断点回归的两大分析框架及Stata应用

陈强
山东大学经济学院
www.econometrics-stata.com

目录

一、引言

- 断点回归设计（regression discontinuity design）是最为流行的准实验因果推断方法之一

- Thistlethwaite and Campbell (1960)首次提出断点回归，并以此研究奖学金对于未来学业成就的影响。

- 由于奖学金由学习成绩决定，而学生无法精确控制其成绩，故成绩刚好达到获奖标准与差点达到的学生具有可比性。

断点回归的两个基本前提

- 首先，个体获得一个得分（score），若此得分超过已知的某个断点（cutoff或threshold），则进入处理组，接受政策处理；反之，则进入控制组。这正是断点回归特有的“处理配置机制”（treatment assignment mechanism）

- 其次，在此断点附近两侧的处理组与控制组个体具有“可比性”（comparability），故可将对方作为“有效的反事实”（valid counterfactuals）。

断点回归的两大分析框架

- 针对断点附近两侧个体的可比性所使用的不同数学表达，文献中出现了两个分析框架。

- 基于连续性的框架（continuity-based framework）假设潜在结果的条件期望在断点处连续，这保证了在断点附近两侧的处理组与控制组个体的特征相近。

- 局部随机化的框架（local randomization framework）则假设在断点附近的小窗口内，个体的驱动变量及处理状态可视为随机分配（as-if randomly assigned）。

二、基于连续性的框架

- 假设数据为横截面的随机样本 $\{X_i, D_i, Y_i\}_{i=1}^{n}$，其中三个变量分别为驱动变量、处理变量与结果变量。

- 断点回归的处理配置规则为 $D_i = 1(X_i \geq c)$，其中 c 为某已知断点。

- 我们仅关注“精确断点回归”（sharp regression discontinuity）

断点回归的直观理解

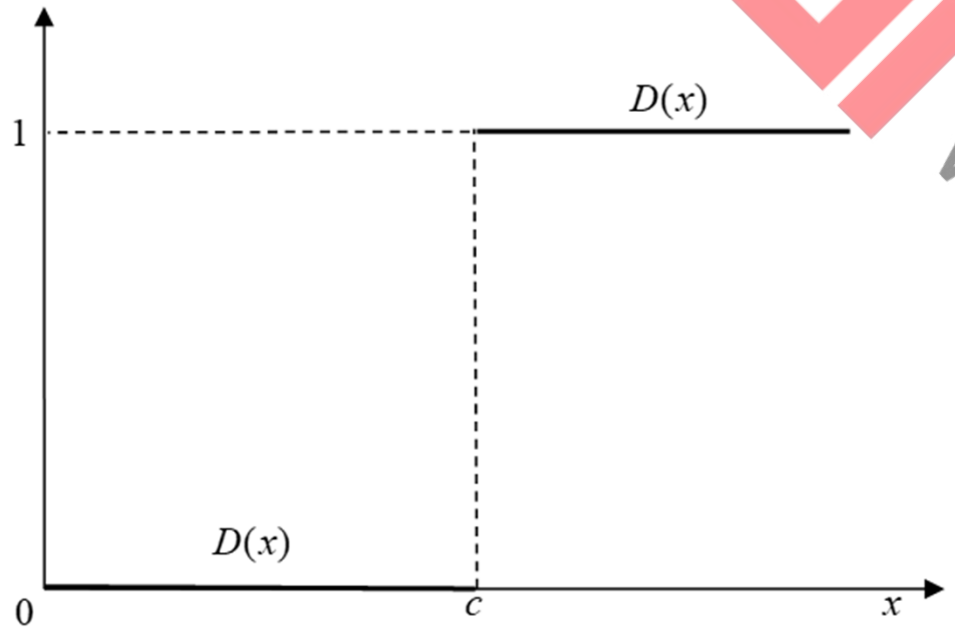


图 16.1 函数 $D(x)$ 的跳跃

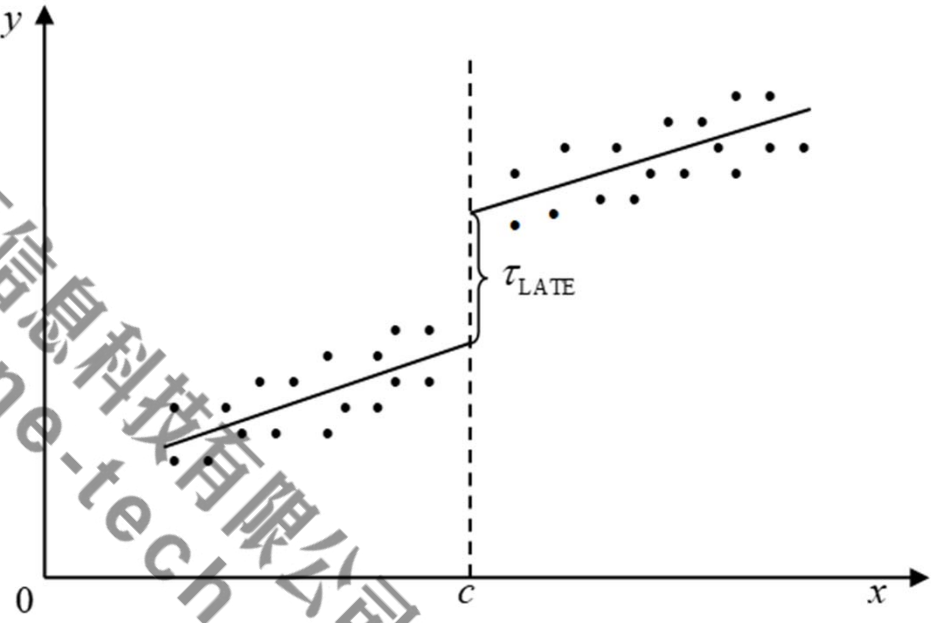


图 16.2 函数 $\mathrm{E}(y \mid x)$ 的跳跃

连续性框架的识别

- Hahn et al. (2001)在潜在结果（potential outcomes）的框架下，首次证明了断点回归的非参数识别（nonparametric identification）条件。

- 记个体 i 的两个潜在结果分别为 $Y_i(0)$（未受处理）与 $Y_i(1)$（受处理），观测结果 $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$

- 感兴趣的“被估量”（estimand）为在断点处的局部平均处理效应（LATE）：

$$\tau \equiv \mathrm{E}[Y_i(1) - Y_i(0) \mid X_i = c]$$

连续性框架的假定

- **假定2.1** (驱动变量在断点处的密度为正)　驱动变量 X_i 为连续型随机变量，且在断点 $X_i = c$ 的密度函数为正数。

- **假定2.2** (潜在结果的条件期望在断点处连续)　条件期望 $\mathrm{E}[Y_i(0) \mid X_i = x]$ 与 $\mathrm{E}[Y_i(1) \mid X_i = x]$ 作为 x 的函数，在断点 c 处连续。

