第 16 章　断点回归

16.1　断点回归的思想

依可测变量选择的一种特殊情形是，有时处理变量D_i完全由某连续变量x_i是否超过某截断点(cutoff point)所决定。

据以进行分组的变量x_i被称为**驱动变量**(running variable 或 forcing variable)或**分配变量**(assignment variable)。

比如，考察上大学对工资收入的影响，并假设上大学与否(D_i)完全取决于由高考成绩x_i是否超过 500 分：

$$D_i = D(x_i) = \begin{cases} 1 & 若 x_i \geq 500 \\ 0 & 若 x_i < 500 \end{cases} \qquad (16.1)$$

其中，D_i是x_i的确定性函数，记为$D(x_i)$。

记不上大学与上大学的两种潜在结果分别为(y_{0i}, y_{1i})。

D_i是x_i的确定性函数，故在给定x_i的情况下，可将D_i视为常数，不可能与任何变量有关系，因此D_i独立于(y_{0i}, y_{1i})，满足条件独立假定(CIA)。

但不宜使用倾向得分匹配法，因为重叠假定并不满足，对于所有处理组成员，都有$x_i \geq 500$；而所有控制组成员都有$x_i < 500$，二者没有交集。

对于高考成绩为 498，499，500 或 501 的考生，可认为他们在各方面(包括可观测变量与不可观测变量)都没有系统差异。

高考成绩的细微差异只是“上帝之手”随机抽样的结果，导致成绩为 500 或 501 的考生上大学(进入处理组)，而成绩为 498 或 499 的考生落榜(进入控制组)。

由于制度原因，仿佛对高考成绩在小邻域$[500-\varepsilon, 500+\varepsilon]$之间的考生进行了随机分组，故可视为**准实验**(quasi experiment)。

由于存在随机分组，故可一致地估计在$x=500$处的**局部平均处理效应**(Local Average Treatment Effect，简记 LATE)，即

$$\tau_{\text{LATE}} \equiv \text{E}(y_{1i} - y_{0i} \mid x = 500) \qquad (16.2)$$

更一般地，断点可以是某常数c，而分配机制(assignment mechanism)为

$$D_i = \mathbf{1}(x_i \geq c) = \begin{cases} 1 & \text{若} x_i \geq c \\ 0 & \text{若} x_i < c \end{cases} \qquad (16.3)$$

