微观计量经济学

阮 睿

中央财经大学
中国财政发展协同创新中心

April 10, 2025

Contents

1 断点回归 (Regression Discontinuity)

- Key reference: Lee and Lemieux (2010, JEL), "Regression Discontinuity Designs in Economics."

1.1 断点回归的工作原理

- Before-after comparison：假定 D 是突然发生的，并且 D 一发生，Y 就发生变化。因此 Y 的变化的原因就可以归结为 D 的变化。

- Before-after comparison 是一种特殊的断点回归：把这种时间上的突变扩展到其他变量上。

$$\begin{aligned}\hat{\tau} = \frac{1}{N}\sum_i (Y_{i1} - Y_{i0}) \to_p & E(Y_1 - Y_0) \\ = & E(Y_1^1 - Y_0^0) \\ = & \underbrace{E(Y_1^1 - Y_1^0)}_{\text{treatment effect on the post-treated}} + E(Y_1^0 - Y_0^0)\end{aligned}$$

  - 第二行：处理组个体在处理之后的 y - 控制组个体在处理之前的 y。（这里所有的个体都被处理了，可以想象有一个控制组。）
  - 第三行：Y_1^0 控制组在事后（处理组被处理，而控制组没被处理）的 Y。
  - 定义 $D_t = \mathbf{1}(t \geq 1)$. 当 D_t 发生跳跃时，$E(Y_t)$ 随之发生跳跃。
  - 识别假设：$E(Y_1^0 - Y_0^0) = 0$，应读作 $E(Y_t^0)$ 在 $t = 1$ 处平滑（连续），时间间隔越短，该假设越可能成立。这个式子的含义是：控制组在事后的 Y 减去控制组在事前的 Y。只有当间隔时间非常短的时候，它俩才能相等。
  - 即，Before-after comparison 是拿处理组在被处理之前的数据，当做自己在事后“如果没被处理”的反事实。
  - 在宏观因果识别上有应用。

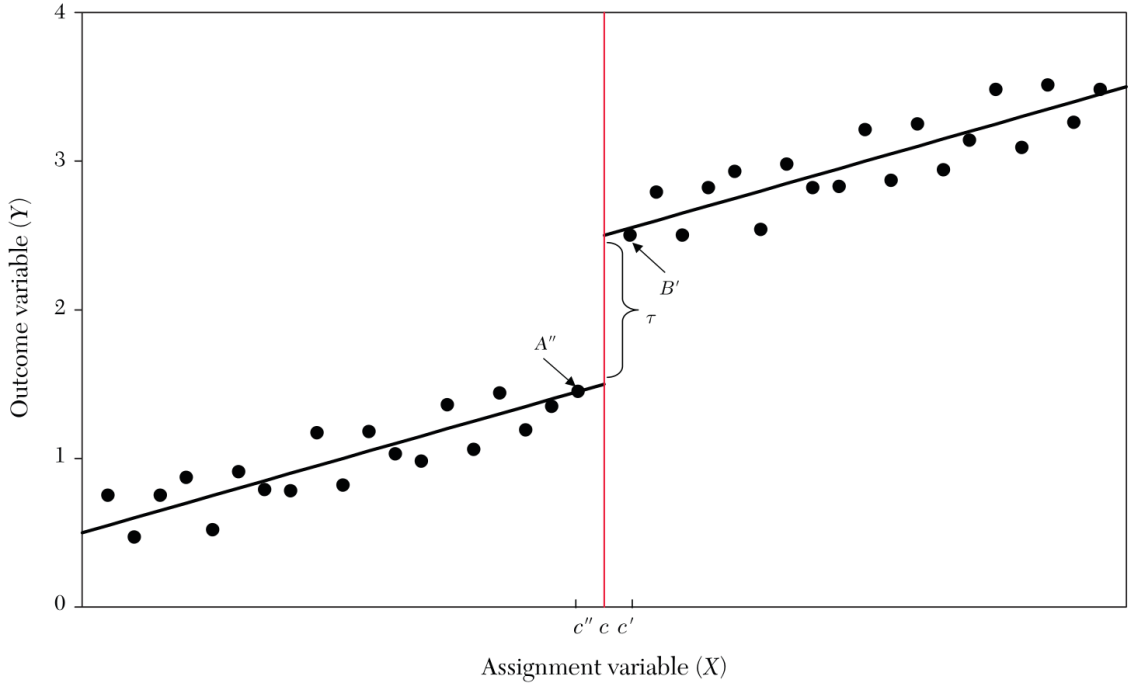
∗Nakamura, E., & Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: the information effect. The Quarterly Journal of Economics, 133(3), 1283-1330.

∗

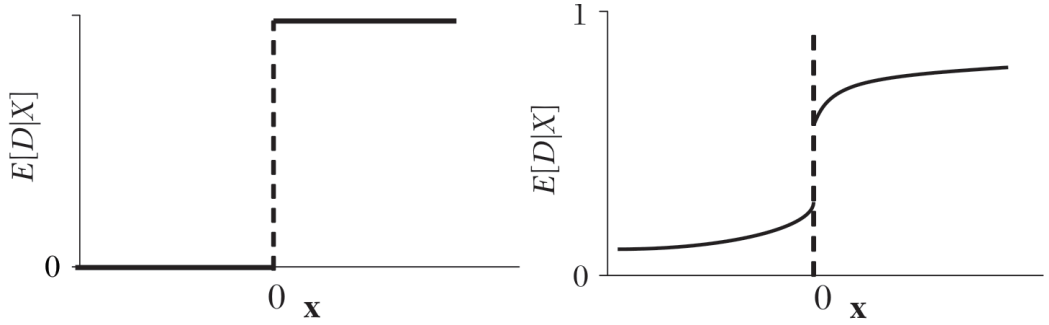
- 推而广之，把处理在时间上的突然变化推广到其他变量引起的处理上，例如，如果是否进入处理组取决于某种“一刀切”标准。

$$D_i = \begin{cases} 0 & X < c \\ 1 & X \geq c \end{cases}$$

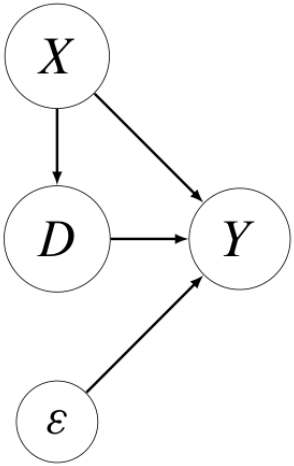
则标准左右两侧 outcome 的差异即可看作 treatment effect.

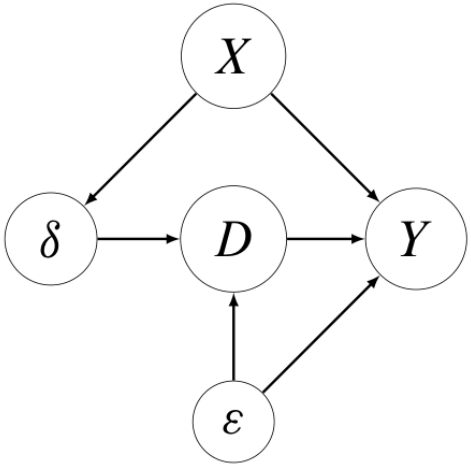

4
3
2
1
0
Outcome variable (Y)
A''
B'
τ
c'' c c'
Assignment variable (X)

- 例子：高考分数线、贫困县、污染警戒线、学区房、退休年龄、赢得选举的票数、出生季度等等。
- 正式地，令 D_i 为 probability of receiving treatment, 如果 $E(D|X) = P(D = 1|X)$ 存在断点，D 影响 Y，则 $E(Y|X)$ 也存在断点。X 被称为 assignment/forcing/running variable.
- 定义 rule of assignment to treatment: $\delta_i = \mathbf{1}(X_i \geq c)$
  - 若 $D = \delta$，则 $E(D|X) = D$，称为 sharp RD.
  - 若 $D \neq \delta$，则 $E(D|X) \neq D$，称为 fuzzy RD.



- 直观上，$E(Y|X)$ 的跳跃与 $E(D|X)$ 的跳跃的比值即为 treatment effect.





- 由图可知，要研究 D 对 Y 的影响，应该控制 X，但要比较相同 X 下的实验组和对照组很困难，因为（几乎）没有 overlap, 因此只能研究 $X \simeq c$ 的那部分样本，这相当于在 before-after comparison 中，$t = 1$ 时所有个体都进入实验组，所以只能以自身的历史作为对照，但这一对照只有当时间间隔很短时才有效。

●RD 的识别。考察如下的结构模型

$$\begin{aligned} Y_i &= \tau D_i + \text{error}_i \\ &= \tau D_i + \underbrace{E(\text{error}|X = X_i)}_{\equiv m(X_i)} + \underbrace{\text{error}_i - E(\text{error}_i|X = X_i)}_{\equiv U_i} \end{aligned}$$

$$E(Y|X) = \tau E(D|X) + m(X)$$

$$\begin{aligned} \lim_{X\downarrow c} E(Y|X) &= \tau \lim_{X\downarrow c} E(D|X) + \lim_{X\downarrow c} m(X) \\ \lim_{X\uparrow c} E(Y|X) &= \tau \lim_{X\uparrow c} E(D|X) + \lim_{X\uparrow c} m(X) \end{aligned}$$

如果

1.$E(D|X)$ 在 $X = c$ 处存在断点：$\lim_{X\downarrow c} E(D|X) \neq \lim_{X\uparrow c} E(D|X)$

2.$m(x)$ 在 $X = c$ 处连续：$\lim_{X\downarrow c} m(X) = \lim_{X\uparrow c} m(X)$.

则

$$\tau^{RD} = \frac{\lim_{X\downarrow c} E(Y|X) - \lim_{X\uparrow c} E(Y|X)}{\lim_{X\downarrow c} E(D|X) - \lim_{X\uparrow c} E(D|X)} \tag{1}$$

●如何从 potential outcomes framework 的角度 justify 这一结构模型？以 sharp RD 为例。

$$\lim_{X\downarrow c} E(D|X) - \lim_{X\uparrow c} E(D|X) = 1$$

$$\begin{aligned}
\tau =& \lim_{X\downarrow c} E(Y^1|X) - \lim_{X\uparrow c} E(Y^0|X) \\
=& \lim_{X\downarrow c} E(Y^1|X) - \lim_{X\downarrow c} E(Y^0|X) \\
& (\text{如果} E(Y^0|X) \text{在} X = c \text{处连续}) \\
=& \underbrace{\lim_{X\downarrow c} E(Y^1 - Y^0|X)}_{\text{treatment effect on the just treated}}
\end{aligned}$$

$$
\begin{aligned}
Y =&(1-D)Y^0 + DY^1 \\
=&(Y^1 - Y^0)D + Y^0 \\
=&\left(\underbrace{E(Y^1 - Y^0|X)}_{\equiv\tau} + \underbrace{Y^1 - Y^0 - E(Y^1 - Y^0|X)}_{\equiv\nu} \right) D + Y^0
\end{aligned}
$$

$$
\begin{aligned}
E(Y|X) =&\tau D + E(\nu D + Y^0|X) \\
=&\tau D + DE(\nu|X) + E(Y^0|X) \\
=&\tau D + E(Y^0|X)
\end{aligned}
$$

所以 $m(X)$ 在 $X = c$ 处连续就是指 $E(Y^0|X)$ 在 $X = c$ 处连续。

●1式还可写作

$$\tau = \frac{E(Y|\delta = 1) - E(Y|\delta = 0)}{E(D|\delta = 1) - E(D|\delta = 0)}$$

在 fuzzy RD 中，此即为 τ 的 Wald 估计量，δ 自动成为 D 的 IV，满足 IV 所需要的三个条件：

–Exclusion: 无法检验，但应该成立。
–Independence: 自动成立。
–Relevance: 可以检验。

关于 IV 估计的所有讨论此处都适用。

- RD 的三个特点：
  - Local randomization.
  - 在 sharp RD 中，D 无内生性之虞；在 fuzzy RD 中，IV 唾手可得。
  - 所有在 $X = c$ 处连续的变量都无需控制。
- 可以把 RD 看作一种特殊的 selection-on-observables 模型。
  - unconfoundedness 条件轻松满足
  - overlap 条件绝不满足，但代之以连续性条件 (of potential outcomes conditional on $X = c$)
- 因此不是把 X 相同的处理组个体和控制组个体进行比较，而是把充分接近 $X = c$ 两侧的处理组个体和控制组个体进行比较。

- Substantializing 识别假设
  - treatment（至少部分地）由可观测变量 X 决定，而不是相反：treatment 不能影响 X，也就是说，X 是 pre-treatment 变量，或者是无法改变的变量。
  - treatment 虽然不是随机的，个体可以通过影响 X 来影响 treatment assignment, 但这种影响不能是精确的——cutoff 附近的个体无法操纵是否落在 cutoff 的哪一侧，也就是说 c 是独立于 X 的（外生的），δ 完全取决于 X 和 c. 只有这样，cutoff 附近的个体才能被看作是 exchangeable or otherwise identical.
  - 除了 treatment 之外，其它变量都是 X 的平滑函数，即不存在其它方式使得邻近 cutoff 两侧的个体出现差异，只有这样才能把 outcome 的跳跃全部归因于 treatment 的跳跃。而且这是可以检验的！

- 强内部有效性，弱外部有效性？不信抬头看，老天放过谁！"The RD estimand can be interpreted as a weighted average treatment effect [across all individuals] where the weights are the relative ex ante probability that the value of an individual's assignment variable will be in the neighborhood of the threshold."

$$
\begin{aligned}
&\lim_{X \downarrow c} E(Y|X=c) - \lim_{X \uparrow c} E(Y|X=c) \\
=&\sum_{u_i} \tau(u_i) \Pr(u_i|X=c) \\
=&\sum_{u_i} \tau(u_i) \frac{f(X=c|u_i)}{f(X=c)} \times \Pr(u_i)
\end{aligned}
$$

- 第一行：定义了清晰 RD (Sharp RD) 的估计量，即结果变量 Y 的条件期望在临界点 c 两侧的极限之差。

- 第二行：利用迭代期望定律，将上述差值表示为个体（或类型 u_i）的处理效应 $\tau(u_i)$ 的加权平均，权重是在临界点 $X = c$ 处个体类型为 u_i 的概率 $\Pr(u_i|X = c)$。这对应了传统上认为 RD 效应只适用于临界点处人群的观点。

- 第三行：用贝叶斯定理 (Bayes' Rule) 将条件概率 $\Pr(u_i|X = c)$ 展开。

$$
\Pr(u_i|X=c) = \frac{f(X=c|u_i) \times \Pr(u_i)}{f(X=c)}
$$

其中，$f(X = c|u_i)$ 是给定个体类型为 u_i 时，其分配变量 X 取值为 c 的条件概率密度；$\Pr(u_i)$ 是个体类型为 u_i 的无条件概率；$f(X = c)$ 是分配变量 X 取值为 c 的边际概率密度。

这个最终表达式显示 RD 估计量是所有个体处理效应 $\tau(u_i)$（按其在总体中的比例 $\Pr(u_i)$ 加权）的再次加权平均。

1.2 断点回归的估计方法

因为不可能真的只使用 $X = c$ 的样本，用于估计的样本中包含距离 c 多远的观测值，面临估计偏误和估计效率的 trade-off, 因此尝试多种估计方法保证结果的稳健性尤为重要。

1.2.1 OLS/2SLS

•OLS for sharp RD

$$Y = \tau D + m(X) + U$$

$$m(X) = \gamma_0 + \gamma_1^+ \delta(X-c) + \gamma_1^- (1-\delta)(X-c) + \gamma_2^+ \delta(X-c)^2 + \gamma_2^- (1-\delta)(X-c)^2$$

•2SLS for fuzzy RD

$$\begin{aligned} Y =& \tau D + m(X) + U \\ D =& \alpha\delta + m(X) + \varepsilon \end{aligned}$$

•注意事项

- 尝试使用全部样本和不同 c 邻域的子样本。
- 尝试 $m(X)$ 的线性形式和二次形式，但不要更高阶了！因为高阶多项式赋予远离 cutoff 的观测值过高权重 (Gelman and Imbens, 2019, Journal of Business & Economic Statistics).

1.2.2 Kernel regression

•Local constant regression.

$$\tau = \frac{E(Y|\delta=1) - E(Y|\delta=0)}{E(D|\delta=1) - E(D|\delta=0)}$$

$$\hat{E}(Y|\delta=1) = \frac{\sum_i K\left(\frac{X_i - c}{h}\right)\delta_i Y_i}{\sum_i K\left(\frac{X_i - c}{h}\right)\delta_i}$$

$$\hat{E}(Y|\delta=0) = \frac{\sum_i K\left(\frac{X_i - c}{h}\right)(1-\delta_i) Y_i}{\sum_i K\left(\frac{X_i - c}{h}\right)(1-\delta_i)}$$

分母类似定义。

•Choices of kernel function.

–Uniform (rectangular) kernel:

$$K(u) = \frac{1}{2}\mathbf{1}(|u| < 1)$$

–Triangular kernel:

$$K(u) = (1 - |u|)\mathbf{1}(|u| < 1)$$

–Gaussian kernel:

$$K(u) = \phi(u)$$

–Epanechnikov kernel:

$$K(u) = 0.75(1 - u^2)\mathbf{1}(|u| < 1)$$

若使用 uniform kernel, 则

$$\hat{E}(Y|\delta=1) = \frac{\sum_i \mathbf{1}(c < X_i < c+h) Y_i}{\sum_i \mathbf{1}(c < X_i < c+h)}$$
$$\hat{E}(Y|\delta=0) = \frac{\sum_i \mathbf{1}(c-h < X_i < c) Y_i}{\sum_i \mathbf{1}(c-h < X_i < c)}$$

此时 $\hat{\tau}$ 等价于

$$Y = \tau D + \gamma_0 + U$$

在 $c-h < X_i < c+h$ 子样本中的 OLS/2SLS，因此被称作 local constant regression estimator.

