

Lecture 9: Regression Discontinuity Design

Introduction to Econometrics, Spring 2023

Zhaopeng Qu

Nanjing University Business School

May 03 2023



- 1 Causal Inference and Regression Discontinuity Design
- 2 Regression Discontinuity Design
- 3 Identification: Model Specification and Bandwidth Selection
- 4 Application: Effect of the Minimum Legal Drinking Age (MLDA) on death rates
- 5 Fuzzy RDD: IV and Application
- 6 Implement of RDD
- 7 A Latest Case

Causal Inference and Regression Discontinuity Design

Review the Basic idea of Causal Inference

- Social science (Economics) theories always ask causal question
- In general, a typical causal question is: The effect of a treatment(D) on an outcome(Y)
 - Outcome(Y): A variable that we are interested in
 - Treatment(D): A variable that has the (causal) effect on the outcome of our interest
- 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 Experiment** (RCT).

Experimental Idea

- In an RCT, researchers can eliminate selection bias by controlling treatment assignment process.
- An RCT randomizes
 - **treatment group** who receives a treatment
 - **control group** who does not
- 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

Causal Inference

- 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”
 - **Instrumental Variable**: Use IVs which are very much alike the endogenous variable but are enough exogenous(randomized) to proxy the treatment and control status.
- **Regression Discontinuity Design**(RDD) is another widely used method to make causal inference which is consider as more reliable and more robust.

Regression Discontinuity Design

Main Ideas

- Regression Discontinuity Design (RDD) exploits the facts that:
 - Some rules are *arbitrary* and generate a *discontinuity* in treatment assignment.
 - The treatment assignment is determined based on whether a unit exceeds some threshold on a variable (**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.

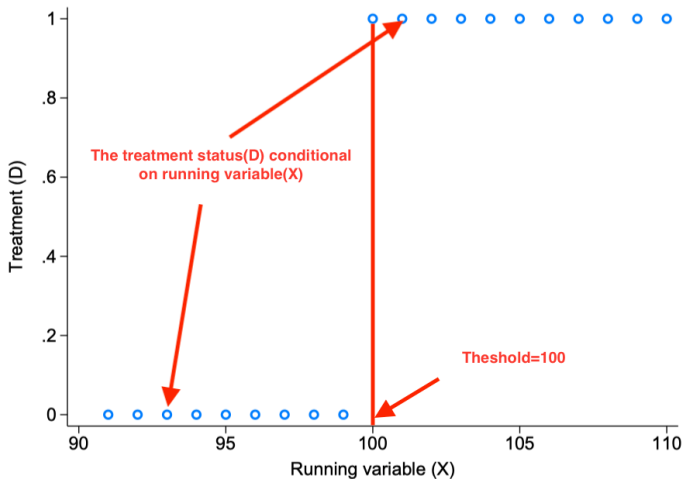
A Motivating Example: Elite University

- Numerous studies have shown that graduates from more **selective** programs or universities earn more than others.
 - e.g Students graduated from **NJU** averagely earn more than those graduated from other ordinary universities like **NUFE**(南京财经大学).
- But it is difficult to know whether the positive earnings premium is due to
 - **true causal impact** of human capital acquired in the academic program.
 - a **spurious correlation** linked to the fact that good students selected in these programs would have earned more no matter what.(**Selection Bias**).
- OLS regression will not give us the right answer for the bias.(Why?)

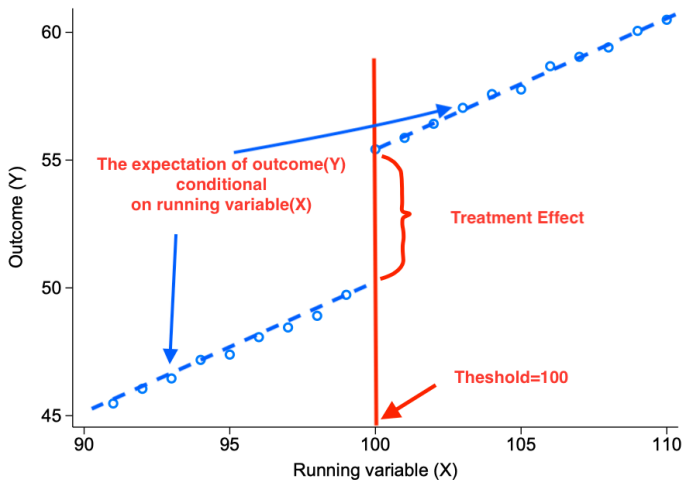
A Motivating Example: Elite University

- But if we could know *National College Entrance Exam Scores* (高考成绩) of all the students. Then we can do something.
- Let us say that the entry cutoff for a score of entrance exam is **600** for NJU.
- Those with scores **590** or even **599** are unlikely to attend NJU, instead attend NUFE(南京财经大学).
- Assume that those get *599* or *595* and those get *600* are **essentially identical**, the different scores can be attributed to *some random events*.
- **RD strategy**: Comparing the long term outcomes(such as earnings in labor market) for the students with 600 (admitted to NJU) and those with the 599 (admitted at NUFE).

Main Idea of RDD: Treatment



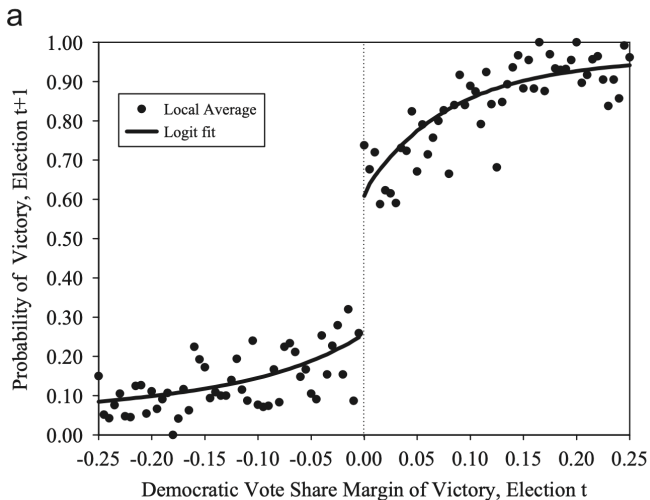
Main Idea of RDD: Outcome



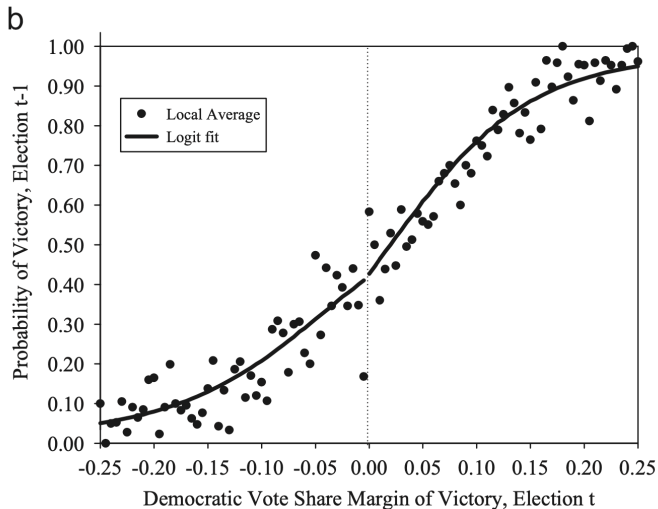
A Classical Example: Lee(2008)

- Important phenomenon in politics: The incumbency advantage
 - Candidates/parties who won the previous election are much more likely to win again.
- By definition, incumbents are candidates who were successful last time. Some or all of incumbency advantage could be due to persistent unobservables.
 - How much of the incumbency advantage is causal?
- Lee (2008) uses an RD design to estimate the causal effect of winning US House elections.

The incumbency advantage



The incumbency advantage



More Cases of RDD

- **Academic test scores:** scholarship, prize, higher education admission, certifications of merit.
- **Poverty scores:** means-tested anti-poverty programs.
- **Land area:** fertilizer program, debt relief initiative for owners of plots below a certain area
- **Date:** age cutoffs for pensions, dates of birth for starting school with different cohorts, date of loan to determine eligibility for debt relief.
- **Elections:** fraction that voted for a candidate of a particular party
- **Geography in policy:** China's Huai River Heating Policy, Spanish's Slavery "Mita" of colonial Peru in sixteen century, and American Air force Bombing in Vietnam War.

RD as Local Randomization

- RD provides “local” randomization if the following assumption holds:
 - Agents have **imperfect** control over the assignment variable X .
- **Assumptions:**
 - the randomness guarantees that the potential outcome curves are smooth (e.g continuous) around the cutoff point.
 - There are no discrete jumps in outcomes at threshold except due to the treat.
 - All observed and unobserved determinants of outcomes are smooth around the cutoff.

RDD and Potential Outcomes: Notations

- Treatment

- assignment variable (running variable): X_i
- Threshold (cutoff) for treatment assignment: c
- Treatment variable: D_i and treatment assignment rule is

$$D_i = 1 \text{ if } X_i \geq c \text{ and } D_i = 0 \text{ if } X_i < c$$

- Potential Outcomes

- Potential outcome for an individual i with treatment, Y_{1i}
- Potential outcome for an individual i without treatment, Y_{0i}

- Observed Outcomes

$$Y_{1i} \text{ if } D_i = 1 (X_i \geq c) \text{ and } Y_{0i} \text{ if } D_i = 0 (X_i < c)$$

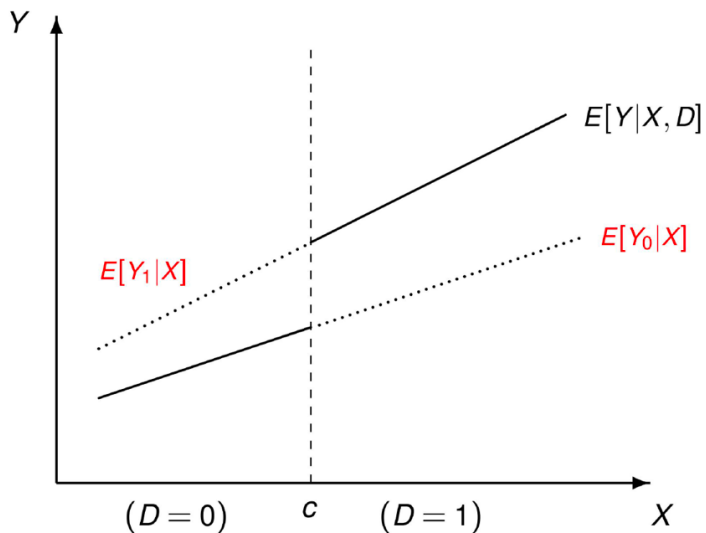
Identification for Sharp RDD

- **Continuity Assumption**

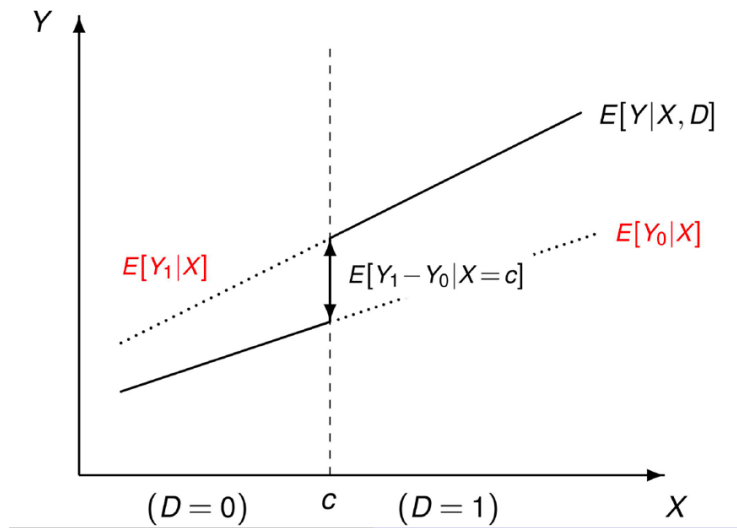
- $E[Y_{1i}|X_i]$ and $E[Y_{0i}|X_i]$ are continuous at $X_i = c$
- Assume potential outcomes do not change at cutoff.
- This means that except treatment assignment, all other unobserved determinants of Y_i are continuous at cutoff c .
- This implies no other confounding factor affects outcomes at cutoff c .
- Any observed discontinuity in the outcome can be attributed to treatment assignment.

$$\rho_{SRD} = \lim_{\varepsilon \rightarrow 0} E[Y_i | X_i = c + \varepsilon] - \lim_{\varepsilon \rightarrow 0} E[Y_i | X_i = c - \varepsilon]$$

Graphical Interpretation



Graphical Interpretation



Continuity Assumption

- **Continuity** is a natural assumption but could be **violated** if:
 - 1 There are differences between the individuals who are just below and above the cutoff that are NOT explained by the treatment.
 - The same cutoff is used to assign some other treatment.
 - Other factors also change at cutoff.
 - 2 Individuals can **fully manipulate** the running variable in order to gain access to the treatment or to avoid it.

