断点回归

赵瑾

主要内容

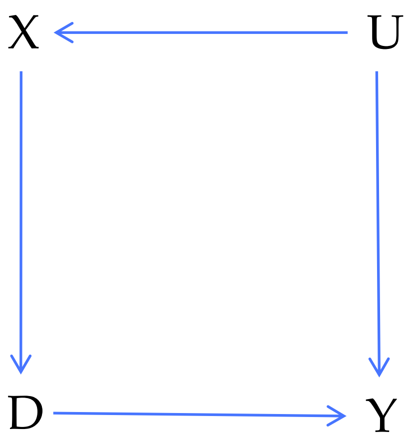
- 精确断点回归
- 模糊断点回归
- 断点回归估计的不同方法
- 带宽选择和滞后阶数
- 模型设定检验

6.1 精确断点回归

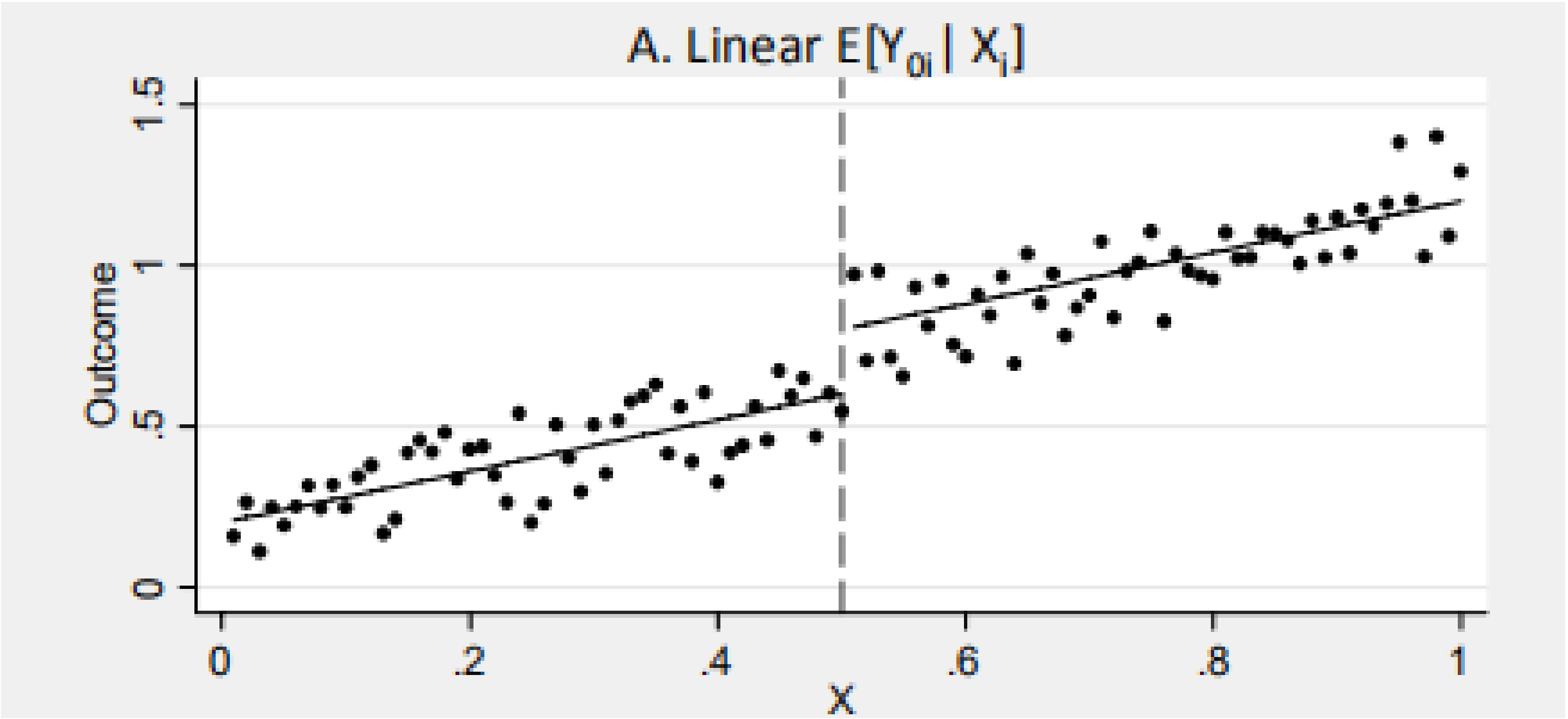
- 定义：干预分配完全由参考变量是否超过临界值决定

$$D_i=\begin{cases}1 & x_i \geqslant x_0\\ 0 & x_i \leqslant x_0\end{cases}$$

- RDD因果图



图形分析



模型设计

- 假设除了分配干预，潜在结果可以被一个线性连续的模型描述（图A）

$$E\left[Y_{0i} \mid x_i\right] = \alpha + \beta x_i$$

$$Y_{1i} = Y_{0i} + \rho$$

- 整理得，回归模型

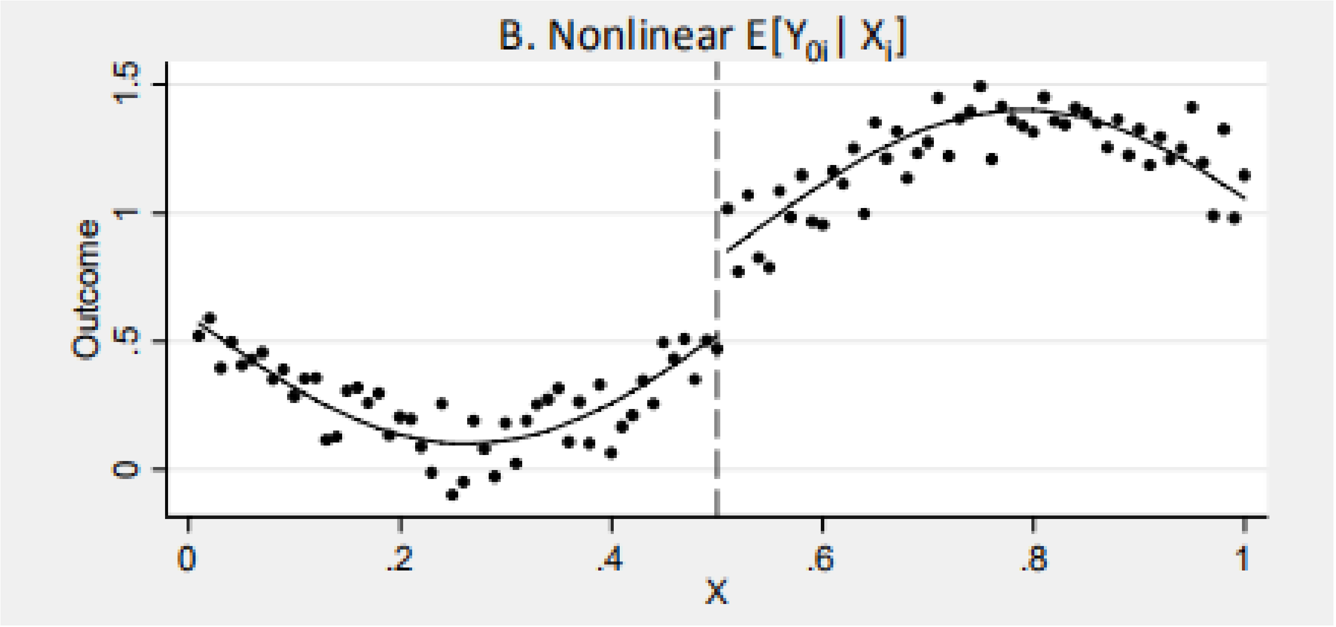
$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i$$