- 在断点处潜在结果的条件期望没有跳跃，故若观测结果的条件期望跳跃，则只能是由于在断点处的处理状态跳跃所致

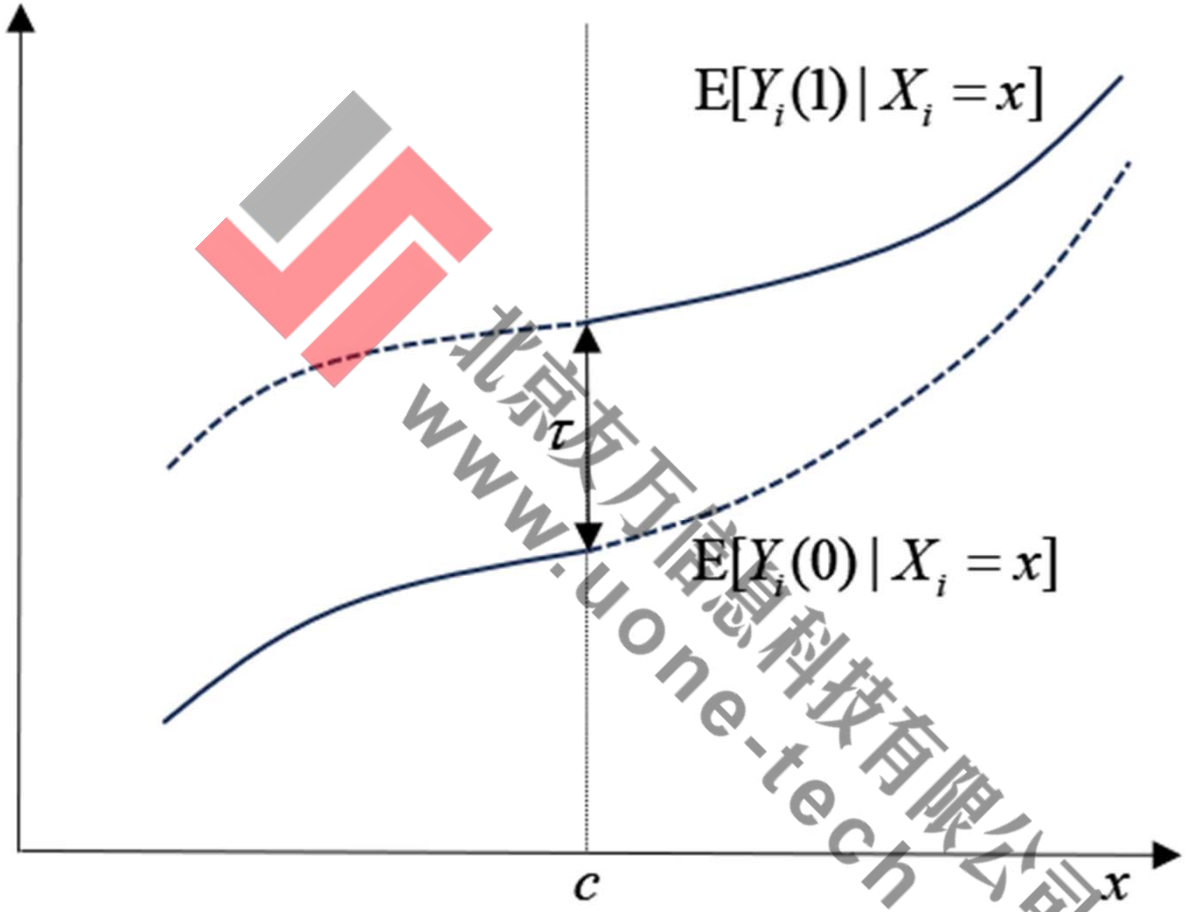


图 1 基于连续性的分析框架

根据假定 2.1，驱动变量 X_i 在断点处的密度为正，故对于任意小的正数 $\delta > 0$，可考察断点两侧个体结果变量的平均差异：

$$\begin{aligned}&\mathrm{E}(Y_i \mid X_i = c+\delta) - \mathrm{E}(Y_i \mid X_i = c-\delta)\\&= \mathrm{E}[Y_i(1) \mid X_i = c+\delta] - \mathrm{E}[Y_i(0) \mid X_i = c-\delta]\end{aligned}$$

在上式中，令 $\delta \to 0$，两边同时求极限可得：

$$\begin{aligned}&\lim_{x \downarrow c} \mathrm{E}(Y_i \mid X_i = x) - \lim_{x \uparrow c} \mathrm{E}(Y_i \mid X_i = x)\\&= \lim_{x \downarrow c} \mathrm{E}[Y_i(1) \mid X_i = x] - \lim_{x \uparrow c} \mathrm{E}[Y_i(0) \mid X_i = x]\\&= \mathrm{E}[Y_i(1) \mid X_i = c] - \mathrm{E}[Y_i(0) \mid X_i = c] = \tau\end{aligned}$$

其中，$\lim_{x \downarrow c} \mathrm{E}(Y_i \mid X_i = x)$ 与 $\lim_{x \uparrow c} \mathrm{E}(Y_i \mid X_i = x)$ 分别为 $\mathrm{E}(Y_i \mid X_i = x)$ 在 $x = c$ 处的右极限与左极限，而倒数第二个等号利用了假定 2.2（潜在结果的条件期望连续）。由此得到识别结果（参见图 1）：

$$\tau = \lim_{x \downarrow c} \mathrm{E}(Y_i \mid X_i = x) - \lim_{x \uparrow c} \mathrm{E}(Y_i \mid X_i = x) \tag{1}$$

连续性框架的估计

- Hahn et al. (2001)建议使用非参数的局部线性回归 (local linear regression)，分别利用断点两侧的数据估计 $\lim_{x\downarrow c}\mathrm{E}(Y_i \mid X_i = x)$ 与 $\lim_{x\uparrow c}\mathrm{E}(Y_i \mid X_i = x)$ 。

- 给定核函数 (kernel function) $K(\cdot)$ 与带宽(bandwidth) h，对于断点右侧的处理组样本，可进行加权最小二乘法(WLS)估计：

$$(\hat{\beta}_{+,0}\ \hat{\beta}_{+,1}) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} \mathbf{1}(X_i \geq c)\left[Y_i - b_0 - b_1(X_i - c)\right]^2 K\left(\frac{X_i - c}{h}\right)$$

连续性框架的估计（续）

- 类似地，对于断点左侧的控制组样本，可进行WLS估计：

$$(\hat{\beta}_{-,0}\ \hat{\beta}_{-,1}) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} \mathbf{1}(X_i < c)\left[Y_i - b_0 - b_1(X_i - c)\right]^2 K\left(\frac{X_i - c}{h}\right)$$

- 断点回归的LATE估计量为 $\hat{\tau} = \hat{\beta}_{+,0} - \hat{\beta}_{-,0}$，即在断点两侧局部线性回归的截距项之差。

灵活参数回归(flexible parametric regression)

- 在早期的断点回归实践中，为更好地拟合数据，常使用高阶多项式回归，且不考虑模型设定误差；而带宽则由研究者主观设定

- **缺陷 #1** 由于数值分析中的龙格现象（Runge's phenomenon），高阶多项式回归在边界处（boundary）的回归拟合值并不稳定。

- **缺陷 #2** 由研究者主观选择的“临时带宽”（ad-hoc bandwidth）缺乏客观性，且易为研究者所操纵。

- **缺陷 #3** 灵活参数回归假设模型为正确设定，在进行统计推断时并不考虑设定误差。

最优带宽

- Imbens and Kalyanaraman (2012)首次提出选择最优带宽 $\hat{h}$，使得估计量 $\hat{\tau}$ 的均方误差（MSE）最小化，但使用“第一代插入法”（first-generation plug-in rule）估计MSE的一阶近似，仍存在偏差。

- Calonico, Cattaneo and Titiunik (2014)利用MSE的更一般展开式，并使用“第二代插入法”（second-generation plug-in rule），得到更为精确的MSE最优带宽（MSE-optimal bandwidth），记为 $\hat{h}_{\mathrm{MSE}}$。

最优带宽（续）

- Calonico, Cattaneo and Farrell (2020)从最优化统计推断的角度选择最优带宽，以最小化置信区间的覆盖误差率（coverage error rate，简记CER）。

- 使用CER最优带宽所得到的点估计并非MSE最优，故实践中一般仍主要使用MSE最优带宽。

- 若数据在断点两侧的稀疏程度不同，也可在断点两侧分别使用不同的带宽。

连续性框架的统计推断

- 由于非参数估计使用了断点附近的观测值，故存在偏差。只要使用最优带宽，此偏差将在大样本下消失，仍为一致估计。

- 在模型误设的情况下，即使用了最优带宽，此偏差仍然对估计量的渐近分布有影响：

$$\frac{\hat{\tau}(\hat{h}_{\mathrm{MSE}})-\tau}{\sqrt{\hat{V}}} \xrightarrow{d} N(B,1)$$

- 其中，$\hat{\tau}(\hat{h}_{\mathrm{MSE}})$ 为使用最优带宽 $\hat{h}_{\mathrm{MSE}}$ 所得到的点估计，$\sqrt{\hat{V}}$ 为标准误，而 $B \neq 0$ 为偏差项

欠光滑(undersmoothing)的解决方法

- 由于渐近分布的中心并不为0，以此进行统计推断将导致误差，使得假设检验的p值不准确，而置信区间的实际覆盖率可能明显小于其95%的置信度

- 解决方法之一为使用比 $\hat{h}_{\mathrm{MSE}}$ 更小的带宽（以更快的速度收敛于0），使得渐近分布的偏差项在大样本下消失

- 但究竟如何选择比 $\hat{h}_{\mathrm{MSE}}$ 更小的带宽则比较主观，且更窄带宽也不再是MSE最优的。

偏差校正(bias correction)的解决方法

- Calonico et al. (2014)提出了“偏差校正”的方法，即先使用MSE最优带宽得到点估计 $\hat{\tau}(\hat{h}_{\mathrm{MSE}})$，然后估计其偏差 $\hat{B}$，再将 $\left(\hat{\tau}(\hat{h}_{\mathrm{MSE}})-\hat{B}\right)$ 作为偏差校正的估计量。其中，偏差估计量 $\hat{B}$ 的表达式为

$$\hat{B}=\hat{\beta}_{+,2}(\hat{h}_{pilot})B_{+}(\hat{h}_{\mathrm{MSE}})-\hat{\beta}_{-,2}(\hat{h}_{pilot})B_{-}(\hat{h}_{\mathrm{MSE}})$$

- 其中，$\hat{\beta}_{+,2}(\hat{h}_{pilot})$ 与 $\hat{\beta}_{-,2}(\hat{h}_{pilot})$ 分别为使用带宽 $\hat{h}_{pilot}$ 在断点右侧与左侧进行局部二次回归所得的二次项系数，而 $\hat{h}_{pilot}$ 为比 $\hat{h}_{\mathrm{MSE}}$更宽的“试验带宽”(pilot bandwidth)。