其中，$\mathbf{1}(\cdot)$为“示性函数”(indicator function)。

作为驱动变量x_i的确定性函数，$D(x_i)$在断点处存在跳跃(jump)，从 0 跳到 1，参见图 16.1。

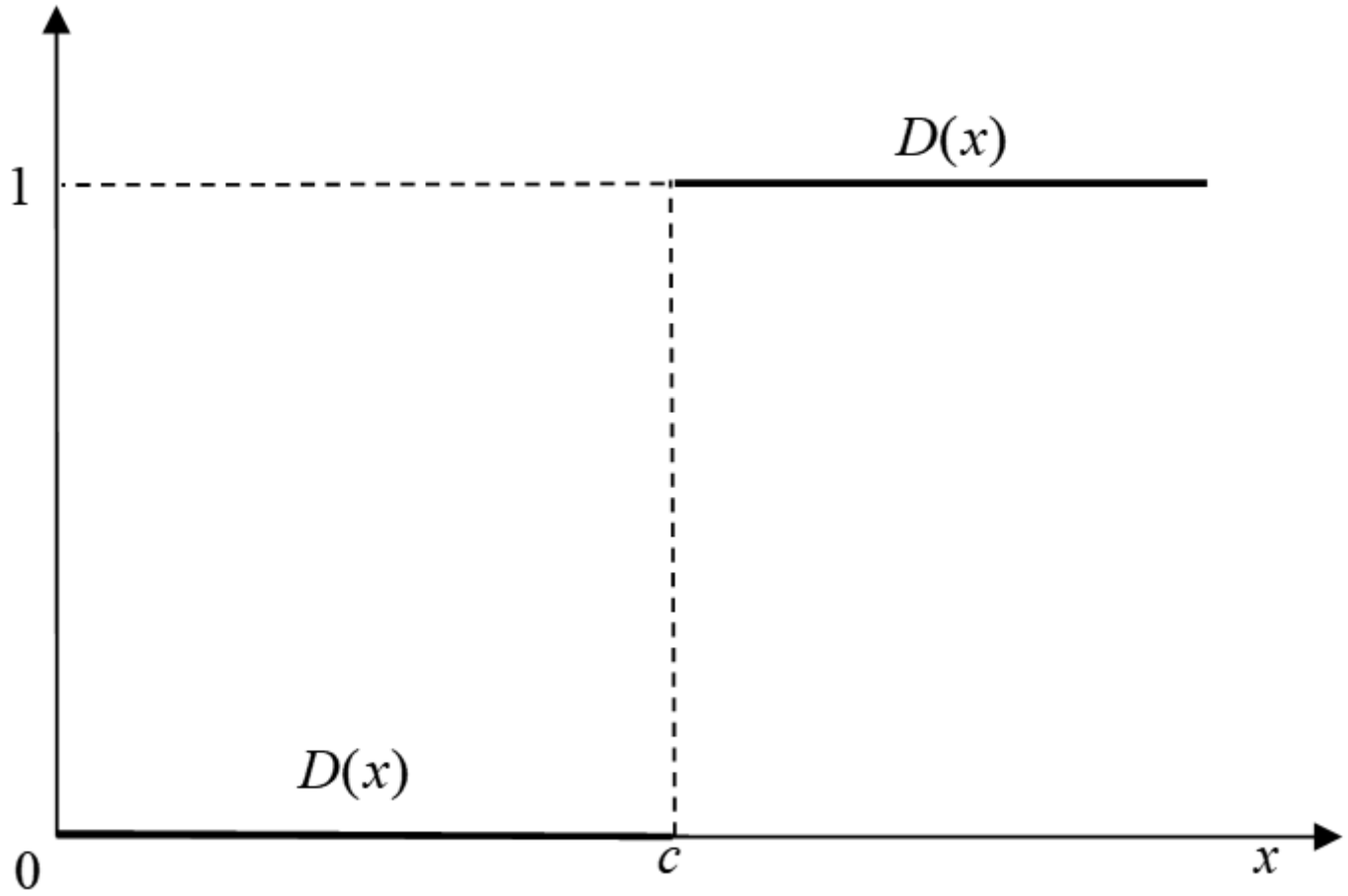

图 16.1 函数$D(x)$的跳跃

由于函数$D(x_i)$在$x=500$处存在一个断点(discontinuity)，这提供了估计D_i对y_i因果效应的机会。

假设在处理前(pretreatment)，结果变量y_i与x_i之间存在如下线性关系：

$$y_i=\alpha+\beta x_i+\varepsilon_i \quad (i=1,\cdots,n) \qquad (16.4)$$

不失一般性，假设D_i对y_i的处理效应为正，则在处理后(posttreatment)，y_i与x_i之间的线性关系在$x=c$处存在一个向上跳跃的断点，参见图 16.2。