ρ为处理效应，D_i是依赖于x_i的一个函数

模型设计

- 假设除了分配干预，潜在结果可以被一个线性连续的模型描述（图A）
- 如果$E[Y_{0i}|x_i]$是非线性的（图B），那么我们假设

$$E\left[Y_{0i} \mid x_i\right] = f\left(x_i\right)$$



模型设计

- 修改后的回归模型

$$Y_i = f\left(x_i\right) + \rho D_i + \eta_i$$

ρ为处理效应，D_i是依赖于x_i的一个函数

- 注：$E[Y_{0i}|x_i]$一定要是连续函数
- $Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \rho D_i + \eta_i \qquad (6.1.4)$

其中 $\tilde{x}_i \equiv x_i - x_0$

模型设计

- 用n阶多项式来拟合

$$E\left[Y_{0i} \mid x_i\right] = f_0\left(x_i\right) = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p}$$

$$E\left[Y_{1i} \mid x_i\right] = f_1\left(x_i\right) = \alpha + \rho + \beta_{11}\tilde{x}_i + \beta_{12}\tilde{x}_i^{\,2} + \cdots + \beta_{1p}\tilde{x}_i^{\,p}$$

其中 $\tilde{x}_i \equiv x_i - x_0$

- $E\left[Y_i \mid x_i\right] = E\left[Y_{0i} \mid x_i\right] + \left(E\left[Y_{1i} \mid x_i\right] - E\left[Y_{0i} \mid x_i\right]\right) D_i$

- 代入得

$$\begin{aligned} E\left[Y_i \mid x_i\right] &= E\left[Y_{0i} \mid x_i\right] + \left(E\left[Y_{1i} \mid x_i\right] - E\left[Y_{0i} \mid x_i\right]\right) D_i \\ &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \\ &\quad \left(\rho + \beta_1^*\tilde{x}_i + \beta_2^*\tilde{x}_i^{\,2} + \cdots + \beta_p^*\tilde{x}_i^{\,p}\right) D_i \end{aligned}$$

其中 $\beta_i^* = \beta_{1i} - \beta_{0i}, i = 1, 2, \ldots, p$

模型设计

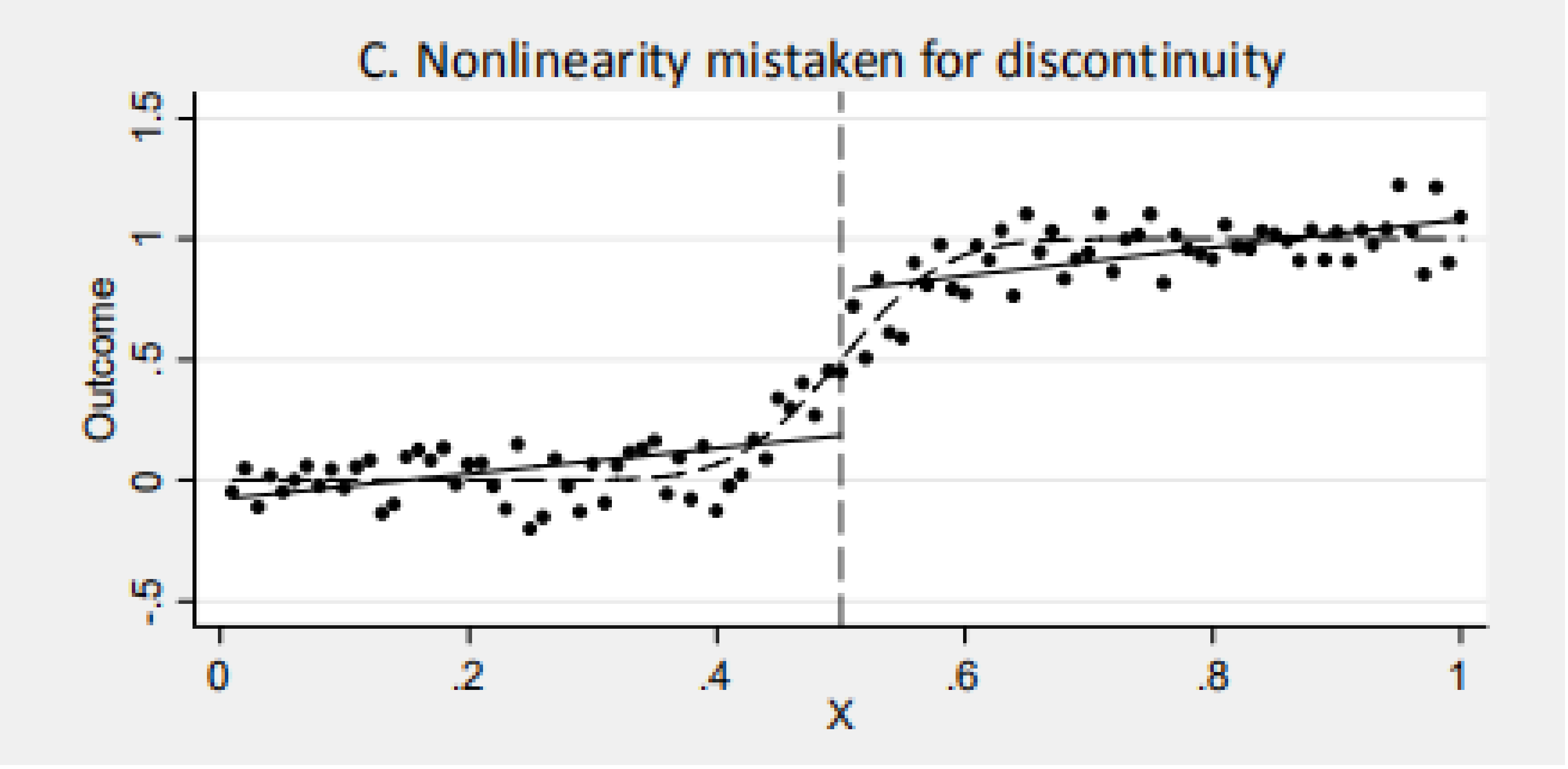
- $$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \left(\rho + \beta_1^{\,*}\tilde{x}_i + \beta_2^{\,*}\tilde{x}_i^{\,2} + \cdots + \beta_p^{\,*}\tilde{x}_i^{\,p}\right)D_i + \eta_i \tag{6.1.6}$$