- $B_{+}(\hat{h}_{\mathrm{MSE}}),\ B_{-}(\hat{h}_{\mathrm{MSE}})$可观测，取值依赖于驱动变量、核函数与带宽 $\hat{h}_{\mathrm{MSE}}$

偏差校正的解决方法（续）

- 对于点估计进行偏差校正本身也会带来不确定性，故偏差校正估计量的方差可写为 $(\hat{V}+\hat{W})$，其中 $\hat{W}$ 为偏差估计量带来的方差

- 由此可得偏差校正稳健（robust bias-corrected）的95%置信区间：

$$I_{\text{RBC}} = \left[\left(\hat{\tau}(\hat{h}_{\text{MSE}}) - \hat{B} \right) \pm 1.96 \cdot \sqrt{\hat{V}+\hat{W}} \right]$$

- 无论模型正确设定或误设，此RBC置信区间均有效，故名“稳健”（robust）。

连续性框架的证伪（falsification）

- 画断点回归图（RD plot）
- 使用“伪断点”(placebo cutoff)进行安慰剂检验
- 使用处理前变量(pretreatment covariates)作为“伪结果变量”(placebo outcome)进行安慰剂检验
- 内生分组检验(密度检验)：McCrary(2008), Cattaneo et al. (2020)
- 甜甜圈法(donut hole approach)：去掉离断点最近的几个观测值作为稳健性检验，因为内生分组最可能发生于离断点最近的个体(Barreca et al., 2011)

三、局部随机化的分析框架

- 实证研究者经常非正式地将断点回归视为“局部随机实验”(local randomized experiment)，或在断点附近“近乎随机分配”(as good as randomly assigned)，此思想最可追溯至Thistlethwaite and Campbell (1960)

- Lee (2008)进一步认为，若个体无法“精确操控”(precisely manipulate)驱动变量，则在断点附近处理状态的分配近似于随机实验

- Cattaneo et al. (2015)首次建立了局部随机化的分析框架。

Matias D. Cattaneo, Brigham R. Frandsen and Rocío Titiunik*

Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate

Abstract: In the Regression Discontinuity (RD) design, units are assigned a treatment based on whether their value of an observed covariate is above or below a fixed cutoff. Under the assumption that the distribution of potential confounders changes continuously around the cutoff, the discontinuous jump in the probability of treatment assignment can be used to identify the treatment effect. Although a recent strand of the RD literature advocates interpreting this design as a local randomized experiment, the standard approach to estimation and inference is based solely on continuity assumptions that do not justify this interpretation. In this article, we provide precise conditions in a randomization inference context under which this interpretation is directly justified and develop exact finite-sample inference procedures based on them. Our randomization inference framework is motivated by the observation that only a few observations might be available close enough to the threshold where local randomization is plausible, and hence standard large-sample procedures may be suspect. Our proposed methodology is intended as a complement and a robustness check to standard RD inference approaches. We illustrate our framework with a study of two measures of party-level advantage in U.S. Senate elections, where the number of close races is small and our framework is well suited for the empirical analysis.

局部随机化框架的识别

- 局部随机化框架的基本假设是，存在一个窗口 $W_0=[c-h,c+h]$ ，在此窗口内，驱动变量 X_i 的取值为随机分配。故可将任何 $X_i \in W_0$ 均视为来自于同一总体的随机抽样，并记此总体的累积分布函数为 $F(x)$ 。

- **假定3.1** (局部随机化) 存在一个窗口 $W_0=[c-h,c+h]$ ，对于任何个体 $X_i \in W_0$ ，都有 $F_{X_i|X_i \in W_0}(x)=F(x)$

- 在此窗口内，驱动变量 X_i 为iid，且独立于潜在结果

局部随机化框架的假定

- 由于 $D_i = \mathbf{1}(X_i \geq c)$，故在窗口 W_0 内，D_i 独立于潜在结果。
- 记个体 i 的潜在结果为 $Y_i(d,x)$，其中 $d=0$或1，为D_i 的实现值；而 x 为 X_i 的实现值。Cattaneo et al. (2015)假设，在窗口 W_0 内，X_i 仅通过影响 D_i 而作用于潜在结果。
- **假定3.2** (排他性约束) 对于任何个体 i 满足 $X_i \in W_0$，都有 $Y_i(d,x) = Y_i(d)$
- 此假定很强（参见图2），Cattaneo et al. (2017) 放松了此假定

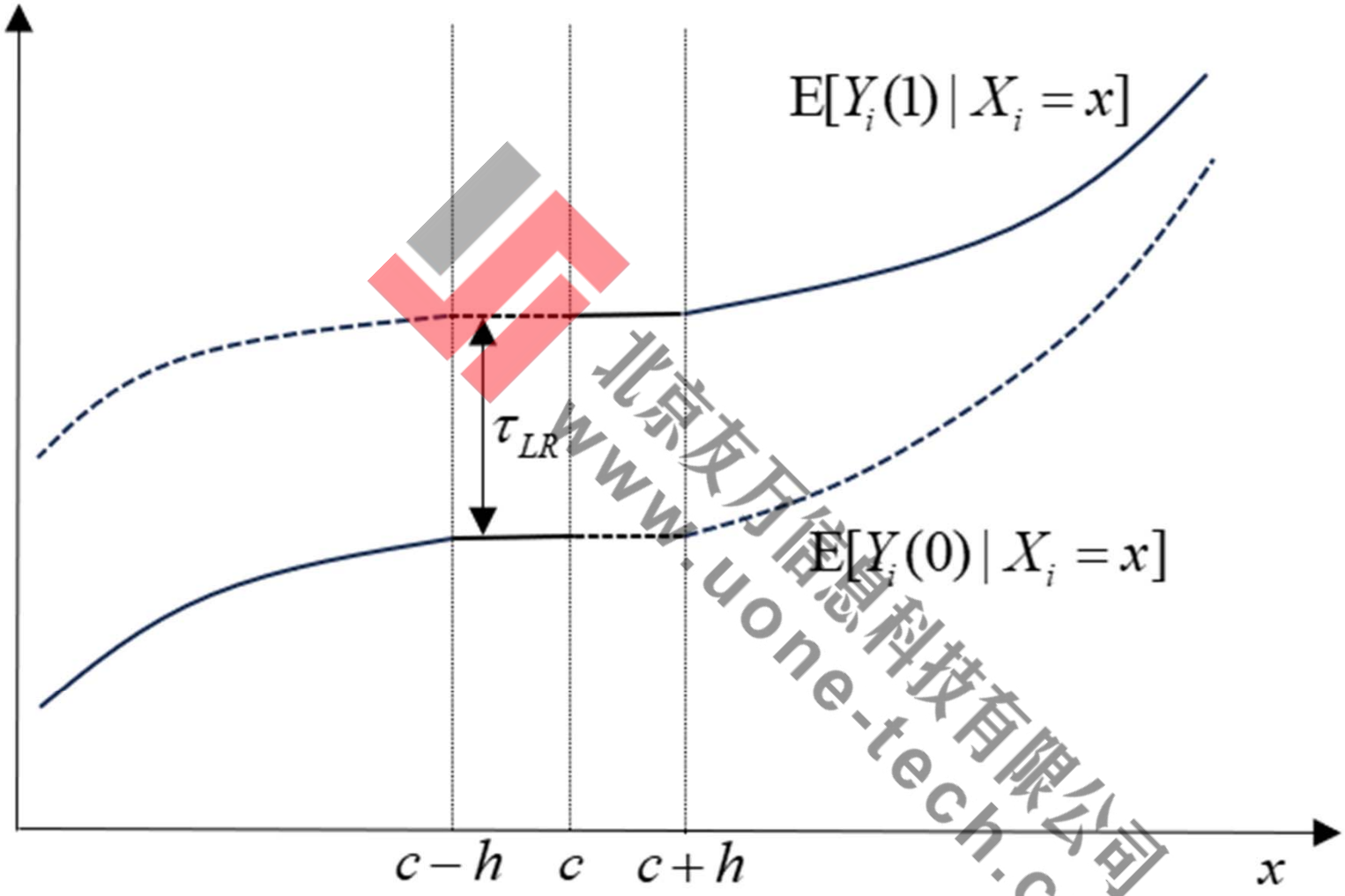


图 2 局部随机化的分析框架

在窗口W_0内，记个体数目为n_{W_0}，处理组个体数目为$n_{W_0}^+$，而控制组个体数目为$n_{W_0}^- = n_{W_0} - n_{W_0}^+$。基于假定 3.1 与 3.2，可识别在窗口$W_0$内的平均处理效应（ATE），记为$\tau_{LR}$：

$$\begin{aligned}
\tau_{LR} &\equiv \frac{1}{n_{W_0}} \sum_{X_i \in W_0} \mathrm{E}\left[Y_i(1, X_i) - Y_i(0, X_i)\right] \\
&= \frac{1}{n_{W_0}} \sum_{X_i \in W_0} \mathrm{E}\left[Y_i(1) - Y_i(0)\right] \\
&= \frac{1}{n_{W_0}} \sum_{X_i \in W_0} \mathrm{E}\left[Y_i(1) \mid D_i = 1\right] - \frac{1}{n_{W_0}} \sum_{X_i \in W_0} \mathrm{E}\left[Y_i(0) \mid D_i = 0\right] \\
&= \frac{1}{n_{W_0}^+} \sum_{X_i \in W_0,\, D_i = 1} \mathrm{E}(Y_i) - \frac{1}{n_{W_0}^-} \sum_{X_i \in W_0,\, D_i = 0} \mathrm{E}(Y_i)
\end{aligned} \tag{6}$$