1.3 估计方法：局部多项式回归

虽然全局多项式回归（在整个数据范围内拟合高阶多项式）有时也被使用，但目前 RD 分析中最常用且推荐的方法是**局部多项式回归 (Local Polynomial Regression)**，特别是**局部线性回归 (Local Linear Regression, LLR)**。

局部线性回归的优势：

- **良好的边界性质：** 相较于局部常数回归（即核回归或简单的分箱均值），LLR 在边界点（即临界点 c）处的偏差更小。
- **直观性：** LLR 本质上是在临界点 c 两侧的一个局部窗口（由带宽 h 定义）内分别拟合线性回归模型。
- **偏差-方差权衡：** LLR 在偏差和方差之间提供了较好的平衡。

实施 LLR：

1. **选择带宽 h：**这是最关键的步骤之一（详见下文）。带宽决定了用于估计临界点处函数值的局部窗口大小。

2. **估计：**

   - **方法一（分别估计）：**

     –对 $c-h \le X_i < c$ 的样本，运行 Y_i 对 $(X_i - c)$ 的 OLS 回归，得到左侧截距 $\hat{\alpha}_L$。
     –对 $c \le X_i \le c+h$ 的样本，运行 Y_i 对 $(X_i - c)$ 的 OLS 回归，得到右侧截距 $\hat{\alpha}_R$。
     –RD 效应估计为 $\hat{\tau} = \hat{\alpha}_R - \hat{\alpha}_L$。

   - **方法二（合并估计）：**运行一个合并样本 $(|X_i - c| \le h)$ 的 OLS 回归：

     $$Y_i = \beta_0 + \tau D_i + \beta_1 (X_i - c) + \beta_2 D_i (X_i - c) + \epsilon_i$$

     其中 $D_i = \mathbf{1}(X_i \ge c)$ 是处理虚拟变量。τ 的估计值 $\hat{\tau}$ 就是 RD 效应。这个方法可以直接得到 τ 的标准误。

3. **核函数 (Kernel)：**上述描述对应于矩形核 (Rectangular Kernel)。也可以使用其他核函数（如三角核 Triangular Kernel, Epanechnikov Kernel），它们会对靠近临界点的观测值赋予更大权重。实践中，核函数的选择通常对结果影响不大，矩形核因其简单直观而被广泛使用。

4. **多项式阶数：**虽然最常用的是局部线性 (p=1)，但也可以使用局部常数 (p=0) 或局部二次 (p=2) 等。局部线性通常是首选。

●Local linear regression. 考察

$$\min \sum_i (Y_i - \rho_0^+ - \rho_1^+ (X_i - c))^2 K\left(\frac{X_i - c}{h}\right) \delta_i$$

$$\min \sum_i (Y_i - \rho_0^- - \rho_1^- (X_i - c))^2 K\left(\frac{X_i - c}{h}\right) (1 - \delta_i)$$

$$\min \sum_i (Y_i - \kappa_0^+ - \kappa_1^+ (X_i - c))^2 K\left(\frac{X_i - c}{h}\right) \delta_i$$

$$\min \sum_i (Y_i - \kappa_0^- - \kappa_1^- (X_i - c))^2 K\left(\frac{X_i - c}{h}\right) (1 - \delta_i)$$

（注意如果令 $\rho_1^+ = \rho_1^- = \kappa_1^+ = \kappa_1^- = 0$, 即为 local constant regression. 当 h 并不严格等于零时，会造成 finite sample bias.）

$$\hat{\tau} = \frac{\hat{\rho}_0^+ - \hat{\rho}_0^-}{\hat{\kappa}_0^+ - \hat{\kappa}_0^-}$$

若使用 uniform kernel，$\hat{\tau}$ 等价于

$$Y = \tau D + \gamma_0 + \gamma_1^+ \delta(X - c) + \gamma_1^- (1 - \delta)(X - c) + U$$

在 $c - h < X_i < c + h$ 子样本中的 OLS/2SLS，因此被称作 local linear regression estimator.
类似地，还有 local polynomial regression.

●注意事项：

–$K(\cdot)$ 的选择不重要。因此参数估计和非参数估计的区别没有想象的那么大。

–Kernel regression 无非就是加权 OLS/2SLS，定义权重以后即可轻松实现。例如，对于 triangular kernel,

```
replace x=x-c
gen d=(x>=0)
gen xd=x*d
gen u=.
gen bw_l = 2
gen bw_r = 2
replace u=x/bw_l if d==0
replace u=x/bw_r if d==1
gen kweight=1-abs(u)
replace kweight=. if x<-bw_l
replace kweight=. if x>bw_r

reg y d x xd [pweight=kweight], robust
```

1.4 断点回归的操作细节

- 作散点图看断点。
  - 将 X 划分成若干个 bin（保证 cutoff 不要落入某个 bin 内部），计算每个 bin 中 Y 的均值，作出该均值与 bin 的中点的散点图。
  - 在 fuzzy RD 中，对 D 也进行类似操作。
  - 同时，还要计算每个 bin 中的观测的个数
  - STATA 实现：binscatter; cmogram
  - cmogram: Draw histogram-style conditional mean or median graph

- More formally, for some bandwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, the idea is to construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1)\, h.$$

The average value of the outcome variable in the bin is

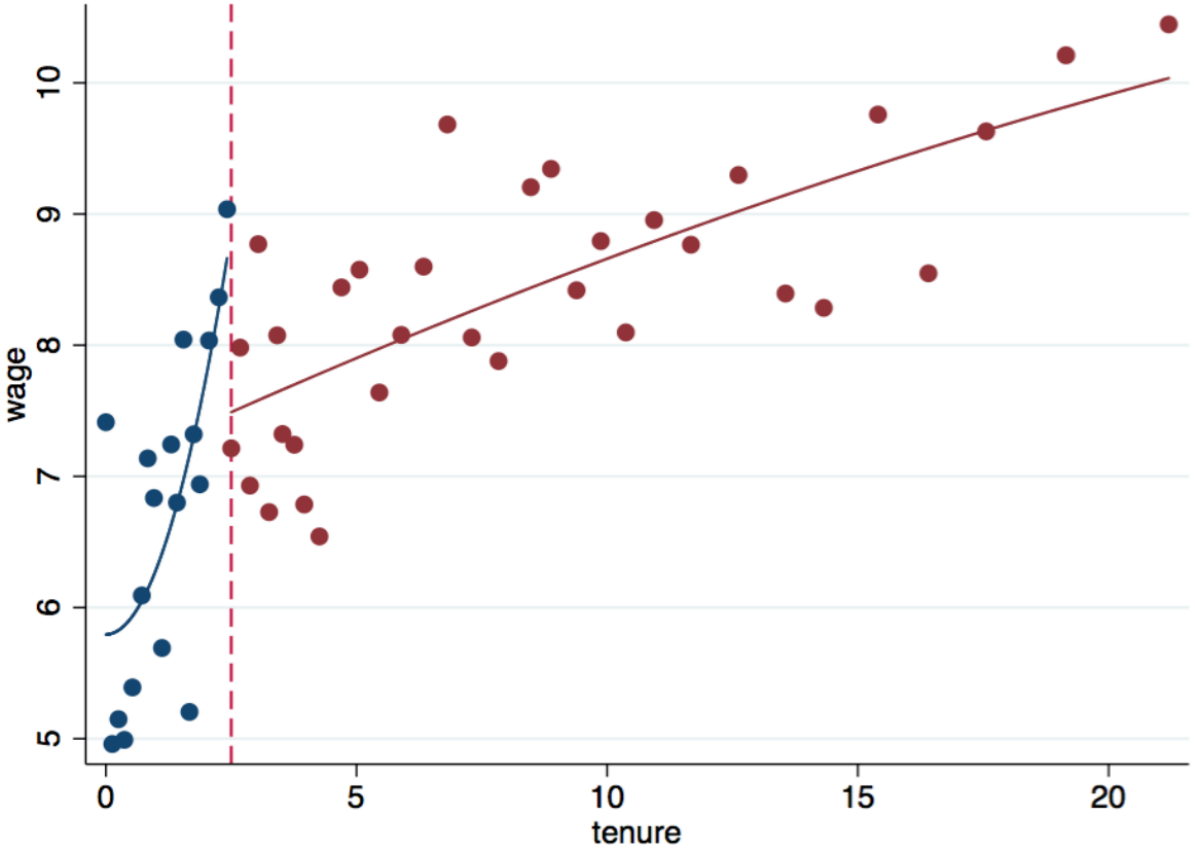
$$\bar{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i 1 \left\{ b_k < X_i \leq b_{k+1} \right\}.$$

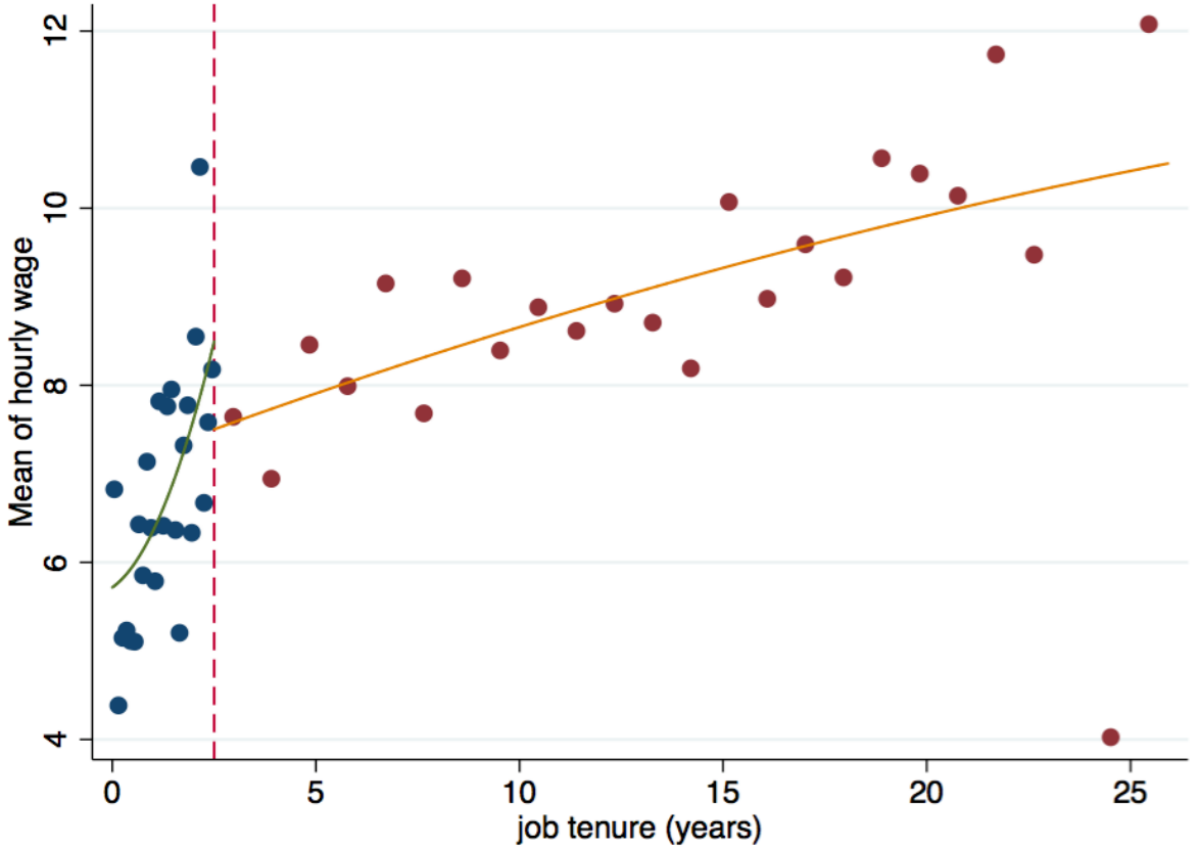
It is also useful to calculate the number of observations in each bin

$$N_k = \sum_{i=1}^{N} 1 \left\{ b_k < X_i \leq b_{k+1} \right\}$$

to detect a possible discontinuity in the assignment variable at the threshold, which would suggest manipulation.

- Advantages in graphing the data this way before starting to run regressions to estimate the treatment effect:
  - The graph provides a simple way of visualizing what the functional form of the regression function looks like on either side of the cutoff point.
  - comparing the mean outcomes just to the left and right of the cutoff point provides an indication of the magnitude of the jump in the regression function at this point, i.e., of the treatment effect.
  - the graph also shows whether there are unexpected comparable jumps at other points.


wage
5
6
7
8
9
10
0
5
10
15
20
tenure


Mean of hourly wage
4
6
8
10
12
0
5
10
15
20
25
job tenure (years)

●如何选择 bandwidth

–主观选择很重要。

–Imbens and Kalyanaraman (2012, RES) 最优 bandwidth.

–Cross-validation: "leave-one-out" estimator for $E(Y|X = X_i)$

$$\hat{E}_{-i}(Y|X = X_i > c) = \frac{\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)\mathbf{1}(X_j > X_i)Y_j}{\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)\mathbf{1}(X_j > X_i)}$$

$$\hat{E}_{-i}(Y|X = X_i < c) = \frac{\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)\mathbf{1}(X_j < X_i)Y_j}{\sum_{j\neq i} K\left(\frac{X_j - X_i}{h}\right)\mathbf{1}(X_j < X_i)}$$

$$h_{CV}^{opt} = \arg\min_h CV(h) \equiv \sum_i (Y_i - \hat{E}_{-i}(Y|X = X_i))^2$$

–用除了 i 以外的个体的 y 的加权平均值作为 y_i 的估计值。随着带宽 h 增加，每个个体 i 的估计差异 $Y_i - \hat{Y}_i$ 会增大，但是个体 i 的个数在增加，所以能够找到一个 h，使得 $CV(h)$ 最小。

–通过以上两种方法计算 $CV(h)$，都可以。但是对于 Fuzzy RD，这两种计算方法有差异[1]。

[1]详见 Matching, Regression Discontinuity, Difference in Differences, and Beyond 的 Regression Discontinuity 章，4.3.3 节

●STATA 实现：

–如前所述，手动做很方便，更可控。
–rd
–rdplot: Regression-discontinuity plots
–rdob: sharp and fuzzy RD by local linear regression with IK optimal bandwidth.
–rdcv: sharp RD with CV bandwidth selection or IK optimal bandwidth

1.4.1 相关检验

- 检验 Y 和 D 在 $X = c$ 处的断点，相当于估计 first-stage equation 和 reduced-form equation.
- 检验 Y 和 D 在 $X \neq c$ 处的断点。
- **检验 $m(X)$ 的连续性**。$m(X)$ 同时包含可观测变量和不可观测变量关于 X 的条件期望。
  - 关于可观测变量 W，检验 $E(W|X)$ 在 $X = c$ 处的连续性，若存在断点，则将 W 作为 regressor 加入 Y 方程。
  - [2]关于不可观测变量 U，$E(U|X)$ 在 $X = c$ 处的连续性无法直接检验，但可以检验其必要条件，$f_X(x)$ 在 $X = c$ 处的连续性。直观地，看 X 是不是被操纵。

$$E(U|X = x) = \int u f_{U|X}(u|x)du = \int u \frac{f_{X|U}(x|u)f_U(u)}{f_X(x)}du$$

[2]详见 Matching, Regression Discontinuity, Difference in Differences, and Beyond 的 Regression Discontinuity 章，4.4.2 节

McCrary (2008, JoE) density test.

*可以先画出 X 的直方图，观察在 $X = c$ 处是否存在跳跃，然后进行正式检验。

***原理：**如果个体能够精确地操纵 (manipulate) 其分配变量 X_i 以便落在临界点的某特定一侧（例如，为了获得处理资格），那么在临界点 c 处，分配变量 X_i 的概率密度函数 $f(x)$ 可能会出现不连续的跳跃或尖峰。而在“无法精确控制”或局部随机化的假设下，密度函数在 c 点应该是连续的。

*检验过程：

1.将 X 划分成 J 个 bin（将中点记作 B_j），计算落入各 bin 的频次（记作 S_j）。

$$\cdots, [c-2b, c-b), [c-b, c), [c, c+b), [c+b, c+2b), \cdots$$

2.使用 local linear regression 对直方图进行平滑化.

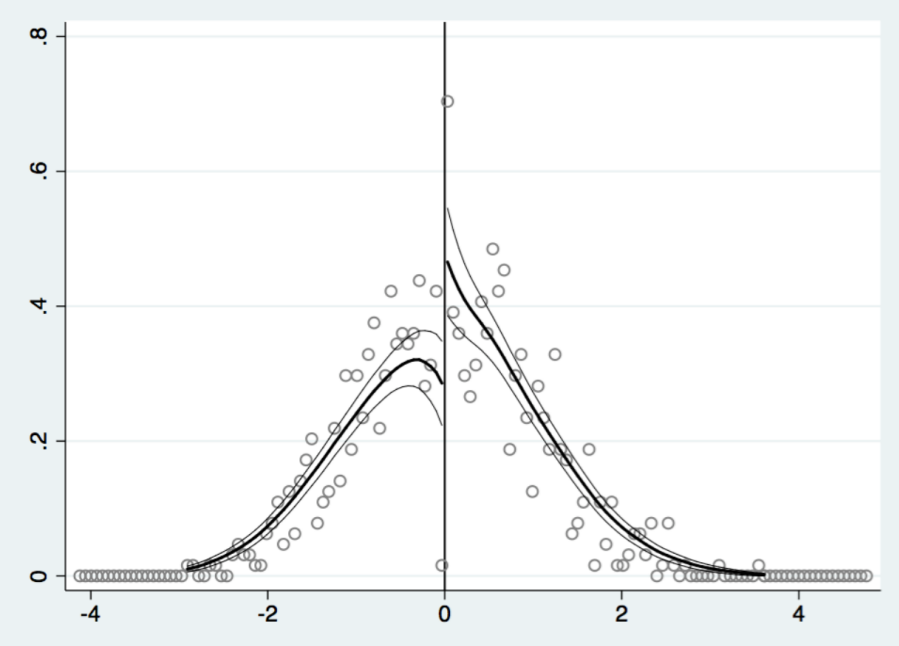
$$\min_{\phi_1,\phi_2} \sum_{j=1}^{J} (S_j - \phi_1 - \phi_2(B_j - x))^2 K\left(\frac{B_j - x}{h}\right) \cdot (\mathbf{1}(B_j > c)\mathbf{1}(x \geq c) + \mathbf{1}(B_j < c)\mathbf{1}(x < c))$$

任意 x 处的密度估计即为 $\hat{f}_X(x) = \hat{\phi}_1$

计算 c 附近的（对数）高度差

$$\phi = \ln \min_{x \downarrow c} f_X(x) - \ln \lim_{x \uparrow c} f_X(x)$$

ϕ 服从渐进正态分布。STATA 实现：DCdensity 报告 $\hat{\phi}$ 及其标准误，并画出 $\hat{f}_X(x)$.


