_i  (Intent-to-Treat)
reg logwage_1718 above gpa_c above_gpa_c if employed_1718, cl(gpa_c
    )
display "Reduced form: " _b[above]

* Manual LATE = Reduced Form / First Stage
quietly reg econ_major above gpa_c above_gpa_c, cl(gpa_c)
local fs_coef = _b[above]
quietly reg logwage_1718 above gpa_c above_gpa_c if employed_1718,
    cl(gpa_c)
display "LATE = " _b[above] / `fs_coef'
```

STATA: Step 4

Fuzzy RDD — 2SLS (LATE)

```
* Fuzzy RDD: 2SLS
* Syntax: (econ_major = above) instruments D_i with Z_i
ivregress 2sls logwage_1718 gpa_c above_gpa_c  /*
*/      (econ_major = above) if employed_1718, vce(cl gpa_c)
display "LATE (2SLS): " _b[econ_major]

* OLS (biased upward due to ability selection -- for comparison)
* Unobserved ability -> both major choice and wages
reg logwage_1718 econ_major gpa_c above_gpa_c  /*
*/      if employed_1718, cl(gpa_c)
* OLS coefficient on econ_major > LATE (upward bias)
```

R: Step 4

First Stage, Reduced Form, 2SLS

```r
library(AER)
# First stage
fs <- lm(econ_major ~ above + gpa_c + above_gpa_c, data = data)
cat("First stage:", round(coef(fs)["above"], 4), "\n")
# Reduced form (ITT)
rf <- lm(logwage_1718 ~ above + gpa_c + above_gpa_c, data = data_
    emp)
cat("Reduced form:", round(coef(rf)["above"], 4), "\n")
cat("Manual LATE:", round(coef(rf)["above"]/coef(fs)["above"], 4),
    "\n")
# Fuzzy IV (2SLS)
fuzzy_iv <- ivreg(
  logwage_1718 ~ econ_major + gpa_c + above_gpa_c |
                 above      + gpa_c + above_gpa_c, data = data_emp)
cat("LATE (2SLS):", round(coef(fuzzy_iv)["econ_major"], 4), "\n")
# OLS (biased -- for comparison)
ols <- lm(logwage_1718 ~ econ_major + gpa_c + above_gpa_c, data =
    data_emp)
cat("OLS:", round(coef(ols)["econ_major"], 4), "(expect > LATE)\n")
```

Empirical Example: Bleemer & Mehta (2022)

Key Findings

- **Large Returns**: Majoring in economics caused:
  - 46% increase in early-career wages ($22,000)
  - Similar effects for male and female students
  - Possibly larger effects for URM students
    - URM: Underrepresented Minority includes Black, Hispanic, and Native American students

Step 5: Nonparametric Estimation

`rdrobust` with `fuzzy()`

- `rdrobust` with `fuzzy()` computes LATE directly
  - Uses Z_i as instrument for D_i automatically
  - Data-driven MSE-optimal bandwidth
  - Bias-corrected, robust confidence intervals
- Run three `rdrobust` calls in the fuzzy case:
  1. First stage: `rdrobust econ_major gpa_econ, c(2.8)`
  2. Reduced form: `rdrobust logwage …, c(2.8)`
  3. LATE: `rdrobust logwage …, c(2.8) fuzzy(econ_major)`

STATA: Step 5

Nonparametric Estimation — rdrobust

```
* First stage (sharp RDD on treatment take-up)
rdrobust econ_major gpa_econ, c(2.8) kernel(tri) all

* Reduced form
rdrobust logwage_1718 gpa_econ if employed_1718, c(2.8) kernel(tri)
     all

* Fuzzy RDD: LATE
* fuzzy() specifies the actual treatment D_i
rdrobust logwage_1718 gpa_econ if employed_1718, /*
*/    c(2.8) fuzzy(econ_major) kernel(tri) all