其中，第 1 个等号为定义式，第 2 个等号使用了假定 3.2，而第 3 个等号使用了假定 3.1（假定 3.1 意味着，在窗口W_0内，D_i独立于$\{Y_i(0), Y_i(1)\}$）。从表达式(6)及图 2 易知，局部随机化框架可识别窗口W_0内所有个体的平均处理效应τ_{LR}，而不仅仅是断点处的局部平均处理效应τ。

局部随机化框架的估计

一个自然的估计量是“差分估计量”（difference-in-means），即在窗口 W_0 内，断点两侧结果变量的平均差异：

$$\hat{\tau}_{LR}(W_0) = \bar{Y}^{+}_{W_0} - \bar{Y}^{-}_{W_0} \tag{7}$$

其中，$\bar{Y}^{+}_{W_0} \equiv \frac{1}{n^{+}_{W_0}} \sum_{X_i \in W_0,\, D_i=1} Y_i$ 与 $\bar{Y}^{-}_{W_0} \equiv \frac{1}{n^{-}_{W_0}} \sum_{X_i \in W_0,\, D_i=0} Y_i$ 分别为在窗口 W_0 内，断点两侧结果变量的样本均值。这等价于在窗口 W_0 内，把 Y_i 对 D_i 作一元回归：

$$Y_i = \gamma_0 + \gamma_1 D_i + u_i \qquad (X_i \in W_0) \tag{8}$$

其中，OLS 估计量 $\hat{\gamma}_1$ 就是差分估计量。

如何选择局部随机化的窗口

- Cattaneo et al. (2015)提出，利用协变量来帮助选择 W_0。这些协变量可以是“处理前协变量”（pretreatment covariates），或已知不受处理影响的“伪结果”（placebo outcomes）。

- 如果个体在窗口 W_0 内确实是随机分组，则处理组与控制组的协变量分布应该没有系统差别，故可通过一系列的“协变量平衡”（covariate balance）检验来确定窗口。

从大到小地选择窗口

- 假设存在协变量 Z_i，在窗口 W_0 内与处理变量 D_i 不相关，但在窗口 W_0 之外，则 Z_i 与 D_i 相关。

- 从可能的最宽 W_0 开始(包括所有观测值)，检验原假设“D_i 对 Z_i 无作用”。若拒绝原假设，则缩小窗口，再次检验。继续此过程，直到在某窗口内，不再拒绝原假设；或达到最小的窗口尺度（minimum window size）。

- 在设定最小窗口时，Cattaneo et al. (2015)建议应大致保证在断点两侧至少各有10个观测值。

从小到大地选择窗口

- 从大到小的窗口选择过程可能较费时。

- Cattaneo et al. (2016)提出从小到大的窗口选择方法，即从最小窗口开始，每次在断点两侧至少增大若干观测值（比如5个或1个观测值），以此逐渐扩大窗口，直到p值低于显著性水平为止。

- 在进行协变量平衡检验时，最简单的检验统计量为“差分估计量”，即在窗口内，断点两侧协变量的样本均值之差 。这等价于在待评估的窗口内，作一元回归：

$$Z_i = \eta_0 + \eta_1 D_i + v_i$$

- 检验原假设 $H_0 : \eta_1 = 0$

设定显著性水平

- 常规的5%显著性水平可能太低，因为我们更担心II型错误（type II error），即未能拒绝错误的原假设 $H_0 : \eta_1 = 0$，导致局部随机化的假定在所选窗口不能成立，致使后续统计推断失效。

- 反之，若发生I型错误（type I error），即拒绝了正确的原假设“ ”，则只是选择更窄的窗口，损失效率而已。

- 为了保守起见，Cattaneo et al. (2015)建议将显著性水平设为15%，以更好地控制II型错误。

使用多个协变量进行检验

- 如果存在多个协变量，比如$\{Z_{1i},\cdots,Z_{Ki}\}$，可对每个协变量分别进行协变量平衡检验

- 记所得的相应 p 值分别为$\{p_1,\cdots,p_K\}$，然后以最小的p值，即$\min\{p_1,\cdots,p_K\}$，进行统计推断。

局部随机化框架的统计推断

- 局部随机化框架的一个优势在于，既然数据可视为局部随机实验，则可使用分析随机实验（analysis of experiments）的经典方法进行统计推断。

- 针对随机实验的分析框架包括“费雪框架”（Fisherian framework）、“内曼框架”（Neyman framework）以及“超总体框架”（superpopulation framework）

- 费雪框架进行有限样本的精确推断，而内曼框架与超总体框架均使用大样本推断。

费雪框架

- 费雪框架(Fisher, 1935)将观测数据本身视为感兴趣的总体，而非来自于更大总体的一个样本；故潜在结果也是固定、非随机的，唯一的随机因素则来自对于处理状态的分配 。

- 原假设为“费雪精确原假设”(Fisherian sharp null hypothesis)，即每位个体的处理效应均为0 (而非平均处理效应为0) 。

- 可得到任何统计量在有限样本下的精确分布，故可进行“有限样本的精确推断”(exact finite-sample inference)，无须使用大样本的渐近分布。

内曼框架与超总体框架

- 在内曼框架下，潜在结果依然被视为固定、非随机的，但原假设为平均处理效应为0，而非每位个体的处理效应均为0。假设固定的潜在结果来自更大的总体，然后进行大样本推断。

- 超总体框架则假设潜在结果为来自更大总体的随机样本（而非固定的），然后针对“平均处理效应为0”的原假设进行大样本推断。

- 内曼框架与超总体框架在概念上不同，但二者在实际操作上通常类似。

随机实验框架的选择

- 选定窗口后，若在窗口内有足够多观测值，可使用内曼或超总体框架进行大样本的统计推断。

- 但满足局部随机化的窗口通常很窄。

- Cattaneo et al. (2015)以局部随机化的断点回归研究美国参议院选举，通过协变量平衡检验得到窗口 $\hat{W}_0 = [-0.75, 0.75]$ ，窗口内仅有37个观测值，不便进行大样本推断。

费雪框架的操作

- 考虑费雪精确原假设，即在窗口 W_0 内，每位个体的处理效应均为0：$H_0: \tau_i^{W_0} = 0 \ (X_i \in W_0)$

- 考虑差分估计量 $\hat{\tau}_{LR} = \bar{Y}_{W_0}^{+} - \bar{Y}_{W_0}^{-}$。对窗口内的观测数据进行 M 次(例如$M=1000$) 随机置换(random permutation)，可得到在原假设下，此统计量的近似分布 $\left\{\hat{\tau}_{LR}^{(1)}, \ldots, \hat{\tau}_{LR}^{(M)}\right\}$。

- 考察估计量相对于此近似分布是否极端，可定义双边 p 值：

$$\text{双边}p\text{值} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{1}\left(\left|\hat{\tau}_{LR}^{(j)}\right| \geq \left|\hat{\tau}_{LR}\right|\right)$$

局部随机化框架的拓展

- 假定3.2(排他性约束)很强。即使 X_i 为随机分配，也无法排除 X_i 对潜在结果的直接效应。

- 若 X_i 对潜在结果有直接效应，则在窗口内的差分估计量不一致，因为存在遗漏变量偏差。

- Cattaneo et al. (2017)放松了假定3.2，转而假定可经过一个变换(transformation)，剔除驱动变量对于潜在结果的直接影响。

结果变量的变换

- **假定3.2a** (结果变量的变换)　存在一个变换 $\phi(\cdot)$，使得对于任何个体 i 满足 $X_i \in W_0$，变换之后的潜在结果仅依赖于处理变量 D_i，即 $\phi\left(Y_i(d,x),d,x\right)=\tilde{Y}_i(d)$。

- 假定3.2a是假定3.2的推广，因为在假定3.2a中，只要令 $\phi(\cdot)$ 为恒等函数 (identity function)，即可得到假定3.2。

- 在实际操作中，可考虑潜在结果的多项式模型。

潜在结果的多项式模型

$$Y_i(d,x)=\begin{cases}\alpha_i(d_i)+\beta_1^-(x_i-c)+\cdots+\beta_p^-(x_i-c)^p & \text{若} d_i=0\\ \alpha_i(d_i)+\beta_1^+(x_i-c)+\cdots+\beta_p^+(x_i-c)^p & \text{若} d_i=1\end{cases}$$

