大数据时代下的管理决策

Lec3: Quasi-Experimental Methods(I)

Zhaopeng Qu

Nanjing University Business School

March 12 2022

Review the previous lecture

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 and Quasi-Experimental Methods

- Although RCTs is a powerful tool for economists, every project or topic can NOT be carried on by it.
- This is the reason why modern econometrics exists and develops. The main job of econometrics is using **non-experimental** data to *making convincing 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"
- If we only have cross-sectional data
  - **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.
- If we have panel data

Instrumental Variable Method

Introduction

- The earliest application involved attempts to estimate demand and supply curve for product.
- A simple but difficult question: **How to find the supply or demand curves?**
- **Difficulty**: We can only observe intersections of supply and demand, yielding pairs.
- **Solution**: Wright(1928) use variables that appear in one equation to shift this equation and trace out the other.
- The variables that do the shifting came to be known as **Instrumental Variables** method.
- It is well-known that IV can address the problems of omitted variable bias, measurement error and reverse causality problems.

Terminology: endogeneity and exogeneity

- An **endogenous variable** is one that both we are interested in and is correlated with u.
- An **exogenous variable** is one that is uncorrelated with u.
- Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality.
- However, this definition is narrow and IV regression can be used to address OVB and errors-in-variable bias, not just to simultaneous causality bias.

Instrumental variables

- Suppose a simple OLS regression like previous equation

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

  Where $E[u_i|X_i] \neq 0$
- The **1st assumption** of OLS is violated,then we can **NOT** obtain a consistent estimate of β
- The idea of IV is quite simple. Intuitively, we can split X_i into two parts:
  1. that is **correlated** with the error term.
  2. that is **uncorrelated** with the error term.
- If we can isolate the variation in X_i into these two parts, then we could use the one that is uncorrelated with u_i to obtain a consistent estimate of β,thus estimate the causal effect of X_i on Y_i.

Instrumental variables

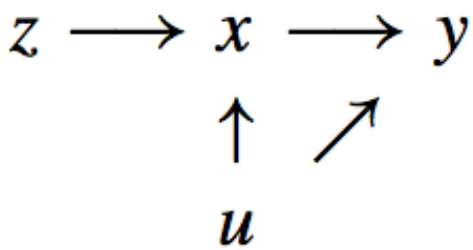
- An instrumental variable Z_i must satisfy the following 2 properties:
  1. **Instrumental relevance**: Z_i should be **correlated** with the casual variable of interest, X_i (endogenous variable),thus

$$Cov(X_i, Z_i) \neq 0$$

.

  2. **Instumental exogeneity**: Z_i is as good as randomly assigned and Z_i only affect on Y_i through X_i affecting Y_i channel.

$$Cov(Z_i, u_i) = 0$$



IV estimator: Jargon

- Our simple OLS regression: Causal relationship of interest

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- **First-Stage** regression: regress *endogenous variable* on IV

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- **Reduced-Form**: regress outcome variable on IV

$$Y_i = \delta_0 + \delta_1 Z_i + e_i$$

IV estimator: Two Steps Least Square (2SLS)

- We can estimate the causal effect of X_i on Y_i in two steps
  1. **First stage**: Regress X_i on Z_i & obtain predicted values of $\hat{X}_i$,if $Cov(Z_i, u_i) = 0$, then $\hat{X}_i$ contains variation in X_i that is uncorrelated with u_i

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

.

  2. **Second stage**: Regress Y_i on $\hat{X}_i$ to obtain the Two Stage Least Squares estimator $\hat{\beta}_{2SLS}$

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{\hat{X}})}{\sum(\hat{X}_i - \bar{\hat{X}})^2}$$

IV estimator: Two Steps Least Square (2SLS)

- Because $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$,then

$$\overline{\hat{X}} = \hat{\pi}_0 + \hat{\pi}_1 \overline{Z}_i$$

- We have

$$\hat{X}_i - \overline{\hat{X}} = \hat{\pi}_1 (Z_i - \bar{Z})$$

- Also because $\hat{\pi}_1$ is the estimating coefficient of Z_i on X_i, then base on simple OLS formula,

$$\hat{\pi}_1 = \frac{\sum (X_i - \bar{X})(Z_i - \overline{Z})}{\sum (Z_i - \overline{Z})^2}$$

IV estimator: Two Steps Least Square (2SLS)

- Then we could obtain

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \overline{\hat{X}})}{\sum(\hat{X}_i - \overline{\hat{X}})^2}$$

IV estimator: Two Steps Least Square (2SLS)

- Then we could obtain

$$\begin{aligned}\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{\hat{X}})}{\sum(\hat{X}_i - \bar{\hat{X}})^2} \\ &= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \bar{Z})}{\sum \hat{\pi}_1^2(Z_i - \bar{Z})^2}\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- Then we could obtain

$$\begin{aligned}\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{\hat{X}})}{\sum(\hat{X}_i - \bar{\hat{X}})^2} \\ &= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \bar{Z})}{\sum \hat{\pi}_1^2(Z_i - \bar{Z})^2} \\ &= \frac{1}{\hat{\pi}_1}\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(Z_i - \bar{Z})^2}\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- Then we could obtain

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\sum(Y_i - \bar{Y})(\hat{X}_i - \bar{\hat{X}})}{\sum(\hat{X}_i - \bar{\hat{X}})^2} \\
&= \frac{\sum(Y_i - \bar{Y})\hat{\pi}_1(Z_i - \bar{Z})}{\sum \hat{\pi}_1^2(Z_i - \bar{Z})^2} \\
&= \frac{1}{\hat{\pi}_1}\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(Z_i - \bar{Z})^2} \\
&= \frac{\sum(Z_i - \bar{Z})^2}{\sum(X_i - \bar{X})(Z_i - \bar{Z})} \times \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(Z_i - \bar{Z})^2}
\end{aligned}$$

IV estimator: Two Steps Least Square (2SLS)

- Which gives the 2SLS IV estimator

$$\hat{\beta}_{2SLS} = \frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})} = \frac{s_{ZY}}{s_{ZX}}$$

- The 2SLS estimator of β_1 is the ratio of *the sample covariance between* Z and Y to *the sample covariance between* Z and X.
- If $Z_i = X_i$, then

$$\hat{\beta}_{2SLS} = \hat{\beta}_{ols}$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$E[\hat{\beta}_{2SLS}] = E\Big[\frac{\sum(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\Big]$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1 + E\left[\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Consider $E[\hat{\beta}_{IV}]$

$$
\begin{aligned}
E[\hat{\beta}_{2SLS}] &= E\left[\frac{\sum(Y_i-\bar{Y})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum[(\beta_0+\beta_1 X_i+u_i)-(\beta_0+\beta_1\bar{X}+\bar{u})](Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= E\left[\frac{\sum\beta_1(X_i-\bar{X})(Z_i-\bar{Z})+\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1 + E\left[\frac{\sum(u_i-\bar{u})(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right] \\
&= \beta_1 + E\left[\frac{\sum u_i(Z_i-\bar{Z})}{\sum(X_i-\bar{X})(Z_i-\bar{Z})}\right]
\end{aligned}
$$

Statistical propertise of 2SLS estimator: Unbiasedness

- Because instrument exogeneity implies $Cov(Z_i, u_i) = 0$,but not $E[u_i|Z_i, X_i] = 0$,then

$$E\Big[\frac{\sum u_i(Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\Big] = E\Big[\frac{\sum E[u_i|X_i, Z_i](Z_i - \bar{Z})}{\sum(X_i - \bar{X})(Z_i - \bar{Z})}\Big] \neq 0$$

- Then we have

$$E[\hat{\beta}_{2SLS}] \neq \beta_1$$

- It means that 2SLS estimator is **biased**.

Statistical propertise of 2SLS estimator: Consistent

- We have a simple regression $Y_i = \beta_0 + \beta_1 X_i + u_i$ and take a covariance of Y_i and Z_i

$$\begin{aligned} Cov(Z_i, Y_i) &= Cov[Z_i, (\beta_0 + \beta_1 X_i + u_i)] \\ &= Cov(Z_i, \beta_0) + \beta_1 Cov(Z_i, X_i) + Cov(Z_i, u_i) \\ &= \beta_1 Cov(Z_i, X_i) \end{aligned}$$

- Thus if the instrument is valid,

$$\beta_1 = \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)}$$

- The population coefficient is the ratio of *the population covariance between* Z and Y to *the popualtion covariance between* Z and X.

Statistical propertise of 2SLS estimator: Consistent

- As discussed in Section 3.7,the sample covariance is a consistent estimator of the population covariance, thus $s_{ZY} \xrightarrow{p} Cov(Z_i, Y_i)$ and $s_{ZX} \xrightarrow{p} Cov(Z_i, X_i)$
- Then the TSLS estimator is **consistent**.

$$\hat{\beta}_{2SLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{Cov(Z_i, Y_i)}{Cov(Z_i, X_i)} = \beta_1$$

Statistical propertise of 2SLS: Statistical Inference

- The variance $\hat{\beta}_{2SLS}$ can be estimated by estimating the variance and covariance terms appearing in Equation (12.8),thus

$$SE(\hat{\beta}_{2SLS}) = \sqrt{\frac{\frac{1}{n}\sum(Z_i - \mu_Z)^2\hat{u}_i^2}{n(\frac{1}{n}\sum(Z_i - \mu_Z)X_i)^2}}$$

- Then the square root of the estimate of $\sigma^2_{\hat{\beta}_{2SLS}}$, thus *the standard error of the IV estimator*, which is a little bit complicated. Fortunately,this is done automatically in TSLS regression commands in econometric software packages.
- Because $\hat{\beta}_{2SLS}$ is normally distributed in large samples, hypothesis tests about β can be performed by computing *the t-statistic*,and a 95% large-sample *confidence interval* is given by

$$\hat{\beta}_{2SLS} \pm 1.96SE(\hat{\beta}_{2SLS})$$

A Good Example: Long live Keju

Chen, Kung and MA(2020)

- Ting Chen, James Kai-sing Kung(龚启圣) and Chicheng Ma(2020),*"Long Live Keju! The Persistent Effects of China's Imperial Examination System"*,*The Economic Journal*, 130 (October), 2030–2064.
- Topic: Long term persistence of human capital:the effect of **Keju**
- Dependent Variable: average schooling years in 2010
- Independent Variable: the density of **jinshi** in the Ming-Qing dynasties
- Data: 272 prefectures in *jinshi*.

Chen, Kung and MA(2020)

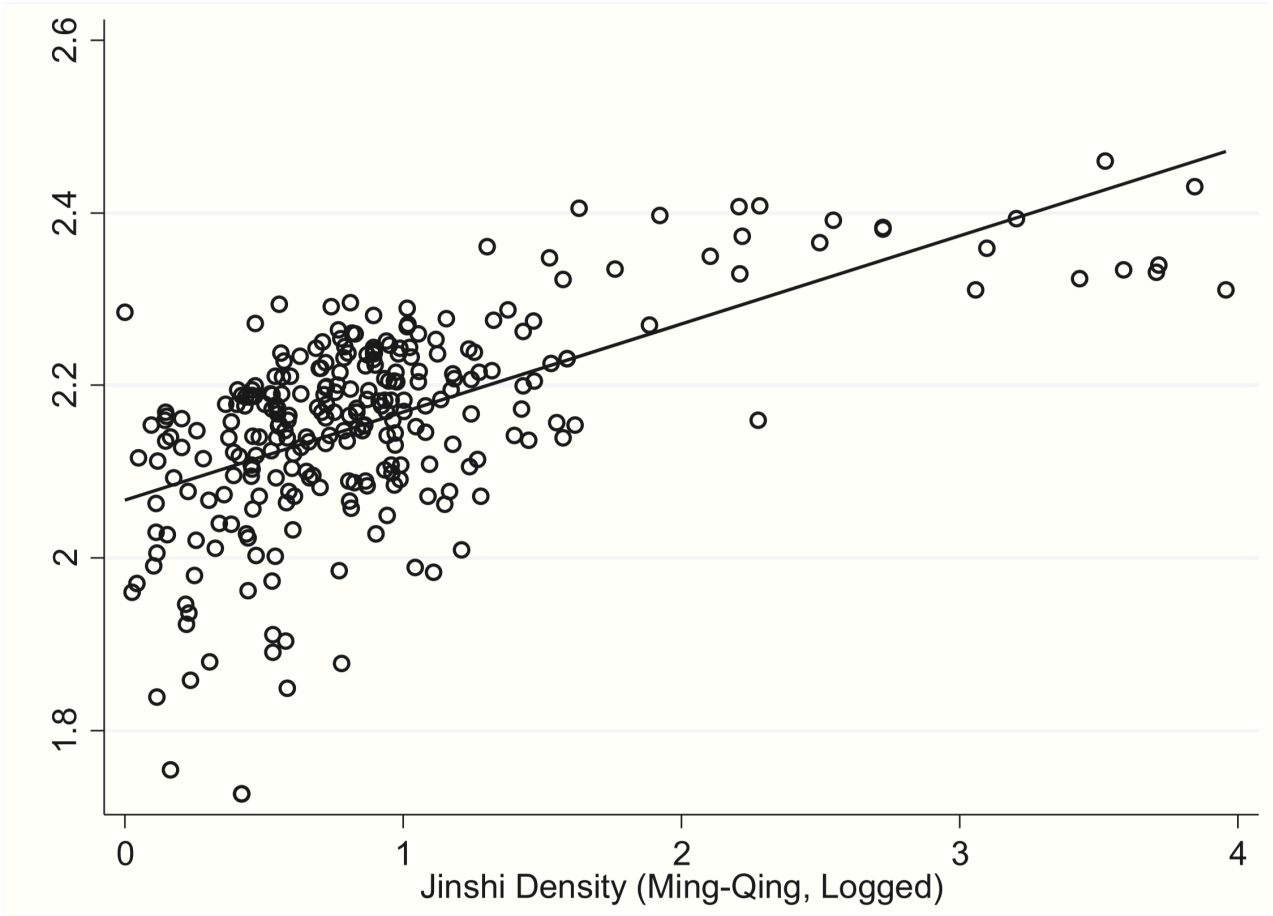


Fig. 1. *Correlation between Historical Success in China's Civil Exam (Keju) and Contemporary Human Capital Outcomes.*

Chen, Kung and MA(2020)