Identification: Model Specification and Bandwidth Selection

Basic Parametric RDD specification

- A simple RD regression is

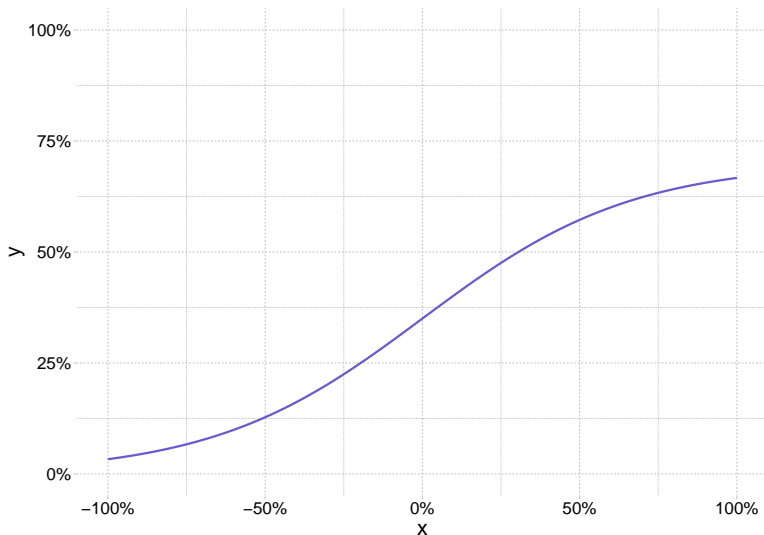
$$Y_i = \alpha + \rho D_i + \gamma(X_i - c) + u_i$$

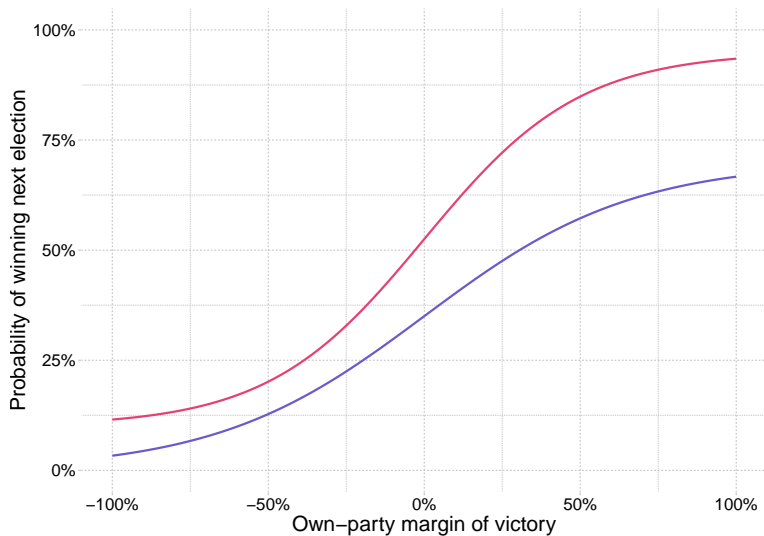
- Y_i is the outcome variable
- D_i is the the treatment variable(indepent variable)
- X_i is the running variable
- c is the value of cut-off
- u_i is the error term including other factors
- **Question:** Which parameter do we care about the most?
- But Linear function form is not enough.

Introduction

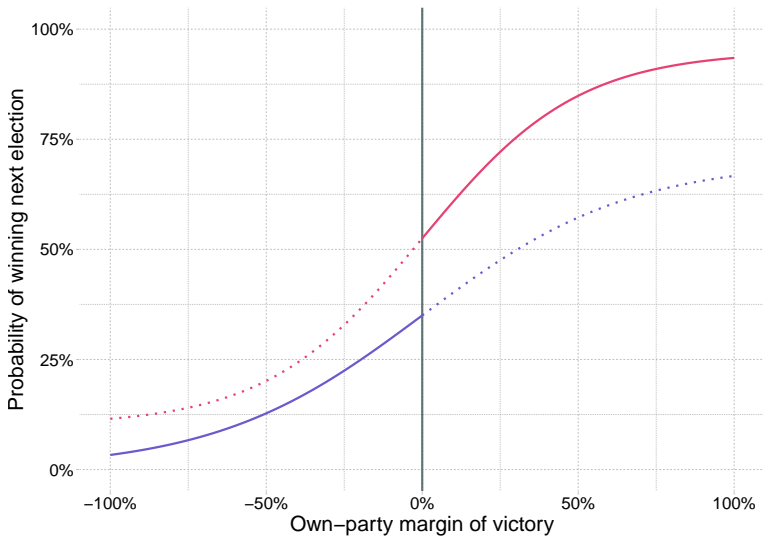
- Two Keys in specifications in RD :
- **Specification:** How should we estimate $E[Y_{1i} | X_i]$ and $E[Y_{0i} | X_i]$?
 - **Parametric:** Estimate treatment effects based on a specific functional form for the outcome and assignment variable relationship.
 - **Nonparametric:** Compare the outcome of treated and untreated observations that lie within specific bandwidth.
- **Bandwidth Selection:** How much data around the cut-off should we use—*i.e.* the widows size
 - **Global:** use all data available.
 - **Local:** only use data with specific bandwidth.

Specification and Bandwidth Selection

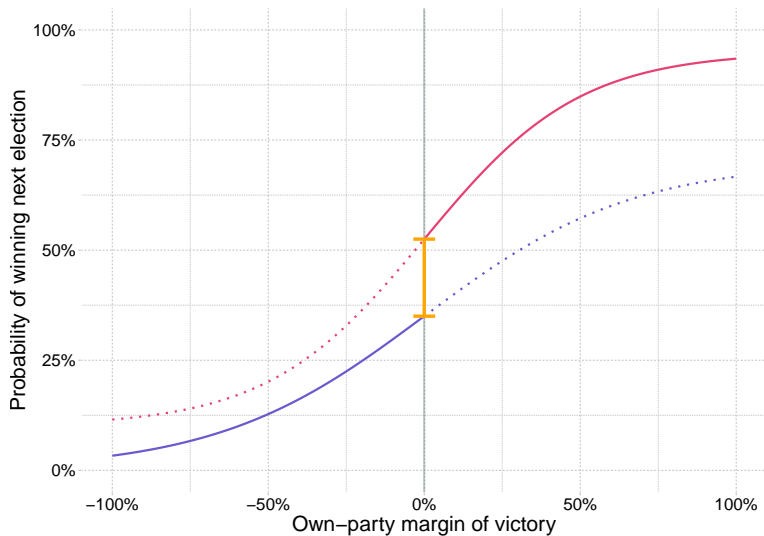


$E[Y_{0i} | X_i]$ and $E[Y_{1i} | X_i]$ 

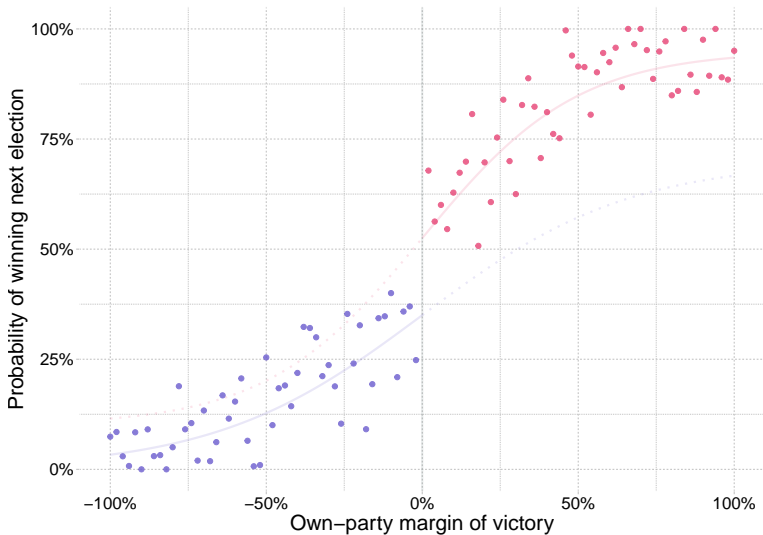
You only win an election if your **margin of victory exceeds zero**.



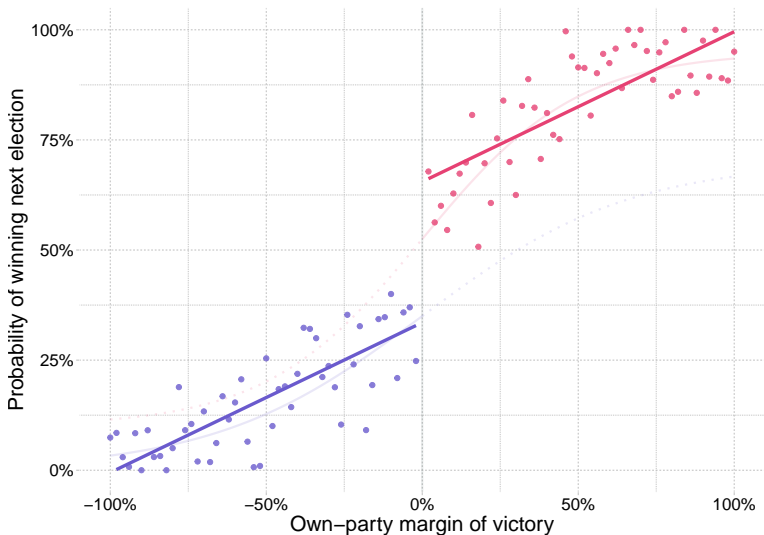
$E[Y_{1i} | X_i] - E[Y_{0i} | X_i]$ at the discontinuity gives ρ_{SRD} .



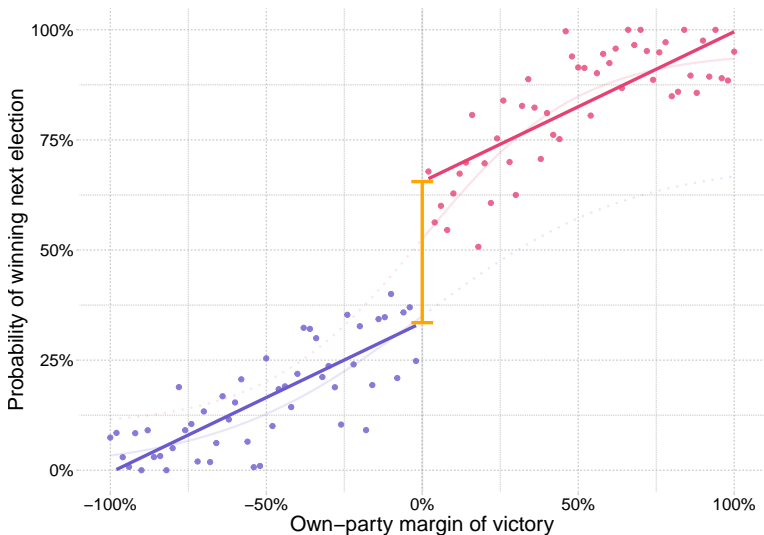
Real data are a bit trickier. We must **estimate** $E[Y_{1i} | X_i]$ and $E[Y_{0i} | X_i]$.

