1917–2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学 100周年校庆

《经济学与金融学实证方法》

断点回归

郭峰
上海财经大学公共经济与管理学院

本讲主要内容

- 断点回归基本原理
- 断点回归识别条件
- 断点回归图形分析
- 断点回归关键问题

- 断点回归基本原理
- 断点回归识别条件
- 断点回归图形分析
- 断点回归关键问题

RD方法的原理

- 随机实验是因果识别理想模式，但观测性研究有时候也可以模拟随机实验场景

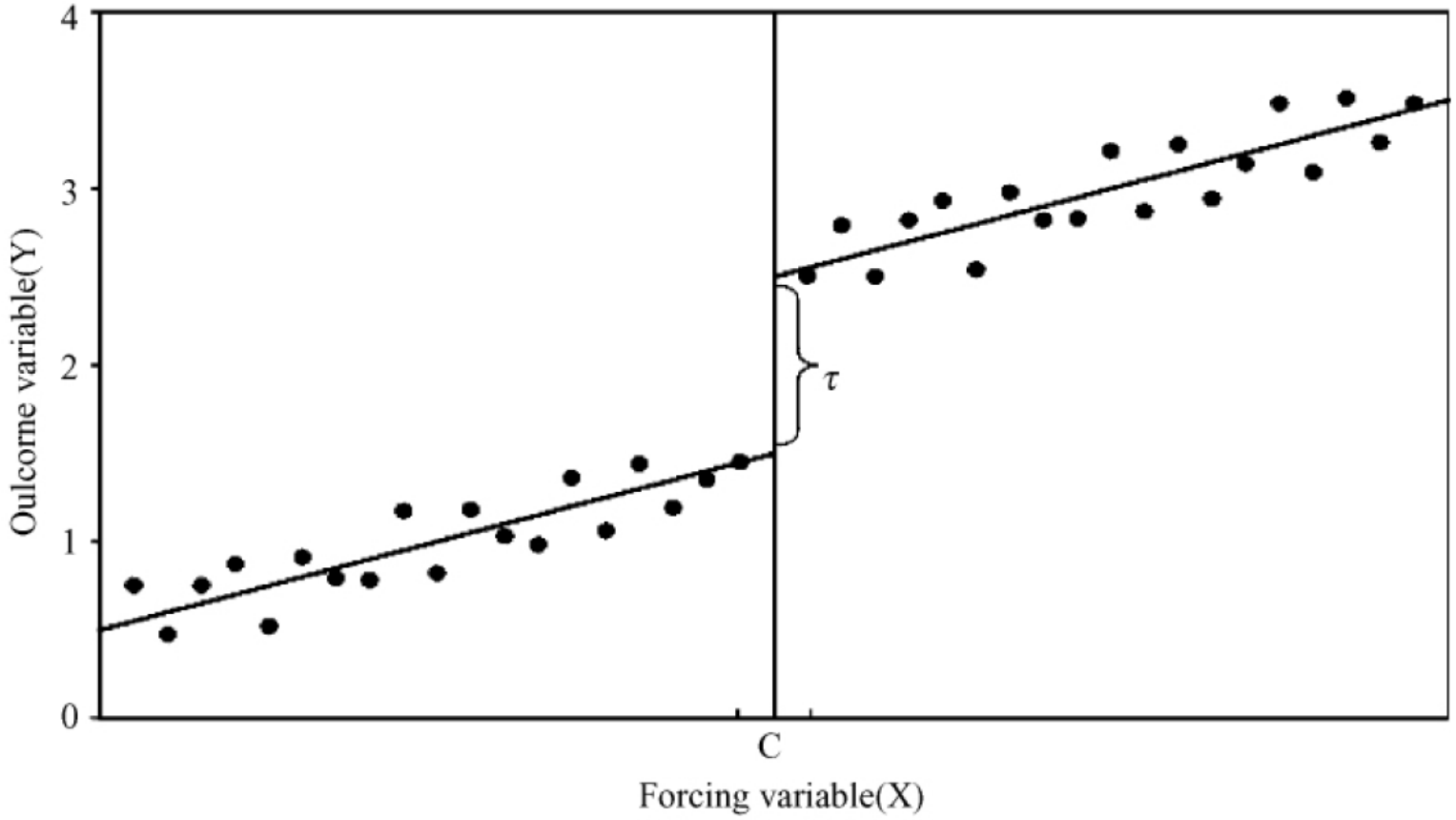
- RD 方法最早由Thistlethwaite 和Campbell 于1960年提出，是在非实验的情况下处理处置效应（treatment effects）的一种方法。

- DID：非随机实验，一部分处理组，一部分对照组，**假定如果没有处理政策发生，处理组和对照组保持相同趋势**

- RD：非随机实验， 当某变量大于临界值时, 个体受处置, 而在该变量小于临界值时, 个体不接受处置，**假定在临界值附近的处理组和对照组可比**

断点设计的基本思想

- **一个处理变量（干预变量，D）完全依赖于一个参考变量（X）**
- 参考变量X本身可以对结果变量Y有影响，也可以没有影响。
- 如果有影响，则Y与X的关系是连续的，其他可能影响Y的因素Z在断点处也是连续的。
- 则Y在断点处的跳跃可以解释为处理变量D的影响。



