断点回归

引言

断点回归（RD）设计作为一种在非实验环境中评估处理效果的方法，由Thistlethwaite和Campbell(1960)首次引入，其中个体是否接受处置取决于观察到的执行变量（在文献中也被称为干预变量或分配变量）是否超过已知的临界点。在最初的断点回归设计应用中，Thistlethwaite和Campbell(1960)分析了奖学金对未来学业成绩的影响，研究设计的主要想法是：分数刚好低于临界值（未获得奖学金）的个体与分数刚好高于临界值（获得奖学金）的个体相比，具有良好的可比性。尽管这种评估策略已经存在了近六十年，但直到最近才引起经济学界的广泛关注。自20世纪90年代末以来，越来越多的研究通过构造准自然实验来评估各种经济背景下的项目效应，相应的利用RDD评估项目效应的研究愈加丰富，如失业保险对劳动力供给的影响、医疗补助对健康的影响、教育补助对学业的影响以及工会对工资和就业的影响。

目前，可能由于如下两个原因推动RDD广泛应用。第一个原因是Hahn.et al(2001)指出RDD与其他非实验方法相比，只需要非常宽松的假设。另一个原因是大多数研究者认为RDD的因果推断可能比典型的“准自然实验”的因果推断结果更可信。Lee（2008）指出，在利用RDD估计时无需假设处置概率与个体独立以使得RDD估计满足随机实验条件，相反，个体在一定程度上可以干预是否处置的状态但无法完全干预处置，从而自然而然具有随机实验性质，这为以上两个原因提供了理论基础。虽然RDD最初被认为是仅使用一些特定的问题的评估方法，但最近的应用状况

与之前的认识有所区别。除了提供一种高度可信和透明的方法来估计项目效果外，RDD还可以在各种情况下使用，涵盖大量重要的经济问题。综上，我们能够解释了为什么RD方法正在迅速成为实证经济学家的常用工具。尽管RDD在经济学中的重要性日益增加，但人们对RD设计的理解仍不够全面，例如什么条件说明RDD是有效的、什么条件下RDD估计结果是稳健的等问题，目前关于RDD的全面总结还较少。此外，标准的计量经济学教科书中也没有涉及到在实证中利用RDD的具体细节，这使得对RDD感兴趣的研究人员很难正确规范进行相关的项目效应评估。广义上讲，本文的主要目标是对RDD全面概述来填补这些空白，勾勒出RDD应用的全貌，为广大实证研究者提供一份应用指南。

本文结构安排如下，第二节主要对断点回归设计思想和基本识别条件进行概述，第三节主要包括断点回归的图形分析，第四节主要包括断点回归的估计，第五节主要包括断点回归的带宽选择，第六节主要包括断点回归的模型设定检验。

断点回归设计

断点回归最早由 Thistlethwait and Camphell (1960)在研究奖学金对学生未来成绩影响的时候提出。当成绩满足某一特定门槛时, 学生将获得奖学金的资助, 成绩在门槛附近两边的学生具有很好的可比性, 因而可以以成绩门槛作为断点来识别奖学金对学生未来成绩的因果影响。但是, 这种方法适用性有限，直到Hahn et al. (2001) 对 RDD 策略的识别条件、估计方法、统计推断进行了理论上的证明, 随后 RDD 在经济学、政治学及社会学等领域广泛应用。本章主要基于Hahn et al. (2001) 的研究介绍精确断点回归设计 (Sharp RDD) 和模糊断点回归设计 (Fuzzy RDD) 的识别条件和估计方法。

断点回归的设计思想

断点回归设计的基本思想是处理变量D依赖于某些连续变量X，而结果变量Y与X以及其他可能影响Y的控制因素集Z的关系是连续的。那么如果Y在断点处发生跳跃，便可看作是处理变量D的影响。现有文献中的断点回归设计主要有两种类型——精确断点回归 (Sharp RDD)和模糊断点回归 (Fuzzy RDD)。

精确断点回归是指处理变量D以一种确定的方式依赖于某些可观察变量X, $D_i = f(X_i)$，其中X的值是连续的，但是函数$f(X_i)$在已知的X_0处不连续。而模糊断点回归中，D不一定是X的特定函数，但是满足条件概率$f(X_i) = E[D_i|X_i = x] = Pr[D_i = 1|X_i = x]$在$X_0$处也是不连续的。上述两种设计的区别可以 Thistlethwait and Campbell (1960) 的研究为例进行说明。首先，如果成绩超过门槛的学生获得奖学金, 反之则不能获得奖学金，在这种情况下, 是否得到奖学金完全由学生分数决定。则精确断点回归适用于该种情形。然而，若断点X_0左右的个体接受奖学金的可能性不同，即除学习成绩外，个体获得奖学金的概率还受到领导能力(无法观测)的影响，则会产生以下现象：存在部分学生成绩超过门槛值，但没有得到奖学金，而有些学生成绩低于门槛值却得到了奖学金，则这种情况适用于模糊断点回归。在下文中，我们介绍断点回归设计的基本识别条件。

假设 (1.1) (断点假设) 假设极限

$$p^+ = \lim_{x \to x_0^+} E\left[D_i \mid X_i = x\right], \quad p^- = \lim_{x \to x_0^-} E\left[D_i \mid X_i = x\right] \tag{1}$$

存在, 并且 $p^+ \neq p^-$。其中$D_i = D(T_i, \varepsilon)$，$T_i = I(X_i \geqslant x_0)$，如果是精确断点设计，则 $D_i = T_i$ 。$p(x) = E[D_i \mid X_i = x] = \Pr[D_i = 1 \mid X_i = x]$ 为倾向指数，表示特征变量为X的个体接受处理的概率。对于精确断点，$p^+ = 1, p^- = 0$，即断点右侧的个体接受处理, 左侧个体为控制组；对于模糊断点，$p^+ = b, p^- = a$ 并且 $0 \leqslant a < b \leqslant 1$，即断点右侧的个体接受处理的概率高于左侧个体。

假设 (1.2) (连续性假设) $E[Y_{0i} \mid X_i = x]$、$E[Y_{1i} \mid X_i = x]$ 是 x 的函数，并且在 x_0 处连续，即

$$\lim_{\varepsilon \to 0} E\left[Y_{ji} \mid X_i = x_0 + \varepsilon\right] = \lim_{i \to 0} E\left[Y_{ji} \mid X_i = x_0 - \varepsilon\right], \quad j = 0, 1 \tag{2}$$

