高级计量经济学

Lecture 14: Regression Discontinuity Design

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室　粤海校区汇文楼1510
E-mail　huangjp@szu.edu.cn
Website　https://huangjp.com

断点回归设计

断点回归设计原理

The Principle of RDD

断点回归设计（简称断点回归，RD）适用于处理水平不是随机分配，且协变量无法全部观测时。DID 和 SC 利用了面板数据消除协变量的影响。RD 针对的是处理变量 W_i 是另一变量的阈值函数的情况。

假设可观测协变量 X_i 可以按照下面定义的方式决定分配结果，

$$W_i = \begin{cases} 1 & X_i \geq c \\ 0 & X_i < c \end{cases},$$

我们称 X_i 为得分变量（score variable）或驱动变量（running variable），c 为断点（cutoff point）或阈值（threshold）。

例如，助学金的发放、政府保障房的申请（W_i）等往往要求家庭收入（X_i）低于某一水准（c）。而奖学金的发放（W_i）则要求成绩或成绩排名（X_i）高于某一水准（c）。

RD 设计的核心假设是：

- 在断点附近的个体无法决定自身的 X_i，此时出现在断点两侧近邻个体的处理分配接近随机。因此 RD 设计不需要假设分配的随机性，其应用场景更加广泛。

RD 设计与平均处理效应：精确 RD

RDD and ATE: Sharp RD

RD 设计包含很多变种，我们从最简单的精确 RD （sharp RD）开始介绍。

Sharp RD 是指 W_i 的取值完全服从定义，换句话说是所有个体都服从对处理的分配。与其相对的是模糊 RD（fuzzy RD），即有些个体不服从对处理的分配。

对于 sharp RD，$X_i = c$ 时的平均处理效应可以定义为

$$\begin{aligned}\tau_{\text{SRD}} &= E[Y_{1i} - Y_{0i} \mid X_i = c] \\ &= \lim_{x \downarrow c} E[Y_i^{\text{obs}} \mid X_i = x] - \lim_{x \uparrow c} E[Y_i^{\text{obs}} \mid X_i = x]\end{aligned}$$

τ_{SRD} 的识别条件是：

- 连续性假设（continuity assumption）：$E[Y_{1i} \mid X_i = x]$ 和 $E[Y_{0i} \mid X_i = x]$ 是 x 的连续函数。（Hahn et al., 2001）

  从连续性假设可知，其他影响潜在结果的协变量在 $X_i = c$ 时不会发生跳跃性变化。否则，我们无法确定处理效应是否源自 W_i。

基于连续性假设，我们可以把断点左侧的平均观测值（对照组）作为断点右侧的平均观测值（处理组）在未接受处理时的潜在结果的估计量。

Hahn, Todd, & van der Klaauw (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69:1, 201-209.

下图取自
Abadie & Catteneo (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10:465-503.

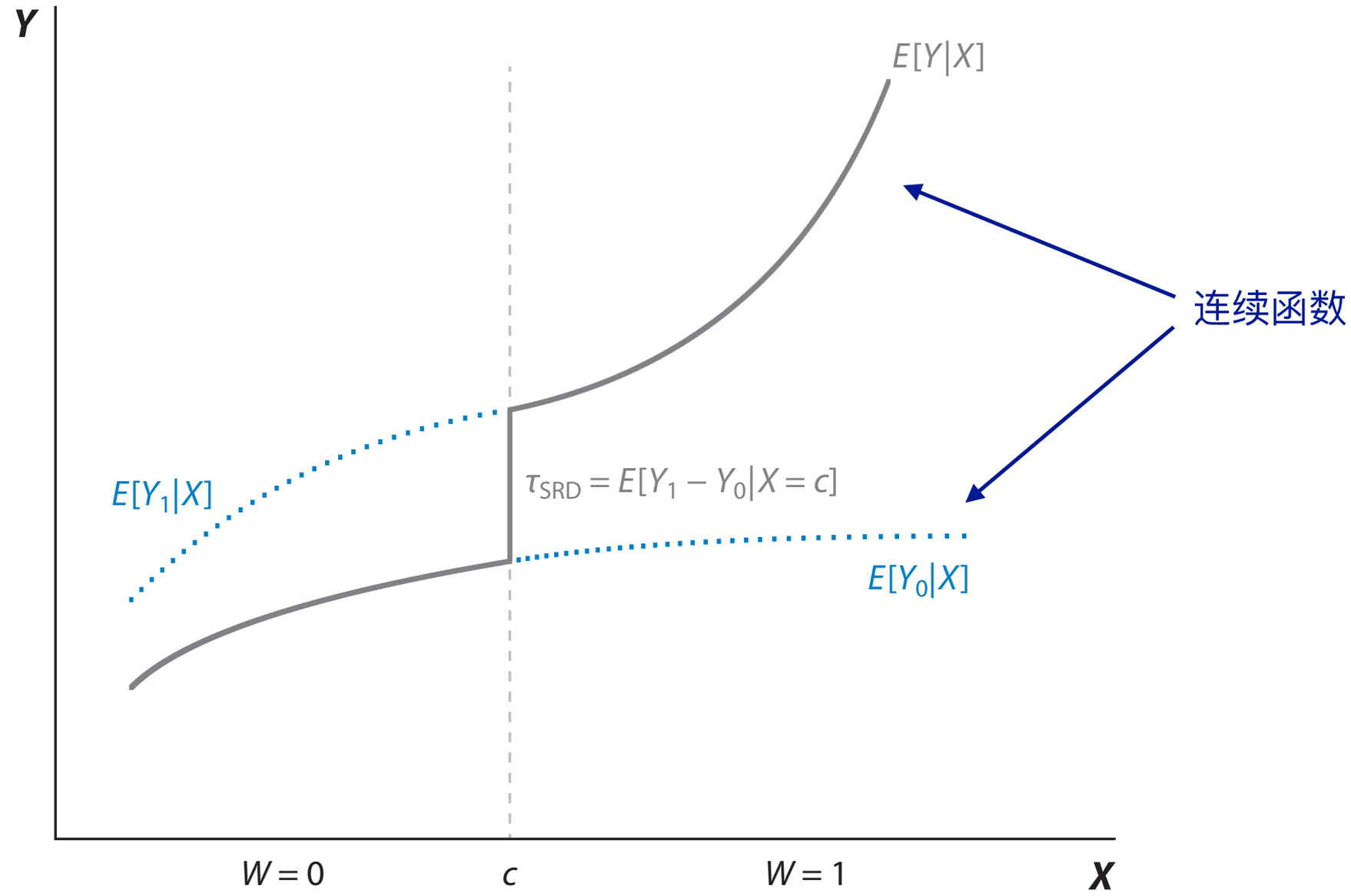


Figure 7

Regression discontinuity design. The outcome of interest is Y; the score is X, and the cutoff point is c. Treatment assignment and status is $W = 1(X \geq c)$. The graph includes the estimable regression functions (*solid lines*) and their corresponding unobservable portions (*dotted lines*).