  其中 $\tilde{x}_i \equiv x_i - x_0$ ，η为误差项

- $Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \rho D_i + \eta_i$ 是(6.1.6)的一个特例

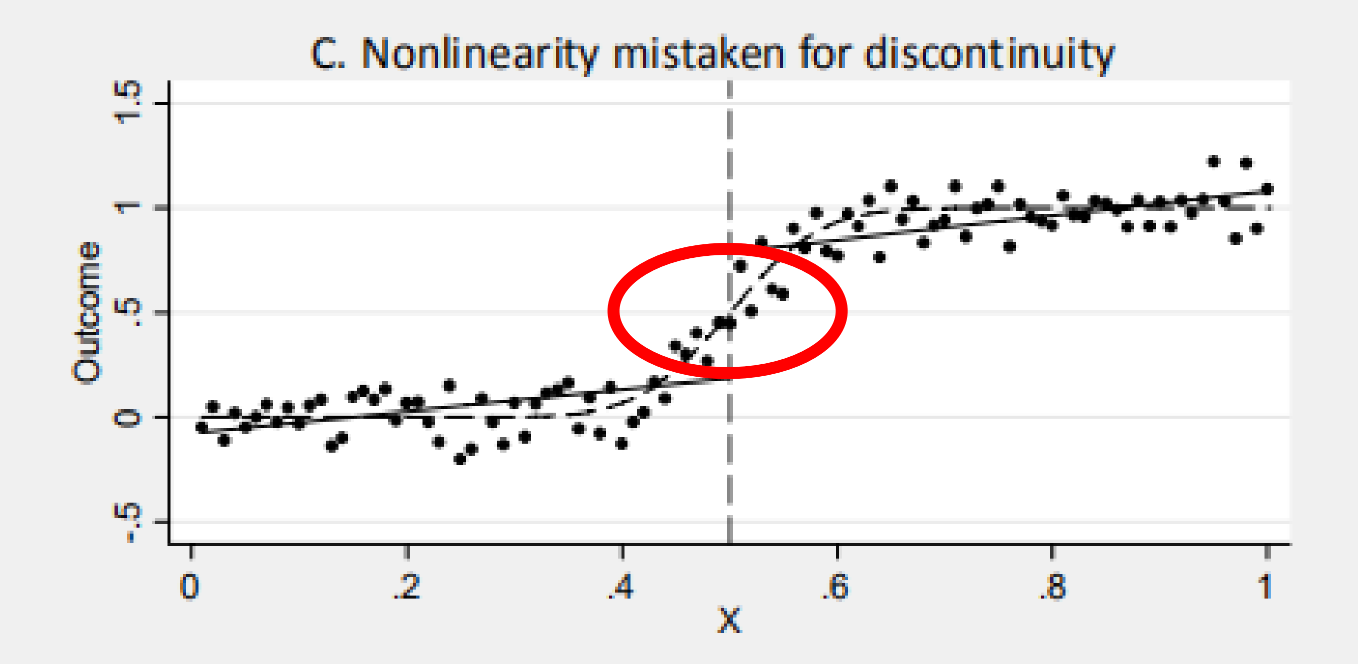
模型设计

- 如图C中所示，如果模型拟合不当，$E[Y_{0i}|x_i]$的一个急剧的转折就会被误认为是一个断点



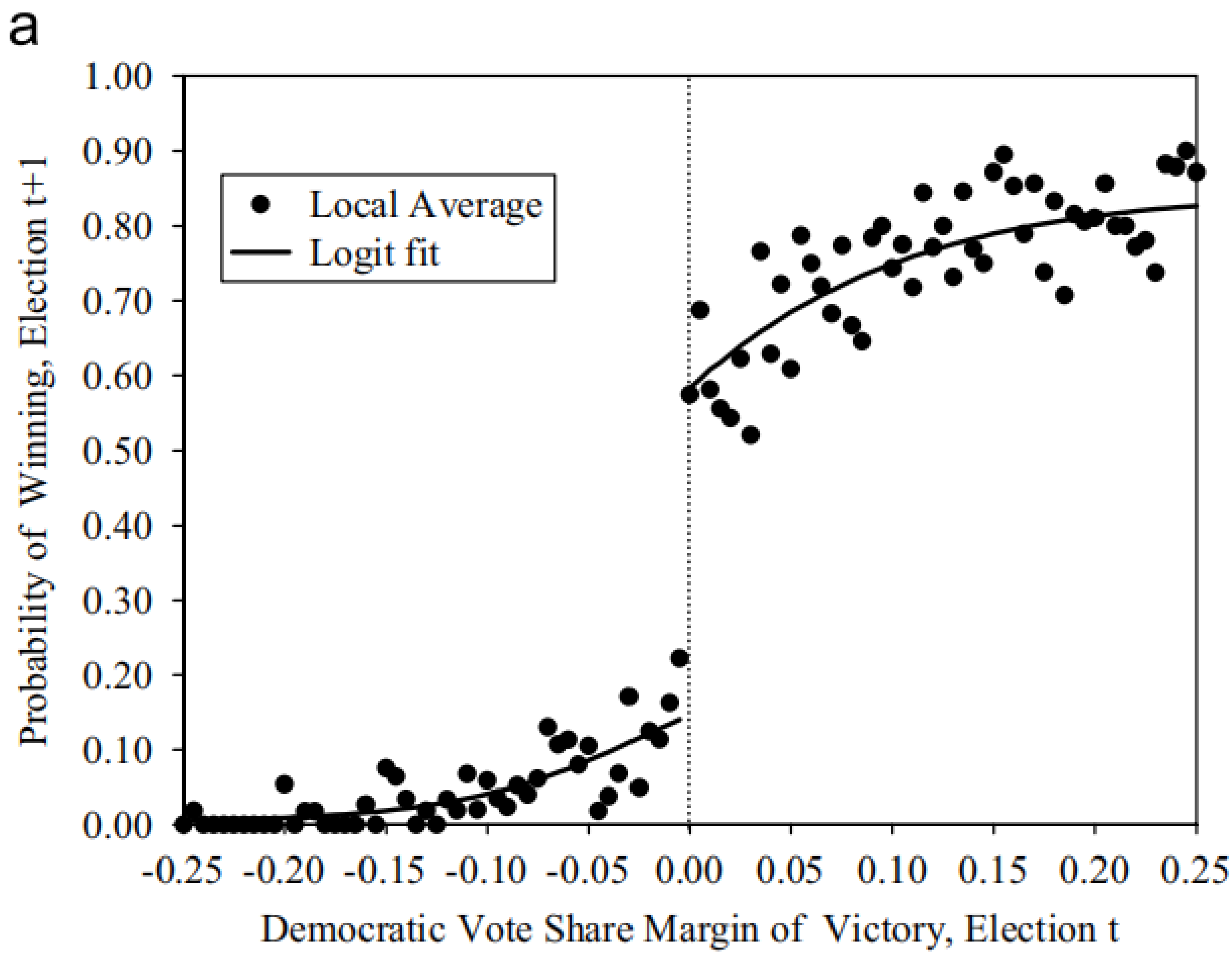