假设(1.1)和(1.2)分别阐述了关于个体分配和个体的结果变量的分布特征。假设(1.1)是说个体的分配概率在临界值左右存在断点，产生跳跃。对于精确断点, 个体状态 D_i 与断点T_i 完全依从；对于模糊断点, 则存在不完全依从，从而 $D_i \neq T_i$。假设(1.2)则表明结果变量 Y_i 在 $X_i = x$ 处是连续的，通常假设两种潜在结果的条件期望函数在所有点上均是连续函数。如果只关心干预组的平均因果数应 (ATT), 那么只需要 $E[Y_{0i} \mid X_i = x]$ 在 x_0 处连续。

除了断点假设和连续性假设之外，断点设计需要满足一项更重要的假设，即一般政策效应评估方法都需满足的随机化实验假设，而断点设计只需个体满足局部随机化。仍以Thistlethwait and Campbell (1960) 为例，局部随机化假设要求个体不能精确控制或操纵X使之超过门槛值，即不能精确地决定自身是否接受处置。学生对学习成绩具有一定的控制能力，会通过努力学习以超过门槛值，但学生不能精确地控制自身的成绩并使其超过门槛值。而只要学生不能精确地控制成绩，那么在临界点附近的学生干预状态的分配就近似于完全随机化实验的结果。局部随机化假设可表述如下：

假设 1.3(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

$$\left(Y_{1i}, Y_{0i}\right) \perp D_i \mid X_i \in \delta\left(x_0\right) \tag{3}$$

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

基于以上假设，关于断点设计的识别可以表述为以下定理：

定理 1.1 若**断点假设 (1.1)** 、**连续性假设 (1.2)** 和**局部随机化假设 (1.3)** 成立, 则有

$E[\tau_i \mid X_i = x_0] = \frac{\mu^+ - \mu^-}{p^+ - p^-}$

其中 $\tau_i = Y_{1i} - Y_{0i}$ 为个体因果效应，$\mu(x) = E[Y_i \mid X_i = x]$，$Y_i = Y_{0i} + \tau_i D_i$，$\mu^+ = \lim_{x \to x*0^+} \mu(x)$，$\mu^- = \lim_{x \to x_0^-} \mu(x))$。

定理 1.1表明若个体接受处理的概率在临界值处存在断点，但潜在的结果变量是特征变量的连续函数，并且个体不能精确地控制自身是否接受处理，从而在断点处近似于局部随机化实验，那么断点处的因果效应就可以被识别。对于精确断点，断点完全决定处置分配状态，则 $p^+ = 1$，$p^- = 0$，从而在断点处的平均因果效应为断点处结果平均值的跳跃，即政策干预的影响可以表达为 $E[\tau_i \mid X_i = x_0] = \mu^+ - \mu^-$。

然而，**定理 1.1**依赖于局部随机化假设, 即假设在断点附近处理变量 D_i 近似于完全随机化实验, 从而排除了个人根据预期收益进行的自选择问题。然而这一假设在现实中很难成立，尤其是在模糊断点情形下，一些个体尽管在断点左侧，仍有可能受激励而进入处置组，也可能存在某些个体的特征变量超过断点却不接受处置。这样，研究者就无法保证处置的分配在断点附近独立于潜在结果变量。个体有可能根据潜在的预期收益决定是否接受处置，这种可能的自选择行为使断点左右的个体不具有可比性，从而违反局部随机化假设。在精确断点情况下, 如果个体不能精确地控制特征变量，局部随机化假设满足。但在模糊断点情况下，局部随机化假设往往不能满足, 下面针对模糊断点识别引入另外的假设替代局部随机化假设。

假设 1.4(独立性假设) 假设潜在结果变量Y_{1i}，Y_{0i}，$D_{1i}(x)$，$D_{0i}(x)$ 在断点附近独立于特征变量 X_i, 即

$$\left(Y_{1i}, Y_{0i}\right), D_{ui}(x), D_{0i}(x) \perp X_i, X_i \in \delta\left(x_0\right) \tag{4}$$

其中

$$D_i = \begin{cases} D_{1i}(x) & x \geqslant x_0 \\ D_{0i}(x) & x < x_0 \end{cases} \tag{5}$$

$D_{1i}(x) = D_i(x)$，$x \geqslant x_0$ 表示特征变量 X 在断点右侧时个体 i 的处置状态。同样地，$D_{0i}(x) = D_i(x))$，$x < x_0$ 表示特征变量 X 在断点左侧时个体 i 的处置状态。

假设 (1.5) (单调性假设) 假设断点对所有个体的影响方向是相同的，即存在$\delta > 0$，使得对于任意$x \in \delta(x_0)$，有

$$D_{1i}(x) \geqslant D_{0i}(x) \tag{6}$$

假设 **1.4** 要求断点独立于所有潜在结果，断点本身不会受到潜在结果或个人选择的影响，断点是外生的。用 $T_i = 1\left(X_i \geqslant x_0\right)$ 表示断点的分配，假设 **1.4** 实际上说明所有潜在结果独立于断点的分配, 即 T_i 的分配近似于完全随机化实验。独立性假设仍然要求个体不能完全精确控制参考变量, 从而在断点附近左右 T_i 分配近似于完全的随机化实验，从而提出假设 **1.5** 。基于以上假设，关于断点设计的识别可以表述为新的定理：

定理 1.2 若断点假设 **(1.1)** 、连续性假设 **(1.2)** 和独立性假设 **(1.4)** 以及单调性假设 **(1.5)** 成立, 则

$$\lim_{x \to x_0} E\left[r_i \mid D_{1i}(x) > D_{0i}(x)\right] = \frac{\mu^+ - \mu^-}{p^+ - p^-} \tag{7}$$