- The effect of Keju on human capital at present
- Run regression

$$lnY_i = \alpha + \beta ln(Keju_i) + \gamma_1 X_i^c + \gamma_2 X_i^h + u_i$$

- Y_i: 2010 年 i 地区 (地级市或" 府") 的平均受教育年限。
- $Keju_i$: 明清时期 i 地区获得进士的人数。
- X_i^c: 控制变量（当代），包括经济繁荣程度 (夜间灯光)；地理因素：该地区到海选距离、地形（免于遭受自然灾害）。
- X_i^c: 控制变量（历史）：历史经济繁荣程度、基础教育设施、社会和政治影响力等等

Chen, Kung and MA(2020): OLS

Table 3. Impact of *Jinshi* Density on Contemporary Human Capital: OLS Estimates

	Average Years of Schooling in 2010 (logged)					
	(1)	(2)	(3)	(4)	(5)	(6)
Jinshi Density (logged)	0.092***	0.065***	0.070***	0.067***	0.058***	
	(0.007)	(0.007)	(0.007)	(0.008)	(0.009)	
	[0.007]	[0.007]	[0.007]	[0.007]	[0.007]	
Jinshi Density (logged, excludes migrant)						0.053***
						(0.019)
						[0.016]
Economic Prosperity						
Population Density (logged)			-0.049***	-0.051***	-0.053***	-0.049***
			(0.016)	(0.016)	(0.015)	(0.015)
			[0.013]	[0.013]	[0.012]	[0.013]
Urbanization Rate			0.062	0.093	0.051	0.234
			(0.163)	(0.156)	(0.164)	(0.180)
			[0.167]	[0.162]	[0.173]	[0.169]
Commercial Center			-0.012	-0.014	-0.020	-0.026*
			(0.014)	(0.014)	(0.013)	(0.014)
			[0.011]	[0.011]	[0.011]	[0.012]
Agricultural Suitability			-0.005	-0.005	-0.003	-0.004
			(0.014)	(0.014)	(0.014)	(0.014)
			[0.009]	[0.009]	[0.009]	[0.009]

Chen, Kung and MA(2020): Potential Bias

- **OVB**: that are simultaneously associated with both historical "jinshi" density and years of schooling today.
- For instance, prefectures that had produced more "jinshi" may be associated with unobserved (natural or genetic) endowments.

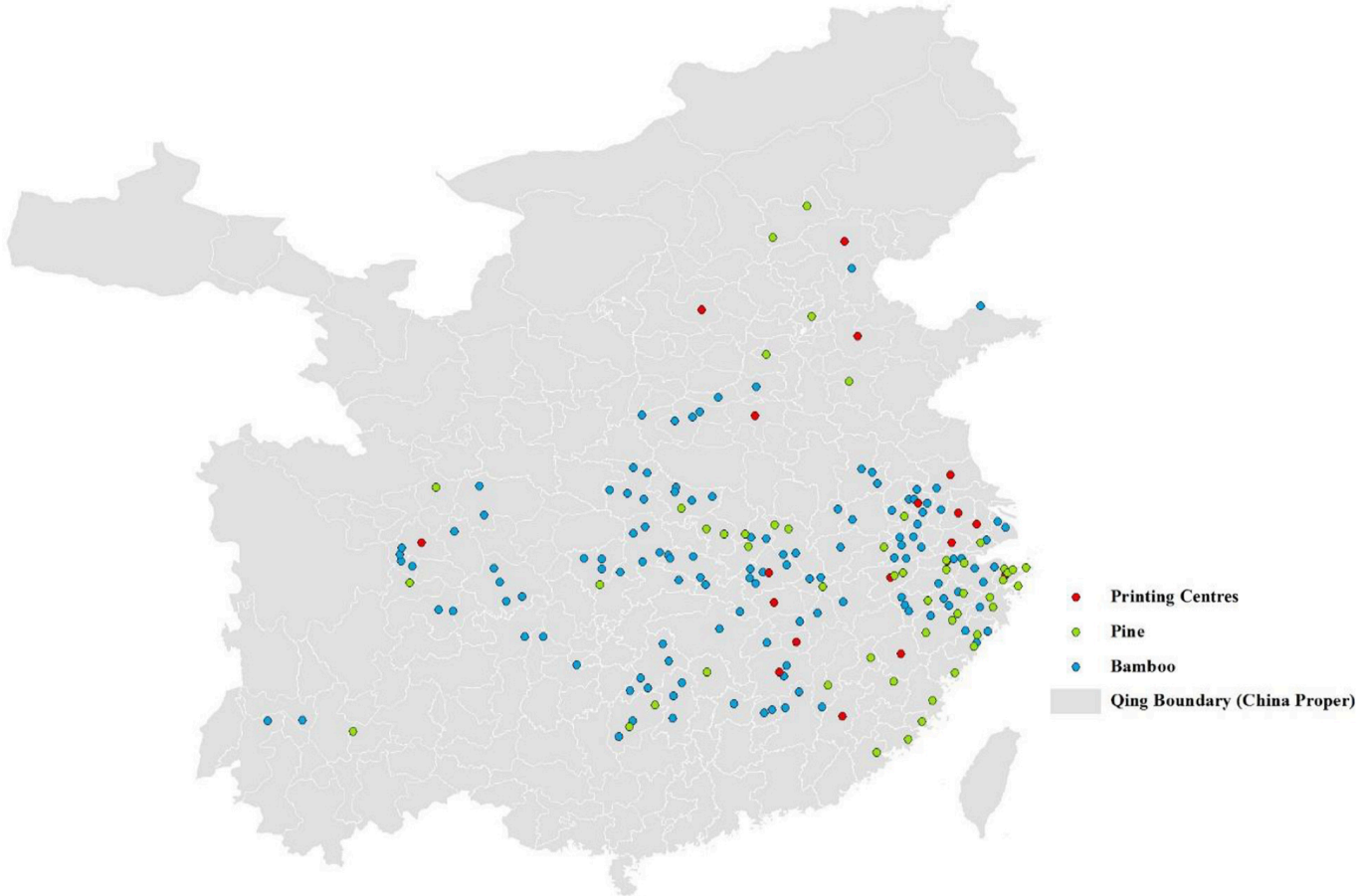
Chen, Kung and MA(2020): Instrumental Variable

- IV: Distance to the Printing Ingredients (Pine and Bamboo) as the Instrumental Variable of "Keju"
- A logic chain:

$$\begin{aligned} \textit{More jinshi} &\Longleftarrow \textit{more references books} \\ &\Longleftarrow \textit{more print in print centers} \\ &\Longleftarrow \textit{centers closer nearby some ingredients} \end{aligned}$$

Chen, Kung and MA(2020): Instrumental Variable

- Only 19 printing centres were distributed across the 278 prefectures, and that these 19 centres accounted for 80% of the 13,050 texts published during that period (Zhang and Han, 2006)



Chen, Kung and MA(2020): Instrumental Variable

- Jianning Fu(建寧府) and Tingzhou FU(汀州府)

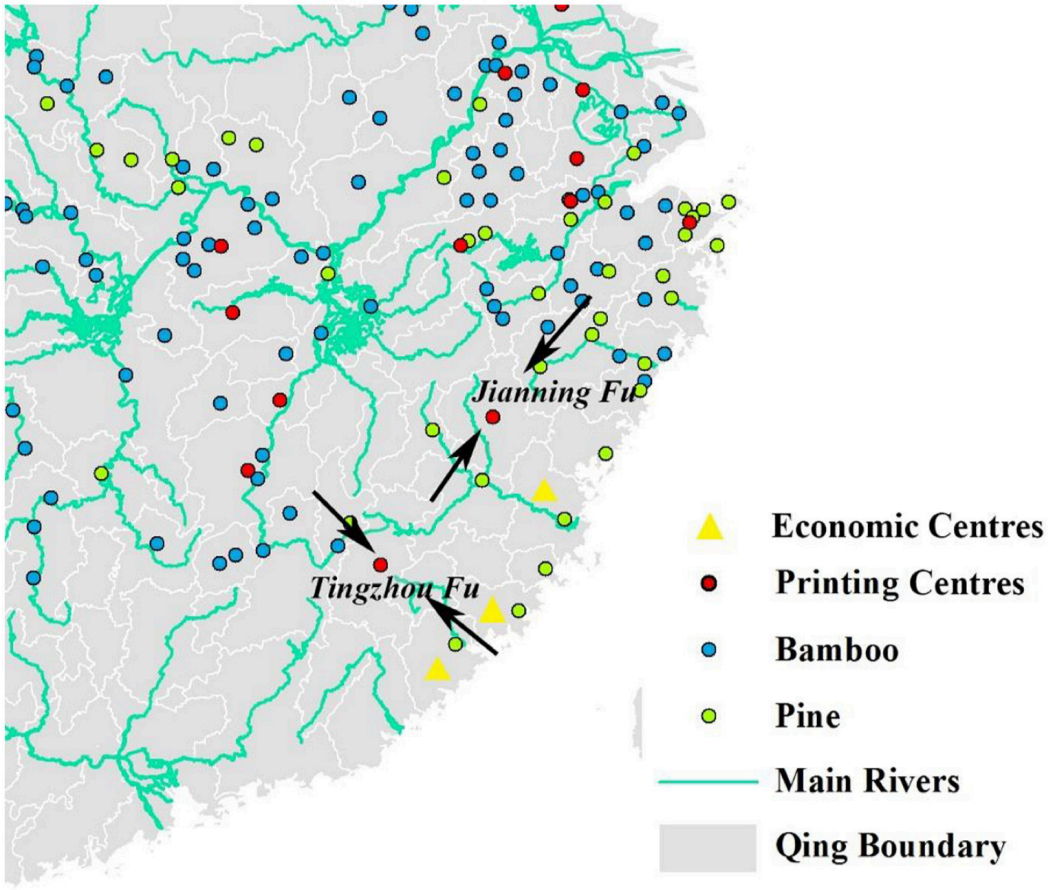


Fig. 3B. *Two Examples of Printing Centres' Proximity to Pine and Bamboo Locations.*

Chen, Kung and MA(2020): First Stage

Table 4. River Distance to Pine and Bamboo Locations, Printing Centers and *Jinshi* Density

	Jinshi Density (logged)		Printing Center		Printed Books (logged)		*Jinshi* Density (logged)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Printed Books (logged)	0.179***	0.170***						
	(0.031)	(0.036)						
River Distance to Pine/Bamboo			-0.017***	-0.017***	-0.092***	-0.084***	-0.102***	-0.099***
			(0.004)	(0.004)	(0.029)	(0.029)	(0.011)	(0.012)
Baseline Control Variables	No	Yes	No	Yes	No	Yes	No	Yes
Provincial Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	274	274	274	274	274	274	274	274
Adj. R-squared	0.323	0.332	0.132	0.131	0.449	0.463	0.526	0.528

More Books, More Jinshis

Closer to Pine/Bamboo, More Center, then More Books

Closer to Pine/ Bamboo, More Jinshis

Notes: All results are OLS estimates. Baseline controls include agricultural suitability, distance to coast, and terrain ruggedness. Robust standard errors adjusted for clustering at the province level are given in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10%, respectively.

Chen, Kung and MA(2020): Reduced-form and 2SLS

Table 7. Impact of *Keju* on Contemporary Human Capital: Instrumented Results

	Reduced-form			2SLS		
	(1)	(2)	(3)	(4)	(5)	(6)
Jinshi Density (logged)				0.104***	0.080***	0.082***
				(0.008)	(0.013)	(0.013)
Distance to Major Navigable Rivers			0.008			0.008
			(0.006)			(0.006)
				First Stage		
River Distance to Bamboo/Pine	-0.011***	-0.006***	-0.006***	-0.011***	-0.006***	-0.006***
	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)
First Stage F-stat				78.04	58.07	57.76
First Stage Partial R-squared				0.392	0.282	0.282
Baseline + Additional Controls	No	Yes	Yes	No	Yes	Yes
Provincial Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of Observations	272	272	272	272	272	272
Adj. R-squared	0.531	0.732	0.735	0.65	0.751	0.752
Cragg-Donald Wald F-statistic				129.156	72.314	72.354

Notes: Baseline controls include nighttime lights in 2010, agricultural suitability, distance to coast, and terrain ruggedness. Additional controls are commercial center, population density, urbanization rate, Confucian academies, private book collections, strength of clan and political elites. Robust standard errors adjusted for clustering at the province level are given in parentheses. ***, **, and * indicate statistical significance at the 1%, 5%, and 10%, respectively.

Chen, Kung and MA(2020): Exclusion Restrictions

- The locations of bamboo and pine geographic distributions were exogenously given. Historians find little if any evidence of planting pine and bamboo intentionally for the purpose of commercial printing.
- But they may be correlated with other omitted variables—most notably **economic prosperity**, which may be correlated with years of schooling today.

Table A3. Exclusion Restrictions

Panel A	Commercial Centers in Ming-Qing	Tea Centers in Ming-Qing	Silk Centers in Ming-Qing	Population Density in Ming (logged)	Population Density in Qing (logged)	Population Density in 1953 (logged)	Urbanization Rate in Ming-Qing	Urbanization Rate in 1920
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
River Distance to Pine/Bamboo (logged)	-0.006	-0.007	-0.008	0.007	-0.020	-0.021	-0.001	-0.020
	(0.005)	(0.005)	(0.007)	(0.045)	(0.019)	(0.017)	(0.001)	(0.024)
Observations	274	274	274	274	274	269	274	274
Adjusted R-squared	0.309	0.216	0.153	0.534	0.624	0.540	0.664	0.296

Where Do Valid Instruments Come From?

Where do we find an IV?

- Generally Speaking
  - “可遇不可求”
- Two main approaches
  1. Economic Theory/Logics
  2. Exogenous Source of Variation in X(natural experiments)

Where do we find an IV?

- Example 1: Does putting criminals in jail reduce crime?
- Run a regression of crime rates(d.v.) on incarceration rates(id.v) by using annual data at a suitable level of jurisdiction(states) and covariates (economic conditions)
- *Simultaneous causality bias*: crime rates goes up, more prisoners and more prisoners,reduced crime.
- IV: it must affect the incarceration rate but be unrelated to any of the unobserved factors that determine the crime rate.
- Levitt (1996) suggested that **lawsuits aimed at reducing prison overcrowding** could serve as an instrumental variable.
- Result: The estimated effect was three times larger than the effect estimated using OLS.

Where do we find an IV?