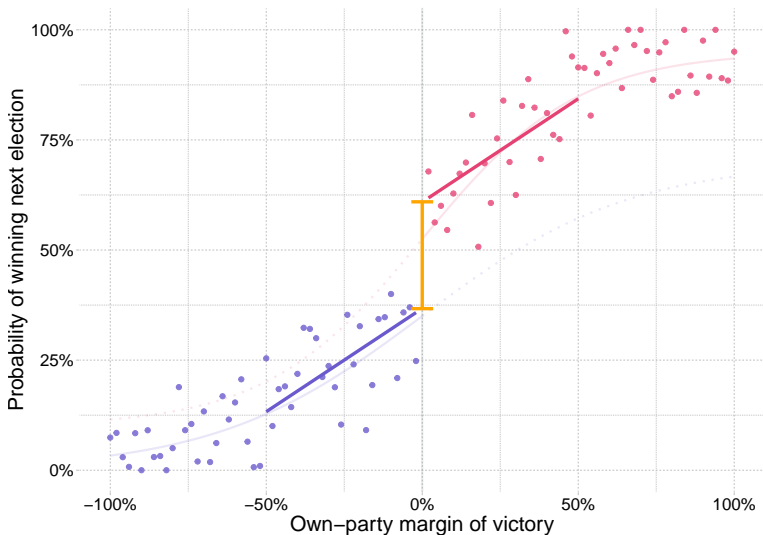


Linear regression with constant slopes (and all data)

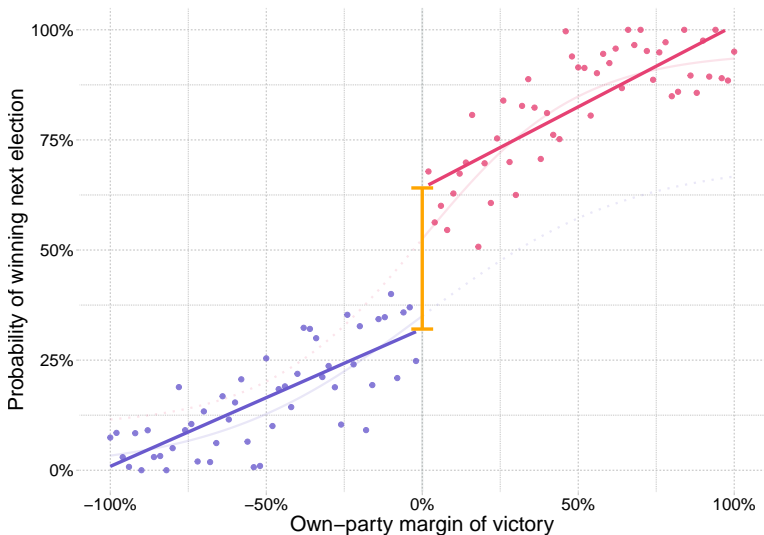


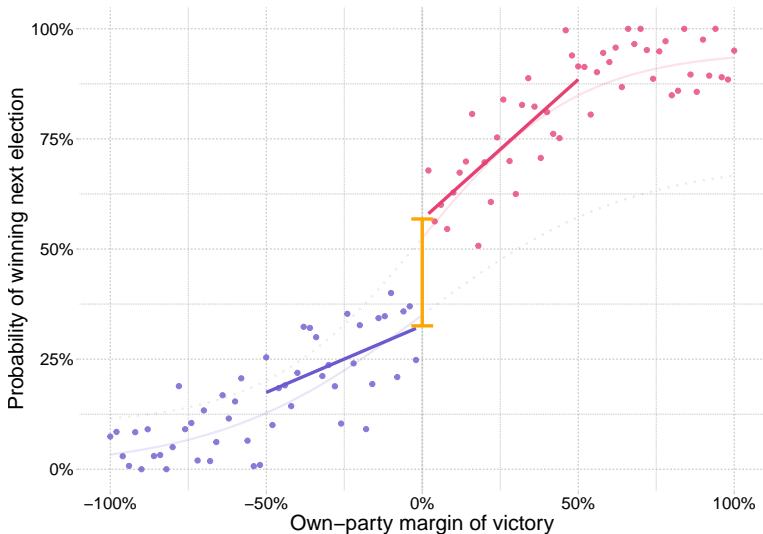
Linear regression with constant slopes (and all data)

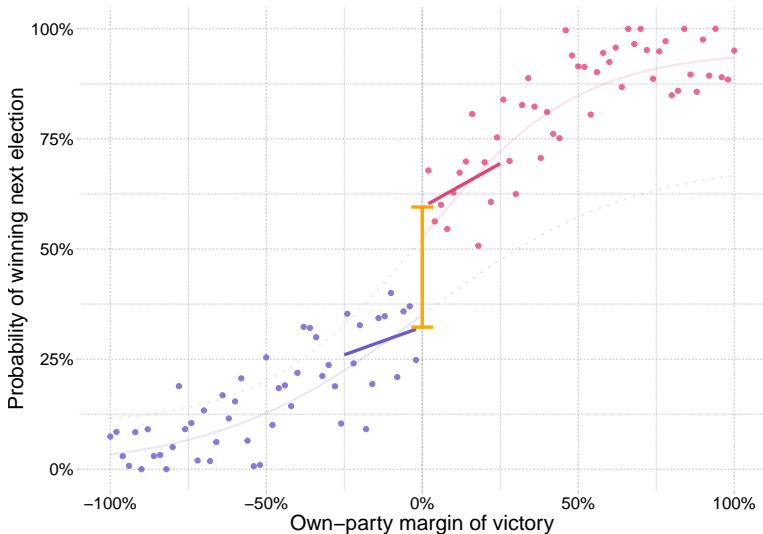


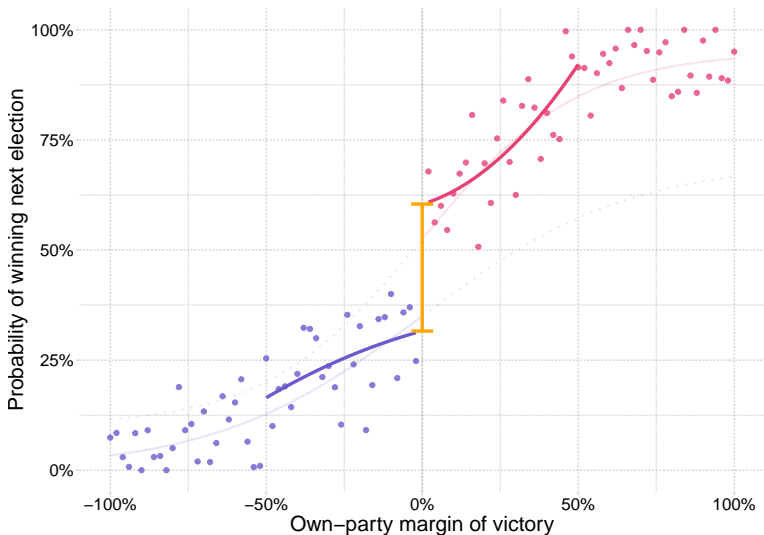
Linear regression with constant slopes; limited to $\pm 50\%$.

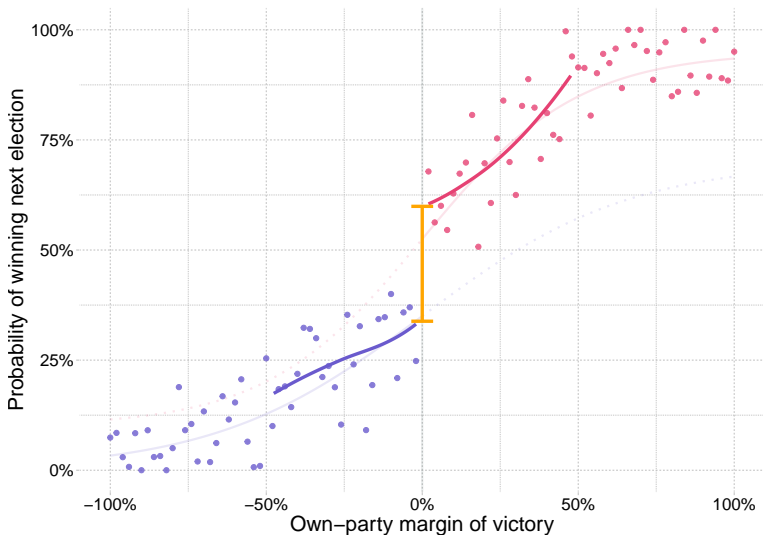
Linear regression with differing slopes (and all data)



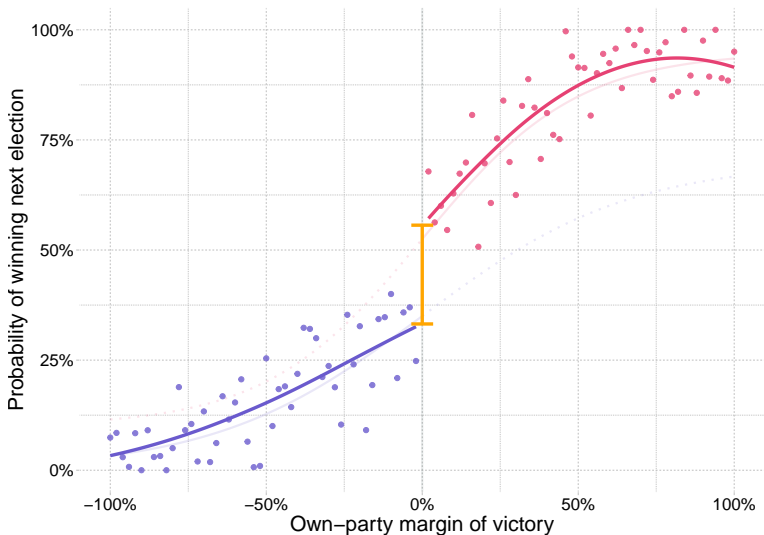
Linear regression with differing slopes; limited to $\pm 50\%$.

Linear regression with differing slopes; limited to $\pm 25\%$.

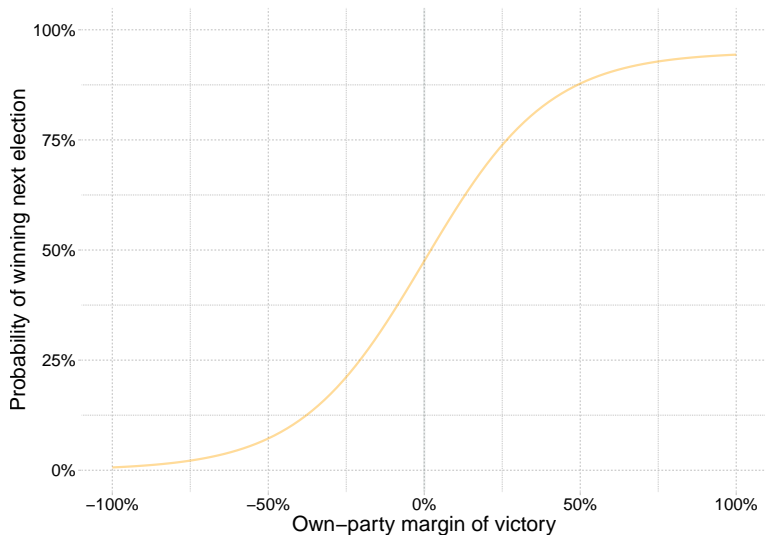
Differing quadratic regressions (limited to $\pm 50\%$).

Differing local (LOESS) regressions (limited to $\pm 50\%$).

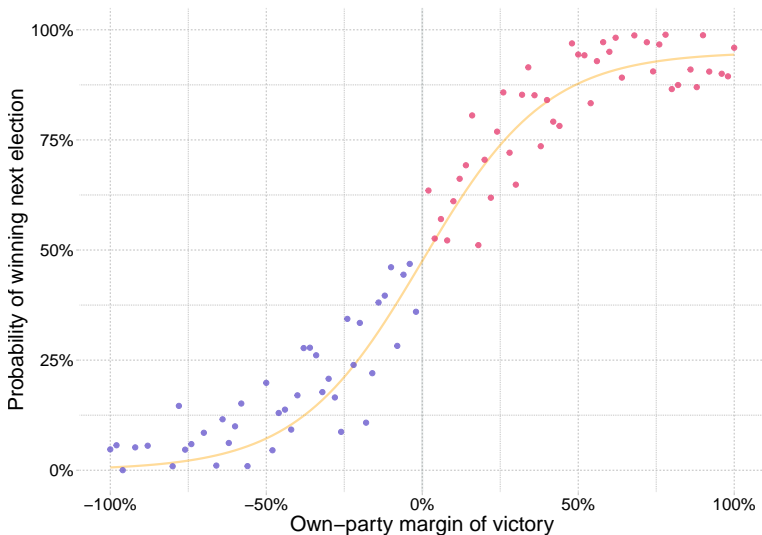
Differing local (LOESS) regressions (all data).



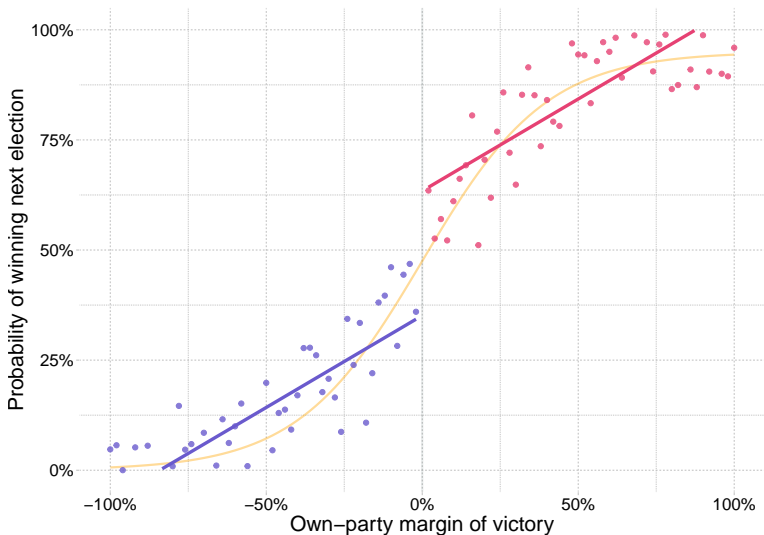
Functional form can be very important.



