

武汉大学金融系2022秋季学期
金融学/金融工程专业博士生方法论专题课

第五讲：断点回归

授课人：刘 岩

2022年10月24日

本讲内容

- 断点回归的基础理论与Stata应用
- 实例1：淮河取暖分界线的健康效应
- 实例2：信用评分、贷款决策与经济后果

断点回归的基础理论与Stata应用

概述

- 断点回归(Regression Discontinuity Design, RD/RDD)是一种研究非随机实验但接近随机实验数据的方法
- 适用：研究某些特定社会科学事件的后果
 - 事件特点：个体是否受到事件的影响取决于某个可观测特征是否大于给定临界值
- 分类
 - 清晰断点回归(Sharp RDD)
 - 模糊断点回归(Fuzzy RDD)

断点回归的直观理解：一个例子

- 假设政府有一个针对低收入人群的医疗福利，政府根据病人收入为其评分，高评分代表高收入；同时病人健康指数越高代表越健康

- 平均接受治疗的潜在健康结果

$$\mathbb{E}(Y(1)|X = x) = f(x)$$

- 平均未接受治疗的潜在健康结果

$$\mathbb{E}(Y(0)|X = x) = g(x)$$

- 两回归曲线间的距离代表给定收入水平的平均治疗效果，例如，对收入 $X_i = 30$ 的病人，平均治疗效果为：

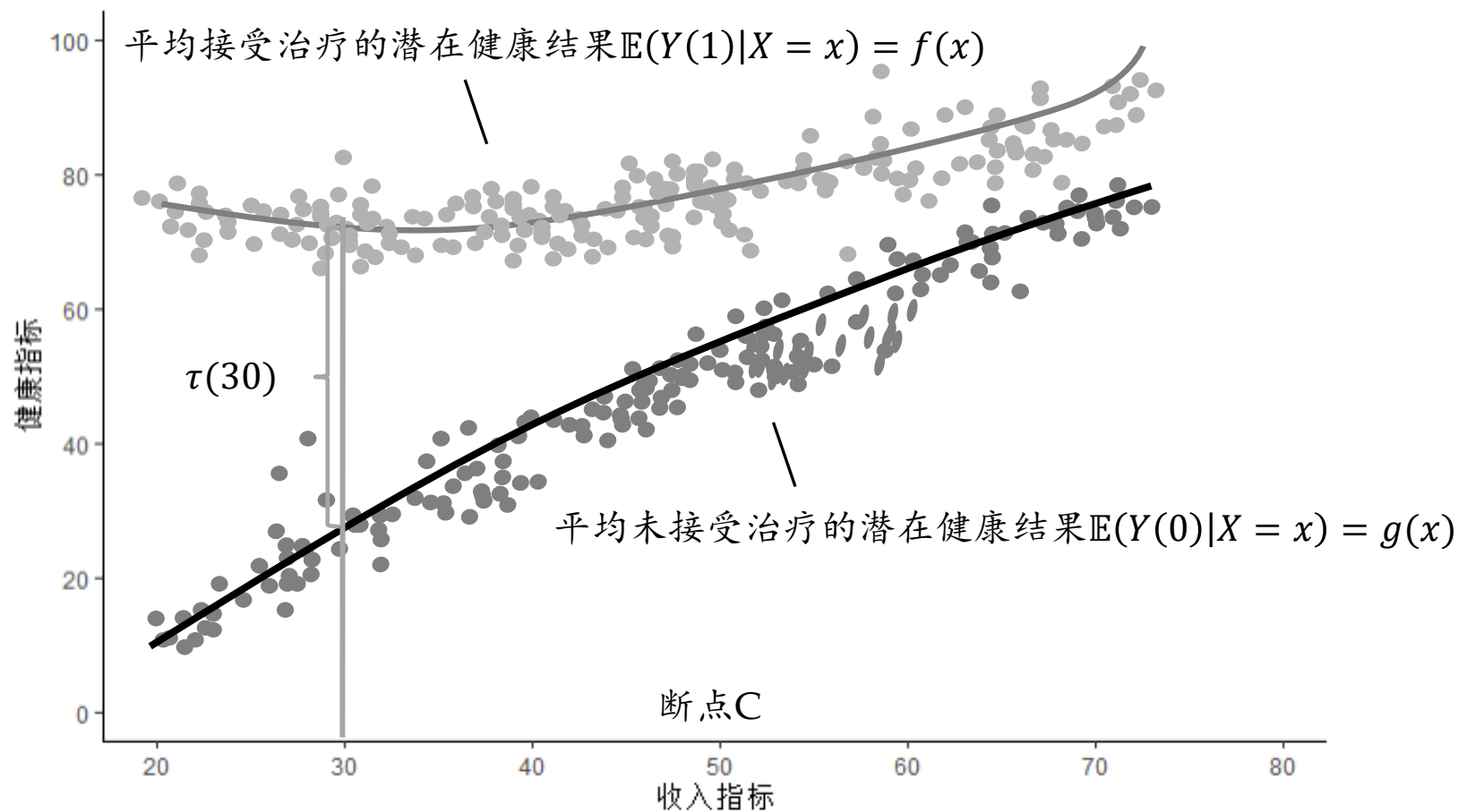
$$\tau(30) = \mathbb{E}(Y_i(1)|X_i = 30) - \mathbb{E}(Y_i(0)|X_i = 30) = f(30) - g(30)$$

- 总体平均治疗效果

$$ATE = \sum_{x=20}^{80} P(x)\tau(x)$$

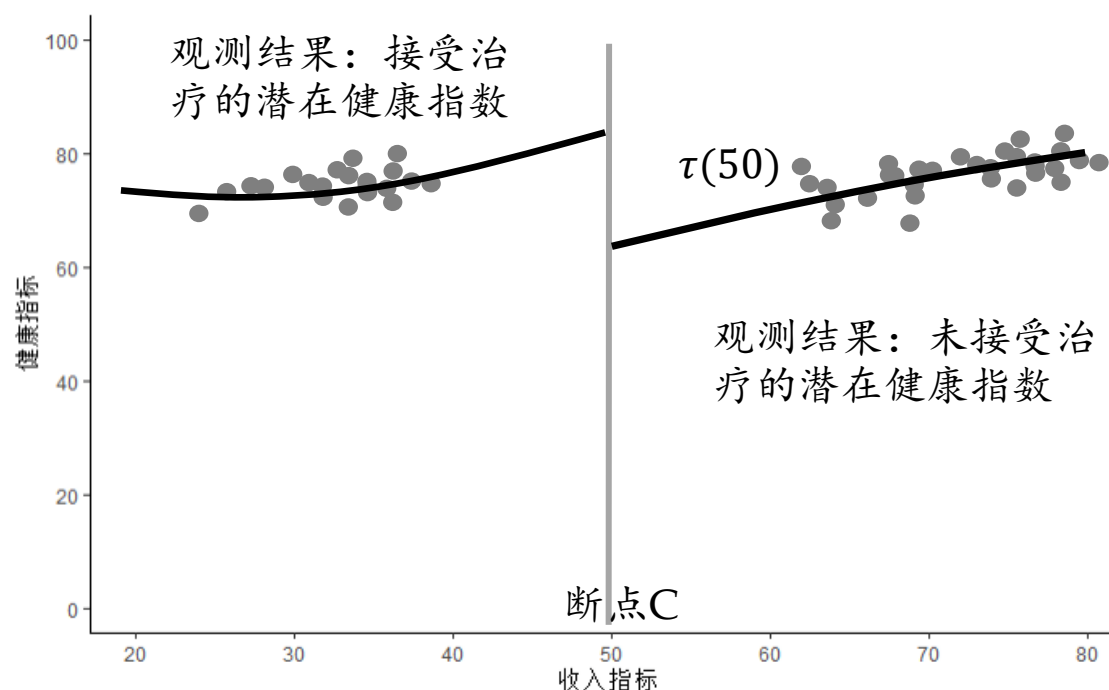
$P(x)$: 不同收入的人数比率

断点回归的直观理解：一个例子



断点回归的直观理解：一个例子

- 实际情况：政府只对收入水平低于或等于50的病人提供治疗，收入高于50的未接受治疗，因此收入50被称为断点C



断点回归的直观理解：一个例子

■ 估计处置效应

□ 用 $X_i = 50$ 的稍左侧接受治疗的平均观测值去近似 $E(Y_i(1)|X_i = 50)$ ，用 $X_i = 50$ 稍右侧未接受治疗的平均观测值去近似 $E(Y_i(0)|X_i = 50)$

□ 治疗效果

$$\begin{aligned}\tau(50) &= E(Y_i(1)|X_i = 50) - E(Y_i(0)|X_i = 50) \\ &\approx E(Y_i(1)|X_i = 49.9) - E(Y_i(0)|X_i = 50.1)\end{aligned}$$

□ 两个理解要点

➤ 连续性

$$\begin{aligned}\tau_c &= E(Y_i(1)|X_i = C) - E(Y_i(0)|X_i = C) \\ &= \lim_{x \rightarrow C^-} E(Y_i(1)|X_i = x) - \lim_{x \rightarrow C^+} E(Y_i(0)|X_i = x)\end{aligned}$$

➤ 局部随机性

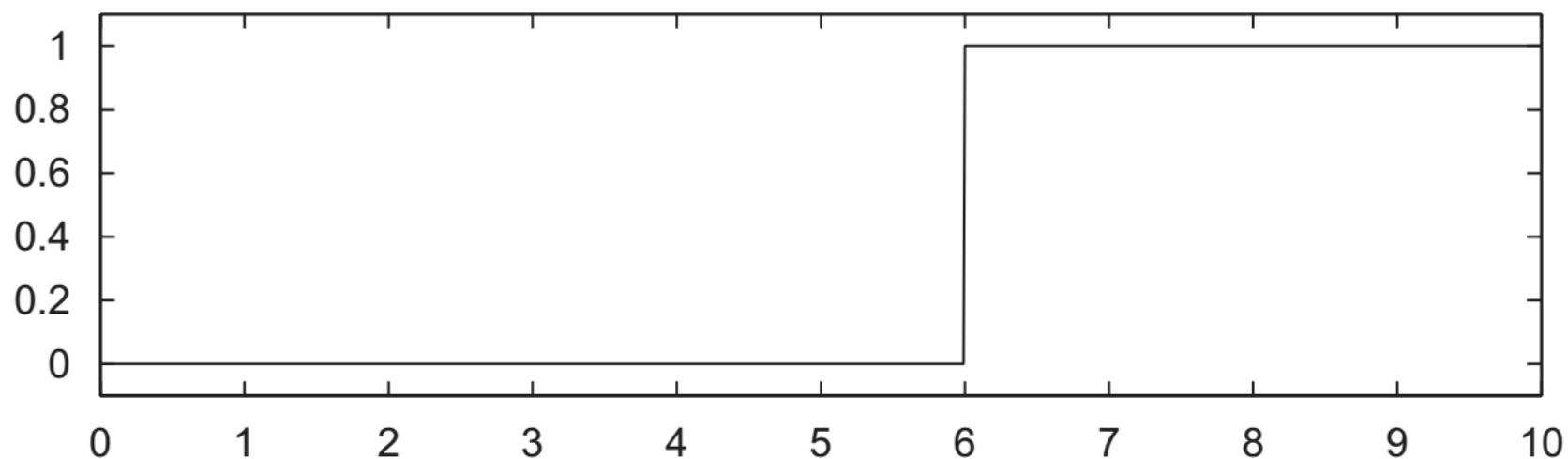
■ 局限：仅适用于断点处的个体，不能推广到其他个体

断点回归的数据要求

- 数据需包含以下三个基本变量
 - 配置(assignment)变量/驱动(driving)变量：连续特征变量
 - 断点：驱动变量临界值
 - 观测结果
- 配置变量的值在断点附近无法被准确操纵
 - 偶然性，局部随机性
- 断点的选择不受配置变量影响
- 除处置状态在断点处发生跳跃式变化外，其他未处置的个体特征变量在断点处无显著变化