* Covariate balance tests (all p-values should be > 0.05)
foreach var in sat_score female urm zip_income {
    rdrobust `var' gpa_econ, c(2.8) kernel(tri)
}
```

R: Step 5

Nonparametric Estimation — rdrobust

```
# First stage
summary(rdrobust(data$econ_major, data$gpa_econ,
                 c = cutoff, kernel = "tri", all = TRUE))
# Reduced form
summary(rdrobust(data_emp$logwage_1718, data_emp$gpa_econ,
                 c = cutoff, kernel = "tri", all = TRUE))
# Fuzzy RDD: LATE (fuzzy= specifies D_i; uses Z_i automatically)
rdd_fuzzy <- rdrobust(y = data_emp$logwage_1718, x = data_emp$gpa_
    econ,
                      fuzzy = data_emp$econ_major,
                      c = cutoff, kernel = "tri", all = TRUE)
summary(rdd_fuzzy)
# Covariate balance
for (v in c("sat_score", "female", "urm", "zip_income")) {
  b <- rdrobust(data[[v]], data$gpa_econ, c = cutoff, kernel = "tri
      ")
  cat(v, ": p =", round(b$pv[1], 4), "\n")
}
```

Step 6: Robustness Checks

- **Bandwidth sensitivity**: LATE should be stable across choices of h
  - Bleemer & Mehta (2022) find estimates stable from $h = 0.3$ to $h = 1.0$
- **Polynomial order**: try quadratic specification
  - Gelman & Imbens (2019): avoid going above second-order polynomial
- **Alternative outcomes**: employment, graduate school enrollment
  - If GPA 2.8 affects wages only through economics major, other outcomes should respond only if they too are affected by major choice

STATA: Step 6

Robustness Checks

```
* Bandwidth sensitivity
foreach h in 0.3 0.4 0.5 0.6 0.8 1.0 {
    quietly rdrobust logwage_1718 gpa_econ if employed_1718, /*
    */    c(2.8) h(`h') fuzzy(econ_major) kernel(tri)
    display "h=`h': LATE=" e(tau_cl) " SE=" e(se_tau_cl)
}

* Quadratic polynomial (2SLS)
ivregress 2sls logwage_1718 gpa_c gpa_c2 above_gpa_c above_gpa_c2
    /*
*/      (econ_major = above) if employed_1718, vce(cl gpa_c)

