

断点回归

慧航

2025年9月



断点回归

断点回归 (Regression discontinuity) 用于政策变量仅仅取决于一个连续变量 x (running variable), 且在某个点 $x = c$ 处, 参与概率有跳跃的情况。

① Sharp RD:

$$w_i = \mathbb{1}(x_i \geq c)$$

② Fuzzy RD

$$\lim_{x \downarrow c} P(w_i | x_i = c) \neq \lim_{x \uparrow c} P(w_i | x_i = c)$$

- 优点: 识别干净
- 缺点: 外部有效性
- 例子: 退休前后的消费情况

常见的断点

- 地理断点：秦岭淮河（Chen等，2013）、南北越（Dell、Lane和Querubin，2018）
- 分数断点：研发（Bronzini和Iachini，2014）
- 日期断点：瑞典2012年1月“double days”（Dahl、Løken和Mogstad，2014；Persson和Rossin-Slater，2024）；
- 政策断点：外包总收益比例（Li、Liu和Sun，2021）、家庭收入（贾男和王赫，2022）

Sharp RD

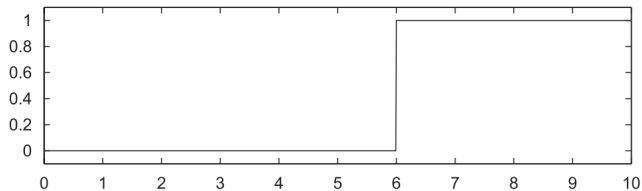


Fig. 1. Assignment probabilities (SRD).

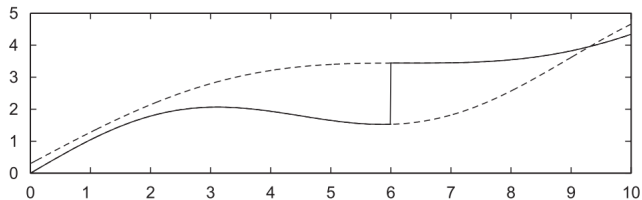


Fig. 2. Potential and observed outcome regression functions.

Sharp RD

政策效应:

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}(y_i | x_i = c) - \lim_{x \uparrow c} \mathbb{E}(y_i | x_i = c)$$

估计:

- ① 参数方法×
- ② 非参数方法:
 - ① 在 $x = c$ 两边取两个小邻域计算均值
 - ② Local constant
 - ③ Local polynomial (Local linear): 使用多项式在 $x = c$ 两边 $((c - h, c + h))$ 对 outcome 进行拟合

Sharp RD

令 $s_i = \mathbb{1}\{x_i \geq c\}$, 对于Sharp RD, 即 $w_i = s_i$, 对于设定:

$$\begin{aligned} y_i &= \alpha + \tau \cdot s_i + s_i \cdot f_r(x_i - c) + (1 - s_i) \cdot f_l(x_i - c) + u_i \\ &= \alpha + \tau \cdot s_i + s_i \cdot f_r(x_i - c) + [1 - s_i] \cdot f_l(x_i - c) + u_i \end{aligned}$$

- 当 $x_i < c$ 时: $y_i = \alpha + f_l(x_i - c) + u_i$
 - 若 $x_i \rightarrow c$, 由于 $f_l(0) = 0$, 从而在 $x_i = c$ 处的截距项为 α
- 当 $x_i > c$ 时: $y_i = \alpha + \tau + f_r(x_i - c) + u_i$
 - 若 $x_i \rightarrow c$, 由于 $f_r(0) = 0$, 从而在 $x_i = c$ 处的截距项为 $\alpha + \tau$
- 因而 τ 就是在 $x = c$ 处的jump, 即此处的处理效应

Fuzzy RD

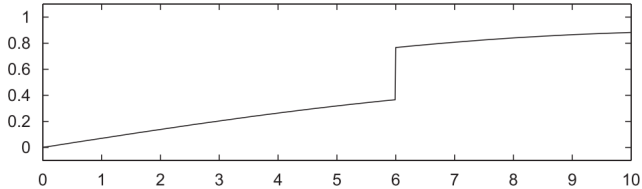


Fig. 3. Assignment probabilities (FRD).