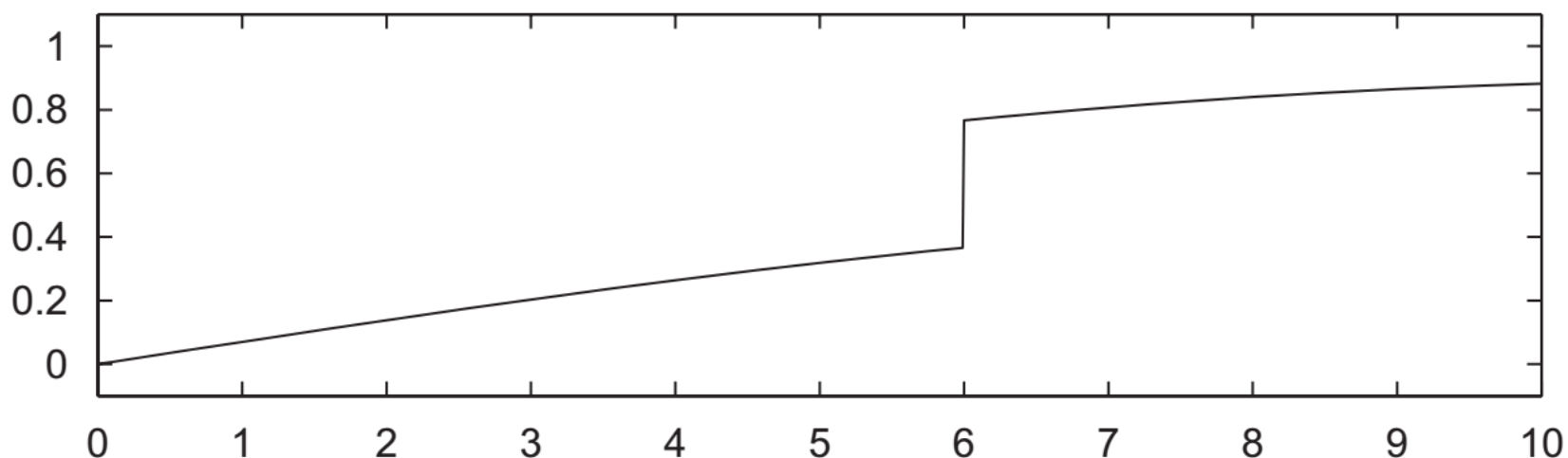
清晰断点

- 处理效应是驱动变量的一个确定函数：断点两侧接受处理效应的概率为0-1区分



模糊断点

- 处理效应是驱动变量的一个随机函数：断点两侧，处理效应的概率有跃升，但并非0-1分类



RDD的估计步骤和相应的Stata命令

■ 常用Stata RDD命令

▣ CCFT编写的非官方RDD命令组

▣ 获取方式: <http://sites.google.com/site/rdpackages>

常用的Stata RDD命令和功能

命令	功能
rdplot	绘制结果变量和配置变量拟合图
rdbwselect	拟合图分区数量选择: 可单独使用或包含在rdplot选项里
rdrobust	RDD局部多项式估计
rdwdensity	局部多项式估计带宽选择: 可单独使用或包含在rdrobust选项里
rddensity	检验配置变量密度函数是否连续

RDD的估计步骤和相应的Stata命令

■ RDD估计步骤

- 讨论配置变量和断点的产生过程，确定配置变量和断点选择是独立的
- 观察结果变量在断点处是否有明显跳跃
 - 散点图显示结果变量和配置变量的关系
 - 拟合图显示上述关系
 - 多项式回归拟合：选择多项式次数
 - 区间均值拟合：
 - ①选择区间分割方式：按配置变量/观测数量
 - ②选择区间数量：手动设置/IMSE最优区间数量
- 检验数据是否符合使用RDD的前提条件
 - 配置变量密度函数在断点处的连续性(rddensity)
 - 非结果特征变量在断点处的连续性(rdplot, rdrobust)
- 估计处置变量在断点处的跳跃程度和显著性-断点处处置效应的点估计
 - 全局多项式回归
 - 局部多项式回归

RDD运用实例

- 目的：考量美国众议院选举是否存在在职优势（一个地区现任政党对在该地区中获得选票的影响）
 - Lee (2008) “Randomized Experiments from Non-random Selection in U.S. House Elections,” *Journal of Econometrics*
 - 配置变量： *margin*
 - 断点： $margin = 0$ ；若 $margin > 0$ ，则选举胜出
 - 虚拟变量： $win = 1$ 若 $margin > 0$ ； $win = 0$ 若 $margin \leq 0$
 - 结果变量： *vote*

RDD运用实例

Name	Label
state	State ID
year	Election Year
vote	Democratic vote share in next election
margin	Democratic margin of victory
class	Senate class
termshouse	Cummulative number of terms served in U.S. House by congress of record
termssenate	Cummulative number of terms served in U.S. Senate by congress of record
win(自己创建)	是否赢得选举

RDD运用实例：描述性统计及其总结

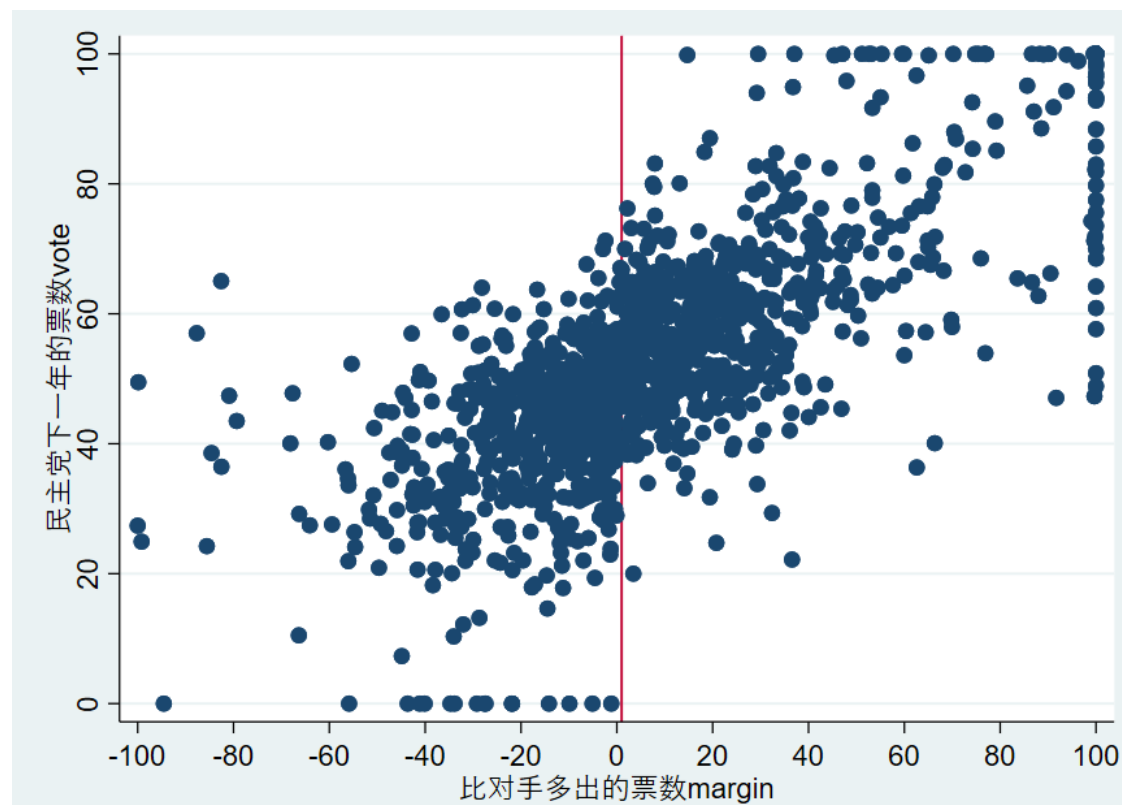
- 数据显示，平均民主党的票数为52.67；*win*的均值为0.5396，显示其赢得了53.96%的选举；样本包含了1914-2010年的选举数据

```
. sum vote win year
```

Variable	Obs	Mean	Std. Dev.	Min	Max
vote	1,297	52.66627	18.12219	0	100
win	1,390	.5395683	.4986113	0	1
year	1,390	1964.63	28.05466	1914	2010