- 其中，$p=1,2,\cdots$ 为多项式的阶数。
- 在应用中，通常令 $p=1$（若 $p=0$ 则回到假定3.2）。
- 此模型允许截距项 $\alpha_i(d_i)$ 因个体而异，以捕捉过滤直接效应之后的处理效应。

变换后的结果

- 变换后的潜在结果为

$$\tilde{Y}_i(d)=\begin{cases}Y_i(d,x)-\beta_1^-(x_i-c)-\cdots-\beta_p^-(x_i-c)^p & \text{若} d_i=0\\ Y_i(d,x)-\beta_1^+(x_i-c)-\cdots-\beta_p^+(x_i-c)^p & \text{若} d_i=1\end{cases}$$

- 为了估计上式的参数，可在窗口内的断点两侧，分别将观测结果对驱动变量进行 p 阶多项式的OLS回归。

- 然后，使用校正结果变量（adjusted outcome）进行费雪推断即可

局部随机化框架的证伪

- 若驱动变量为离散变量，则无法通过驱动变量的密度检验来考察是否存在内生分组。
- Cattaneo et al. (2017)提出在窗口内进行“二项检验”(binomial test)，无论驱动变量连续或离散均适用。
- 假定在窗口内，个体进入处理组的概率为 q，其中 $0 < q < 1$ 。若无额外信息，一般令 $q = 0.5$。
- 在窗口内，受处理个体的数目服从二项分布(binomial distribution)。若观测到的窗口 内受处理个体数目相对于此二项分布为极端值，则可拒绝其服从二项分布的原假设，认为存在个体完全操纵驱动变量的可能性。

留一稳健性检验

- 由于到窗口可能很窄，使得窗口内的观测值较少，故局部随机化框架的估计与推断可能受“离群值”(outliers)较大影响。解决方法之一是，通过画窗口内的散点图，直观考察是否存在离群值，但不严格。

- 本文提出使用“留一估计”(leave-one-out estimation)进行稳健性检验，即每次去掉窗口内的一个观测值，重新进行局部随机化的估计与推断。

- 若某留一估计量 在数值上与窗口内的全样本估计相差很大，则说明局部随机化估计不够稳健，受到极端值很大影响。

四、断点回归两大框架的比较

- 连续性框架始于Hahn et al. (2001)的非参数识别，而成熟于Calonico et al. (2014)的偏差校正稳健估计，其技术已十分完善，是目前断点回归的主流方法。

- 局部随机化框架起步较晚，始于Cattaneo et al. (2015)，经过Cattaneo et al. (2017)的拓展，也趋于成熟。

- 局部随机化框架目前应用较少，仍主要作为稳健性检验或替补方法。

连续性框架隐含的外生性假定

- 在进行线性回归时，一致估计的最基本要求是解释变量为外生变量，即解释变量与扰动项不相关。此结论对于非参数回归依然成立。

- 在连续性框架下，一致估计要求驱动变量在所选最优带宽内为外生变量。

- 假定2.1（驱动变量的连续性）与假定2.2（结果变量条件期望的连续性）均无法保证驱动变量的外生性。

连续性框架隐含的外生性假定（续）

- 连续性框架的最优带宽通常较宽，难以满足局部随机实验的假设。
- 在美国参议院选举的经典案例中（Cattaneo et al., 2015），使用三角核进行局部线性回归所得的MSE最优带宽为[−17.754, 17.754]。
- 并非势均力敌的选举(close election)，难以满足局部随机实验的假设
- 但研究者经常非正式地将最优带宽内的断点回归一概视为局部随机实验（无论此带宽有多宽），而不担心驱动变量可能的内生性。

连续性框架的外生性假定

- **假定2.3** (驱动变量的外生性) 驱动变量 在所选最优带宽内为外生变量，与局部多项式回归方程的扰动项不相关。

- 假定2.3长期为文献所忽视，而实证研究者则通常将其看作理所当然的隐含假定。

- 实证研究者也担心存在内生分组的可能，一般通过密度检验考察在断点处存在个体完全操纵的可能性。但即使密度检验通过，依然无法保证驱动变量在整个最优带宽内的外生性。

外生性假定对实践的影响

- 由于学界对于假定2.3充满信心，以至于经常以一元回归进行断点回归，而不在回归方程中加入任何协变量。
- 即使加入协变量，也认为只是改进了估计效率，而不影响估计的一致性。
- 在使用面板数据进行断点回归时，一般也认为不必考虑个体固定效应，尽管控制个体固定效应可以提高估计效率。

实践建议

- 在使用连续性框架进行断点回归时，应重视引入协变量，以解决可能存在的遗漏变量偏差。

- 若在使用面板数据进行断点回归时，则建议控制个体固定效应，以缓解内生性偏差的顾虑。

- 局部随机化框架在通过协变量平衡检验选择带宽时，已充分考虑满足局部随机化的假定，且所选带宽通常更窄，故假定2.3在局部随机化框架下更易满足，从而在源头上避免了可能的内生性偏差。

离散的驱动变量

- 连续性框架假设驱动变量为连续型随机变量，且在断点处密度为正数（假定2.1）。
- 若驱动变量离散，则使用连续性框架进行断点回归可能会遇到困难。
- 无论驱动变量连续或离散，局部随机化框架均同样适用，无须额外假定。

表 1、连续性框架与局部随机化框架的对比

	连续性框架	局部随机化框架
基本假设	假定 2.1（驱动变量连续，且在断点处密度为正）；假定 2.2（潜在变量的条件期望在断点处连续）；假定 2.3（在所选带宽内，驱动变量为外生变量）	假定 3.1（局部随机化）； 假定 3.2（排他性约束，即驱动变量不直接影响潜在结果）； 假定 3.2a（结果变量的变换）
被估量	在断点处的局部平均处理效应（LATE）	所选带宽内所有个体的平均处理效应（ATE）
带宽选择	选择最优带宽，以最小化均方误差（MSE），或最小化覆盖错误率（CER）	通过一系列协变量平衡检验选择带宽，以满足假定 3.1 的局部随机化要求
估计方法	局部多项式回归，通常为局部线性回归	在假定 3.2 下，使用差分估计量；在假定 3.2a 下，对结果变量变换后，使用差分估计量
统计推断	构建偏差校正的稳健置信区间，进行大样本的统计推断	若带宽内观测值足够多，进行大样本推断；反之，则使用适用于小样本的费雪推断法
证伪验证	（1）画断点回归图；（2）伪断点；（3）以处理前变量或伪结果替代结果变量；（4）驱动变量的密度检验	（1）画断点回归图；（2）伪断点；（3）以处理前变量或伪结果替代结果变量；（4）驱动变量的密度检验或二项检验；（5）留一稳健性检验
适用场景	（1）驱动变量为连续变量（若驱动变量离散，需增加额外假设）；（2）不担心驱动变量在最优带宽内的内生性	（1）无论驱动变量连续或离散，均可适用； （2）担心驱动变量在连续性框架最优带宽内的内生性

五、蒙特卡罗模拟

- 使用蒙特卡罗模拟比较断点回归两大分析框架的差异，特别着重于连续性框架可能存在的内生性偏差 。

- 参考Imbens and Kalyanaraman (2012)与Calonico et al. (2014)，考虑如下包含复合扰动项 $(u_i + \varepsilon_i)$ 的数据生成过程：

$$y_i = 0.42 + \mathbf{1}(x_i \geq 0) + 0.84x_i + u_i + \varepsilon_i \tag{15}$$

其中，真实的处理效应为 1，驱动变量 $x_i \sim 20\mathcal{B}(2,4) - 10$，$\mathcal{B}(p_1, p_2)$ 表示服从参数为 p_1 与 p_2 的 Beta 分布。由于 $\mathcal{B}(p_1, p_2)$ 在 $[0,1]$ 区间取值，故 x_i 的取值范围是 $[-10,10]$。扰动项 $\varepsilon_i \sim N(0,1)$，而不可观测的 u_i 满足下式：

$$u_i = x_i \cdot \mathbf{1}\left(|x_i| > 1\right) \tag{16}$$

方程(16)表明，如果 $|x_i| \leq 1$，则 $u_i = 0$。因此，若限制仅使用满足 $|x_i| \leq 1$ 的样本，则回归方程(15)无内生性。反之，若 $|x_i| > 1$，则 $u_i = x_i$，故一般而言回归方程(15)存在内生性，无法得到一致估计。

蒙特卡罗模拟的过程

- 在进行模拟时，分别使用均匀核、三角核与二次核进行断点回归，并汇报传统点估计（conventional estimate）与偏差校正估计（bias-corrected estimate）。

- 若使用连续性框架，将带宽设为MSE最优带宽

- 若使用局部随机化框架，将带宽设为1，并对结果变量进行线性变换。

- 将样本容量设为1000，而模拟的重复次数设为10,000。

表 2、传统点估计的模拟结果

分析框架	核函数	带宽	均值	偏差	标准差	均方误差
连续性	均匀核	MSE 最优	1.772	0.772	0.502	0.848
连续性	三角核	MSE 最优	1.330	0.330	0.574	0.438
连续性	二次核	MSE 最优	1.436	0.436	0.591	0.540
局部随机化	均匀核	1	0.997	-0.003	0.367	0.135
局部随机化	三角核	1	0.998	-0.002	0.401	0.160
局部随机化	二次核	1	0.998	-0.002	0.388	0.151

