# Lecture 6: Regression Discontinuity Design

*Applied MicroEconometrics,Fall 2021*

**Zhaopeng Qu**

**Nanjing University Business School**

November 17 2021

# 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.

# Main Idea of Regression Discontinuity Design

- 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.

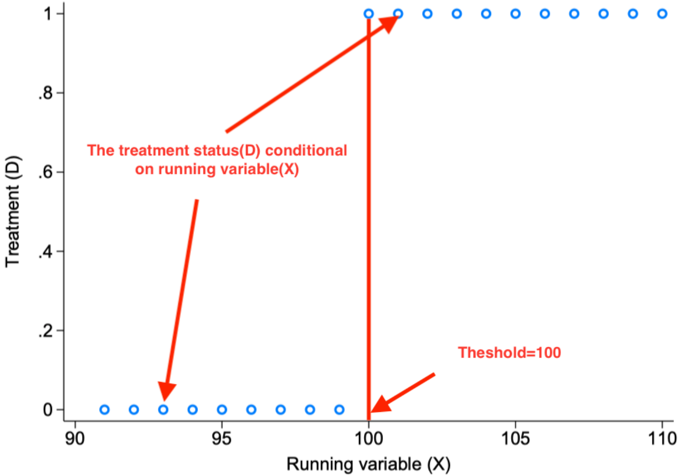
# A Motivating Example: Elite University

- Numerous studies have shown that graduates from more selective programs or schools 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.(Because?)

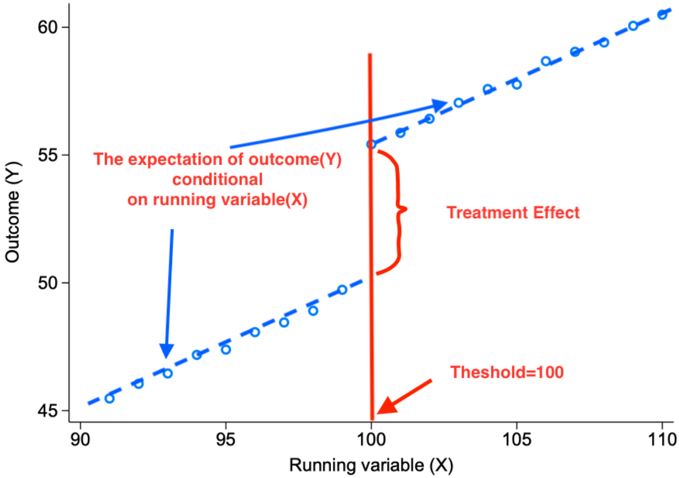
# A Motivating Example: Elite University

- 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 **100** for NJU.
- Those with scores **95** or even **99** are unlikely to attend NJU, instead attend NUFE(南京财经大学).
- Assume that those get *99* or *95* and those get *100* 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: Graphics



# Main Idea of RDD: Graphics



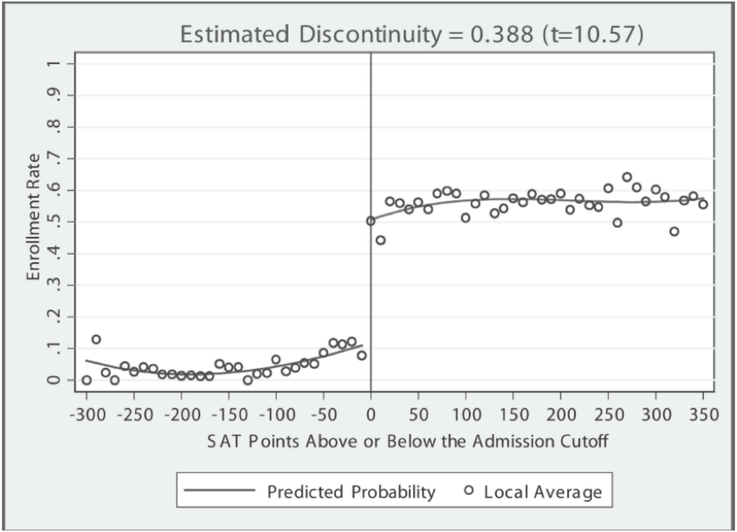
# A Motivating Example: Elite University

- Mark Hoekstra (2009) "The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach" Review of Economics and Statistics
- The paper demonstrates RD idea by examining the economic return of attending the most selective public state university.
- In the United States, most schools used SAT (or ACT) scores in their admission process.
- For example, the flagship state university considered here uses a strict cutoff based on SAT score and high school GPA.

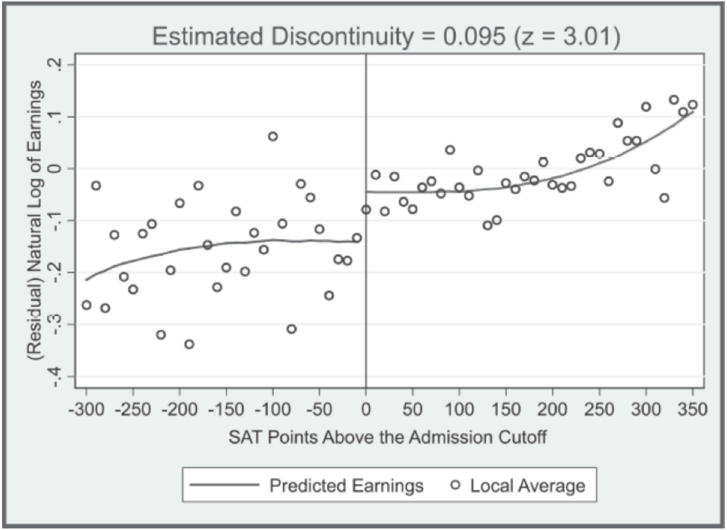
# A Motivating Example: Elite University

- For the sake of simplicity, we just focuses on the SAT score.
- The author is then able to match (using social security numbers) students applying to the flagship university in 1986-89 to their administrative earnings data for 1998 to 2005.