RDD运用实例：结果变量与配置变量关系-直观

- 讨论配置变量和断点的产生过程
 - 断点0由法律规定，不会受得票率的影响
- 用散点图显示结果变量和配置变量的关系



RDD运用实例：结果变量与配置变量关系-直观

```
. ttest vote, by(win)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	595	40.92053	.4989536	12.17079	39.9406	41.90045
1	702	62.6217	.6147484	16.28793	61.41474	63.82867
combined	1,297	52.66627	.5032002	18.12219	51.67909	53.65344
diff		-21.70118	.810499		-23.29121	-20.11114

diff = mean(0) - mean(1) t = -26.7751
Ho: diff = 0 degrees of freedom = 1295

Ha: diff < 0
Pr(T < t) = 0.0000

Ha: diff != 0
Pr(|T| > |t|) = 0.0000

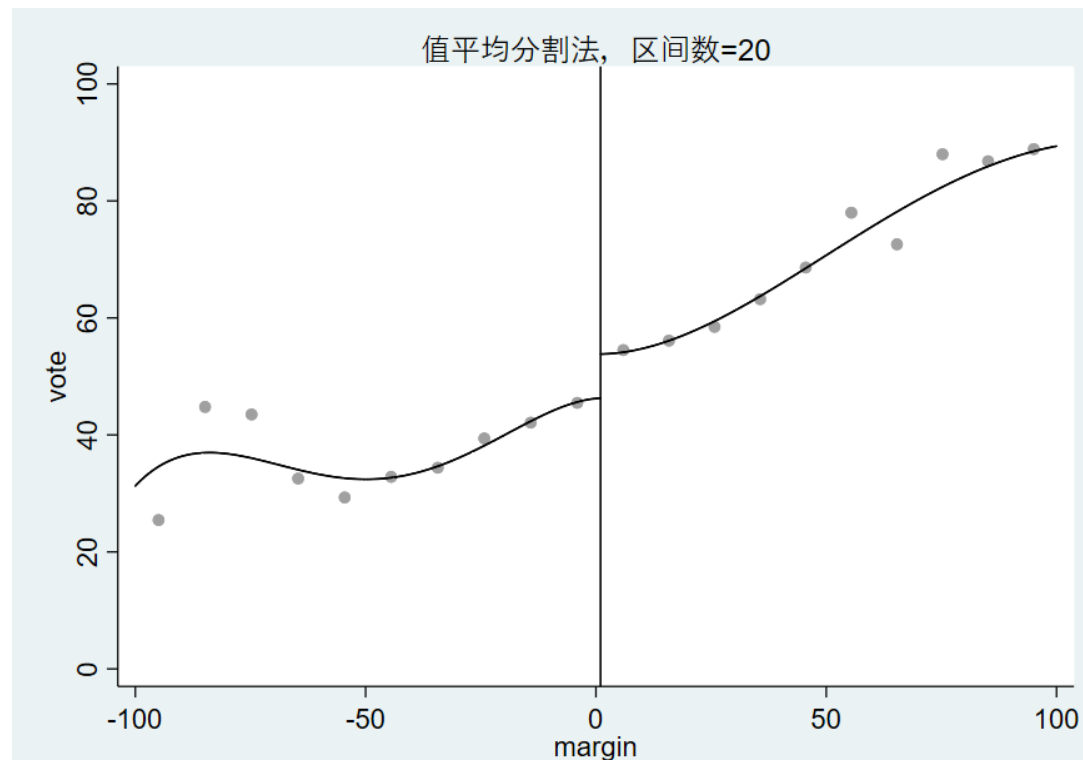
Ha: diff > 0
Pr(T > t) = 1.0000

RDD运用实例：结果变量与配置变量关系-直观

- 用拟合图显示结果变量和配置变量的关系
 - 多项式回归拟合：绘图过程中不必太严谨，多取2-4次
 - 分区均值拟合
 - 区间分割方法 $\left\{ \begin{array}{l} \text{值平均分割}(ES) \\ \text{数量平均分割}(QS) \end{array} \right.$
 - 区间分割数量：IMSE最优区间数量-原理：如果区间增多，区间平均值估计量偏差（bias）降低，但由于区间内观测值较少，使得估计方差(variance)增加
 - IMSE最优区间数量指权衡这二者的因素，通过最小化估计量的IMSE得到的区间数量
 - 通过图形考察结果变量和配置变量的关系是RDD的关键第一步；如果平滑拟合后的图形没有明确显示结果变量在断点处有跳跃，即使统计方法发现了显著的处置效果，稳健性也是可疑的

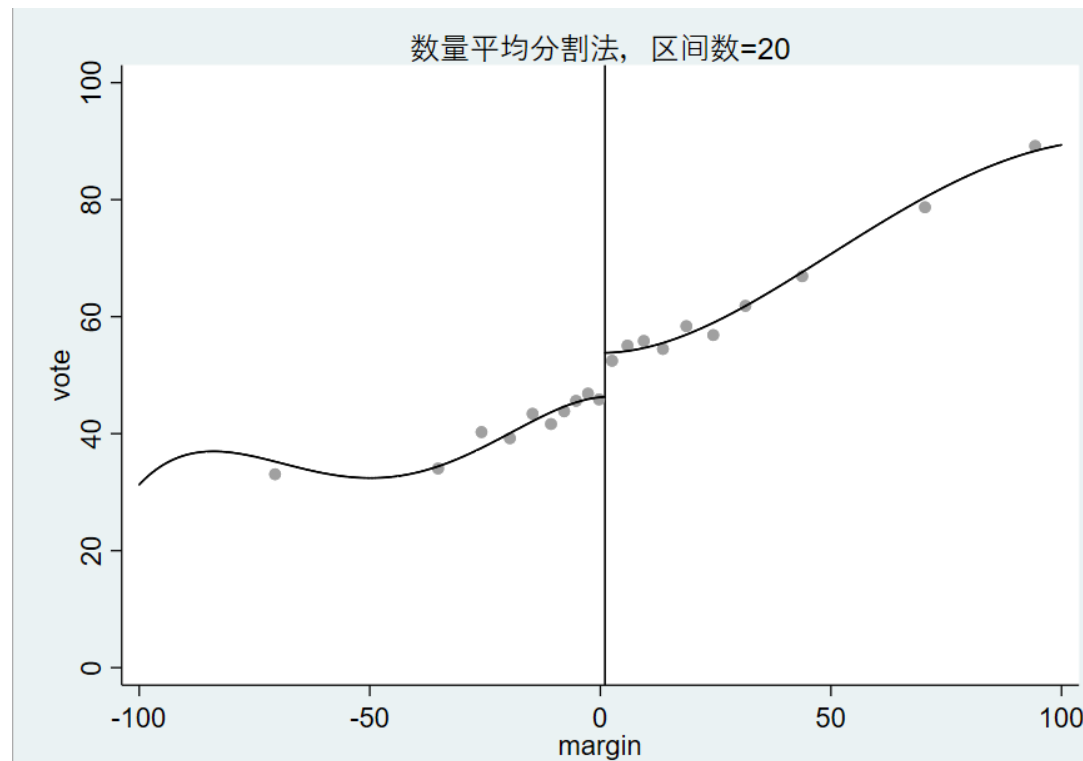
RDD运用实例：结果变量与配置变量关系-直观

- `rdplot vote margin, nbins(10 10) binselect(es) c(1) p(4)`
`graph_options(legend(off) xtitle("margin") ytitle("vote"))`
`ylabel(0(20)100,nogrid) subtitle("值平均分割法，区间数=20")`
`legend(off))`



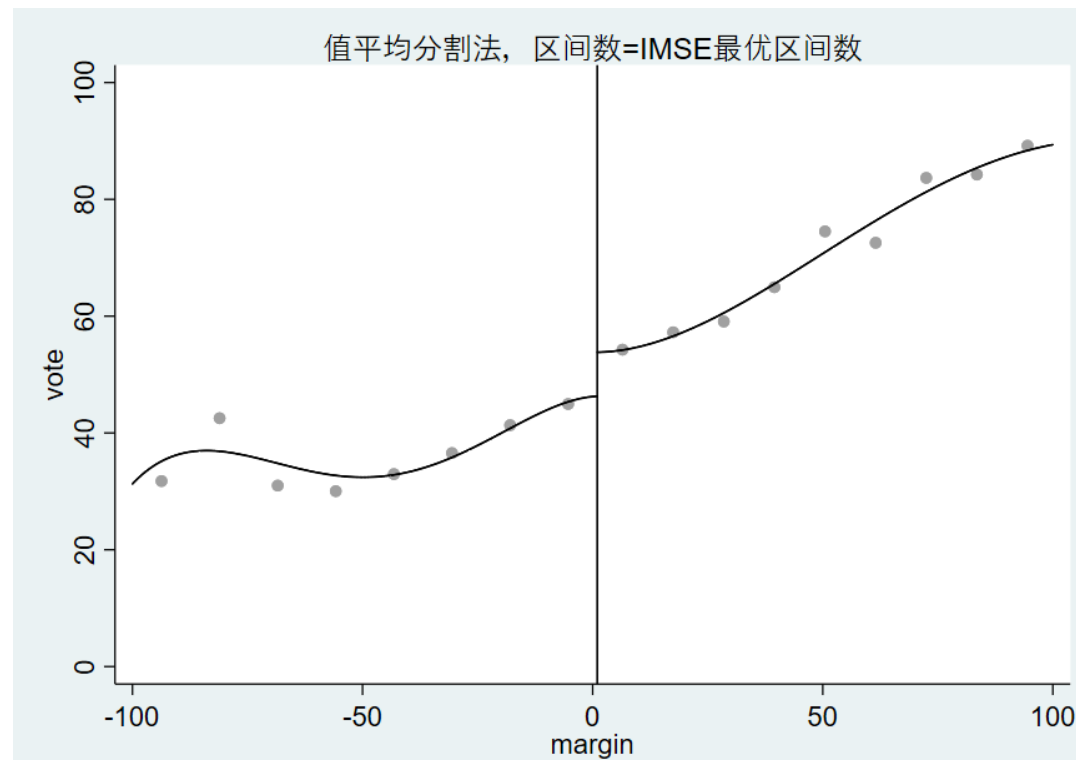
RDD运用实例：结果变量与配置变量关系-直观

- `rdplot vote margin, nbins(10 10) binselect(qs) c(1) p(4)`
`graph_options(legend(off) xtitle("margin") ytitle("vote"))`
`ylabel(0(20)100,nogrid) subtitle("数量平均分割法，区间数=20")`
`legend(off))`



RDD运用实例：结果变量与配置变量关系-直观

- `rdplot vote margin, [去掉nbins参数] binselect(es) c(1) p(4)`
`graph_options(legend(off) xtitle("margin") ytitle("vote"))`
`ylabel(0(20)100,nogrid) subtitle("值平均分割法，区间数=IMSE最优区间数") legend(off))`



RDD运用实例：验证RDD的有效性

■ 检验配置变量的概率分布连续性

□ 方法1

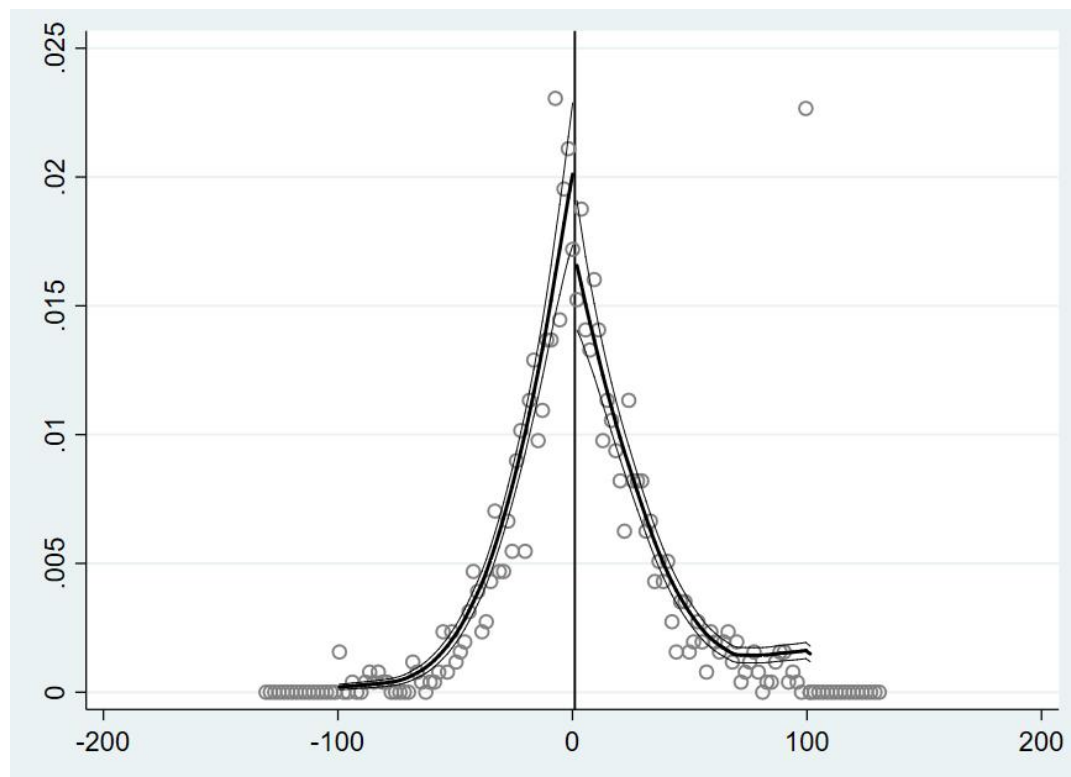
方法	命令	原理	缺陷
McCrary	DCdensity	在断点两边各选定一定的区间 将概率值对配置变量运行两个线性回归 分别得到回归直线在断点左边和右边的截距值 检验log(截距差)的显著性	需要选择局部线性回归的区间

□ 方法2：Cattaneo, Jansson and Ma (2018, Stata J.), rddensity 统计性质更优越

RDD运用实例：验证RDD的有效性

```
. DCdensity margin, breakpoint(1) generate(Xj Yj r0 fhat se fhat)  
Using default bin size calculation, bin size = 1.84133021  
Using default bandwidth calculation, bandwidth = 30.3968964
```

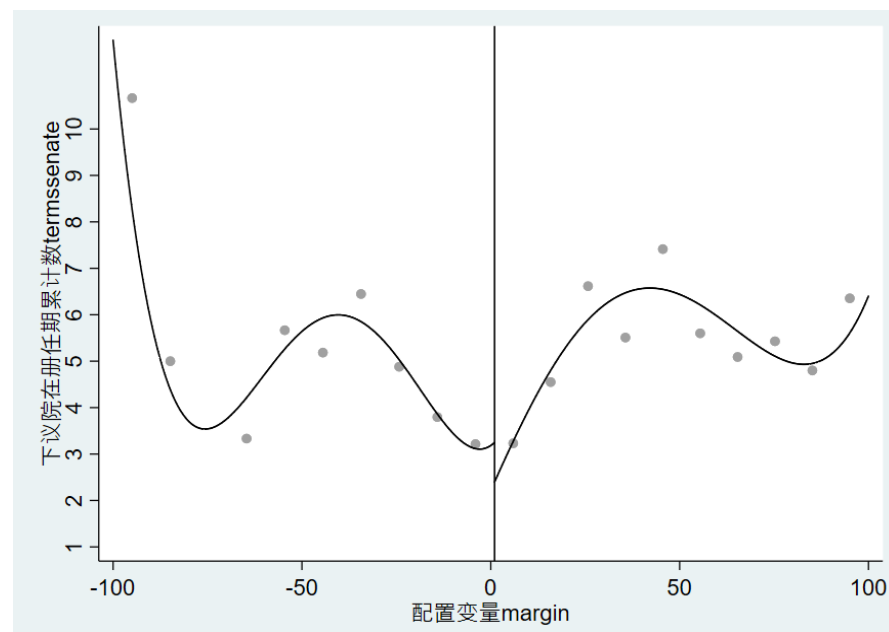
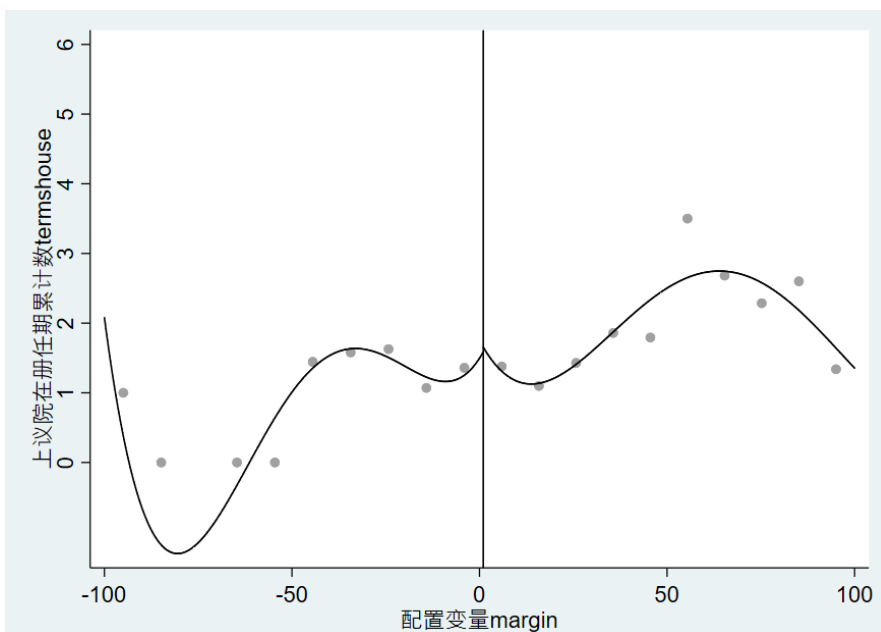
```
Discontinuity estimate (log difference in height): -.194383955  
(.110478306)
```



RDD运用实例：验证RDD的有效性

■ 检验个体特征变量连续性

- 使用rdplot绘制显示特征变量和配置变量关系的平滑拟合图，并观测在断点处是否连续（注：这些变量必须是在事件发生前的值）



RDD运用实例：断点处置效应估计

■ 全局多项式回归

□ 将断点左边与右边数据分别回归

$$Y_i = \alpha_{l0} + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \dots + \alpha_{lp_{left}}(x_i - C)^{p_{left}} + u_i$$

$$Y_i = \alpha_{r0} + \alpha_{r1}(x_i - C) + \alpha_{r2}(x_i - C)^2 + \dots + \alpha_{rp_{right}}(x_i - C)^{p_{right}} + v_i$$

□ 截距项为多项式在断点的截距

□ 断点处的处置效应=处置边多项式截距-未处置边多项式截距

RDD运用实例：断点处置效应估计

- 将上述两个方程合并为一：引入虚拟变量 $T = win$ ，反映处理效应

$$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \dots + \alpha_{lk}(x_i - C)^k \\ + (\alpha_{r1} - \alpha_{l1})T(x_i - C) + (\alpha_{r2} - \alpha_{l2})T(x_i - C)^2 + \dots \\ + (\alpha_{rk} - \alpha_{lk})T(x_i - C)^k + e_i$$

- T 为 treatment 虚拟变量
- 方程次数的选择：AIC/尝试报告一系列的次数（1-6次）来观察结果是否对次数选择很敏感
- 优点：数据多，方差小；缺点：结果对多项式方程设置敏感

RDD运用实例：断点处置效应估计