断点回归设计的图形分析

断点回归设计与其他识别方法不同之处在于其设计的透明性和清晰性，根据上文的理论分析，RDD的基本识别条件是干预分配概率在临界点会有跳跃，相应的结果变量(outcome variable)在临界点也会有跳跃，而其他影响结果变量的因素在临界点没有跳跃，从而可以将结果变量的跳跃归咎于干预变量。因此，在进行进一步详细的断点回归实证分析之前，通常可以画出干预分配概率与执行变量(running variable)之间的关系图，判断是适用精确断点回归还是模糊断点回归。然后，可以画出结果变量与执行变量之间的关系图，观测结果变量在断点处是否有跳跃，另外也可以观测在执行变量的其他位置是否也存在着跳跃，从而作为一种证伪检验，即如果在非断点位置，仍发现结果变量有跳跃，则说明可能是其他因素引起的，那么，我们就有理由认为在断点处的跳跃可能混杂了执行变量之外的其他因素所影响，则断点回归设计就可能存在问题。另外，为了验证断点回归设计的有效性，如果存在其他影响结果的可观测协变量，则可以画出这些协变量与执行变量之间的关系图，检验它们在临界点处是否存在跳跃，如果其他协变量在间断点处有跳跃，那么，结果的跳跃有可能是这些协变量造成的，就不能完全将结果的跳跃归因为干预变量的影响，从而，断点回归的实证结果可能会存在问题，相反如果其他协变量在断处是连续的，则断点回归结果更为可信。最后，为了保证断点回归设计的有效性，个体不能精确控制执行变量，一种图形检验方法是画出执行变量的分布图，如果个体不能精确控制执行变量，其分布在断点处应该是连续的，如果发现执行变量集中于断点的一侧，则个体可能可以精确控制执行变量。

Lee（2008）分析了美国各地区众议员选举中在位党在竞选中是否具有优势。美国具有两大党派，一党所获选票份额如果超过竞争对手，那么，该党在该地区将成为在位党，Lee以民主党所获选票份额与共和党所获选票份额的差额作为执行变量，间断点为0，只要上次竞选中，执行变量大于临界点0即意味着民主党在位，否则共和党在位。由于两党选票份额差额不可能大于1，作者将执行变量限制在断点左右0.5范围内，因为0.5之外的样本点较少。结果变量为未来竞选中民主党所获得的选票份额、选举成功概率、成为候选人概率，各党的竞选经费、候选人质量等是竞选选票份额的其他协变量。由于民主党份额超过共和党，民主党就成为在位党，不存在不依从的情况，因而Lee（2008）利用精确断点回归分析在位党派的在位优势。下面，根据Lee(2008)的研究，探讨断点回归中的图形分析。

执行变量和结果变量关系图分析

首先可以画出结果变量与执行变量之间的关系图, 看看结果变量是否在间断点处跳跃, 但避免直接利用原始数据画图, 原始数据中噪音太多, 可以通过适当的平均后再画图。通常可以将执行变量划分为一系列区间, 区间的笼度相同, 并且保证断点左边和右边分别在不同的区间里, 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, 可以得到结果变量相对于执行变量的关系图, 可以通过多项式分别对断点两边的点进行拟合, 同时将拟合的曲线描在图上。

具体而言, 选择某一带宽 h，相应断点左右所划分的区间数分别为 K_0 和 K_1, 目的是构造一系列区间 $(b_k, b_{k+1}], k = 1, \cdots, K = K_0 + K_1$，其中

$$b_k = x_0 - (K_0 - k + 1)h \tag{2.1}$$

区间 k 中个体结果的平均值为:

$$\bar{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i \cdot 1\left(b_k < X_i \leqslant b_{k+1}\right) \tag{2.2}$$

其中 N_k 为区间 k 中个体的数量, 可以表示为:

$$N_k = \sum_{i=1}^{N} 1\left(b_k < X_i \leqslant b_{k+1}\right) \tag{2.3}$$

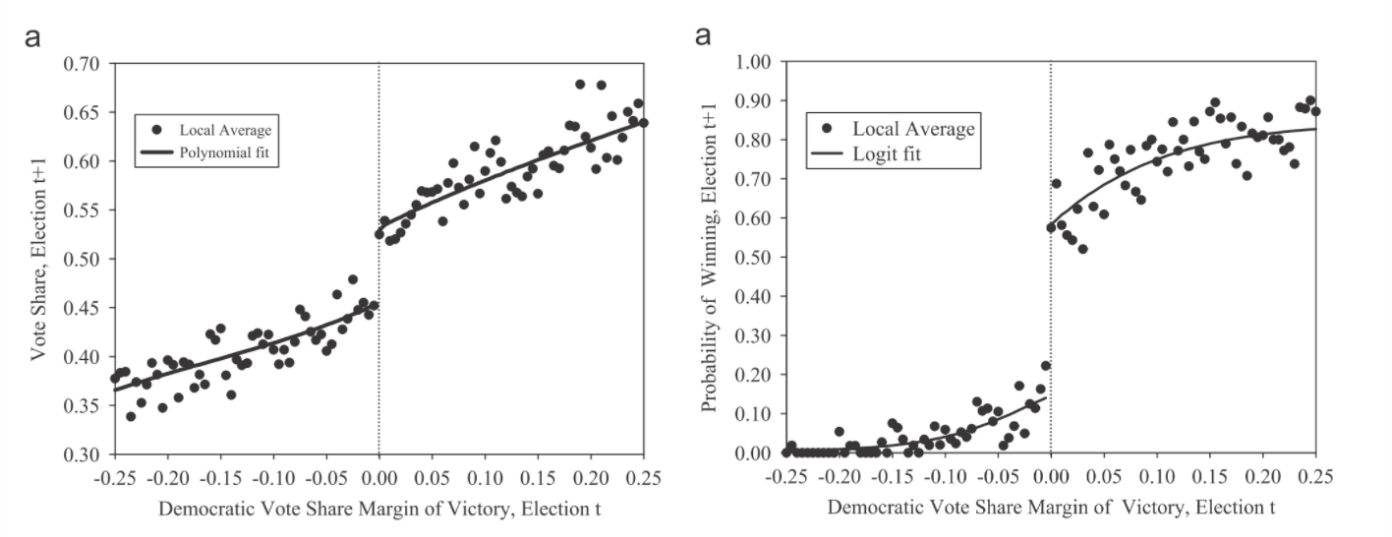

图3-1

图3-1选择0.02的带宽，将执行变量分成了50个区间，图中的点为每个区间中结果变量的平均值，横坐标对应于相应区间的中点，图中曲线是利用4阶多项式分别对断点左右的点拟合而得的。3-1左边图显示，结果变量在断点处有一个非常明显的跳跃，在断右侧，民主党选票份额超过50，而在断点左侧，民主党选票份额为45％左右，相差8％左右，并且在执行变量的其他位置没有发现明显的跳跃。如果将下一次民主党竞选成功的概率作为结果变量，断点则更加明显；3-1右边图显示，断点右侧民主党竞选成功的概率超过60％，而断点左侧民主党竞选成功的概率仅有20％，相差接近40％，即如果民主党为在位党，那么民主党在下一次众议员选举中成功的概率会提高40％，在执行变量的其他位置没有发现明显的跳跃。结果变量与参考变量之间的关系图可以帮助我们观测结果变量在临界点处是否出现间断。另外，也可以帮助我们检测在参考变量的其他位置结果变量是否出现间断。正是由于这一点，使得断点回归方法与其他识别方法相比，更加清晰透明，避免研究者在研究设计上有意或无意的主观偏差，从而使断点回归的识别结果更加可信。