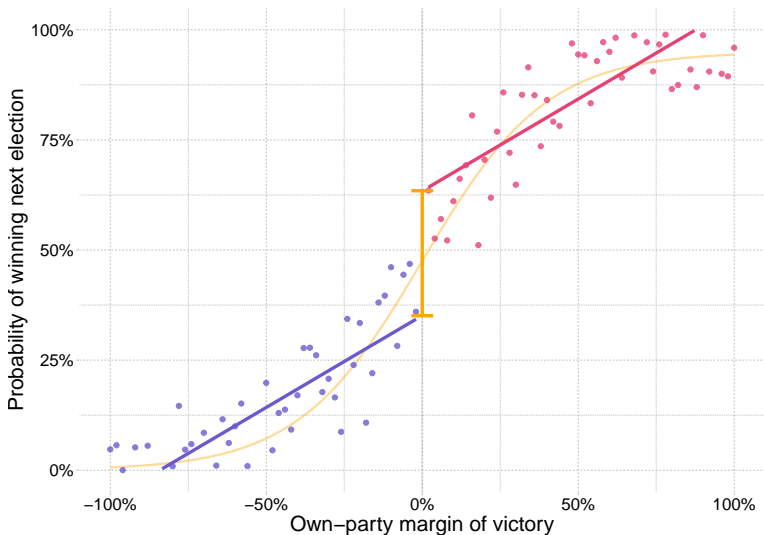
Functional form can be very important.



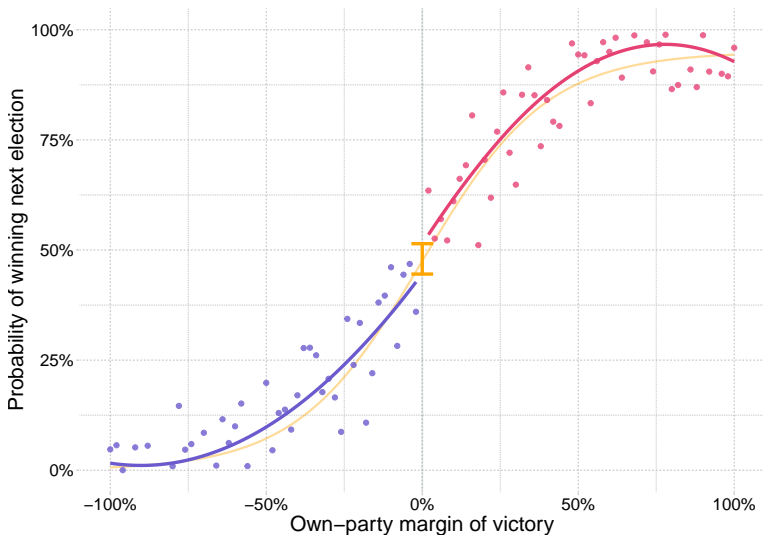
Functional form can be very important



Functional form can be very important



Functional form can be very important.



RDD Estimation

- There are 2 types of strategies for correctly specifying the functional form in a RDD:
 - 1 **Parametric**/global method: Use all available observations and 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.

Parametric/Global method

- Suppose that in addition to the assignment mechanism above, potential outcomes can be described by some reasonably smooth function $f(X_i)$

$$E[Y_{i0}|X_i] = \alpha + f(X_i)$$

$$Y_{i1} = Y_{i0} + \rho$$

- Simply, we can construct RD estimates by fitting

$$Y_i = \alpha + \rho D_i + f(X_i) + u_i$$

Specification in RDD

- More generally, we could also estimate two separate regressions for each side respectively.

$$Y_i^b = \beta^b + f(X_i^b - c) + u_i^b$$

$$Y_i^a = \beta^a + g(X_i^a - c) + u_i^a$$

- Continue Assumption: $f(\cdot)$ and $g(\cdot)$ be any continuous function of $(x_i^{a,b} - c)$, and satisfy

$$f(0) = g(0) = 0$$

- We estimate equation using only data above c and only data below c .
- Then the treatment effect is $\rho = \beta^b - \beta^a$

Specification in RDD

- Can do all in one step; just use all the data at once and estimate:

$$Y_i = \alpha + \rho D_i + f(X_i - c) + D_i \times h(X_i - c) + u_i$$

where D_i is a dummy variable for treated status.

- Then when $D_i=0$, thus

$$Y_i = \alpha + f(X_i - c) + u_i$$

- Then when $D_i=1$, let $g(X_i - c) = f(X_i - c) + h(X_i - c)$, then

$$Y_i = \alpha + \rho + g(X_i - c) + u_i$$

Nonlinear Function forms

- Use a flexible polynomial (p th order polynomial) regression to estimate $f(X_i)$
- Let $f(X_i) = \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_p X_i^p$
- In a simple case: a flexible polynomial (p th order polynomial) regression to estimate $f(x_i)$ and $g(x_i)$

$$Y_i = \alpha + \rho D_i + \beta_1 X_i + \beta_2 X_i^2 + \dots + \beta_p X_i^p + \eta_i$$

- How to decide which polynomial to use?
 - start with the **eyeball test**, similar to OLS regression

Parametric Approach

- Let

$$\begin{aligned} f(X_i - c) &= f(\tilde{X}_i) \\ &= \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \dots + \beta_p \tilde{X}_i^p \end{aligned}$$

$$\begin{aligned} h(X_i - c) &= h(\tilde{X}_i) \\ &= \beta_1^* \tilde{X}_i + \beta_2^* \tilde{X}_i^2 + \dots + \beta_p^* \tilde{X}_i^p \end{aligned}$$

- In a comprehensive case, the regression model which we estimate is then

$$\begin{aligned} Y_i &= \alpha + \rho D_i + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \dots + \beta_p \tilde{X}_i^p \\ &\quad + \beta_1^* D_i \tilde{X}_i + \beta_2^* D_i \tilde{X}_i^2 + \dots + \beta_p^* D_i \tilde{X}_i^p + u_i \end{aligned}$$

- The treatment effect at c is ρ

How to Select Select Polynomial Order

- F-Test
 - use F-test in OLS regression to test the order
- AIC approach
 - Akaike information criterion (AIC) procedure
- BIC approach
 - Bayesian information criterion(BIC) procedure

Gelman and Imbens (2018)

Gelman and Imbens (2018) on functional form:

- **controlling for global high-order polynomials** is a flawed approach with three major problems:
 - it leads to noisy estimates, sensitivity to the degree of the polynomial, and poor coverage of confidence intervals.
- Recommending researchers instead use estimators based on **local linear** or **quadratic polynomials** or **other smooth functions**.

Nonparametric/Local Approach

- Recall we can construct RD estimates by fitting

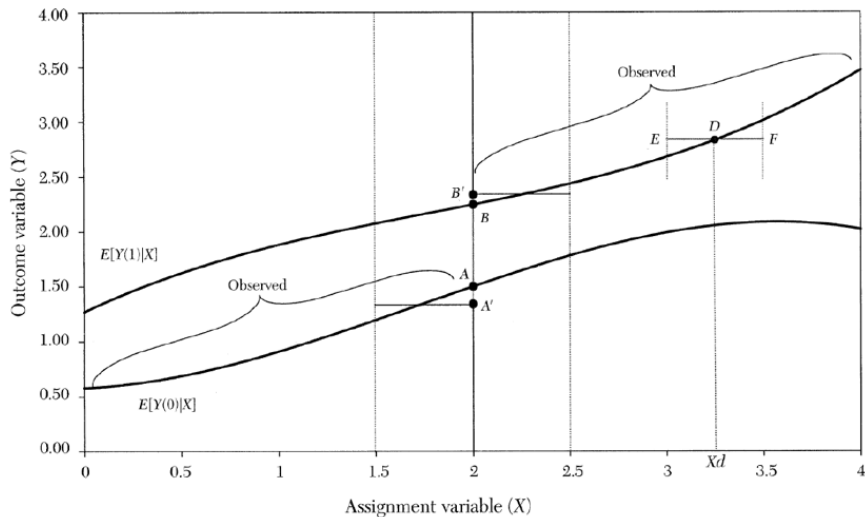
$$Y_i = \alpha + \rho D_i + f(x_i) + u_i$$

- Nonparametric approach does **NOT** specify particular functional form of the outcome and the assignment variable, thus $f(x_i)$
- Instead, it uses only data within a small neighborhood (known as **bandwidth**) to estimate the discontinuity in outcomes at the cutoff:
 - Compare means in the two bins adjacent to the cutoff (treatment v.s. control groups)
 - Local linear polynomial regression (a formal nonparametric regression method)

Nonparametric/Local Approach

- However, comparing means in the two bins adjacent to the cutoff is generally **biased** in the neighborhood of the cutoff.
- This is called **boundary bias**.

Nonparametric/Local Approach: boundary bias



Nonparametric/Local Approach

- The most often used nonparametrics method is **local linear polynomial regression**, which is linear smoother within a given interval.
- Thus we estimate the following weighted linear regression within a given window of width h :

$$Y_i = \alpha + \rho D_i + \beta_1 \tilde{X}_i + \beta_1^* D_i \tilde{X}_i + u_i$$

- Here we often use some nonparametric functions (such as **kernel**) as the weight, which measures the “distance” to the cut-off.
- *The detail is a little bit beyond the scope of this course. You could refer to Li and Racine (2006) or other nonparametric econometric textbooks.*

Nonparametric/Local Approach: boundary bias

- The main challenge of nonparametric approach is to **choose a bandwidth**.
- There is essentially a trade-off between bias and precision
- Use a larger bandwidth:
 - Get more **precise** treatment effect estimates since more data points are used in the regression.
 - But the linear specification is less likely to be accurate and the estimated treatment effect could be biased.

How to Choose Bandwidth

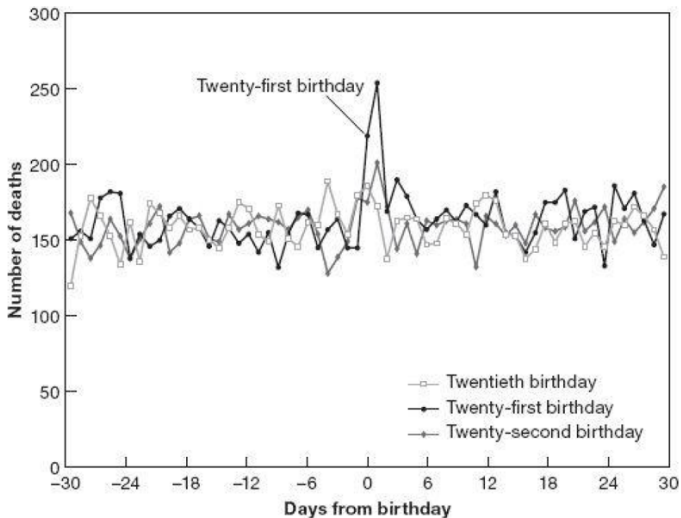
- **Bias/variance trade-off:** Smaller bandwidth reduces bias from using points away from the boundary, but also reduces precision.
- The optimal bandwidth: use
 - **Cross-Validation Procedure:** Choose bandwidth h that produces the best fit for the relationship of outcome and assignment variable.
- Usually, we would present the RD estimates by different choices of bandwidth.

Application: Effect of the Minimum Legal Drinking Age (MLDA) on death rates

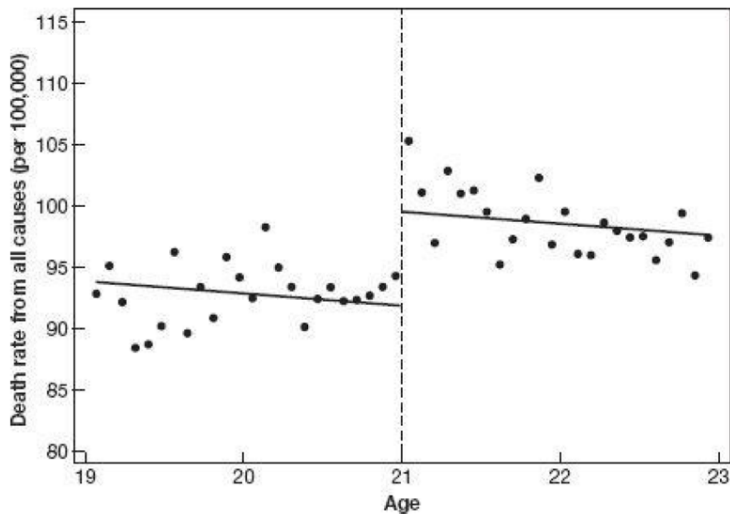
Introduction

- Carpenter and Dobkin (2009)
- Topic: Birthdays and Funerals
- In American, **21th birthday** is an very important milestone. Because over-21s can drink legally.
- Two Views:
 - A group of American college presidents have lobbied states to return the minimum legal drinking age (MLDA) to the Vietnamera threshold of 18.
 - They believe that legal drinking at age 18 discourages binge drinking and promotes a culture of mature alcohol consumption.
 - MLDA at 21 reduces youth access to alcohol, thereby preventing some harm.
- Which one is right?