注：横坐标（X）代表考试成绩，纵坐标（Y）代表学术成就。

- **研究问题：**学习上的荣誉奖励(原因)是否能够提升学生未来的学术成就(结果)？
- **自变量设计：**这里的荣誉奖励是根据考试成绩而定的：当考试成绩x超过一定分数c，则给予奖励(D=1)，否则(x＜c时)则没有奖励(D=0)。
- 自变量D在x=c处产生“中断”，随后如果学生的学术成就也发生了类似的中断(例如考试成绩在c以下学生的学术成就低于考试成绩在c以上的学生的学术成就)，则可以认为奖励和学术成就之间有因果关系。

精确断点（sharp）

- 干预分配完全由参考变量是否超过临界值决定。

$$D = 1(X \geq x_0)$$

- 超过临界值的个体均接受干预，纳入实验组；未超过临界值的个体均未接受干预，纳入对照组；

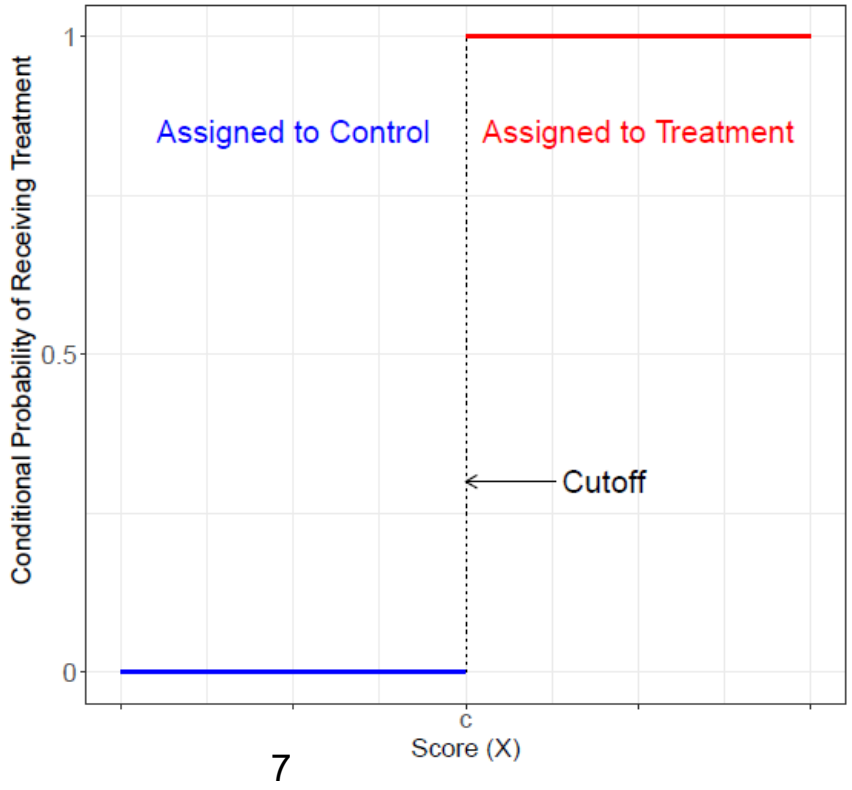
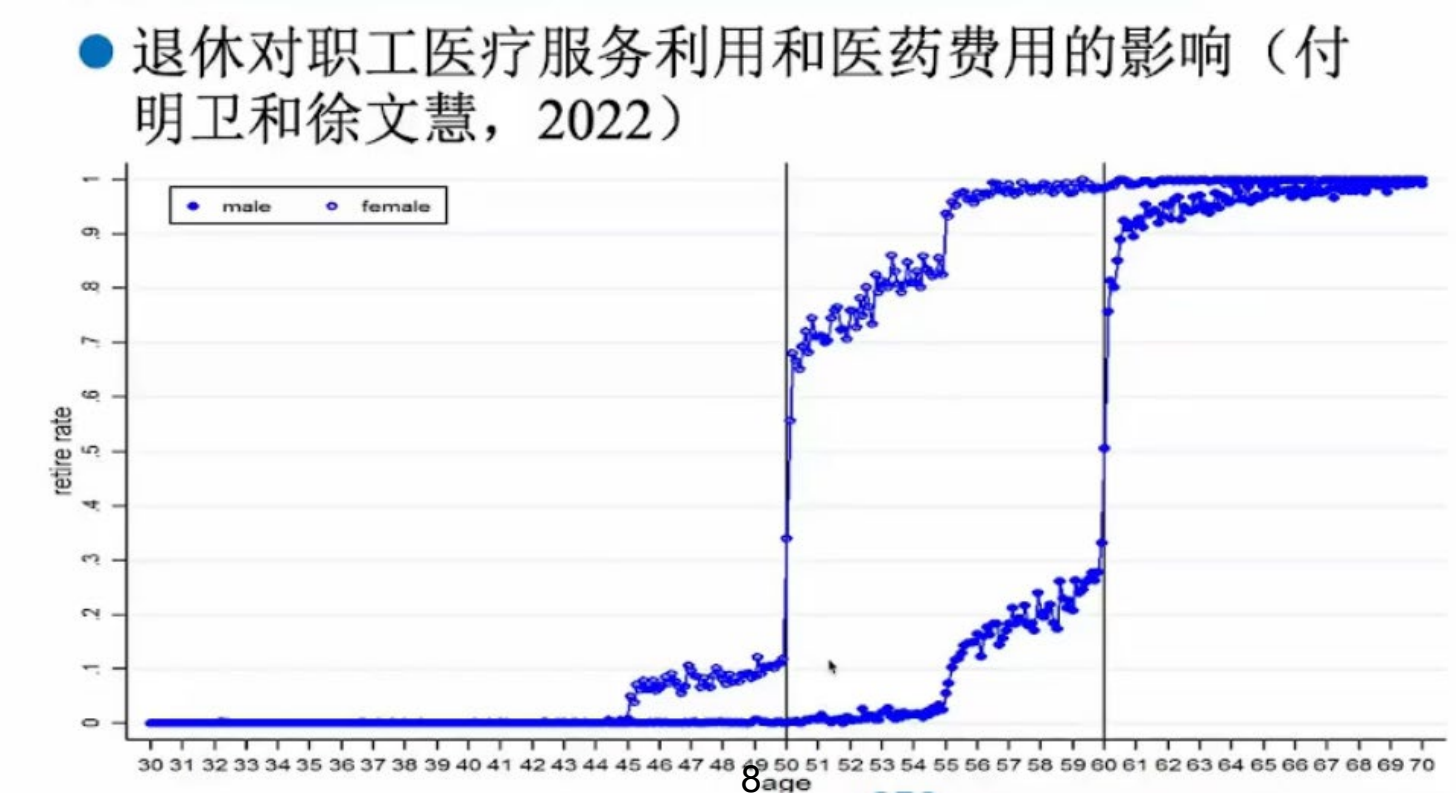