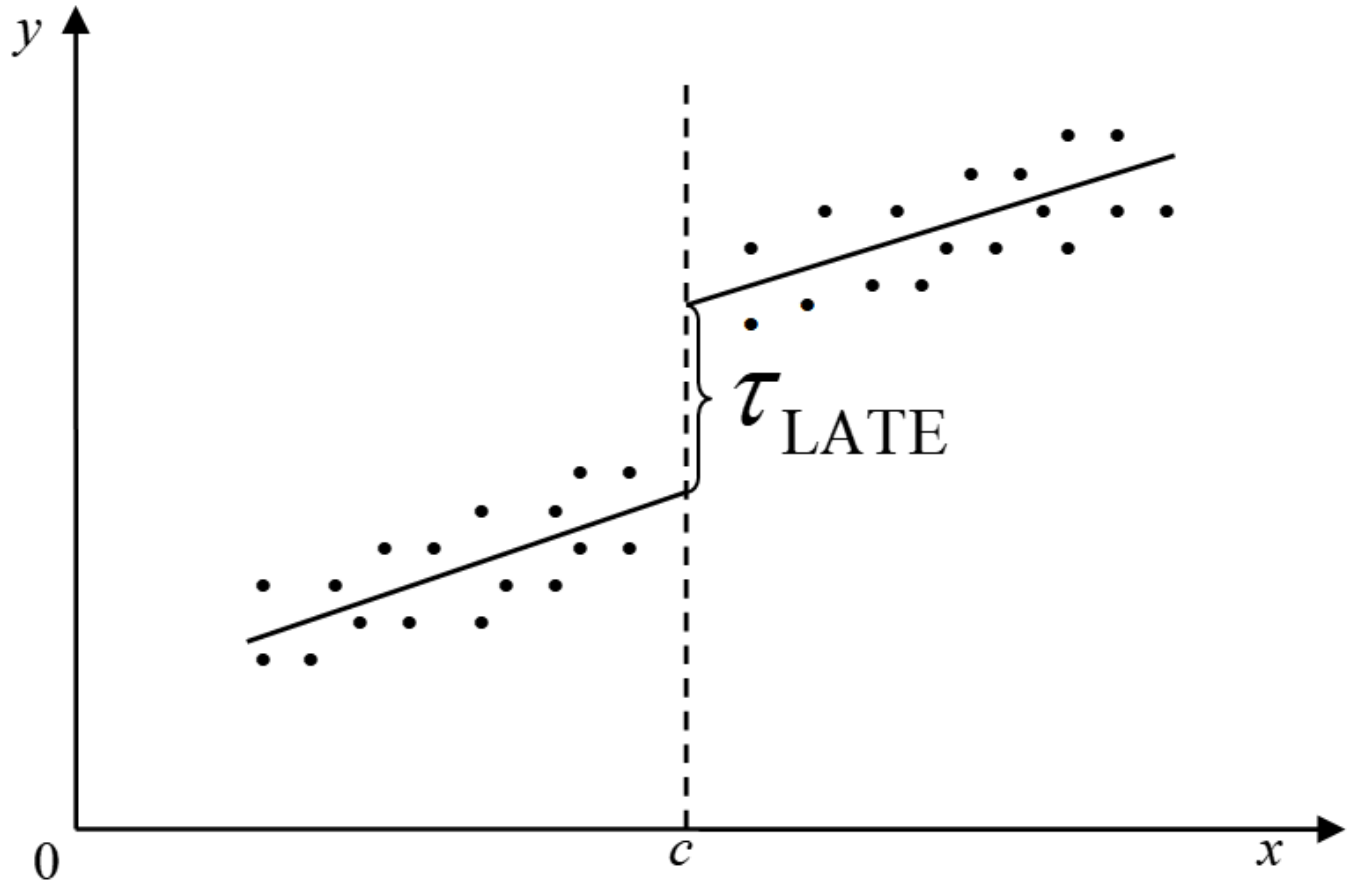

图 16.2 函数$\mathrm{E}(y \mid x)$的跳跃

由于在$x=c$附近，个体在各方面均无系统差别(相当于随机分组)，故造成条件期望函数$\mathrm{E}(y\mid x)$在此跳跃的唯一原因只可能是D_i的处理效应。

可将此跳跃视为在$x=c$处D_i对y_i的因果效应。

将条件期望函数$\mathrm{E}(y\mid x)$在断点c处的右极限(right limit)减去其左极限(left limit)，即可估计此跳跃的高度。

假定潜在结果的条件期望函数$\mathrm{E}(y_{1i}\mid x)$与$\mathrm{E}(y_{0i}\mid x)$在$x=c$处连续，则局部平均处理效应(LATE)可写为

$$
\begin{aligned}
\tau_{\text{LATE}} &\equiv \mathrm{E}(y_{1i} - y_{0i} \mid x = c) && \text{(LATE的定义)} \\
&= \mathrm{E}(y_{1i} \mid x = c) - \mathrm{E}(y_{0i} \mid x = c) && \text{(期望为线性运算)} \\
&= \lim_{x \to c^{+}} \mathrm{E}(y_{1i} \mid x) - \lim_{x \to c^{-}} \mathrm{E}(y_{0i} \mid x) && \text{(潜在结果的条件期望连续)} \\
&= \lim_{x \to c^{+}} \mathrm{E}(y_{i} \mid x) - \lim_{x \to c^{-}} \mathrm{E}(y_{i} \mid x) && \text{(断点回归的分配机制)}
\end{aligned}
$$

(16.5)

其中，$\lim_{x \to c^{+}}$与$\lim_{x \to c^{-}}$分别表示从断点c的右侧与左侧取极限。

由于条件期望函数$\mathrm{E}(y_{1i} \mid x)$在$x = c$处连续，故其函数值等于右极限，即$\mathrm{E}(y_{1i} \mid x = c) = \lim_{x \to c^{+}} \mathrm{E}(y_{1i} \mid x)$。

同理，$\mathrm{E}(y_{0i} \mid x = c) = \lim_{x \to c^{-}} \mathrm{E}(y_{0i} \mid x)$。

根据断点回归的分配机制可知，当$x \to c^{+}$时，$y_{1i} = y_i$；而当$x \to c^{-}$时，$y_{0i} = y_i$。

应如何估计τ_{LATE}？

我们通常将断点标准化为 0，即以$(x_i - c)$作为驱动变量。

为了估计图 16.2 中的跳跃，可将方程(16.4)改写为：

$$y_i = \alpha + \beta(x_i - c) + \delta D_i + \gamma(x_i - c)D_i + \varepsilon_i \quad (i = 1, \cdots, n) \qquad (16.6)$$

引入交互项$\gamma(x_i - c)D_i$是为了允许在断点两侧的回归线斜率可以不同。

对方程(16.6)进行 OLS 回归，所得$\hat{\delta}$就是在$x=c$处局部平均处理效应(LATE)的估计量。

由于此回归存在一个断点，故称为**断点回归**(Regression Discontinuity，简记 RD)或**断点回归设计**(Regression Discontinuity Design，简记 RDD)。