结果变量与执行变量关系图、干预变量与执行变量关系图使我们看到结果变量和干预变量在断点处的行为，但是为了建立结果变量与干预变量之间的因果关系，我们还需要其他影响结果变量的因素在断点处连续变化，因而，我们可以画出其他协变量与执行变量的关系图。

可观测协变量和执行变量关系图分析

为了检验连续性假设是否成立，利用可观测协变量与执行变量之间的关系图来度量。Lee（2008）采用以前民主党在众议员席位竞选中的选票份额来反映民主党的竞争力相应的关系图3-2如下。

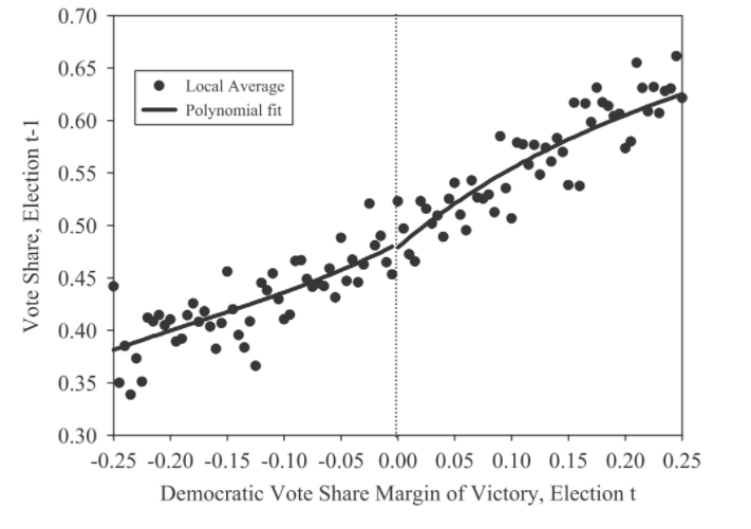

图3-2

图3-2显示，民主党候选人的总体竞争力在临界点处并没有明显的跳跃，从而说明民主党竞选前的竞争力是连续的。如果还可以观测到其他的协变量，比如竞选经费，也可以画出类似的图形检验是否在断点处连续。当然，我们期望其他未观测因素在断点处也是连续的，但无法进行检验，只能利用那些所有可能影响结果的可观测因素进行检测，以期望未观测因素也是符合连续性要求的。如果发现有些协变量在临界点处有间断，那么，说明断点回归设计可能存在问题，结果变量在断点的跳跃有可能是由这一观测因素的跳跃造成的，而不完全是干预的影响。

执行变量的分布图

Lee（2008）基于2001年hahn的文献，提出之前断点回归的缺点，之前断点回归的缺陷体现在两方面，一是假设条件过于严格，没有考虑参与者的自选择问题，二是在检验方面有所欠缺，无法检验RD的有效应。基于这两点，作者首先放松了严格假设条件，在之前的RD分析框架中加入了自选择的假设；其次，提出了检验RD有效性的方法——检验执行变量连续性。因此，建立相对宽松的假设条件，使得断点回归(RD)的因果推断可信度能与随机实验相同。因此，Lee（2008）给出了更宽松条件下仍能有效识别的RDD，其分配机制纳入了个体自选择的因素，其具体步骤如下：首先给出随机实验的分配机制，其次扩展随机实验的分配机制，纳入自我选择因素形成新的断点回归分配机制。

随机实验的分配机制如下：

- 从群体中随机抽取单个个体；
- 以恒定的概率p_0将处置分配给该个体；
- 测量所有变量，包括感兴趣的结果。

根据以上的分配机制，将产生一系列的可观测的随机变量，由Y, X, D组成，其中Y代表结果变量，X代表所有的前定变量，即处置分配之前已经确定的变量，D代表处置状态的虚拟变量。因果推断是基于潜在结果框架进行的，所以随机实验分配机制将产生Y_1, Y_0, X的数据形式。进一步，基于以上分配机制得出相应的条件和命题如下：

条件1a：令 (W,D) 为一对随机变量，其中 W 不可观测，并且令 $Y_1 \equiv y_1(W),\quad Y_0 \equiv y_0(W),\quad X \equiv x(W)$, 其中 $y_1(\cdot), y_0(\cdot)x(\cdot)$ 都是真值函数。
条件1b: 对 W 中所有的 w 都有 $\Pr[D=1 \mid w] = p_0$ 。

根据这两个条件可以推得如下 $(a)(b)(c)$ 三个命题：

$$(a) \quad \Pr[W \leqslant w \mid D=1] = \Pr[W \leqslant w \mid D=0] = \Pr[W \leqslant w] \tag{8}$$
$$(b) \quad \mathrm{E}[Y \mid D=1] - \mathrm{E}[Y \mid D=0] = \mathrm{E}\left[Y_1 - Y_0\right] = ATE \tag{9}$$
$$(c) \quad \Pr\left[X \leqslant x_0 \mid D=1\right] = \Pr\left[X \leqslant x_0 \mid D=0\right], \quad \forall x_0 \tag{10}$$

纳入自我选择断点回归的分配机制如下：

- 在个体作出最优选择之后，从总体中随机抽取单个个体；
- 分配一个分数V，从一个非退化、足够平滑的个体特定概率分布中抽取；
- 处置状态取决于分数V，$D = 1\left[V \geq v_0\right]$；
- 测量所有变量，包括感兴趣的结果。

条件2a：令 (W,V) 为一对随机变量，其中 W 不可观测，V 可观测；并且令 $Y_1 \equiv y_1(W), Y_0 \equiv y_0(W),$ $X \equiv x(W)$, 其中 $y_1(\cdot), y_0(\cdot)x(\cdot)$ 都是真值函数。令 $D = 1\left[V \geq v_0\right], G[\cdot]$ 为 W 的边缘分布函数。
条件2b: $F(v \mid w)$ 是 V 条件于 W 的概率分布函数，$0 < F(0 \mid w) < 1$ ，且当 $v=0$ 时对任意 w 是连续可微的。$f(\cdot)$ 和 $f(\cdot \mid \cdot)$ 分别是 V 和 V 条件于 W 的密度函数。