# SAT Score and Enrollment



# SAT Score and Earnings



# More Cases of RDD

- Academic test scores: scholarship, prize, higher education admission, certifications of merit.
- Poverty scores: (proxy-) means-tested anti-poverty programs(generally: any program targeting that features rounding or cutoffs)
- 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
- Graphically 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.
- Intuition: 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: Theory and Application

# 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

$$D_i = 1\, if\ X_i \geq c\ and\ D_i = 0\, if\ X_i < c$$

# RDD and Potential Outcomes:Notations

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

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

# Sharp RDD and Fuzzy RDD

- In general, depending on enforcement of treatment assignment, RDD can be categorized into two types:
  1. **Sharp RDD**: nobody below the cutoff gets the "treatment", everybody above the cutoff gets it
     - Everyone follows treatment assignment rule (all are compliers).
     - Local randomized experiment with perfect compliance around cutoff.
  2. **Fuzzy RDD**: the probability of getting the treatment jumps discontinuously at the cutoff (NOT jump from 0 to 1)
     - Not everyone follows treatment assignment rule.
     - Local randomized experiment with partial compliance around cutoff.
     - Using initial assignment as an instrument for actual treatment.

# Identification for Sharp RDD

- **Deterministic Assumption**

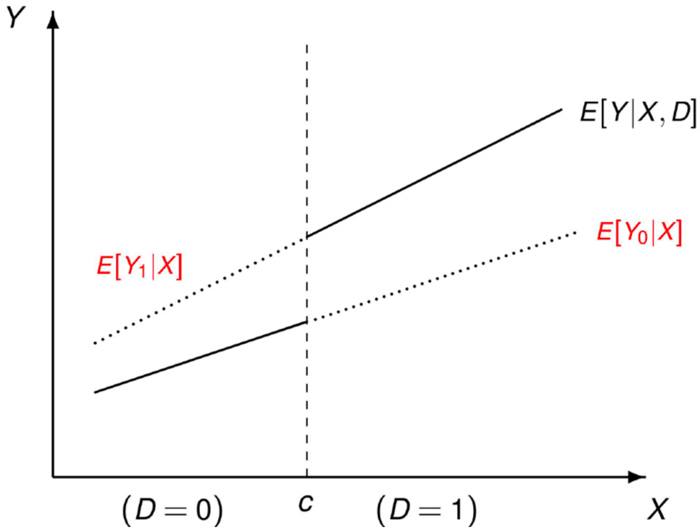
$$D_i = 1(X_i \geq c)$$

- Treatment assignment is a deterministic function of the assignment variable $X_i$ and the threshold $c$.

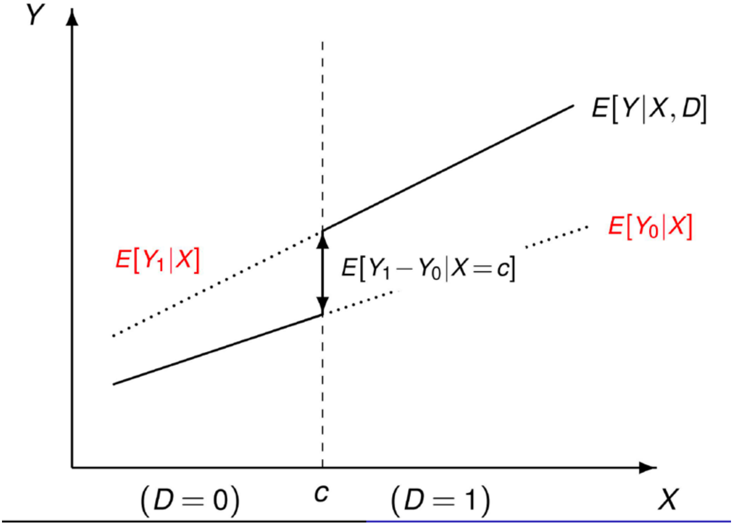
# Identification for Sharp RDD

- **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.