由于在断点附近仿佛存在随机分组，故一般认为断点回归是**内部有效性**(internal validity)比较强的准实验。

在某种意义上，断点回归可视为**局部随机实验**(local randomized experiment)。

可通过考察协变量在断点两侧的分布是否有显著差异来检验此随机性。

断点回归仅度量在断点处的因果关系，一般并不能推广到其他样本值，故**外部有效性**(external validity)受局限。

即使我们知道在高考成绩$x=500$处的处理效应$\tau|_{x=500}$，一般也无法推知在$x=600$处的处理效应$\tau|_{x=600}$。

断点回归最早由心理学家Thistlewaite and Campbell (1960)提出，但直到20世纪90年代末才引起经济学家的重视。

例 奖学金的效应。Thistlewaite and Campbell (1960)使用断点回

归研究奖学金对于学生的影响。

样本由1957年参加美国“全国奖学金竞赛”(national scholarship competition)的高中学生组成，其中5,126名获得“优秀奖”，而另外2,848名仅收到“鼓励信”。

学生获得优秀奖或鼓励信，完全取决于CEEB SQT考试的成绩。

成绩刚好达到获奖标准与差点达到的学生具有可比性。

结果发现，优秀奖得者更可能得到其他渠道的奖学金，但对于其对学术的态度以及职业生涯计划无显著影响。

例 扶贫政策的效应。

在 1994-2000 年，中国实施了“国家八七扶贫攻坚计划”，力争在 20 世纪的最后 7 年，基本解决全国农村 8000 万贫困人口的温饱问题。

该计划将 1992 年农村人均纯收入不足 700 元的 592 个贫困县作为扶贫对象，提供信贷、财税、经济开发三大方面的优惠政策。

Meng (2013)使用县级面板数据，以 1992 年农村人均纯收入为驱动变量，将贫困线 700 元作为断点，进行断点回归。

结果发现，在 1994-2000 年间，八七扶贫攻坚计划使得贫困县农村人均纯收入增长约 38.4%，成效显著。

例 济南户口的价值。

在 2008-2017 年，济南实施的购房落户政策要求，在济南市区购置建筑面积 90 平米以上商品住宅并取得房产证(若为二手房，须取得房产证两年以上)，即可获得济南户籍。

Chen, Shi and Tang (2019)使用 2017 年 6-7 月济南市区 26,031 笔房屋交易数据，以住宅建筑面积为驱动变量，将落户标准 90 平米作为断点，进行断点回归。

结果发现，购房者所获落户权利使房屋单价上升约 1000-1400 元/平米，故济南户口的价值约为 90,000-126,000 元。

16.2 精确断点回归

断点回归可分为两种类型。

一种类型是上节介绍的**精确断点回归**(Sharp Regression Discontinuity，简记 SRD)，其特征是在断点$x=c$处，个体得到处理的概率从 0 跳跃为 1。

另一种类型为**模糊断点回归**(Fuzzy Regression Discontinuity，简记 FRD)，其特征是在断点$x=c$处，个体得到处理的概率从a跳跃为b，其中$0<a<b<1$。