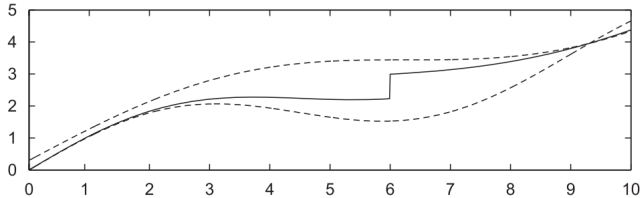


Fig. 4. Potential and observed outcome regression (FRD).

Fuzzy RD

对于Fuzzy RD, $s_i \neq w_i$ 。思想: s_i 作为 w_i 的工具变量:

- Intention-to-treat (ITT):

$$y_i = \alpha_1 + \tau_{\text{ITT}} \cdot s_i + s_i \cdot f_r(x_i - c) + [1 - s_i] \cdot f_l(x_i - c) + u_i$$

- 政策概率:

$$w_i = \alpha_2 + \delta \cdot s_i + s_i \cdot g_r(x_i - c) + [1 - s_i] \cdot g_l(x_i - c) + v_i$$

- 结构方程:

$$y_i = \alpha + \tau \cdot w_i + s_i \cdot h_r(x_i - c) + [1 - s_i] \cdot h_l(x_i - c) + \epsilon_i$$

Fuzzy RD

估计:

- ① 根据以上可知:

$$\tau = \frac{\tau_{ITT}}{\delta}$$

其中 τ_{ITT} 和 δ 都可以通过Local polynomial估计得到

- ② 2SLS: 直接进行工具变量回归
 - 本质上RD的识别是一个LATE
 - 而LATE使用2SLS估计



断点回归

- Stata: rdrobust
- 多项式阶数选取
 - cross validation: 对临近断点的观测进行预测
 - 一般不宜高过2, 过高的多项式阶数导致不稳定
- 窗框选取
 - cross validation: 对临近断点的观测进行预测
 - Calonico, Cattaneo and Titiunik(2014)



断点回归

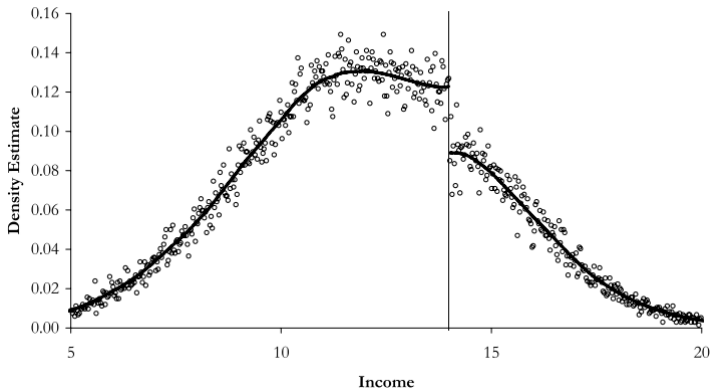
一般步骤：

- ① 画图
- ② 检验manipulation
- ③ 推断
- ④ 稳健性检验



断点回归

检验manipulation: McCrary(2008)



Stata: rddensity

实例：意大利R&D

Bronzini and Iachini(2012) 研究了研发补贴对于企业的影响。背景：

- 2003年，意大利Emilia-Romagna政府为了激励研发，实施了针对研发的补贴计划
- 能够被补贴的资质：通过打分确定，大于75分则可以接受补贴
- 断点：分数

聚束

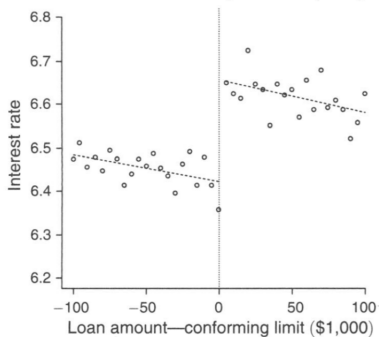
如果存在操纵，是不是完全没有办法呢？

- 操纵本身也是理性选择的结果
- 操纵会带来密度函数在断点处的不连续
- 一种特殊的“操纵”：聚束（bunching）可能帮助我们识别一些关键参数
- 两种不同的聚束：
 - kink（拐点、弯折）
 - norch（切口？）