# Graphical Interpretation



# Graphical Interpretation



# Identification for Sharp RDD

- Intuitively, we are interested in the discontinuity in the outcome at the discontinuity in the treatment assignment.
- We can use sharp RDD to investigate the behavior of the outcome around the threshold

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

# 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.

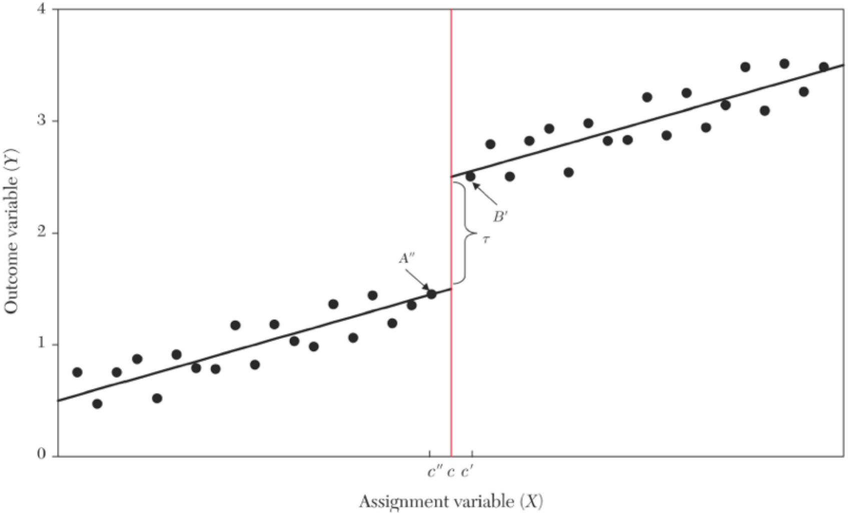
# Sharp RDD specification

- 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?

# Linear Specification

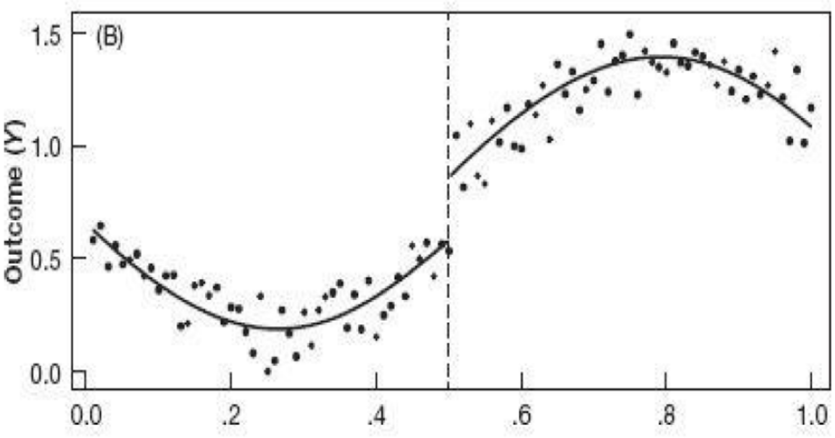


# Specification

- The validity of RD estimates depends crucially on the function forms, which should provide an adequate representation of $E[Y_{0i}|X]$ and $E[Y_{1i}|X]$
- If not what looks like a jump may simply be a non-linear in $f(X_i)$ that the polynomials have not accounted for.

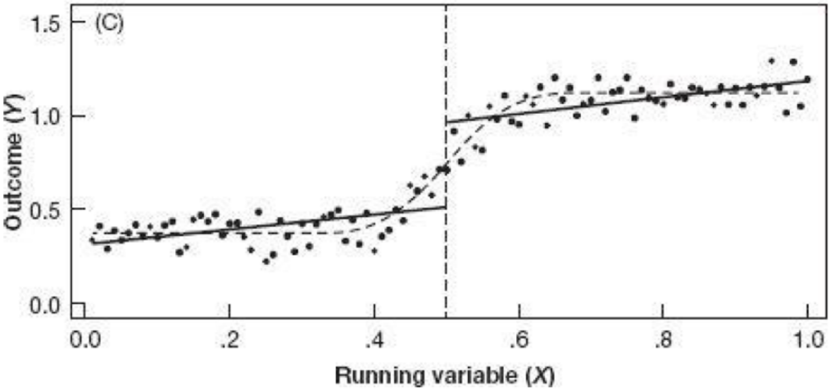
# Nonlinear Case

- What if the Conditional Expectation Function is **nonlinear**?



# Nonlinear Case

- The function form is very important in RDD.



# Sharp 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

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

# Parametric/Global method

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

$$Y_i = \alpha + \rho D_i + \beta_1 X_i + \beta_1 X_i^2 + ... + \beta_p X_i^P + \eta_i$$

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

# Parametric/Global Approach

- Let

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

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

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

- The treatment effect at c is $\rho$

# 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

# F-Test for Polynomial Functions

To implement F-Test, one can complete the following steps

1. Create a set of indicator variables for $K-2$ of the bins used to graphically depict the data.
2. Exclude any two of the bins to avoid having a model that is collinear which is default term as dummy variables.(normally they are the first and the last bins)
3. Add the set of bin dummies $B_k$ to the polynomial regression and jointly test the significance of the bin dummies

$$\begin{aligned} Y_i = \alpha + \rho D_i + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + ... + \beta_p \tilde{X}_i^p \\ + \beta_1^* D_i \tilde{X}_i + \beta_2^* D_i \tilde{X}_i^2 + ... + \beta_p^* D_i \tilde{X}_i^p \\ + \sum_{k=2}^{K-1} \phi_k B_k + ... + \varepsilon_i \end{aligned}$$

# F-Test for Polynomial Functions

To implement F-Test, one can complete the following steps(continued)

4.Test the null hypothesis

$$\phi_2 = \phi_2 = ... = \phi_{K-1} = 0$$

which means that there is no significant gap between any $B_k$ and $B_1$ or $B_K$

5.Repeat test for $p = 1, 2, ..., p$ until the F-test of coefficients($\phi$) before bin dummies $B_k$ are **no longer jointly significant**.