根据这两个条件可以推得如下 $(a)(b)(c)$ 三个命题。

$$(a)\Pr[W \leqslant w \mid V=v]\text{在}v=0\text{处连续，}\forall w \tag{11}$$

$$\begin{aligned}(b)\mathrm{E}[Y \mid V=0] - \lim_{\Delta \to 0^-} \mathrm{E}[Y \mid V=\Delta] &= \mathrm{E}\left[Y_1 - Y_0 \mid V=0\right] \\ &= \int_{-\infty}^{\infty} (y_1(w) - y_0(w)) \frac{f(0 \mid w)}{f(0)} \mathrm{d}G(w) \\ &= ATE^*\end{aligned} \tag{12}$$

$$(c)\Pr\left[X \leqslant x_0 \mid V=v\right]\text{在}v=0\text{处连续，}\forall x_0 \tag{13}$$

连续性条件2b对命题$(b)(c)$的局部随机分配结果至关重要，容易看出，如果对于总体中的部分群体来说，V密度函数在截断点不连续，那么命题$(b)(c)$一般不会成立。一般而言，若V密度函数在断点处不连续，通常意味着断点处V的分布集中于一侧，那么说明主体很可能可以完全操纵执行变量，使得断点回归失效。因此，执行变量的连续性表明了主体无法完全操纵执行变量，保证了随机实验的外生性。Lee(2008)的参考变量分布图如下图3-3所示，可以看出，以选票份额作为执行变量时，其在断点两侧基本连续。

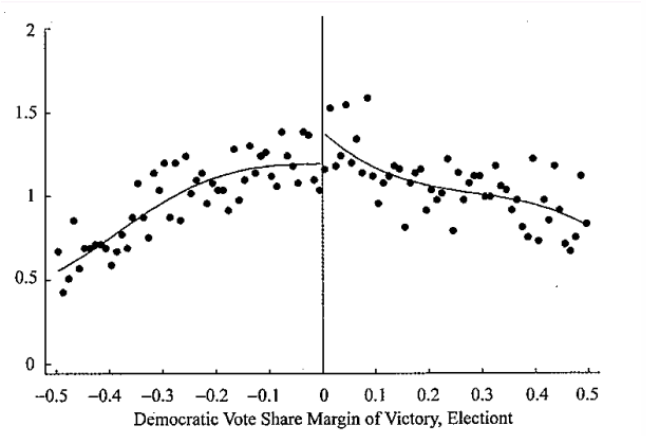

图3-3

断点回归设计的估计

RDD 的估计方法主要有边界非参数回归 (nonparametric regression at the boundary)，局部线性回归 (Local Linear Regression, LLR) 和局部多项式回归 (Local Polynomial Regression, LPR)，由于非参数回归在边界上收敛速度比较慢, 在断点处的估计并不理想, Hahn et al. (2001)， Imbens and Lemieux (2008)等建议采用非参数局部线性回归方法 (LLR)，从而避免在边界上收玫速度慢的问题。下面我们分别对三种估计方法 进行介绍。

边界非参数回归

以精确断点回归的估计为例, 要估计的因果效应参数为断点处的平均因果效应:

$$\tau_{\mathrm{ATE}}^{\mathrm{SRD}}=E\left[Y_{1i}-Y_{0i} \mid X_i=x_0\right]=\mu^{+}-\mu^{-} \tag{14}$$

可以利用标准的非参数回归估计 μ^{+}和 μ^{-}, 假设我们使用核 $K(u)$ 满足 $\int K(u)=1$，则在点 x 处的非参数回归函数可以写成:

$$\hat{\mu}^{-}(x)=\frac{\sum_{i, X_i<x} Y_i \cdot K\left(\left(X_i-x\right) / h\right)}{\sum_{i, X_i<x} K\left(\left(X_i-x\right) / h\right)} \tag{15}$$

$$\hat{\mu}^{+}(x)=\frac{\sum_{i, X_i \geqslant r} Y_i \cdot K\left(\left(X_i-x\right) / h\right)}{\sum_{i, X_i \geqslant x} K\left(\left(X_i-x\right) / h\right)} \tag{16}$$

其中 h 为带宽。
例如，利用矩形核函数 $K(u)=1/2 \cdot 1(|u|<1)$, 相应估计量可以写为:

$$\begin{aligned}
\hat{\mu}^{-}(x) =& \frac{\sum_{i=1}^N Y_i \cdot 1\left(x-h \leqslant X_i<x\right)}{\sum_{i=1}^N 1\left(x-h \leqslant X_i<x\right)} \\
\dot{\mu}^{+}(x) =& \frac{\sum_{i=1}^N Y_i \cdot 1\left(x<X_i \leqslant x+h\right)}{\sum_{i=1}^N 1\left(x<X_i \leqslant x+h\right)}
\end{aligned} \tag{17}$$

则 RDD 估计量为:

$$\hat{\tau}_{\mathrm{ATP}}^{\mathrm{SRD}}=\hat{\mu}^{+}\left(x_0\right)-\hat{\mu}^{-}\left(x_0\right) \tag{18}$$

对于矩形核函数, RDD 估计量实际上是断点左右 h 范围内观测结果平均值之差，即断点右边 $\left[x_0, x_0+h\right]$ 结果变量 Y 的平均值与断点左边 $\left[x_0-h, x_0\right]$ 结果变 量 Y 平均值之差。

局部线性回归

由于非参数回归在边界点上估计效果不佳，在断点回归设计的估计中，通常建议利用局部线性回归方法(LLR), 局部线性回归方法可以避免边界问题 (Hahn et al., 2001)。简单而言，刚才利用矩形核实际上是对断点两边 h 范围内个体进行局部平均，现在分别在断点左右两边 h 范围内利用线性回归进行拟合，利用回归调整执行变量不同而造成的可能偏差。无论真正的潜在结果与执行变量之间是什么样的函数形式，即使是高度非线性的，只要带宽 h 足够小，线性回归函数都将是条件期望函数非常好的近似，这一回归调整在执行变量 X 也会影响结果变量的时候尤其重要。具体地，利用断点左右 h 范围内的样本分别估计下列两个线性回归模型:

$$\min_{a_i,b_i}\sum_{i=1}^{N}\left(Y_i-a_i-b_i\cdot\left(X_i-x_0\right)\right)^2\cdot K\left(\frac{X_i-x_0}{h}\right)\cdot 1\left(X_i<x_0\right) \tag{19}$$

$$\min_{a_i,b_i}\sum_{i=1}^{N}\left(Y_i-a_i-b_i\cdot\left(X_i-x_0\right)\right)^2\cdot K\left(\frac{X_i-x_0}{h}\right)\cdot 1\left(X_i\geqslant x_0\right) \tag{20}$$