τ_{SRD} 的非参估计

Nonparametric Estimation of τ_{SRD}

- 核密度估计（kernel density estimation）

  选取一个带宽（bandwidth）$h > 0$，然后选取 $X_i \in [c-h, c+h]$ 内的个体 i，

$$\hat{\tau}_{\text{SRD}} = \frac{\sum_{i:X_i \in [c,c+h]} Y_i^{\text{obs}}}{\sum_{i:X_i \in [c,c+h]} 1} - \frac{\sum_{i:X_i \in [c-h,c)} Y_i^{\text{obs}}}{\sum_{i:X_i \in [c-h,c)} 1}$$

  这个估计量是边界核密度估计的一个特殊形式（采用长方形核密度函数，详见 Imbens & Lemieux, 2008）。此估计量的估计偏差是 h 的线性函数，收敛很慢。

- 局部线性回归（local linear regression）

  在断点 c 左右两侧的带宽内，用 $(X_i - c)$ 对 Y_i^{obs} 进行回归：

$$\text{断点左侧：} \quad Y_i^{\text{obs}} = \alpha_\ell + \beta_\ell (X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c-h, c)$$
$$\text{断点右侧：} \quad Y_i^{\text{obs}} = \alpha_r + \beta_r (X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c, c+h]$$

  则 $\hat{\tau}_{\text{SRD}} = \hat{\alpha}_r^{\text{OLS}} - \hat{\alpha}_\ell^{\text{OLS}}$。

Imbens & Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615-635.

τ_{SRD} 的回归估计

Estimation of τ_{SRD} Using Regression

我们也可以用下面的全域线性回归模型对 τ_{SRD} 进行估计

$$Y_i^{\text{obs}} = \alpha + \beta(X_i - c) + \tau W_i + \gamma(X_i - c) \times W_i + \varepsilon_i$$

在这个模型中，

$$\alpha = E[Y_{0i} \mid X_i = c]$$

$$\tau = E[Y_{1i} - Y_{0i} \mid X_i = c]$$

β 和 $\beta + \gamma$ 分别是断点左右两侧的斜率。因此，$\hat{\tau}_{\text{SRD}} = \hat{\tau}_{\text{OLS}}$。

此处的回归模型和局部线性回归的区别为是否要把观测值限定在带宽之内，而从方法论上说，回归（不限定带宽）属于参数估计，局部线性回归（限定带宽）属于非参估计。

有时回归的线性假设可能太强，这时候也可以加入 $X_i - c$ 的高次项。但是需要注意的是，加入三次以上的项弊大于利，具体可参考 Gelman & Imbens (2017)。

关于如何选择最优带宽，可参考 Imbens & Kalyanaraman (2012)。

Gelman & Imbens (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 37:3, 447-456.

Imbens & Kalyanaraman (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *Review of Economic Studies*, 79, 933-959.

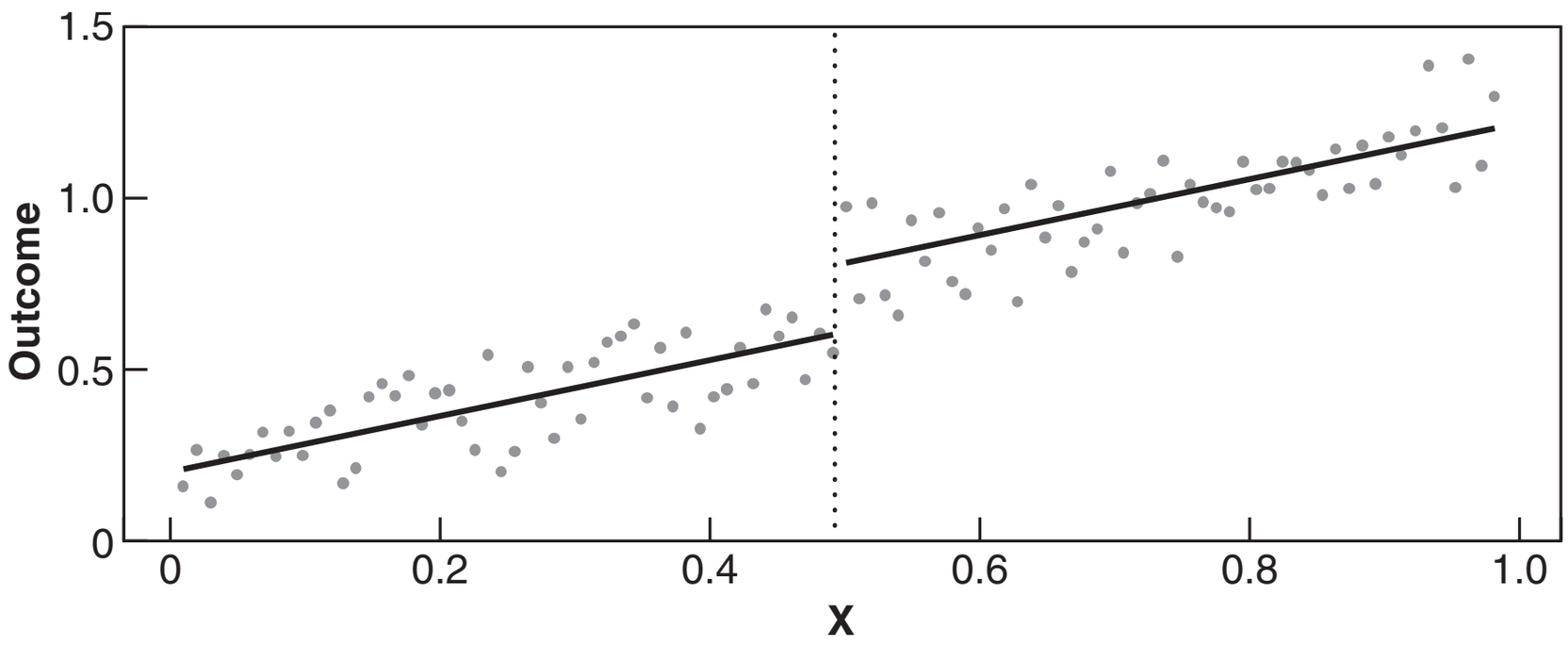
图形分析

Graphical Analysis

RD 设计相比其他方法的优势之一是可以通过一些图形分析法方便地判断平均处理效应的有无。

最基本的图形分析是 RD plot。这是一种类似直方图的数据整合方法，目的是将观测结果 Y_i^{obs} 在一些区间中的均值以散点图的形式展现。具体方法是：

- 在断点 c 左右两侧各确定区间个数 K_ℓ 和 K_r，确定区间宽度，并计算每个区间 k 的左右端点；
- 针对 X_i 取值在区间 k 中的个体 i，计算其观测结果 Y_i^{obs} 的均值 $\bar{Y}_k^{\text{obs}}$；
- 将每个 $\bar{Y}_k^{\text{obs}}$ 画在区间 k 的中点处。



此图取自
Angrist & Krueger (2009).
Mostly Harmless Econometrics.