- In other words, the orders of polynomial function in this case *fit the data well*.

# Akaike information criterion (AIC) Approach

- Conceptually,AIC describes the trade-off between the goodness of fit of the model and the simplicity of the model.
- These two terms move in opposite directions as the model becomes more complex.
  - the estimated residual variance $\hat{\sigma}_b^2$ should decrease
  - the number of parameters $p$ used to increase
- In a regression context, the AIC is given by

$$AIC = Nln(\hat{\sigma}_b^2) + 2P$$

- The set of models are then ranked according to their AIC values, and the model with **the smallest AIC value** is deemed the optimal model among the set of candidates.

# Nonparametric/Local Approach

- Recall we can construct RD estimates by fitting

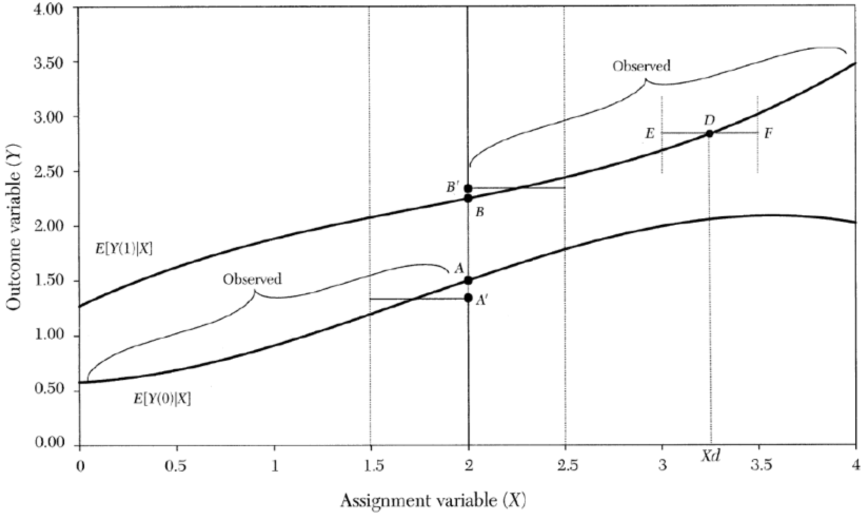
$$Y_i = \alpha + \rho D_i + f(x_i) + u_i$$

- 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 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: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.

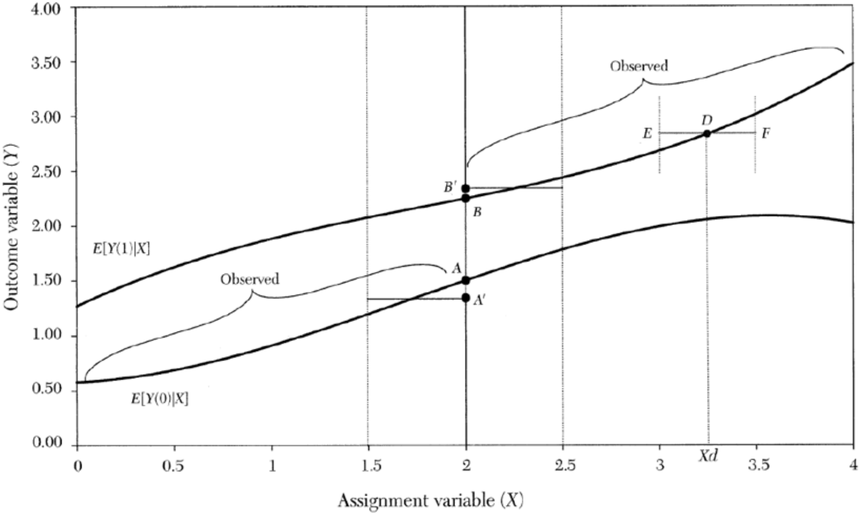
# Nonparametric/Local Approach

- The standard solution to reduce the boundary bias is to run **local linear regression**.It is a nonparametric method which is linear smoother within a given bandwidth (window) of width h around the threshold.
- Thus we estimate the following linear regression within a given window of width h around the cutoff:

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

- So the bandwidth is a key.

# Nonparametric/Local Approach:boundary bias



# 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 use more date points far from cutoff,then the estimated treatment effect could be biased.

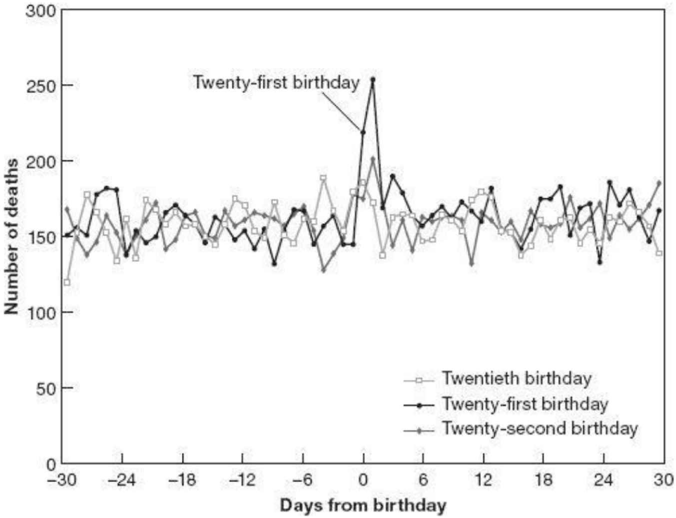
# How to Choose Bandwidth

- Cross-Validation Procedure:Choose bandwidth $h$ that produces the best fit for the relationship of outcome and assignment variable.
- Plug-In Procedure:
    - Solve for the optimal bandwidth formula in terms of minimizing mean square error.
    - Plug the parameters into formula to get optimal bandwidth.
- Usually, we would present the RD estimates by different choices of bandwidth.