Figure 1: Conditional Probability of Receiving Treatment in the Sharp RD Design

模糊断点回归(fuzzy)

- 干预分配不完全由参考变量决定，还受到其他未观测因素的影响。
- $D_i = D(T_i, \varepsilon)$,其中 $T_i = 1(X_i = x_0)$完全由参考变量X决定；
- ε是影响干预的其他未观测因素，也可能同时影响结果变量Y。
- 一定比例的合格样本未进入实验组，同时，一定比例的不合格样本进入实验组。



RD方法的原理

- 给定x条件下平均因果关系：

$$E[Y_i(1)|X_i=x]-E[Y_i(0)|X_i=x]$$

- 因果识别的根本性问题：潜在结果无法观察

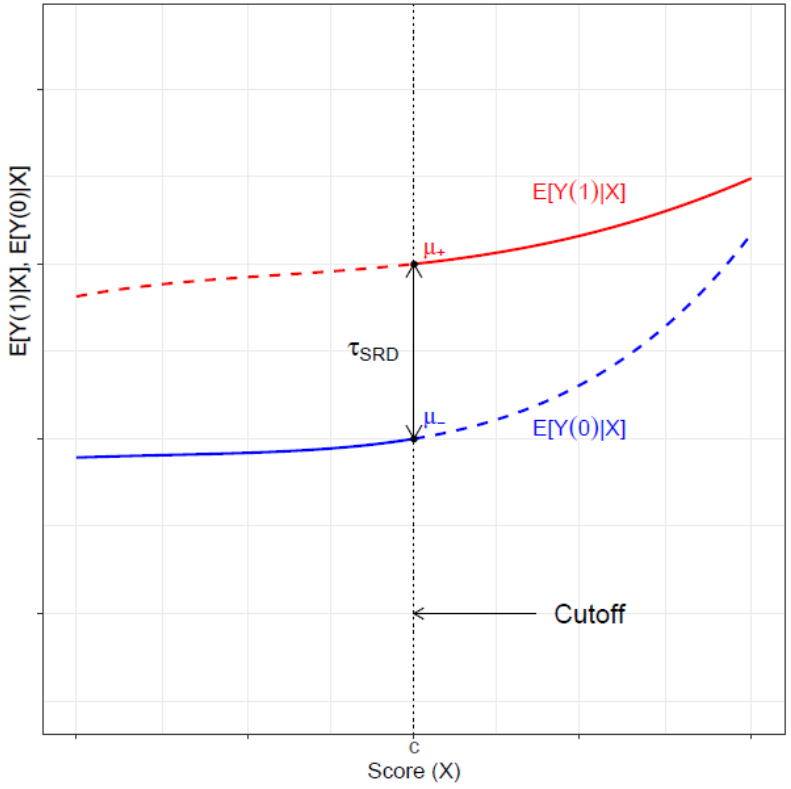


Figure 2: RD Treatment Effect in Sharp RD Design

$$\tau_{\text{SRD}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c].$$

RD方法主要应用场景

- 教育收入评估（一分定乾坤）
- 地理边界效应（内外大不同）
- 政策冲击评估（上下一刀切）
- 选举后果分析（一票定天下）

本讲主要内容

- 断点回归基本原理
- **断点回归识别条件**
- 断点回归图形分析
- 断点回归关键问题

回归断点设计的识别条件

假设1（断点假设）：

- 假设极限 $p^+ = \lim_{x \to x_0^+} E[D_i|X_i = x]$，$p^- = \lim_{x \to x_0^-} E[D_i|X_i = x]$

  存在，并且 $p^+ \neq p^-$

  其中：$D_i = D(T_i, \varepsilon)$，$T_i = 1(X_i = x_0)$

- 如果是**精确断点**，则 $D_i = T_i$，$p^+ = 1$，$p^- = 0$，
  即断点右侧个体都进入干预组，左侧个体进入控制组。
- 如果是**模糊断点**，则 $D_i = D(T_i, \varepsilon)$，$D_i \neq T_i$
  但要求断点右侧个体接受干预的概率高于断点左侧的概率。