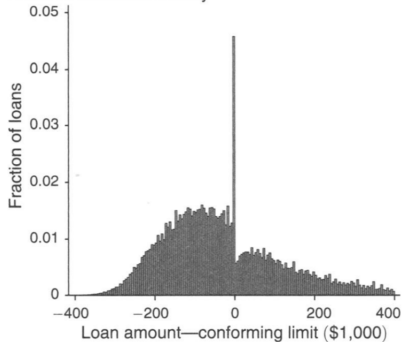


数据中的聚类

Panel A. Mean interest rate by loan size (2006)



Panel B. Loan size density



Kinks

Kink最常见的例子：累进税制

- 比如对于个人所得税，如果税前收入 z 在 z^* 的左右两边税率不同：

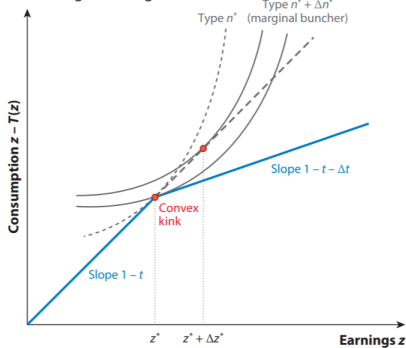
$$T(z) = \begin{cases} tz & z < z^* \\ tz + \Delta t(z - z^*) & z \geq z^* \end{cases}$$

即当收入在 z^* 以下时，边际税率为 t ，当收入超过 z^* 时，边际税率为 $t + \Delta t$

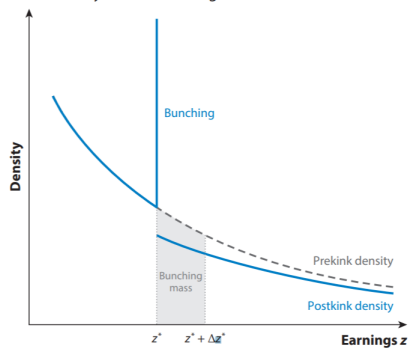
- 收入稍微超过 z^* 时，边际税率增加，激励不同，会导致一部分人（比如本来收入应该在 $(z^*, z^* + \Delta z^*)$ 的人）选择赚取 z^* 即可
- 效用最大化选择如下图
 - 会导致 z^* 左侧出现bunching现象
 - 右侧密度仍然不等于0，是由于右侧激励低，本来部分应该在更右侧的向左移动了

Kinks

a Budget set diagram



b Density distribution diagram



弹性

- Saez(2010)指出, 对于税前收入为 $z^* + \Delta z^*$ 的“边际人”(marginal buncher), 其效用函数曲线应该相切于 z^* 后的折线
- 收入的补偿弹性:

$$e = \frac{\frac{\Delta z^*}{z^*}}{\frac{\Delta t}{1-t}}$$

其中: $\Delta t / (1 - t)$ 是额外的税收百分比; $\Delta z^* / z^*$ 是收入增加百分比, “边际人”对于这两者无差异

Bunching的大小

- 如果假设税前收入有一个分布： $h(z)$
- 那么，选择 z^* 的人数：

$$B = \int_{z^*}^{z^* + \Delta z^*} h(z) dz \approx h(z^*) \Delta z^*$$

- 如果考虑 e 存在异质性，那么考虑 z, e 的联合分布函数： $h(z, e)$ ，那么：

$$B = \int_e \int_{z^*}^{z^* + \Delta z_e^*} h(z, e) dz de \approx \int_e h(z^*, e) \Delta z_e^* de = h(z^*) \mathbb{E}(\Delta z_e^*)$$