使用线性方程(16.6)来估计精确断点回归，存在两个问题。

首先，如果回归函数包含高次项，比如二次项$(x-c)^2$，则会导致遗漏变量偏差，参见图 16.3。

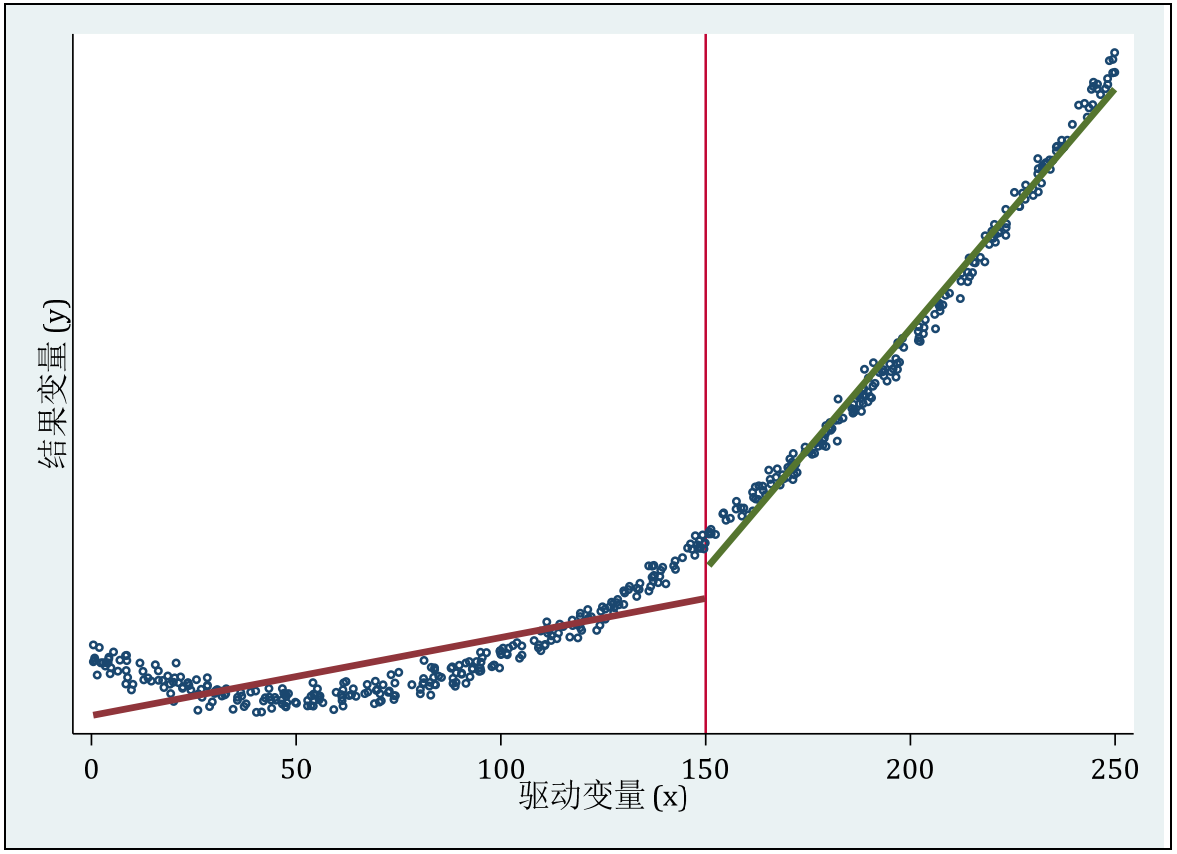

图 16.3 遗漏二次项所导致的偏差

其次，既然断点回归为局部随机实验，则原则上只应使用断点附近的观测值，但方程(16.6)却使用了整个样本。

为解决这两个问题，可在方程(16.6)中引入高次项(比如二次项)，并限定 x 的取值范围为$(c-h, c+h)$：

$$y_i = \alpha + \beta_1(x_i - c) + \delta D_i + \gamma_1(x_i - c)D_i + \beta_2(x_i - c)^2 + \gamma_2(x_i - c)^2 D_i + \varepsilon_i \qquad (c-h < x < c+h) \tag{16.7}$$

其中，$\hat{\delta}$为 LATE 的估计量，而h称为**带宽**(bandwidth)。

但上式并未确定带宽h，且仍依赖于具体的(二次)函数形式。上式是否还遗漏了更高阶项，比如三次或四次项？

在早期的断点回归研究中，常使用三阶、四阶甚至更高阶的多项式进行回归。

高阶多项式的回归结果并不稳定(尤其在边界部分)，故 Gelman and Imbens (2019)建议使用线性或二次的局部回归。研究者开始转向**非参数回归**(nonparametric regression)。

比如，使用**局部线性回归**(local linear regression)，即不假设回归函数的整体形式，而在每个小区间使用线性回归来近似非线性的回归函数(相当于一阶泰勒展开)。

非参数回归的最大优点在于不依赖于具体的函数形式，而且可以选择最优带宽。

局部线性回归最小化加权的残差平方和：

$$\min_{\{\alpha,\beta,\delta,\gamma\}} \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i \underbrace{\left[y_i - \alpha - \beta(x_i - c) - \delta D_i - \gamma(x_i - c) D_i \right]^2}_{\text{残差平方}} \tag{16.8}$$

其中，$w_i = K(z_i) \equiv K\left[(x_i - c)/h\right]$为观测值$i$的权重(weight)，$K(\cdot)$为**核函数**(kernel function)，本质上为权重函数(weighting function)

$z_i \equiv (x_i - c)/h$为x_i离开断点c的标准化距离(衡量此距离为带宽h的几倍)

$\hat{\delta}$为 LATE 的估计量。

局部线性回归的实质是，在一个小邻域$(c-h, c+h)$内进行“加权最小二乘法”(weighted least squares)的估计

权重由核函数$K(\cdot)$来计算，通常离断点c越近的点权重越大，而在区间$(c-h, c+h)$之外则权重降为 0。

核函数类似于随机变量的概率密度函数，也要求曲线下的面积积分为 1。

例如，可用标准正态的概率密度作为核函数，即

$K(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$，称为**高斯核**(Gaussian kernel)。

但断点回归更常用的核函数为三角核、均匀核与二次核。

也可进行**局部二次回归**(local quadratic regression)，只要将(16.7)式中的二次回归函数代入(16.8)式即可。

由于使用断点附近的观测值来估计在断点处的函数值，故非参数回归存在偏差(bias)。

Calonico, Cattaneo and Titiunik (2014)提出了偏差校正(Bias-Corrected，简记 BC)的估计量：

$$\hat{\delta}_{BC} = \hat{\delta} - \widehat{bias}(\hat{\delta}) \qquad (16.9)$$