回归断点设计的识别条件

假设2（连续性假设）：

- $E[Y_{0i}|X_i = x]$，$E[Y_{1i}|X_i = x]$ 是x的函数，并且在x_0处是连续的，

  即

$$\lim_{\varepsilon \to 0} E[Y_{ji}|X_i = x_0 + \varepsilon] = \lim_{\varepsilon \to 0} E[Y_{ji}|X_i = x_0 - \varepsilon], \qquad j = 0,1$$

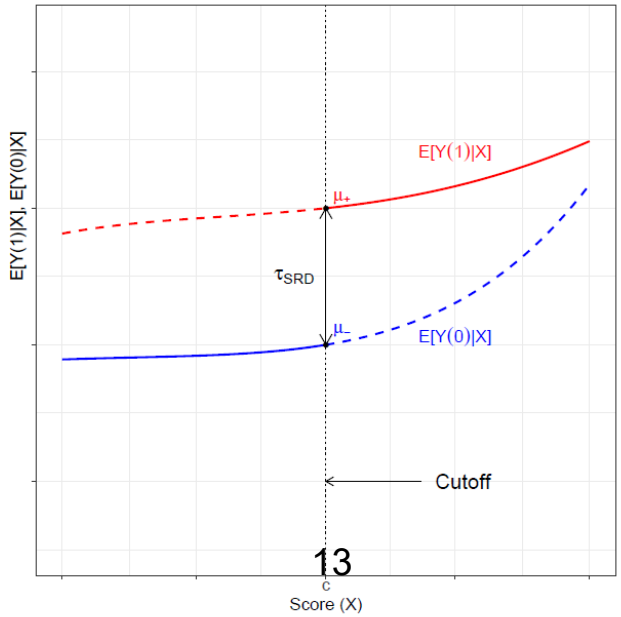


Figure 2: RD Treatment Effect in Sharp RD Design

回归断点设计的识别条件

假设3（局部随机化假设）：

- 假设在断点附近近似于完全随机化实验，即

$$(Y_{1i}, Y_{0i}) \perp D_i | X_i \in \delta(x_0)$$

其中 $\delta(x_0) = (x_0 - \delta, x_0 + \delta)$ 为x0的δ邻域，δ>0为任意小的正数。

- 局部随机化假设要求个体不能精确控制或操纵参考变量X，使之超过临界值。

- 如果个体能精确控制参考变量X，则RDD方法失效。

- 局部随机化假设是RDD策略有效的关键假设之一，可以利用参考变量X分布在断点处是否连续进行判断。

- 比如学生不能通过讨好老师或其他方式修改成绩

回归断点设计的识别条件

	基于连续性的RDD：rdrobust	基于局部随机性的RDD：rdlocrand
基本假设	观测样本是从无穷大的整体中随机抽取的； 在断点处，回归函数是连续的； 允许配置变量除影响处理状态外，对结果变量也产生影响。	观测窗口内，样本点的分布是随机的； 结果变量仅与处理状态有关，与配置变量无关； 观测窗口内样本是否受处理是完全随机的，而且结果变量的变化完全由处理状态决定。
估计方法	多项式参数估计； 非参数估计	控制组和处理组均值之差； 使用多项式变换来排除X_i对Y_i的影响
统计推断	选择更小带宽； 稳健偏差校正法	费雪推断法（小样本）； 依分布收敛于正态分布（大样本）
带宽/观测窗口选择	选择带宽h_{MSE}，使得断点回归估计的均方误差最小化； 选择带宽h_{CER}，使得覆盖误差率最小化。	选择在观测窗口外与配置变量相关、且不受处理状态影响的协变量，寻找令协变量与配置变量无关的最大观测窗口。
适用场景	断点附近的样本量较多，且较为连续。	断点附近的样本量较少； 配置变量为离散变量； 作为基于连续性的RDD的稳健性检验。

模糊断点作为IV

案例：读研对收入的影响

根据学生综合成绩是否超过一个临界值来确定推免资格

学生是否读研，除推免资格外，还有一些其他未观测因素影响，这些因素即影响是否读研，又影响个体收入，从而是未观测混淆因素

推免资格是学校按照标准给定的，满足外生性假设，断点独立于学生个体的的潜在结果，而且推免资格也明显影响个体读研行为（但不完全）

无论是否有保研资格都会读研（总是参与者）；无论是否有保研资格都不会读研（从不参与者）

有资格就读研，没资格就不断言（依从者）

断点作为工具变量估计出的就是依从着的局部平均处理效应

本讲主要内容

- 断点回归基本原理
- 断点回归识别条件
- 断点回归图形分析
- 断点回归关键问题

回归断点设计的图形分析

- RD识别的图形分析非常重要，没有绘图，绝不可能通过审稿流程
- RD识别的基本条件是：干预分配概率在临界点会有跳跃，结果变量在临界点也会有跳跃，而其他影响结果的变量在临界点没有跳跃。
- 绘制结果变量与参考变量的关系图，判断结果变量在断点处是否有跳跃，以及在非断点处是否有跳跃。
- 绘制干预分配概率与参考变量的关系图，判断是适用精确断点回归，还是模糊断点回归。
- 绘制协变量与参考变量的关系图，检验其在临界点处是否有跳跃。