其中 $K(u)$ 为核函数, 如果是矩形核, 上述两个方程可以写为:

$$\min_{a_i,b_i}\sum_{i=1}^{N}\left(Y_i-a_i-b_l\cdot\left(X_i-x_0\right)\right)^2\cdot 1\left(x_0-h\leqslant X_i<x_0\right) \tag{21}$$

$$\min_{a_r,b_r}\sum_{i=1}^{N}\left(Y_i-a_r-b_r\cdot\left(X_i-x_0\right)\right)^2\cdot 1\left(x_0\leqslant X_i\leqslant x_0+h\right) \tag{22}$$

估计上述两个方程得到相应的估计值:

$$\begin{aligned}\hat{\mu}^{-}\left(x_0\right)&=\dot{a}_l+\dot{b}_l\left(x_0-x_0\right)=\dot{a}_l\\ \hat{\mu}^{+}\left(x_0\right)&=\dot{a}_l+\dot{b}_r\left(x_0-x_0\right)=\hat{a}_r\end{aligned} \tag{23}$$

从而得到在断点处的平均因果效应其实是两条局部线性回归曲线在断点处的截距之差, 即

$$\hat{\tau}_{\mathrm{ATE}}^{\mathrm{SRD}}=\hat{\mu}^{+}\left(x_0\right)-\hat{\mu}^{-}\left(x_0\right)=\hat{a}_r-\dot{a}_l \tag{24}$$

另外, 也可以利用下面的回归方程得到断点处平均因果效应的直接估计 :

$$\min_{a,b,\tau,\gamma}\sum_{i=1}^{N}1\left(x_0-h\leqslant X_i\leqslant x_0+h\right)\cdot\left(Y_i-a-b\left(X_i-x_0\right)-\tau D_i-\gamma D_i\left(X_i-x_0\right)\right)^2 \tag{25}$$

其中系数 τ 的回归估计量就是 $\hat{\tau}_{ATE}^{\mathrm{SRD}}$ 。

局部多项式回归

如果断点附近样本量太少，为了得到相对比较精确的参数估计，有时我们不得不选择较大的带宽。当带宽较大时，线性近似所造成的偏差可能就会增大，这时局部多项式回归可以捕捉结果变量与执行变量之间的高阶非线性关系，可以得到更好的拟合，从而降低估计偏差。与局部线性回归类似， 局部多项式回归也是利用断点左右的样本分别估计下列模型:

$$\min_{b_l}\sum_{i=1}^{N}\left(Y_i-b_l'x\right)^2K\left(\frac{x_i}{h}\right)1\left(x_i<0\right) \tag{26}$$

$$\min_{b_r}\sum_{i=1}^{N}\left(Y_i-b_r'x\right)^2K\left(\frac{x_i}{h}\right)1\left(x_i\geqslant 0\right) \tag{27}$$

其中，$x_i=X_i-x_0$ 是标准化后的参考变量，临界点变为 0, $b_j=\left(b_{0j},b_{1j},\cdots,b_{pj}\right)'$, $j=l,r, x=\left(1,x_i,x_i^2,\cdots,x_i^p\right)'$ 。相应的 RDD 估计量为:

$$\hat{\tau}_{\mathrm{ATE}}^{\mathrm{SRD}}=\dot{\mu}^{+}\left(x_0\right)-\dot{\mu}^{-}\left(x_0\right)=\hat{b}_{0r}-\hat{b}_{0l} \tag{28}$$

也可以利用下列模型直接估计 RDD 估计量:

$$\min_{b,c}\sum_{i=1}^{N}\left(Y_i-b'x-D_i\cdot c'x\right)^2K\left(x_i/h\right) \tag{29}$$

其中 $b=(b_0,b_1,\cdots,b_p),c=(\tau,\gamma_1,\gamma_2,\cdots,\gamma_p)$, 参数 τ 的回归估计量就是相应的 RDD 估计量。

模糊断点估计

在模糊断点回归中，仍然可以采用上文提到的三种方法，不过需要针对结果变量和干预变量分别对断点进行回归，两个回归得到的参数的比率就是模糊断点回归估计量。下面以局部线性回归为例进行说明，首先利用结果变量对断点进行局部线性回归, 得到估计量 $\hat{\tau}_Y$:

$$\min_{a_r,b_\gamma\cdot\tau_\gamma\cdot\gamma_Y}\sum_{i=1}^{N}K\left(\frac{X_i-x_0}{h}\right)\cdot\left(Y_i-a_Y-b_Y\left(X_i-x_0\right)-\tau_YT_i-\gamma_YT_i\left(X_i-x_0\right)\right)^2 \tag{30}$$

其中 $T_i=1\left(X_i\geqslant x_0\right)$, 然后利用原因变量对断点进行局部线性回归, 得到估计量 $\hat{\tau}_D$:

$$\min_{a_D,b_D\cdot\tau_D\cdot\gamma_D}\sum_{i=1}^{N}K\left(\frac{X_i-x_0}{h}\right)\cdot\left(Y_i-a_D-b_D\left(X_i-x_0\right)-\tau_DT_i-\gamma_DT_i\left(X_i-x_0\right)\right)^2 \tag{31}$$

相应的模糊断点回归估计量为:

$$\dot{\tau}_{\mathrm{LATE}}^{\mathrm{FRD}}=\frac{\hat{\tau}_{\mathrm{Y}}}{\hat{\tau}_D} \tag{32}$$

如果上面方程式采用矩形核进行估计, 并且采用相同的带宽，那么模糊断点回归的估计量也可以用T_i作为D_i的工具变量，利用两阶段最小二乘法估计下列模型:

$$Y_i=\alpha+\tau D_i+\beta(X_i-x_0)+\delta T_i(X_i-x_0)+\epsilon_i \tag{33}$$

其中第一阶段回归为:

$$D_i=\alpha_D+\tau_DT_i+b_D(X_i-x_0)+\gamma_DT_i(X_i-x_0)+\mu_i \tag{34}$$

断点回归设计的最优宽带选择

最优带宽选择的方法概述

根据上述讨论，RDD 的估计显然还依赖于带宽h的选择。若带宽比较小，则断点附近的个体特征差异较小，估计偏差较小。同时，带宽较小意味着回归包含的样本容量较小，估计量方差较大，估计精度较低。因此，带宽的选择通常存在着估计偏差和估计方差的权衡。