模型设计

- 为了避免这种情况，我们仅关注断点附近的一些点

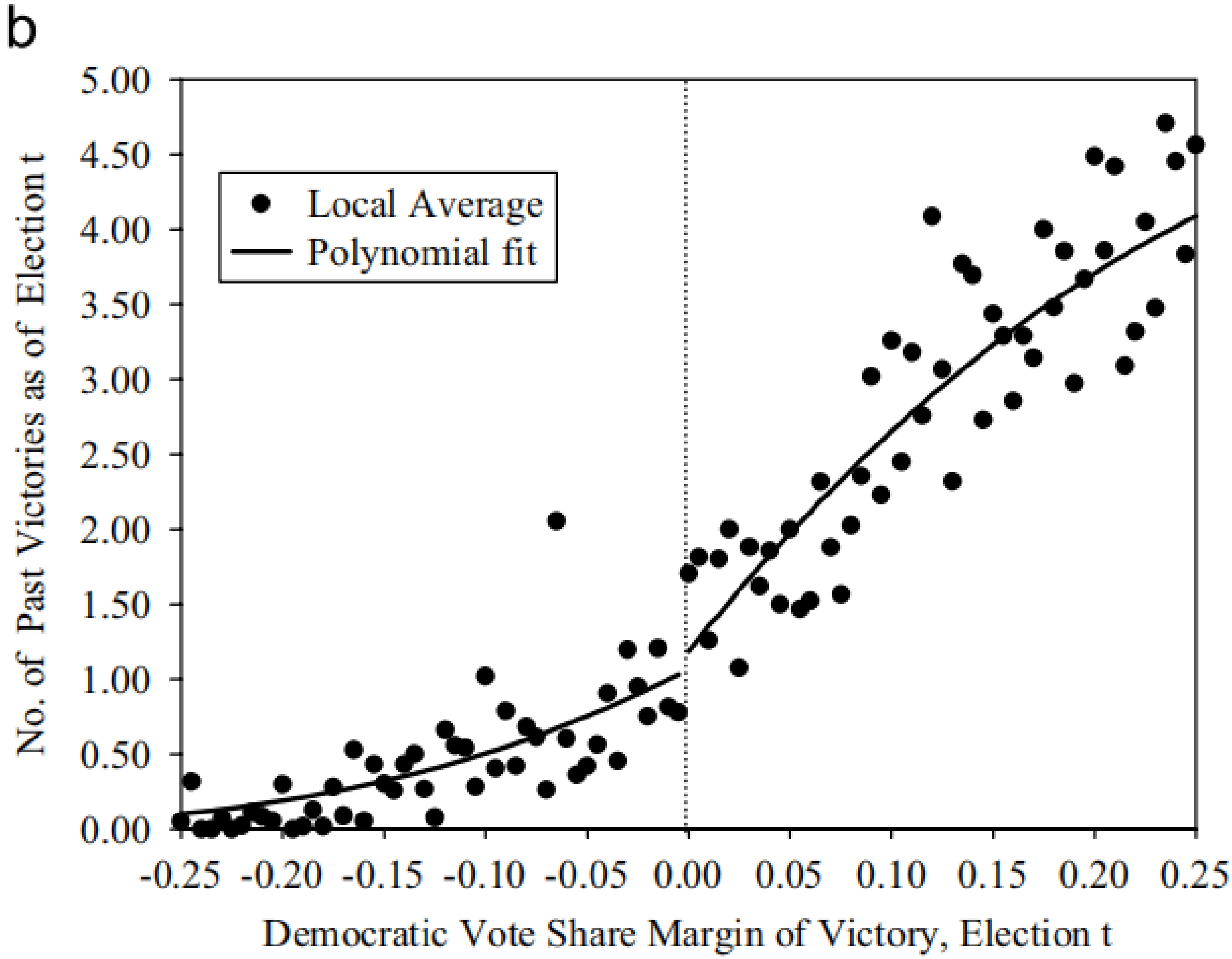


- $$\lim_{\Delta \to 0} E\left[Y_i \mid x_0 \leqslant x_i < x_0 + \Delta\right] - E\left[Y_i \mid x_0 - \Delta \leqslant x_i < x_0\right]$$
  $$= E\left[Y_{1i} - Y_{0i} \mid x_i = x_0\right]$$