- 图中也可加入回归曲线。
- 也可针对其它可观测协变量作图，目的是确认连续性假设是否成立。

模糊 RD

Fuzzy RD

在 sharp RD 的设定下，所有个体都服从对处理的分配，即给出 X_i 的取值时，$W_i = 1$ 的概率或为 0 或为 1。

Fuzzy RD 设计对以上设定做出让步，即允许一部分人不服从分配，但要满足下面的条件

$$\lim_{x \downarrow c} \Pr(W_i = 1 \mid X_i = x) \neq \lim_{x \uparrow c} \Pr(W_i = 0 \mid X_i = x)$$

Sharp RD 可以看作 fuzzy RD 的一种特殊情况，即上式两边的概率之差为 1。

Fuzzy RD 设计中 $X_i = c$ 时的局部平均处理效应（LATE）定义为

$$\begin{aligned} \tau_{\text{FRD}} &= E[Y_{1i} - Y_{0i} \mid X_i = c, i \text{ is a complier}] \\ &= \frac{\lim_{x \downarrow c} E[Y_i^{\text{obs}} \mid X_i = x] - \lim_{x \uparrow c} E[Y_i^{\text{obs}} \mid X_i = x]}{\lim_{x \downarrow c} E[W_i \mid X_i = x] - \lim_{x \uparrow c} E[W_i \mid X_i = x]} \end{aligned}$$

此处需要假设单调性条件：

- 单调性（monotonicity）：$W_i(c)$ 为 c 的非增加函数（不存在 defier）。
  Complier：当 $c \leq X_i$ 时，$W_i = 1$；当 $c > X_i$ 时，$W_i = 0$
  Always taker：$W_i = 1$
  Never taker：$W_i = 0$

τ_{FRD} 的估计

Estimation of τ_{FRD}

为了估计 τ_{FRD}，只需找到其定义中四个极限条件期望的估计量。在 sharp RD 的估计中我们已经找到了分子的非参估计量（核密度估计），这里也可以用同样的方法估计分母。但是这种方法偏差较大。

更好的方法是采用局部线性回归进行估计。和前面一样，我们在断点 c 左右两侧的带宽内考虑四个回归：

$$\begin{aligned}
&\text{断点左侧结果变量：} \quad Y_i^{\text{obs}} = \alpha_{y\ell} + \beta_{y\ell}(X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c-h, c) \\
&\text{断点右侧结果变量：} \quad Y_i^{\text{obs}} = \alpha_{yr} + \beta_{yr}(X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c, c+h] \\
&\text{断点左侧处理变量：} \quad W_i = \alpha_{w\ell} + \beta_{w\ell}(X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c-h, c) \\
&\text{断点右侧处理变量：} \quad W_i = \alpha_{wr} + \beta_{wr}(X_i - c) + \varepsilon_i \quad \text{for } X_i \in [c, c+h]
\end{aligned}$$

则 τ_{FRD} 的估计量是 $\hat{\tau}_{\text{FRD}} = \dfrac{\hat{\alpha}_{yr}^{\text{OLS}} - \hat{\alpha}_{y\ell}^{\text{OLS}}}{\hat{\alpha}_{wr}^{\text{OLS}} - \hat{\alpha}_{w\ell}^{\text{OLS}}}$

也可以考虑下面的回归模型：

$$Y_i^{\text{obs}} = \alpha_{y\ell} + \beta_{y\ell} 1_{\{X_i < c\}}(X_i - c) + \beta_{yr} 1_{\{X_i \geq c\}}(X_i - c) + \tau W_i + \varepsilon_i$$

采用 $1_{\{X_i \geq c\}}$ 作为 W_i 的工具变量对 τ 进行 2SLS 估计也可以得到相同的 $\hat{\tau}_{\text{FRD}}$。

实证研究

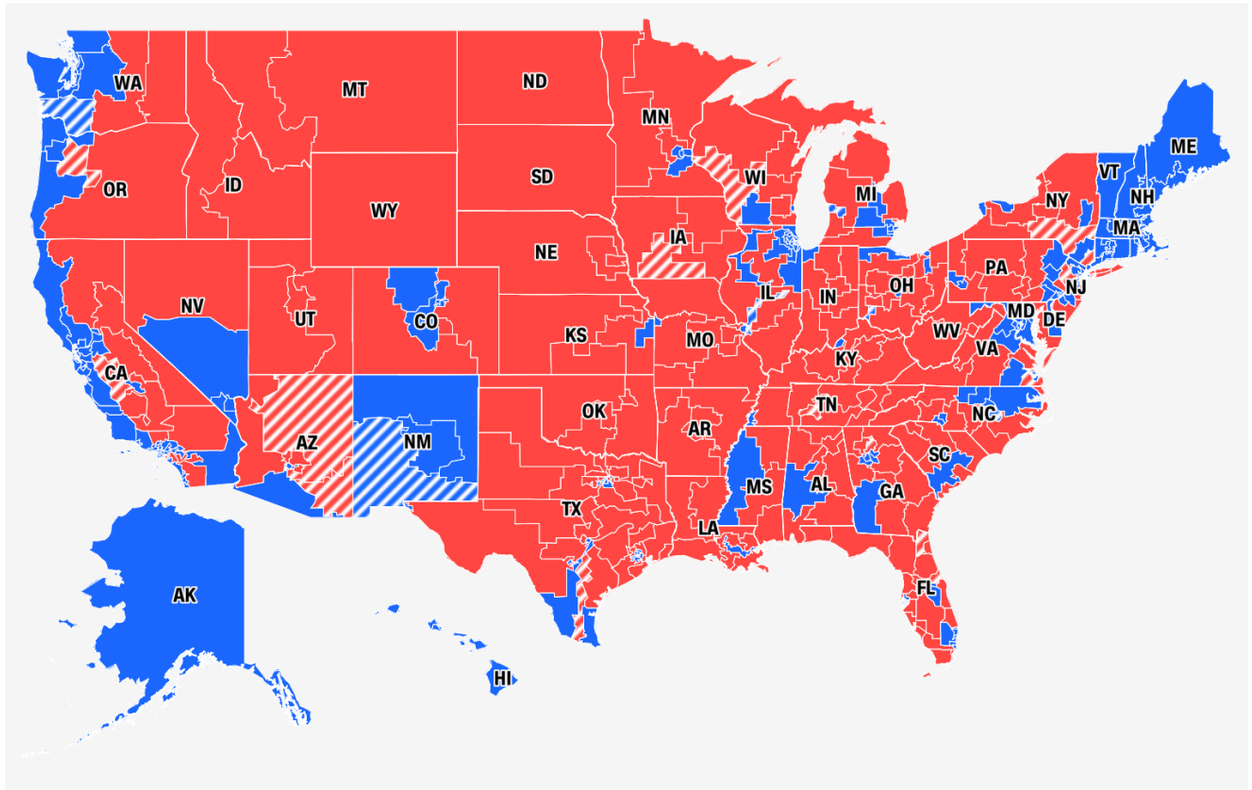
Lee (2008)