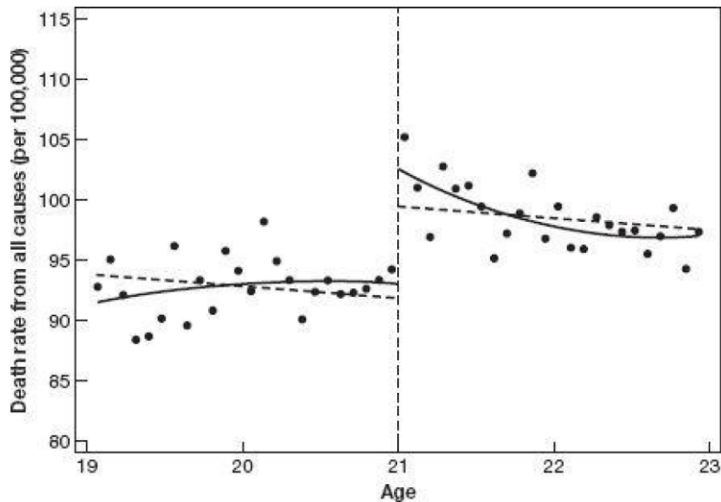
Application: MLDA on death rates



Application: MLDA on death rates



Application: MLDA on death rates



Application: MLDA on death rates:

- The cut off is age 21, so estimate the following regression with cubic terms

$$Y_i = \alpha + \rho D_i + \beta_1(x_i - 21) + \beta_2(x_i - 21)^2 + \beta_3(x_i - 21)^3 \\ + \beta_4 D_i(x_i - 21) + \beta_5 D_i(x_i - 21)^2 + \beta_6 D_i(x_i - 21)^3 + u_i$$

- The effect of legal access to alcohol on mortality rate at age 21 is ρ

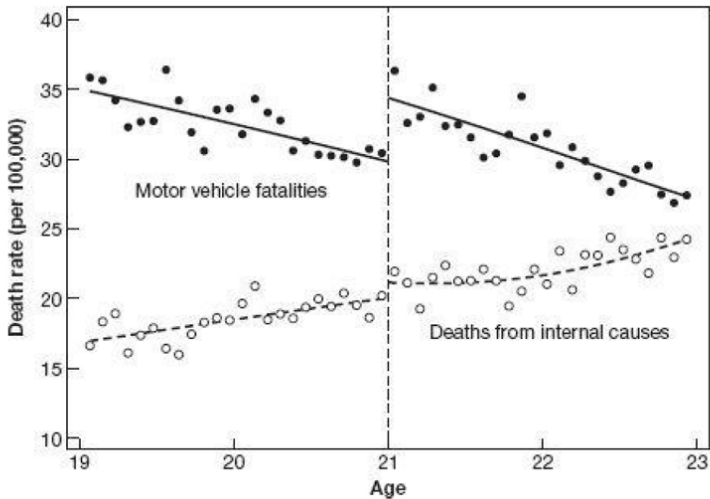
Application: MLDA on death rates

TABLE 4—DISCONTINUITY IN LOG DEATHS AT AGE 21

	(1)	(2)	(3)	(4)
<i>Deaths due to all causes</i>				
Over 21	0.096 (0.018)	0.087 (0.017)	0.091 (0.023)	0.074 (0.016)
Observations	1,460	1,460	1,460	1,458
R ²	0.04	0.05	0.05	
Prob > Chi-Squared		0.000	0.735	
<i>Deaths due to external causes</i>				
Over 21	0.110 (0.022)	0.100 (0.021)	0.096 (0.028)	0.082 (0.021)
Observations	1,460	1,460	1,460	1,458
R ²	0.06	0.08	0.08	
Prob > Chi-Squared		0.000	0.788	
<i>Deaths due to internal causes</i>				
Over 21	0.063 (0.040)	0.054 (0.040)	0.094 (0.053)	0.066 (0.031)
Observations	1,460	1,460	1,460	1,458
R ²	0.10	0.10	0.10	
Prob > Chi-Squared		0.000	0.525	
Covariates	N	Y	Y	N
Quadratic terms	Y	Y	Y	N
Cubic terms	N	N	Y	N
LLR	N	N	N	Y

Notes: See Notes from Table 1. The dependent variable is the log of the number of deaths that occurred x days from the person's twenty-first birthday. External deaths include all deaths with mention of an injury, alcohol use, or drug use. The Internal Death category includes all deaths not coded as external. Please see Web Appendix C for the ICD codes for each of the categories above. The first three columns give the estimates from polynomial regressions on age interacted with a dummy for being over 21.

Application: MLDA on death rates



Fuzzy RDD: IV and Application

Fuzzy RDD

- In sharp RDD not the treatment assignment but **the probability of treatment** jumps at the threshold.
 - **Sharp RDD:**
 - the probability of treatment jumps at the threshold from 0 to 1.
 - Nobody below the cutoff gets the “treatment”, everybody above the cutoff gets it.

Fuzzy RDD

- Treatment Assignment:

$P(D_i = 1|x_i) = p_1(X_i)$ if $x_i \geq c$, the probability assign to treatment group

$P(D_i = 1|x_i) = p_0(X_i)$ if $x_i < c$, the probability assign to control group

- **Fuzzy RDD**: Some individuals *above cutoff* do **NOT** get treatment and some individuals *below cutoff* do receive treatment.
- The result is a research design where the discontinuity becomes an **instrumental variable** for treatment status instead of deterministically switching treatment on or off.

Fuzzy RD v.s Sharp RD

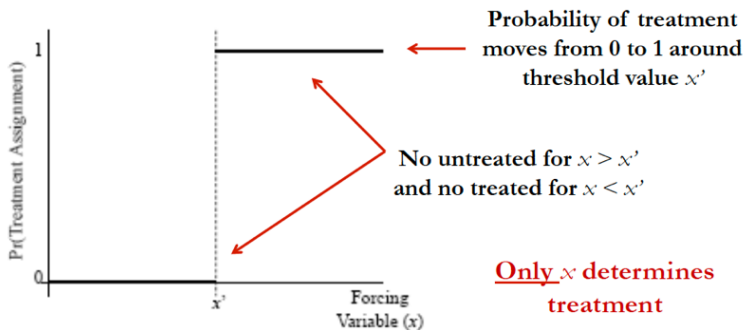
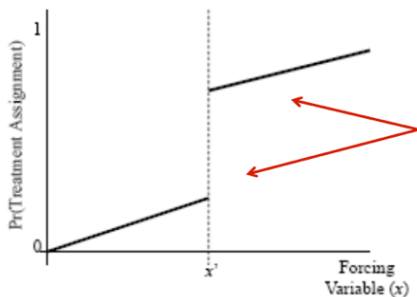


Figure is from Roberts and Whited (2010)

Fuzzy RD v.s Sharp RD



Treatment probability
increases at x'

Some untreated for $x > x'$
and some treated for $x < x'$

Treatment is not
purely driven by x

Figure is from Roberts and Whited (2010)

Identification in Fuzzy RD

- Encourage Variable:

$Z_i = 1$ if assign to treatment group

$Z_i = 0$ if assign to control group

- The relationship between the probability of treatment and X_i

$$P(D_i = 1|x_i) = p_0(x_i) + [p_1(x_i) - p_0(x_i)]Z_i$$

Identification in Fuzzy RD

- Recall in SRD, we estimate

$$Y_i = \alpha + \rho D_i + f(x_i - c) + D_i \times g(x_i - c) + u_i$$

- Then the **First Stage** of FRD regression:

$$P(D_i = 1|x_i) = \alpha_1 + \phi Z_i + f(x_i - c) + Z_i \times g(x_i - c) + \eta_{1i}$$

- Recall IV terminology: Which one is
 - endogenous variable?**
 - instrumental variable?**

Identification in Fuzzy RD

- The **second stage** regression is

$$Y_i = \alpha_2 + \delta \hat{D}_i + f(x_i - c) + \hat{D}_i \times g(x_i - c) + \eta_{2i}$$

- The **reduced form** regression in FRD is

$$Y_i = \alpha_3 + \beta Z_i + f(x_i - c) + Z_i \times g(x_i - c) + \eta_{3i}$$

- You can also add covariates in every equations to making further controls.

Fuzzy RDD

- Still 2 types of strategies for correctly specifying the functional form in a FRD:
 - 1 **Parametric**/global method:
 - 2 **Nonparametric**/local method

Application: Air pollution in China

- Chen et al(2013),“Evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River policy”,PNSA,vol.110,no.32.
- Ebenstein et al(2017),“New evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River Policy”,PNSA,vol.114,no.39.
- Topic: Air pollution and Health
- A Simple OLS regression

$$Health_i = \beta_0 + \beta_1 Air\ pollution_i + \gamma X_i + u_i$$

- Potential bias?

Application: Air pollution in China

- More elegant method: SRD and FRD in Geography
- Natural experiment: “Huai River policy” in China
- Result:
 - Life expectancies (预期寿命) are about **5.5** year lower in the north owing to an increased incidence of cardiorespiratory(心肺) mortality.
 - the PM₁₀ is the causal factor to shorten lifespans and an additional $10 \mu\text{g}/\text{m}^3$ PM10 reduces life expectancy by **0.86** years.

Application: Air pollution in China



Fig. 1. The cities shown are the locations of the Disease Surveillance Points. Cities north of the solid line were covered by the home heating policy.

Application: Air pollution in China

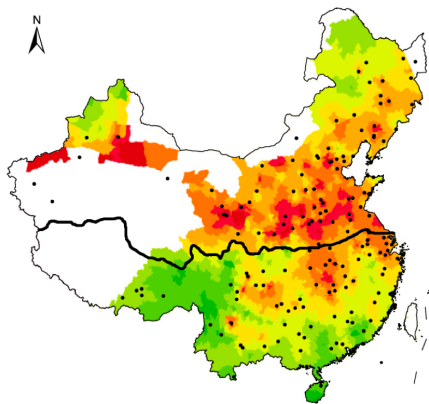


Fig. 1. China's Huai River/Qinling Mountain Range winter heating policy line and PM_{10} concentrations. Black dots indicate the DSP locations. Coloring corresponds to interpolated PM_{10} levels at the 12 nearest monitoring stations, where green, yellow, and red indicate areas with relatively low, moderate, and high levels of PM_{10} , respectively. Areas left in white are not within an acceptable range of any station.

Application: Air pollution in China: Chen et al(2013)

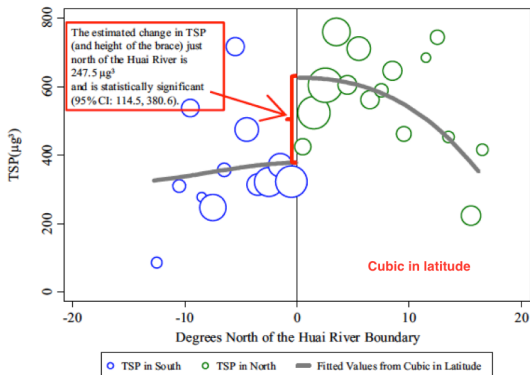


Fig. 2. Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. The plotted line reports the fitted values from a regression of TSPs on a cubic polynomial in latitude using the sample of DSP locations, weighted by the population at each location.