```
. gen X1 = margin-0  
  
. gen X2 = X1^2  
  
. gen X3 = X1^3  
  
. gen X4 = X1^4  
  
. gen win_X1 = X1*win  
  
. gen win_X2 = X2*win  
  
. gen win_X3 = X3*win  
  
. gen win_X4 = X4*win
```

RDD运用实例：断点处置效应估计

```
. reg vote win X1 X2 X3 X4 win_X1 win_X2 win_X3 win_X4, cluster(class)
```

```
Linear regression                               Number of obs   =       1,297
                                                F(1, 2)         =           .
                                                Prob > F         =           .
                                                R-squared        =       0.5958
                                                Root MSE        =      11.562
```

(Std. Err. adjusted for 3 clusters in class)

vote	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
win	9.407075	1.556146	6.05	0.026	2.71152	16.10263
X1	-.31181	.1971207	-1.58	0.255	-1.159952	.536332
X2	-.0371986	.0072708	-5.12	0.036	-.0684824	-.0059147
X3	-.0007185	.0001245	-5.77	0.029	-.0012541	-.0001829
X4	-3.92e-06	7.32e-07	-5.35	0.033	-7.07e-06	-7.69e-07
win_X1	.3682319	.114532	3.22	0.085	-.1245594	.8610231
win_X2	.0456877	.0139022	3.29	0.081	-.0141285	.105504
win_X3	.0006644	.000093	7.15	0.019	.0002644	.0010644
win_X4	3.91e-06	1.37e-06	2.87	0.103	-1.96e-06	9.79e-06
_cons	43.93729	1.71457	25.63	0.002	36.5601	51.31449

RDD运用实例：断点处置效应估计

■ 局部多项式回归

□ 原理：将回归区域限制在断点附近的区域，通常使用低次多项式（例如，1次/2次多项式）

□ 模型次数选择：局部线性回归

$$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + (\alpha_{r1} - \alpha_{l1})T(x_i - C) + \varepsilon_i$$

□ 核函数权重选择：建议三角核函数(triangular kernel function)，给越接近断点处的值赋予更高权重

➤ 与最优带宽一起使用时可以得到均值平方误差(MSE)最优的系数估计量 $\hat{\tau}$

□ 带宽（“局部”的范围）选择

➤ 带宽越大，受多项式方程设置和数据异常值影响越大，偏差越大；带宽越小，可用数据越少，估计方差越大

➤ 常使用MSE最优带宽，原理：

➤ $MSE(\tau) = Bias^2(\tau) + Variance(\tau)$

RDD运用实例：断点处置效应估计

```
. rdrobust vote margin, c(1) kernel(triangular) p(1) bwselect(mserd) vce(cluster class)
```

Sharp RD estimates using local polynomial regression.

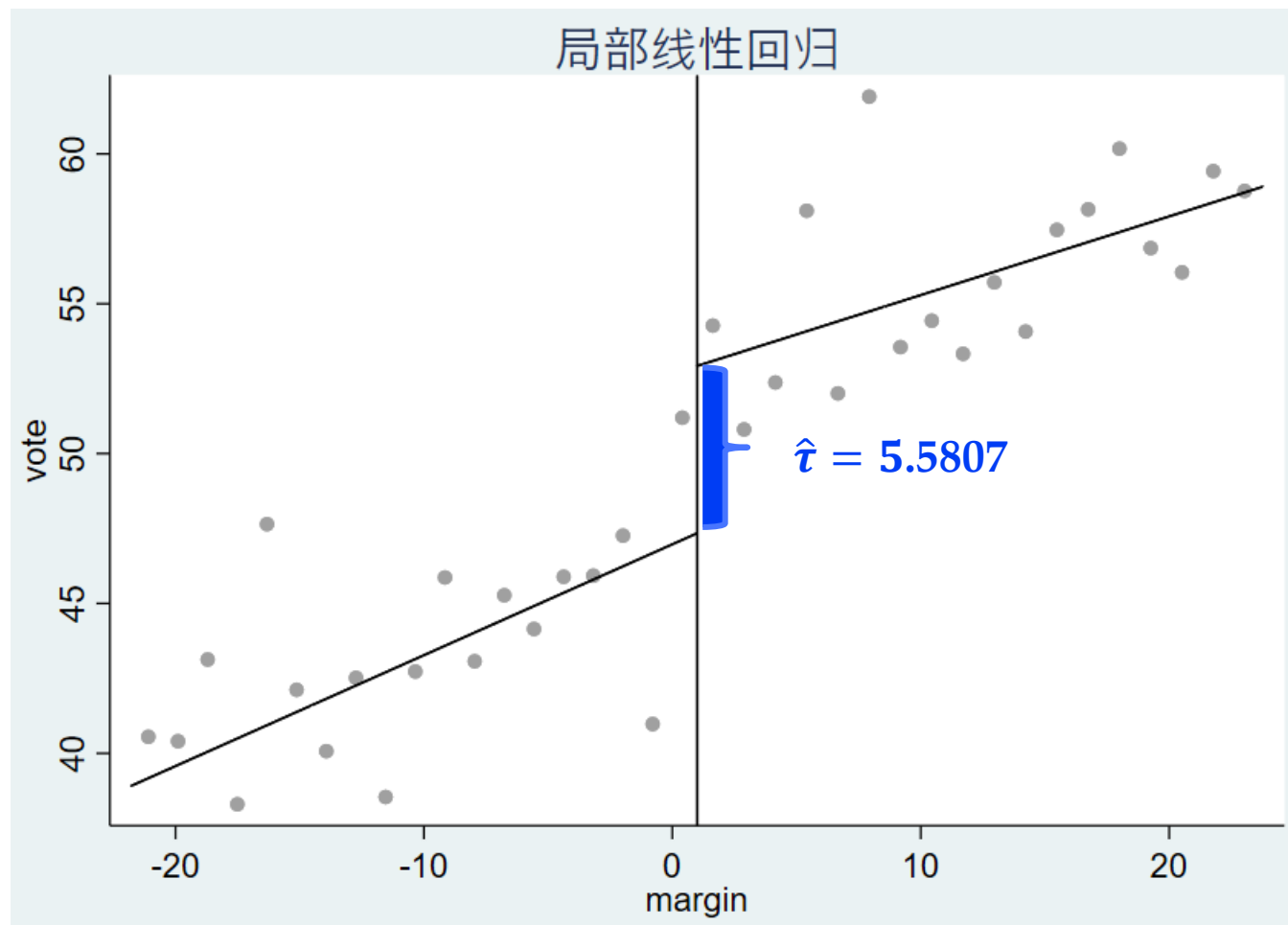
Cutoff c = 1	Left of c	Right of c		
Number of obs	620	677	Number of obs =	1297
Eff. Number of obs	426	363	BW type =	mserd
Order est. (p)	1	1	Kernel =	Triangular
Order bias (q)	2	2	VCE method =	Cluster
BW est. (h)	22.778	22.778		
BW bias (b)	32.080	32.080		
rho (h/b)	0.710	0.710		
Number of clusters	3	3		

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Conventional	5.5807	1.7028	3.2774	0.001	2.24328	8.91805
Robust	-	-	2.6198	0.009	1.25916	8.73903

Std. Err. adjusted for clusters in class

RDD运用实例：断点处置效应估计



RDD运用实例：断点处置效应估计

■ 局部多项式回归

□ 点估计的置信区间

➤ 全局回归

$$95\% \text{ Conf. Interval} = (\hat{\tau} - 1.96\sqrt{\text{Variance}}, \hat{\tau} + 1.96\sqrt{\text{Variance}})$$

➤ 局部线性回归：偏差调整稳健置信区间

95% Conf. Interval

$$= [(\hat{\tau} - \widehat{Bias}) - 1.96\sqrt{\text{Variance}_{bc}}, (\hat{\tau} - \widehat{Bias}) + 1.96\sqrt{\text{Variance}_{bc}}]$$

➤ Bias: 与带宽选择相关的估计值偏差

➤ Variance_{bc} : robust bias-corrected 方差

RDD运用实例：断点处置效应估计

```
. rdrobust vote margin, c(1) kernel(triangular) p(1) bwselect(mserd) vce(cluster class)
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 1	Left of c	Right of c	Number of obs =	1297
			BW type =	mserd
Number of obs	620	677	Kernel =	Triangular
Eff. Number of obs	426	363	VCE method =	Cluster
Order est. (p)	1	1		
Order bias (q)	2	2		
BW est. (h)	22.778	22.778		
BW bias (b)	32.080	32.080		
rho (h/b)	0.710	0.710		
Number of clusters	3	3		

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Conventional	5.5807	1.7028	3.2774	0.001	2.24328	8.91805
Robust	-	-	2.6198	0.009	1.25916	8.73903

Std. Err. adjusted for clusters in class

RDD运用实例：断点处置效应估计

■ 结果稳健性检验

- 考虑在多项式中加入其他特征变量 Z 作为自变量

$$Y_i = \alpha_{l0} + \tau T + \alpha_{l1}(x_i - C) + \alpha_{l2}(x_i - C)^2 + \cdots + \alpha_{lk}(x_i - C)^k \\ + (\alpha_{r1} - \alpha_{l1})T(x_i - C) + (\alpha_{r2} - \alpha_{l2})T(x_i - C)^2 + \cdots + (\alpha_{rk} - \alpha_{lk})T(x_i - C)^k \\ + \alpha_{r,k+1}T(x_i - C)^{k+1} + \cdots + \alpha_{r,k+w}T(x_i - C)^{k+w} + \mathbf{Z}'\boldsymbol{\beta} + e_i$$

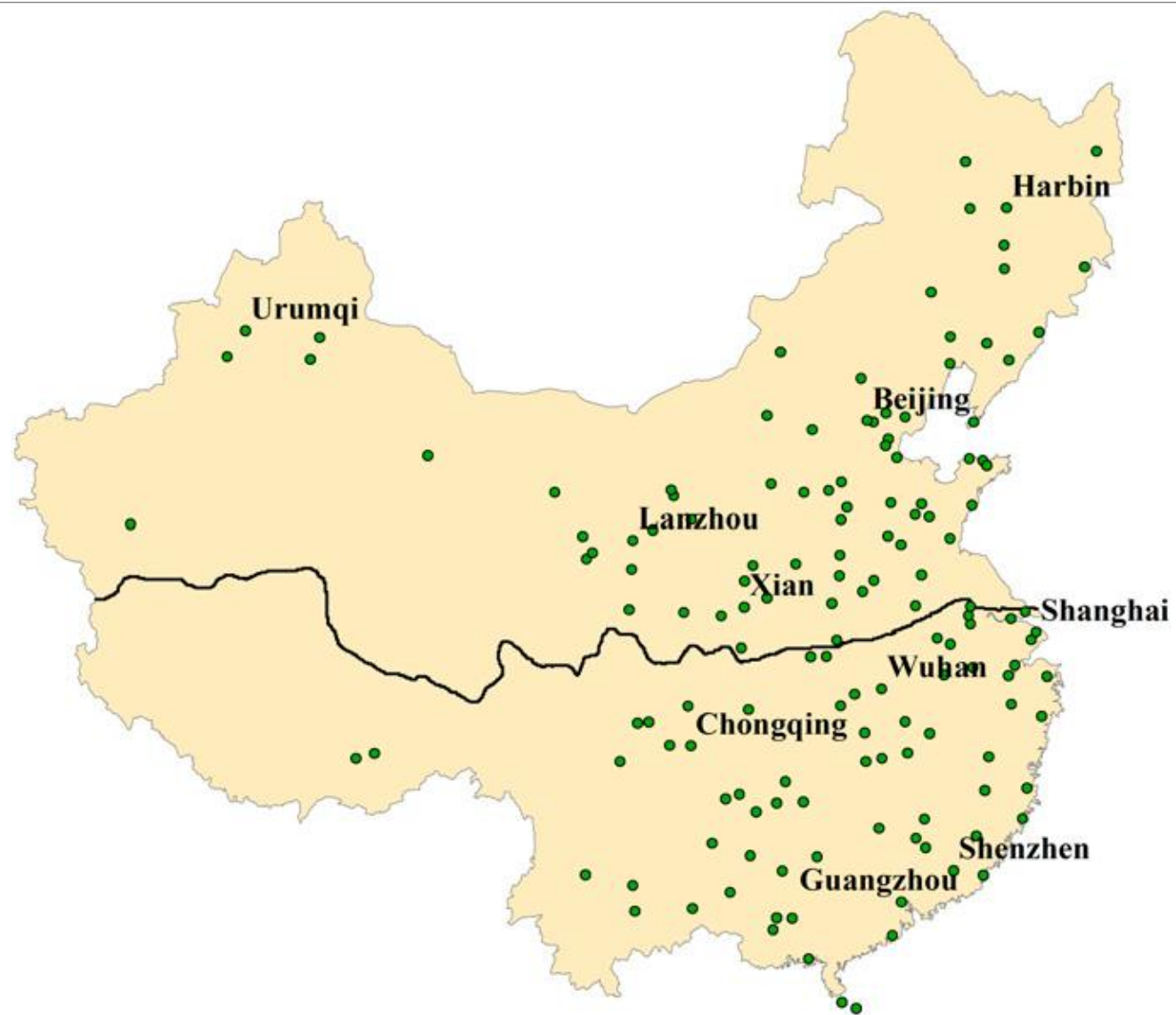
- 加入控制变量后，标准误应该降低；若其增大，则说明其他特征变量也影响了结果变量在断点处的跳跃
- 安慰剂检验，即检验一个不应该受处置事件影响的结果变量在断点处是否也有跳跃