核心问题：在美国的众议院选举中是否存在执政党优势

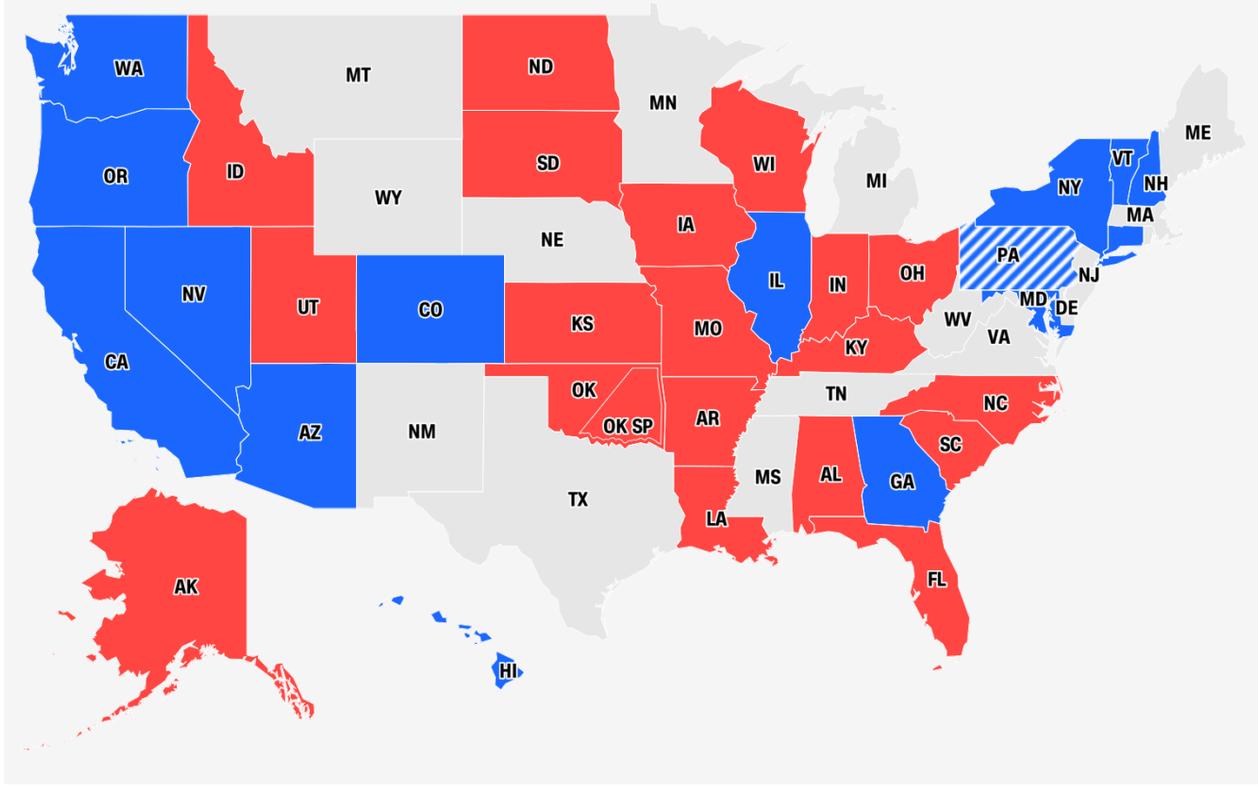
背景：

- 美国议会（Congress）分为众议院（House of Representatives）和参议院（Senate）
  - 众议院按人口比例划分选区（每个州至少有一个选区，现在共有435个选区），从每个选区选出一名议员。议员任期为两年，每两年进行一次全体改选。
- 参议院以州为单位进行选举，从每个周选出两名议员（共有100个议席）。议员任期为六年，每两年有 1/3 的议员需要改选（每个州每次最多改选一名）。
- 州有权决定州内的选举方式。例如，佐治亚州和路易斯安那州要求过半数的候选人才能当选（若没有，则在得票数前两位的候选人间进行第二轮选举，得票数多者当选）；其他州要求得票数最多的候选人当选（不要求过半数）。
- 2022年改选结果：
  - 参议院：民主党 51 v.s. 共和党 49
  - 众议院：民主党 213 v.s. 共和党 222

2022 年美国议会选举结果



众议院



参议院

众议院的议席总数固定在 435，但是选区会根据人口变化而动态调整。
美国每十年会进行一次人口普查，根据人口普查结果，每十年向各州重新分配议席数。
例如在2022年选举中，得克萨斯州增加了两个议席，另外有五个州各增加了一个议席。加利福尼亚州等七个州各减少了一个议席。

美国的总统选举（题外话）

- 美国的总统选举采用“选举人团”模式，是一种两阶段间接选举方法。
  - 第一阶段-普选：以州为单位对总统和副总统候选人进行公民投票。投票人选择自己支持的候选人进行投票。
  - 第二阶段-选举人团选举：在州内普选中获胜的候选人所属党派会按事先分配的名额选出一定数量的“选举人”。全国的选举人在一起投票选出总统和副总统。
  - 对于大多数州来说，赢得该州普选的总统候选人会赢得该州所有选举人的支持（“赢者通吃”）。内布拉斯加州和缅因州会分配选举人给不同的总统候选人。

- 选举人的人数
  - 每个州的选举人人数 = 该州的众议院议席数 + 2（参议院议席数）
  - 华盛顿哥伦比亚特区拥有 3 个选举人
  - 现在全国共有 435 + 100 + 3 = 538 个选举人。总统和副总统候选人需要获得过半数的 270 票才能当选。如果低于这个票数，应由众议院投票选出。众议院也无法选出时，由参议院选出。

- 选举人原则上需要忠于自己的党派（忠于州内的普选结果），但也有背叛者出现（2016年特朗普对希拉里克林顿的选举中，有7名选举人背叛了自己的党派）。

变量与模型

本文中的执政党优势（incumbency advantage）指在选举中，候选人和现任议员同属一个党派给选举结果带来的正向影响。因此样本中的个体为选区。

数据显示，在 1948-1998 年间，执政党（选区内议员所属的党）所属候选人在下一次选举中当选的概率一直保持在 90% 左右。现任议员再次参选并连任的概率也保持在80%左右。但这些数字并不能直接说明执政党优势的因果关系，因为有太多协变量对其进行干扰。

我们无法在控制其他变量的同时随机改变执政党。但是幸运的是，执政党由得票率决定，因此可以用 RDD 对因果效应进行估计。

假设每个地区只有两个党派，在选举中得票率超半数的党派为执政党。令 v_{it} 为民主党候选人在 i 选区的 t 年度选举中的得票率（$t \in 1,2$），d_{i2} 是民主党在 i 选区的下一届选举中是否是现任执政党的虚拟变量，$\boldsymbol{w}_{i1}$ 是影响现任议员选举结果的协变量（选区选民的政治倾向、政党的政治资源、候选人资质等）。作者考虑下面的回归模型：

$$v_{i2} = \boldsymbol{w}_{i1}^{\top}\boldsymbol{\alpha} + \beta v_{i1} + \gamma d_{i2} + e_{i2}, \qquad d_{i2} = 1_{\{v_{i1} \geq 0.5\}}$$

并假设条件密度函数 $f_{i1}(v_{i1} = v \mid \boldsymbol{w}_{i1} = \boldsymbol{w})$ 关于 v 连续，$E[e_{i2} \mid \boldsymbol{w}_{i1}, v_{i1}] = 0$。