结果变量与参考变量的关系图

- 用于观察结果变量是否在间断点处有跳跃。但避免直接利用原始数据绘图，原始数据中噪音太多。

- 可以通过适当平均后绘图：通常将参考变量划分为一系列区间，区间的宽度相同，并且保证断点左边和右边分别在不同区间内，避免将处于不同干预状态的个体混在同一区间。然后将所有区间里个体结果变量的平均值与区间的中点进行描点。

- 可以通过多项式分别对断点两边的点进行拟合，并将拟合曲线描在图上。

- 可观测协变量与参考变量的关系图也参照上述方法绘制。

结果变量与参考变量的关系图

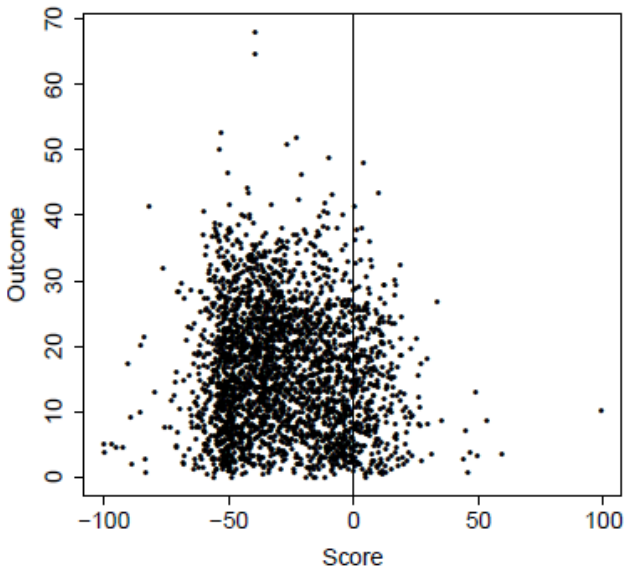


Figure 5: Scatter Plot (Meyersson Data)