其中，$\widehat{bias}(\hat{\delta})$为对$bias(\hat{\delta})$(即$\hat{\delta}$的偏差)的估计。

经过偏差校正后，估计量的标准误也会发生变化。

一般称$\hat{\delta}_{BC}$的标准误为**稳健标准误**(robust standard errors)。

在估计偏差$bias(\hat{\delta})$时，仍须使用带宽，通常记为b；此带宽可以不同于估计$\hat{\delta}$所用的带宽h。

16.3 核函数

在进行断点回归时，较常用的核函数包括

(1)“均匀核”(uniform kernel)，也称“矩形核”(rectangular kernel)；
(2) “三角核” (triangular kernel)；

(3) “二次核” (quadratic kernel)，也称 “伊番科尼可夫核”

(epanechnikov kernel)。

(1) **均匀核**的函数形式为

$$K(z)=\frac{1}{2}\cdot\mathbf{1}(|z|<1) \qquad (16.10)$$

其中，$\mathbf{1}(\cdot)$为示性函数。均匀核的图形参见图 16.4。

如果使用均匀核，则为标准 OLS 回归(但仅使用带宽内的样本数据)，等价于上文的参数回归(16.6)。

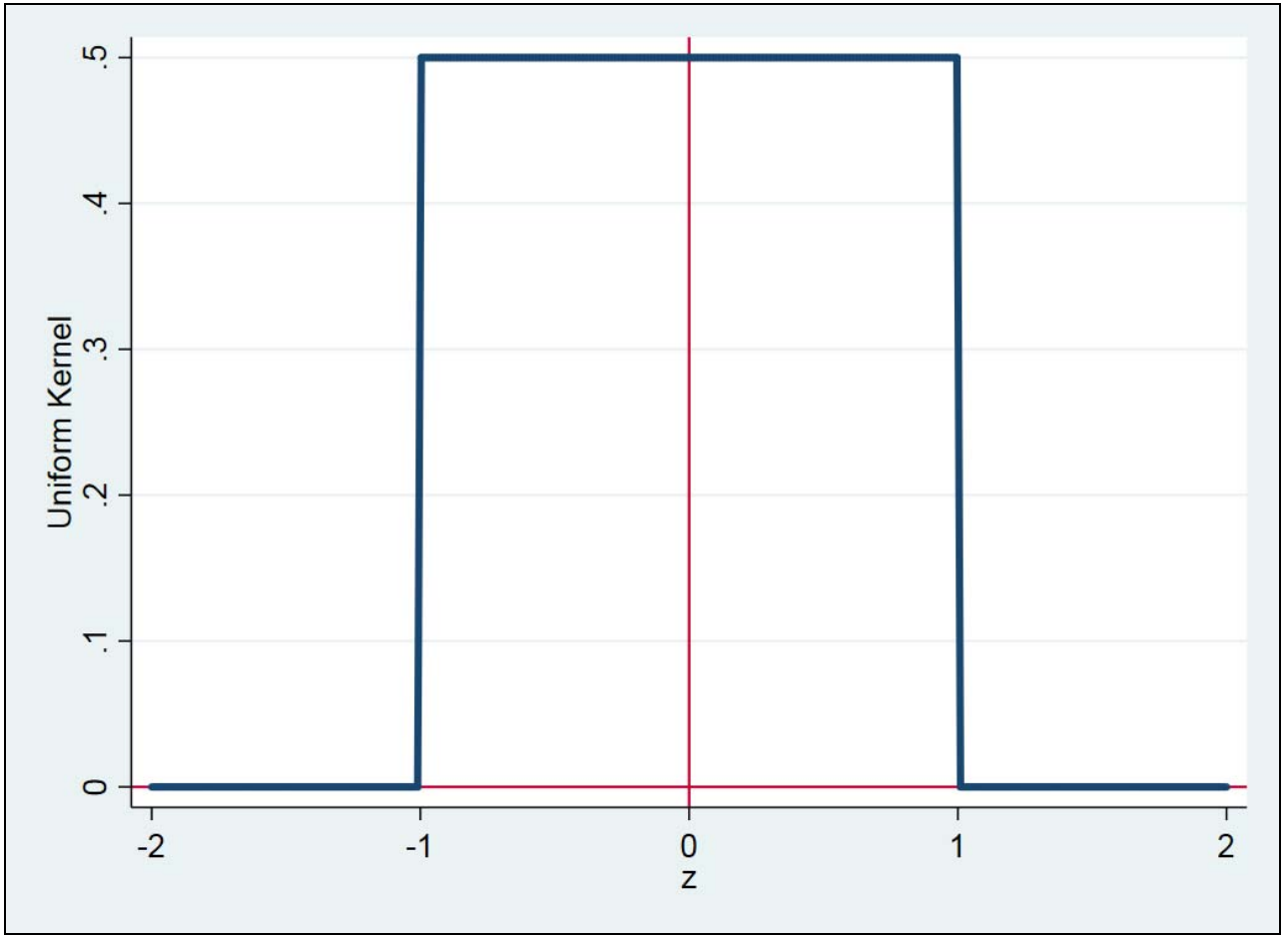

图 16.4 均匀核(矩形核)