注：真实的处理效应为 1，样本容量为 1000，模拟次数为 10,000。

表 3、偏差校正估计的模拟结果

分析框架	核函数	带宽	均值	偏差	标准差	均方误差
连续性	均匀核	MSE 最优	1.756	0.756	0.571	0.897
连续性	三角核	MSE 最优	1.221	0.221	0.646	0.466
连续性	二次核	MSE 最优	1.351	0.351	0.671	0.573
局部随机化	均匀核	1	1.000	0.000	0.559	0.313
局部随机化	三角核	1	1.000	0.000	0.599	0.359
局部随机化	二次核	1	1.000	0.000	0.584	0.341

注：真实的处理效应为 1，样本容量为 1000，模拟次数为 10,000。

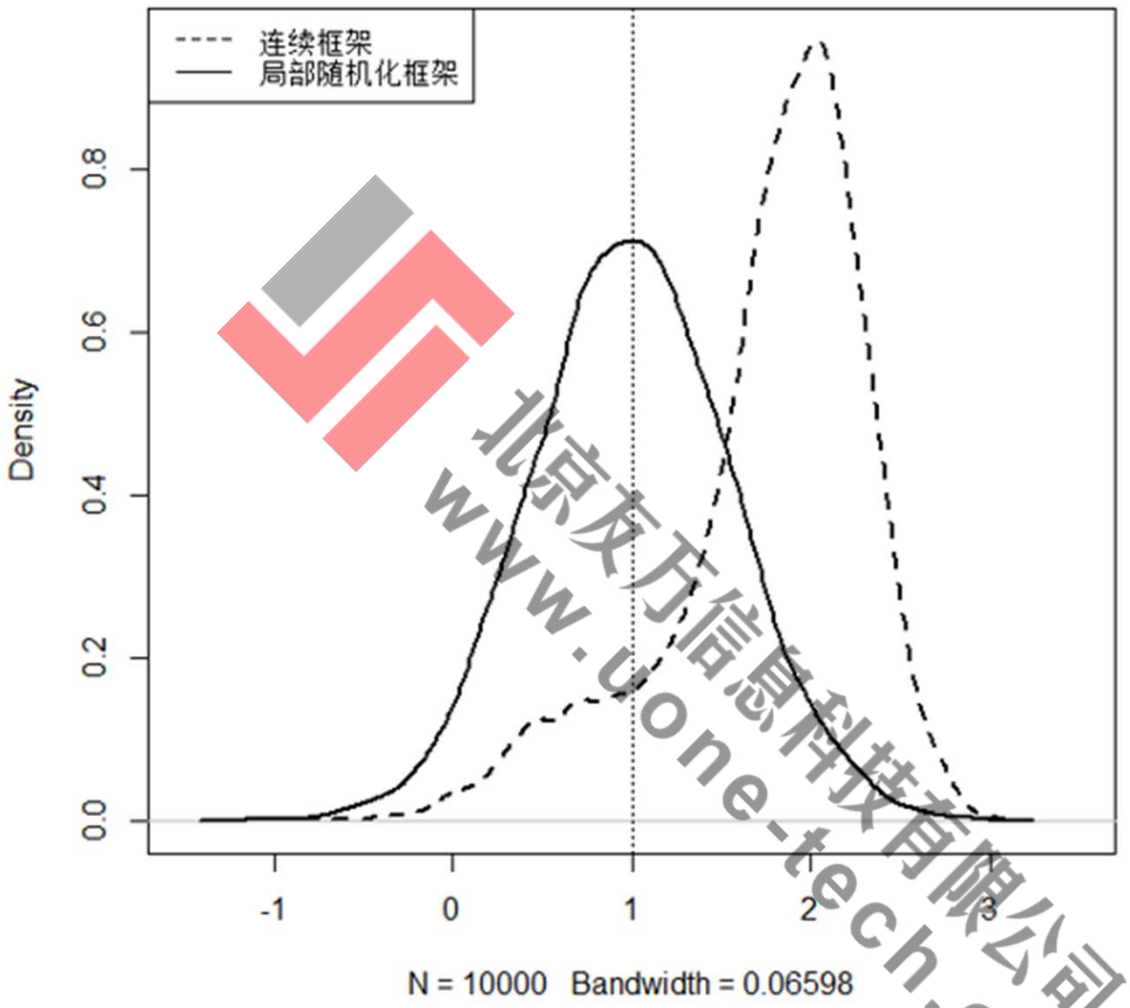


图 3、断点回归估计量的分布

注：估计量为使用均匀核的偏差校正估计。

六、Stata案例

- 使用美国参议院选举的经典案例（Cattaneo et al., 2015），演示断点回归两大框架的不同操作过程与细节。该案例研究美国参议院选举中当选政党的在位优势（incumbent-party advantage）。

- Stata命令包`rdlocrand`自带的数据集rdlocrand_senate.dta包含1914年至2010年美国各州共1390次参议院选举的信息。

变量说明

- 驱动变量 *demmv*（Democratic margin of victory）表示民主党在本次(第*t*期)选举中，得票领先于最强对手的百分数（取值介于-100至100）。如果 *demmv* 取值为正，则民主党获胜当选，故断点在0处。

- 结果变量 *demvoteshfor2*（Democratic vote share in the following election for the same Senate seat）表示共和党在同一参议院议席下次（第*t*+2期）选举的得票百分数（取值介于0至100）。

连续性框架的实证分析

- 本小节的实证分析使用Stata命令rdrobust。

- 采用MSE最优带宽，分别以均匀核、三角核与二次核（伊番科尼可夫核），进行局部线性与局部二次回归。

表 4、连续性框架的估计结果（MSE 最优带宽）

阶数	核函数	带宽	观测值	偏差校正估计	稳健标准误	95%置信区间		区间宽度
1	均匀核	11.597	506	7.594***	1.852	3.963	11.224	7.261
1	三角核	17.754	683	7.507***	1.741	4.094	10.919	6.825
1	二次核	16.104	633	7.300***	1.768	3.835	10.766	6.931
2	均匀核	18.765	717	8.177***	2.102	4.057	12.297	8.240
2	三角核	22.256	779	8.317***	2.093	4.214	12.419	8.205
2	二次核	20.251	739	8.171***	2.129	3.999	12.343	8.344

注：***、**、*分别表示在 1%、5%、10%的显著性水平。

表 5、连续性框架的估计结果（CER 最优带宽）

阶数	核函数	带宽	观测值	偏差校正估计	稳健标准误	95%置信区间		区间宽度
1	均匀核	8.104	376	6.317***	1.982	2.433	10.202	7.769
1	三角核	12.407	532	7.682***	1.841	4.074	11.289	7.215
1	二次核	11.254	495	7.415***	1.861	3.768	11.062	7.294
2	均匀核	12.459	535	8.632***	2.367	3.993	13.271	9.278
2	三角核	14.776	604	9.238***	2.333	4.665	13.810	9.145
2	二次核	13.445	564	9.287***	2.400	4.583	13.991	9.408