一个例子：在位党在竞选中是否具有优势？



一个例子：在位党在竞选中是否具有优势？



小结：精确断点回归

- 模型：

$$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \left(\rho + \beta_1^{\,*}\tilde{x}_i + \beta_2^{\,*}\tilde{x}_i^{\,2} + \cdots + \beta_p^{\,*}\tilde{x}_i^{\,p}\right)D_i + \eta_i$$

$$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^{\,2} + \cdots + \beta_{0p}\tilde{x}_i^{\,p} + \rho D_i + \eta_i$$

- 重要假设
  - 断点假设
  - 连续性假设
  - 局部随机化假设

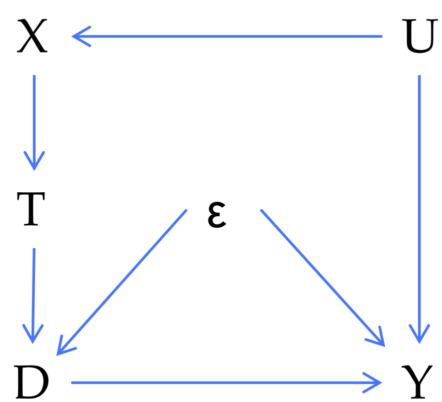
6.2 模糊断点回归

- 定义：干预分配不完全由参考变量决定，还受到其他观测不到的因素的影响

$$P(D_i = 1 \mid x_i) = \begin{cases} g_1(x_i) & x_i \geq x_0 \\ g_0(x_i) & x_i \leq x_0 \end{cases}$$

$$g_1(x_i) \neq g_0(x_i)$$

- RDD因果图



- 方法：2SLS（将T作为工具变量）

模型设计

- 假设$g_1(x_i) \geq g_0(x_i)$，即$x_i \geq x_0$在时，个体更容易进入处理组

$$E[D_i \mid x_i] = P(D_i = 1 \mid x_i) = g_0(x_i) + [g_1(x_i) - g_0(x_i)]T_i$$

$$T_i = 1(x_i \geq x_0)$$

- 假设$g_1(x_i)$和$g_0(x_i)$都是p阶多项式

$$g_0(x_i) = \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^2 + \cdots + \gamma_{0p}x_i^p$$

$$g_1(x_i) = \gamma_{10} + \gamma_{11}x_i + \gamma_{12}x_i^2 + \cdots + \gamma_{1p}x_i^p$$

此处的x_i没有被中心化

模型设计

- 代入得

$$E\left[D_i \mid x_i\right] = g_0\left(x_i\right) + \left[g_1\left(x_i\right) - g_0\left(x_i\right)\right]T_i$$

$$= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^{\ 2} + \cdots + \gamma_{0p}x_i^{\ p}$$

$$+\left[\pi + \gamma_1 x_i^* + \gamma_2 x_i^{*2} + \cdots + \gamma_p x_i^{*p}\right]T_i$$

$$= \gamma_{00} + \gamma_{01}x_i + \gamma_{02}x_i^{\ 2} + \cdots + \gamma_{0p}x_i^{\ p}$$

$$+\pi T_i + \gamma_1 x_i^* T_i + \gamma_2 x_i^{*2} T_i + \cdots + \gamma_p x_i^{*p} T_i$$

- 为了简化，我们仅选用工具变量T_i，忽略交互项
- 去掉期望符号得,第一阶段的回归模型为

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^{\ 2} + \cdots + \gamma_p x_i^{\ p} + \pi T_i + \xi_{1i}$$

模型设计

- 第二阶段回归

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + \cdots + \kappa_p x_i^p + \pi_i T_i + \xi_{2i}$$

其中 $\mu = \alpha + \rho\gamma_0$, $\kappa_j = \beta_j + \rho\gamma_j, j = 1, \ldots, p$

- 证明：将 $D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \cdots + \gamma_p x_i^p + \pi T_i + \xi_{1i}$

代入 $Y_i = \alpha + \beta_{01} x_i + \beta_{02} x_i^2 + \cdots + \beta_{0p} x_i^p + \rho D_i + \eta_i$

得 $$\begin{aligned} Y_i &= \alpha + \beta_{01} x_i + \beta_{02} x_i^2 + \cdots + \beta_{0p} x_i^p \\ &\quad + \rho\left(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \cdots + \gamma_p x_i^p + \pi T_i + \xi_{1i}\right) + \eta_i \\ &= \mu + \kappa_1 x_i + \kappa_2 x_i^2 + \cdots + \kappa_p x_i^p + \pi_i T_i + \xi_{2i} \end{aligned}$$

模型设计

- 如果对x_i进行中心化处理，即 $\tilde{x}_i \equiv x_i - x_0$

那么

$$D_i = \gamma_{00} + \gamma_{01}\tilde{x}_i + \gamma_{02}\tilde{x}_i^{\,2} + \cdots + \gamma_{0p}\tilde{x}_i^{\,p} + \left(\pi + \gamma_1^{*}\tilde{x}_i + \gamma_2^{*}\tilde{x}_i^{\,2} + \cdots + \gamma_p^{*}\tilde{x}_i^{\,p}\right)T_i + \xi_{1i}$$

参数估计

- 边界非参数回归

$$E[Y_i \mid x_0 \le x_i < x_0 + \Delta] - E[Y_i \mid x_0 - \Delta < x_i < x_0] \cong \rho\pi$$

$$E[D_i \mid x_0 \le x_i < x_0 + \Delta] - E[D_i \mid x_0 - \Delta < x_i < x_0] \cong \pi$$

- 因此，

$$\rho = \lim_{\Delta \to 0} \frac{E[Y_i \mid x_0 \le x_i < x_0 + \Delta] - E[Y_i \mid x_0 - \Delta < x_i < x_0]}{E[D_i \mid x_0 \le x_i < x_0 + \Delta] - E[D_i \mid x_0 - \Delta < x_i < x_0]}$$

- 上式是Wald估计量的一种，测度了x_0附近的处理效应
- 在小范围内，该参数估计同样适用于精确断点估计

一个例子：班级规模对成绩的影响Angrist and Lavy(1999)

利用以色列教育系统的一项制度（Maimonides' rule）进行断点回归；该制度限定班级规模的上限为40名学生，一旦超过40名学生（比如41名学生），则该班级被一分为二