(2) **三角核**的函数形式为

$$K(z)=(1-|z|)\cdot \mathbf{1}(|z|<1) \qquad (16.11)$$

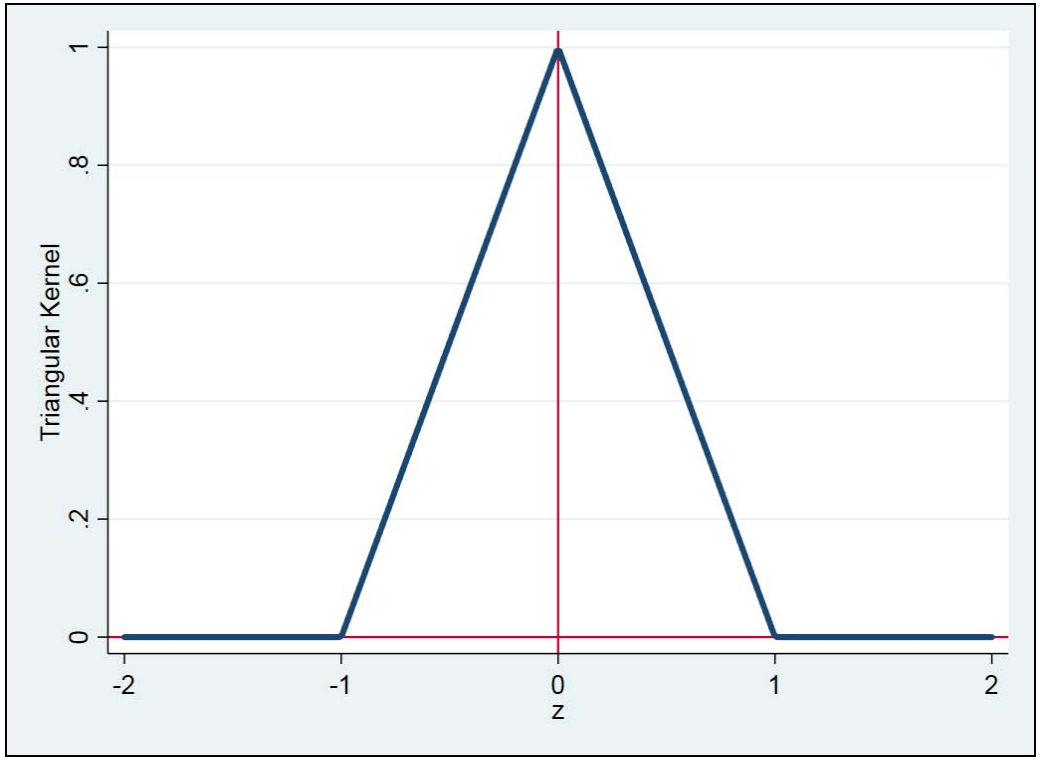

图 16.5 三角核

(3) **二次核**的函数形式为

$$K(z)=\frac{3}{4}(1-z^2)\cdot\mathbf{1}(|z|<1) \qquad (16.12)$$

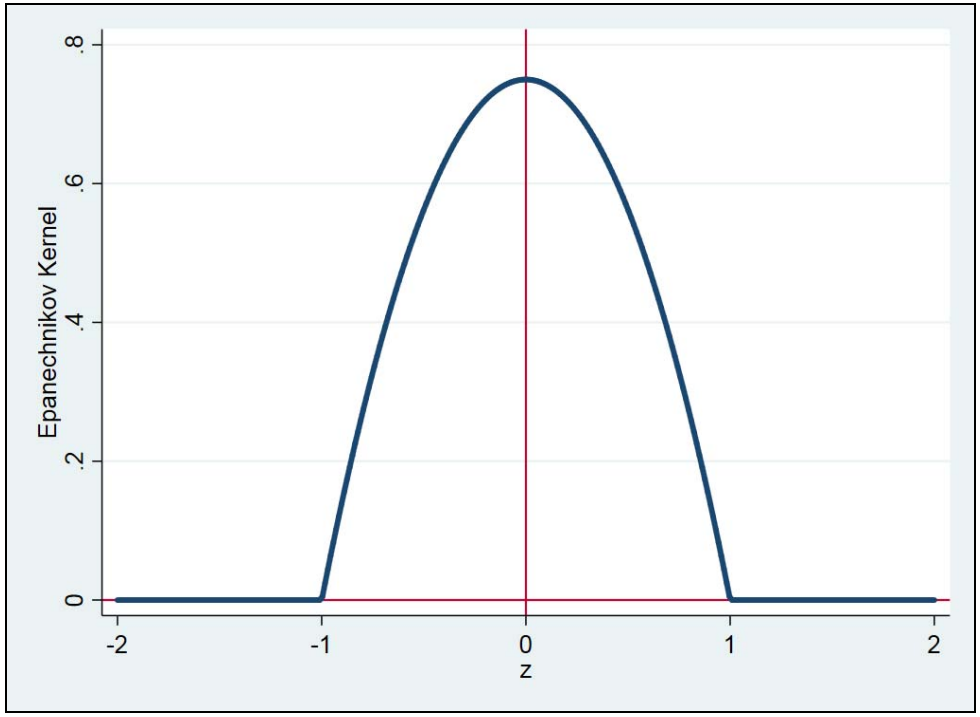

图 16.6 二次核

16.4 带宽

使用不同的核函数，对于非参数回归的影响很小；但带宽的影响则较大。

带宽h越小，则偏差(bias)越小，但离$x=c$很近的点可能很少，导致方差变大。

带宽h越大，则带宽内的观测值越多，故方差越小，但由于包含了离$x=c$较远的点而导致偏差变大。

选择最优带宽涉及**偏差与方差的权衡**(bias and variance trade-off)。

Imbens and Kalyanaraman (2012)提出通过最小化估计量的**均方误差**(MSE)来选择最优带宽h。

Calonico, Cattaneo, Farrell and Titiunik (2020)提出通过最小化置信区间的**覆盖误差率** (Coverage Error Rate，简记 CER)来选择最优带宽h。

覆盖误差(coverage error)定义为“实际覆盖率”减去“名义覆盖率”(即置信度)。

例如，若置信度为 95%，而实际覆盖率仅为 85%，则覆盖误差为 0.85 – 0.95 = –0.1。

随着样本容量趋于无穷，覆盖误差将收敛于 0。最小化 CER 意味着，我们希望覆盖误差以最快的速度(rate)收敛于 0。

由于数据在断点两侧的稀疏程度可能不同，也可在断点两侧使用不同的最优带宽。

如果使用最小化 MSE 的带宽，可得到最优的点估计，即该估计量的 MSE 最小。

如果使用最小化 CER 的带宽，则在统计推断方面更有优势，比如构造覆盖率最接近于置信度的置信区间。

16.5 协变量的作用