1. Institutional Background

- Angrist(1990)-draft lottery: Vietnam veterans were randomly designated based on birth day used to estimate the wage impact of a shorter work experience.
- Acemoglu, Johnson, and Robinson(2001): the dead rate of some diseases in some areas to estimate the impact of institutions to economic growth.
- Li and Zhang(2007),Liu(2012): "One Child policy"

Where do we find an IV?

2. Natural conditions(geography,weather,disaster)

   - the Rainfall,Hurricane,Earthquake,Tsunami...
   - the number of Rivers: Hoxby(2000)
   - the distance to print ingredients: Chen,Gong and Ma(2020)

Where do we find an IV?

3. Economic theory and Economic logic

- study the alcohol consumption and income relationship. alcohol price change by government's tax in a local market may be as a instrument of alcohol consumption.
- Angrist & Evans(1998): have same sex or different sex children used to estimate the impact of an additional birth on women labor supply.

Warp Up

- IV is a very powerful and popular tool to make causal inference.
- Making assumptions solid(convincing) plays a key role in using IV.
- Don't forget about the limitation: what we get in IV estimation is just a LATE.

A Business Case: Email Promotions

- Suppose an alcohol company(like 洋河) want to find the optimal strategy of emails promotion to increase the sales.
- Experimental Design: we randomly select customers into two groups, send emails to one group, and compare the outcome with the other, which is whether the customer buy or not at least one bottle of alcohol.
- Question: What is the outcomes?
  - **buy a bottle at least**
- Question: What is the treat?
  - **receive an email**?
  - or **receive and open(read) the email**?

A Business Case: Email Promotions

- If we think the treat is **receiving and opening(read) the email**, then selection bias may arise
  - Because those who chose to open the email may be different from those who didn't choose to open the email.
- To avoid the selection bias, we may compare the control group with the entire treatment group regardless of email opens.
  - However,in that case the estimated effect would be diluted because some of the people in the treatment group were not actually treated.

A Business Case: Email Promotions

- We can use the framework of IV to estimate the effect of actually receiving the treatment on the outcomes by treating the group assignment as an IV. Thus
  - The outcome is Y, thus "buy or not"
  - The endogenous variable is X, thus "read(open)the email or not"
  - The instrumental variable is Z, thus "receive the email or not"
- Then we can use 2SLS or other methods to recover the causal effect of X on Y.

Regression Discontinuity Design

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.

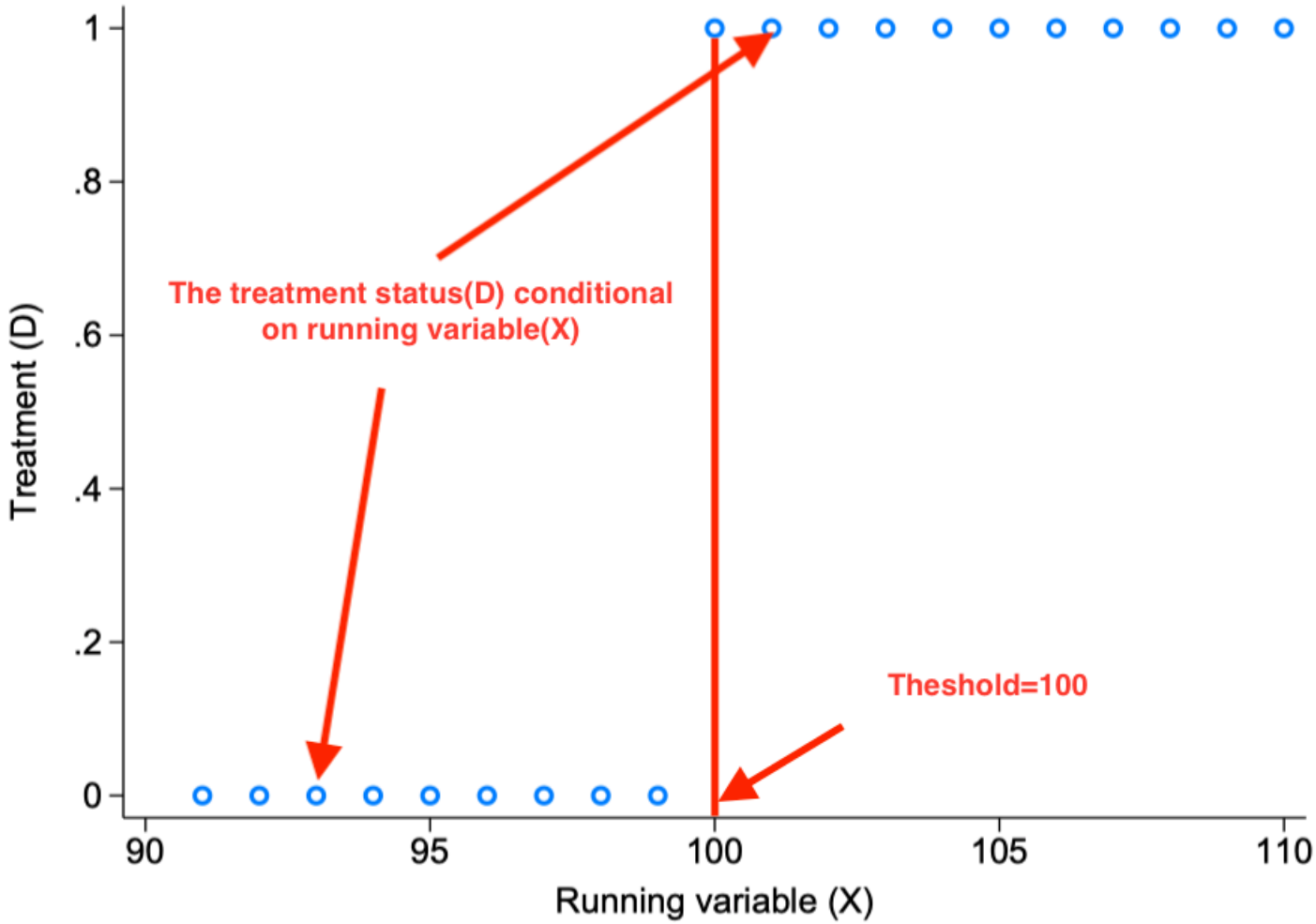
A Motivating Example: Elite University

- 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?)

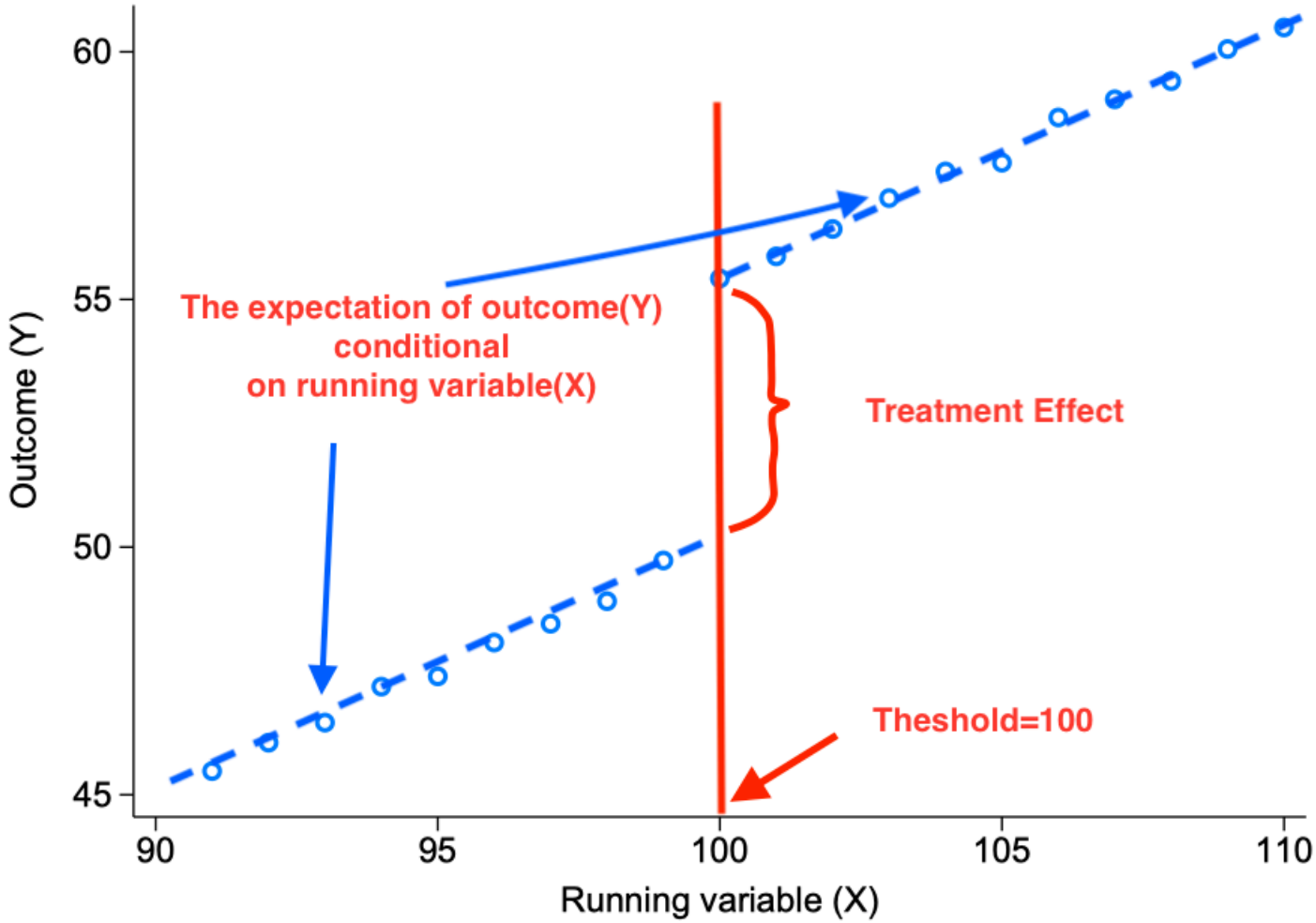
A Motivating Example: Elite University

- 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



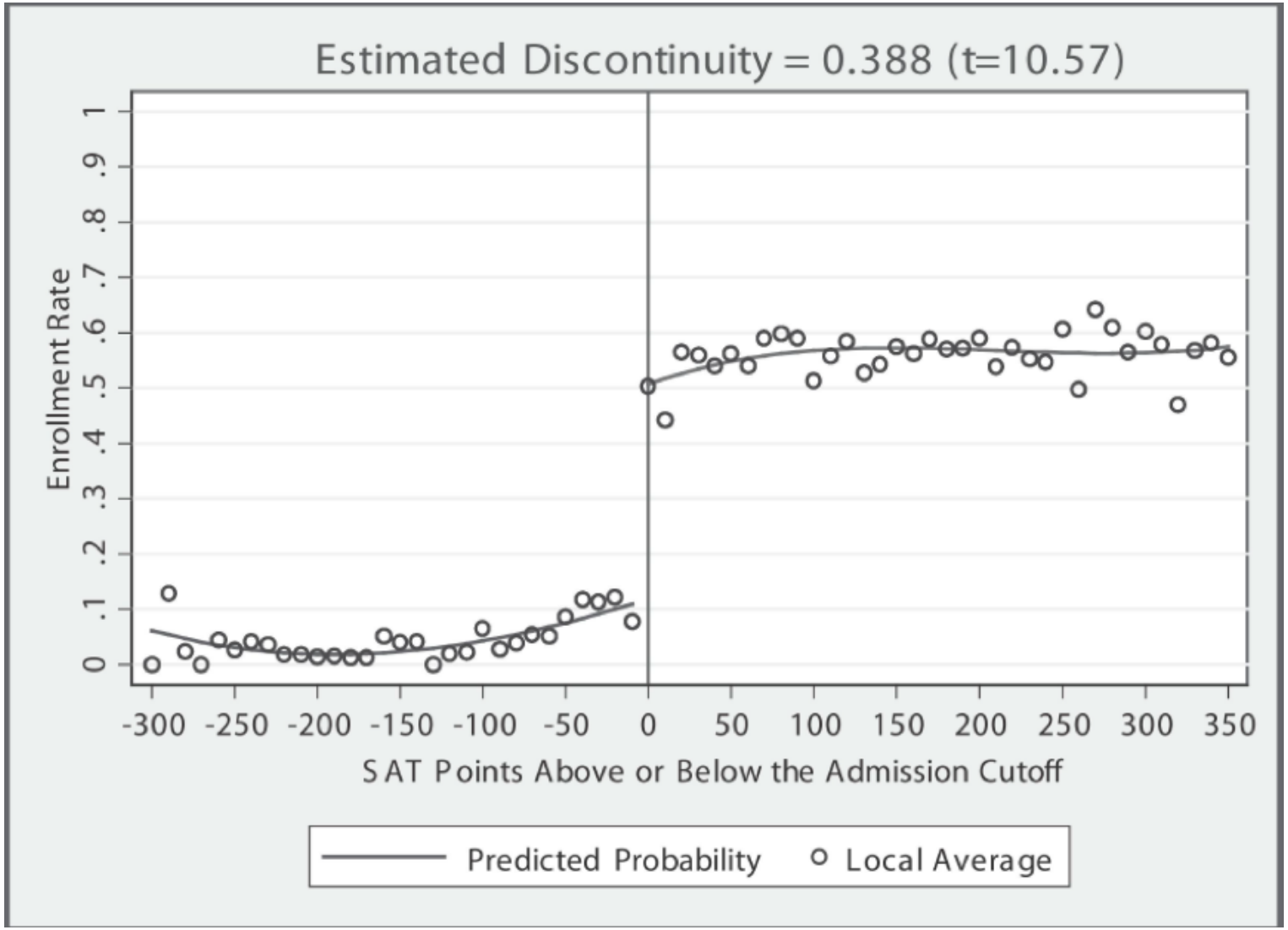
A Motivating Example: Elite University

- 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.