.8
.6
.4
.2
0
-4
-2
0
2
4

敏感性分析

- **原理：** RD 估计结果（特别是跳跃的大小和显著性）不应过度依赖于模型设定的具体选择，如带宽大小或多项式阶数。

- **实施：**

  1.改变带宽：使用一系列不同的带宽值（例如，从较小到较大）重新估计 RD 模型，观察结果是否保持稳定。

  2.改变多项式阶数：使用不同阶数的多项式（例如，线性、二次、三次等）来拟合 $\mathbb{E}[Y|X=x]$ 函数，观察 RD 估计量是否稳健。

  3.包含/排除协变量：比较包含和不包含基线协变量时的 RD 估计结果。理论上，如果 RD 有效，包含协变量主要影响精度而非点估计本身。

- **意义：** 如果结果对模型设定非常敏感，可能意味着模型设定不当（例如，函数形式错误）或数据本身不支持清晰的断点效应。

1.5 RD + Matching

- 关键假设：
  - 均值独立
$$E(Y^d|X,W) = E(Y^d|W), d = 0, 1$$
  - Overlap
$$0 < P(\delta = 1|W) < 1$$
- 识别（以 ATT 为例）

$$\begin{aligned} & E(Y^1 - Y^0|\delta = 1) \\ =& E\{E(Y^1 - Y^0|W, \delta = 1)|\delta = 1\} \\ =& E\{E(Y|W, \delta = 1) - E(Y|W, \delta = 0)|\delta = 1\} \\ =& E(Y|\delta = 1) - E\{E(Y|W, \delta = 0)|\delta = 1\} \end{aligned}$$

- 条件于可观测变量 W，处理组和控制组之间是可比的。

- 检验均值独立假定：寻找 W

$$E\left(Y^1 \mid X, W, X \geq c\right) = E\left(Y^1 \mid W\right) = E\left(Y^1 \mid W, X \geq c\right)$$

因此

$$E(Y \mid X, W, \delta = 1) = E(Y \mid W, \delta = 1)$$

同理

$$E(Y \mid X, W, \delta = 0) = E(Y \mid W, \delta = 0)$$

可以在 cutoff 两侧，将 Y 对 X 和 W 回归，若 X 不显著，则该假定很可能成立。

- 可以估计任何形式的 treatment effect, 特别是远离 cutoff 处的 treatment effect. 以 ATT 为例，若使用 OLS 方法，则

$$Y = \beta_0 + \beta_1\delta + \beta_2 W + \beta_3\delta \cdot W + \varepsilon$$
$$\tau_1 = \beta_1 + \beta_3 E(W \mid \delta = 1)$$

也可以使用其他匹配方法。

示例应试录取型学校能否提高学习成绩？(Abdulkadiroğlu et al., 2014, ECMA; Angrist and Rokkanen, 2015, JASA)

- X: 按统一入学考试成绩和学分绩计算的标准分
- Y :10 年级数学和英语成绩
- 三所 exam schools (7-12 年级), 按录取分数线高低依次为 BLS, BLA 和 O'Bryant.
- 两个录取批次，7 年级和 9 年级。
- W : 除人口统计学变量以外，对于 7 年级批次，还能观测到 4 年级的数学和英语成绩，对于 9 年级批次，还能额外观测到 8 年级的数学和 7 年级的英语成绩。
- RD around the cutoff: local linear regression, triangular kernel

1.6 Regression Kink

•Key references: Card et al. (2012, NBER WP; 2015, ECMA)

$$Y = \tau D + m(X) + U$$

•D 在 $X = c$ 处连续但左右斜率不同。

$$\lim_{X \downarrow c} \frac{\partial D(X)}{\partial X} \neq \lim_{X \uparrow c} \frac{\partial D(X)}{\partial X}$$

•$m(X)$ 在 $X = c$ 处连续可微，即 $m'(X)$ 在 $X = c$ 处连续。
（注意，在 RD 中，通常 $\lim_{X \downarrow c} m'(X) \neq \lim_{X \uparrow c} m'(X)$.）

$$\lim_{X \downarrow c} \frac{\partial m(X)}{\partial X} = \lim_{X \uparrow c} \frac{\partial m(X)}{\partial X}$$

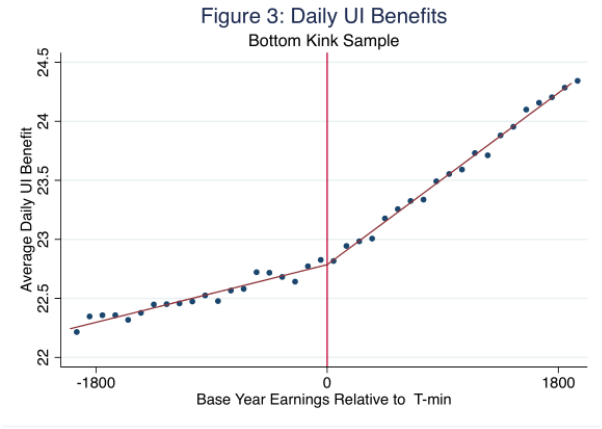


Figure 3: Daily UI Benefits
Bottom Kink Sample
Average Daily UI Benefit
22
22.5
23
23.5
24
24.5
-1800
0
1800
Base Year Earnings Relative to T-min

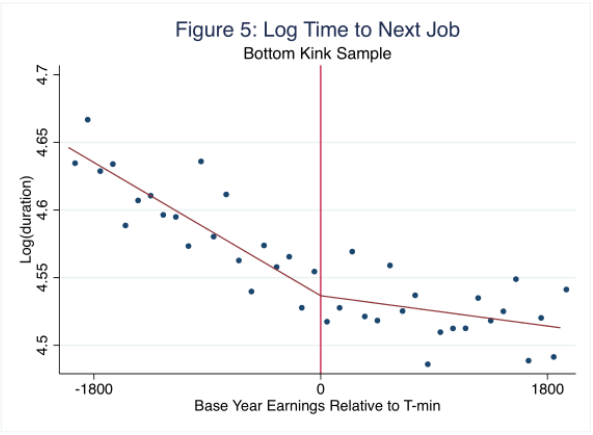


Figure 5: Log Time to Next Job
Bottom Kink Sample
Log(duration)
4.5
4.55
4.6
4.65
4.7
-1800
0
1800
Base Year Earnings Relative to T-min

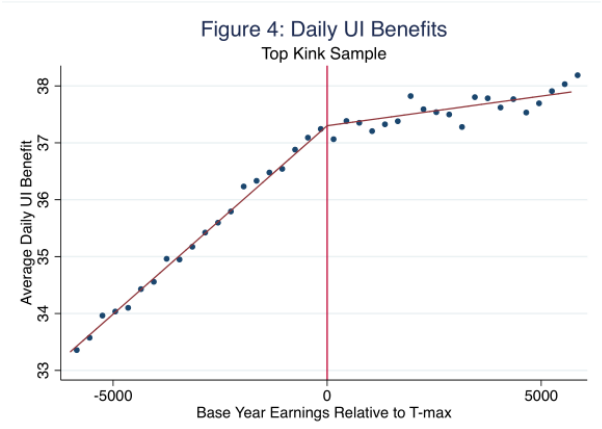


Figure 4: Daily UI Benefits
Top Kink Sample
Average Daily UI Benefit
33
34
35
36
37
38
-5000
0
5000
Base Year Earnings Relative to T-max

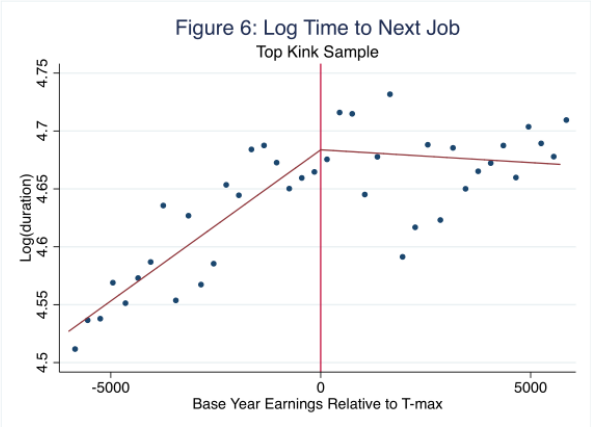


Figure 6: Log Time to Next Job
Top Kink Sample
Log(duration)
4.5
4.55
4.6
4.65
4.7
4.75
-5000
0
5000
Base Year Earnings Relative to T-max

●RK 的识别

$$\lim_{X\downarrow c}\frac{\partial E(Y|X)}{\partial X}=\lim_{X\downarrow c}\left(\tau\frac{\partial D(X)}{\partial X}+\frac{\partial m(X)}{\partial X}\right)$$
$$\lim_{X\uparrow c}\frac{\partial E(Y|X)}{\partial X}=\lim_{X\uparrow c}\left(\tau\frac{\partial D(X)}{\partial X}+\frac{\partial m(X)}{\partial X}\right)$$

$$\tau^{RK}=\frac{\lim_{X\downarrow c}\frac{\partial E(Y|X)}{\partial X}-\lim_{X\uparrow c}\frac{\partial E(Y|X)}{\partial X}}{\lim_{X\downarrow c}\frac{\partial E(D|X)}{\partial X}-\lim_{X\uparrow c}\frac{\partial E(D|X)}{\partial X}}$$

●Card et al. (2012, WP) 的建议：

–Fuzzy RK.

$$Y = \tau D + \gamma_0 + \gamma_1(X - c) + U$$

用 $\delta(X - c)$ 作为 D 的 IV，估计在 $c - h < X_1 < c - h$ 子样本中的 2SLS。

–Sharp RK，分子是如下 reduced-form 回归

$$Y = \gamma_0 + \gamma_1(X - c) + \tau\delta(X - c) + U'$$

的 OLS 估计量 $\hat{\tau}$, 分母是可以直接计算的常数。

–由于 RK 对于 $m(X)$ 的非线性形式极为敏感，Ganong and Jager (2015, WP) 建议对没有 policy kink 的地区进行同样的估计，然后根据 placebo effects 的经验分布进行统计推断。

–例子：Nielson et al. (2010, AEJ: Policy); Rothstein and Rouse (2011, JPubE); Dahlberg et al. (2008, JPubE); Simonsen et al. (2016, JAE).

–最后的评论：很多 IV 文章都可以当成 RD/RK 来理解，反之亦然。

RD 案例

美国的青少年禁酒法案。美国人合法饮酒年龄（the minimum legal drinking age, MLDA）是 21 岁.

这个图：X 轴是死亡日期相对于生日的天数。例如，一个人出生在 1990 年 9 月 18 日，死亡日期是 2012 年 9 月 19 日，那么他的相对日期为 +1. Y 轴是 10 万人死亡人数。

- 只有在 21 岁死亡组里，相对日期为 1 的人数有一个明显的跳跃。
- 这应该不是一般的开 party 玩 high 了。否则 22 岁和 20 岁都应该有这样的跳跃。

数学上表示最佳饮酒年龄的 treatment 为：

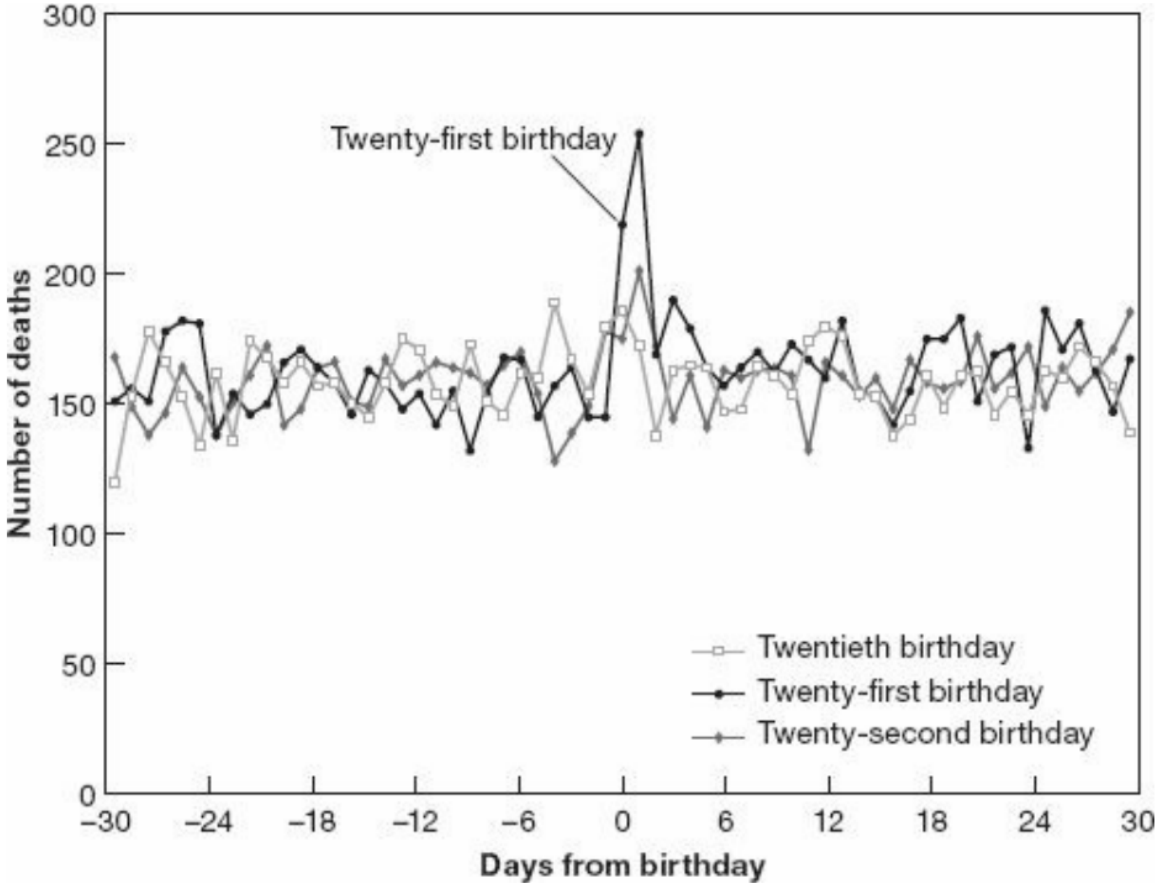
$$D_a = \begin{cases} 1 & \text{if} a \geq 21 \\ 0 & \text{if} a < 21 \end{cases}$$

$$\bar{M}_a = \alpha + \rho D_a + \gamma a + e_a$$

M_a 是 a 月的死亡率。D_a 是根据月份的 treatment dummy。a 代表线性趋势。ρ 是 21 岁的死亡率跳跃。

为什么这个模型可以没有 observables？没有可观察变量是 inside information 的 payoff. 尽管 treatment 不是随机分配的，但是知道 treatment 的来源，即完全来自 running variable。因此，因果识别的问题转变为 running variable 和 outcomes 之间的关系是否可以被包含年龄的回归所决定。

尽管 RD 使用回归方法估计因果效应，但 RD 和那些带有控制变量的回归方法以及匹配方法不同。匹配和回归方法是基于处理组和控制组的比较（条件于协变量），而 RD 的有效性依赖于对 running variable 的


300
250
200
150
100
50
0
Number of deaths
Twenty-first birthday
Twentieth birthday
Twenty-first birthday
Twenty-second birthday
−30
−24
−18
−12
−6
0
6
12
18
24
30
Days from birthday

推算。

RD Specifies

RD 并不保证产生可靠的因果估计。在 Panel A, running variable (X) 和 outcome (Y) 之间的关系是线性的。有一个清晰的跳跃。而在 Panel B, X 和 Y 虽然有清晰的跳跃，但是之间的关系并不是线性的。而 Panel C 是最常遇见的情形，有一个清晰的转变，但是连续的。如果使用线性模型，就会出现偏误。

有两种方法可以减少 RD 的偏误。第一种方法是直接使用非线性模型。第二种方法是只关注 cutoff 周围的观测。

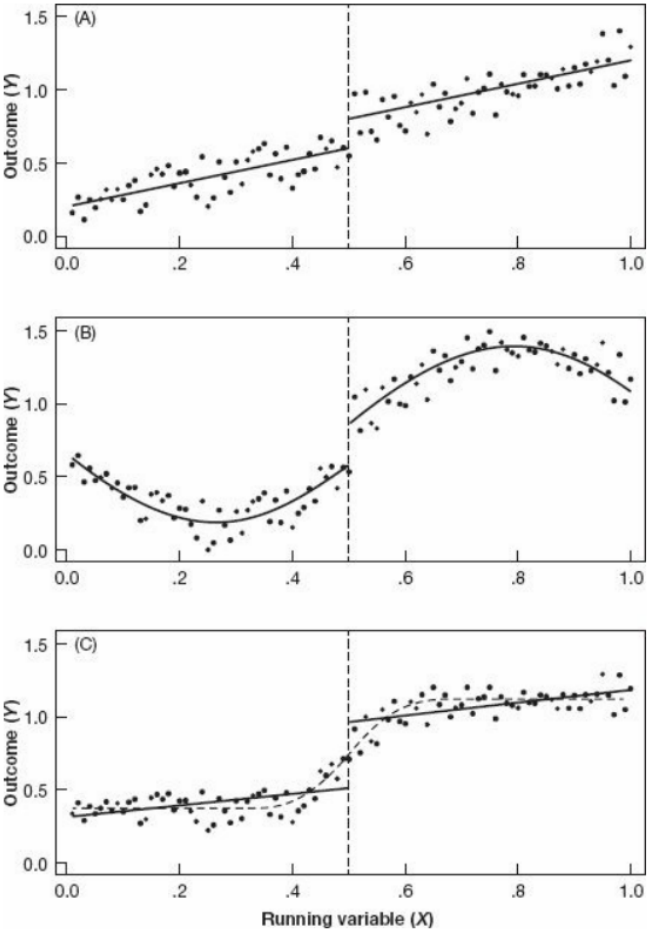
$$\bar{M}_a = \alpha + \rho D_a + \gamma_1 a + \gamma_2 a^2 + e_a$$

$$\bar{M}_a = \alpha + \rho D_a + \gamma_1 (a - a_0) + \delta[(a - a_0) D_a] + e_a$$

$$\begin{aligned} \bar{M}_a = \alpha + \rho D_a + \gamma_1 (a - a_0) + \gamma_2 (a - a_0)^2 \\ + \delta_1[(a - a_0) D_a] + \delta_2[(a - a_0)^2 D_a] + e_a \end{aligned}$$

哪一种模型更好，必须要通过观察数据进行判断。

RD in action, three ways



1.7 把 RD 视为一种研究设计，而非研究方法

1.8 论文案例分析

THE PERSISTENT EFFECTS OF PERU’S MINING *MITA*

秘鲁和玻利维亚在 1573 年至 1812 年间实行的“Mita”制度的长期影响的研究。

MITA Mita 是一种由西班牙政府强制征召的矿业劳动制度。Mita 制度始于 1573 年，由西班牙殖民政府设立，目的是为波托西（Potosi）的银矿和万卡韦利卡（Huancavelica）的汞矿提供所需的劳动力。这些矿产对于西班牙帝国的财富积累至关重要。