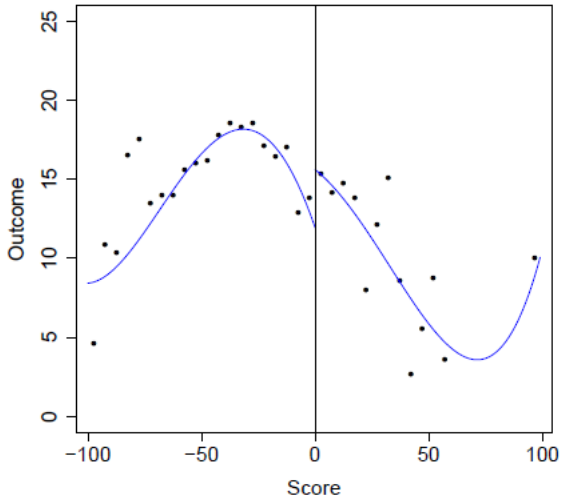


Figure 6: RD Plot for Meyersson Data Using 40 Bins of Equal Length

结果变量与参考变量的关系图

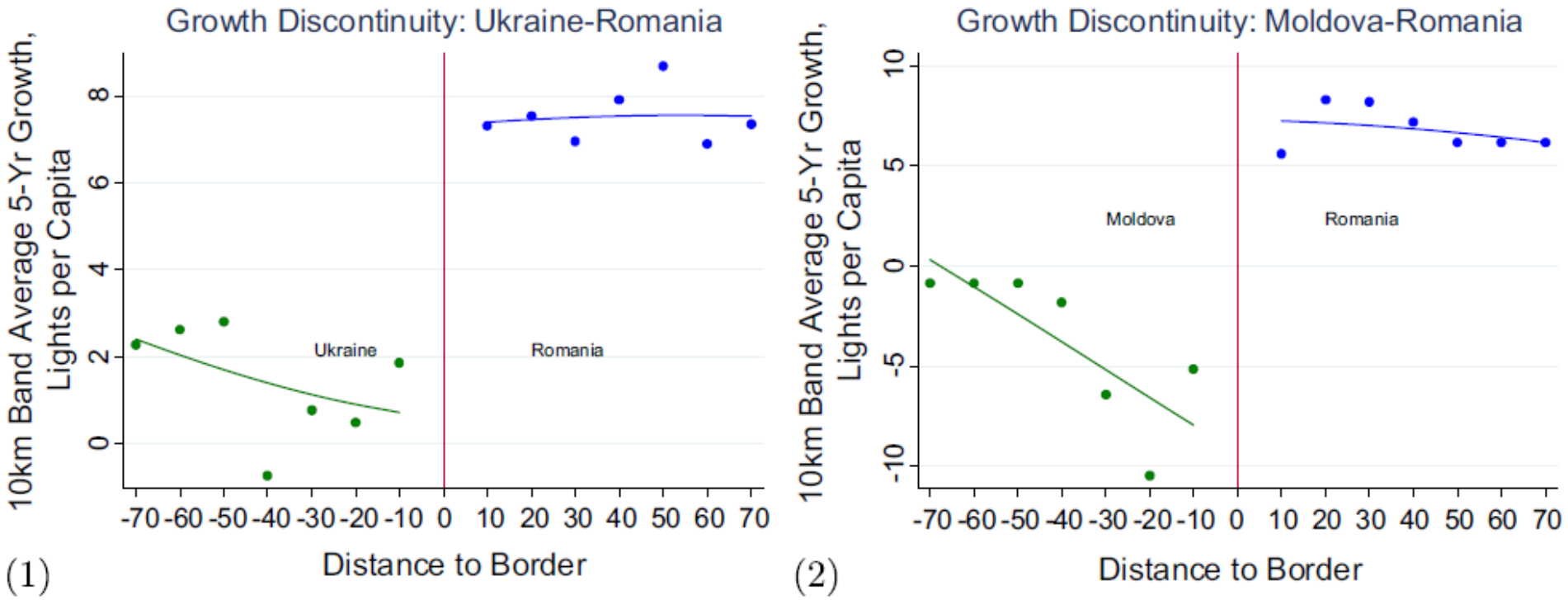
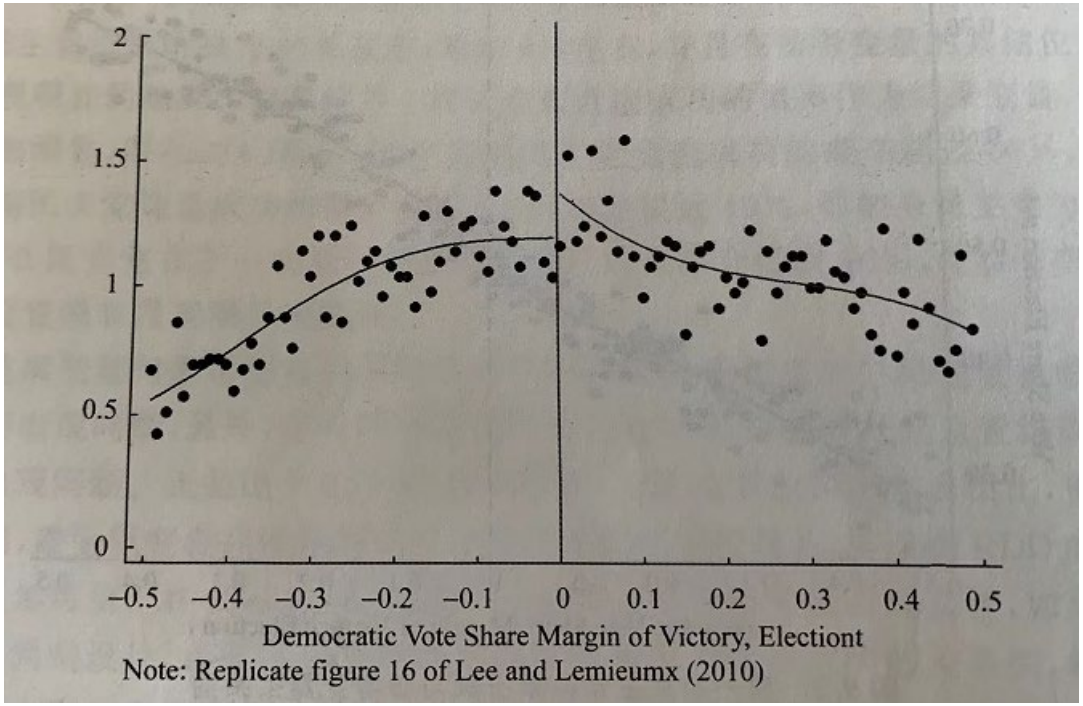


Fig. 3 Discontinuity plots: growth in light density across selected Eastern European borders

参考变量分布图

- RDD的另一个关键识别条件是个体不能精确地控制或操纵临界点，如果个体能够精确操纵可能导致断点左右个体分别差异很大

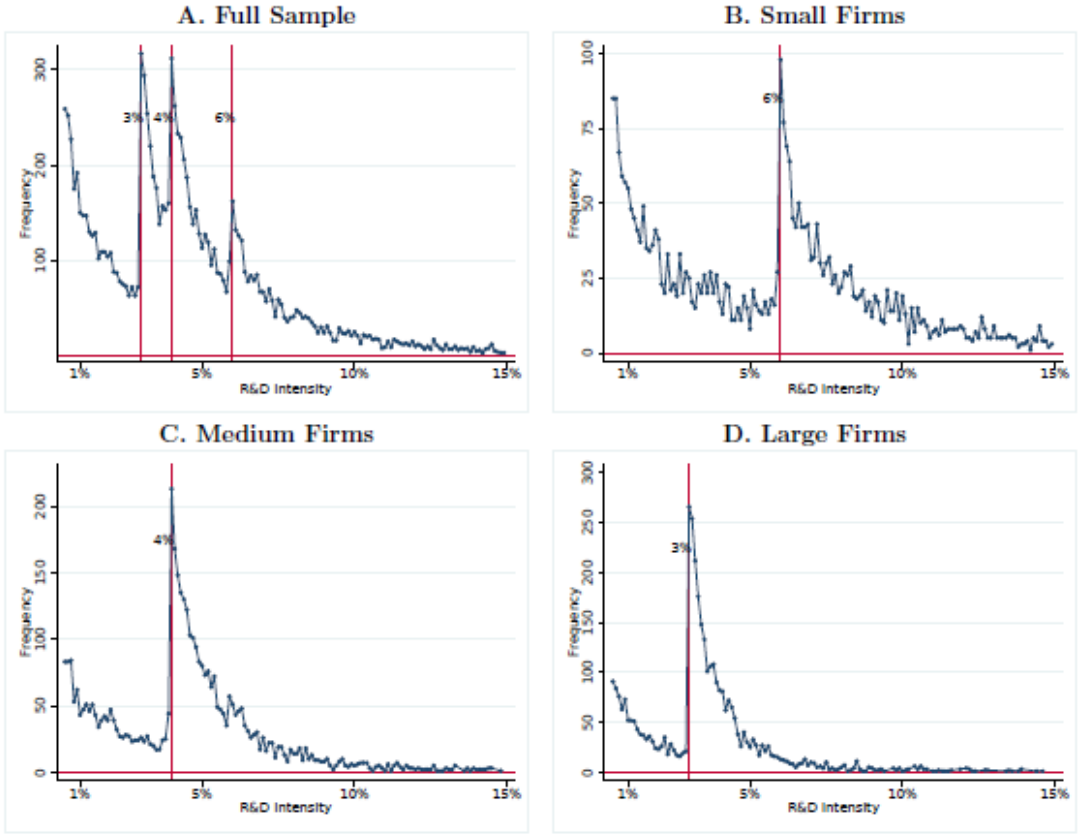
- 因此可以使用参考变量分布图进行检验，如果参考变量在断点处是联系的，说明个体没有操纵断点，否则，可能意味着断点附近存在操纵