实例1：淮河取暖分界线的健康效应

文献介绍

- Chen, Ebenstein, Greenstone, and Li (2013), “Evidence on the impact of sustained exposure to air pollution on life expectancy from China’s Huai River policy,” *PNAS*
- 以秦岭-淮河为分界，识别空气污染对居民预期寿命影响

地理分界

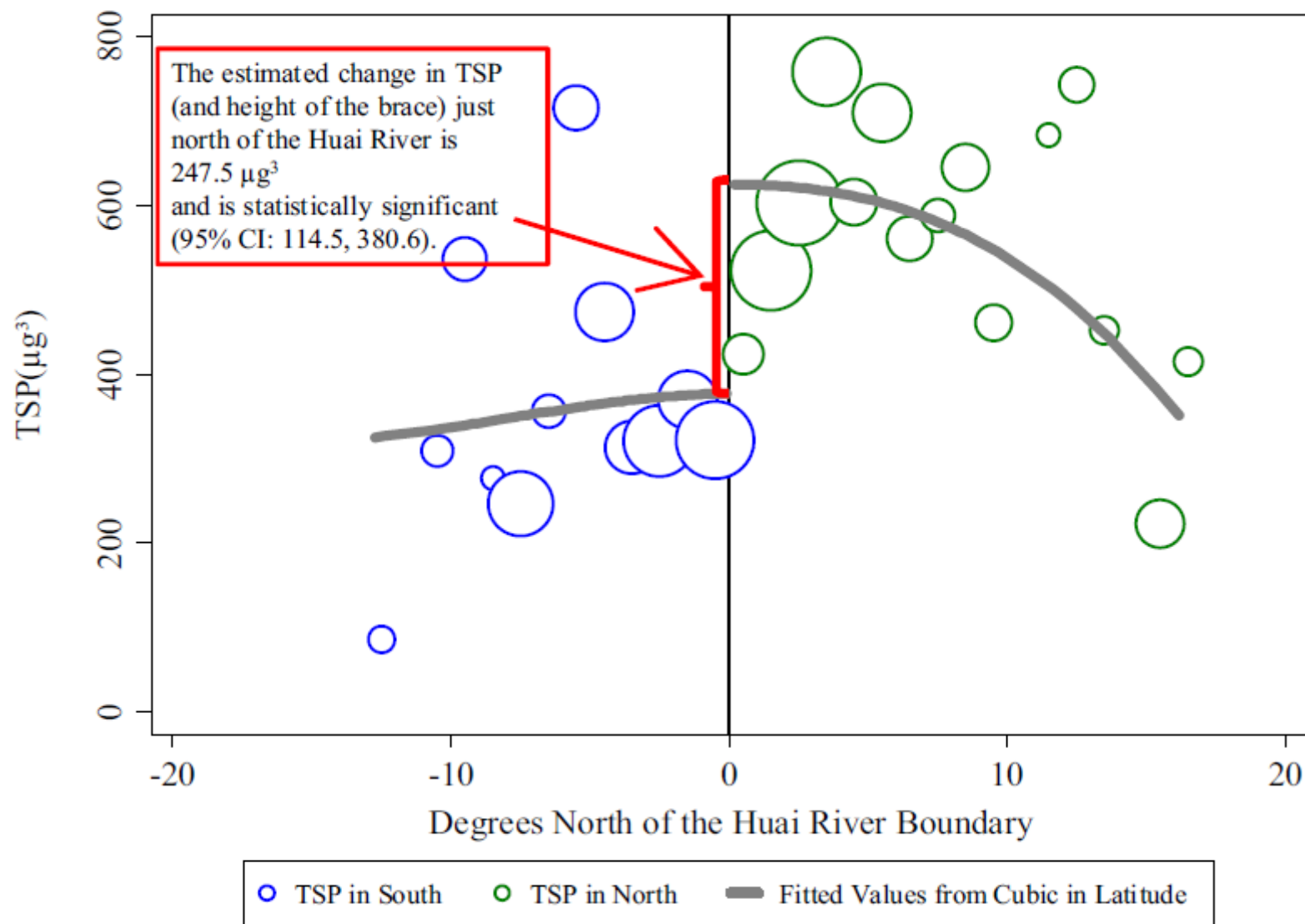


描述性统计

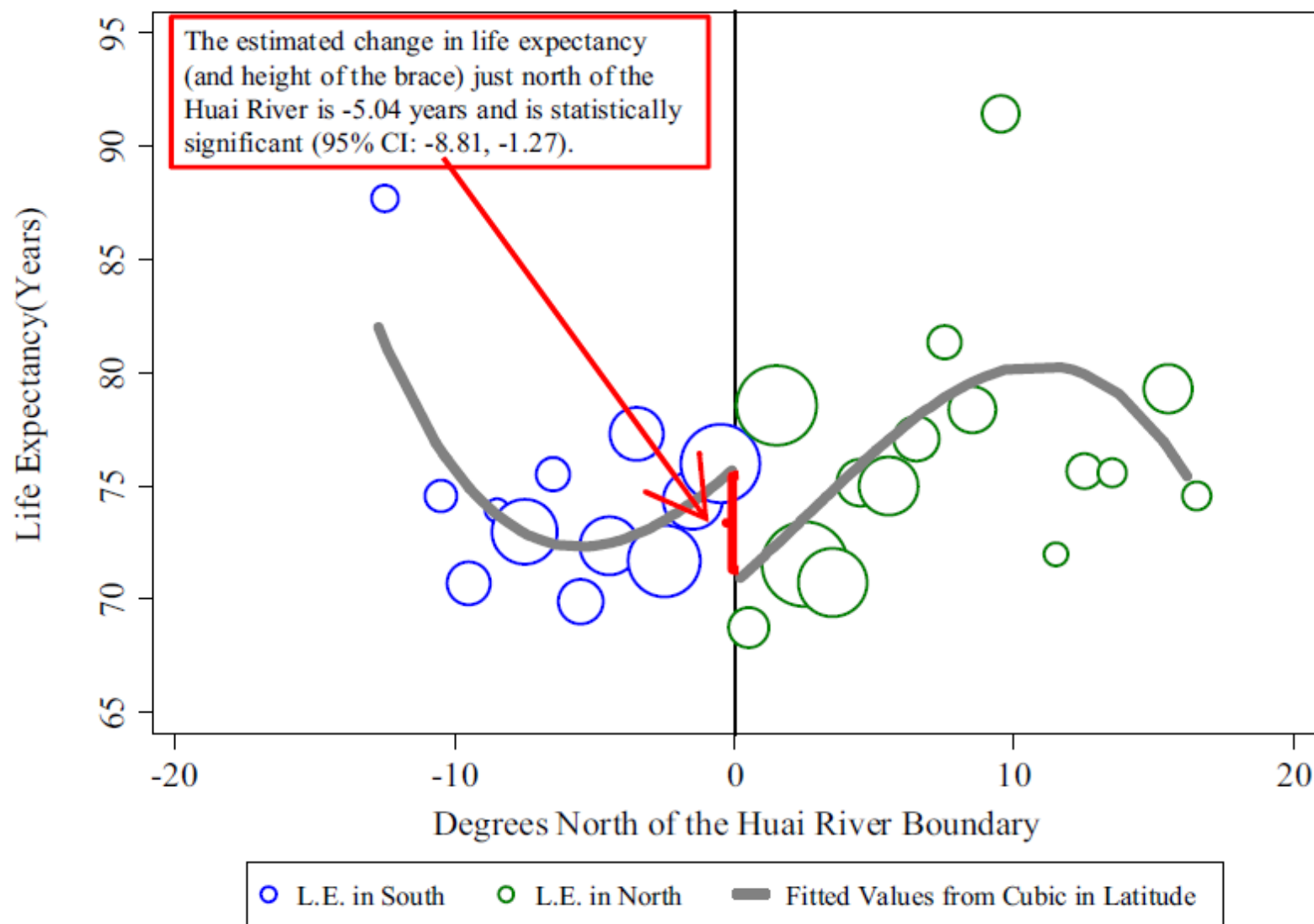
Table 1. Summary statistics

Variable	South (1)	North (2)	Difference in means (3)	Adjusted difference in means (4)	P value (5)
Panel 1: Air pollution exposure at China's Disease Surveillance Points					
TSPs, $\mu\text{g}/\text{m}^3$	354.7	551.6	196.8***	199.5***	<0.001/0.002
SO ₂ , $\mu\text{g}/\text{m}^3$	91.2	94.5	3.4	-3.1	0.812/0.903
NO _x , $\mu\text{g}/\text{m}^3$	37.9	50.2	12.3***	-4.3	<0.001/0.468
Panel 2: Climate at the Disease Surveillance Points					
Heating degree days	2,876	6,220	3,344***	482	<0.001/0.262
Cooling degree days	2,050	1,141	-910***	-183	<0.001/0.371
Panel 3: Demographic features of China's Disease Surveillance Points					
Years of education	7.23	7.57	0.34	-0.65	0.187/0.171
Share in manufacturing	0.14	0.11	-0.03	-0.15***	0.202/0.002
Share minority	0.11	0.05	-0.05	0.04	0.132/0.443
Share urban	0.42	0.42	0.00	-0.20*	0.999/0.088
Share tap water	0.50	0.51	0.02	-0.32**	0.821/0.035
Rural, poor	0.21	0.23	0.01	-0.33*	0.879/0.09
Rural, average income	0.34	0.33	0.00	0.24	0.979/0.308
Rural, high income	0.21	0.19	-0.02	0.27	0.772/0.141
Urban site	0.24	0.25	0.01	-0.19	0.859/0.241
Predicted life expectancy	74.0	75.5	1.54***	-0.24	<0.001/0.811
Actual life expectancy	74.0	75.5	1.55	-5.04**	0.158/0.044