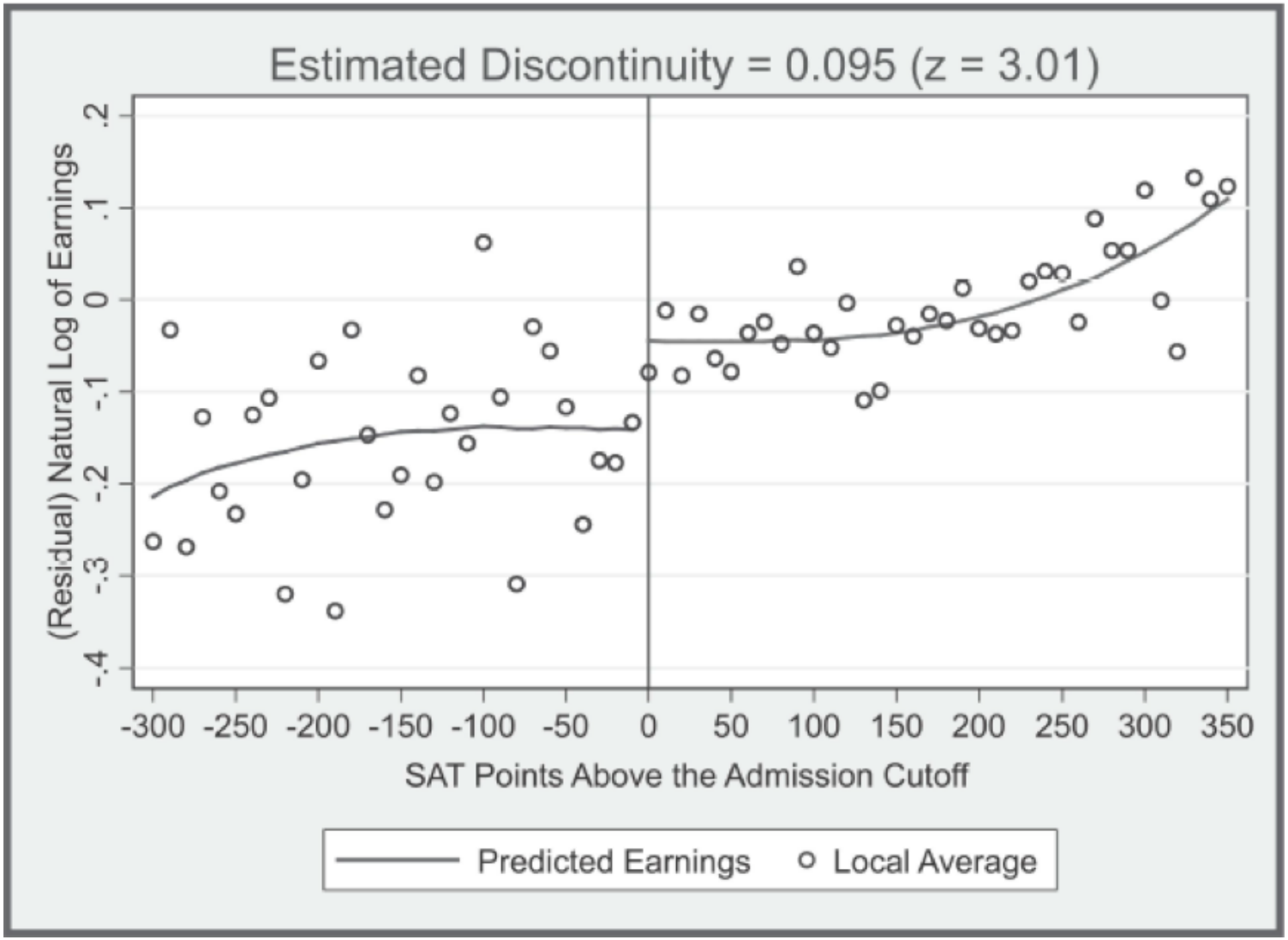
A Motivating Example: Elite University

- 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

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

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**

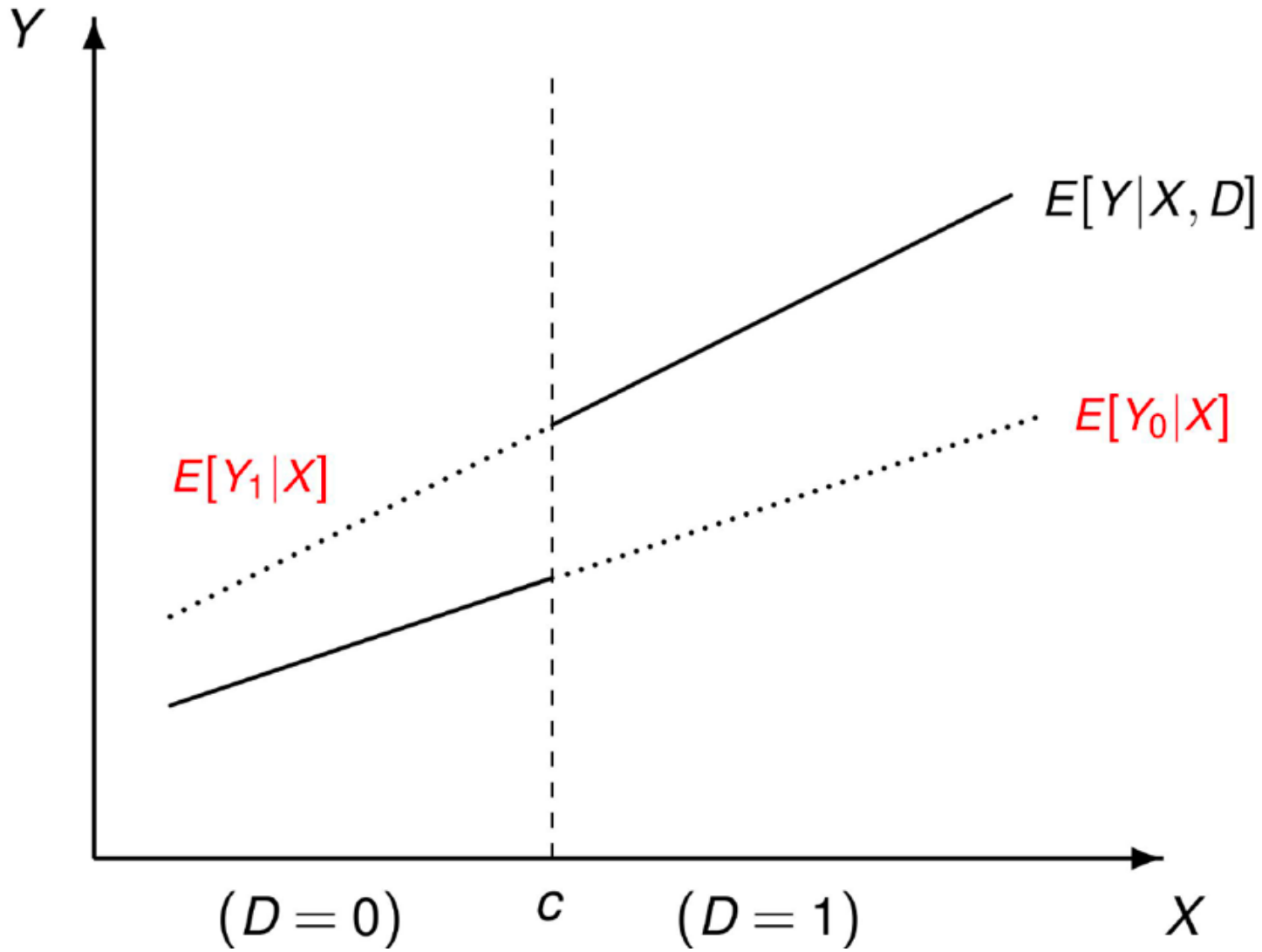
$$D_i = 1(X_i \geq c)$$

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

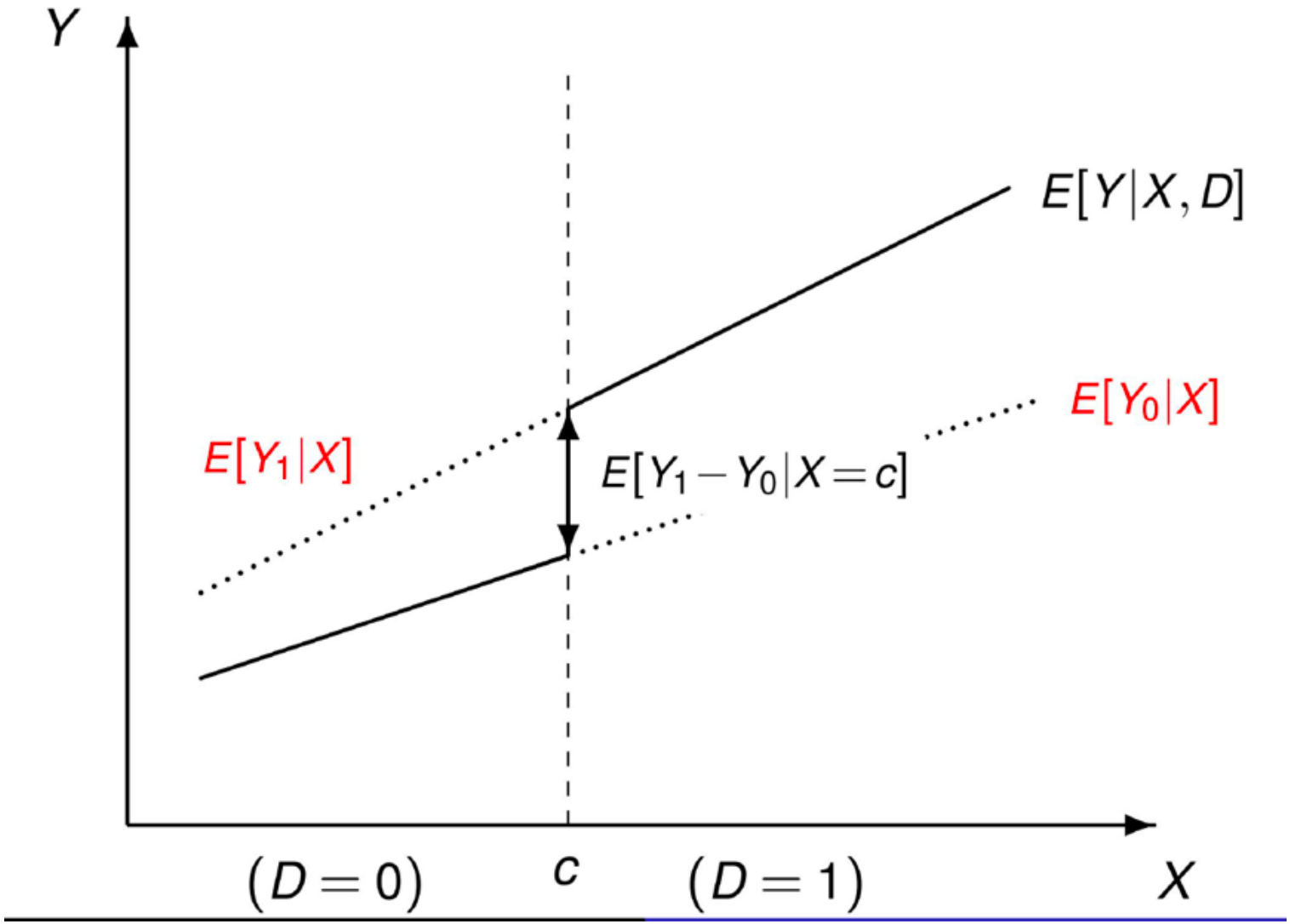
Identification for Sharp RDD

- **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.