- 可以使用直方图、DCdensity检验、rddensity检验等方法

数据操纵案例

Figure 2: Bunching at Different Thresholds of R&D Intensity (2011)



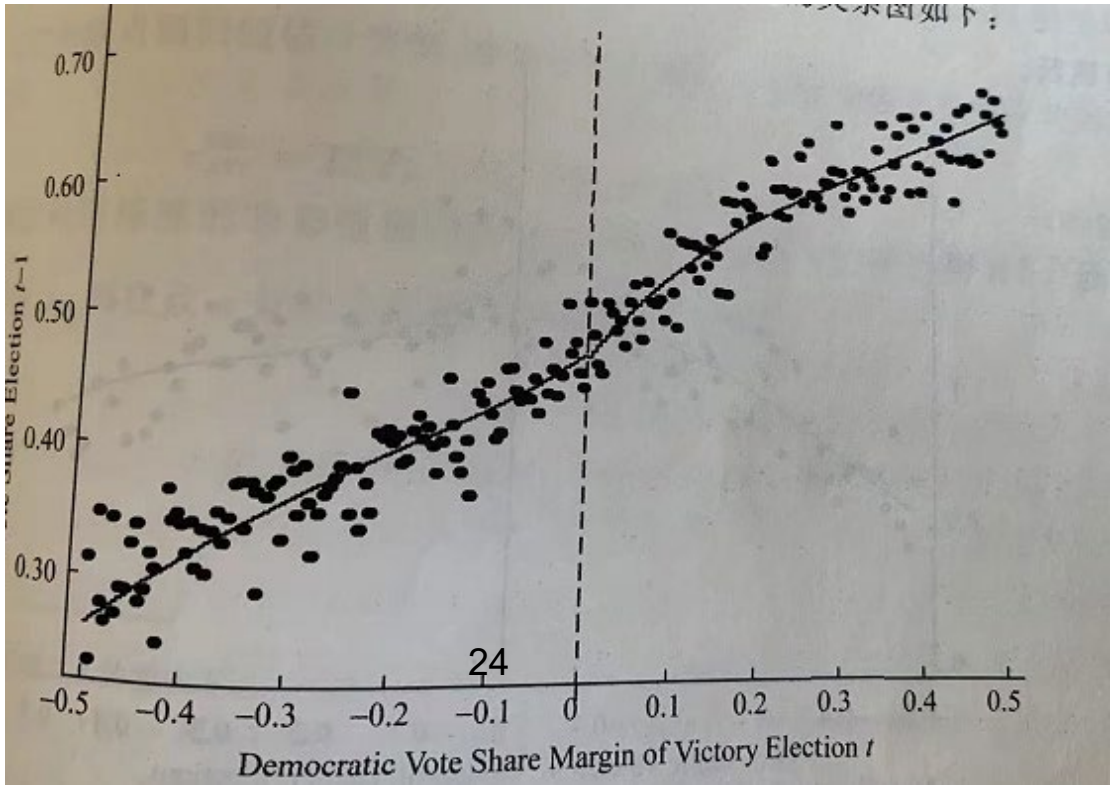
Notes: This figure plots the empirical distribution of R&D intensity for all manufacturing firms with R&D intensity between 0.5% and 15% in the Administrative Tax Return Database. Panel A reports the pooled data distribution with all sizes of firms. Panels B, C, and D report the R&D intensity distribution of small, medium, and large firms, respectively. Note that large fractions of the firms bunch at the thresholds (6% for large, 4% for medium, and 3% for large) at which they qualify to apply for the InnoCom certification. Source: Administrative Tax Return Database. See Section 3.1 for details.

Chen, Z.， Liu, Z., Suárez Serrato, J.,Xu, Y., "**Notching R&D Investment with Corporate Income Tax Cuts in China**", *American Economic Review， 2021，4.*