- 根据 Mita 制度，超过 200 个土著社区被迫贡献其成年男性人口的七分之一去矿山工作。这些被征召的劳工被称为″mitayos″。
- Mita 劳动是一种轮换制度，意味着每个社区的男性居民大约每七年需要服役一次。这种轮换旨在减少对任何单一社区的长期影响。
- Mita 制度的实施基于地理分配，特定区域内的社区需要提供劳动力，而其他区域则被免除。这种分配导致了明显的社会经济差异。
- 随着西班牙殖民统治的衰落和独立运动的兴起，Mita 制度最终在 1812 年被废除。然而，其长期影响在受影响地区仍然持续存在。

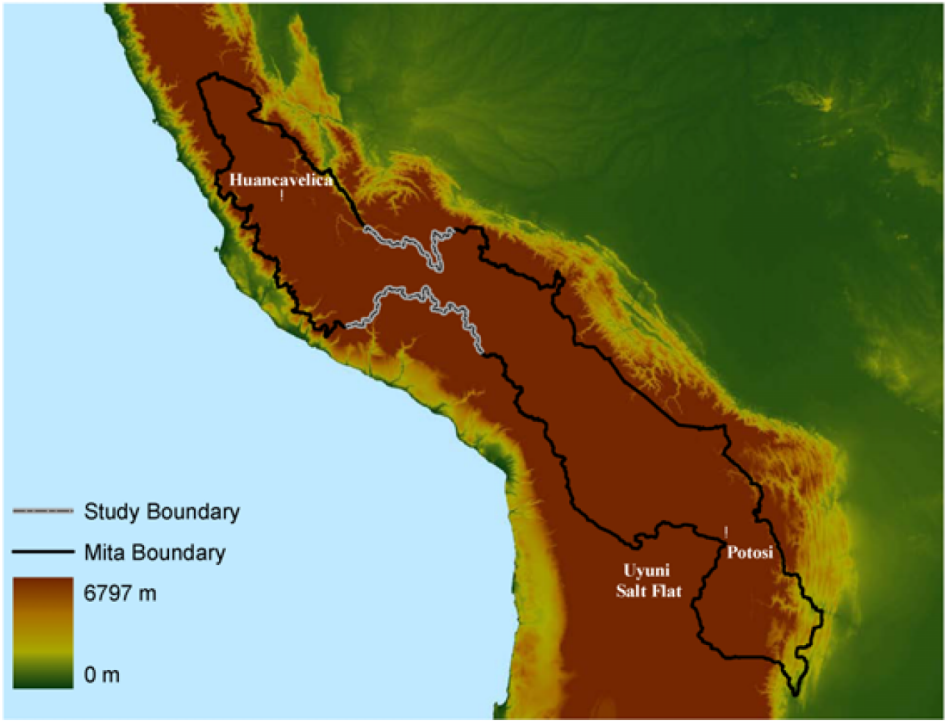


FIGURE 1.—The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.

注意，作者只选取了一段边界进行研究，而不是用全部的 mita 边界。

- 其他部分的 Mita 边界紧贴着安第斯山脉的陡峭悬崖，因此在海拔和人口的种族分布上呈现出明显的不连续性。

- 为了确保 RD 方法的有效性，需要在边界两侧控制其他可能影响结果的变量。作者发现，在所关注的边界段上，这些变量在统计上是一致的，这减少了其他因素可能引起的干扰。

Estimation framework

$$c_{idb} = \alpha + \gamma \text{ mita }_d + X'_{id}\beta + f\left(\text{ geographic location }_d\right) + \phi_b + \varepsilon_{idb},$$

where c_{idh} is the outcome variable of interest for observation i in district d along segment b of the mita boundary, and mita $_d$ is an indicator equal to 1 if district d contributed to the mita and equal to 0 otherwise; X_{id} is a vector of covariates that includes the mean area weighted elevation and slope for district d and (in regressions with equivalent household consumption on the left-hand side) demographic variables giving the number of infants, children, and adults in the household; f (geographic location $_d$) is the RD polynomial, which controls for smooth functions of geographic location. Various forms will be explored. Finally, ϕ_b is a set of boundary segment fixed effects that denote which of four equal length segments of the boundary is the closest to the observation's district capital.

作者使用半参方法估计，体现在这一项上：$f\left(\text{ geographic location }_d\right)$，展开就是：

$$\gamma_1^+\delta(X-c) + \gamma_1^-(1-\delta)(X-c) + \gamma_2^+\delta(X-c)^2 + \gamma_2^-(1-\delta)(X-c)^2$$

对应到 Estimation Results，X 分别是 Latitude and Longitude、Distance to Potosí、Distance to Mita Boundary。

Latitude and Longitude，经纬度，其实是两个变量，所以这篇文章涉及到双变量的多项式估计。Letting x denote longitude and y denote latitude, this polynomial is $x + y + x^2 + y^2 + xy + x^3 + y^3 + x^2y + xy^2$.

RD 的关键识别假设：1、除了处理之外的所有相关因素在 Mita 边界处必须平滑变化。That is, letting c_1 and c_0 denote potential outcomes under treatment and control, x denote longitude, and y denote latitude, identification requires that $E[c_1 \mid x, y]$ and $E[c_0 \mid x, y]$ are continuous at the discontinuity threshold. 然后作者检查了下面几个变量：elevation, terrain ruggedness, soil fertility, rainfall, ethnicity, preexisting settlement patterns, local 1572 tribute (tax) rates, and allocation of 1572 tribute revenues.

方括号和圆括号里面是标准误。T 统计量需要用 Inside 减 Outside，再除以标准误。

"< 25 km" 组的结果都是不显著的。

TABLE I

SUMMARY STATISTICS[a]

	Sample Falls Within											
	<100 km of *Mita* Boundary			<75 km of *Mita* Boundary			<50 km of *Mita* Boundary			<25 km of *Mita* Boundary		
	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.
GIS Measures												
Elevation	4042	4018	[188.77] (85.54)	4085	4103	[166.92] (82.75)	4117	4096	[169.45] (89.61)	4135	4060	[146.16] (115.15)
Slope	5.54	7.21	[0.88]* (0.49)***	5.75	7.02	[0.86] (0.52)**	5.87	6.95	[0.95] (0.58)*	5.77	7.21	[0.90] (0.79)*
Observations	177	95		144	86		104	73		48	52	
% Indigenous	63.59	58.84	[11.19] (9.76)	71.00	64.55	[8.04] (8.14)	71.01	64.54	[8.42] (8.43)	74.47	63.35	[10.87] (10.52)
Observations	1112	366		831	330		683	330		329	251	
Log 1572 tribute rate	1.57	1.60	[0.04] (0.03)	1.57	1.60	[0.04] (0.03)	1.58	1.61	[0.05] (0.04)	1.65	1.61	[0.02]* (0.03)

(*Continues*)

TABLE I—*Continued*

	Sample Falls Within											
	<100 km of *Mita* Boundary			<75 km of *Mita* Boundary			<50 km of *Mita* Boundary			<25 km of *Mita* Boundary		
	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.	Inside	Outside	s.e.
% 1572 tribute to												
Spanish Nobility	59.80	63.82	[1.39]***	59.98	63.69	[1.56]**	62.01	63.07	[1.12]	61.01	63.17	[1.58]
			(1.36)***			(1.53)**			(1.34)			(2.21)
Spanish Priests	21.05	19.10	[0.90]**	21.90	19.45	[1.02]**	20.59	19.93	[0.76]	21.45	19.98	[1.01]
			(0.94)**			(1.02)**			(0.92)			(1.33)
Spanish Justices	13.36	12.58	[0.53]	13.31	12.46	[0.65]	12.81	12.48	[0.43]	13.06	12.37	[0.56]
			(0.48)*			(0.60)			(0.55)			(0.79)
Indigenous Mayors	5.67	4.40	[0.78]	4.55	4.29	[0.26]	4.42	4.47	[0.34]	4.48	4.42	[0.29]
			(0.85)			(0.29)			(0.33)			(0.39)
Observations	63	41		47	37		35	30		18	24	

[a]The unit of observation is 20×20 km grid cells for the geospatial measures, the household for % indigenous, and the district for the 1572 tribute data. Conley standard errors for the difference in means between *mita* and non-*mita* observations are in brackets. Robust standard errors for the difference in means are in parentheses. For % indigenous, the robust standard errors are corrected for clustering at the district level. The geospatial measures are calculated using elevation data at 30 arc second (1 km) resolution (SRTM (2000)). The unit of measure for elevation is 1000 meters and for slope is degrees. A household is indigenous if its members primarily speak an indigenous language in the home (ENAHO (2001)). The tribute data are taken from Miranda (1583). In the first three columns, the sample includes only observations located less than 100 km from the *mita* boundary, and this threshold is reduced to 75, 50, and finally 25 km in the succeeding columns. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

Estimation Results

TABLE II

LIVING STANDARDS[a]

	Dependent Variable						
	Log Equiv. Household Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
	Panel A. Cubic Polynomial in Latitude and Longitude						
Mita	−0.284	−0.216	−0.331	0.070	0.084*	0.087*	0.114**
	(0.198)	(0.207)	(0.219)	(0.043)	(0.046)	(0.048)	(0.049)
R^2	0.060	0.060	0.069	0.051	0.020	0.017	0.050
	Panel B. Cubic Polynomial in Distance to Potosí						
Mita	−0.337***	−0.307***	−0.329***	0.080***	0.078***	0.078***	0.063*
	(0.087)	(0.101)	(0.096)	(0.021)	(0.022)	(0.024)	(0.032)
R^2	0.046	0.036	0.047	0.049	0.017	0.013	0.047
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary						
Mita	−0.277***	−0.230**	−0.224**	0.073***	0.061***	0.064***	0.055*
	(0.078)	(0.089)	(0.092)	(0.023)	(0.022)	(0.023)	(0.030)
R^2	0.044	0.042	0.040	0.040	0.015	0.013	0.043
Geo. controls	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes
Clusters	71	60	52	289	239	185	63
Observations	1478	1161	1013	158,848	115,761	100,446	37,421

[a]The unit of observation is the household in columns 1–3 and the individual in columns 4–7. Robust standard errors, adjusted for clustering by district, are in parentheses. The dependent variable is log equivalent household consumption (ENAHO (2001)) in columns 1–3, and a dummy equal to 1 if the child has stunted growth and equal to 0 otherwise in columns 4–7 (Ministro de Educación (2005a)). ***Mita*** is an indicator equal to 1 if the household's district contributed to the ***mita*** and equal to 0 otherwise (Saignes (1984), Amat y Juniet (1947, pp. 249, 284)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the ***mita*** boundary. All regressions include controls for elevation and slope, as well as boundary segment fixed effects (F.E.s). Columns 1–3 include demographic controls for the number of infants, children, and adults in the household. In columns 1 and 4, the sample includes observations whose district capitals are located within 100 km of the ***mita*** boundary, and this threshold is reduced to 75 and 50 km in the succeeding columns. Column 7 includes only observations whose districts border the ***mita*** boundary. 78% of the observations are in ***mita*** districts in column 1, 71% in column 2, 68% in column 3, 78% in column 4, 71% in column 5, 68% in column 6, and 58% in column 7. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

TABLE III

SPECIFICATION TESTS[a]

	Dependent Variable						
	Log Equiv. Household Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
Alternative Functional Forms for RD Polynomial: Baseline I							
Linear polynomial in latitude and longitude							
Mita	−0.294***	−0.199	−0.143	0.064***	0.054**	0.062**	0.068**
	(0.092)	(0.126)	(0.128)	(0.021)	(0.022)	(0.026)	(0.031)
Quadratic polynomial in latitude and longitude							
Mita	−0.151	−0.247	−0.361	0.073*	0.091**	0.106**	0.087**
	(0.189)	(0.209)	(0.216)	(0.040)	(0.043)	(0.047)	(0.041)
Quartic polynomial in latitude and longitude							
Mita	−0.392*	−0.324	−0.342	0.073	0.072	0.057	0.104**
	(0.225)	(0.231)	(0.260)	(0.056)	(0.050)	(0.048)	(0.042)
Alternative Functional Forms for RD Polynomial: Baseline II							
Linear polynomial in distance to Potosí							
Mita	−0.297***	−0.273***	−0.220**	0.050**	0.048**	0.049**	0.071**
	(0.079)	(0.093)	(0.092)	(0.022)	(0.022)	(0.024)	(0.031)
Quadratic polynomial in distance to Potosí							
Mita	−0.345***	−0.262***	−0.309***	0.072***	0.064***	0.072***	0.060*
	(0.086)	(0.095)	(0.100)	(0.023)	(0.022)	(0.023)	(0.032)
Quartic polynomial in distance to Potosí							
Mita	−0.331***	−0.310***	−0.330***	0.078***	0.075***	0.071***	0.053*
	(0.086)	(0.100)	(0.097)	(0.021)	(0.020)	(0.021)	(0.031)
Interacted linear polynomial in distance to Potosí							
Mita	−0.307***	−0.280***	−0.227**	0.051**	0.048**	0.043*	0.076***
	(0.092)	(0.094)	(0.095)	(0.022)	(0.021)	(0.022)	(0.029)
Interacted quadratic polynomial in distance to Potosí							
Mita	−0.264***	−0.177*	−0.285**	0.033	0.027	0.039*	0.036
	(0.087)	(0.096)	(0.111)	(0.024)	(0.023)	(0.023)	(0.024)

TABLE III—*Continued*

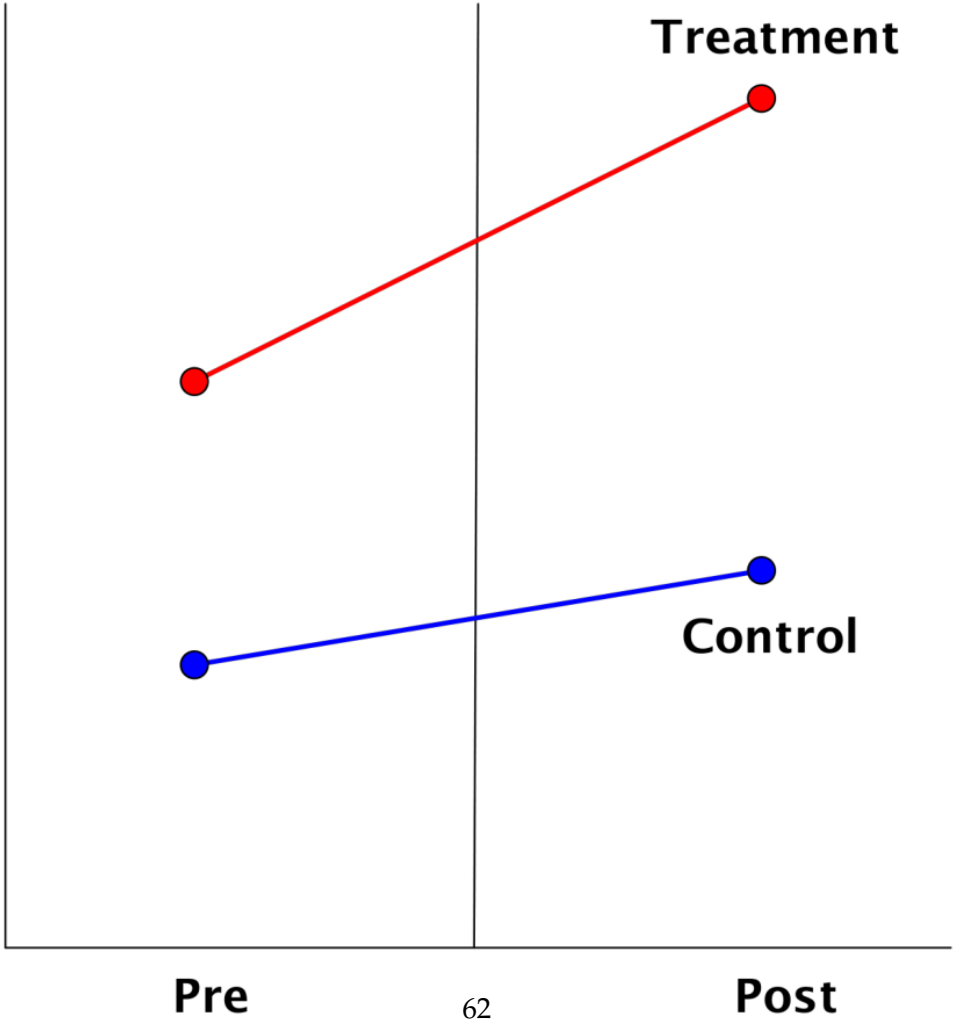
Sample Within:	Dependent Variable: Log Equiv. Household Consumption (2001) <100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	Dependent Variable: Stunted Growth, Children 6–9 (2005) <100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
Alternative Functional Forms for RD Polynomial: Baseline III							
Linear polynomial in distance to *mita* boundary							
Mita	−0.299***	−0.227**	−0.223**	0.072***	0.060***	0.058**	0.056*
	(0.082)	(0.089)	(0.091)	(0.024)	(0.022)	(0.023)	(0.032)
Quadratic polynomial in distance to *mita* boundary							
Mita	−0.277***	−0.227**	−0.224**	0.072***	0.060***	0.061***	0.056*
	(0.078)	(0.089)	(0.092)	(0.023)	(0.022)	(0.023)	(0.030)
Quartic polynomial in distance to *mita* boundary							
Mita	−0.251***	−0.229**	−0.246***	0.073***	0.064***	0.063***	0.055*
	(0.078)	(0.089)	(0.088)	(0.023)	(0.022)	(0.023)	(0.030)
Interacted linear polynomial in distance to *mita* boundary							
Mita	−0.301*	−0.277	−0.385*	0.082	0.087	0.095	0.132**
	(0.174)	(0.190)	(0.210)	(0.054)	(0.055)	(0.065)	(0.053)
Interacted quadratic polynomial in distance to *mita* boundary							
Mita	−0.351	−0.505	−0.295	0.140*	0.132	0.136	0.121*
	(0.260)	(0.319)	(0.366)	(0.082)	(0.084)	(0.086)	(0.064)
Ordinary Least Squares							
Mita	−0.294***	−0.288***	−0.227**	0.057**	0.048*	0.049*	0.055*
	(0.083)	(0.089)	(0.090)	(0.025)	(0.024)	(0.026)	(0.031)
Geo. controls	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes
Clusters	71	60	52	289	239	185	63
Observations	1478	1161	1013	158,848	115,761	100,446	37,421