* Secondary outcomes
rdrobust employed_1718 gpa_econ, c(2.8) fuzzy(econ_major) kernel(
    tri)
rdrobust grad_school   gpa_econ, c(2.8) fuzzy(econ_major) kernel(
    tri)
```

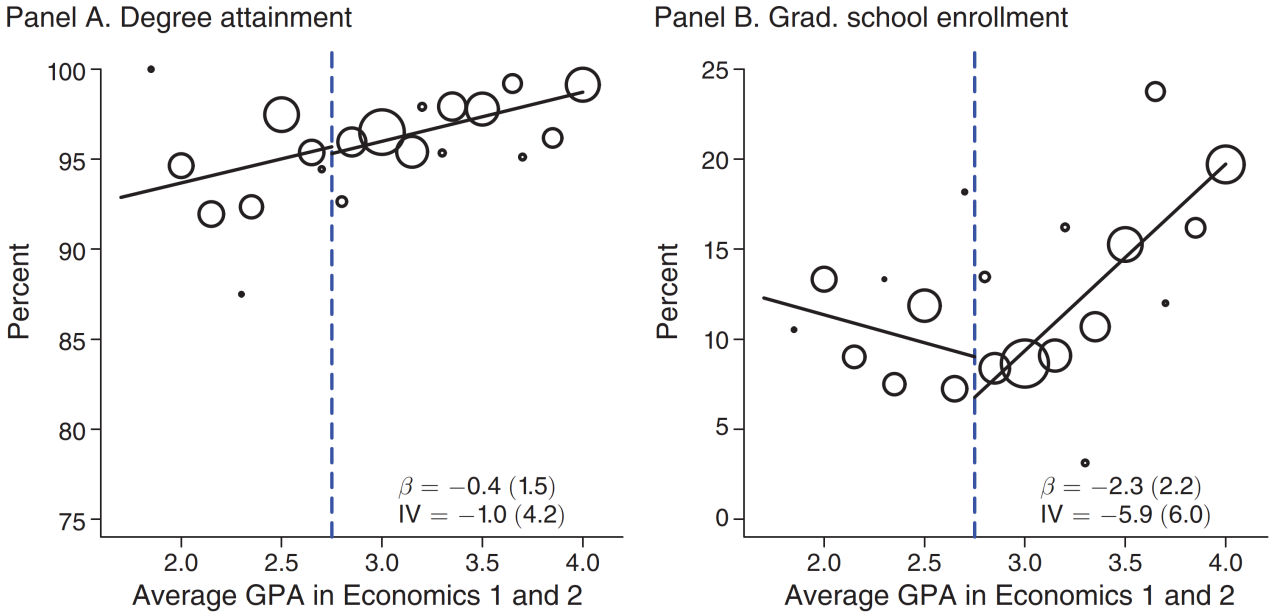
R: Step 6

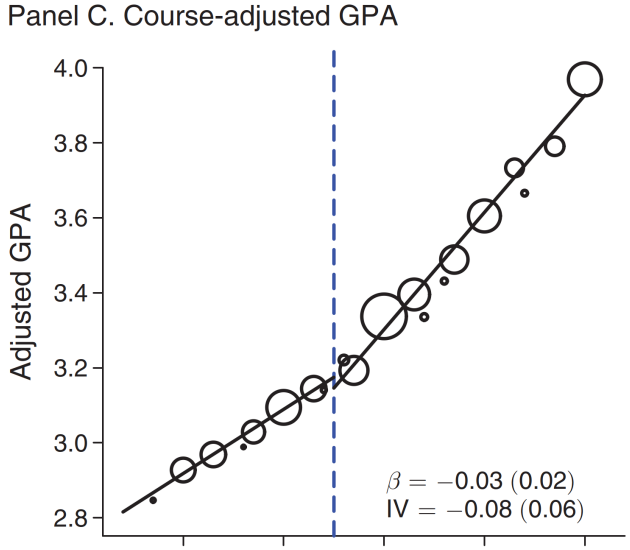
Robustness Checks