Ludwig and Miller (2007) 和 Imbens and Lemieux (2008) 提出了一种选择最优带宽的交叉验证方法 (cross validation)。交叉验证的基本思想是在所有可能的带宽下, 选择使拟合的均方 误差最小的带宽。具体地, 对于给定带宽 (h), 在 (x) 处的回归估计为:

$$\dot{\mu}(x)=\begin{cases}\hat{a}_l(x) & x<x_0\\ \dot{a}_r(x) & x\geqslant x_0\end{cases} \tag{35}$$

其中 $\hat{a}_t(x)$，$\hat{a}_r(x)$ 分别对应于 4.2 中的解 $\hat{a}_t$，$\hat{a}_r$ 。只利用 $x-h<X_j<x+h$，$x=X_i$，$j\neq i$ 估计回归模型, 然后利用估计的回归参数得到在 $x=X_i$ 处的拟合值 $\hat{\mu}\left(X_i\right)$ 。交叉验证标准为:

$$\mathrm{CV}_{\mathrm{Y}}(h)=\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\dot{\mu}\left(X_i\right)\right)^2 \tag{36}$$

相应的最优带宽是最小化上述标准, 即

$$h_{\mathrm{CV}}^{\mathrm{opt}}=\arg\min_{h}\mathrm{CV}_{Y}(h) \tag{37}$$

为提高搜索速度，用 x_δ 表示参考变量在断点左侧的 δ 分位数，$x_{1-\delta}$ 表示参考变量在断点右侧的 $(1-\delta)$ 分位数，则可以将最优带宽搜索限制在 $x_i \leqslant X_i \leqslant x_{1-\delta}$ 范围之内, 即

$$h_{\mathrm{CV}}^{\delta\mathrm{opt}}=\arg\min_{h}\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\hat{\mu}\left(X_i\right)\right)^2\cdot 1\left(x_\delta \leqslant X_i \leqslant x_{1-\delta}\right) \tag{38}$$

比如 Ludwig and Miller (2007) 在计算最优带宽时，将样本限制在断点左右 5% 的数据范围内。对于模糊断点回归，RDD 估计量分子分母需要进行两次最优带宽选择, 可以采用类似于交叉验证标准估计分母的最优带宽，

$$\mathrm{CV}_{D}(h)=\frac{1}{N}\sum_{i=1}^{N}\left(D_i-\dot{\mu}\left(X_i\right)\right)^2 \tag{39}$$

Imbens and Lemieux(2008)建议分子分母采用相同的带宽, 因而, 可以选择结果方程和选择方程两个最优带宽中最小的那个作为共同的最优带宽, 即

$$h_{\mathrm{CV}}^{\mathrm{opt}}=\min_{h}\left\{\arg\min_{h}\mathrm{CV}_{Y}(h),\arg\min_{h}\mathrm{CV}_{D}(h)\right\} \tag{40}$$

Imbens and KALYANARAMAN(IK,2012)

概念设定

这一部分详细介绍 Imbens and KALYANARAMAN(2012) 关于断点回归最优带宽的研究。他们针对局部线性回归的估计，并得到了在平方误差损失下的渐近最优带宽。具体来说，这种最佳带宽依赖于数据分布，为了获得显式的带宽算法，该研究为这些数据分布的泛函提出了简单和一致的估计，并证明了其对带宽的估计是最优的。

首先定义相关概念：

$$Y_i=Y_i\left(D_i\right)=\begin{cases}Y_i(0) & \text{if } D_i=0\\ Y_i(1) & \text{if } D_i=1\end{cases} \tag{41}$$

$$m(x)=\mathbb{E}\left[Y_i \mid X_i=x\right] \tag{42}$$

其中 Y_i 代表个体结果变量，$D_i=I(X_i\geq c)$ 代表处置变量，c 是门槛值，X_i 是特征变量。然后定义 Y_i 的条件期望 $m(x)=E[Y_i|X_i=x]$，则在断点处的平均处理效应可以表示为

$$\tau_{\mathrm{ATE}}=\mathbb{E}\left[Y_i(1)-Y_i(0) \mid X_i=c\right]=\mu_{+}-\mu_{-} \tag{43}$$

其中$\mu_{+}=\lim_{x\downarrow c}m(x)$ ，$\mu_{-}=\lim_{x\uparrow c}m(x)$。基于局部线性回归，有以下结果：

$$\hat{m}_h(x)=\begin{cases}\hat{\alpha}_{-}(x) & \text{if } x<c\\ \hat{\alpha}_{+}(x) & \text{if } x\geq c\end{cases} \tag{44}$$

其中

$$\left(\hat{\alpha}_{-}(x), \hat{\beta}_{-}(x)\right) = \arg\min_{a,\beta} \sum_{i=1}^{N} \mathbf{1}_{X_i<x} \cdot \left(Y_i - \alpha - \beta\left(X_i - x\right)\right)^2 \cdot K\left(\frac{X_i - x}{h}\right) \tag{45}$$

$$\left(\hat{\alpha}_{+}(x), \hat{\beta}_{+}(x)\right) = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} \mathbf{1}_{X_i>x} \cdot \left(Y_i - \alpha - \beta\left(X_i - x\right)\right)^2 \cdot K\left(\frac{X_i - x}{h}\right). \tag{46}$$

则τ_{ATE} 可以估计为

$$\hat{\tau}_{\text{SRD}} = \hat{\mu}_{+} - \hat{\mu}_{-}, \tag{47}$$

$$\hat{\mu}_{-} = \lim_{x \uparrow c} \hat{m}_h(x) = \hat{\alpha}_{-}(c) \quad \text{and} \quad \hat{\mu}_{+} = \lim_{x \downarrow c} \hat{m}_h(x) = \hat{\alpha}_{+}(c). \tag{48}$$

而本节主要介绍如果选择最优的h。

误差准则与不可行的最佳带宽选择

现有研究通常通过交叉验证或特殊方法来选择带宽，使平均积分平方误差准则(MISE)的近似最小化来选择带宽：

$$MISE(h) = \mathbb{E}\left[\int_x \left(\hat{m}_h(x) - m(x)\right)^2 f(x) dx\right] \tag{49}$$