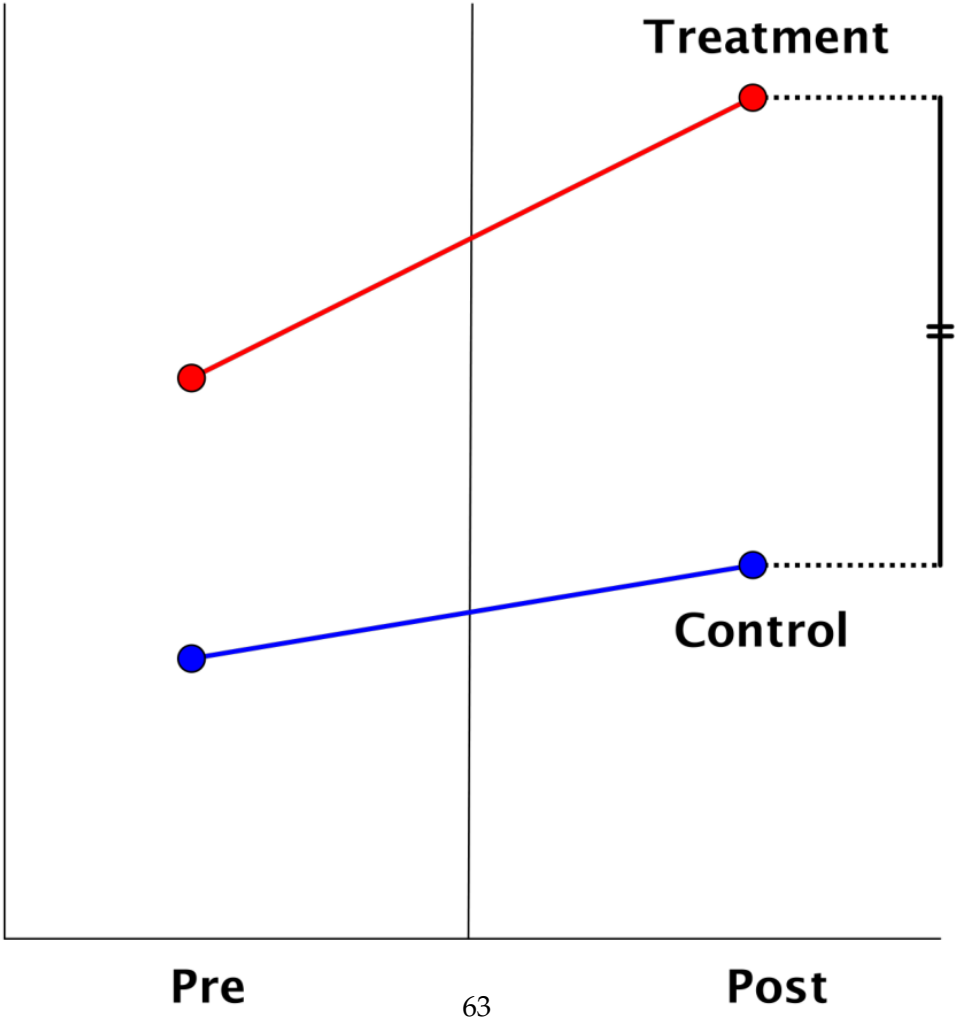
[a]Robust standard errors, adjusted for clustering by district, are in parentheses. All regressions include geographic controls and boundary segment fixed effects (F.E.s). Columns 1–3 include demographic controls for the number of in-

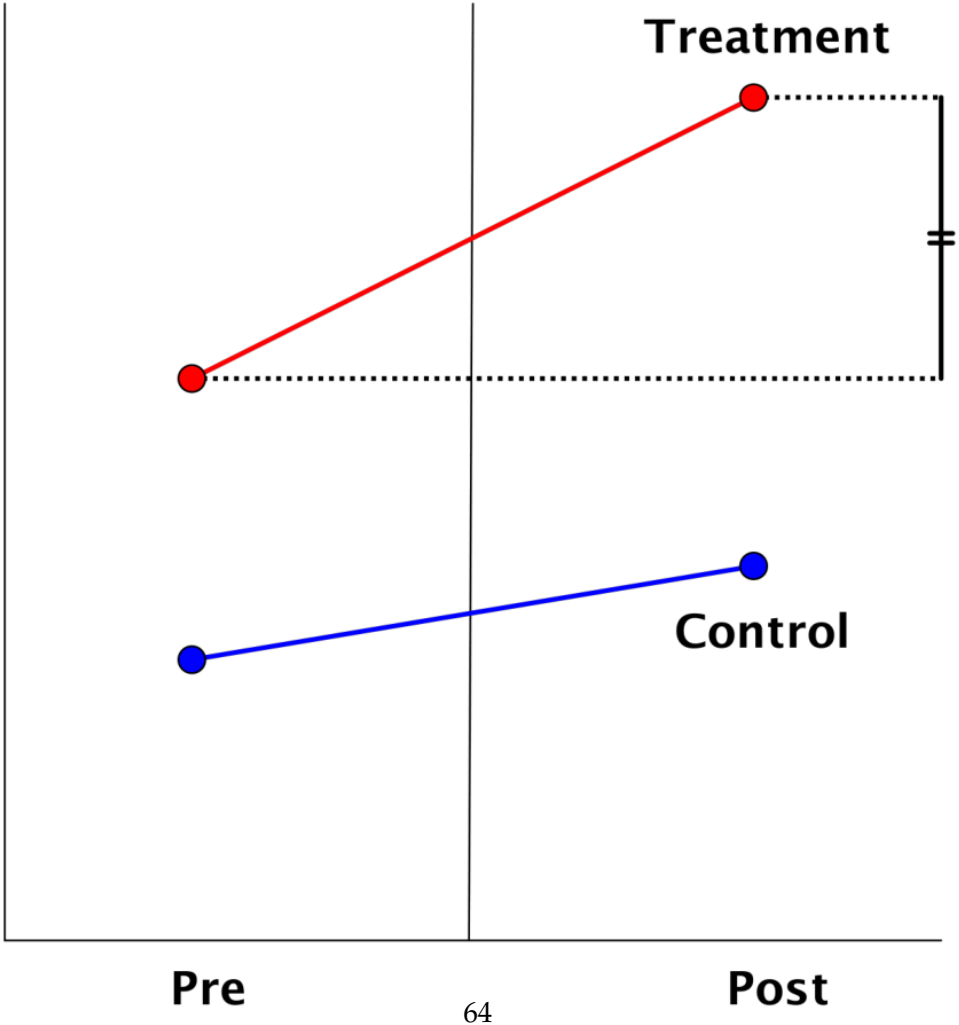
1.9 对 RD 操作过程的批评

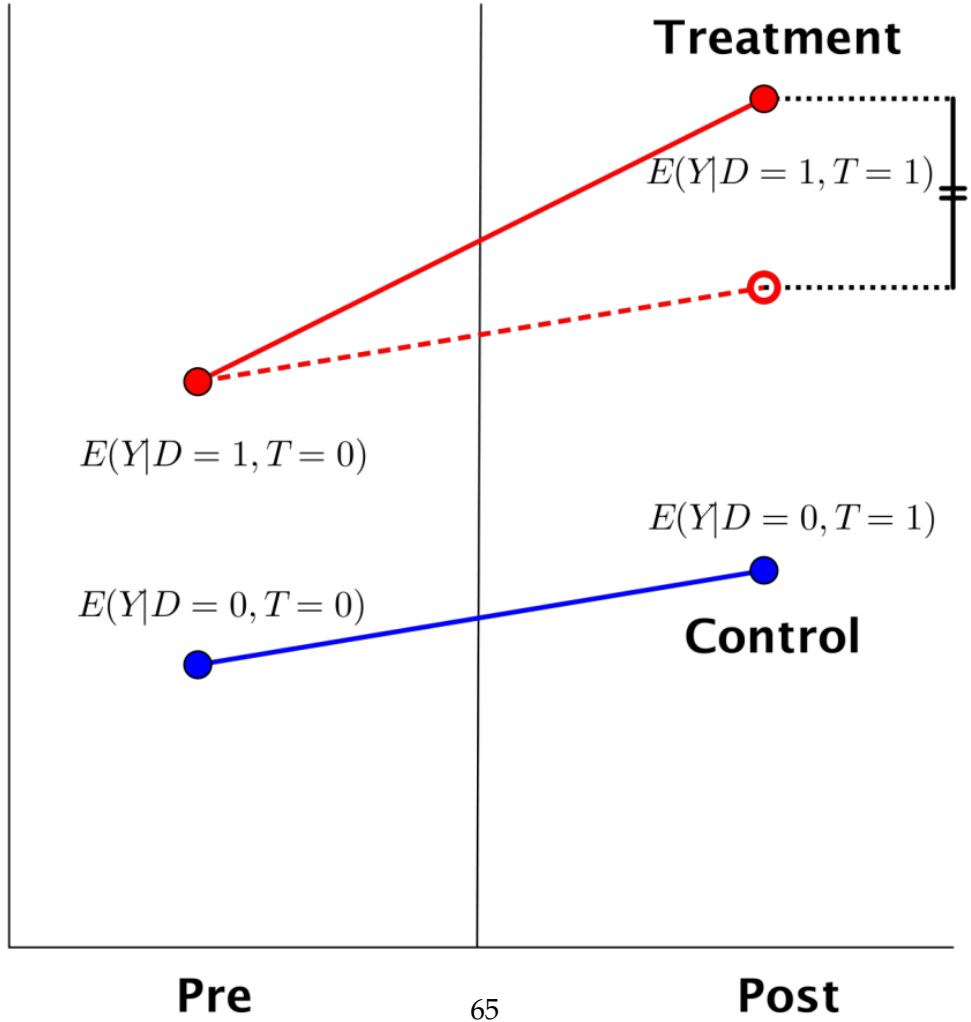
Visual Inference and Graphical Representation in Regression Discontinuity Designs QJE(2023)

2　双重差分：DID


Treatment
Control
Pre
Post


Treatment
Control
Pre
Post


Treatment
Control
Pre
Post


Treatment
$E(Y|D=1,T=1)$
$E(Y|D=1,T=0)$
$E(Y|D=0,T=1)$
$E(Y|D=0,T=0)$
Control
Pre
Post

2.1 双重差分的工作原理

•DID 的识别

$$\begin{aligned}\tau(X) =& E(Y|X, D = 1, T = 1) - E(Y|X, D = 1, T = 0) \\ & - [E(Y|X, D = 0, T = 1) - E(Y|X, D = 0, T = 0)] \\ =& E(Y^1|X, D = 1, T = 1) - E(Y^0|X, D = 1, T = 0) \\ & - [E(Y^0|X, D = 0, T = 1) - E(Y^0|X, D = 0, T = 0)] \\ =& \left[E\left(Y^1 \mid X, D = 1, T = 1\right) - E\left(Y^0 \mid X, D = 1, T = 1\right)\right] \\ & + \left[E\left(Y^0 \mid X, D = 1, T = 1\right) - E\left(Y^0 \mid X, D = 1, T = 0\right)\right] \\ & - \left[E\left(Y^0 \mid X, D = 0, T = 1\right) - E\left(Y^0 \mid X, D = 0, T = 0\right)\right]\end{aligned}$$

因此识别条件是

$$\begin{aligned}&\left[E\left(Y^0 \mid X, D = 1, T = 1\right) - E\left(Y^0 \mid X, D = 1, T = 0\right)\right] \\ =& \left[E\left(Y^0 \mid X, D = 0, T = 1\right) - E\left(Y^0 \mid X, D = 0, T = 0\right)\right]\end{aligned}$$

（如果右边等于零，那么 DID 等于 before-after comparison）也可写作

$$\begin{aligned}&[E\left(Y^0 \mid X, D = 1, T = 1\right) - E\left(Y^0 \mid X, D = 0, T = 1\right)] \\ =& [E\left(Y^0 \mid X, D = 1, T = 0\right) - E\left(Y^0 \mid X, D = 0, T = 0\right)]\end{aligned}$$

（如果右边等于零，那么 DID 等于 cross-sectional group difference）

•DID 要求随机分组么？不需要

随机分组意味着：

$$E(Y^0|X, D = 1, T = 0) = E(Y^0|X, D = 0, T = 0)$$
$$E(Y^0|X, D = 1, T = 1) = E(Y^0|X, D = 0, T = 1)$$

而实际上，DID 的识别假设可以简写作

$$E(\Delta Y^0|X, D = 1) = E(\Delta Y^0|X, D = 0)$$

也就是说给定 X，ΔY^0 均值独立于 D，即关于 ΔY^0 是随机分组的！

•在此条件下，DID 估计量识别了 ATT at the post-treatment period.

$$\tau_{DID} = \int \tau(X) dF(X \mid D = 1, T = 1) = E\left(Y^1 - Y^0 \mid D = 1, T = 1\right)$$

•线性模型视角下的识别。考察下面的线性模型（简便起见，省略 X）

$$\begin{aligned} Y_{d0}^0 &= \beta_0 + \beta_1 D + U_0 \\ Y_{d1}^0 &= \beta_0 + \beta_1 D + \beta_2 + U_1 \\ Y_{d1}^1 &= \beta_0 + \beta_1 D + \beta_2 + \beta_3 + U_1 \end{aligned}$$

β_3 为 treatment effect

$$\begin{aligned} Y_{dt} &= (1-T)Y_{d0}^0 + T\left[(1-D)Y_{01}^0 + DY_{11}^1\right] \\ &= \beta_0 + \beta_1 D + \beta_2 T + \beta_3 D \cdot T + [(1-T)U_0 + TU_1] \end{aligned}$$

之前的识别条件可写作

$$\begin{aligned} &E\left(U_1 \mid D=1, T=1\right) - E\left(U_0 \mid D=1, T=0\right) \\ =\ &E\left(U_1 \mid D=0, T=1\right) - E\left(U_0 \mid D=0, T=0\right) \end{aligned}$$

此时

$$\begin{aligned} \tau(X) =& E(Y \mid X, D=1, T=1) - E(Y \mid X, D=1, T=0) \\ &- [E(Y \mid X, D=0, T=1) - E(Y \mid X, D=0, T=0)] \\ =& \beta_3 \end{aligned}$$

•DID 就是 before-after comparison + matching.

•所有交互项模型都可以从 DID 的角度来理解。

2.2 Regression DiD

$$y_{it} = \beta_0 + \beta_1 \text{ Post }_t + \beta_2 \text{ Treat }_i + \beta_3 \text{ Treat }_i \times \text{ Post }_t + \gamma X_{it} + e_{it}$$

	Treatment Group ($T_i = 1$) (1)	Control Group ($T_i = 0$) (2)	Difference (1) – (2)
Post-Treatment Period ($P_t = 1$) (a)	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1$	$\beta_2 + \beta_3$
Pre-Treatment Period ($P_t = 0$) (b)	$\beta_0 + \beta_2$	β_0	β_2
Difference (a) – (b)	$\beta_1 + \beta_3$	β_1	$\boldsymbol{\beta_3}$

Figure 1: Treatment Effect

TWFE:

$$y_{it} = \beta_0 + \alpha_i + \delta_t + \beta_3 d_{it} + \gamma X_{it} + e_{it}$$

α_i : Individual fixed effects that change across individuals (state-specific characteristics, individual's gender, etc.)

δ_t : Time fixed effects that change across time (e.g. year dummies to allow intercept to vary across different years)

d_{it} : dummy variable which equals 1 if the unit of observation is in the post-treatment period (in contrast to Pt equals 1 in the second time period)

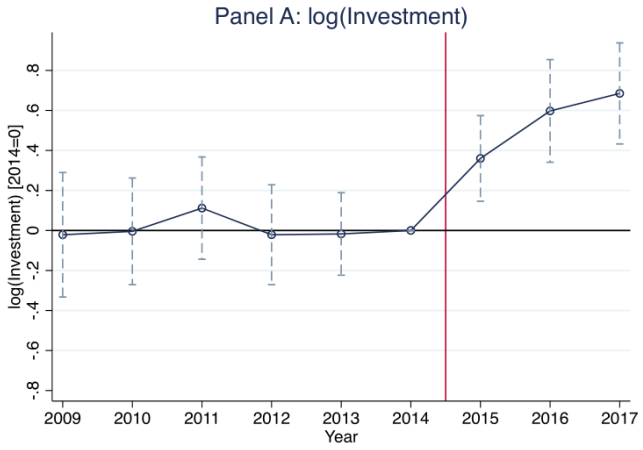
对 DID 的威胁

- $D \cdot T$ 反映的可能是其它 treatment effect（不可解决）
- 不好的控制组，检验平行趋势很重要!
- D 不稳定，即处理组和控制组存在 compositional changes. 换句话说，是否进入处理组可操纵，例如 program application, policy threshold, migration 等等。当 treatment 针对的是个体，但实施是在地区层面时尤其要注意，控制个体特征一定程度上能缓解这一问题。

平行趋势检验示例以及 Event study

$$Y_{it} = \alpha_i + \lambda_t + \sum_{r \neq -1} \beta_r \times 1\left[D_i = 1\right] \times 1[t = r] + e_{it}.$$

Figure 3: Tax Effects on Real Outcomes of Affected and Unaffected Firms



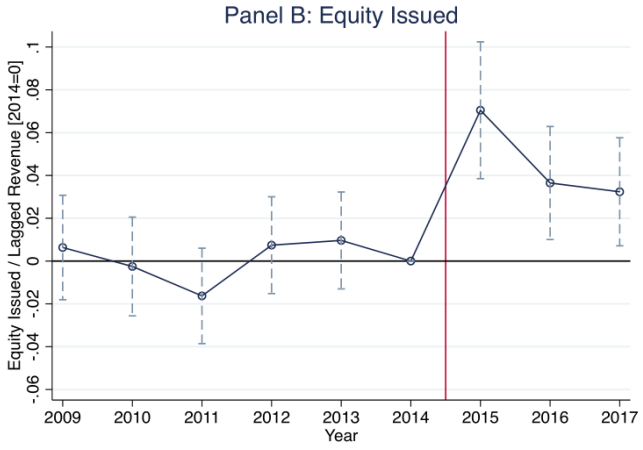
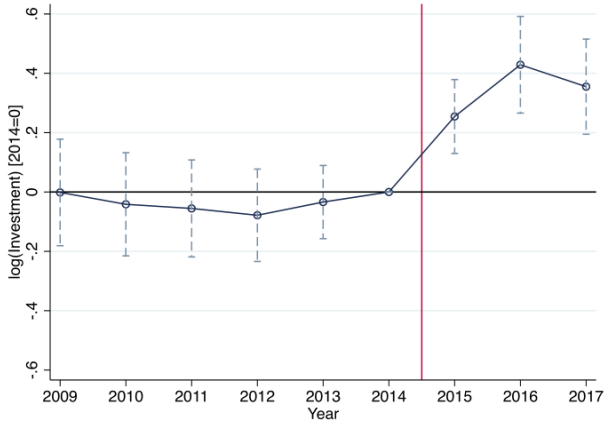
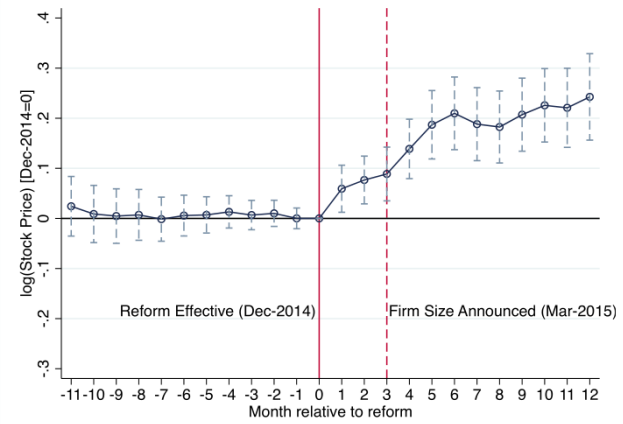


Figure 4: Tax Effects on *ln*(Investment): Listed and Private Firms



Notes: This figure shows the coefficients on *Treated* × *Time* for firms' investment, defined as log(expenditures on physical capital assets), in equation (7) using both publicly listed and private firms. The dashed lines indicate 95% confidence intervals for these coefficient estimates. The solid vertical line indicates the reform year.