可观测协变量与参考变量关系图

- 为了检验连续性假设是否成立，可以采用同样的方法，绘出可观测协变量与参考变量直接的关系图

- 如果发现有些协变量在临界点有间断，说明RD设计可能存在问题，结果变量在断点的跳跃有可能是由这一观测因素跳跃造成的，而不一定是想要的政策干预结果

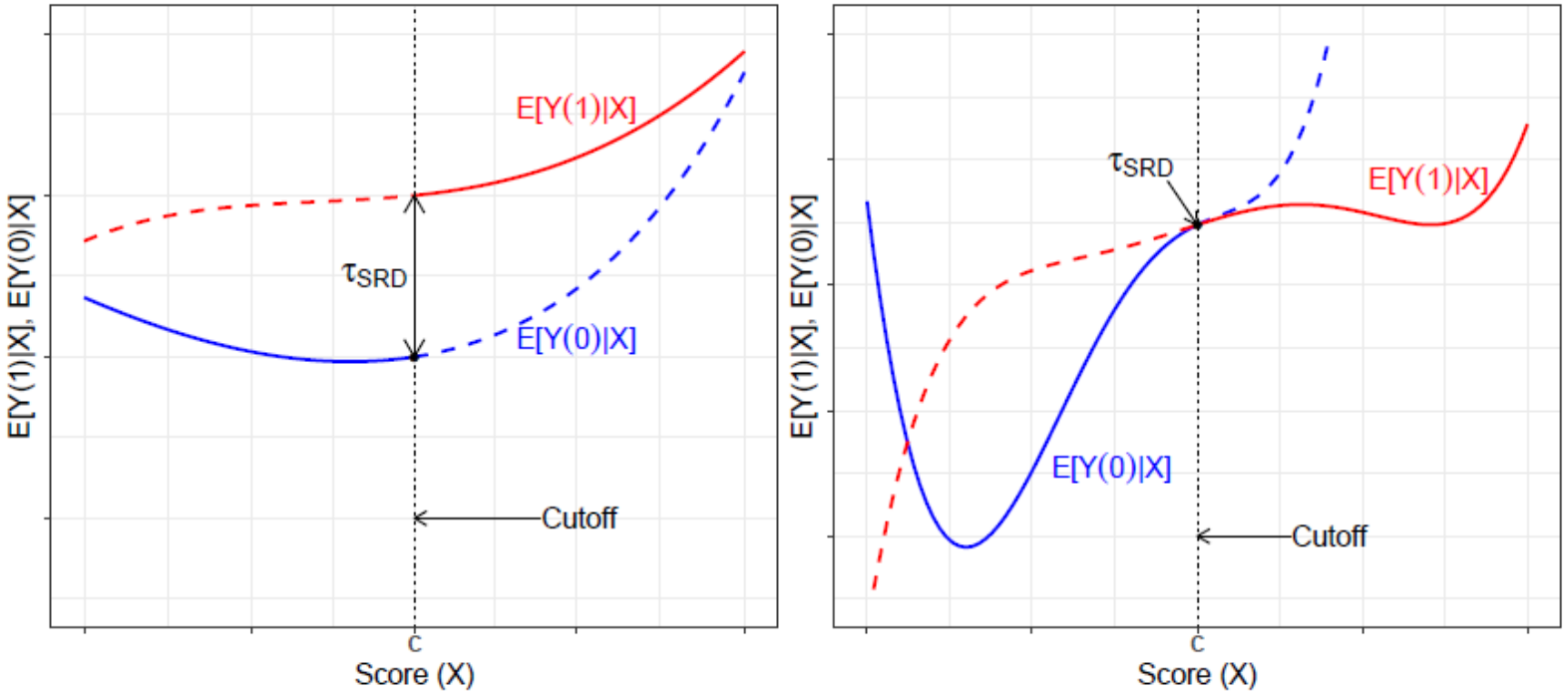


本讲主要内容

- 断点回归基本原理
- 断点回归识别条件
- 断点回归图形分析
- **断点回归关键问题**

理解RD估计结果的局部特征

- RD估计的是局部平均处理效应，其外部有效性一般不得而知



(a) Mild Heterogeneity (b) Severe Heterogeneity

Figure 4: Local Nature of the RD Effect

带宽选择

- RDD的参数估计依赖于一个重要参数—带宽h的选择。

- 带宽比较小时，断点左右的个体特征差异较小，估计偏差较小。但是，带宽小意味着断点左右h范围内的样本容量较小，估计量的方差较大，估计精度较低。

- 带宽比较大时，断点左右h范围内的样本容量较大，估计量的方差较小，估计精度较高。但是，较大的带宽意味着有些个体特征差异较大，相似度降低，估计偏差较大。

- 实操中有很多方法：CV、IK、CCT

- 建议：没有统一的最优方法，多尝试，最好是多种方法结论一致

多项式选择

在利用局部多项式进行RDD估计时，需要选择滞后阶数P：

- 早期文献建议：AIC标准，AICC标准，或BIC标准。
- 带宽越大时，需要选择的滞后阶数越大；带宽越小时，滞后阶数越小。
- 最新的建议：尽可能简单，不要搞太复杂的多项式

稳健性分析

- 不同带宽
- 不同多项式
- 去除断点附近值
- 其他假断点处是否跳跃

参考文献(要参考最新的文献)

- Cattaneo, M, D., Idrobo, N., and Titiunik, R., A Practical Introduction to Regression Discontinuity Designs: Foundations, Working Paper, 2019.

- Cattaneo, M, D., and Titiunik, R., Regression Discontinuity Designs, Working Paper 2022.

- 赵西亮，《基本有用的计量经济学》，北京：北京大学出版社，2017年7月。

- RD实操：https://zhuanlan.zhihu.com/p/100524478

- 张川川: https://www.bilibili.com/video/BV1bC4y1b79b/?spm_id_from=333.337.search-card.all.click

1917—2017
100th Anniversary
Shanghai University of Finance and Economics
上海财经大学 100周年校庆

Thank You!

公众号：经济数据勘探小分队