其中 $h(z^*) = \int_e h(z, e) de$ 。

- 只要计算出 B 和 $h(z^*)$ ，就可以计算出 $\Delta z^* (\mathbb{E}(\Delta z_e^*))$

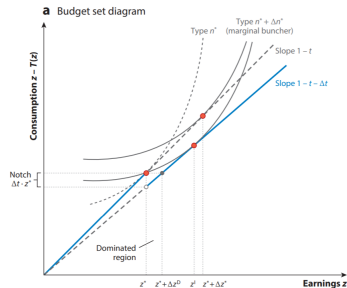
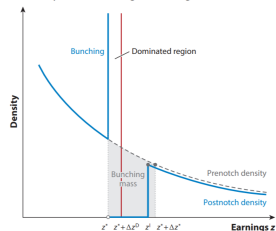
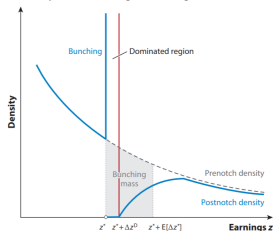
Norches

- norch与kink类似，区别在于：
 - kink在拐点处是连续的，norch在拐点处是离散的
 - norch造成了平均税率的不连续变化
 - norch可能会造成一部分被严格占优的区域
- 比如，以下这种非累进的、突然变化的税率：

$$T(z) = \begin{cases} tz & z < z^* \\ tz + \Delta tz & z \geq z^* \end{cases}$$

- 例子：Kleven和Waseem(2013)

Norches

**b** Density distribution diagram: homogenous elasticities**c** Density distribution diagram: heterogenous elasticities

Norches

- 以上区域中, $(z^*, z^* + \Delta z^D)$ 理论上应该是没有样本的——被占优的区域
- 然而, 由于优化摩擦 (optimization frictions) 等的存在, 实际数据中仍然可能会有少量样本
- 优化摩擦: 经济个体没能达到最优的选择 (未能达到kink或者norch的地方)
 - 调整成本 (Chetty等人, 2011)
 - 注意力成本
- 此外, 行为经济学中的参照点 (reference point) 也可能会混淆bunching的估计
 - 比如退休年龄 (Seibold, 2020)

Bunching的估计

- 在公式

$$B = \int_{z^*}^{z^* + \Delta z^*} h(z) dz \approx h(z^*) \Delta z^*$$

中，需要估计两个部分：

- Bunching B
- 反事实的密度函数 $h(z^*)$
- 实际上，只要估计出 $h(z)$ ，以上两个问题就都解决了
 - Bunching B 无非就是实际数据频率高出密度函数的部分

估计方法

不失一般性，我们假设Bunching的点 $z^* = 0$

- ① 类似于直方图，将数据按照 z 的取值进行分组，假设组中值为 $m_j, j = -J, \dots, L, \dots, 0, \dots, U, \dots, J$
 - 比如，按照1的组距进行分组：
组： $\dots (-4, -3], (-3, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], \dots$ 对应的 组中值为 $-3.5, -2.5, -1.5, -0.5, 0.5, 1.5$
- ② 确定排除区域 (excluded region) : $m_L < 0 < m_U$ ，即收到bunching影响的区域
- ③ 计算落入到每个组别的样本个数 n_j
- ④ 使用如下多项式回归拟合密度函数：

$$n_j = \sum_{k=0}^p \beta_k m_j^k + \sum_{t=L}^U \gamma_t 1 \{m_t = m_j\} + \epsilon_j$$

其中 p 为多项式阶数

估计方法

- 根据以上方法, γ_t 就度量了反事实的密度函数, 从而:

$$\hat{n}_j = \sum_{k=0}^p \hat{\beta}_k m_j^k$$

- 从而bunching:

$$\hat{B} = \sum_{t=L}^0 \hat{\gamma}_t$$

- 而0右侧损失的密度为:

$$\hat{M} = \sum_{t=0}^U \hat{\gamma}_t$$

- 0处的密度函数:

$$\hat{h}(z^*) = \hat{h}(0) = \frac{\sum_{j=L}^0 \hat{n}_j}{\frac{m_0 - m_L}{L}}$$

超参数和标准误

需要决定的超参数:

- 多项式阶数 p
- 组距
- 排除区域: L 和 U
 - 固定一个 L , 通过迭代的方法确定 U , 使得 $|\hat{B} - \hat{M}|$ 最小

标准误: 使用bootstrap

例子: The interest rate elasticity of mortgage demand: evidence from bunching at the conforming loan limit