Application: Air pollution in China:Chen et al(2013)

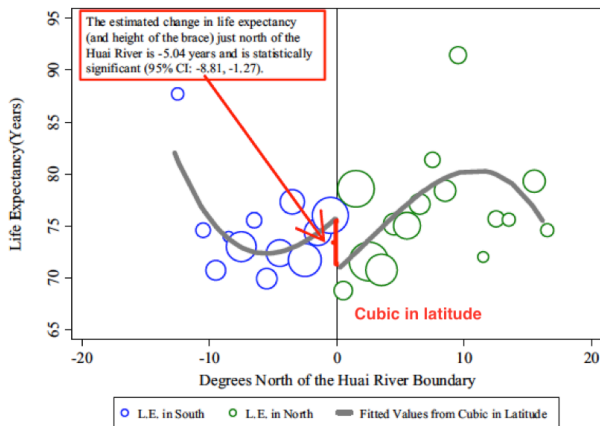


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Application: Air pollution in China: Chen et al(2013)

Table 2. Impact of TSPs ($100 \mu\text{g}/\text{m}^3$) on health outcomes using conventional strategy (ordinary least squares)

Dependent variable	(1)	(2)
ln(All cause mortality rate)	0.03* (0.01)	0.03** (0.01)
ln(Cardiorespiratory mortality rate)	0.04** (0.02)	0.04** (0.02)
ln(Noncardiorespiratory mortality rate)	0.01 (0.02)	0.01 (0.02)
Life expectancy, y	-0.54** (0.26)	-0.52** (0.23)
Climate controls	No	Yes
Census and DSP controls	No	Yes

$n = 125$. Each cell in the table represents the coefficient from a separate regression, and heteroskedastic-consistent SEs are reported in parentheses. The cardiorespiratory illnesses are heart disease, stroke, lung cancer and other respiratory illnesses. The noncardiorespiratory-related illnesses are violence, cancers other than lung, and all other causes. Models in column (2) include demographic controls and climate controls reported in Table 1. Regressions are weighted by the population at the DSP location. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Application: Air pollution in China: Chen et al(2013)

Table 2. Impact of TSPs ($100 \mu\text{g}/\text{m}^3$) on health outcomes using conventional strategy (ordinary least squares)

Dependent variable	(1)	(2)
ln(All cause mortality rate)	0.03* (0.01)	0.03** (0.01)
ln(Cardiorespiratory mortality rate)	0.04** (0.02)	0.04** (0.02)
ln(Noncardiorespiratory mortality rate)	0.01 (0.02)	0.01 (0.02)
Life expectancy, y	-0.54** (0.26)	-0.52** (0.23)
Climate controls	No	Yes
Census and DSP controls	No	Yes

$n = 125$. Each cell in the table represents the coefficient from a separate regression, and heteroskedastic-consistent SEs are reported in parentheses. The cardiorespiratory illnesses are heart disease, stroke, lung cancer and other respiratory illnesses. The noncardiorespiratory-related illnesses are violence, cancers other than lung, and all other causes. Models in column (2) include demographic controls and climate controls reported in Table 1. Regressions are weighted by the population at the DSP location. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Application: Air pollution in China: Chen et al (2013)

- Sharp RDD

$$Y_j = \delta_0 + \delta_1 N_j + \delta_2 f(L_j) + X_j' \phi + u_j$$

Table 3. Using the Huai River policy to estimate the impact of TSPs (100 $\mu\text{g}/\text{m}^3$) on health outcomes

Dependent variable	(1)	(2)	(3)
Panel 1: Impact of "North" on the listed variable, ordinary least squares			
TSPs, 100 $\mu\text{g}/\text{m}^3$	2.48*** (0.65)	1.84*** (0.63)	2.17*** (0.66)
ln(All cause mortality rate)	0.22* (0.13)	0.26* (0.13)	0.30* (0.15)
ln(Cardiorespiratory mortality rate)	0.37** (0.16)	0.38** (0.16)	0.50*** (0.19)
ln(Noncardiorespiratory mortality rate)	0.00 (0.13)	0.08 (0.13)	0.00 (0.13)
Life expectancy, y	-5.04** (2.47)	-5.52** (2.39)	-5.30* (2.85)
Panel 2: Impact of TSPs on the listed variable, two-stage least squares			
ln(All cause mortality rate)	0.09* (0.05)	0.14** (0.07)	0.14* (0.08)
ln(Cardiorespiratory mortality rate)	0.15** (0.06)	0.21** (0.09)	0.23** (0.10)
ln(Noncardiorespiratory mortality rate)	0.00 (0.05)	0.04 (0.07)	0.00 (0.06)
Life expectancy, y	-2.04** (0.92)	-3.00** (1.33)	-2.44 (1.50)
Climate controls	No	Yes	Yes
Census and DSP controls	No	Yes	Yes
Polynomial in latitude	Cubic	Cubic	Linear
Only DSP locations within 5° latitude	No	No	Yes

The sample in columns (1) and (2) includes all DSP locations ($n = 125$) and in column (3) is restricted to DSP locations within 5° latitude of the Huai River boundary ($n = 69$). Each cell in the table represents the coefficient from a separate regression, and heteroskedastic-consistent SEs are reported in parentheses. Models in column (1) include a cubic in latitude. Models in column (2) additionally include demographic and climate controls reported in Table 1. Models in column (3) are estimated with a linear control for latitude. Regressions are weighted by the population at the DSP location. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Application: Air pollution in China: Chen et al (2013)

- Fuzzy RDD
 - First Stage:

$$TSP_j = \alpha_0 + \alpha_1 N_j + \alpha_2 f(L_j) + X_j' \kappa + v_j$$

- Second Stage:

$$Y_j = \beta_0 + \beta_1 \widehat{TSP}_j + \beta_2 f(L_j) + X_j' \gamma + \varepsilon_j$$

Table 3. Using the Huai River policy to estimate the impact of TSPs (100 $\mu\text{g}/\text{m}^3$) on health outcomes

Dependent variable	(1)	(2)	(3)
Panel 1: Impact of "North" on the listed variable, ordinary least squares			
TSPs, 100 $\mu\text{g}/\text{m}^3$	2.48*** (0.65)	1.84*** (0.63)	2.17*** (0.66)
ln(All cause mortality rate)	0.22* (0.13)	0.26* (0.13)	0.30* (0.15)
ln(Cardiorespiratory mortality rate)	0.37** (0.16)	0.38** (0.16)	0.50*** (0.19)
ln(Noncardiorespiratory mortality rate)	0.00 (0.13)	0.08 (0.13)	0.00 (0.13)
Life expectancy, y	-5.04** (2.47)	-5.52** (2.39)	-5.30* (2.85)
Panel 2: Impact of TSPs on the listed variable, two-stage least squares			
ln(All cause mortality rate)	0.09* (0.05)	0.14** (0.07)	0.14* (0.08)
ln(Cardiorespiratory mortality rate)	0.15** (0.06)	0.21** (0.09)	0.23** (0.10)
ln(Noncardiorespiratory mortality rate)	0.00 (0.05)	0.04 (0.07)	0.00 (0.06)
Life expectancy, y	-2.04** (0.92)	-3.00** (1.33)	-2.44 (1.50)
Climate controls	No	Yes	Yes
Census and DSP controls	No	Yes	Yes
Polynomial in latitude	Cubic	Cubic	Linear
Only DSP locations within 5° latitude	No	No	Yes

The sample in columns (1) and (2) includes all DSP locations ($n = 125$) and in column (3) is restricted to DSP locations within 5° latitude of the Huai River

Air pollution in China: Ebenstein et al(2017)

- More accurate measures of pollution particles(PM_{10})
- More accurate measures of mortality from a more recent time period(2004-2012)
- More samples size(eight times than previous one)
- More subtle functional form: Local Linear Regression

Air pollution in China: Ebenstein et al(2017)

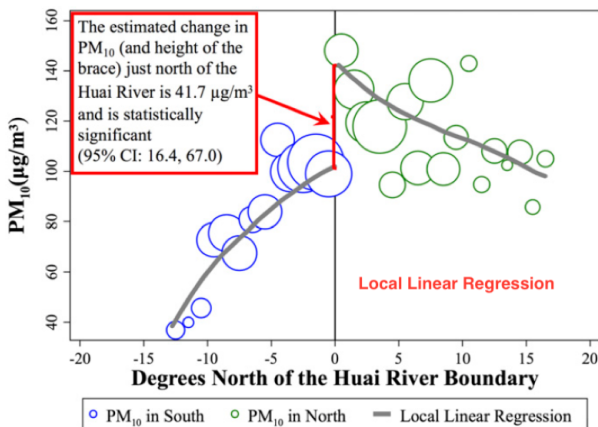


Fig. 2. Fitted values from a local linear regression of PM_{10} exposure on distance from the Huai River estimated separately on each side of the river.

Air pollution in China: Ebenstein et al(2017)

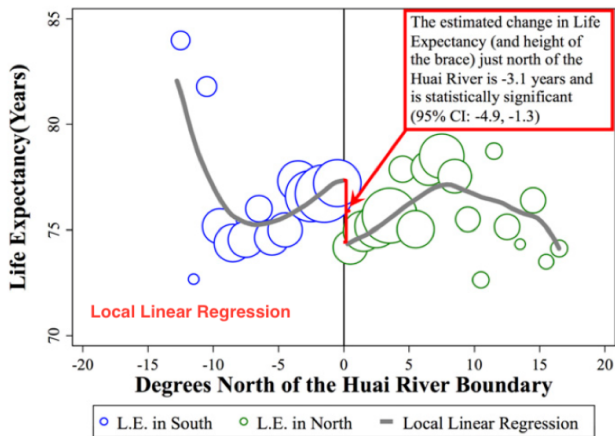


Fig. 3. Fitted values from a local linear regression of life expectancy (L.E.) on distance from the Huai River estimated in the same manner as in Fig. 2.

Air pollution in China: Ebenstein et al(2017)

- Sharp RD

$$Y_j = \delta_0 + \delta_1 N_j + f(L_j) + N_j f(L_j) + X_j' \phi + u_j$$

- Fuzzy RD

- First Stage

$$PM_j^10 = \alpha_0 + \alpha_1 N_j + f(L_j) + N_j f(L_j) + X_j' \gamma + u_j$$

- Second Stage

$$Y_j = \beta_0 + \beta_1 \widehat{PM_j^10} + f(L_j) + N_j f(L_j) + X_j' \phi + \varepsilon_j$$

Air pollution in China: Ebenstein et al(2017)

Table 2. RD estimates of the impact of the Huai River Policy

Outcome	[1]	[2]	[3]
Pollution and life expectancy			
PM ₁₀	27.4*** (9.5)	31.8*** (9.1)	41.7*** (12.9)
Life expectancy at birth, y	-2.4** (1.0)	-2.2* (1.1)	-3.1*** (0.9)
Cause-specific mortality (per 100,000, log)			
Cardiorespiratory	0.30** (0.14)	0.22* (0.13)	0.37*** (0.11)
Noncardiorespiratory	0.06 (0.10)	0.08 (0.09)	0.13 (0.08)
RD type	Polynomial	Polynomial	LLR
Polynomial function	Third	Linear	
Sample	All	5°	

Column [1] reports OLS estimates of the coefficient on a north of the Huai River dummy after controlling for a polynomial in distance from the Huai River interacted with a north dummy using the full sample ($n = 154$) and the control variables from [SI Appendix, Table S1](#). Column [2] reports this estimate for the restricted sample ($n = 79$) of DSP locations within 5° of the Huai River. Column [3] presents estimates from local linear regression (LLR), with triangular kernel and bandwidth selected by the method proposed by Imbens and Kalyanaraman (14).

Air pollution in China: Ebenstein et al(2017)

Table 2. RD estimates of the impact of the Huai River Policy

Outcome	[1]	[2]	[3]
Pollution and life expectancy			
PM ₁₀	27.4*** (9.5)	31.8*** (9.1)	41.7*** (12.9)
Life expectancy at birth, y	-2.4** (1.0)	-2.2* (1.1)	-3.1*** (0.9)
Cause-specific mortality (per 100,000, log)			
Cardiorespiratory	0.30** (0.14)	0.22* (0.13)	0.37*** (0.11)
Noncardiorespiratory	0.06 (0.10)	0.08 (0.09)	0.13 (0.08)
RD type	Polynomial	Polynomial	LLR
Polynomial function	Third	Linear	
Sample	All	5°	

Column [1] reports OLS estimates of the coefficient on a north of the Huai River dummy after controlling for a polynomial in distance from the Huai River interacted with a north dummy using the full sample ($n = 154$) and the control variables from [SI Appendix, Table S1](#). Column [2] reports this estimate for the restricted sample ($n = 79$) of DSP locations within 5° of the Huai River. Column [3] presents estimates from local linear regression (LLR), with triangular kernel and bandwidth selected by the method proposed by Imbens and Kalyanaraman (14).