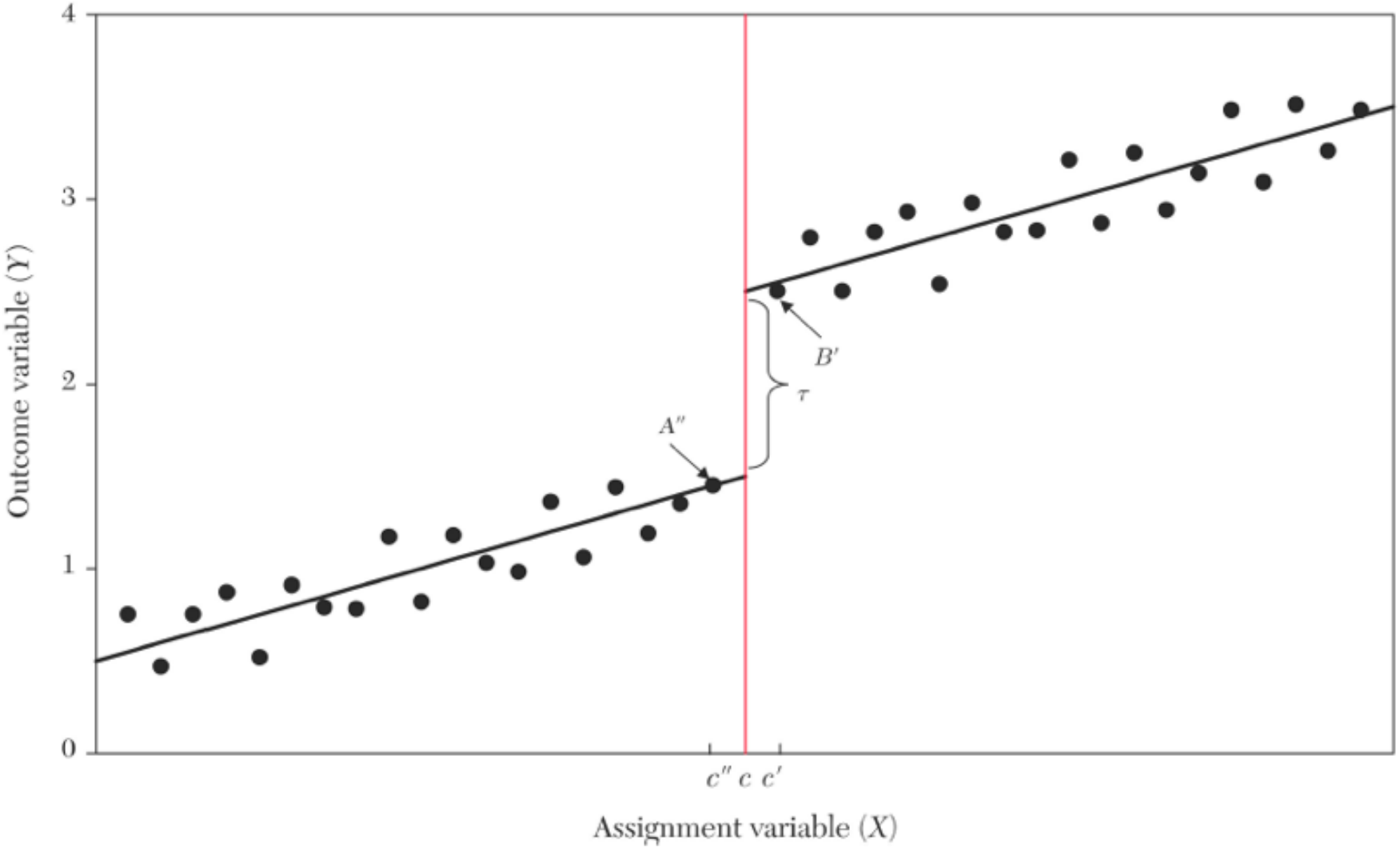
Sharp RDD specification

- 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

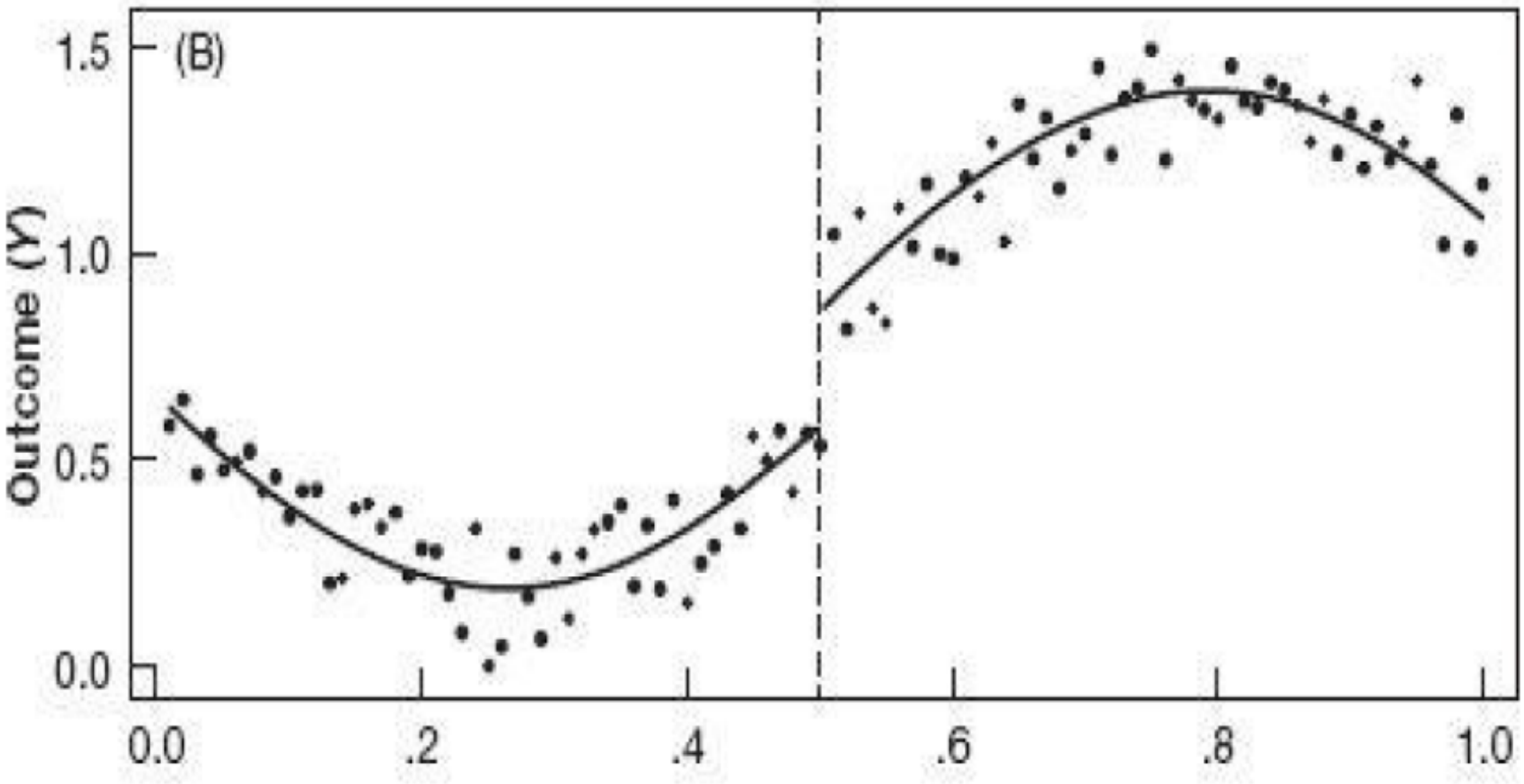


Specification

- 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.

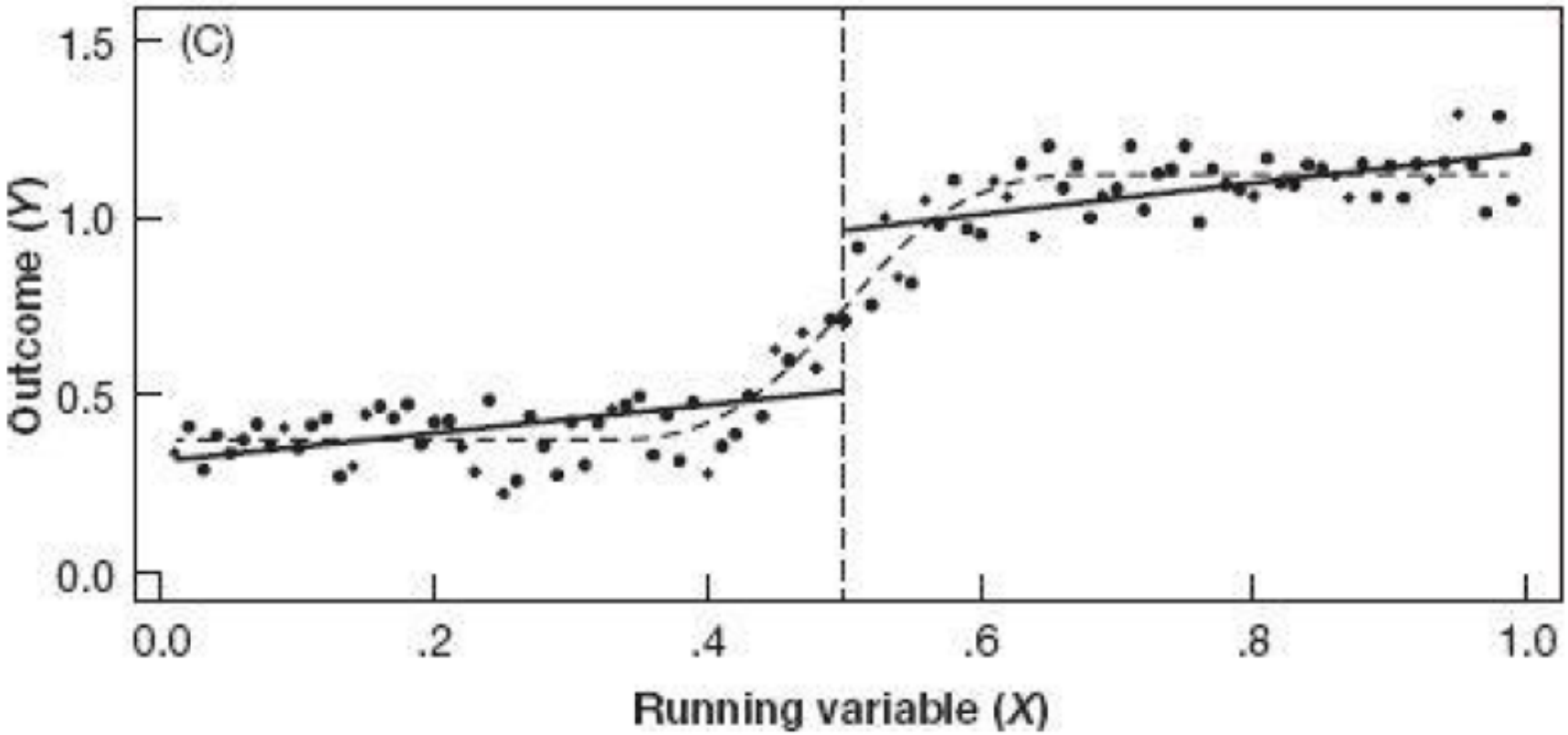
Nonlinear Case

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



Nonlinear Case

- The function form is very important in RDD.



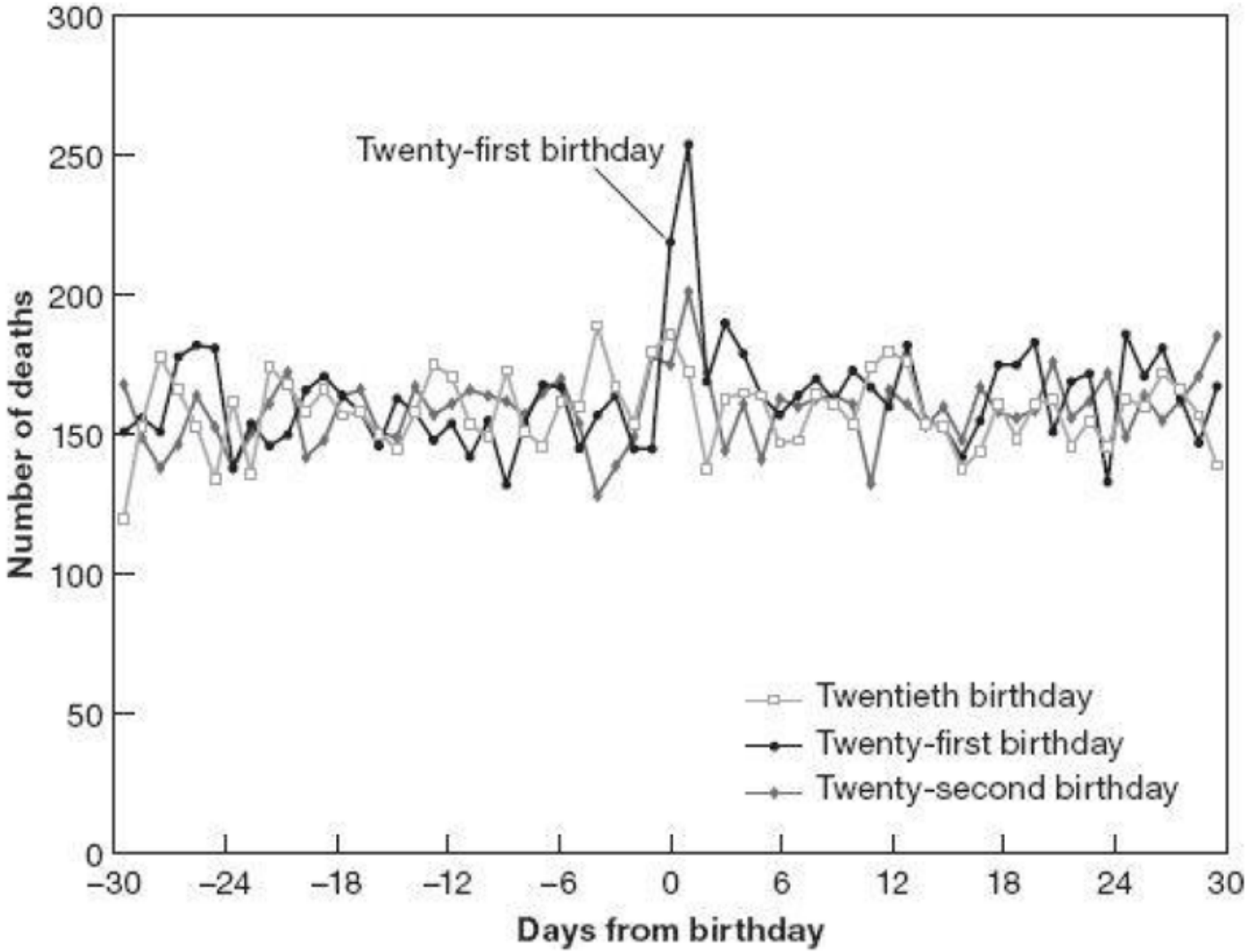
Sharp RDD Estimation

- 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.

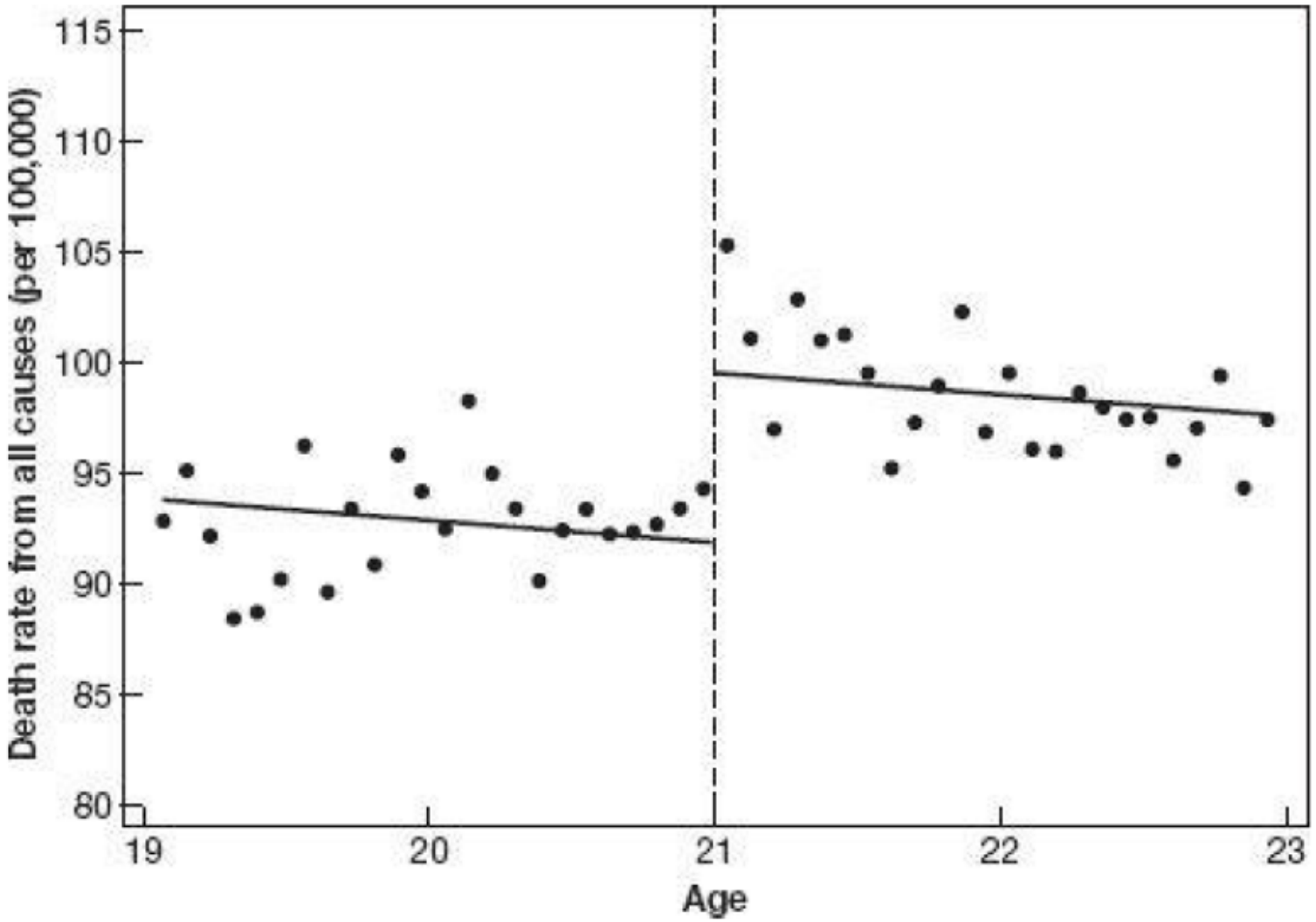
Application: Effect of the Minimum Legal Drinking Age

- 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?