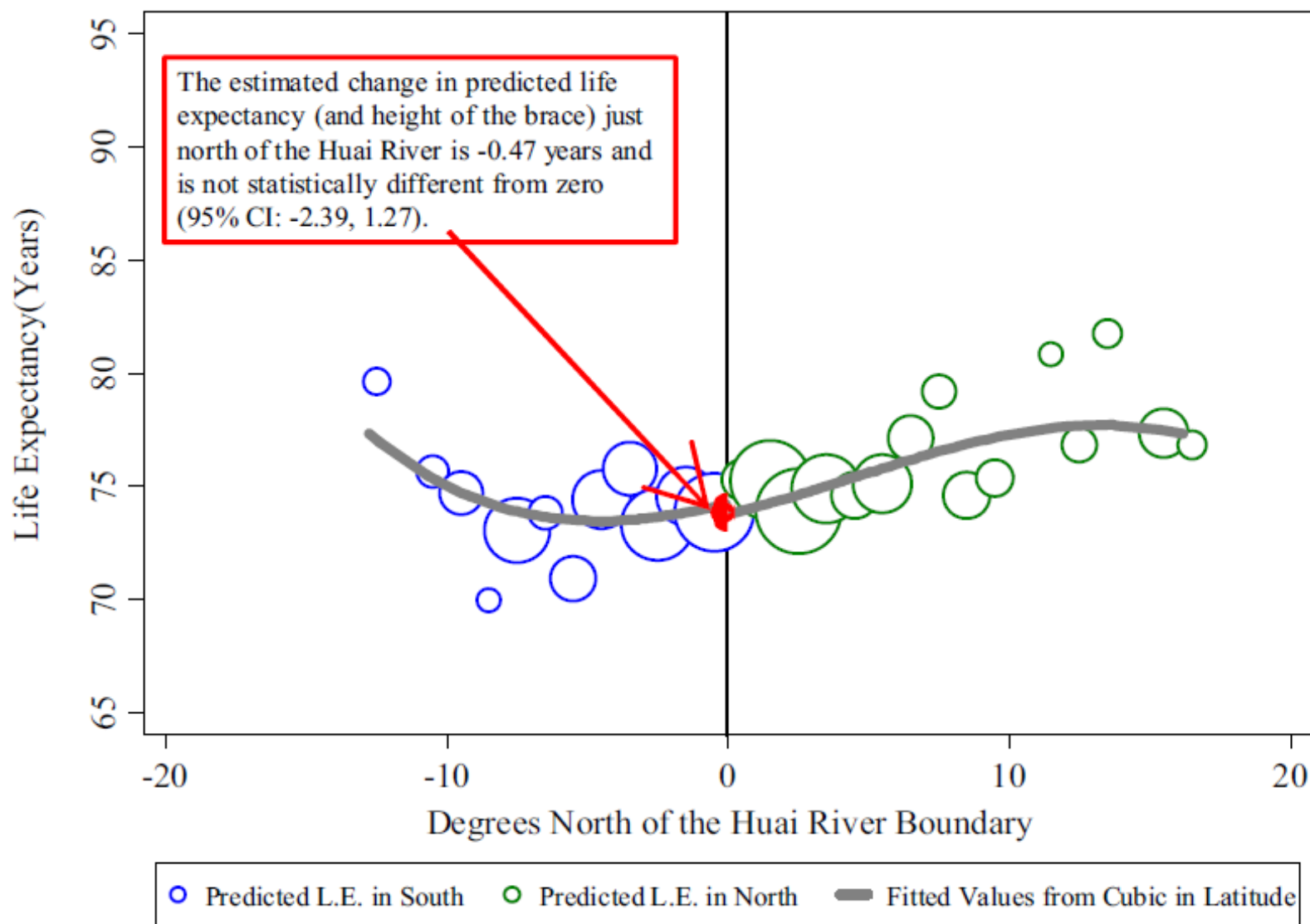
污染排放差别



预期寿命差别



安慰剂检验：去除污染物后的回归结果，没有差异



回归方程设定

■ 模型1：RD回归

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

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

- 其中 TSP (total suspended particulates) 为总悬浮颗粒物，污染指标
- N_j 为北方虚拟变量， L_j 为城市距离淮河的纬度， X_j 为城市特征控制变量
- $f(\cdot)$ 为一个3阶多项式

■ 模型2：2SLS

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

$$Y_j = \beta_0 + \beta_1 TSP_j + \beta_2 f(L_j) + X_j \Gamma + \varepsilon_j$$

回归结果

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

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

RD常用设定之一：地理分界

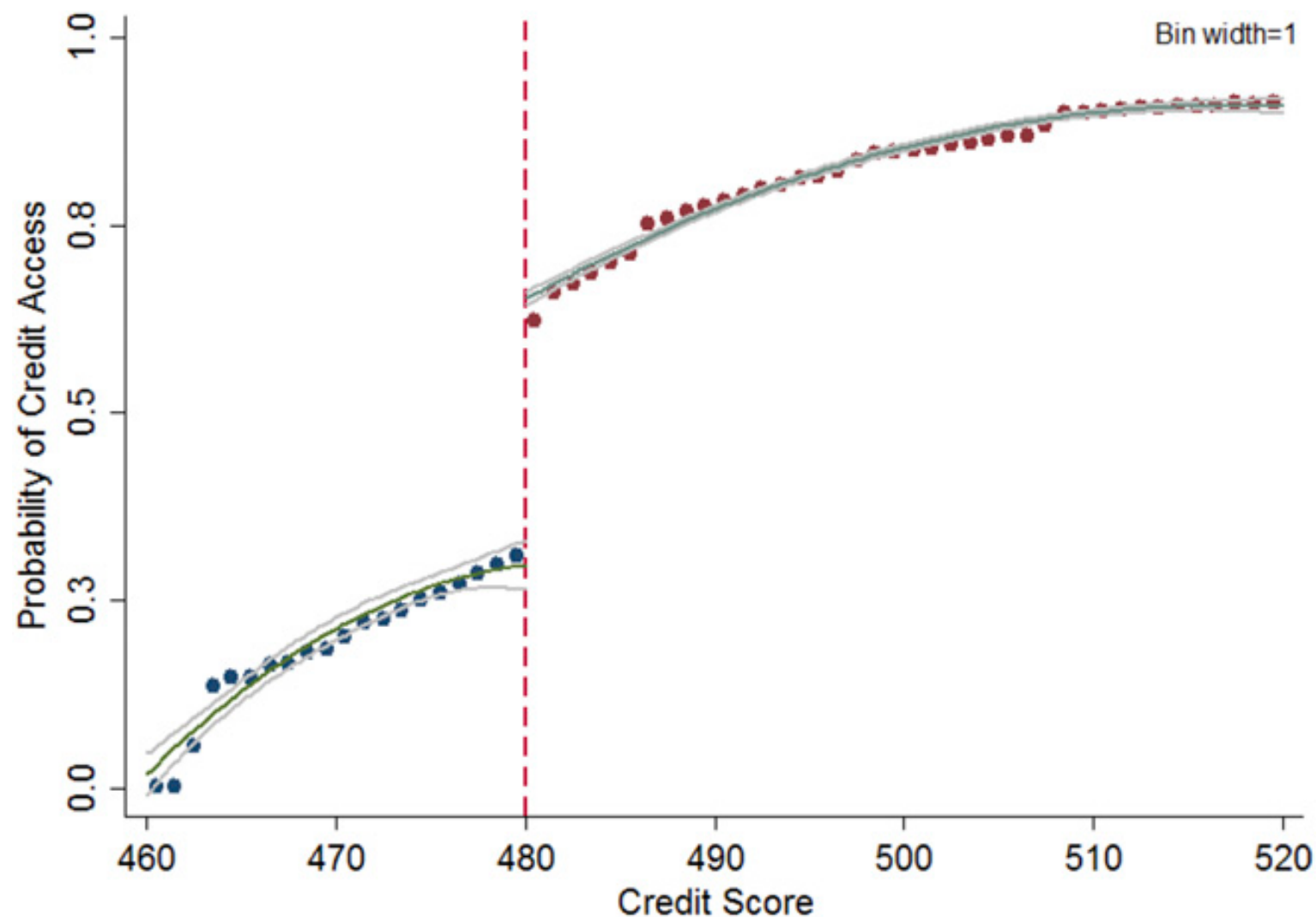
- 地理分界线如果有一定的随机性，则两侧结果变量的差异，就更容易归结为地理分界线的随机影响；如果分界线同时对某一机制（解释）变量带来随机差异，则结果变量的差别更容易归因到机制变量的差别上
- 在反事实因果推断框架下，所有因果识别本质上都在利用样本观测变量（解释变量）在某些取值范围局部的随机性
 - IV变量随机取值（至少外生于可观测解释变量），DID里要求处理组随机分配，RD两侧驱动变量随机取值
- 不过现实中，地理界限两侧差异未必随机
 - 重要的是了解地理界限，或者任何“外生于解释变量”的变动的成因
 - IV变量的驱动因素，地理分界的渊源，处理组的成因

信用评级、贷款决策与经济后果

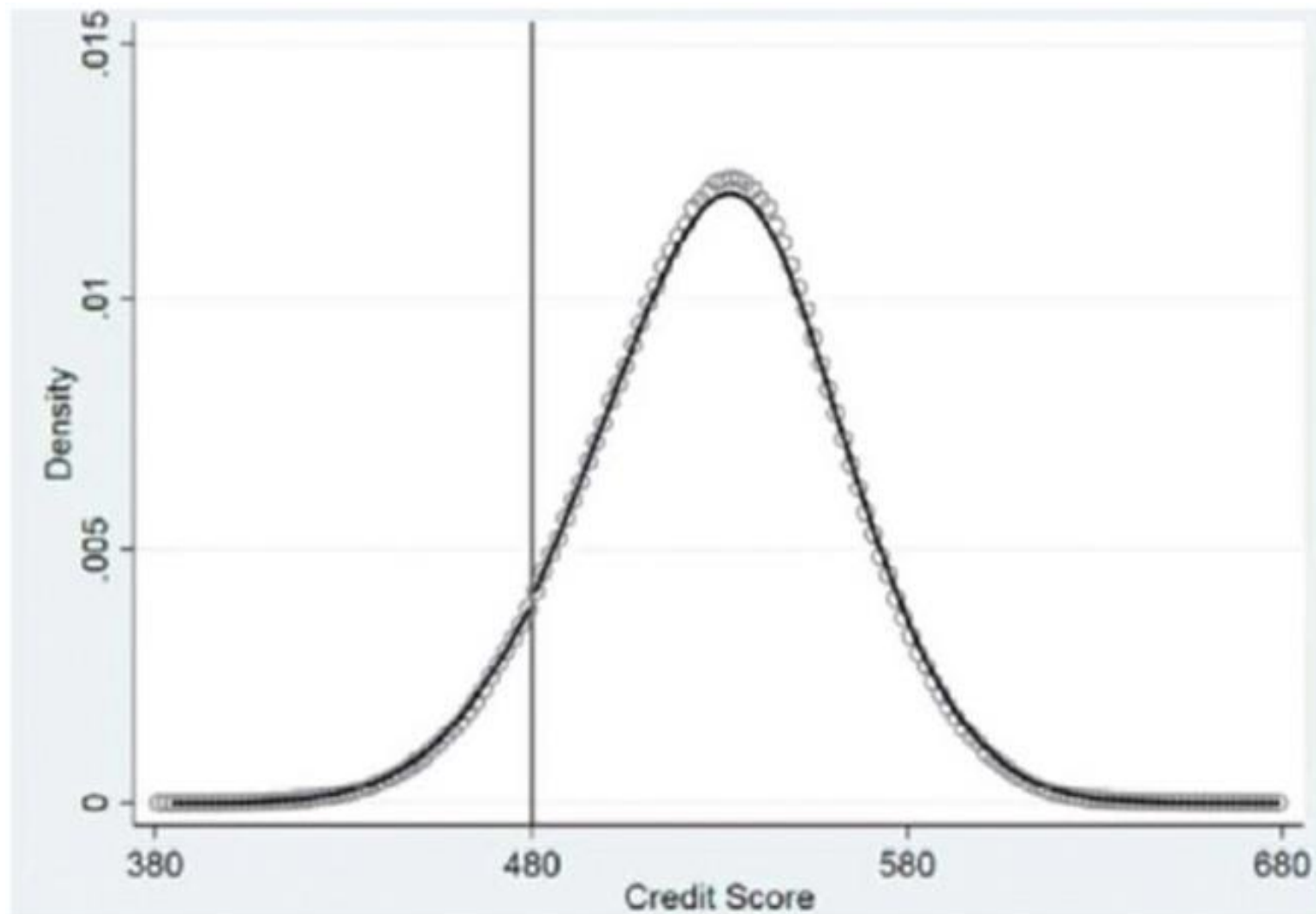
文献介绍

- Tao Chen, Yi Huang, Chen Lin, Zixia Sheng (2022), “Finance and Firm Volatility: Evidence from Small Business Lending in China,” *Management Science*, 68(3):2226-2249
- 数据：蚂蚁金服金融数据及淘宝商户业务数据
 - 蚂蚁金服为符合信用评分条件的淘宝商户提供商业授信(credit line)，这样商户需要融资时，就可以从授信中按照一定利率、期限获得贷款
 - 样本期：2014年11月至2015年6月，1,898,180商户，8,848,251商户-月度样本，非平衡面板
 - 截止2016年8月，蚂蚁金服在5年时间内累计提供的小微贷款总额为7000亿，贷款户数超过400万；平均贷款规模为2万，平均不良率低于3%，期限主要为半年至一年
 - 样本期蚂蚁金服的信贷评分标准线为480，评分高于此线，即可自动获得授信服务
 - 自然提供了一个断点设计场景