注：***、**、*分别表示在 1%、5%、10%的显著性水平。

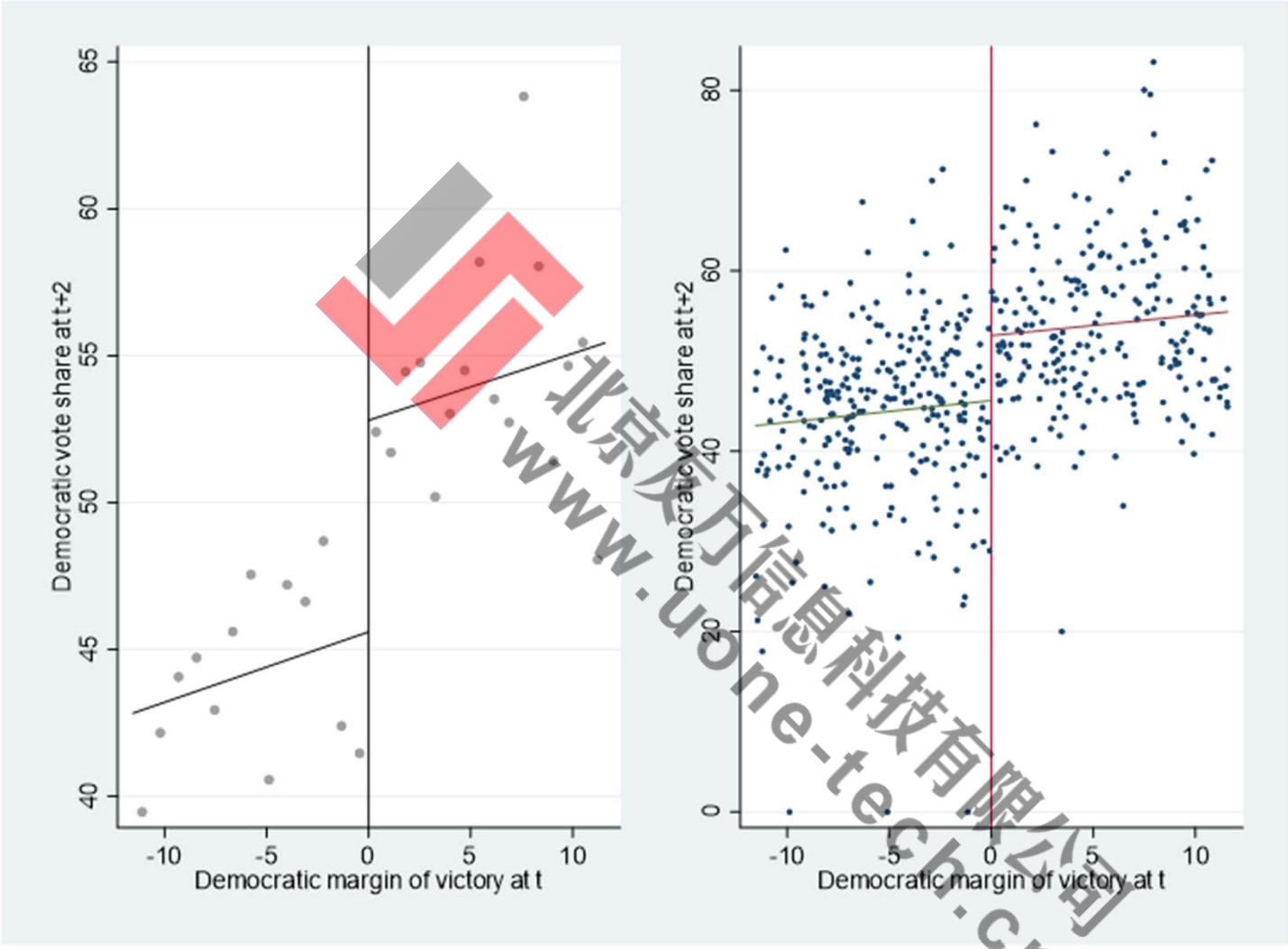


图 4、使用均匀核的散点图与局部线性回归

注：左图为分仓散点图，右图为实际散点图。

局部随机化的实证分析

- 使用Stata命令包rdlocrand，包含4个子命令，即rdwinselect，rdrandinf，rdsensitivity以及rdrbounds

- Stata命令包rdlocrand的下载网址为：https://github.com/rdpackages/rdlocrand

- 在Stata内安装此包：net install rdlocrand, from(https://raw.githubusercontent.com/rdpackages/rdlocrand/master/stata) replace

通过协变量平衡检验选择窗口

- 使用9个协变量进行一系列的协变量平衡检验，包括：

- *presdemvoteshlag1* (上次总统大选所在州的民主党得票百分数)，*population* (州人口数)，*demvoteshlag1* (第*t*-1期参议院选举民主党得票百分数)，*demvoteshlag2* (第*t*-2期参议院选举民主党得票百分数)，*demwinprv1* (民主党是否在第*t*-1期参议院选举获胜)，*demwinprv2* (民主党是否在第*t*-2期参议院选举获胜)，*dopen* (在第*t*期是否有空缺参议院议席)，*dmidterm* (第*t*期是否有中期选举)，*dpresdem* (第*t*期总统是否为民主党)。

局部随机化框架的Stata操作

- sysuse rdlocrand_senate.dta,clear
- global cov presdemvoteshlag1 population demvoteshlag1 demvoteshlag2 demwinprv1 demwinprv2 dopen dmidterm dpresdem

- rdwinselect demmv $cov, cutoff(0) wobs(1) plot

- 选择项“wobs(1)”表示每次在断点两侧至少各增加1个观测值将窗口逐步放大，默认为“wobs(5)”

Window selection for RD under local randomization

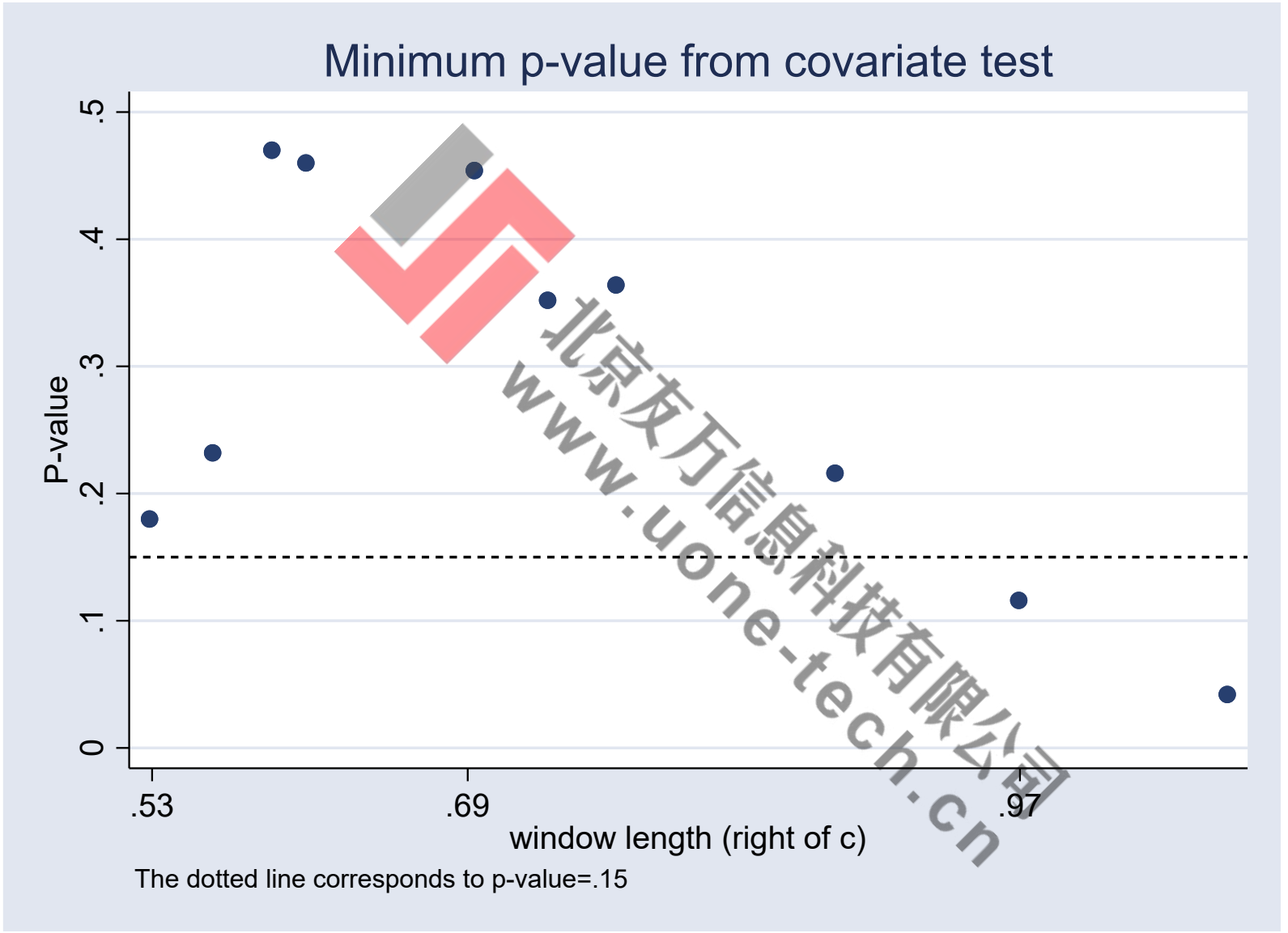
Cutoff c = 0.00	Left of c	Right of c
Number of obs	640	750
1st percentile	6	7
5th percentile	32	37
10th percentile	64	75
20th percentile	128	150

Number of obs	=	1390
Order of poly	=	0
Kernel type	=	uniform
Reps	=	1000
Testing method	=	rdrandinf
Balance test	=	diffmeans

Window		Bal. test p-value	Var. name (min p-value)	Bin. test p-value	Obs<c	Obs>=c
-0.529	0.529	0.180	demvoteshlag2	0.327	10	16
-0.561	0.561	0.232	demvoteshlag2	0.345	11	17
-0.591	0.591	0.470	dopen	0.362	12	18
-0.608	0.608	0.460	dopen	0.377	13	19
-0.693	0.693	0.454	dopen	0.311	14	21
-0.731	0.731	0.352	demvoteshlag1	0.256	15	23
-0.765	0.765	0.364	demvoteshlag1	0.154	15	25
-0.876	0.876	0.216	dopen	0.164	16	26
-0.969	0.969	0.116	dopen	0.135	17	28
-1.075	1.075	0.042	dopen	0.111	18	30

Variable used in binomial test (running variable): demmv
Covariates used in balance test: presdemvoteshlag1 population demvoteshlag1 demv
> oteshlag2 demwinprv1 demwinprv2 dopen dmidterm dpresdem

Recommended window is [-0.876; 0.876] with 42 observations (16 below, 26 above).