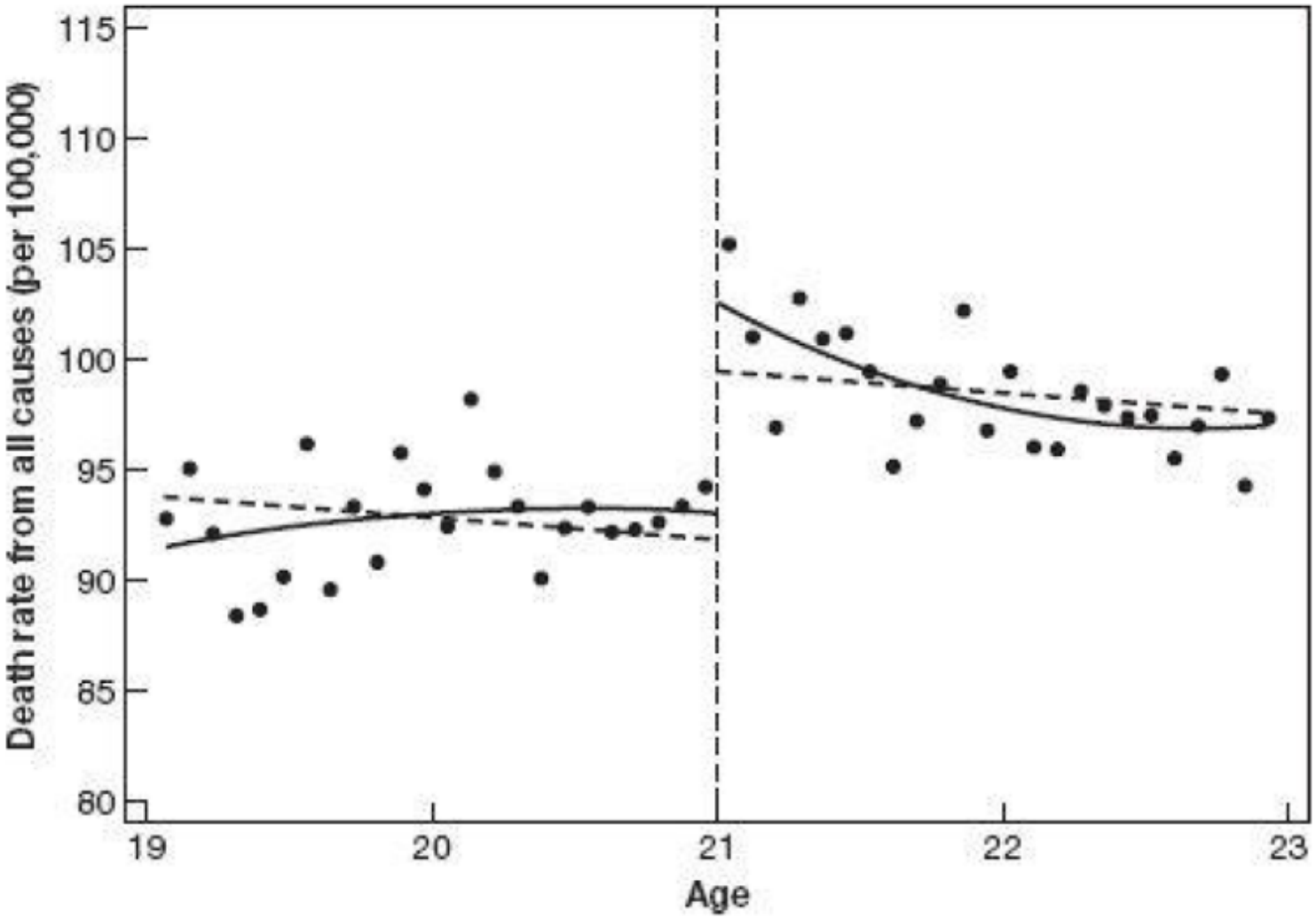
Application: MLDA on death rates



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

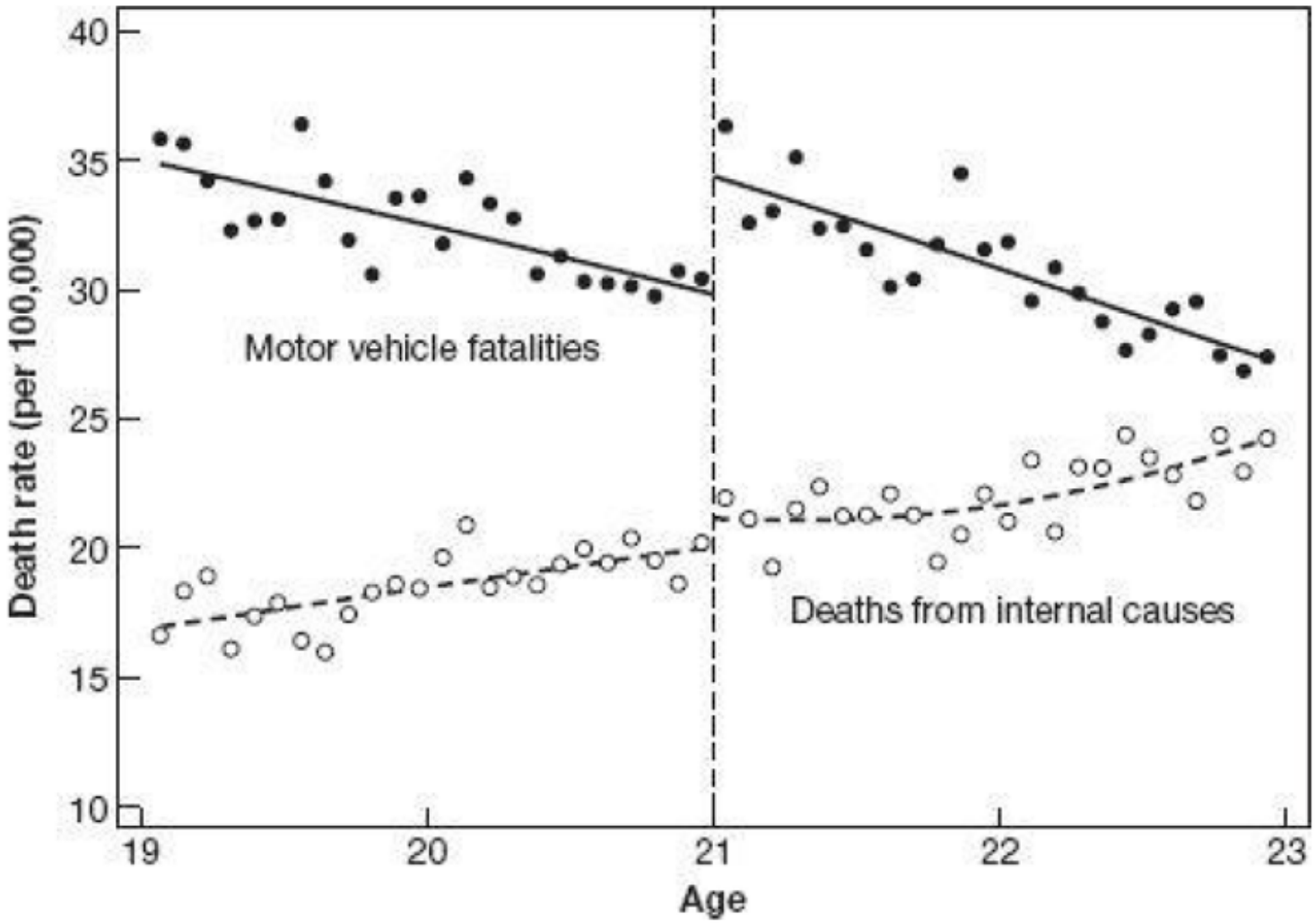
Application: MLDA on death rates

TABLE 4—DISCONTINUITY IN LOG DEATHS AT AGE 21

	(1)	(2)	(3)	(4)
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.

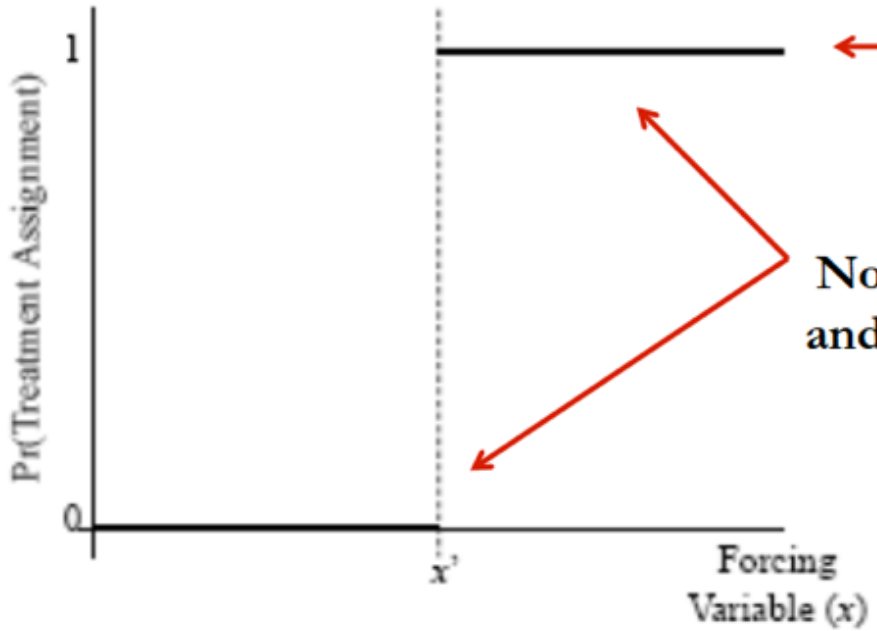
Fuzzy RDD

- Treatment Assignment:

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



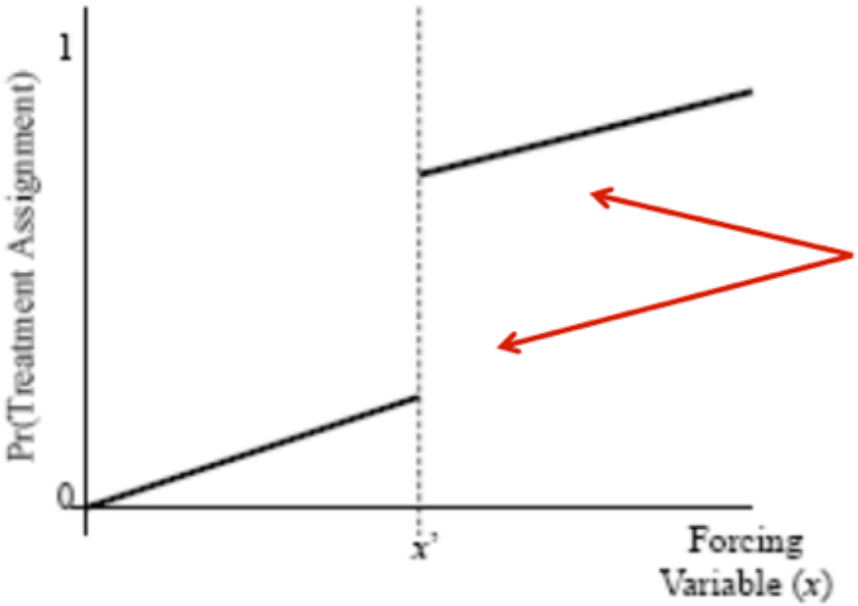
Probability of treatment moves from 0 to 1 around threshold value x'

No untreated for $x > x'$ and no treated for $x < x'$

Only x determines treatment

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:

$$Z_i = 1 \textit{ if assign to treatment group}$$
$$Z_i = 0 \textit{ if assign to control group}$$