信用评分断点480附近的信贷可得性：模糊断点



断点附近驱动变量信用评分的连续性



主要的实证设定：RDD-IV

- 1阶段：断点附近贷款获得概率的“外生”变化

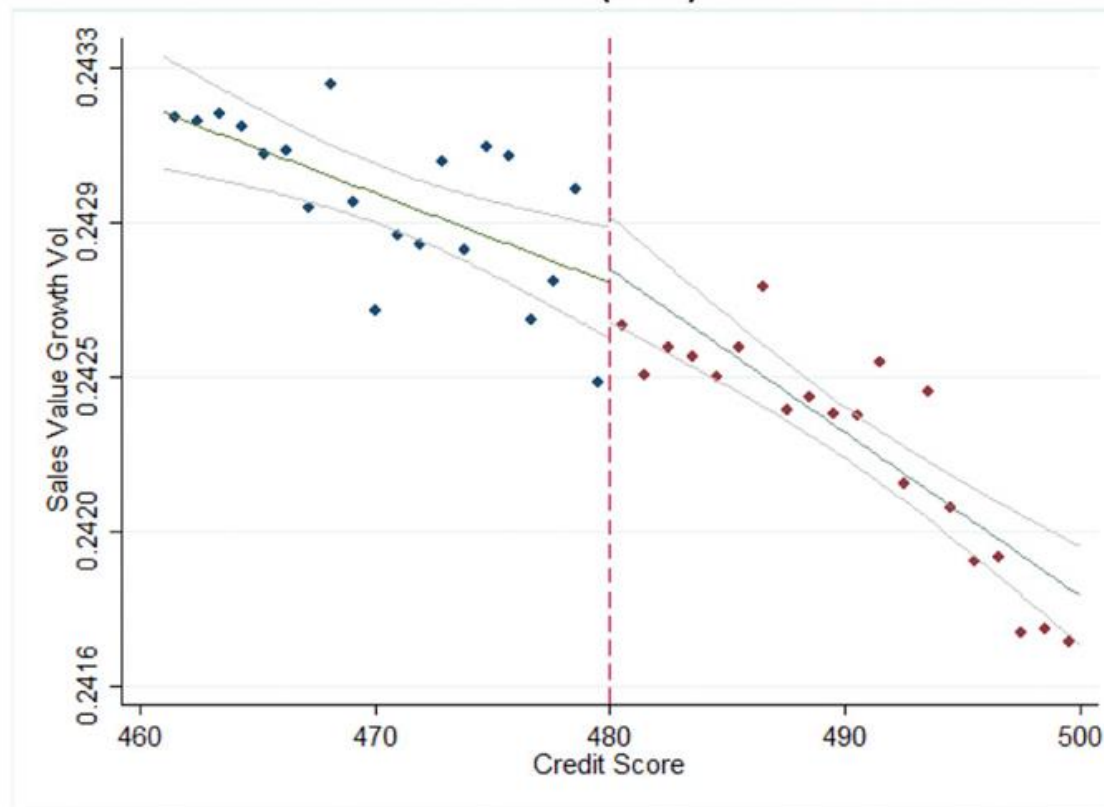
$$D_{it} = \alpha + \pi T_{it} + \sum_{k=1}^K \rho_k (s_{it} - s^*)^k + T_{it} \sum_{k=1}^K \sigma_k (s_{it} - s^*)^k + \varphi_j + \theta_t + \mu_{it}$$

- 2阶段：断点附近，贷款获得概率的跃升，如何影响商户销售增速的波动率（考虑滞后期）

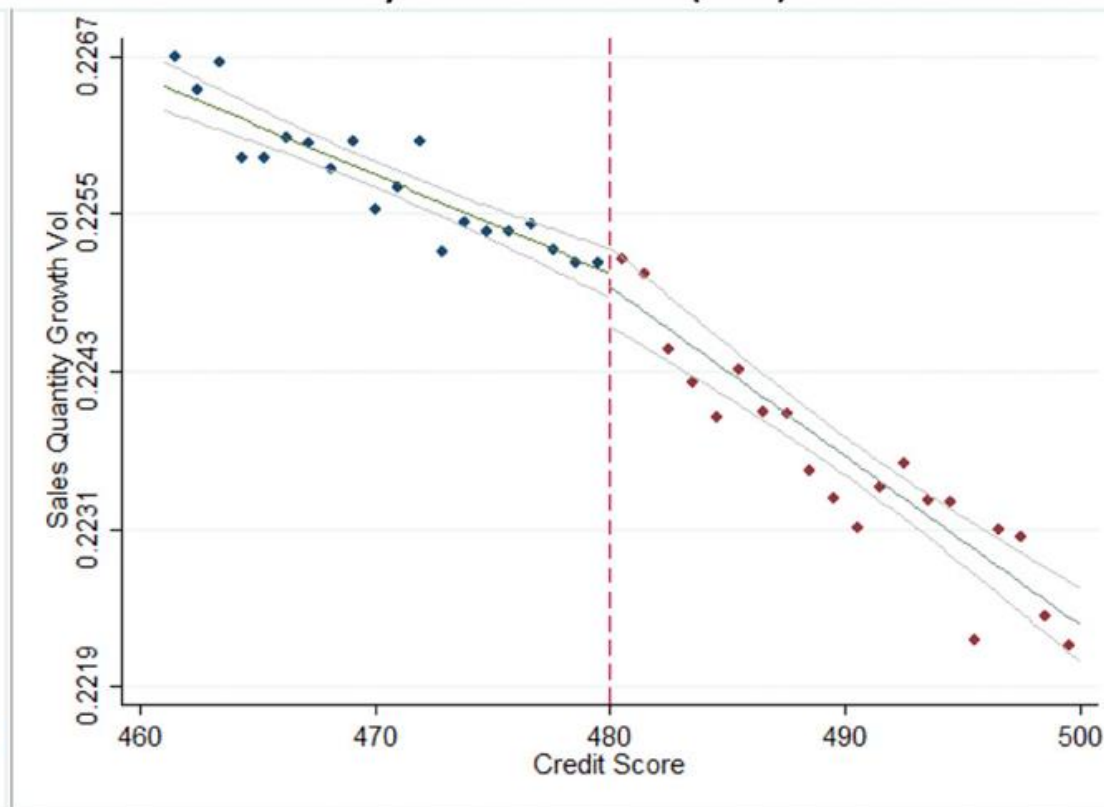
$$\begin{aligned} & SalesGrVol_{it+n} \\ &= \alpha + \beta \hat{D}_{it} + \sum_{k=1}^K \gamma_k (s_{it} - s^*)^k + T_{it} \sum_{k=1}^K \delta_k (s_{it} - s^*)^k + \varphi_j + \theta_t + \mu_{it} \end{aligned}$$

获得贷款前：断点无影响

(c) Sales Value Growth Vol ($T-1$)

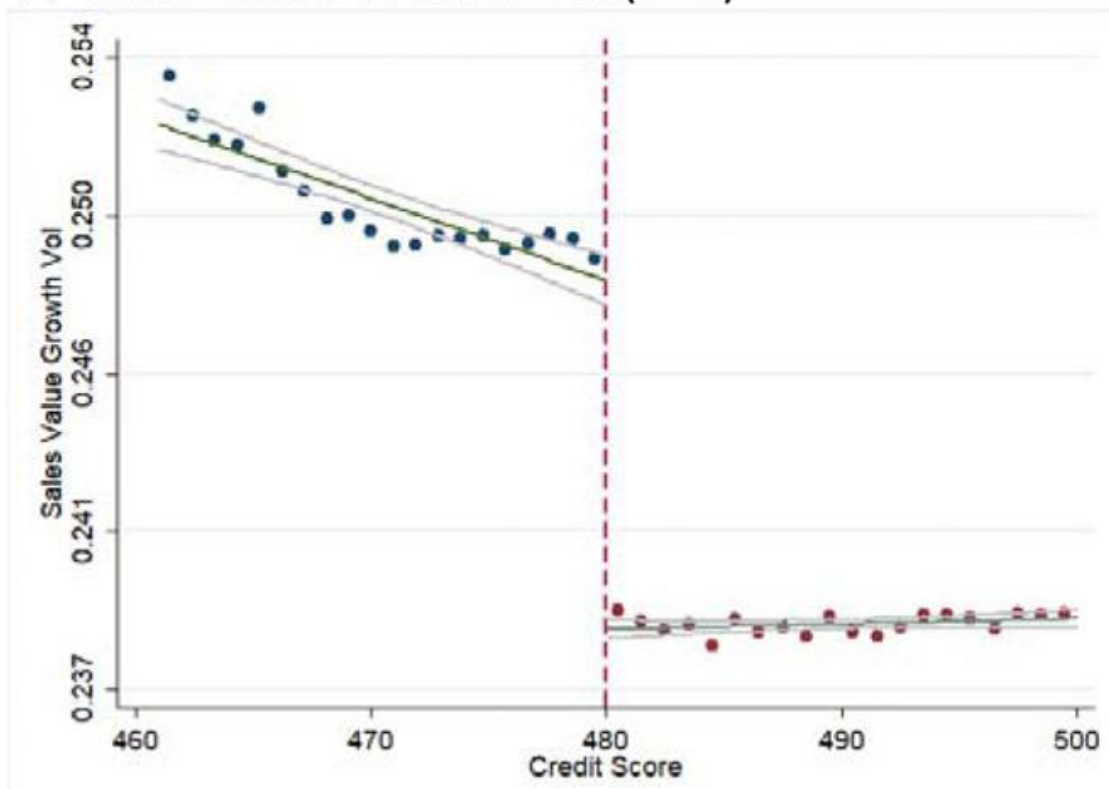


(d) Sales Quantity Growth Vol ($T-1$)

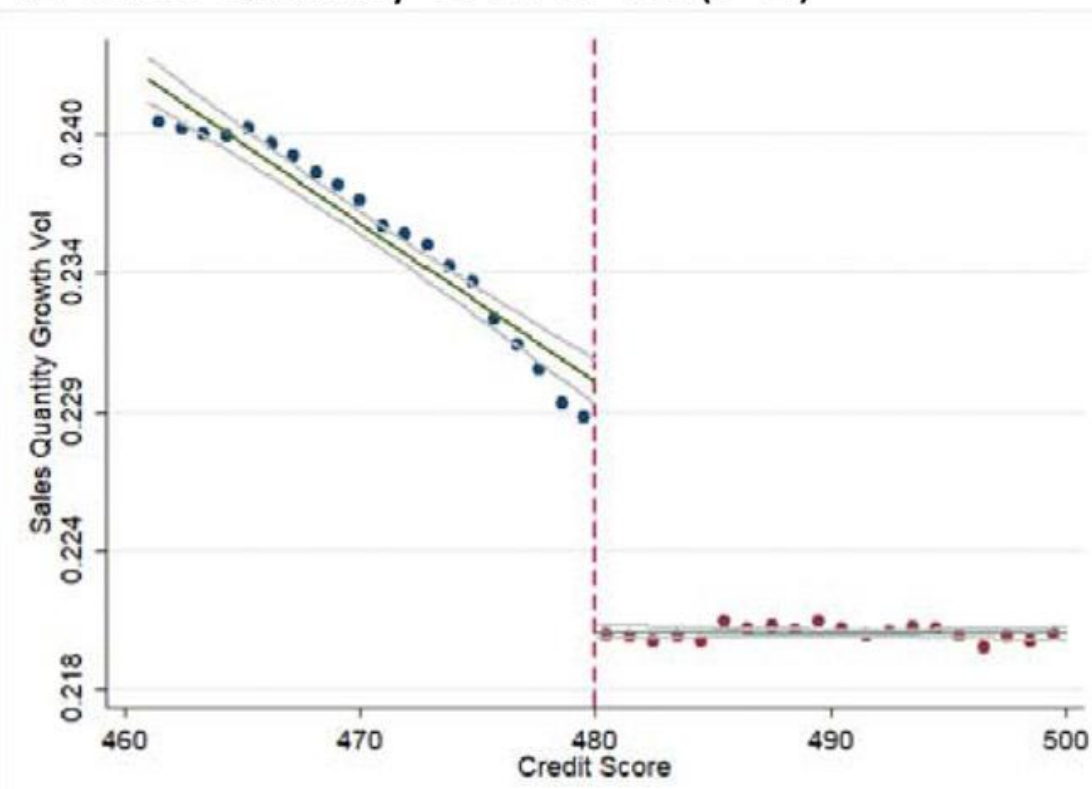


获得贷款后：断点有显著影响

(a) Sales Value Growth Vol ($T+1$)

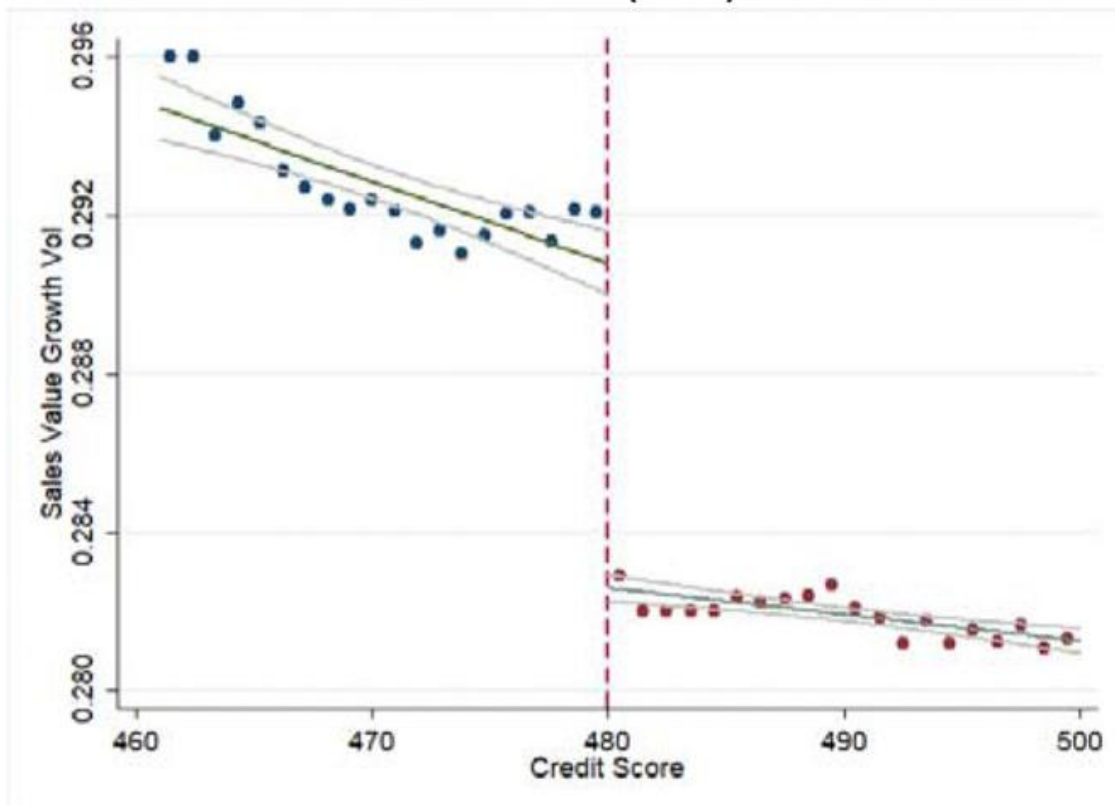


(b) Sales Quantity Growth Vol ($T+1$)