此模型中的 γ 代表执政党在下一届选举中的优势。$\boldsymbol{w}_{i1}$ 通常包含不可观测变量，因此 OLS 估计不准确。

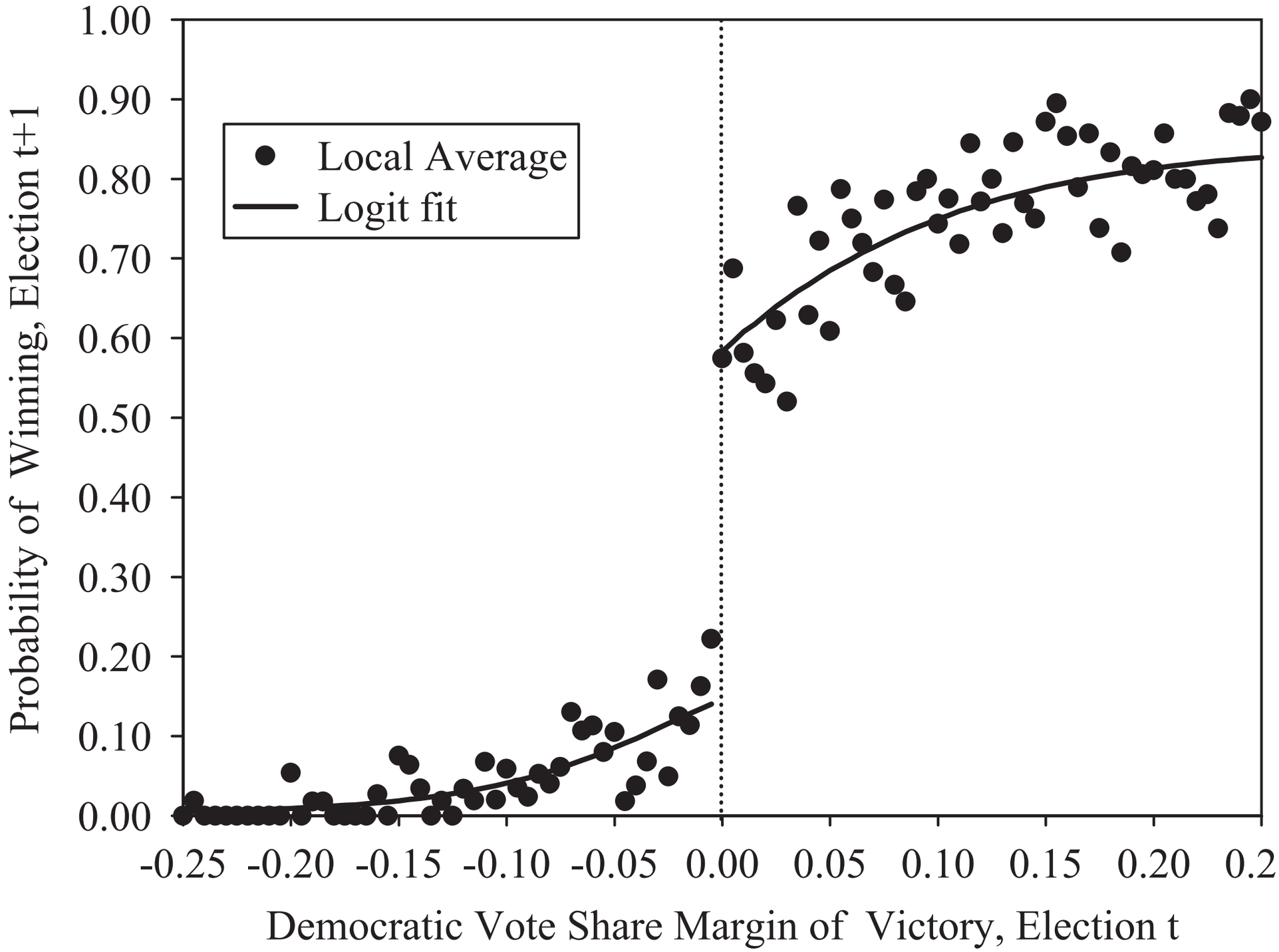
基于 RDD 的估计

作者采用了另一种 RDD 的解释，即在断点附近，对处理的分配接近随机。通过假设选民在投票时有随机外生因素对其决策产生影响，我们可以认为断点附近有近似随机的处理分配。

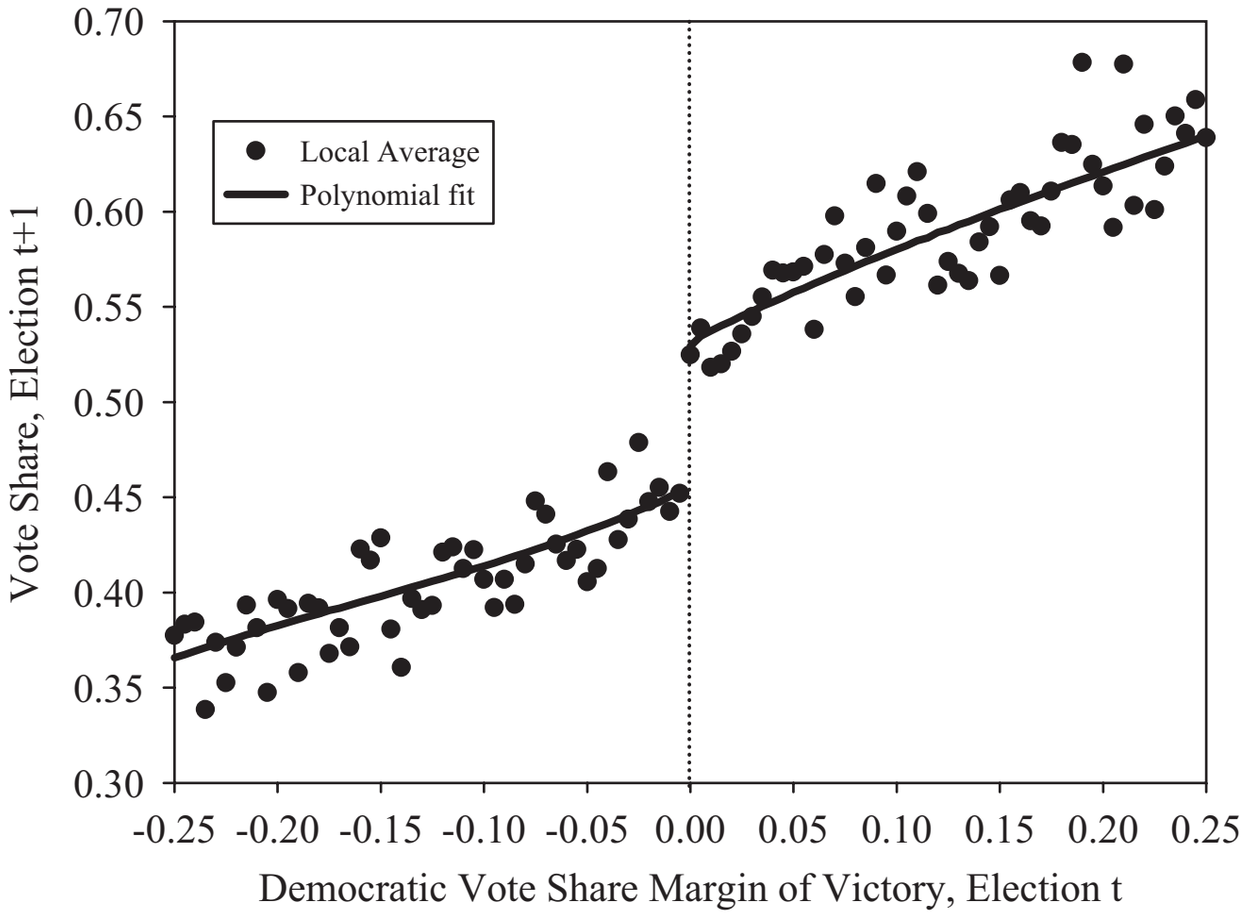
将现任选举中（$t = 1$）民主党以校服优势当选的州作为处理组，以小幅劣势落选的州作为对照组，通过比较它们在下一届选举中（$t = 2$）的平均得票率之差可以估计平均处理效应。

本文中的样本包含1946-1998年间的选举结果，将其中1946-1996年间的数据作为现任选举（$t = 1$），1948-1998年间的数据作为下一届选举（$t = 2$）。由于每十年一次的选区变更，加上独立变量包含两届连续选举中的数据，因此作者将个位为 0 和 2 的年份从数据中剔除。