- 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

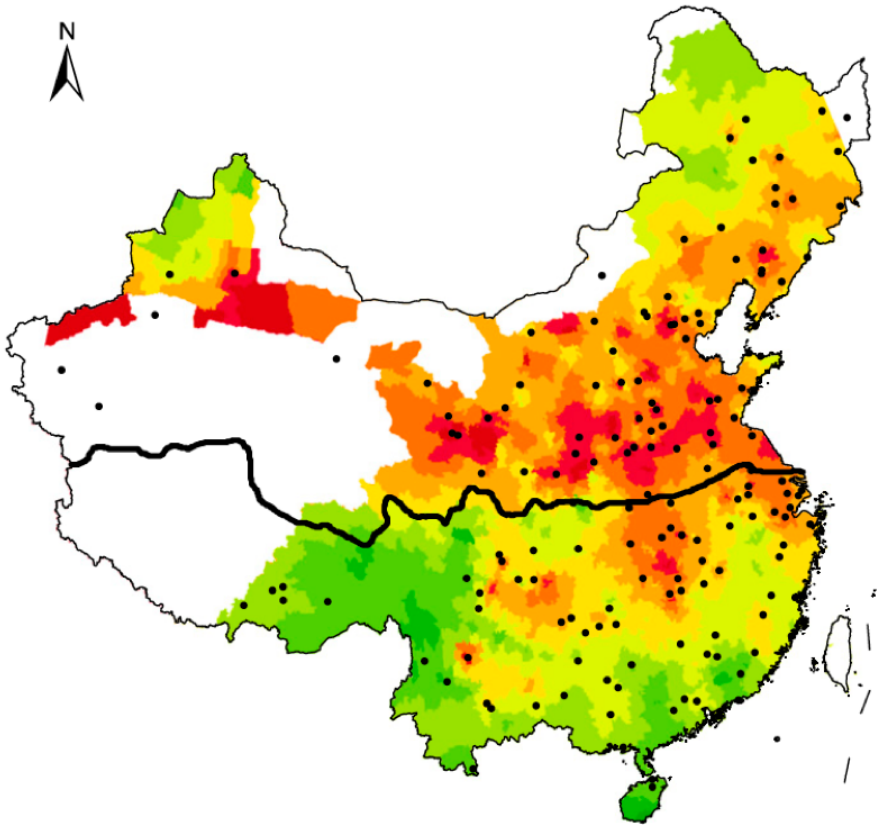


Fig. 1. China's Huai River/Qinling Mountain Range winter heating policy line and PM_{10} concentrations. Black dots indicate the DSP locations. Coloring corresponds to interpolated PM_{10} levels at the 12 nearest monitoring stations, where green, yellow, and red indicate areas with relatively low, moderate,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

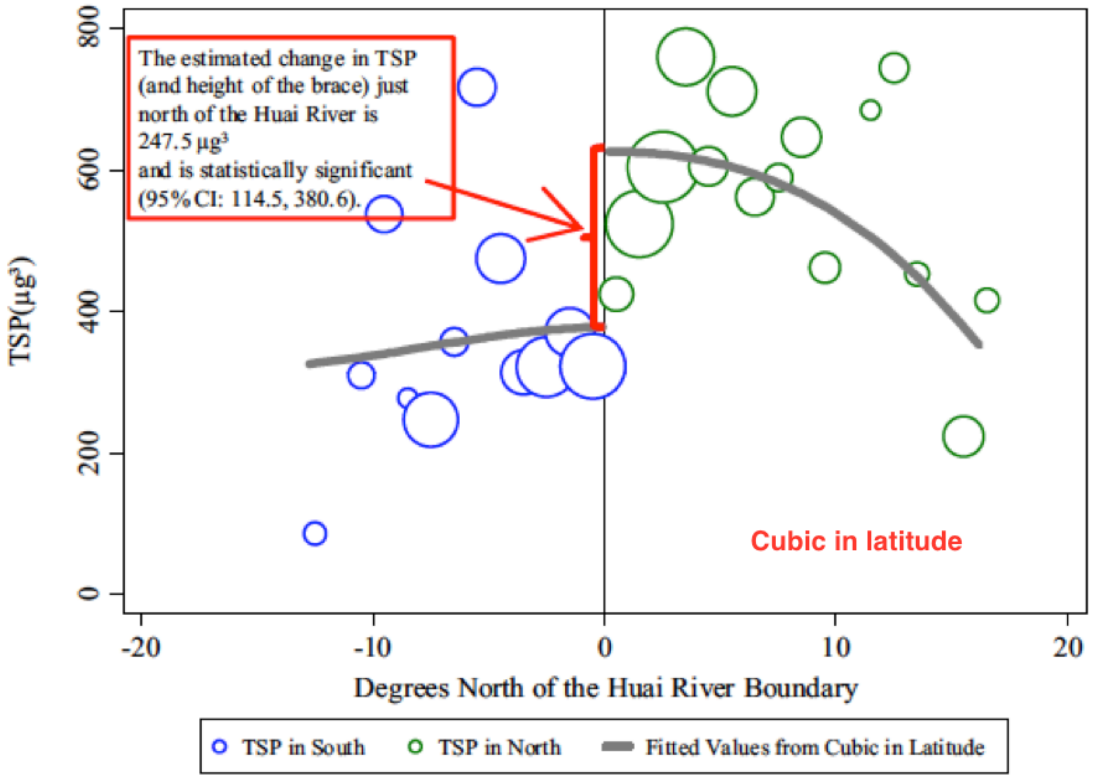


Fig. 2. Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. The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

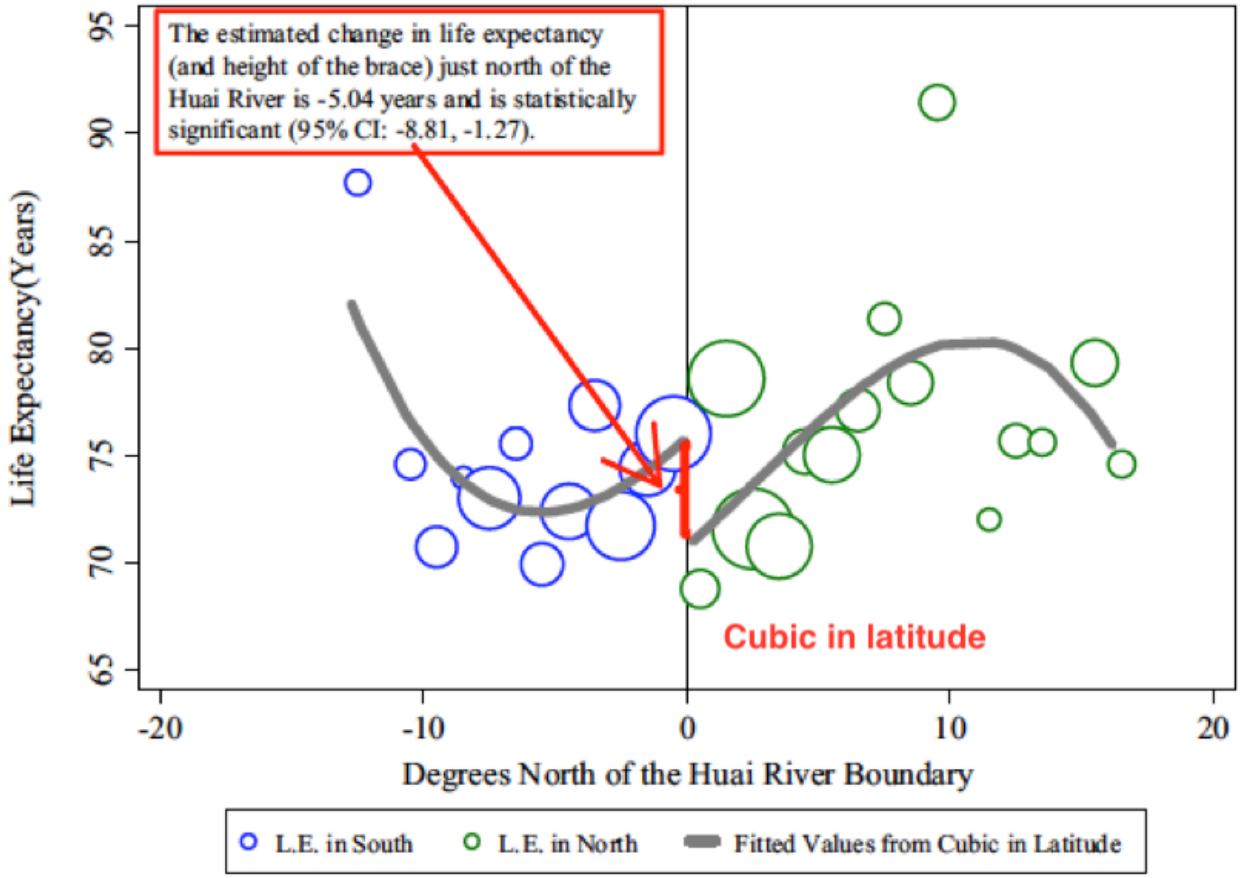


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

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/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 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)

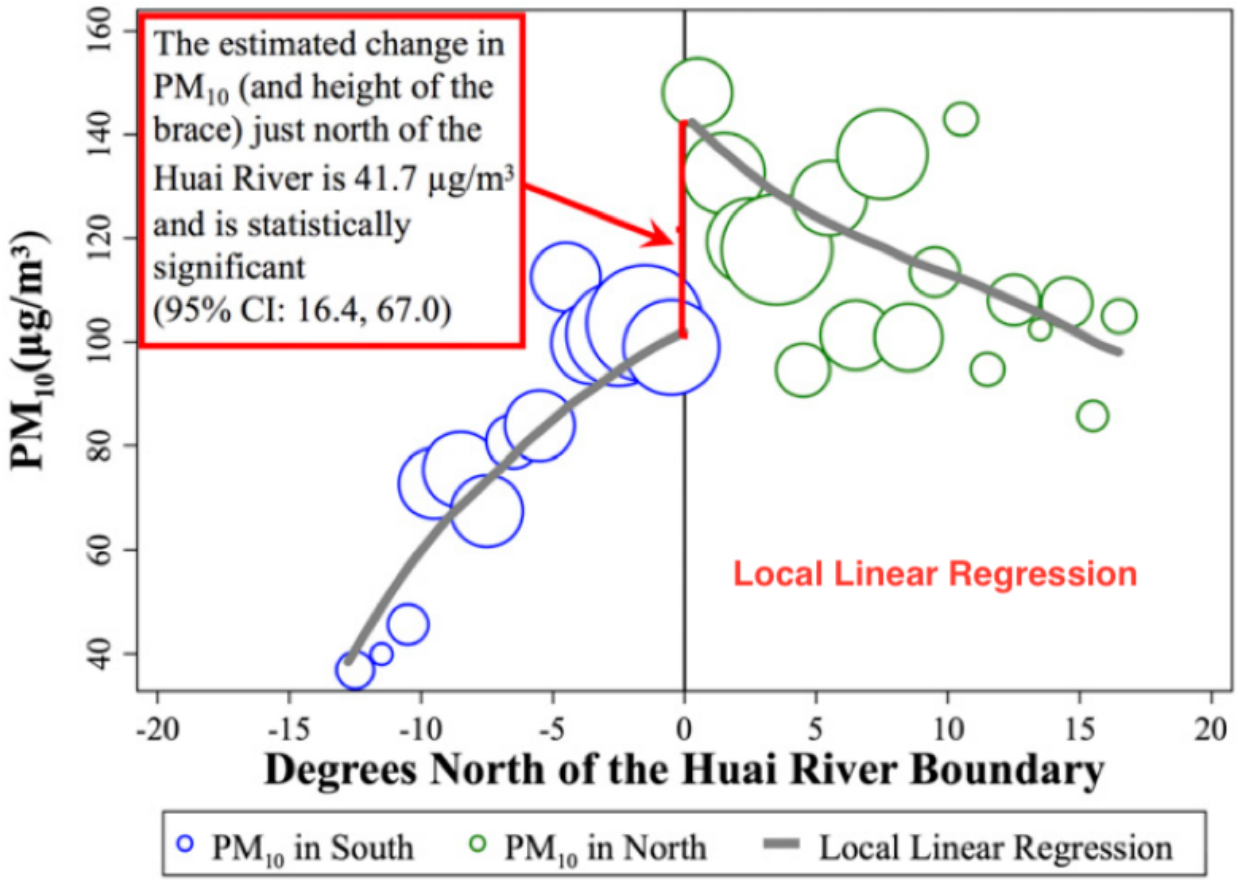


Fig. 2. Fitted values from a local linear regression of PM_{10} exposure on distance from the Huai River estimated separately on each side of the river.

Air pollution in China: Ebenstein et al(2017)

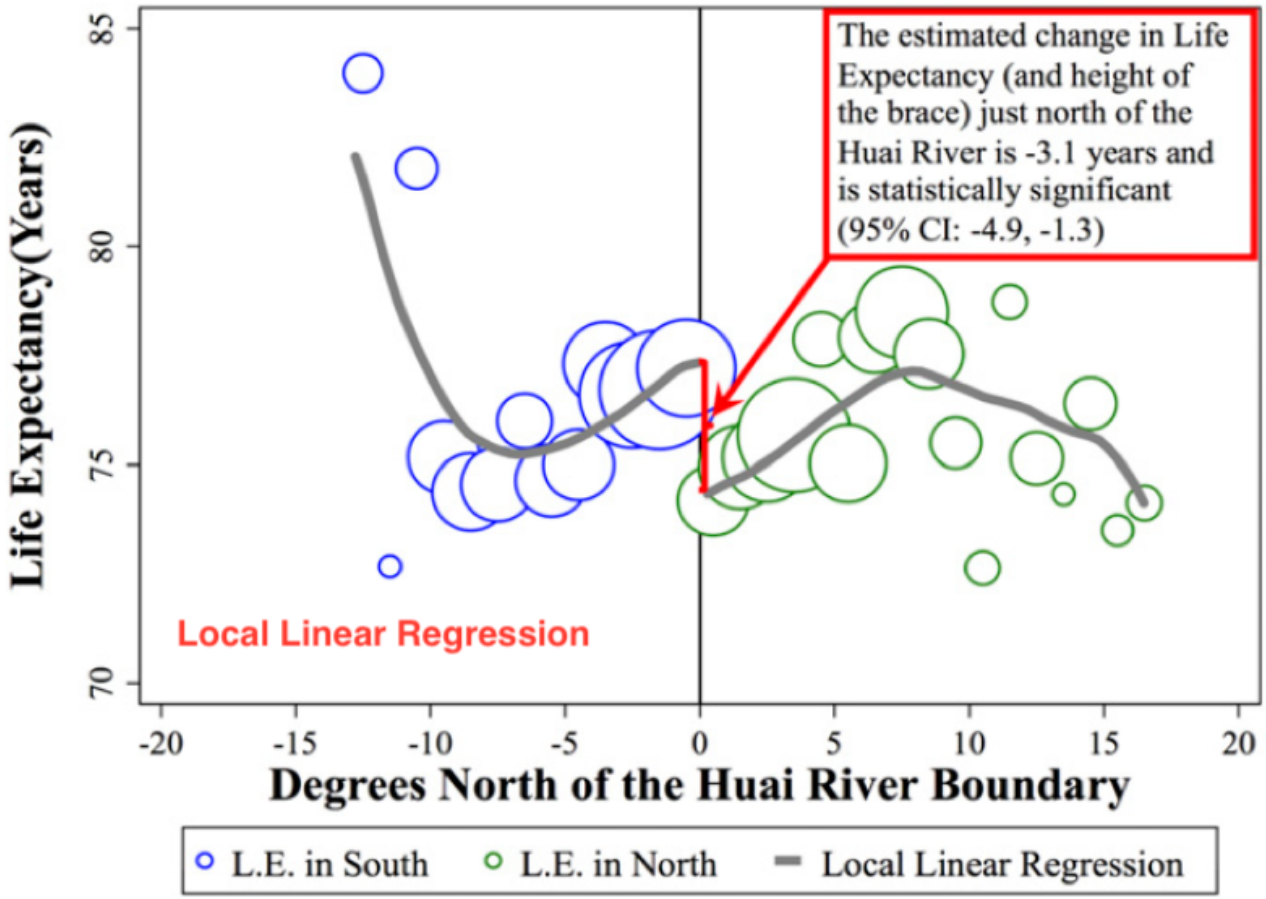


Fig. 3. Fitted values from a local linear regression of life expectancy (L.E.) on distance from the Huai River estimated in the same manner as in Fig. 2.