Figure 5: Tax Effects on *ln*(Stock Prices): Listed Firms



Notes: This figure shows the coefficients on *Treated* × *Time* for each time period (month). The outcome variable is log(Stock Prices). The dashed lines indicate the 95% confidence intervals for those coefficient estimates. The solid vertical line indicates the month in which the reform was implemented, and the dash vertical line indicates the month in which the firm size was publicly announced through the annual audit reports. The sample periods are from the beginning of 2014 to the end of 2015. The sample is restricted to publicly listed companies, where I can observe their stock prices at the monthly frequency.

示例 魁北克取消对职工医疗保险的税收补贴 (Finkelstein, 2002, JPubE).

- 背景：如果企业为员工交健康保险，这部分钱当然不用企业交个人所得税。加拿大各州这部分保费也不算做员工的收入，不用交联邦所得税和工薪税。但是，在除了魁北克以外的各州，计算州所得税时，这部分保费也不算做员工收入，不用交州所得税。在魁北克 93 年时，把保费算作员工收入，要征收州所得税。

- 政策：1993 年，魁北克对雇主提供的补充健康保险的税收补贴减少了近 60‰

- 发现:

  – 雇主提供的补充保险覆盖面减少了约五分之一。

  – 提供保险的税收价格弹性为-0.5

$$\begin{aligned} INSURANCE =& \beta_0 + \beta_1(QUEBEC \times AFTER) \\ &+ \beta_2 QUEBEC + \beta_3 AFTER + X\beta_4 + \varepsilon \end{aligned}$$

Compositional changes in the treatment group relative to the control group in characteristics that are correlated with coverage by employer-provided supplementary health insurance could drive differences in trends in coverage between the treatment and control group over the sample period. For example, if union membership decreases in Quebec relative to the control provinces, and union members are more likely to be covered by employer-provided supplementary health insurance than non-union members, this change in union membership could drive differences in the trends in insurance coverage in the two regions. I therefore control flexibly for this possibility by including the X matrix of covariates in Eq. (3). These covariates consist of dummies for union membership, marital status, gender, spousal employment (conditional on marriage), age, educational attainment, number of children under 25, self-reported health

status, full time versus part time work, full year versus part year work, personal income, and occupation. I also include the provincial unemployment rate as a covariate to control for any differences in the macroeconomic cycle across provinces.

●工具变量法计算保险的税收价格弹性：

$$\begin{aligned}\text{INSURANCE} =& \beta_0 + \beta_1 \text{ TAXPRICE} + \beta_2 \text{ PROVINCE} + \beta_3 \text{ AFTER} \\ &+ \text{X}\beta_4 + \varepsilon\end{aligned}$$

解释 TAXPRICE: 如果雇主没有帮雇员买保险，把钱发给雇员，就需要交各种税，企业的利润要交企业所得税，发给员工之后员工要交联邦所得税、州所得税和薪酬税，雇员实际上也拿不到全额的节省下来的保险费用。τ_{cons} 是消费税。

$$\begin{aligned}&\text{TAXPRICE} = \\ &\left\{(1+\tau_{\text{cons}}) * \left(1-\tau_{\text{fed}} - \tau_{\text{prov}} - \tau_{\text{payroll, worker}}\right)\right\} / \left\{1+\tau_{\text{payroll, firm}}\right\}\end{aligned}$$

在魁北克改革以前，雇员需要在 1 美元的额外保险和 57 美分的税后消费之间选择，而在改革以后，需要在 1 美元额外保险和 82 美分的税后消费之间选择。

$TAXPRICE$ is then instrumented for using $QUEBEC \times AFTER$. The variation in the individual's tax price used to identify the relationship between the tax price and coverage by workplace-based insurance therefore comes from the Quebec reform. This approach allows me to translate the estimated effect of the Quebec reform into a parameterized estimate of the relationship between the tax price of employer-provided supplementary health insurance and coverage by such insurance. This parameterized estimate can then be compared to estimates that have used other sources of variation to estimate the relationship between the tax price of insurance and coverage by such insurance.

Patterns from patchwork 1975 年，Alabama 把最低饮酒年龄减低到 19 岁。

$$\begin{aligned} Y_{st} = \alpha + \beta TREAT_s + \gamma POST_t \\ + \delta_{rDD}(TREAT_s \times POST_t) + e_{st} \end{aligned}$$

$TREAT_s$ 是 Alabama 的 dummy variable，$POST$ 是表示 1975 年以后的 dummy variable。

Probing DID Assumptions

$$Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} + \sum_{k=Alaska}^{Wyoming} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} + e_{st}$$

包含多个个体和时间的样本允许放松 common trends assumption，即允许个体在没有 treatment effect 时有非平行的趋势：

$$Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} + \sum_{k=Alaska}^{Wyoming} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} + \sum_{k=Alaska}^{Wyoming} \theta_k (STATE_{ks} \times t) + e_{st}$$

这个模型假设了如果没有 treatment effect，state k 的死亡率相比于线性趋势会偏离 θ_k。

2.3 基于队列构造 DID：

DID+IV

Jackson, Johnson, Persico, The Effects of School Spending on Educational and Economic Outcomes —— Evidence from School Finance Reforms，QJE2016

文章的核心内容是研究公共学校支出对学生的教育和经济成果的影响，特别是通过分析 20 世纪 70 年代初至 80 年代加速进行的美国学校财政改革所引起的变化。

主要发现包括：

- 每名学生每年在 12 年的公立学校教育中增加 10% 的支出，可以使学生完成的教育年限增加 0.31 年，工资提高约 7%，并减少成年贫困发生率 3.2 个百分点。
- 对于来自低收入家庭的儿童，支出增加的效果更为显著。
- 外生支出增加与学校投入的显著改善有关，包括降低师生比例、提高教师工资和延长学年。

制度背景

在大多数州，20 世纪 70 年代之前，K-12 教育的大部分资源是通过地方财产税筹集的。由于地方财产税基础通常在房屋价值较高的地区更高，并且存在高度的按社会经济地位居住的种族隔离，过分依赖地方融资有助于富裕地区能够为每个学生花费更多。为了回应富裕/高收入和贫困地区之间每个学生支出的州内差异，1971 年至 2010 年间，州最高法院推翻了 28 个州的学校财政体系，并许多州实施了立法改革，这些改革导致了公共教育资金的重要变化。

一旦现行的学校财政体系被认定违反宪法，大多数 SFRs 改变了支出公式的参数，以减少学校支出的不平等，并削弱教育支出水平与学区的财富和收入水平之间的关系（Card 和 Payne，2002）。

然而，各州为实现这些目标而设计的州公式差异很大。正如 Hoxby（2001）所指出的，SFR 对学校支出的影响取决于:

1. 改革引入的学校资金公式类型

2. 资金公式与学区的特定特征如何相互作用。

遵循 Jackson、Johnson 和 Persico（2014）所概述的分类，并将改革归类为几种主要类型:

1. 基础计划保证每个学生的学校支出基础水平，并旨在增加最低支出学区的每个学生支出。

2. 支出限制禁止每个学生支出水平超过某个预定金额。这类计划往往会减少高支出和更富裕学区的支出，并可能在长期内减少所有学区的支出。

3. 努力奖励计划将当地为教育筹集的资金与额外的国家资金相匹配（通常对于低收入地区的匹配率更高）。这类计划将倾向于增加所有学区的支出，对于低收入地区的学区增加更多。

4. 最后，均等化计划旨在通过向所有学区征税并重新分配资金给低财富和低收入学区来均等化支出水平。

数据

收集并整合了有关学校支出的数据，并将其与描述各种学校财政改革（SFRs）的数据库以及一个追踪个体从童年到成年的全国代表性纵向数据集相连接。教育资金数据来自多个来源，将其组合成一个 1967 年以及 1970 年至 2010 年每年美国学区的每名学生支出的面板数据集。为了避免混淆名义上的变化与随时间的实际支出变化，使用劳工统计局的消费者价格指数将所有年份的学校支出转换为 2000 年的美元。使用 1969 年有效的学区边界将学区与县关联起来，并从 1970 年的人口普查中获取县级家庭收入中位数数据。然后将这些支出数据与 1972 年至 2010 年之间的改革数据库相连接。

关于长期成果的数据来自收入动态小组研究（Panel Study of Income Dynamics，PSID），该研究将个体与其童年时期的人口普查区块相关联。样本包括 1955 年至 1985 年间出生的 PSID 样本成员，并跟踪至 2011 年成年。这些队列跨越了首批法院命令的 SFRs（第一个法院命令是在 1971 年）的时期，并且到 2011 年也已经完成了正规学校教育。样本中有三分之二的人在 1971 年至 2000 年间所在的州接受了法院命令的 SFRs。将最早的可用儿童居住地址与 1969 年的学区边界匹配，以避免由于学区边界随时间变化而产生的内生性问题。匹配算法在在线附录 E 中有详细说明。每个记录都与他们学校年代对应的学区层面的学校支出和前述学校财政变量的数据合并。最后，合并了 1962 年政府普查和 1970 年人口普查的县级特征数据，以及其他关键政策变化的信息（在第二节中描述），允许进行异常丰富的控制。

最终样本包括来自 1,409 个学区、1,031 个县和所有 50 个州及哥伦比亚特区的 15,353 名个体的 93,022 个成人-年观测值。为了描述童年家庭环境，计算了 12 至 17 岁期间父母收入和教育的平均值，并在出生时测量家庭结构。

分析框架

文章的目标是识别儿童时期每名学生公共学校支出对成人成果的因果效应。由于一个地区每名学生支出与在这些学校就读的学生成人成果之间的相关性很可能受到其他因素的干扰（由于居住隔离、Tiebout 排序、补偿性支出增加等），寻找每名学生支出的外生变化。为此，我们仅使用由于法院命令的学校财政改革（SFRs）引起的学校支出变化。

如第二节所述，SFRs 的目标是在低支出地区增加支出水平，并减少学区间每名学生学校支出水平的差异。按设计，一些学区经历了支出增加，而其他学区则经历了减少（Murray, Evans, and Schwab 1998; Hoxby 2001; Card and Payne 2002）。利用 SFRs 目标引起的学校支出变化，即在低支出地区增加资金并在学区间减少资金差异水平。将这种变化视为外生的，并使用由此产生的自然实验来估计每名学生支出对成人成果的因果效应。

政策实验：

- 在州法院命令 SFR 的年份满 17 岁的个人应该在改革实施时已完成中等教育。这样的队列应该不受改革的影响，因此我们将其归类为未暴露的。

- 在州法院命令 SFR 的年份满 16 岁或更小的个人可能在改革实施时正在上小学或中学。称这些队列为暴露的。

- 可以通过比较来自同一学区的暴露和未暴露出生队列之间的成果变化，来估计个人在成人成果上的暴露效应。为了考虑不同出生队列之间的任何潜在差异，可以使用同一出生队列在非改革学区的成果变化作为比较。暴露队列与未暴露队列在处理学区的结果差异减去同一出生队列在比较学区的结果差异，可以得到该学区成果上的双重差分（DiD）估计。

- 关键假设是，学区内由改革引起的支出变化与可能直接影响成人成果的其他学区级变化无关。在这一假设

下，检验儿童时期每名学生支出对成人成果因果效应的一个检验是，来自同一学区的暴露和未暴露队列之间的成果差异（即暴露效应）是否倾向于对那些经历更大改革引起的每名学生支出增加的学区更大（即剂量-反应效应）。

- 一个额外的检验是，是否观察到那些经历了更多年的支出增加的个体在成人成果上取得了更大的改善。使用两阶段最小二乘（2SLS）双重差分回归模型来实现这些直观的测试，其中儿童时期的每名学生支出是感兴趣的内生处理变量。

由于法院命令的 SFR 而在学区内发生的每名学生支出变化为 dosage。使用暴露于法院命令的 SFR 的度量以及暴露于法院命令的 SFR 与剂量预测因子的交互项，来预测学区间跨队列的看似外生的支出变化。原则上，可以使用仅暴露的效应来预测支出变化的水平。然而，仅使用暴露作为工具，可能违反了有效工具的单调性假设，因为一些 SFRs 导致支出增加，而其他一些则导致减少。即使暴露本身是一个有效的工具，这种方法将排除所有由于在同一州内通过 SFR 而发生在学区内部的变化。实际上，仅使用暴露变化并忽略剂量变化会导致第一阶段弱工具（F 统计量为 5.7）。

按照 Card 和 Krueger（1992）的方法，衡量儿童时期学校支出的指标是个体在其童年学区预期学龄年（5-17 岁）的平均学校支出（以下简称支出），PPE5–17。为了仅依赖于由于法院命令的 SFR 而产生的学区跨队列的支出变化，通过 2SLS 估计以下形式的方程组。

$$\begin{aligned}\ln\left(P\bar{P}E_{5-17}\right)_{idb} &= \pi_1\left(\text{Exp}_{idb}\times \text{Dosage }_d\right)+\pi_2\left(\text{Exp}_{idb}\right)\\ &\quad +\Pi C_{idb}+\rho_d+\rho_b+\xi_{idb}\\ Y_{idb} &= \delta\cdot\ln\left(\widehat{P\bar{P}E_{5-17}}\right)_{idb}+\Phi C_{idb}+\theta_d+\theta_b+\varepsilon_{idb}.\end{aligned}$$

其中，内生处理变量 ln(PPE5–17)idb 是个体 i 在出生队列 b 的学区 d 的预期学龄年的平均学校支出的

自然对数。控制了学区固定效应 ρ_d 和 θ_d。为了解释出生队列之间的一般潜在差异（无论是否暴露），包括出生队列固定效应 ρ_b 和 θ_b。通过出生队列固定效应，估计的改革学区跨队列的变化都是相对于那个时期没有实施改革的学区中相同出生队列的变化。暴露度量 Exp_{idb} 是个体 i 在出生队列 b 来自学区 d，在州法院命令 SFR 通过后发生的学龄年数。Exp_{idb} 在州-出生队列级别变化。这个变量从 0（对于那些在州法院命令年份满 17 岁或以上的人）到 12（对于那些在州法院命令年份满 5 岁或更小的人）。为了捕捉暴露条件下剂量的变化，交互项 Exp_{idb} 与 $Dosaged$。$Dosaged$ 是学区 d 由于法院命令的 SFR 引起的支出变化的学区级度量。这个双重差分 2SLS 模型比较了来自同一学区的暴露和未暴露队列之间的成果差异，这些队列在暴露时间和暴露程度上有所不同。如果由于改革引起的支出增加而暴露的队列在成人成果上也倾向于随后经历更大的改善（相对于未暴露的队列），那么方程（2）中的系数 δ 将是正的（对于像工资这样更大的正值更好的成果）。然而，如果暴露于更大的改革引起的支出变化（跨队列）与成果的变化无关，那么系数 δ 将是 0。只要法院命令的 SFRs 的时间与影响成果的学区级变化无关（无论剂量如何），系数 δ 应该揭示学校支出对成人成果的因果效应。重要的是要注意，由于改革之前儿童学区可能并不总是个体实际就读的学区（由于改革后居住流动性的变化），δ 是一个意向处理估计，量化了在个体童年学区增加每名学生学校支出的政策效应。

关键变量之一是 Dosaged，即由于法院命令的 SFR，暴露队列在学区 d 中经历的支出变化。尽管 $Dosaged$ 没有直接观察到，但可以使用剂量的代理变量（或预测因子）在方程（1）中。尽管使用学区在改革后实际经历的支出变化作为剂量的代理似乎是合理的，但不采用这种方法，以避免内生性偏差。学区在改革后实际经历的支出变化将包括所有与法院命令时间相一致的支出变化。为了避免使用可能由于学区其他政策变化或可能直接影响成果的其他本地变化而发生的学校支出变化，仅使用改革前学区的特征来预测剂量。这排除了由于学区其他变化可能直接影响学生成果而可能导致的学校支出变化。作者提出了两种改革前剂量预测方法。为了确保分离了由法院命令的 SFRs 引起的变化，在 C_{idb} 中包括了各种额外的个体、家庭和童年县级控制变量。这些包括父母的教育和职业地位、父母的收入、孩子出生时母亲的婚否、出生体重、儿童健康保险覆盖以及性别。C_{idb} 还包括种族-普查分区-出生队列固定效应，以及与 1960 年童年县级的各种特征（贫困率、黑人百分比、平均教育水平、城市化百分比和人口规模）相互作用的出生队列线性趋势。最后，为了避免将文章希望估计的效果与研究期间其他政策的影响相混淆，C_{idb} 包括了对县级按年份的学校去种族隔离、医院去种族隔离、社区卫生中心、国家对幼儿园的资金支持、人均起步资金支出、标题 I 学校资金、税收限制政策

的实施、儿童食品券平均支出、对有依赖儿童的家庭援助、医疗补助和失业保险的控制（Chay, Guryan, and Mazumder 2009; Hoynes, Schanzenbach, and Almond 2012; Johnson 2011）。标准误差在学校区级别聚类。