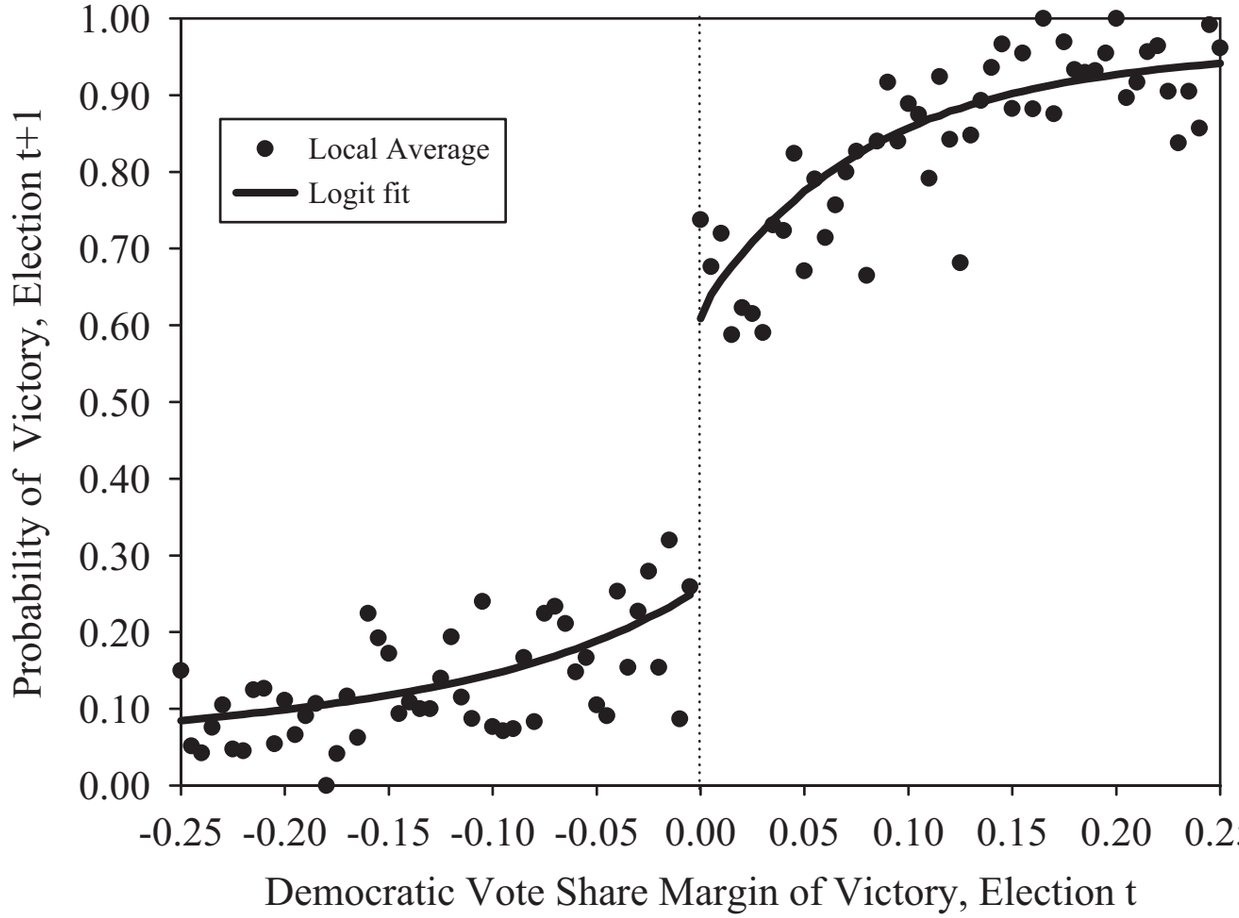
美国并不是严格意义上的两党制，有些州中第三方候选人的得票率也不可忽视，导致真正的胜出线不在 50%。因此，作者用民主党候选人的得票率减去其他候选人中的最高得票率作（Democratic vote share margin）为得分变量，并将断点值设为 0 。



图中的点是区间内民主党议员是否参加并赢得下一届选举的样本均值，区间宽度为 0.005（即 5‰）。RD 估计显示，民主党执政期的连任概率比非执政期高约 45%。



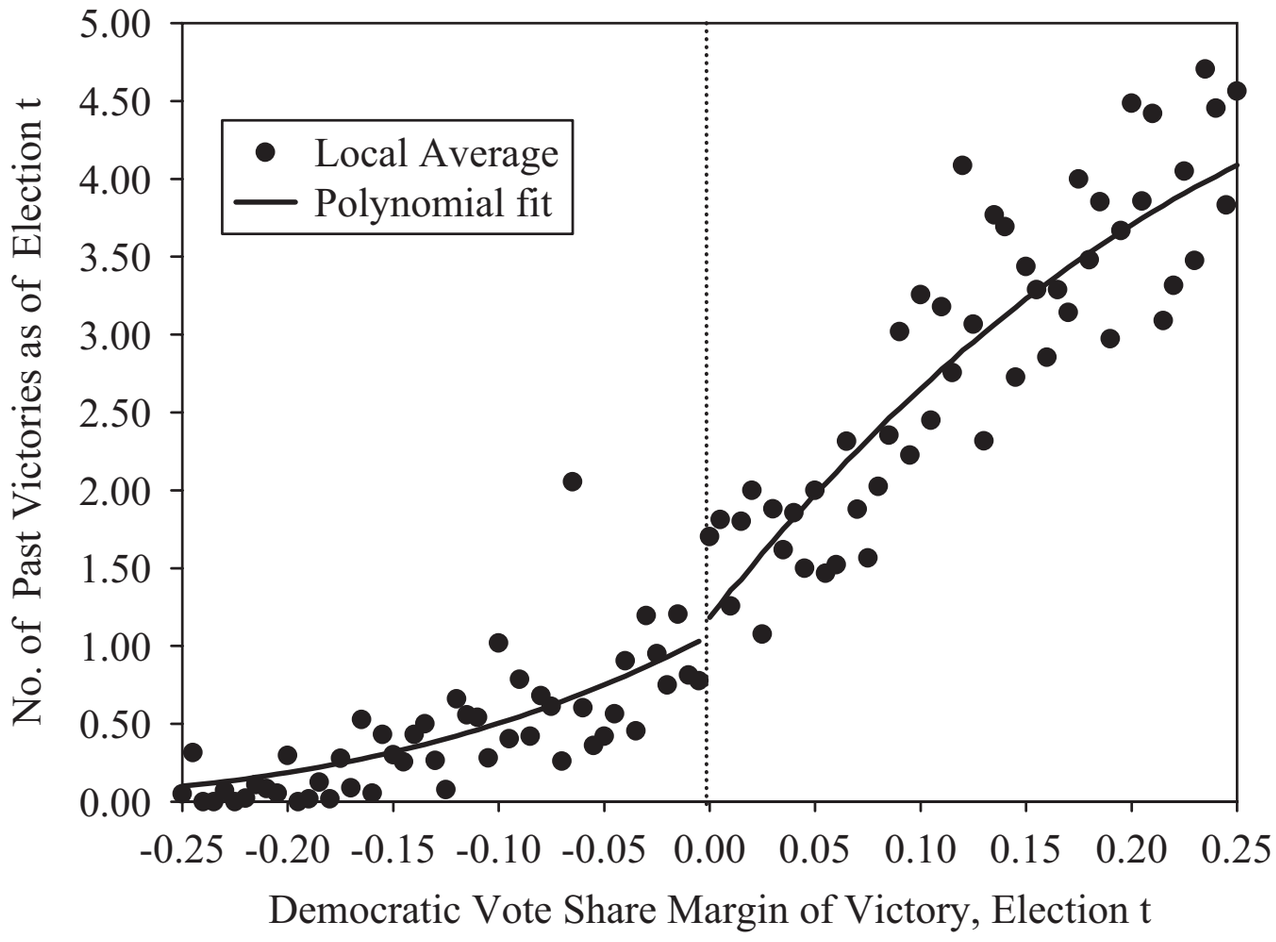
民主党在下一届选举中的得票率。
执政党优势约为 7%-8%。



民主党在下一届选举中的当选率。
执政党优势约为 35%。

对连续性假设的检验

如果连续性假设成立，则任何处理前协变量都应满足连续性。此时针对协变量进行 RD 估计不会得到有效的结果。



针对民主党议员连任次数进行 RD 估计，结果并不显著，侧面验证了连续性假设。

Hoekstra (2009)