```r
# Bandwidth sensitivity
for (h in c(0.3, 0.4, 0.5, 0.6, 0.8, 1.0)) {
  rdd_h <- rdrobust(y = data_emp$logwage_1718, x = data_emp$gpa_econ,
                    fuzzy = data_emp$econ_major,
                    c = cutoff, h = h, kernel = "tri")
  cat(sprintf("h=%.1f: LATE = %.4f\n", h, rdd_h$coef[1]))
}
# Quadratic polynomial (2SLS)
fuzzy_quad <- ivreg(
  logwage_1718 ~ econ_major + gpa_c + gpa_c2 + above_gpa_c + above_gpa_c2 |
                 above      + gpa_c + gpa_c2 + above_gpa_c + above_gpa_c2,
  data = data_emp)
cat("Quadratic LATE:", round(coef(fuzzy_quad)["econ_major"], 4), "\n")
```

Empirical Example: Bleemer & Mehta (2022)

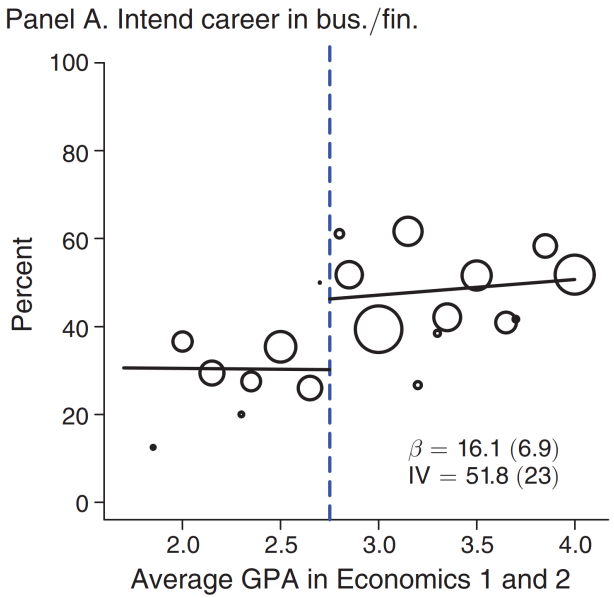
Mechanisms: Educational Investment

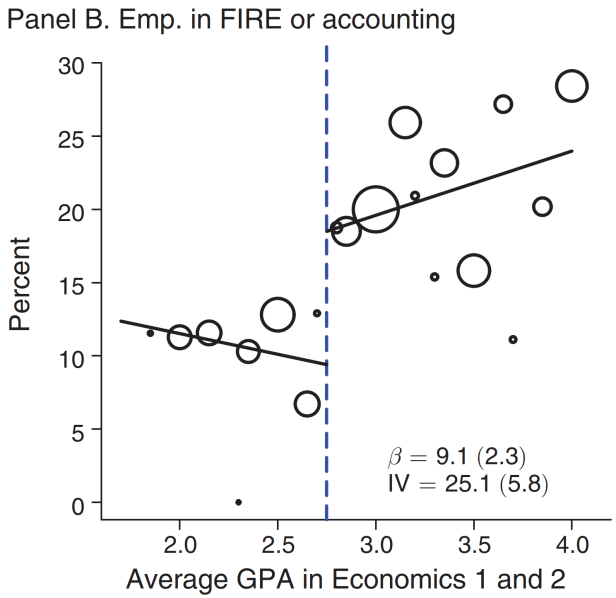


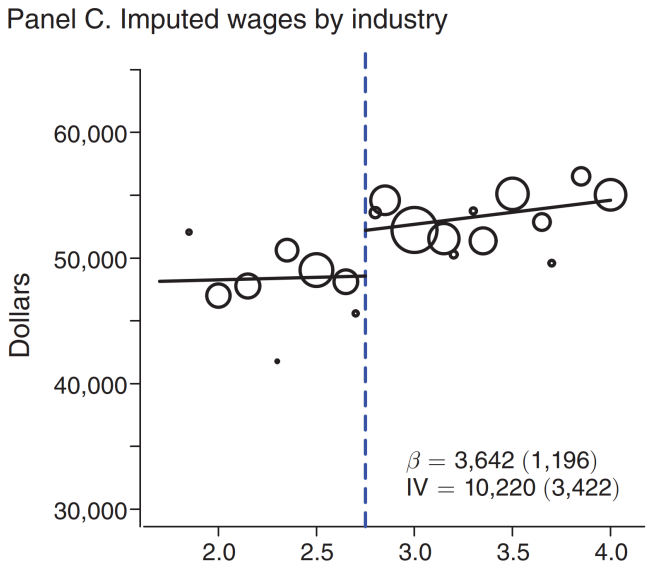


Empirical Example: Bleemer & Mehta (2022)

Mechanisms: Industry Effects







Empirical Example: Bleemer & Mehta (2022)

Mechanisms

- **Educational Investment**:
  - No effect on graduation rates or graduate school enrollment
  - No change in study time or course-adjusted grades
  - 13 more economics courses, 9 fewer other social science courses
- **Industry Effects**:
  - 52 pp more likely to prefer business/finance careers
  - 25 pp more likely to work in FIRE or accounting
  - About half ($10,220) of wage effect explained by industry sorting

Suggested Readings

- Chapter 4, *Mastering 'Metrics: The Path from Cause to Effect*
- Chapter 6, *Mostly Harmless Econometrics: An Empiricist's Companion*

- Chapter 6, *Causal Inference: The Mixtape*