Implement of RDD

Three Steps

- 1 Graph the data for visual inspection
- 2 Estimate the treatment effect using regression methods
- 3 Run checks on assumptions underlying research design

RDD graphical analysis

- First, divide X into bins, making sure no bin contains c as an interior point

- if x ranges between 0 and 10 and $c = 5$, then you could construct 10 bins:

$$[0, 1), [1, 2), \dots, [9, 10]$$

- if $c = 4.5$, you may use 20 bins, such as

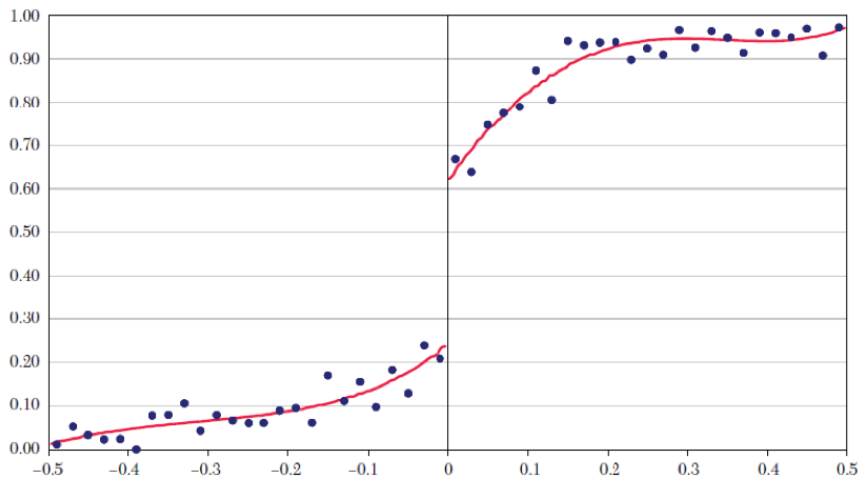
$$[0, 0.5), [0.5, 1), \dots, [9.5, 10]$$

- Second, calculate average y in each bin, and plot this above midpoint for each bin.
- Third, plot the forcing variable X_i on the horizontal axis and the average of Y_i for each bin on the vertical axis. (Note: You may look at different bin sizes)
- Fourth, plot predicted line of Y_i from a flexible regression
- Fifth, inspect whether there is a discontinuity at c and there are other unexpected discontinuities.

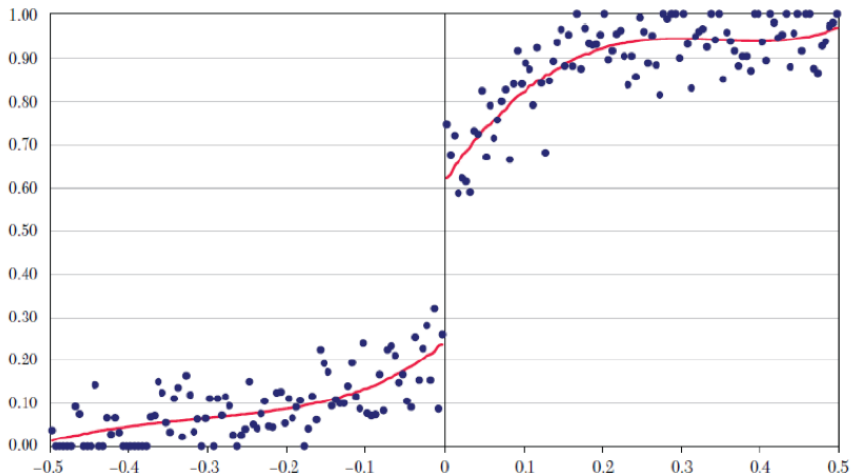
RDD graphical analysis: Select Bin Width

- What is optimal # of bins (i.e. bin width)?
- Choice of bin width is subjective because of tradeoff between precision and bias
 - By including more data points in each average, wider bins give us more precise estimate.
 - But, wider bins might be biased if $E[y|x]$ is not constant within each of the wide bins.
- Sometimes software can help us.

Graphical Analysis in RD Designs: different bin size



Graphical Analysis in RD Designs: different bin size



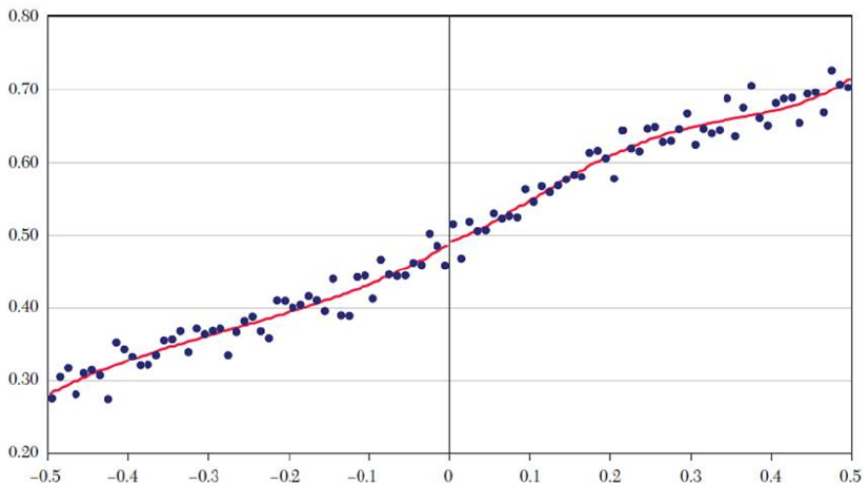
Estimate the treatment effect using regression methods

- It is probably advisable to report results for both estimation types:
- Polynomials in X .
 - In robustness checks you also want to show that including higher order polynomials does not substantially affect your findings.
 - But quadratic(at most Cubic) is enough,higher-order polynomial may hurt and should not be used.(Gelman and Imbens,2019)
- Local linear regression or other nonparametric estimation
 - Your results are not affected if you vary the window(bandwidth)around the cutoff.
 - Standard errors may go up but hopefully the point estimate does not change.

Testing the Validity of the RDD

- ④ Test involving covariates(Nonoutcome Variable):
 - Test whether other covariates exhibit a jump at the discontinuity. (Just re-estimate the RD model with the covariate as the dependent variable).
 - Construct a similar graph to the one before but using a covariate as the “outcome”.
 - There should be no jump in other covariates

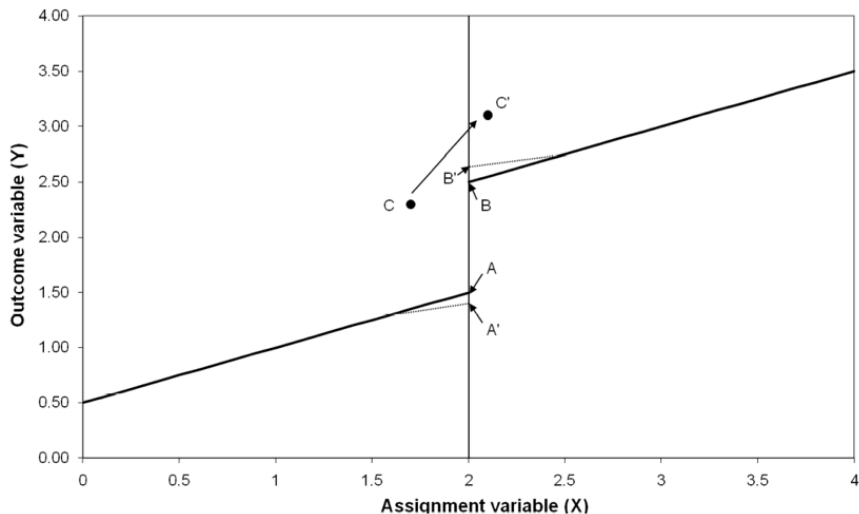
Graphical: Example Covariates by Forcing Variable



Testing the Validity of the RDD

- ② Test sorting behavior
 - Individuals may invalidate the continuity assumption if they strategically **manipulate assignment variable X** to be just above or below the cutoff
 - Recall a key assumption of RD is that agents can **NOT perfect** control over the assignment variable X.
 - That is, people just above and just below the cutoff are no longer comparable.

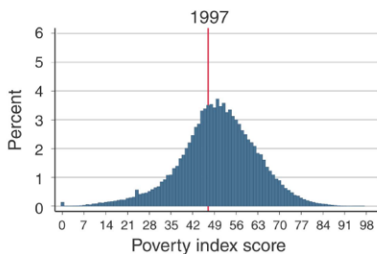
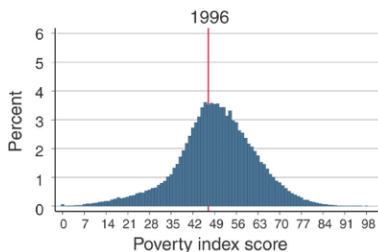
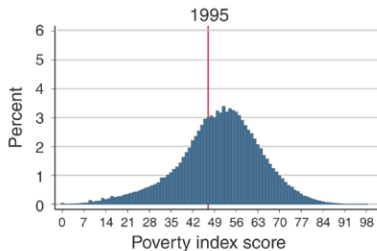
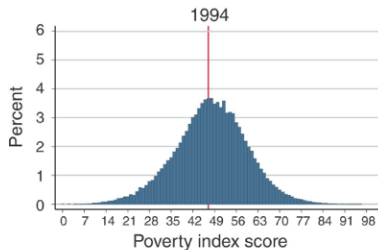
Sorting behavior



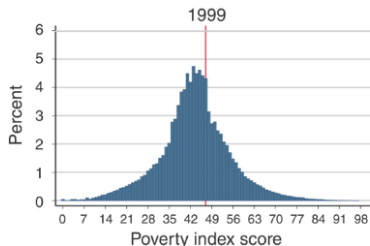
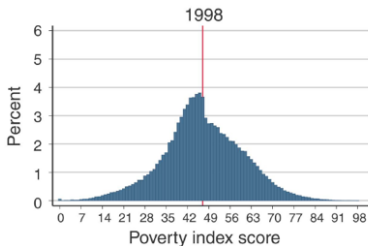
Manipulation of a poverty index in Colombia

- Adriana Camacho and Emily Conover (2011) “Manipulation of Social Program Eligibility” AEJ: Economic Policy
- 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: Manipulation of a poverty index in Colombia



Manipulation of a poverty index in Colombia



Testing the Validity of the RDD

- Testing discontinuity in the density of assignment variable X
 - 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.
- Also a more formal test: McCrary(2008) test.

Testing the Validity of the RDD

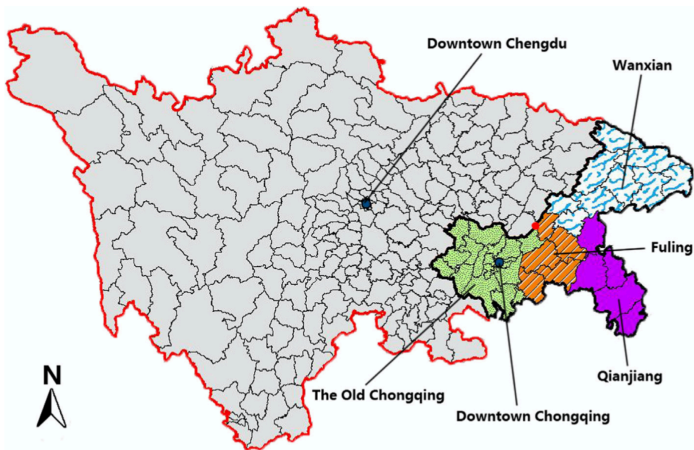
- **Falsification Tests I:** testing for jumps at non-discontinuity points
- If threshold x only existed in certain c or for certain types of observations...
- Make sure no effect in c where there was no discontinuity or for agents where there isn't supposed to be an effect.

A Latest Case

Jia,Liang and Ma(2021)

- “Political hierarchy and regional economic development: Evidence from a spatial discontinuity in China”, Journal of Public Economics Volume 194, February 2021
- Topic: Political hierarchy and regional economic development
- Background: In 1997, the prefecture-level Chongqing city was elevated to a province-level municipality, splitting off from Sichuan province.
- It consequently gained a substantial increase in decision-making power for administrative, personnel, and fiscal affairs.