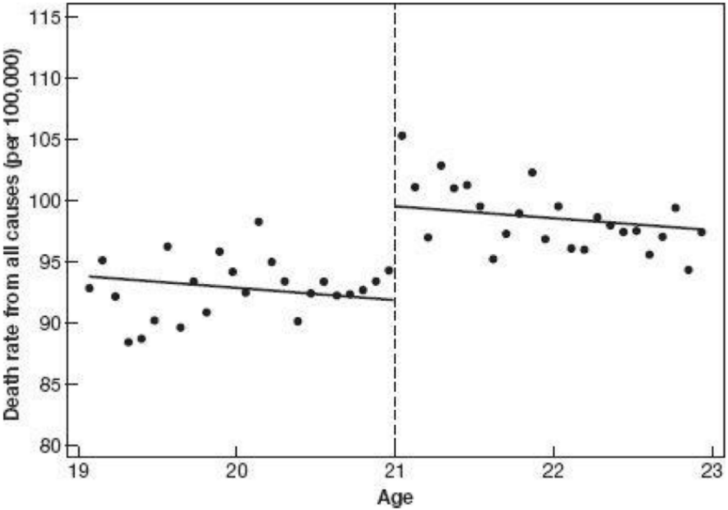
# Application: Effect of the Minimum Legal Drinking Age (MLDA) on death rates

- 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?

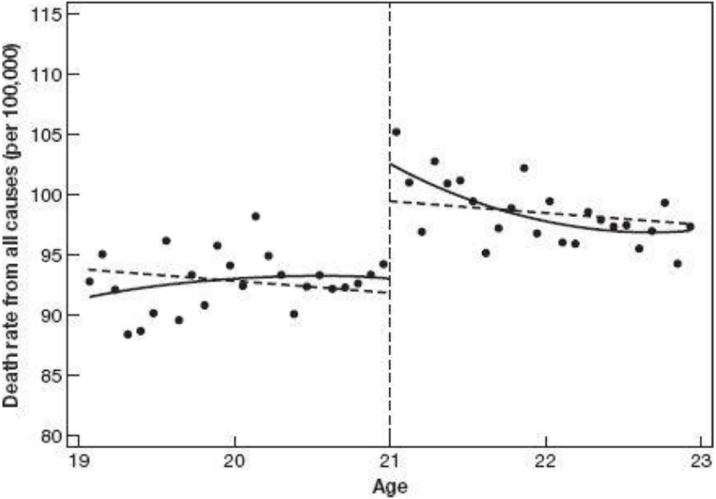
# Application: MLDA on death rates



# 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

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

- The effect of legal access to alcohol on mortality rate at age 21 is $\rho$

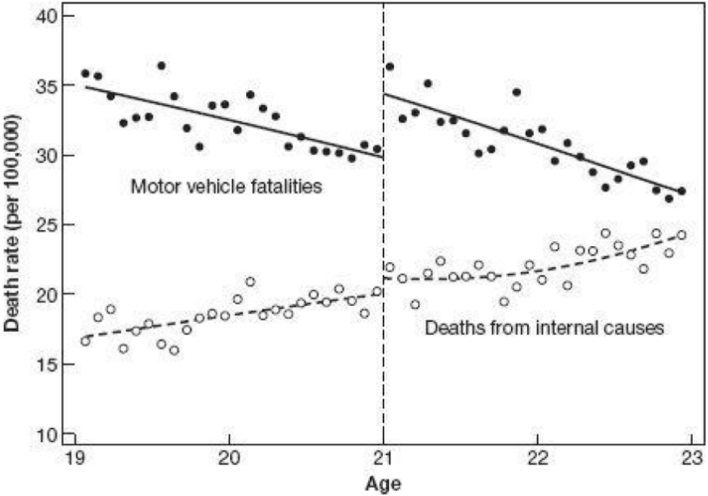
# Application: MLDA on death rates

TABLE 4—DISCONTINUITY IN LOG DEATHS AT AGE 21

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| *Deaths due to all causes* | | | | |
| Over 21 | 0.096 | 0.087 | 0.091 | 0.074 |
| | (0.018) | (0.017) | (0.023) | (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 | |
| *Deaths due to external causes* | | | | |
| Over 21 | 0.110 | 0.100 | 0.096 | 0.082 |
| | (0.022) | (0.021) | (0.028) | (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 | |
| *Deaths due to internal causes* | | | | |
| Over 21 | 0.063 | 0.054 | 0.094 | 0.066 |
| | (0.040) | (0.040) | (0.053) | (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 Nt 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 polyNmial 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:

$$P(D_i = 1|x_i) = p_1(X_i)\ if\ x_i \geq c, the\ probability\ assign\ to\ treatment$$
$$P(D_i = 1|x_i) = p_0(X_i)\ if\ x_i < c, the\ probability\ assign\ to\ control\ gro$$