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

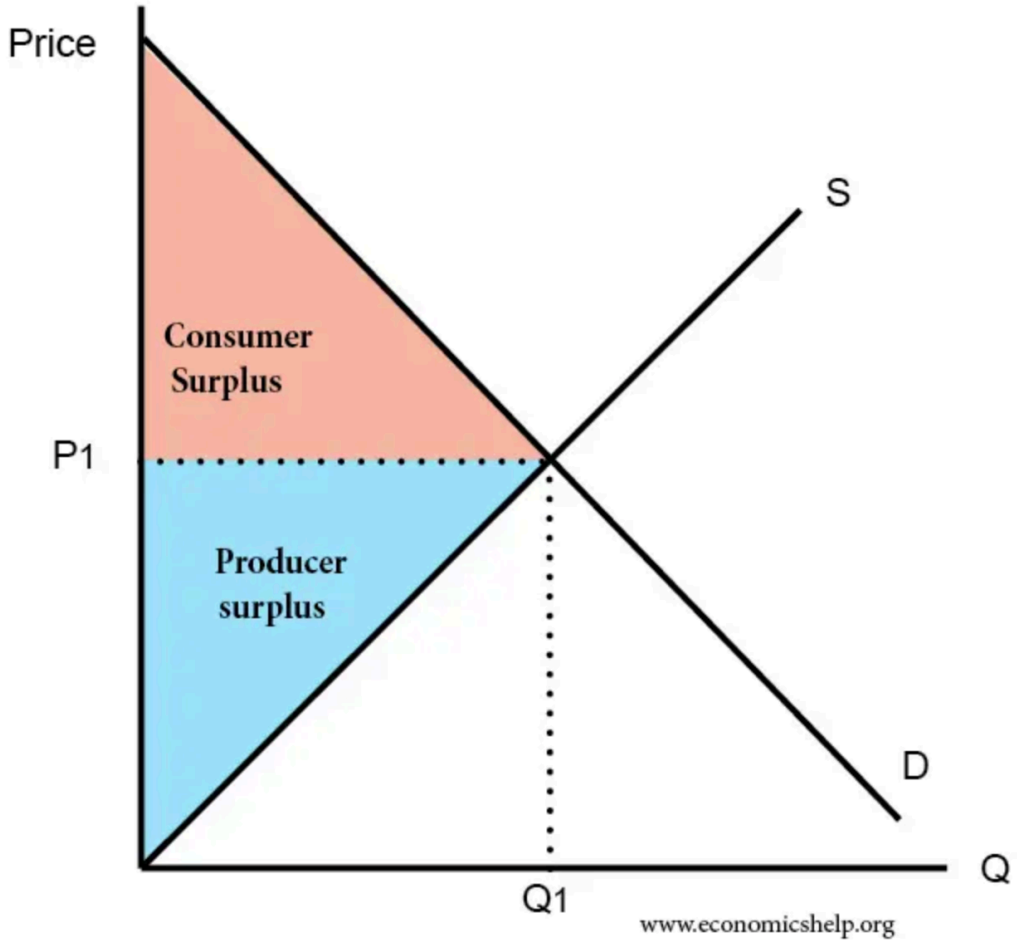
Outcome	[1]	[2]	[3]
Pollution and life expectancy			
PM_{10}	27.4*** (9.5)	31.8*** (9.1)	41.7*** (12.9)
Life expectancy at birth, y	−2.4** (1.0)	−2.2* (1.1)	−3.1*** (0.9)
Cause-specific mortality (per 100,000, log)			
Cardiorespiratory	0.30** (0.14)	0.22* (0.13)	0.37*** (0.11)
Noncardiorespiratory	0.06 (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).

The Big Data Case

Using Big Data to Estimate Consumers' Demand

- Consumer surplus（消费者剩余）: the difference between the price that consumers pay actually and the price that they are willing to pay.



- In practice obtaining convincing empirical estimates of consumer

Using Big Data to Estimate Consumers' Demand

- Uber uses real-time pricing ("surge" pricing) to equilibrate local, short-term supply and demand.
- "surge" pricing means that a consumer wishing to take a particular trip can face prices ranging from the base price (what we call the "no surge" or "1.0x" price) to five or more times higher, depending on local market conditions.
- Importantly, authors can observe detailed information not only for every trip taken using Uber, but also, critically, when a consumer searches for a ride using Uber without ultimately deciding to make a request. thus, observe the price offered to the consumer, and whether she or he accepts or rejects that offer.
- If the degree of surge pricing faced by a consumer on a given trip were generated **at random**, then all that would be required to trace out a demand curve would be to compute the share of Uber searches culminating in a ride at each level of surge pricing.

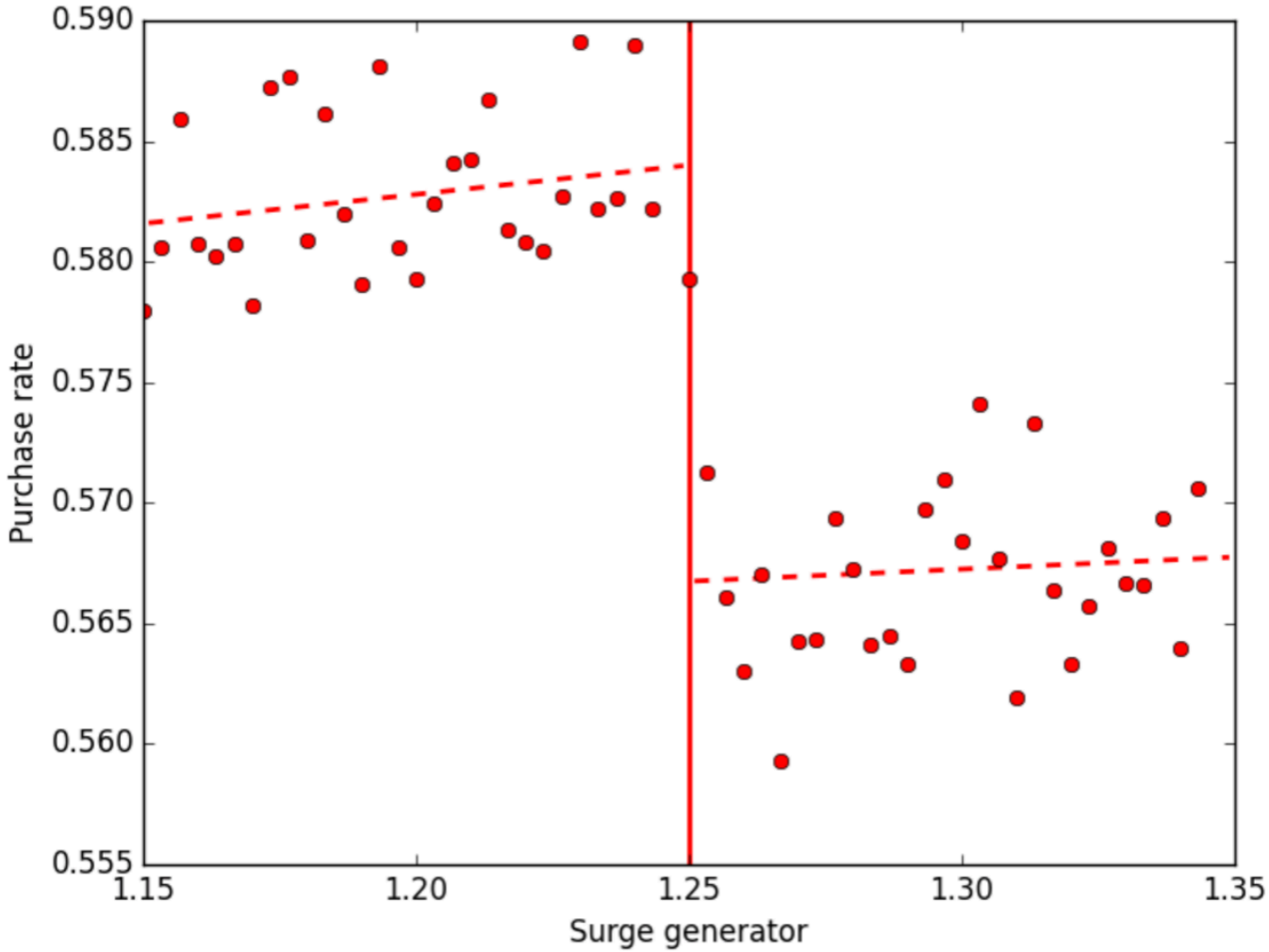
Using Big Data to Estimate Consumers' Demand

- In practice, the surge price that consumers face is not random; it reflects local demand and supply conditions and calculated by Uber's algorithm. Then we can not use the information to calculate the demand curve directly.
- Uber calculates each surge price to an arbitrary number of decimal places, but consumers are presented with discrete price increments to facilitate a simple, easy user experience.
  - For example, Uber might estimate that the appropriate multiplier is 1.61809, but for easy interpretation, they would charge the customer 1.6x.

Where is the discontinuity?

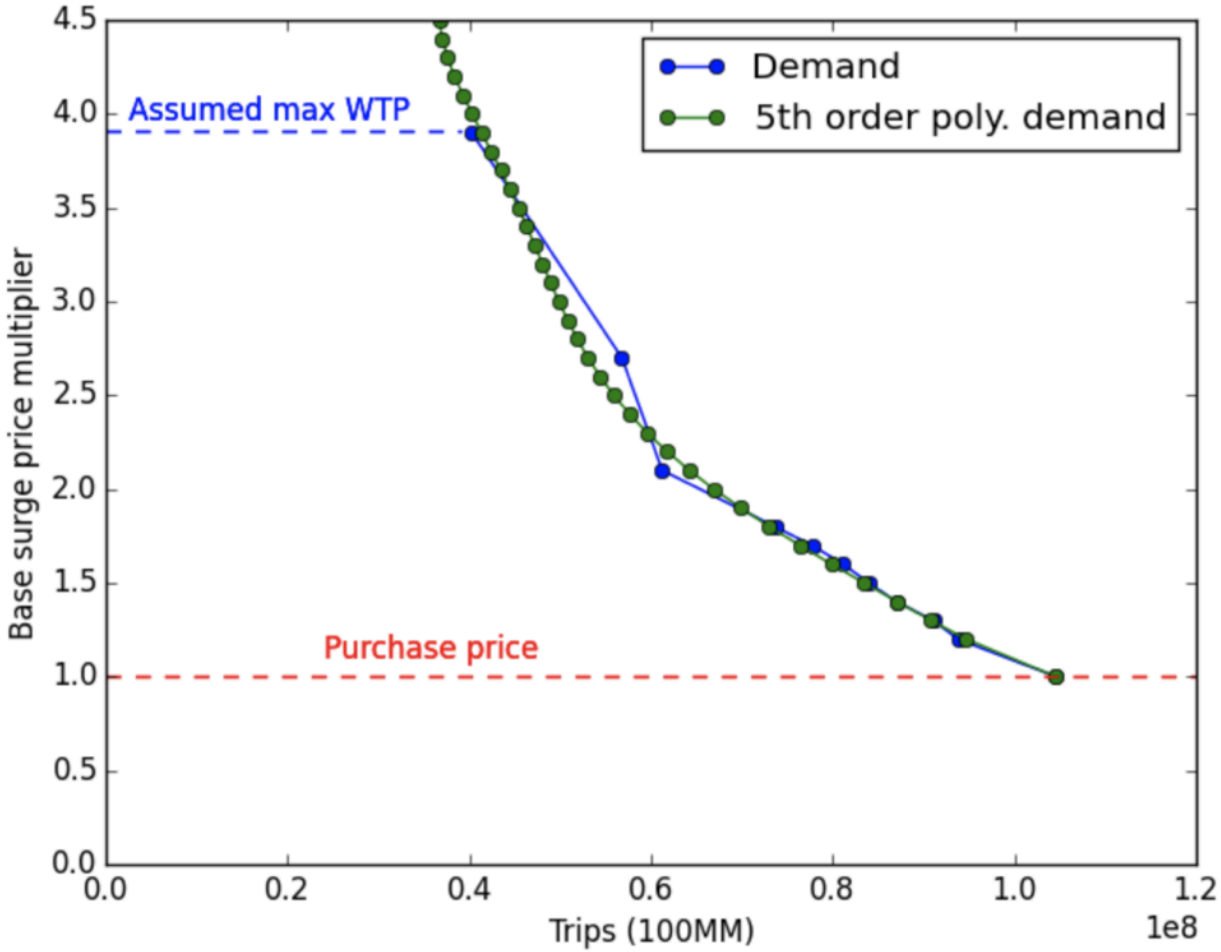
- When the value calculated by Uber's algorithm is very close to a cut-off such as 1.2,
  1. it suggests a surge of 1.251x, then consumers face a 1.3x surge.
  2. it suggests a surge of 1.249x, then consumers face a 1.2x surge.
- which means that the market conditions are nearly identical(the supply is fixed), but consumer may make different decisions.
- It provide a special discontinuity allow authors to local elasticities of demand across the full range of surge prices.

Regression Discontinuity at 1.25X



Using Big Data to Estimate Consumers' Demand

- Further to estimate at every cut-off $(1.15, 1.25, 1.35, 1.45, ...)$



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