The Effect of Attending the Flagship State University on Earnings: A Discontinuity-Based Approach
***Review of Economics and Statistics*, 91:4, 717-724.**

核心问题：大学质量对收入的影响

作者针对是否就读一流公立大学（flagship state university）对毕业后收入的影响进行了RDD 研究。

数据：

- 某一流公立大学（匿名）的申请人信息（白人学生，1986-1989年间申请该学校）
  包括：社会保险号码、性别、申请入学学期、SAT得分、该学生是否入学、该学生高中阶段的GPA。（这类数据通常非常难以获得）
- 该大学所在州的失业保险缴税信息
  匹配了上面数据中的申请人在1998Q1-2005Q2间的季度收入。

作者选择了样本中白人男性在其28-33岁间（高中毕业后10-15年）的收入作为分析对象。文中的断点为入学录取条件（与SAT和GPA有关，后面详述），因此作者舍弃了样本中成绩特别好和特别差的申请人，最终保留了12189个申请人，其中有8424个在其28-33岁间有至少一年的收入数据。

大学录取中的断点

在文中讨论的时代，大学入学录取是基于 SAT 得分和 GPA 分数的非线形函数。每一个 GPA 分数都有一个对应的 SAT 录取线，而该录取线随 GPA 分数的升高而降低。

作者将录取规则整合为一个一维函数：

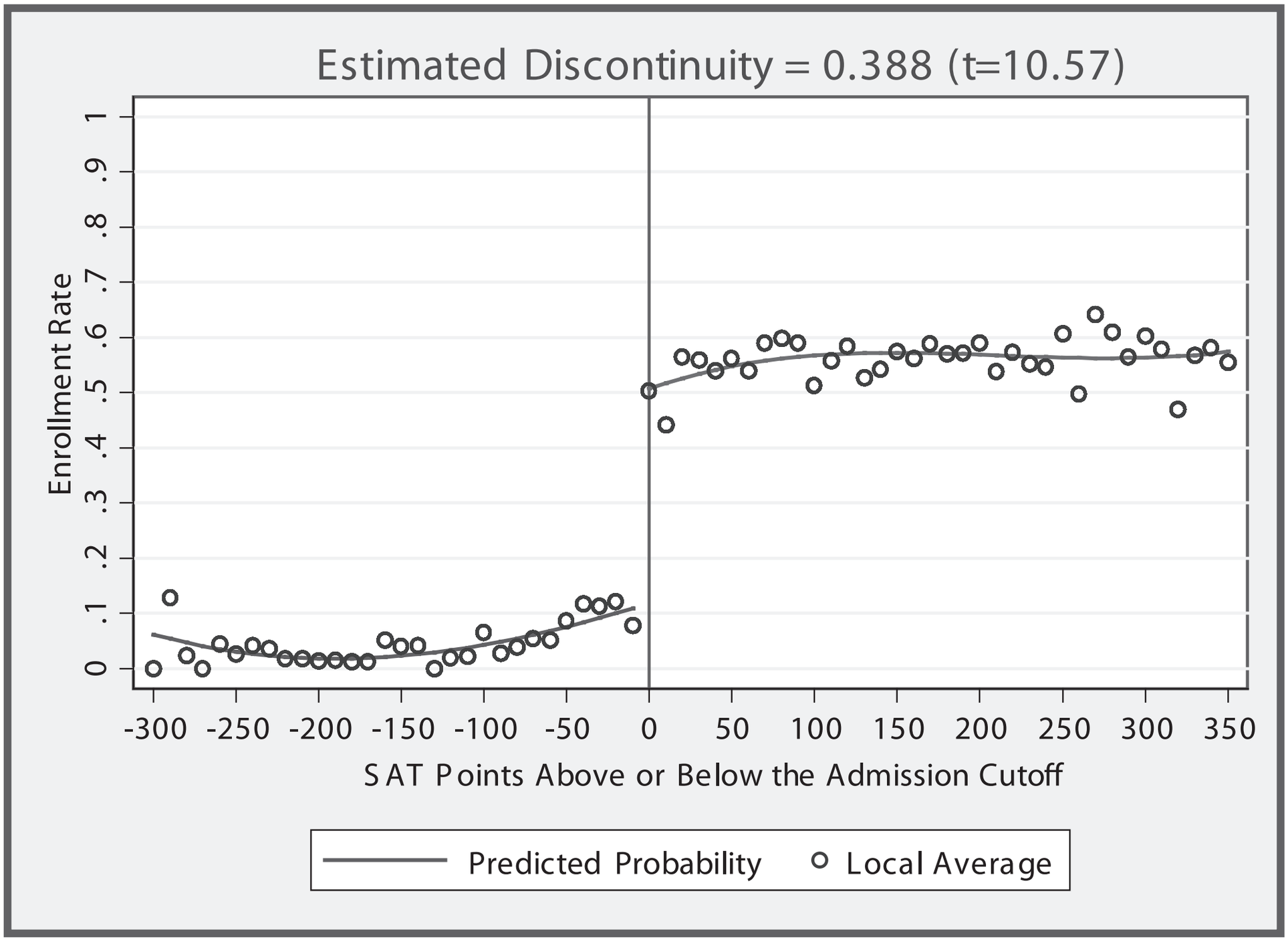
- 针对每一个 GPA 分数，从申请人的 SAT 得分中减去该 GPA 对应的录取分数线，得到调整后的 SAT 得分（adjusted SAT）。
- 这样做意味着所有申请人的 adjusted SAT 录取线都为 0，而 adjusted SAT 得分为正数者被录取。

录取分数线的缺失与估计：

- 大学没有记录具体的 SAT 录取分数线！因此作者用下面的方法对每个组别（申请入学学期 × GPA 分数）的录取分数线进行了估计。
- 假设录取结果服从回归模型 $1_{\text{Acceptance}} = \beta_0 + \beta_1 1_{\text{SAT} \geq \text{cutoff}} + \varepsilon$。作者尝试了所有可能的 cutoff 进行 OLS 估计，从中选出令 R^2 最大的 cutoff 值作为该组别的 SAT 录取分数线。