Chongqing v.s Sichuan



- Chongqing contains 30 million people in 43 counties and 933 towns
- The remaining Sichuan contains 85 million people in 20 prefectures,

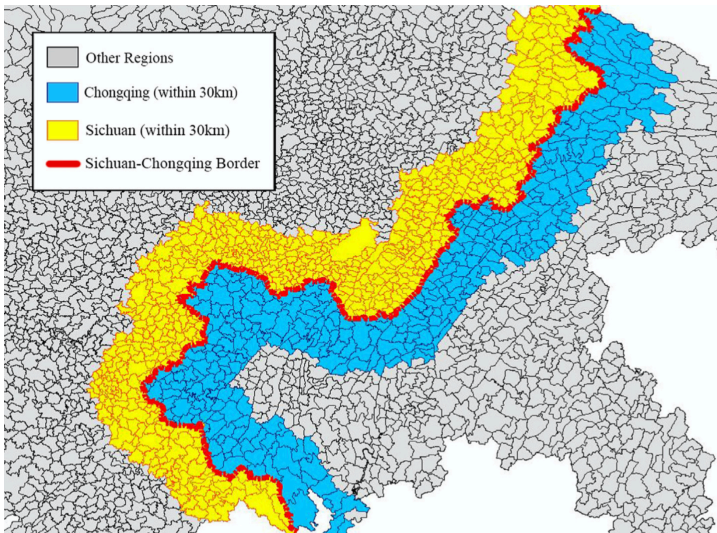
Empirical Strategy

- Regression Equation

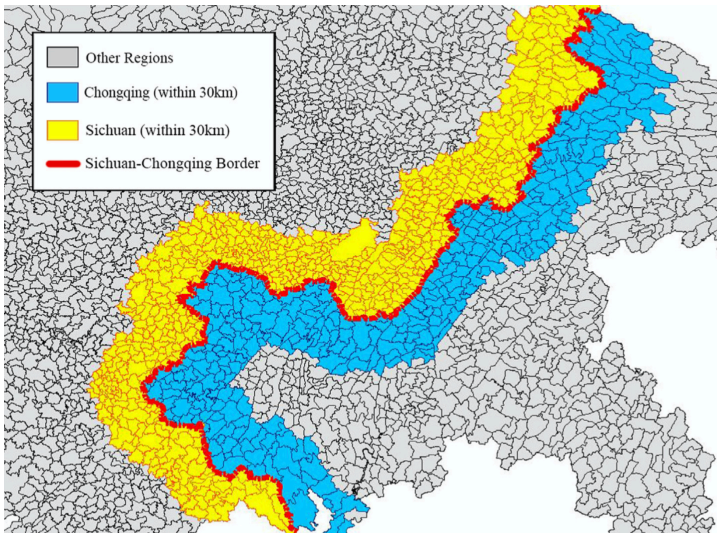
$$Y_i = \beta_0 + \beta_1 \text{Chongqing}_i + f(L) + \varepsilon_i$$

- Y_i as outcome variable of interest in town i , thus the economic growth rate.
- *Chongqing* is a binary indicator.
- $f(L)$ control for a two-dimensional polynomial in a town centroid's longitude and latitude.
- β_1 captures the Chongqing promotion treatment effect on economic growth

Empirical Strategy



Empirical Strategy



Pretreatment Balance

Table 2
Balance test.

	Chongqing	Sichuan	Mean Difference (s.e.)
	(1)	(2)	(3)
Mean values			
<i>Panel A. town-level variables within 30 km bandwidth</i>			
Light intensity in 1996	0.704	0.764	-0.060 (0.218)
Elevation (meter)	505.530	458.497	47.033 (80.209)
Slope (%)	9.205	8.310	0.894 (2.001)
Distance to Chongqing Downtown (km)	126.794	142.989	-16.196 (25.513)
Distance to Chengdu Downtown (km)	285.987	254.320	31.667 (31.663)
Ethnic minority population share	0.003	0.012	-0.009 (0.011)
Observations	279	467	
<i>Panel B. county-level variables for full sample</i>			
Per capita GDP in 1996 (yuan, in logarithm)	8.098	8.012	0.086 (0.100)
Per capita industrial output in 1996 (yuan, in logarithm)	7.580	7.346	0.235 (0.201)
Per capita fiscal revenue in 1996 (yuan, in logarithm)	4.722	4.713	0.009 (0.099)
Urbanization rate in 1996 (%)	78.495	81.663	-3.168 (3.006)
Observations	43	178	

Notes: *, **, and *** indicate statistical significance at 10%, 5%, and 1% levels, respectively. The county-level clustered standard errors are reported in parentheses

Pretreatment Discontinuity

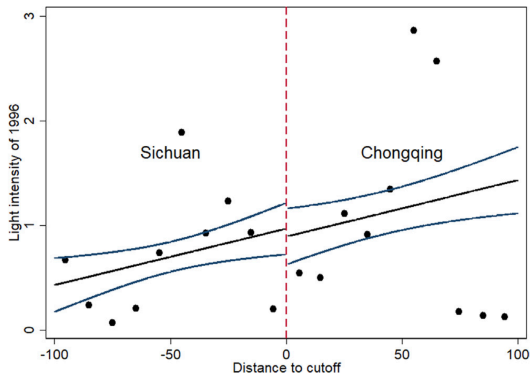


Fig. 3. Balance of initial development level across the border. *Notes:* This figure shows the single dimension RD graphs. The x-axis denotes the distance from a town centroid to the Chongqing–Sichuan border, where negative numbers refer to the control group (Sichuan). The dark dots show growth rates averaged over 10 km wide bins in distance from the border. The black lines fit local linear regressions within 100 km bandwidth on both sides of the boundary and the blue lines denote 95 percent confidence interval.

Baseline Discontinuity

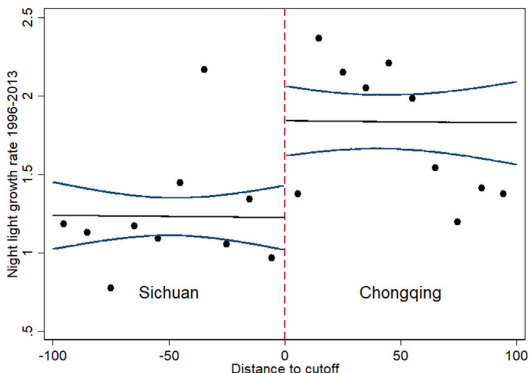


Fig. 5. Discontinuity in growth rate of light intensity from 1996 to 2013. *Notes:* The x-axis denotes the distance from a town centroid to the Chongqing–Sichuan border, where negative numbers refer to the control group (Sichuan). The dark dots show growth rates averaged over 10 km wide bins in distance from the border. The black lines fit local linear regressions within 100 km bandwidth on both sides of the boundary, and the blue lines denote 95 percent confidence interval.

Baseline results

Table 3
Baseline RD results.

Sample within	Dependent variable: light intensity growth from 1996-2013					
	Local linear approach		Local quadratic approach		Global polynomial approach	
	<30 km (1)	<50 km (2)	<30 km (3)	<50 km (4)	Full Sample (5)	Full Sample (6)
Chongqing	1.038*** (0.291)	1.170*** (0.295)	1.028*** (0.287)	1.199*** (0.308)	1.036*** (0.234)	1.022*** (0.239)
Polynomial	Linear	Linear	Quadratic	Quadratic	Cubic	Quartic
Observations	746	1,188	746	1,188	5,088	5,088
R-squared	0.104	0.087	0.117	0.094	0.034	0.033

Notes: The dependent variable is $\ln(0.01 + LightIntensity_{i,2013}) - \ln(0.01 + LightIntensity_{i,1996})$. All regressions include two-dimensional geographic controls. The county-level clustered standard errors are reported in parentheses. *, **, and *** indicate statistical significance at 10%, 5%, and 1% levels, respectively.

Parallel Trends and Dynamic effects

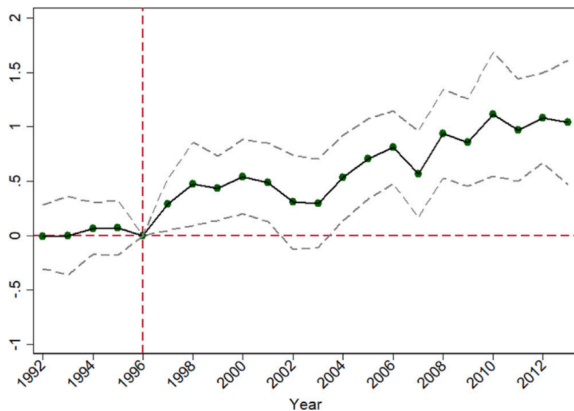


Fig. 4. Dynamics of the effects on light intensity growth. *Notes:* Point estimates are reported under alternative time windows. The basic line is for 1996. The solid line plots the point estimate of a separate estimation of β_1 in Eq. (1) and the dash lines denote 95 percent confidence interval.

Robustness checks

- alternative bandwidths and specifications

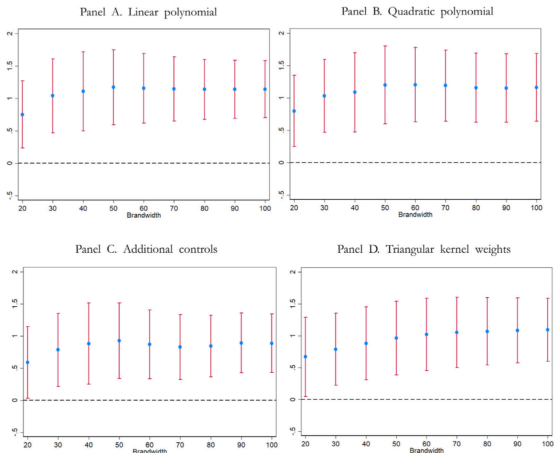


Fig. 6. Robustness to alternative bandwidths and model specifications. Notes: Each point plots the point estimate of a separate estimation of β_1 in Eq. (1) along with the 95 percent confidence interval, ranging from 20-km to 100-km bandwidths. Panel A plots estimates using linear polynomials in latitude and longitude. Panel B plots estimates from equivalent regressions but using second-order polynomials in latitude and longitude. Panel C plots estimates additionally controlling for elevation, slope, and a series of segment dummies. Panel D shows results using triangular kernel weights to give higher weight to observations that are closer to the boundary.

Placebo tests

Table 4

Placebo tests.

Sample within	Dependent variable: light intensity growth from 1996-2013	
	Move the true boundary 30 kilometers westward (1)	Move the true boundary 30 kilometers eastward (2)
East of the falsified border	-0.086 (0.249)	0.256 (0.404)
Observations	881	517
R-squared	0.069	0.068

Notes: The dependent variable is $\ln(0.01 + \text{LightIntensity}_{i,2013}) - \ln(0.01 + \text{LightIntensity}_{i,1996})$. We set a 30 km bandwidth. All regressions include two-dimensional geographic controls. The county-level clustered standard errors are reported in parentheses. *, **, and *** indicate statistical significance at 10%, 5%, and 1% levels, respectively.

Displacement effects

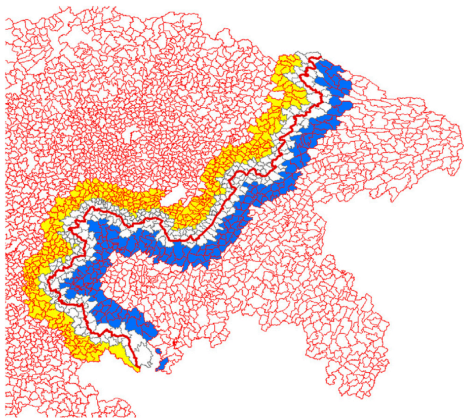


Fig. 8. Spatial exclusion approach. *Notes:* The red line marks the border between Sichuan and Chongqing. The blue and yellow shaded areas are towns in our boundary sample that belong to treated and non-treated areas, respectively. We eliminate towns within 10 kilometers of the boundary, which are represented by the blank area.

Displacement effects

Table 5
Test on displacement effects.

Dependent variable: light intensity growth from 1996-2013						
Sample within	<30 km			<50 km		
	Baseline (1)	Exclude towns within 2*10 km across boundary (2)	Exclude towns within 2*5km across boundary (3)	Baseline (4)	Exclude towns within 2*10 km across boundary (5)	Exclude towns within 2*5km across boundary (6)
Chongqing	1.038*** (0.291)	1.418*** (0.342)	1.199*** (0.275)	1.170*** (0.295)	1.506*** (0.314)	1.315*** (0.276)
Observations	746	476	622	1,188	918	1,064
R-squared	0.104	0.126	0.113	0.087	0.095	0.091
Test on equality with the baseline estimate		p = 0.2731	p = 0.5609		p = 0.2886	p = 0.6012

Notes: The dependent variable is $\ln(0.01 + LightIntensity_{i,2013}) - \ln(0.01 + LightIntensity_{i,1996})$. All regressions include two-dimensional local linear geographic controls. The last row reports the p-value of the Wald test on equality with the baseline estimate. The county-level clustered standard errors are reported in parentheses. *, **, and *** indicate statistical significance at 10%, 5%, and 1% levels, respectively.

Summary

RDD in the toolkit of Causal Inference

- It is so called the **nearest** method to RCT which identify causal effect of treatment on outcome.
- RDD needs a arbitrary cut-off and agents can **imperfect** manipulate the treatment.
- Two types
 - Sharp RD
 - Fuzzy RD
- Assumption: continued at the cut-off
- Concerns:
 - Functional form
 - Bandwidth selection
 - Bin selection