- **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

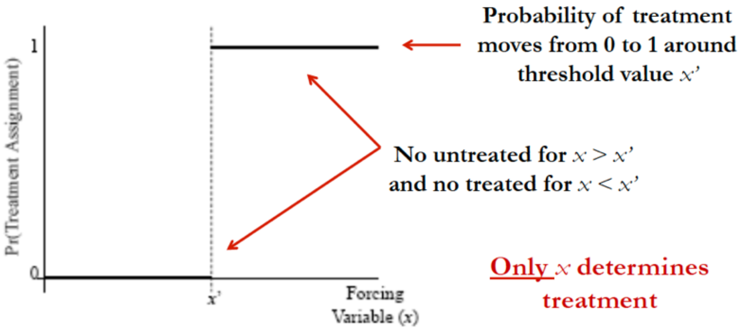



Figure is from Roberts and Whited (2010)

# Fuzzy RD v.s Sharp RD

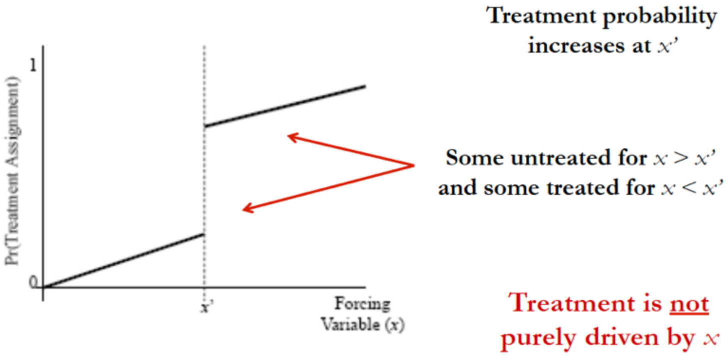



Figure is from Roberts and Whited (2010)

# 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 varaible**?
  - **instrumetal varaible**?

# 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 PM_10 is the causal factor to shorten lifespans and an additional 10 $\mu g/m^3$ PM10 reduces life expectancy by **0.86** years.

# Application: Air pollution in China

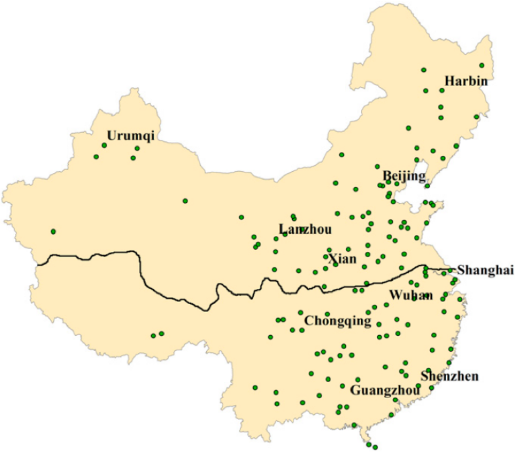



**Fig. 1.** The cities shown are the locations of the Disease Surveillance Points. Cities north of the solid line were covered by the home heating policy.

# Application: Air pollution in China

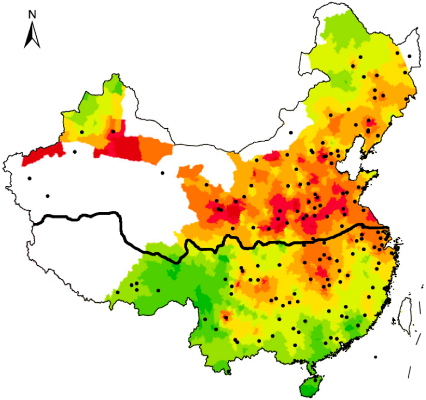



**Fig. 1.** China's Huai River/Qinling Mountain Range winter heating policy line and $PM_{10}$ concentrations. Black dots indicate the DSP locations. Coloring corresponds to interpolated $PM_{10}$ levels at the 12 nearest monitoring stations, where green, yellow, and red indicate areas with relatively 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)

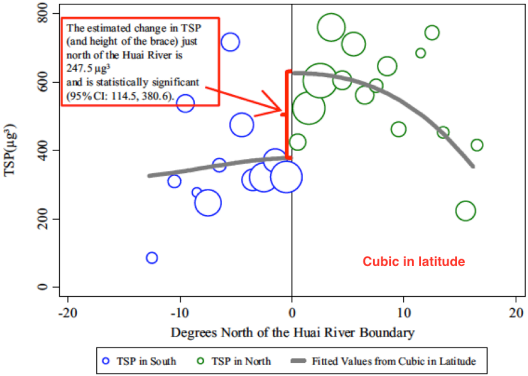



**Fig. 2.** Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. 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)

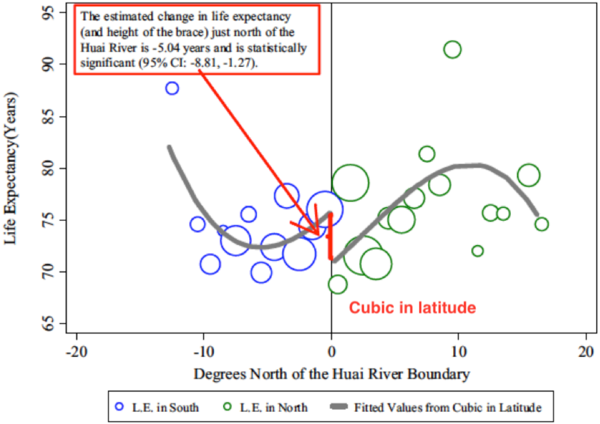



Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

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

**Table 2. Impact of TSPs (100 μg/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 μg/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 μg/m³) on health outcomes**

| Dependent variable | (1) | (2) | (3) |
|---|---|---|---|
| Panel 1: Impact of "North" on the listed variable, ordinary least squares | | | |
| TSPs, 100 μg/m³ | 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 μg/m$^3$) on health outcomes**

| Dependent variable | (1) | (2) | (3) |
|---|---|---|---|
| Panel 1: Impact of "North" on the listed variable, ordinary least squares | | | |
| TSPs, 100 μg/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)