这个 2SLS 的双重差分模型，预测了暴露队列在学区中更积极的支出变化，这些变化与更高的剂量有关。研究设计的可信度取决于这样一个假设：法院命令的 SFRs 的时间与直接影响成果的学区层面的变化无关（无论剂量如何）。

度量 Dosage

方法一：估计了一个灵活的事件研究模型。

法院命令的 SFRs 倾向于通过增加以前低支出地区的支出来使各州内部的每名学生支出平均化，并对以前高支出的地区影响较小。因此，法院出了 SFR 之后，不同学区的教育支出增加可能多可能少，这一差异更大程度上来自这个相对外生的 SFR 政策冲击。

$$\ln\left(P\bar{P}E_{5-17}\right)_{idb} = \sum_{Q_{ppe}=1}^{4}\sum_{T=-20}^{20}\left(I_{T_{idb}=T}\times I_{Q_{ppe72,d}=Q_{ppe}}\right)\cdot\alpha_{T,Q_{ppe}} + \Pi C_{idb} + \theta_d + \theta_{brg} + v_{idb} \tag{3}$$

$I_{T_{idb}=T}$ 指的是 i 这个人在 17 岁时距离学区教育支出改革的相对年份的虚拟变量，即，如果他所在学区改革时他 18 岁，$T=-1$；如果他所在学区改革时他 37 岁，$T=-20$；如果他所在学区改革时他 16 岁，$T=1$。

$I_{Q_{ppe72,d}=Q_{ppe}}$ 指的是学区 d 在 72 年时的教育支出的所在四分位数的虚拟变量。

系数 $\alpha_{T,Q_{ppe}}$ 是改革引起的学校各学龄学生支出的变化。

方法二，

基于在其他州的观察上相似的学区所经历的估计剂量。通过观察相的在改革前具有相同的相对支出水平、相同的相对收入水平，并面临由于法院命令而产生的相同资金公式变化的学区。步骤：

1. 排除了包含学区 d 的州的所有数据。

2. 将每个法院命令的 SFR 与法院命令引起的改革类型相关联。使用上述五种关键改革类型：基础计划、支出限制、努力奖励计划、均等化计划和公平案例。对于每个法院命令，通过将法院命令后三年内的任何公式变化（可能有多于一个的变化）与该法院命令关联起来，来确定改革引入的资金公式类型。

3. 我们使用与每个学区相关联的县 1969 年的中等家庭收入作为对学区收入的度量。使用这个度量，计算了每个学区 1969 年在州内中等收入分布中的四分位数。

4. 在方程（3）中增加年份暴露指标，与 1969 年学区中等收入四分位数相互作用，每个都与五个改革类型中每一个是否作为法院命令的结果被引入相互作用。

5. 使用来自步骤 1 中排除的州的每个学区的估计方程 4，计算了那些在初始法院命令年份介于 10 岁和 15 岁之间的暴露队列的平均支出变化（相对于未暴露队列），这只基于 (a) 法院命令引入的改革类型/资金公式，(b) 改革前学区在州内每名学生支出分布中的四分位数，以及 (c) 改革前学区在州内中等家庭收入分布中的四分位数。

6. 对每个州重复步骤 1 到 5。

$$\ln\left(P\bar{P}E_{5-17}\right)_{idb} = \sum_{Q_{ppe}=1}^{4} \sum_{T=-20}^{20} \left(I_{T_{idb}=T} \times I_{Q_{ppe72,d}=Q_{ppe}}\right) \cdot \alpha_{T,Q_{ppe}}$$

$$+ \sum_{F=1}^{5} \sum_{Q_{\text{inc}}=1}^{4} \sum_{T=-20} \left(I_{T_{idb}=T} \times I_{Q_{\text{inc }69.d}=Q_{inc}} \times I_{F,d}\right) \cdot \alpha_{T,Q_{\text{inc}},F} + \Pi C_{idb} + +\theta_d + \theta_{brg} + v_{idb} \quad (4)$$

$Spend_d$ 指的是学区特定剂量预测器，是基于其他州的观察上相似的学区所经历的改革引起的学龄支出变化的估计，这些估计是基于那些在初始法院命令改革年份介于 10 岁和 15 岁之间的人，其中预测变化是基于与学区 d 具有相同的改革前相对收入水平、相同的改革前相对支出水平，并面临相同类型的改革的学区的经验。

因为不同类型的改革可能以不同的方式影响学区（Hoxby 2001），并且许多资金公式是基于学区的支出和收入水平（Card and Payne 1999），使用关于改革类型和学区收入水平的额外信息来预测剂量可能会导致预测剂量在具有相同改革前支出水平的学区之间有额外的变化。为了说明这种额外预测剂量变化的潜在好处，在图 II 中绘制了类似于方程（3）的事件研究估计，其中将四个 $Q_{ppe72,d}$,d 指示变量替换为一个单一的二元变量，如果 $Spend_d$ 大于 0 则等于 1，否则等于 0。这个指标表示，根据在其他州面临相同类型改革的观察上相似的学区的经验，一个学区是否预计会由于改革而经历支出增加。大约有三分之二的改革州的学区预计会由于法院命令的 SFRs 而增加支出。

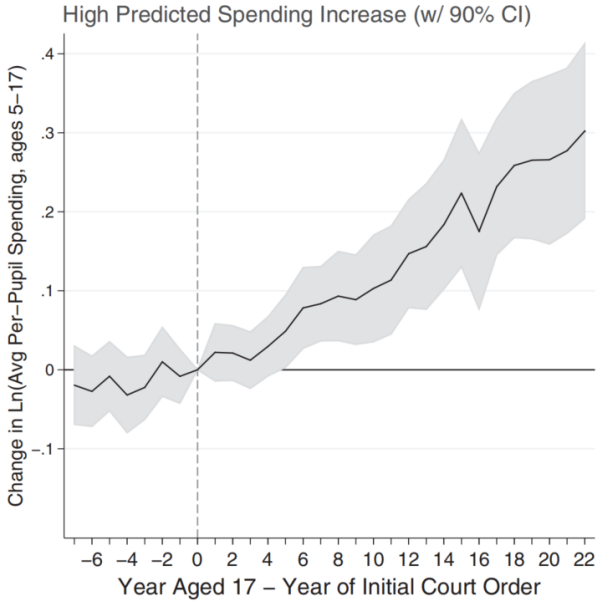


FIGURE II

Data: PSID geocode data (1968–2011), matched with childhood school and district characteristics. Analysis sample includes all PSID individuals born 1955–1985, followed into adulthood through 2011 ($N = 15{,}353$ individuals from 1,409 school districts [1,031 child counties, 50 states]). Sampling weights are used so that the results are nationally representative.

Models: The event study plot is based on indicator variables for the number of school-age years of exposure to a court-ordered SFR interacted with whether the district is predicted to experience a spending increase due to reforms ($Spend_{d.} > 0$) or not. Results are based on nonparametric event study models that include school district fixed effects, race × census division × birth cohort fixed effects, and additional controls.

主要结果

TABLE III

OLS VERSUS 2SLS ESTIMATES OF COURT-ORDERED SCHOOL FINANCE REFORM INDUCED EFFECTS OF PER PUPIL SPENDING ON EDUCATIONAL ATTAINMENT: BY CHILDHOOD POVERTY STATUS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent variable:							
	Years of Education				Prob(High School Graduate)			
	OLS	2SLS 1	2SLS 2		OLS	2SLS 1	2SLS 2	
$\text{Ln}(PPE_d)_{(\text{age 5-17})}$	−0.0763	3.1600***	3.1488***		0.0216	0.5910**	0.7053***	
	(0.1474)	(1.1327)	(0.7906)		(0.0278)	(0.2858)	(0.1436)	
$\text{Ln}(PPE_d)_{(\text{age 5-17})} \times$ low income				4.5899***				0.9878***
				(1.2072)				(0.2744)
$\text{Ln}(PPE_d)_{(\text{age 5-17})} \times$ Nonpoor				0.7156				0.2470
				(1.3193)				(0.1757)
Number of individuals	15,353	15,353	15,353	15,353	15,353	15,353	15,353	15,353
Number of childhood families	4,586	4,586	4,586	4,586	4,586	4,586	4,586	4,586
First-stage F-statistic	N/A	15.62	20.25	20.25	N/A	15.62	20.25	20.25

3 线性回归更多专题

3.1 稳健标准误

- 大数定律：随机变量 x_i 满足 $E(x_i)=\mu$，考虑随机变量序列 $x_1, x_2, \cdots$，则有

$$\bar{x}_n \to_p \mathbb{E}(\bar{x}_n) = \mu$$

- 中心极限定理：独立同分布的随机变量 x_i 满足 $E(x_i)=\mu, Var(x_i)=\Sigma$，则有

$$\sqrt{n}\left(\frac{\bar{x}_n - \mu}{\sqrt{\Sigma}}\right) \to_d N(0,1)$$

或可记作

$$\bar{x}_n \to_d N\left(\mathbb{E}(\bar{x}_n), \mathbf{Var}(\bar{x}_n)\right) = N\left(\mu, \frac{\Sigma}{n}\right)$$

- 独立同分布并不是中心极限定理成立所必需的的假设，只不过在 x_i 互相相关的情形中，$Var(\bar{x}_n)$ 将不但包含 x_i 的方差项，还包含其协方差项。

●考察线性回归（为论述的方便，y 和 x 均已作中心化处理）

$$y_i = \beta x_i + \varepsilon_i$$

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \beta + \frac{\frac{1}{n}\sum x_i \varepsilon_i}{\frac{1}{n}\sum x_i^2}$$

$$\frac{1}{n}\sum x_i^2 \to_p \mathbb{E}x_i^2$$

$$\frac{1}{n}\sum x_i \varepsilon_i \to_d N\left(0, \frac{1}{n^2}\mathrm{Var}\left(\sum x_i \varepsilon_i\right)\right)$$

$$b \to_d N\left(0, \frac{1}{n^2}\frac{\mathrm{Var}\left(\sum x_i \varepsilon_i\right)}{\left(\mathbb{E}x_i^2\right)^2}\right)$$

●可以看出，系数估计标准误 $s.e.(b)$ 的正确形式取决于 $x_i \varepsilon_i$ 之间是否相关。

- 当假定 $Var(\varepsilon_i|x_i)=\sigma_2$ 为常数，且 $Cov(\varepsilon_i,\varepsilon_j)=0$ 时，相应的标准误称为同方差标准误。**这个假设太强，通常不采用**。

$$\frac{1}{n^2}\text{Var}\left(\sum x_i\varepsilon_i\right)=\frac{1}{n}\sigma^2\mathbb{E}x_i^2$$

$$\text{Var}(b)=\frac{\sigma^2}{n\mathbb{E}x_i^2},\widehat{\text{Var}(b)}=\frac{s^2}{\sum x_i^2}$$

- 当假定 $\text{Var}\left(\varepsilon_i\mid x_i\right)$ 可以各不相同，且 $\text{Cov}\left(\varepsilon_i,\varepsilon_j\right)=0$ 时，相应的标准误称为**异方差稳健标准误**。

$$\frac{1}{n^2}\text{Var}\left(\sum x_i\varepsilon_i\right)=\frac{1}{n}\mathbb{E}\left(x_i^2\varepsilon_i^2\right)$$

$$\text{Var}(b)=\frac{1}{n}\frac{\mathbb{E}\left(x_i^2\varepsilon_i^2\right)}{\left(\mathbb{E}x_i^2\right)^2},\widehat{\text{Var}(b)}=\frac{\sum x_i^2e_i^2}{\left(\sum x_i^2\right)^2}$$

- 当 $\text{Cov}\left(\varepsilon_i,\varepsilon_j\right)\neq 0$ 时，需要进一步假定扰动项在同一类内可能相关，但在不同类间不相关，此时称之为**聚类稳健标准误**。

- 大样本理论在类的数目足够大时成立。

- 聚类稳健标准误通常比异方差稳健标准误更大，采用聚类稳健标准误之后，系数估计的统计显著性更难得到。

- 研究者实际无从知道观测个体在哪个层面上互相相关，将类定义得过小，类间不相关的条件不容易满足，相应的标准误估计将不是真实的标准误的一致估计；将类定义得过大，标准误估计的一致性虽然得以保证，但毕竟标准误的形式只有当类的数目足够大时才成立，基于此的标准误估计可能对系数估计量的有限样本标准误近似得很差 (有限样本偏误大)。

- 异方差稳健标准误可以看作每个个体独自为一类的“聚类稳健标准误”。合适的聚类层级需要依研究情境和数据特征而异。一个经验法则是：当核心解释变量的数据层级高于被解释变量时，标准误应聚类到核心解释变量所在层级。

●注意：不要混淆固定效应的类别划分和标准误的聚类层级，这是完全不同的两件事!

- 固定效应的类别划分得越细致，结果越稳健。这个稳健性是针对因果识别而言，因为固定效应控制下的识别假设是，固定效应所定义的类别内部，核心解释变量和扰动项不相关，所以固定效应的类别划分得越细致，识别假设就越容易成立。
- 相反，标准误的聚类层级越高，结果越稳健。这个稳健性是针对统计推断而言——稳健标准误是否可信，即其背后关于观测个体的相关性假设是否可信。

●考虑 ε_{icp}，i 表示企业，c 表示城市，p 表示省份，如果标准误聚类到城市层面，所隐含的对扰动项方差协方差结构的假设是

$$
\begin{pmatrix}
 & \varepsilon_{111} & \varepsilon_{211} & \varepsilon_{321} & \varepsilon_{421} & \varepsilon_{532} & \varepsilon_{632} & \varepsilon_{742} & \varepsilon_{842} \\
\varepsilon_{111} & \times & \times & & & & & & \\
\varepsilon_{211} & \times & \times & & & & & & \\
\varepsilon_{321} & & & \times & \times & & & & \\
\varepsilon_{421} & & & \times & \times & & & & \\
\varepsilon_{532} & & & & & \times & \times & & \\
\varepsilon_{632} & & & & & \times & \times & & \\
\varepsilon_{742} & & & & & & & \times & \times \\
\varepsilon_{842} & & & & & & & \times & \times
\end{pmatrix}
$$

●如果标准误聚类到省份层面，所隐含的对扰动项方差协方差结构的假设是

$$
\begin{pmatrix}
 & \varepsilon_{111} & \varepsilon_{211} & \varepsilon_{321} & \varepsilon_{421} & \varepsilon_{532} & \varepsilon_{632} & \varepsilon_{742} & \varepsilon_{842} \\
\varepsilon_{111} & \times & \times & \times & \times & & & & \\
\varepsilon_{211} & \times & \times & \times & \times & & & & \\
\varepsilon_{321} & \times & \times & \times & \times & & & & \\
\varepsilon_{421} & \times & \times & \times & \times & & & & \\
\varepsilon_{532} & & & & & \times & \times & \times & \times \\
\varepsilon_{632} & & & & & \times & \times & \times & \times \\
\varepsilon_{742} & & & & & \times & \times & \times & \times \\
\varepsilon_{842} & & & & & \times & \times & \times & \times
\end{pmatrix}
$$

可见，聚类层级越高，所隐含的假设越弱，标准误估计更稳健。

- 考虑数据 ε_{icd}，其中 c 表示城市，d 表示行业，此时同一城市内部不同个体的扰动项之间可能相关，同一行业内部不同个体的扰动项之间也可能相关，此时有必要使用双向聚类 (two-way clustering) 稳健标准误，其隐含的假设是

$$
\begin{pmatrix}
 & \varepsilon_{111} & \varepsilon_{211} & \varepsilon_{312} & \varepsilon_{412} & \varepsilon_{521} & \varepsilon_{621} & \varepsilon_{722} & \varepsilon_{822} \\
\varepsilon_{111} & \times & \times & \times & \times & \times & \times & & \\
\varepsilon_{211} & \times & \times & \times & \times & \times & \times & & \\
\varepsilon_{312} & \times & \times & \times & \times & & & \times & \times \\
\varepsilon_{412} & \times & \times & \times & \times & & & \times & \times \\
\varepsilon_{521} & \times & \times & & & \times & \times & \times & \times \\
\varepsilon_{621} & \times & \times & & & \times & \times & \times & \times \\
\varepsilon_{722} & & & \times & \times & \times & \times & \times & \times \\
\varepsilon_{822} & & & \times & \times & \times & \times & \times & \times
\end{pmatrix}
$$

  要注意其与“聚类到城市 × 行业层面的稳健标准误”的区别。

- 聚类稳健标准误的构造思路是定义“类”，类内相关，类外不相关。另一种稳健标准误的构造思路是定义“距离”，两个观测单位的相关性随距离衰减。这方面最典型的例子是适用于时间序列数据的 Newey-West 标准误，假定两期之间的自相关性和间隔的期数负相关。

- Conley (1999, JoE) 类似地构造了反映横截面相关性 (cross-sectional dependence) 的稳健标准误。根据研究情境不同，两个横截面观测单位之间的相关性可能与地理距离有关，也可能与经济距离有关。

3.2 中介效应

- **什么是中介效应 (mediation)**？D 可能直接影响 Y，也可能影响 X 继而通过 X 影响 Y，这时称 X 是 D 影响 Y 的中介变量 (mediator)。如果 D 对 Y 没有直接影响，称为完全中介，如果 D 对 Y 有直接影响，称为不完全中介。

$$
\begin{aligned}
Y &= \alpha_0 + \alpha_1 D + \varepsilon_{Y_1} \\
Y &= \beta_0 + \beta_1 D + \beta_2 X + \varepsilon_{Y_2} \\
X &= \gamma_0 + \gamma_1 D + \varepsilon_X
\end{aligned}
$$

因此

$$\alpha_1 = \beta_1 + \beta_2 \cdot \gamma_1$$

– 中介效应分析在心理学、流行病学、政治学、社会学、组织行为学等领域应用非常广泛。尤其对社会心理学研究而言几乎是必不可少的操作。但**主要应用于随机实验研究**。

– 中介效应检验：D 对 Y 的总体效应 α_1，和 D 对 Y 的直接效应 β_1，它们的统计显著性可以方便地得到。但中介效应分析的重点是间接效应 $\beta_2 \cdot \gamma_1$，或等价地，$\alpha_1 - \beta_1$，它的统计显著性不能只看单个系数 β_2 或 γ_1 的统计显著性，或肉眼比较 α_1 和 β_1 的相对大小，但有一些正式的检验手段，比如 Sobel 检验、自助法置信区间等。