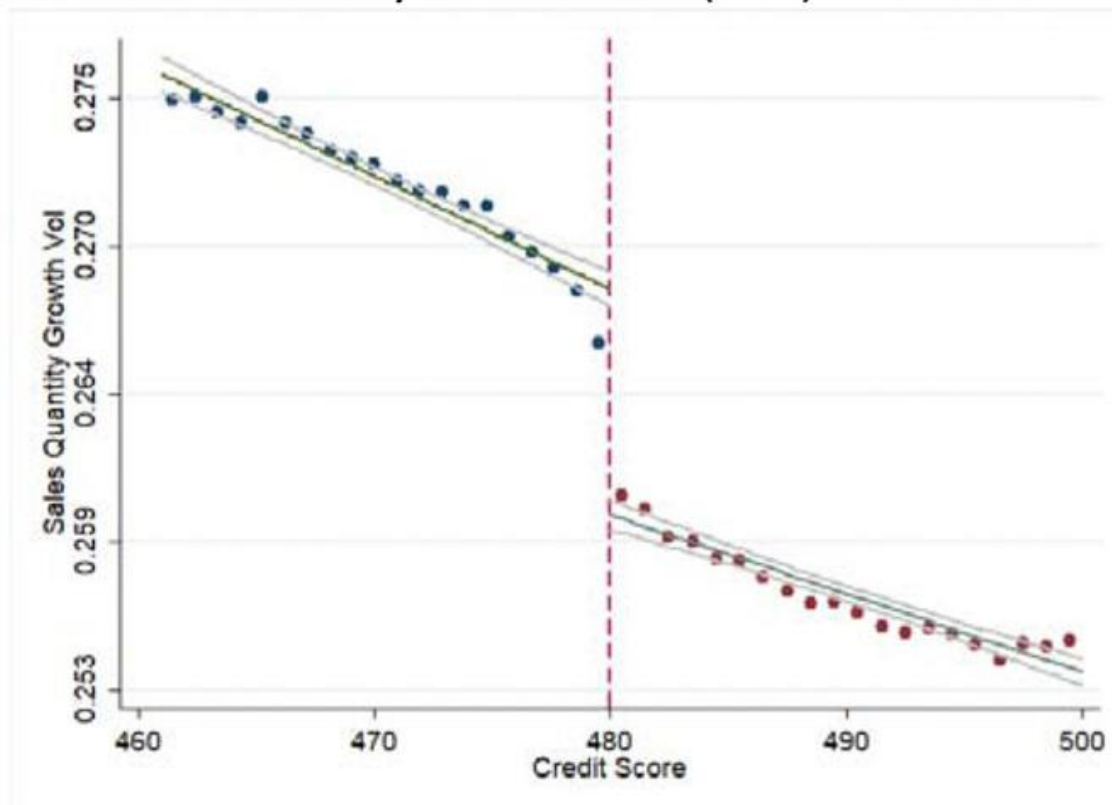


获得贷款后：断点有显著影响

(c) Sales Value Growth Vol ($T+2$)

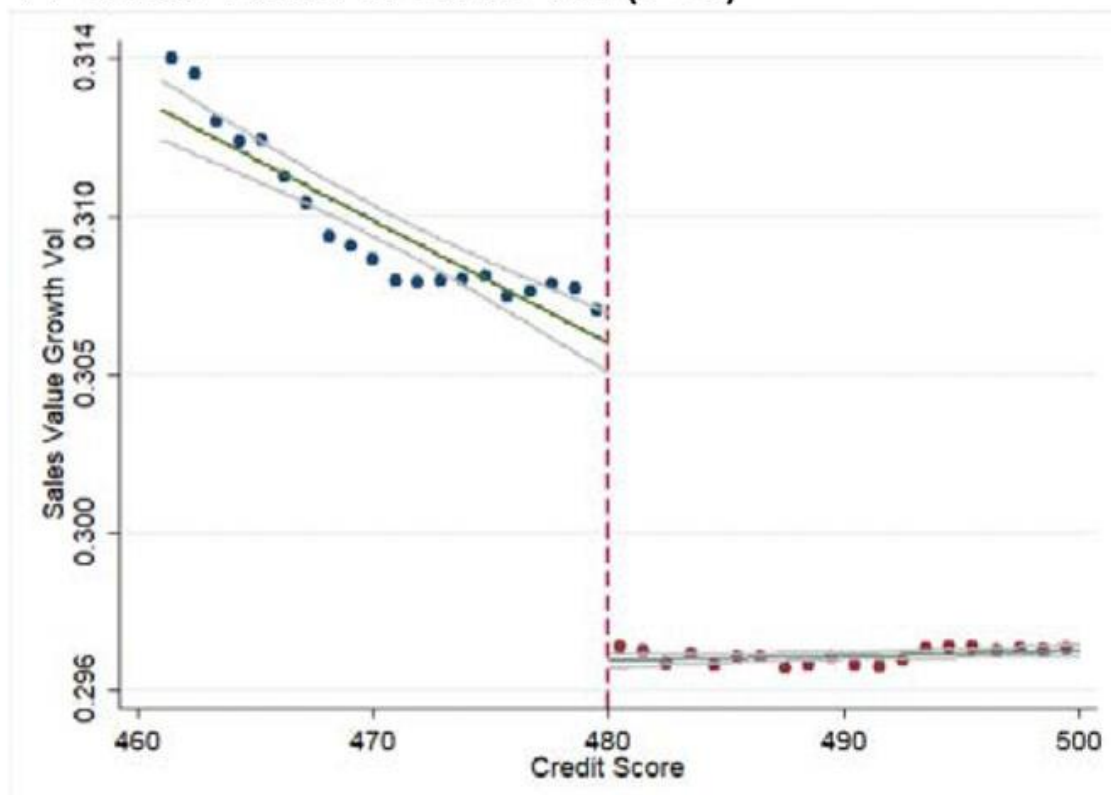


(d) Sales Quantity Growth Vol ($T+2$)

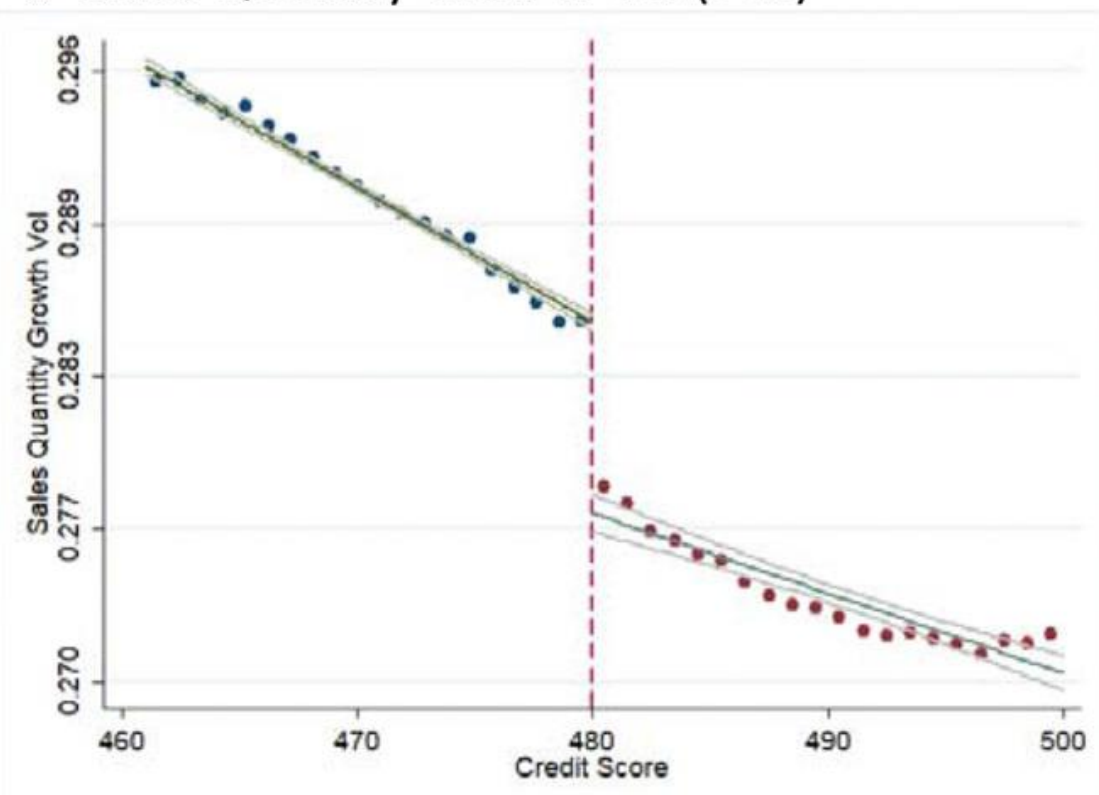


获得贷款后：断点有显著影响

(e) Sales Value Growth Vol ($T+3$)



(f) Sales Quantity Growth Vol ($T+3$)



回归结果：1阶段

Panel A. First Stage	
	Dependent variable
	<i>D</i> [Credit access]
	(1)
<i>T</i> [Credit score ≥ 480]	0.2274*** (104.7387)
Industry fixed effects	Yes
Time fixed effects	Yes
Adjusted R^2	0.3594
N	561,313

回归结果：2阶段

Panel B: Second Stage

	Dependent variable					
	<i>Sales value growth vol</i>			<i>Sales quantity growth vol</i>		
	<i>T + 1</i>	<i>T + 2</i>	<i>T + 3</i>	<i>T + 1</i>	<i>T + 2</i>	<i>T + 3</i>
	(1)	(2)	(3)	(4)	(5)	(6)
$\hat{D}$ [Predicted credit access]	−0.0382*** (−11.2082)	−0.0412*** (−12.7195)	−0.0391*** (−12.6505)	−0.0485*** (−10.0561)	−0.0428*** (−13.5527)	−0.0348*** (−11.3221)
Industry fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.0306	0.0317	0.0297	0.0319	0.0399	0.0404
N	561,313	561,313	561,313	561,313	561,313	561,313

安慰剂检验：更换断点

Panel A: First Stage (Placebo Cutoffs)		
	Dependent variable	
	<i>D</i> [Credit access]	
	(1)	(2)
<i>T</i> [Credit score $\geq$ 460 or 500]	0.03141 (0.0038)	-0.0128 (-1.3214)
Placebo cutoff	460	500
Score range	[440, 480]	[480, 520]
Industry fixed effects	Yes	Yes
Time fixed effects	Yes	Yes
Adjusted R^2	0.1585	0.1351
N	222,659	1,298,741

安慰剂检验：更换断点

Panel B: Second Stage (Placebo Cutoffs)

	Dependent variable			
	<i>Sales value growth vol</i>	<i>Sales quantity growth vol</i>	<i>Sales value growth vol</i>	<i>Sales quantity growth vol</i>
	(1)	(2)	(3)	(4)
$\hat{D}$ [Predicted credit access]	-0.0042 (-0.7490)	0.0011 (0.1946)	0.0144 (0.3764)	-0.0016 (-0.1577)
Placebo cutoff		460		500
Score range		[440, 480]		[480, 520]
Industry fixed effects	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes
Adjusted R^2	0.0311	0.0296	0.0347	0.0345
N	222,659	222,659	1,298,741	1,298,741

分值断点

- 以分值为断点进行识别，对数据依赖度极高
 - 通常的政策断点都是针对精细的微观指标，数据本身需要包括潜在的驱动变量
 - 财政、税收政策经常会有各类断点：规模、收入等
 - 公开数据较易获得，相应的RDD类文章较多：陆毅、陈钊、吕冰洋等
- 更重要的问题是对研究对象的深入理解：详细的制度背景、历史沿革、实操规律等