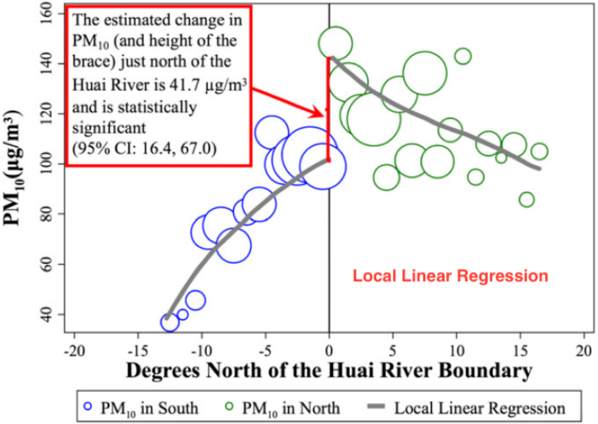



**Fig. 2.** Fitted values from a local linear regression of $PM_{10}$ exposure on distance from the Huai River estimated separately on each side of the river.

# Air pollution in China: Ebenstein et al(2017)

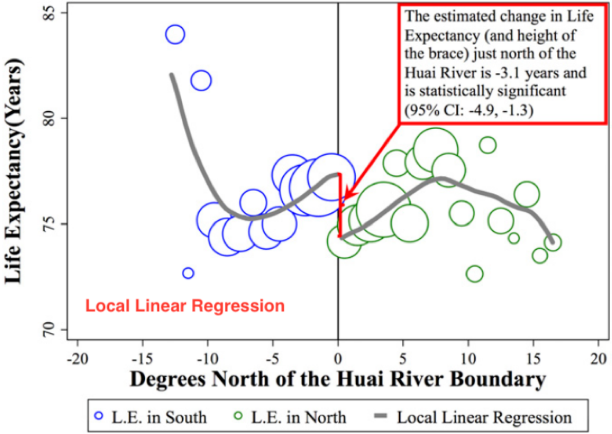



**Fig. 3.** Fitted values from a local linear regression of life expectancy (L.E.) on distance from the Huai River estimated in the same manner as in Fig. 2.

# 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

$$PM_j^1 0 = \alpha_0 + \alpha_1 N_j + f(L_j) + N_j f(L_j) + X_j' \gamma + u_j$$

  - Second Stage

$$Y_j = \beta_0 + \beta_1 \widehat{PM_j^1 0} + f(L_j) + N_j f(L_j) + X_j' \phi + \varepsilon_j$$

# 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

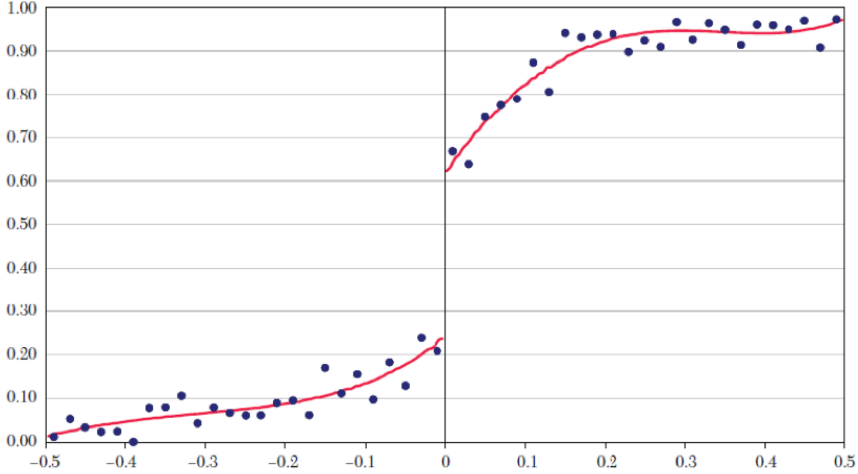
# RDD graphical analysis

- 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:
$$[0, 1), [1, 2), ..., [9, 10]$$
  - if $c = 4.5$, you may use 20 bins, such as
$$[0, 0.5), [0.5, 1), ..., [9.5, 10]$$
- 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.