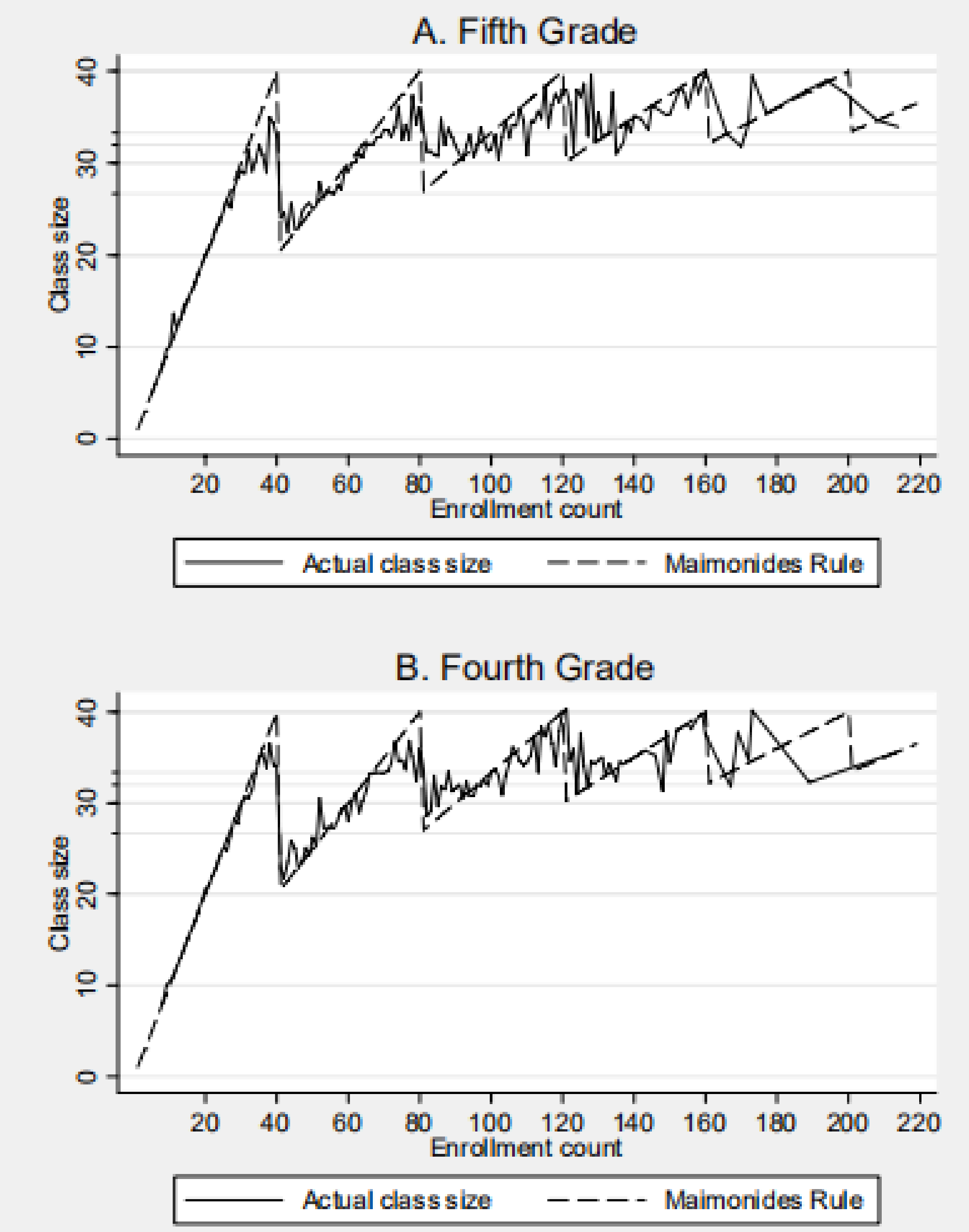
- Maimonides' rule

$$m_{sc} = \frac{e_s}{\mathrm{int}\left[\frac{(e_s - 1)}{40}\right] + 1}$$

- 模型

$$Y_{isc} = \alpha_0 + \alpha_1 d_s + \beta_1 e_s + \beta_2 e_s^2 + \cdots + \beta_p e_s^p + p n_{ns} + \eta_{isc}$$

  - m_{sc}指预测的s学校c班级的班级规模(T)
  - e_s指入学人数(x)
  - n_{sc}指实际的s学校c班级的班级规模(D)



A. Fifth Grade
Class size
0
10
20
30
40
20
40
60
80
100
120
140
160
180
200
220
Enrollment count
Actual class size
Maimonides Rule
B. Fourth Grade
Class size
0
10
20
30
40
20
40
60
80
100
120
140
160
180
200
220
Enrollment count
Actual class size
Maimonides Rule

Table 6.2.1: OLS and fuzzy RD estimates of the effects of class size on fifth grade math scores

	OLS			2SLS				
				Full sample		Discontinuity samples		
						+/- 5		+/- 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean score		67.3		67.3		67.0		67.0
(s.d.)		(9.6)		(9.6)		(10.2)		(10.6)
Regressors								
Class size	.322	.076	.019	-.230	-.261	-.185	-.443	-.270
	(.039)	(.036)	(.044)	(.092)	(.113)	(.151)	(.236)	(.281)
Percent disadvantaged		-.340	-.332	-.350	-.350	-.459	-.435	
		(.018)	(.018)	(.019)	(.019)	(.049)	(.049)	
Enrollment			.017	.041	.062		.079	
			(.009)	(.012)	(.037		(.036)	
Enrollment squared/100					-.010			
					(.016)			
Segment 1								-12.6
(enrollment 36-45)								(3.80)
Segment 2								-2.89
(enrollment 76-85)								(2.41)
Root MSE	9.36	8.32	8.30	8.40	8.42	8.79	9.10	10.2
R-squared	.048	.249	.252					
N		2,018		2,018		471		302

Notes: Adapted from Angrist and Lavy (1999). The table reports estimates of equation (6.2.6) in the text using class averages. Standard errors, reported in parentheses, are corrected for within-school correlation.

The impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

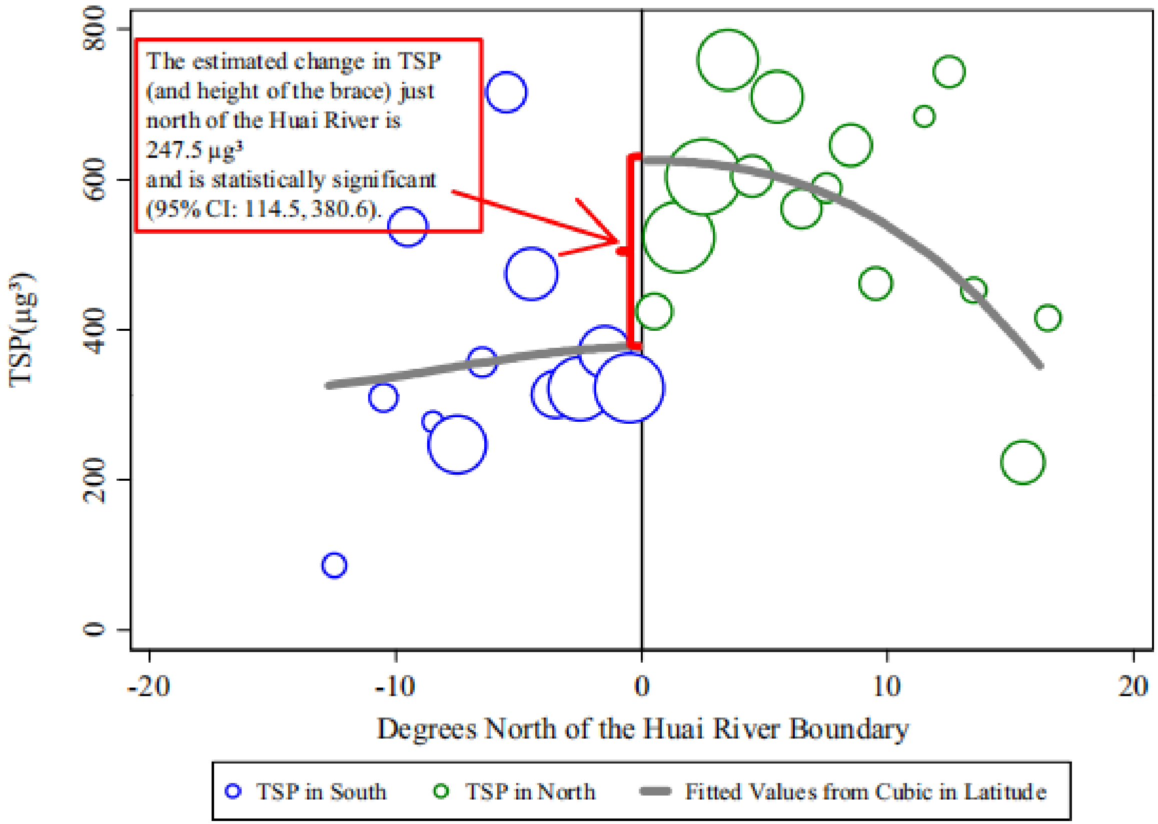


Fig. 2. Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. The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

The impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

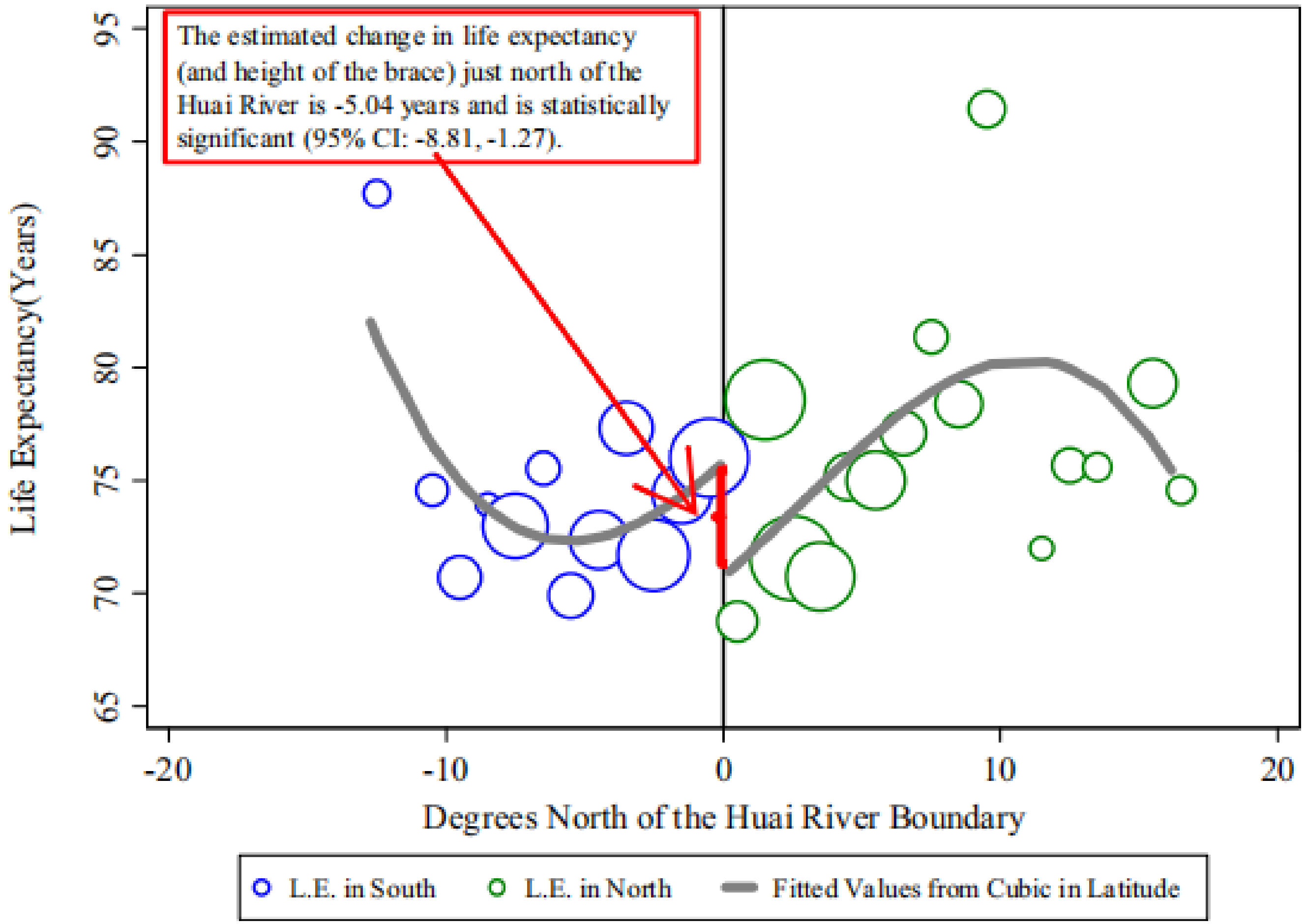


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

The impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

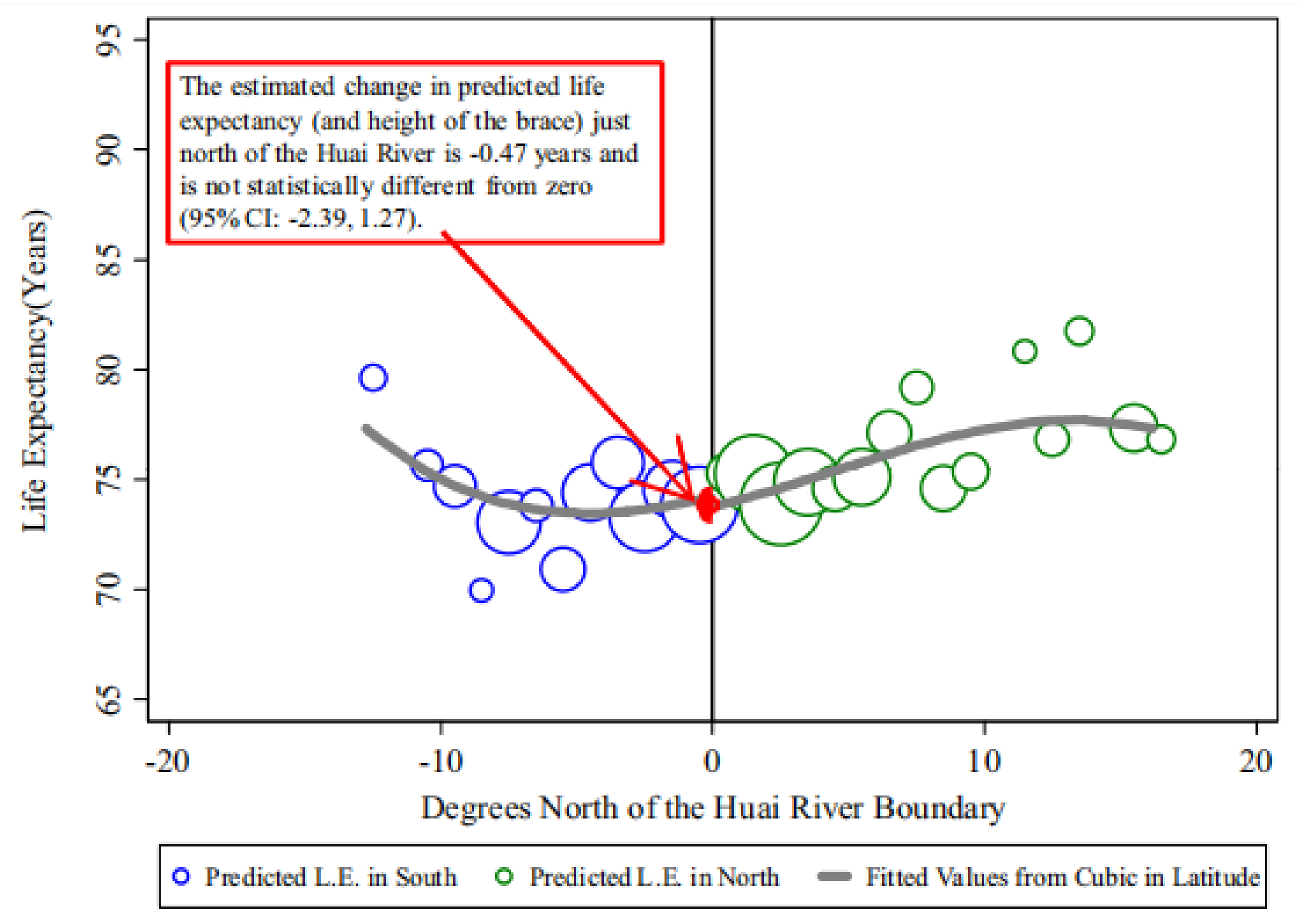


Fig. 4. The plotted line reports the fitted values from a regression of predicted life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location. Predicted life expectancy is calculated by OLS using demographic and meteorological covariates (excluding TSPs).

The impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

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).

小结：模糊断点回归

- 模型：

$$D_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^{\ 2} + \cdots + \gamma_p x_i^{\ p} + \pi T_i + \xi_{1i}$$

$$Y_i = \mu + \kappa_1 x_i + \kappa_2 x_i^2 + \cdots + \kappa_p x_i^p + \pi_i T_i + \xi_{2i}$$

- 重要假设
  - 断点假设
  - 连续性假设
  - 独立性假设
  - 单调性假设
- 参数估计

$$\rho = \lim_{\Delta \to 0} \frac{E\left[Y_i \mid x_0 \leqslant x_i < x_0 + \Delta\right] - E\left[Y_i \mid x_0 - \Delta < x_i < x_0\right]}{E\left[D_i \mid x_0 \leqslant x_i < x_0 + \Delta\right] - E\left[D_i \mid x_0 - \Delta < x_i < x_0\right]}$$

估计处理效应的不同方法

- ✓ 边界非参数回归
  - 参数实际上是断点左右两边结果变量的平均值之差
  - 在边界点上表现不好
- ✓ 局部线性回归
  - 线性回归函数是条件期望函数非常好的近似
  - 可以引入其他的协变量Z_i
- ✓ 局部多项式回归
  - 有时断点附近样本量太少，我们不得不选择较大的带宽，线性估计会造成较大偏差

带宽选择

- 太小
  - 断点左右的个体特征差异较小，估计偏差较小
  - 样本容量可能较小，估计精度降低
- 太大
  - 样本容量较大，估计精度提高
  - 个体特征差异较大，估计偏差降低
- 交叉验证方法

滞后阶数

- AIC标准、AICC标准、BIC标准
- 一般而言，带宽越大，需要选择的滞后阶数越大；带宽越小，滞后阶数越小

模型设定检验

- 协变量连续性检验（伪结果检验）
- 参考变量分布连续性检验
  - 连续，意味着个体没有精准操控参考变量的能力
- 伪断点检验
- 带宽选择的敏感性检验