随机化框架的统计推断

- `rdrandinf demvoteshfor2 demmv, wl(-0.876) wr(0.876) interfci(.05)`
- `rdrandinf demvoteshfor2 demmv, wl(-0.876) wr(0.876) p(1) interfci(.05)`
- `rdrandinf demvoteshfor2 demmv, wl(-0.876) wr(0.876) p(2) interfci(.05)`

- 选择项“`p(1)`”与“`p(2)`”分别表示对结果变量进行一阶或二阶变换。

- 选择项“`interfci(.05)`”表示显著性水平为5%的置信区间(允许个体间有溢出效应)

对结果变量不做变换

Cutoff c = 0.00	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	17	23
Mean of outcome	41.468	52.420
S.D. of outcome	7.627	7.573
Window	-0.876	0.876

Number of obs =	1297
Order of poly =	0
Kernel type =	uniform
Reps =	1000
Window =	set by user
H0: tau =	0.000
Randomization =	fixed margins

Outcome: demvoteshfor2. Running variable: demmv.

	Finite sample		Large sample	
Statistic	T	P>|T|	P>|T|	Power vs d = 3.81
Diff. in means	10.953	0.000	0.000	0.348

Confidence interval under interference

Statistic	[95% Conf. Interval]	
Diff. in means	5.196	16.643

对结果变量做一阶变换

Cutoff c = 0.00	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	17	23
Mean of outcome	41.468	52.420
S.D. of outcome	7.627	7.573
Window	-0.876	0.876

Number of obs =	1297
Order of poly =	1
Kernel type =	uniform
Reps =	1000
Window =	set by user
H0: tau =	0.000
Randomization =	fixed margins

Outcome: demvoteshfor2. Running variable: demmv.

	Finite sample		Large sample	
Statistic	T	P>|T|	P>|T|	Power vs d = 3.81
Diff. in means	10.994	0.000	0.152	0.079

Confidence interval under interference

Statistic	[95% Conf.	Interval]
Diff. in means	5.612	16.764

对结果变量做二阶变换

Cutoff c = 0.00	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	17	23
Mean of outcome	41.468	52.420
S.D. of outcome	7.627	7.573
Window	-0.876	0.876

Number of obs =	1297
Order of poly =	2
Kernel type =	uniform
Reps =	1000
Window =	set by user
H0: tau =	0.000
Randomization =	fixed margins

Outcome: demvoteshfor2. Running variable: demmv.

	Finite sample		Large sample	
Statistic	T	P>|T|	P>|T|	Power vs d = 3.81
Diff. in means	27.593	0.000	0.033	0.060

Confidence interval under interference

Statistic	[95% Conf.	Interval]
Diff. in means	18.239	36.844

对比一阶与二阶变换的拟合

- twoway (scatter demvoteshfor2 demmv if demmv>=-0.876 & demmv<=0.876,msize(vsmall) xline(0) ytitle(Democratic vote share at t+2) legend(off)) (lfit demvoteshfor2 demmv if demmv>=0 & demmv<=0.876) (lfit demvoteshfor2 demmv if demmv<=0 & demmv>=-0.876), name(rawscatter_linear,replace)
- graph save rawscatter_linear.gph,replace

- twoway (scatter demvoteshfor2 demmv if demmv>=-0.876 & demmv<=0.876,msize(vsmall) xline(0) ytitle(Democratic vote share at t+2) legend(off)) (qfit demvoteshfor2 demmv if demmv>=0 & demmv<=0.876) (qfit demvoteshfor2 demmv if demmv<=0 & demmv>=-0.876), name(rawscatter_quadratic,replace)
- graph save rawscatter_quadratic.gph,replace

- graph combine rawscatter_linear.gph rawscatter_quadratic.gph

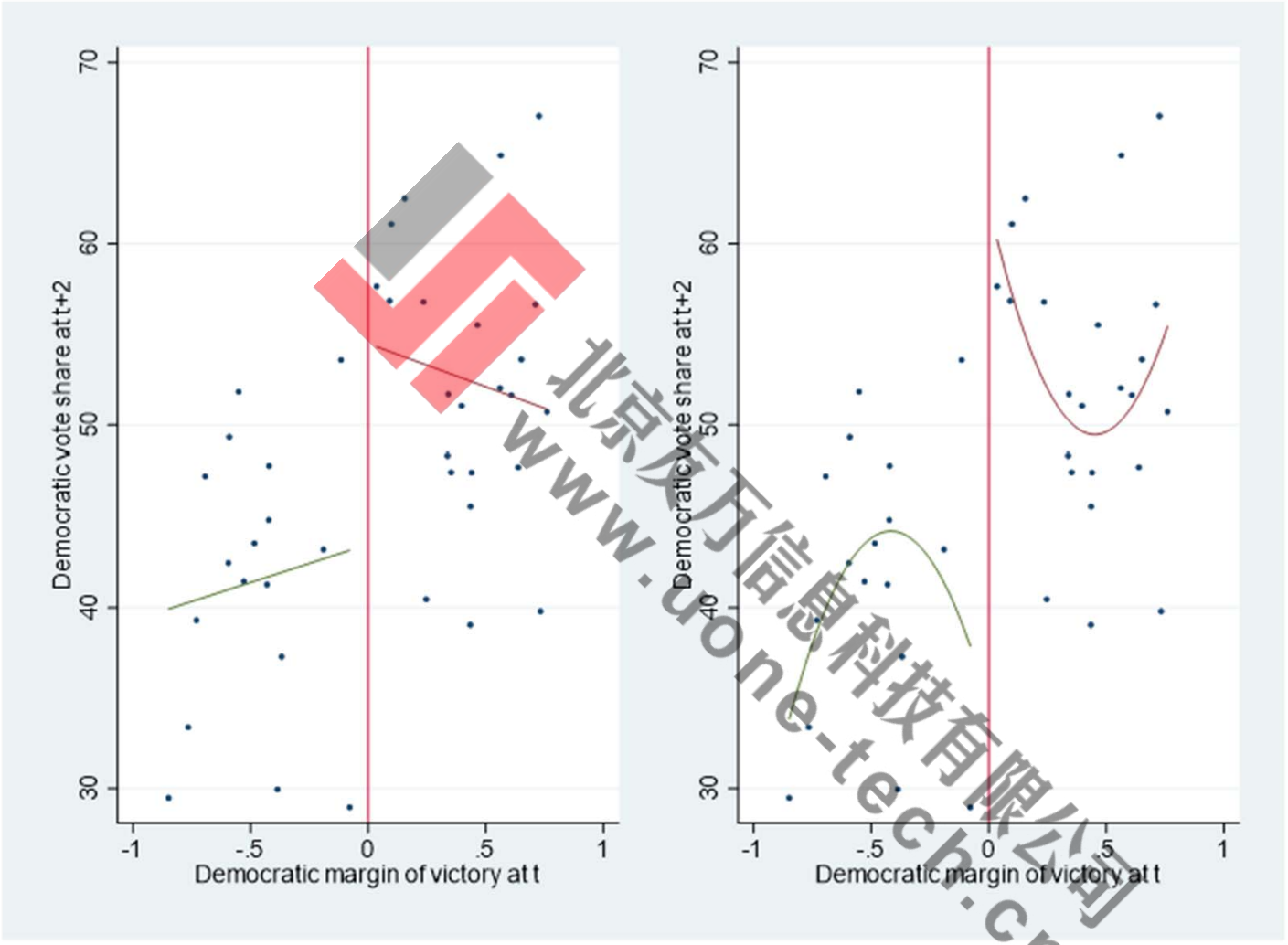


图 6、局部随机化框架的散点图

注：左图为线性拟合，右图为二次拟合。

七、结论

- 当前断点回归的主流方法使用连续性框架，而仅以局部随机化框架作为替补或稳健性检验。

- 连续性框架通过最小化MSE选择最优带宽，但无论带宽多宽，实证研究者均非正式地将其视为局部随机实验，而不考虑驱动变量可能的内生性。

- 局部随机化框架通过一系列协变量平衡检验选择带宽，以满足局部随机化的假定，故所选带宽通常更窄，在源头上排除了驱动变量可能的内生性。

结论（续）

- 连续性框架假定驱动变量连续，故在离散驱动变量的情况下，需要额外假定才能适用。无论驱动变量连续或离散，局部随机化框架均同样适用。

- 连续性框架的被估量仅为断点处的局部平均处理效应，故外部有效性不强；而局部随机化框架可识别在带宽内所有个体的平均处理效应。

- 局部随机化框架的最大缺陷在于，由于所选带宽通常很窄，导致有效样本容量大幅下降，故一般须使用费雪法进行小样本的精确推断。

欢迎评论，谢谢！