– 为了和调节效应区别，此处把 X 称作 D 影响 Y 的渠道 (channel)。

•中介效应检验和因果推断的关系？

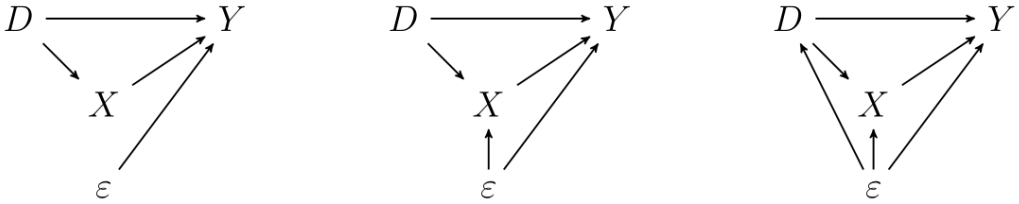
– 中介效应检验在中文经济学界中大有攻城略地之势，这得益于队伍中出现的一些“食不精、脍不细”的技术至上主义的作者和审稿人，以及这些“劣币”对“良币”的绑架。

– 他们没有看到，与此同时，国际学术界却正在对不严谨的中介效应分析的泛滥进行反思 (Bullock et al, 2010, Journal of Personality and Social Psychology)。

– 假定 D 是一种随机干预，

$$\hat{\beta}_2^{OLS} \to_p \beta_2 + \frac{\text{Cov}\left(\varepsilon_{Y_2}, \varepsilon_X\right)}{\text{Var}(\varepsilon_X)}$$
$$\hat{\beta}_1^{OLS} \to_p \beta_1 - \gamma_1 \frac{\text{Cov}\left(\varepsilon_{Y_2}, \varepsilon_X\right)}{\text{Var}(\varepsilon_X)}$$

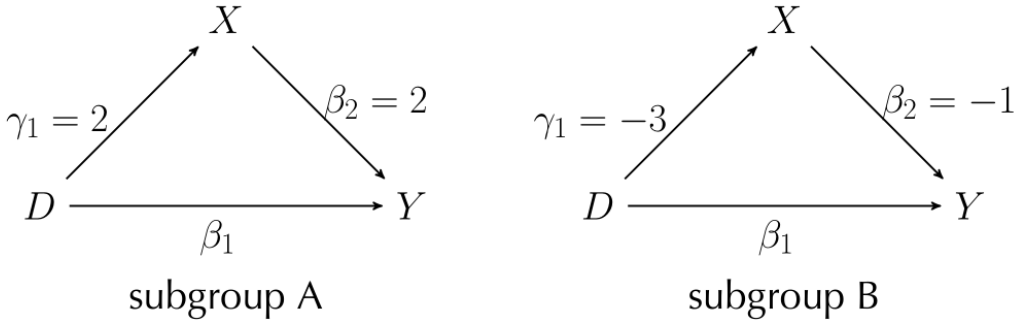
- 换言之，如果 $Cov(\varepsilon_{Y_2}, \varepsilon_X) \neq 0$，则 β_1 和 β_2 都无法通过 OLS 得到一致估计。但在实际研究中，这是很经常发生的情形——只要存在同时影响 Y 和 X 的不可观测因素。即使在随机实验中，研究者往往只能对 D 进行随机干预，却无法对 X 进行随机干预。更遑论观测性研究。



随机实验 I　　随机实验 II　　观测性研究 III

- 这其实是所谓“坏控制”的典型例子：当 D 是随机干预的，不加任何控制原本可以得到 D 对 Y 的因果效应，控制 X 则适得其反，无法一致估计因果效应。

- 即使在“随机实验 I”中（中介变量 X 也是随机干预的），如果存在因果效应的异质性，也很难估计间接效应。



- 组 A 和组 B 的间接效应均为 4。若两组规模相当，则 $\hat{\gamma}_1 = -.5$，$\hat{\beta}_2 = .5$，间接效应的 OLS 估计为 $\hat{\beta}_2 \cdot \hat{\gamma}_1 = -.25$。

- 一些中介效应研究举例
  - Newheiser and Barreto (2014): Concealing a stigmatized identity $(D) \rightarrow$ perceived self-disclosure $(X) \rightarrow$ evaluation of the interaction (Y)
  - Dubois-Comtois et al (2013): maternal psychosocial distress $(D) \rightarrow$ quality of mother-child interactions $(X) \rightarrow$ child-reported behavior problems (Y)
  - An et al (2016): natural elements exposure $(D) \rightarrow$ depressed mood $(X) \rightarrow$ job satisfaction (Y)
  - de Moor (2015): politicians' perceived willingness and ability to act $(D) \rightarrow$ perceived effectiveness of political participation $(X) \rightarrow$ political participation (Y)
  - Goodboy et al (2016): school bus driver bullying $(D) \rightarrow$ job stress $(X) \rightarrow$ anxious driving (Y)

- 由此可见，中介效应研究，确实存在着**“普遍的掩耳盗铃”**的症状！请注意，不是说不应该研究中介效应，而是说中介效应很难研究。但目前这种粗糙的研究手段竟然在自诩为因果识别最考究的中文经济学界流行，不得不说是十足吊诡和令人遗憾的事情。但更恐怖的是，错误地进行中介效应分析还不是中文经济学界所实际出现的症状的全部。

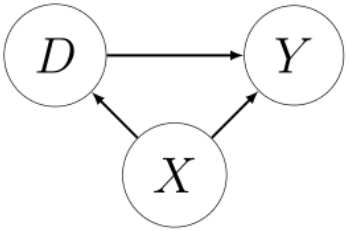
- 中介效应检验的目的是考察 D 对 Y 的因果关系是否通过 X 这个渠道发生作用，这是对 D 对 Y 的因果关系研究的一个扩展，但它**不能用来论证 D 对 Y 的因果关系！**换句话说，在 Y 对 D 的回归中加入了中介变量 X，不论是否发现 D 的系数发生了变化，这个结果本身都无法使得 D 对 Y 的因果关系变得更加可信。所以，中介效应检验（如果做对了的话）有它独立存在的价值，但**认为中介效应检验是因果关系的一种稳健性检验，这是错上加错！**

- **什么时候应该做中介效应检验？如何做？**

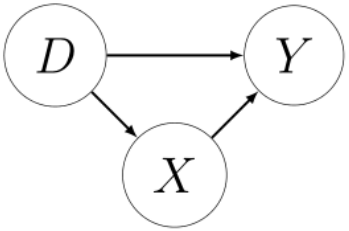
  - 中介效应检验的适用性前提是，识别 D 对 X 和 Y 的因果关系比较容易，以及，X 对 Y 的因果关系在理论上比较直观（逻辑上和时空关系上都比较接近）且识别比较容易。而基于第三方数据的观测性研究，真实的数据生成过程纷繁复杂，找到合适的研究情境研究 D 对 Y 的因果关系已属不易，研究中介效应更是困难，这就是为什么中介效应检验历来在经济学文献中很少见到的原因。

– 试想，一项严肃的中介效应研究，除了要正式探讨 D 对 Y 的因果关系之外，还要正式探讨 D 对 X 的因果关系，以及 X 对 Y 的因果关系，这不是一篇文章的工作量，而是三篇文章的工作量！而且，即使 X 成了一个合格的中介变量，如果发现 D 对 Y 的影响除了通过 X 发生的间接影响之外，还存在直接影响，即 X 是不完全中介，这是好事么？不是好事！因为这暴露出你对 D 如何影响 Y 有相当一部分是无知的。

- 从技术上说，研究中介效应就是在回归中加入 X，但在观测性研究中，加入 X 的目的主要是为了解决 D 的选择性，X 被称作控制变量。一会儿把控制 X 当作一种因果效应的识别策略，一会儿又把它当作因果效应的作用渠道，这只能反映出研究者逻辑的混乱。相反，在随机实验研究中，不需要额外控制 X 来帮助识别，所以讨论中介效应成为可能，但如前所述，即使如此，讨论也很困难。



作为控制变量的 X



作为中介变量的 X

- **不做中介效应检验，不是说不研究中介效应**。经济学研究中讨论影响渠道时，一种常见的做法是找一个或一些 X，这些 X 和 Y 的关系不言自明，以至于不必采用正式的因果推断方法来研究两者之间的关系，然后只看 D 对 X 的影响（即把 X 回归在 D 上），仅此而已，而避免去正式区分在间接效应之外是否还有无法解释的直接效应。这样的例子比比皆是，比如 Dell (2010, ECMA) 发现 16-19 世纪秘鲁的 mita 徭役制度 (D) 导致当代居民家庭消费降低以及儿童发育迟缓 (Y)。紧随其后的文章第 4 节 “持续影响的作用渠道 (Channles of Persistence)”，将被解释变量依次替换为土地所有权 (land tenure)、教育和道路等公共品供给、劳动供给和市场参与等消费的直接决定因素 (X)，即视为达到了检验渠道的目的，至于这些 X 如何影响 Y 以及 D 是否还会在除了 X 之外直接影响 Y 就不再着墨了。

- **有些看起来像中介效应检验的做法其实在做什么？**
  - 有些文献会先做 Y 对 D 的回归，然后做 Y 对 D 和 X 的回归，然后发现 D 依然显著（用中介效应检验的术语，意味着 D 对 Y 有直接影响），以此来论证 D 对 Y 的因果关系，这应该如何理解？
  - 这种做法尽管跟中介效应检验很类似，但研究设计的出发点是截然不同的。再看 Nunn and Wantchekon (2011)。

*奴隶贸易影响的可能是人们的内心行为规范：不愿意信任他人（不信任成为最优的经验决策法则）。
*但也有可能是外部环境——国家、制度和法律结构的衰败：他人不值得信任。
*检验方法：将竞争性解释在回归中控制起来，比较控制前后的系数变化，看核心解释变量的效果是否完全被竞争性解释所吸收。
*检验 1：控制政府质量（此时用对当地政府的信任作为被解释变量）
  · 主观感受：称职、腐败、倾听。
  · 客观指标：是否提供各类公共服务。
*检验 2：控制所在地其他种族历史上受奴隶贸易的影响，从而间接控制周围他人的可信任程度（此时用族群间信任作为被解释变量）。
*检验 3：控制所在地历史上所居种族受奴隶贸易的影响，从而间接控制所在地外部环境的可信任程度。

– 总结一下，这类做法首先对 D 如何影响 Y 有一个基准理论，然后再有一个竞争性理论，然后构造控制变量 X 放入主回归，发现竞争性理论不能完全解释 D 和 Y 的相关性（存在“直接影响”），由此说明基准理论很可能是对的。在这里 D 会影响 X 么？不会！关于 D 的故事和关于 X 的故事是两个互相竞争的故事。这时，Y 同时对 D 和 X 的回归被形象地称作“horse race”。这种“看起来像中介效应检验的做法”，实际上是用来强化因果关系论证。

3.3 调节效应

- **什么是调节效应 (moderation)？**
  - D 对 Y 的边际影响随 X 的变化而变化，称 X 为 D 影响 Y 的调节变量。调节效应模型就是交互项模型：将 Y 回归在 D、X 和 $D \cdot X$ 上，看交互项系数的显著性。
  - 为了和中介效应区别，此处把 X 称作 D 影响 Y 的机制 (mechanism)。

•调节效应检验和异质性分析、因果推断的关系？

– 调节效应就是异质性分析。

– 最简单的理解，当调节变量 X 是 $0-1$ 变量时，相当于把全样本分为 $X=0$ 和 $X=1$ 两个组，交互项 $D\cdot X$ 的系数就是分组进行的 Y 对 D 的回归中 D 的系数的差异。当 X 是连续变量时，这种理解的本质并没有变化。

– 大家很习惯做异质性分析，但很少问为什么要做异质性分析。也许异质性本身是重要的，比如我想研究教育回报率，除了得到一个全样本的点估计之外，还想看女性的教育回报率是不是显著高于或者低于男性，由此得出政策含义。

- 将异质性作为主要卖点的文章通常具有一个特点，全样本的显著性只在一个子样本中存在，在另一个子样本中不存在，这样的结论才好卖。比如有人研究小额信贷 (D) 对家庭财务状况 (Y) 的影响，发现总体上两者呈现出反直觉的负相关——借了钱更容易陷入窘境。但对家庭按理财素养 (X) 进行分组以后发现，这个负面效应只在理财素养低的家庭中存在，在理财素养高的家庭中不存在，这样极端的结果往往才有意思。

- 但在中文世界里，大家往往只是为了凑够版面，在主回归之外出于某种八股本能，按地区、按规模、按所有制做一些异质性分析，反正这样的分析很安全，有没有异质性都有话可说，却忘了文章的重点是因果识别，文章的每一字每一句都应该为这个大目标服务。而异质性分析更重要的作用正是**通过分析因果关系的作用机制从而强化因果关系论证！**

- 我们发现了 D 与 Y 的相关性，并且想要主张 D 是 Y 的原因，可以通过检验 D 影响 Y 的某个具体机制来对从 D 到 Y 的因果关系进行论证，论证的逻辑如下：

  1.提出一个 D 影响 Y 的理论 T。根据这个理论，D 通过某个机制 M 影响 Y ，并且可以识别出 M 在某些子总体 (subpopulation) 中存在，在另一些子总体中不存在，令 $X = 1$ 表示存在 M，$X = 0$ 表示不存在 M。

  2.在 $X = 1$ 组，D 与 Y 的相关性继续存在，而在 $X = 0$ 组，D 与 Y 的相关性不复存在。

  3.可能导致 D 与 Y 相关的竞争性解释包括 Y 影响 D 的反向因果理论 R，或者有混淆因素同时影响 D 和 Y 的遗漏变量理论 C。**如果无法想象理论 R 或理论 C 发挥作用的机制在 $X = 1$ 和 $X = 0$ 组存在显著差异**，则很可能理论 R 或理论 C 不成立。否则，我们应该在 $X = 0$ 组也观察到 D 与 Y 的相关性。

  4.有时候，两个组中 D 与 Y 的相关性都存在，但在 $X = 1$ 组这种相关性更强，表现在对 Y 的回归中，D 的系数估计在 $X = 1$ 组更大，且组间差异在统计上显著。这时我们至少可以说，D 与 Y 的相关性不全是理论 R 或理论 C 所带来的，否则这种相关性应该在 $X = 1$ 和 $X = 0$ 组无差异。

- 在 Rajan and Zingales (1998) 中，金融发展水平 (D) 与经济增长 (Y) 强相关，文章想说金融发展是经济增长的原因，并检验了金融发展通过缓解企业的外部融资约束 (M) 从而促进了企业成长这一理论 (T)。文章将行业分成两组，一组是外部融资依存度 (X) 较高的行业（存在 M），另一组是外部融资依存度较低的行业（不存在 M），发现在对行业增长的回归中，金融发展水平与外部融资依存度的交互项显著，表明金

融发展水平与行业增长之间的相关性在外部融资依存度不同的组间存在显著差异。

- 金融与增长之间的相关性可能是因为增长影响金融，高增长引发了融资需求从而导致金融市场发展（理论 R），也可能是因为某个混淆因素（例如节俭传统）同时影响金融发展和经济增长（理论 C），那么除非理论 R 和理论 C 在外部融资依存度不同的组间发挥作用的程度不同，否则就证明了理论 T。

- 好的 X 本身应该比较稳定，或者其变动是外生的，尤其是 X 不受 D 或 Y 影响。[3]

[3]坏的 X 相当于在双重差分研究中，处理组和控制组的构成一直在变化 (compositional change)，并且导致这种变化的因素和 Y 相关（隐藏在扰动项之中），那么处理组和控制组的平行趋势假定就不再成立。

- 此前我们处理内生性/选择性的思路一直聚焦在寻找合适的 conditioning variables 和 conditioning strategy，即找到导致内生性/选择性的原因，然后正式地刻画它、测量它、控制它。异质性分析则提供了另一种处理内生性/选择性的思路，即尝试挖掘因果模型的可验证含义——被解释变量和核心解释变量之间更丰富的相关性，如果这种相关性是其它因果故事所不能解释的，那么即使此时内生性/选择性或许仍然存在，但至少证明我们所感兴趣的因果关系是存在的，否则这种更丰富的相关性不会出现。

- 在超级明星效应的例子中，无法完全控制“赛事难度”这一导致核心解释变量“伍兹是否参赛”存在内生性/选择性的因素（尽管已经最大限度地控制了赛事级别、场地质量、奖金总额等），作者转而去挖掘更丰富的相关性：伍兹是否参赛导致的比赛杆数差异对于高水平选手而言是否更大？这个事实可以被超级明星效应所解释，但不能被伍兹参加的都是高难度赛事所解释，因果论证的目的就达到了。

- 同样地，在金融促进增长的例子中，无法完全控制反向因果或第三方混淆因素（尽管已经控制了不变的行业特征和国家特征，但仍然可能存在同时随行业和国家而变的选择性变量），作者转而去挖掘更丰富的相关性：金融和增长的正相关性在外部融资依存度更高的行业是否更强？这个事实可以被金融通过缓解企业外部融资约束从而促进增长的故事所解释，但不能被其它故事合理解释，因果论证的目的就达到了。

- 这两个例子启发我们，异质性分析是一种很重要的因果论证手段，但在使用这种手段之前，首先要发展出一个说得通的理论：如果超级明星效应这个故事成立，那么就应该看到不同人的效应大小不同（因为激励强度不同）；如果金融促进增长这个故事成立，那么就应该看到不同行业的效应大小不同（因为对金融的需求不同），如此之类。然后再构建相应的交互项模型去验证这个理论。这就是因果推断理论先行 (theory driven) 的含义。

被解释变量为对数时，系数 β 的经济解释：

$$\begin{aligned} \log(y+\Delta y)-\log y &= \beta\Delta x \\ \log\left(\frac{y+\Delta y}{y}\right) &= \beta\Delta x \end{aligned}$$

当 $\Delta y \to 0$ 时，根据泰勒公式，

$$\log\left(\frac{y+\Delta y}{y}\right) = \log\left(1+\frac{\Delta y}{y}\right) \simeq \left(\frac{\Delta y}{y}\right)$$

所以才有 x 增加 1，对应 y 增加 $100\times\beta\%$。
如果被解释变量是 $log(1+y)$，则推导结果为：

$$\log\left(\frac{y+1+\Delta y}{y+1}\right) \simeq \left(\frac{\Delta y}{y+1}\right)$$

所以 β 的经济含义应解释成 $y+1$ 增加了 $100\times\beta\%$。

我们这个研究里，y 的平均值为 0.136，所以 $y+1$ 增加的 $100\times\beta\%$，不能近似等于 y 增加了 $100\times\beta\%$。