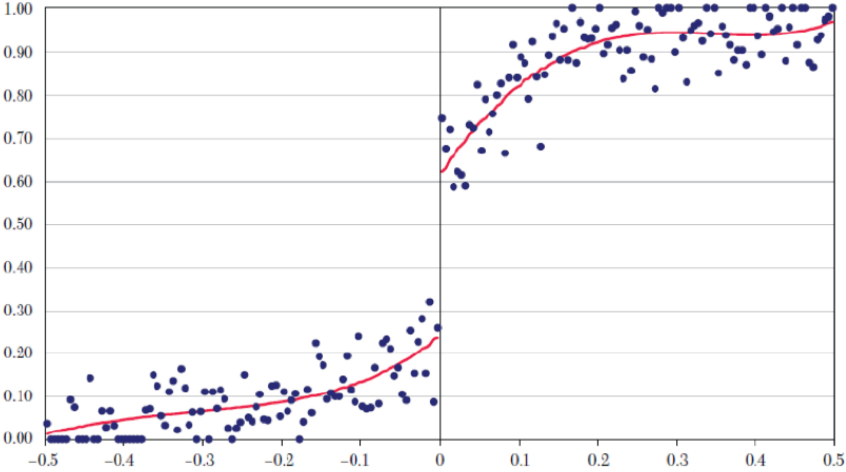
# RDD graphical analysis: Select Bin Width

- 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

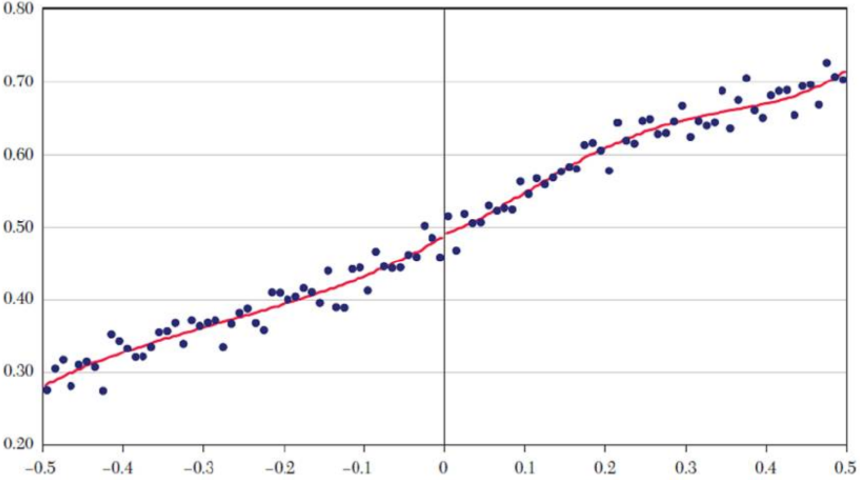
- It is probably advisable to report results for both estimation types:

1. 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 hurm and should not be use.(Gelman and Imbens,2019)
2. 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

1. 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

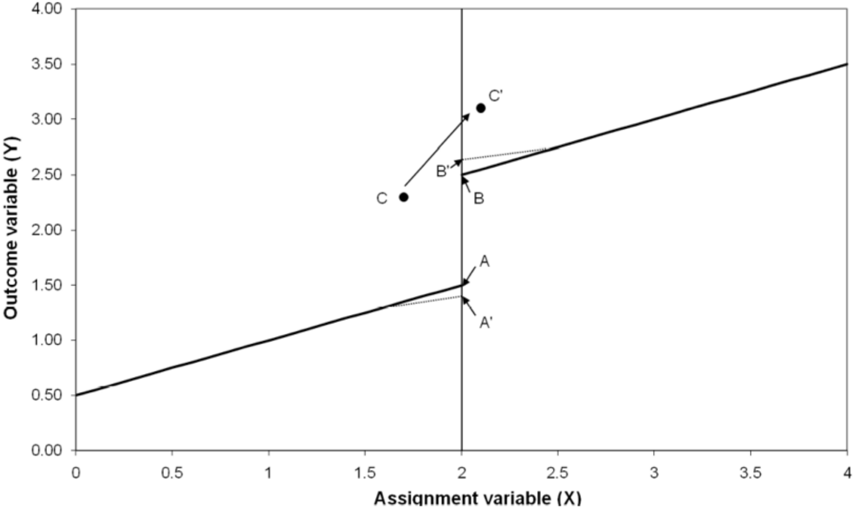
# Graphical:Example Covariates by Forcing Variable



# 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

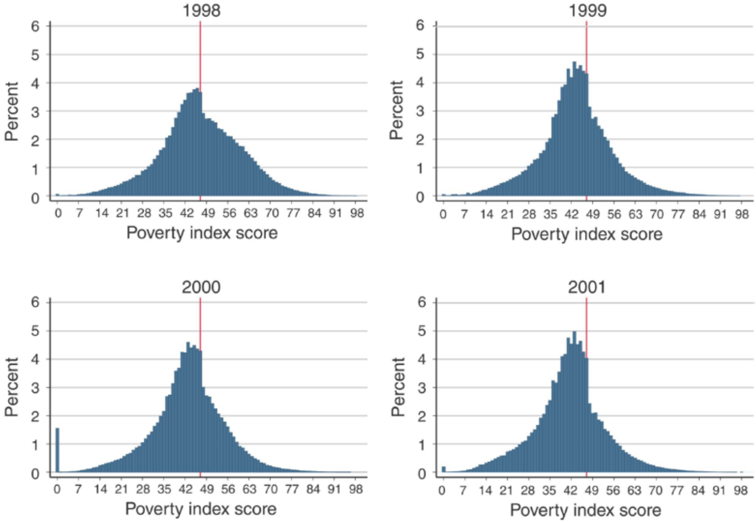


# Sorting behavior: Manipulation of a poverty index in Colombia

- 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

# Sorting behavior: 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: 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.

# 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