其中 $f(x)$ 是 X 的密度函数。上述的选择标准与处理效应的估计没有直接的关系：估计中有两个在MISE标准中没有捕捉到的特殊特性。第一，τ_{ATE} 仅取决于 $m(x)$ 的差值。其次，这两个值都是边界值。而**IK**提出的准则是基于平方误差$(\hat{\tau}_{ATE} - \tau_{ATE})^2$在 $h = 0$ 附近渐近展开的期望。定义:

$$\text{MSE}(h) = \mathbb{E}\left[\left(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}\right)^2\right] = \mathbb{E}\left[\left(\left(\hat{\mu}_{+} - \mu_{+}\right) - \left(\hat{\mu}_{-} - \mu_{-}\right)\right)^2\right] \tag{50}$$

$$h^* = \arg\min_{h} \text{MSE}(h) \tag{51}$$

这个标准很难直接使用。问题是，在许多情况下，即使样本容量变得无限，最优带宽h^*也不会收敛于零。这是因为偏离阈值的回归函数的不同部分的偏差可能会被抵消。在这种情况下，最佳带宽h^*可能对实际的分布和回归函数非常敏感。此外，当识别是局部时，似乎不适合基于全局标准的估计。因此，**IK**专注于使MSE(h)的一阶近似最小化的带宽，称之为渐近均方误差或AMSE(h)。

误差的渐近展开式

本小节是推导出MSE(h)的渐近展开式，并正式定义渐近逼近AMSE(h)：

$$\text{AMSE}(h) = C_1 \cdot h^4 \cdot \left(m_{+}^{(2)}(c) - m_{-}^{(2)}(c)\right)^2 + \frac{C_2}{N \cdot h} \cdot \left(\frac{\sigma_{+}^2(c)}{f(c)} + \frac{\sigma_{-}^2(c)}{f(c)}\right) \tag{52}$$

其中 C_1 和 C_2 是核函数：

$$C_1 = \frac{1}{4}\left(\frac{v_2^2 - v_1 v_3}{v_2 v_0 - v_1^2}\right)^2 \quad \text{and} \quad C_2 = \frac{v_2^2 \pi_0 - 2 v_1 v_2 \pi_1 + v_1^2 \pi_2}{\left(v_2 v_0 - v_1^2\right)^2}, \tag{53}$$

$$v_j = \int_0^\infty u^j K(u) du \text{ and } \pi_j = \int_0^\infty u^j K^2(u) du. \tag{54}$$

可行的最佳带宽选择

IK 提出了一种简单接入带宽：

$$\tilde{h}_{\text{opt}} = C_K \cdot \left(\frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \cdot \left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2} \right)^{1/5} \cdot N^{-1/5} \tag{55}$$

为避免计算出不良带宽，**IK** 使用正则化思想来修改带宽估计量，此外还提供了一种具体的算法来实现带宽估计（详见Imbens and KALYANARAMAN(IK,2012)）。

模型设定检验

模型估计完成后，可以进行下列模型设定检验，以判断估计结果的稳健性。

- 协变量连续性检验，也称为伪结果检验（pseudo outcome)。以协变量作为伪结果，利用与前面相同的方法，检验相应的RDD估计量是否显著，如果显著说明这些协变量不符合连续性假设，上文的RDD估计量可能存在问题。
- 参考变量分布连续性检验。如果参考变量分布连续，意味着在断点处个体没有精确操纵参考变量的能力，局部随机化假设成立，从而保证断点附近左右样本能够代表断点处的总体。如果个体能够操纵参考变量，我们将能观测到断点左右个体数量有较大差别，比如很多个体通过操纵到了断点的右侧，那么，在断点右侧的区间中个体数量可能将大大超过断点左侧区间中个体的数量。
- 伪断点检验（pseudo cutoff point)。在参考变量的其他位置，比如断点左右两侧中点位置作为伪断点，利用同样的方法估计RDD估计量，我们知道在伪断点干预效应为零，如果发现伪断点的RDD估计量不为零，则说明我们的RDD设计可能有问题，可能混杂了其他未观测因素的影响，得到的因果效应可能是由其他未观测混杂的跳跃造成的，而不完全是干预的影响。
- 带宽选择的敏感性检验。选择不同的带宽对RDD估计量进行重新估计，检验估计结果是否有较大的变量，如果差异较大，尤其是影响方向有变化，则说明RDD设计可能有问题。

小结

我们讨论了断点回归设计，由于断点回归设计与完全随机化实验非常相似，是完全随机化实验的近亲，RDD策略所得到的因果效应参数是最为可信的，因而，RDD策略成为目前经济学实证工具箱中最受欢迎的识别策略。自从20世纪90年代末经济学家将RDD策略重新挖掘出来，特别是Hahn et al.（2001）从理论上严格证明了RDD策略的识别条件和估计方法，RDD策略的理论及应用文献大量涌现。本章介绍了最基本的断点回归设计，对弯折回归设计，分位数断点回归设计等没有进行讨论，后续研究者可以对这些方面总结。断点回归设计近似于完全随机化实验，具有很强的内部有效性，估计结果具有很强的可信性。但是，也与完全随机化实验一样，RDD策略得到的估计往往只是断点处的平均因果效应，不能简单地推广到其他位置，外部有效性较弱，这是断点回归设计的主要限制，在使用RDD策略的时候，需要将这一点考虑进来。有些时候断点处的因果效应就是我们感兴趣的，RDD策略是回答这类问题的最好工具。但是，如果关心断点之外地方的因果效应，就需要引人一定的假设才能外推，外推有效性依赖于假设是否成立。

参考文献:

[1]Hahn J, Todd P, Van der Klaauw W. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design[J]. Econometrica, 2001, 69(1): 201-209.

[2]Imbens G W, Lemieux T. Regression discontinuity designs: A guide to practice[J]. Journal of Econometrics, 2008, 142(2): 615-635.

[3]Imbens G, Kalyanaraman K. Optimal Bandwidth Choice for the Regression Discontinuity Estimator[J]. The Review of economic studies, 2012, 79(3): 933-959.

[4]Lee D S. Randomized experiments from non-random selection in U.S. House elections[J]. Journal of Econometrics, 2008, 142(2): 675-697.

[5]Thistlethwaite D L. Rival hypotheses for explaining the effects of different learning environments[J]. Journal of Educational Psychology, 1962, 53(6): 310-315.