入学率在断点附近的变化

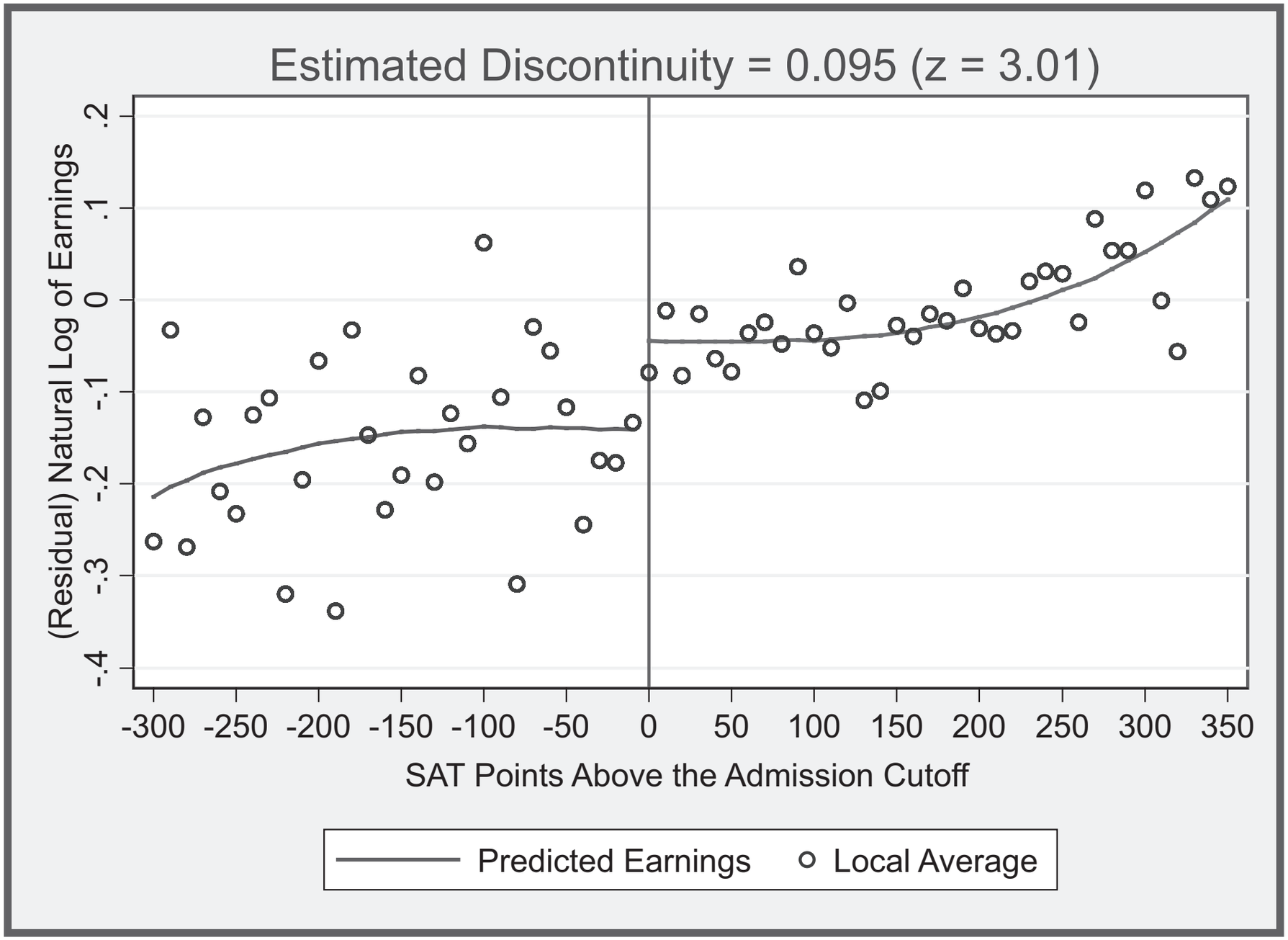
FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



注意：录取 ≠ 入学
因此作者采用的是
fuzzy RD。

收入在断点附近的变化

FIGURE 2.—NATURAL LOG OF ANNUAL EARNINGS FOR WHITE MEN TEN TO FIFTEEN YEARS AFTER HIGH SCHOOL GRADUATION (FIT WITH A CUBIC POLYNOMIAL OF ADJUSTED SAT SCORE)



敏感性分析

TABLE 1.—EARNINGS DISCONTINUITIES AND CORRESPONDING INTENT-TO-TREAT AND ENROLLMENT ESTIMATES FOR WHITE MEN

Regression Specification	Function of Adjusted SAT	Flexible Polynomial?	Additional Controls	Discontinuity: Estimated Earnings Discontinuity	Treatment Effect: Intent-to-Treat Effect	Treatment Effect: Enrollment Effect
(1) Plotted in Figure 2	Cubic	No	No	0.095***	0.135***	0.223***
				(0.032)	(0.046)	(0.079)
				[0.003]	[0.004]	[0.005]
(2)	Cubic	No	Yes	0.092***	0.131***	0.216***
				(0.033)	(0.048)	(0.081)
				[0.005]	[0.006]	[0.008]
(3)	Quadratic	Yes	Yes	0.111**	0.170**	0.281**
				(0.045)	(0.073)	(0.121)
				[0.014]	[0.019]	[0.021]
(4) (includes only applicants within 200 points of cutoff)	Quadratic	No	Yes	0.081**	0.116**	0.192**
				(0.038)	(0.056)	(0.094)
				[0.034]	[0.038]	[0.041]
(5) (includes only applicants within 100 points of cutoff)	Linear	No	Yes	0.074**	0.110*	0.181*
				(0.038)	(0.058)	(0.099)
				[0.050]	[0.060]	[0.067]

Notes: Bootstrapped standard errors are in parentheses; p-values are given in brackets. Additional controls include (residual) SAT score and (residual) high school GPA. "Flexible polynomial" indicates whether the estimated coefficients of the adjusted SAT score polynomial were allowed to differ on each side of the admission cutoff. *, **, and ***: statistical significance at the 10%, 5%, and 1% levels, respectively. Intent-to-treat and enrollment effects are estimated using two-stage least squares.

在改变回归函数、控制变量及带宽的情况下，断点附近的收入差的估计值在 7.4%-11.1% 之间。

作者用 2SLS 对 LATE 进行了估计（读一流公立大学对收入的局部平均处理效应），其结果在 18.1%-28.1%之间。

读一流大学为什么会增加收入？

因果推断方法只能估计因果效应的大小，却无法回答为什么会产生因果效应的问题。

在本文中，作者讨论了在录取线附近没有入学的学生会做什么。

- 没有入学的学生如果选择工作（全职），其收入会明显高于在读的学生。然而作者并没发现这样的证据。
- 在本文中的一流公立大学所在州还有另外 7 所公立大学（学费水平和文中的大学基本相似），这给没有入学该一流大学的学生提供了替代机会。同时，有调查结果显示，美国 85% 的大学生都选择在自己出生的州读大学。

因此，作者认为没有入学一流大学的学生去了其他大学。这样的话，读一流大学带来的收入增加主要会通过两种途径实现：人力资本的形成（教育的效果）、信号传递（学生自身素质）。

- 8 所公立大学的人均经费投入非常接近，因此尤其引起的教育效果差异应该不大。
- 其他因素也会带来教育效果的差异，我们很难将它们和学生自身素质的差异区分开。

课外阅读

- Bertrand, M., Duflo, E., & Mullainathan, S. (2004).
  How Much Should We Trust Differences-in-Differences Estimates?
  The Quarterly Journal of Economics, 119:1, 249-275.
  https://www.jstor.org/stable/25098683

- Stephens-Davidowitz, S., Varian, H., & Smith, M. D. (2017).
  Super returns to Super Bowl ads?
  Quantitative Marketing and Economics, 15, 1-28.
  https://link.springer.com/article/10.1007/s11129-016-9179-0

- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001).
  The Colonial Origins of Comparative Development: An Empirical Investigation.
  American Economic Review, 91:5, 1369-1401.
  https://www.aeaweb.org/articles?id=10.1257/aer.91.5.1369

- Angrist, J. & Lavy, V. (1999).
  Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement.
  The Quarterly Journal of Economics, 114:2, 533-575.
  https://www.jstor.org/stable/2587016