在回归方程(16.8)中，也可以加入影响结果变量y_i的其他协变量$\mathbf{w}_i$(Calonico, Cattaneo, Farrell and Titiunik, 2019)。

由于断点回归可视为局部随机实验，故是否包括协变量$\mathbf{w}_i$并不影响断点回归估计量的一致性。

加入协变量$\mathbf{w}_i$的好处在于，如果这些协变量对于被解释变量y_i有解释力，则可以减少扰动项方差，使得估计更为准确。

若所加入的协变量为无关变量，则可能使得估计更不准确。

若加入的协变量为内生变量，与扰动项相关，则反而会干扰对 LATE 的估计。

如果协变量$\mathbf{w}$的条件期望函数$\mathrm{E}(\mathbf{w} \mid x)$在$x = c$处也存在跳跃，则可能也是导致结果变量$y$在断点处跳跃的原因。

断点回归的一个隐含假设是，协变量$\mathbf{w}$的条件期望$\mathrm{E}(\mathbf{w} \mid x)$在$x = c$处连续。

为了检验此假设，可将$\mathbf{w}$中每个变量作为被解释变量，进行断点回归，考察其条件期望是否在$x = c$处有跳跃。

16.6 内生分组

在进行断点回归时，还应注意可能存在**内生分组**(endogenous sorting)的情形。

如果个体事先知道分组规则，并可通过自身努力而完全控制分配变量(complete manipulation)，则可自行选择进入处理组或控制组，导致在断点附近的内生分组，致使断点回归失效。

如果个体事先不清楚分组规则，或只能部分地控制分组变量(partial manipulation)，则一般不存在此担忧。

回到高考成绩的例子，如果考生能够完全控制分配变量x的取值(比如通过自身努力)，则断点回归将失效。

考生无法精确地控制成绩。故在断点附近的考生，成绩大于或小于断点的概率大约都是二分之一，形成局部的随机分组。

如果驱动变量为财务指标或统计指标，则存在人为操作甚至造假的可能性。

若使用地理分界线作为断点，则尤应关注因人口移民而导致内生分组的可能性。

对于内在分组的可能性，可从理论上讨论，也可根据数据进行检验。

假设存在内生分组，则个体将自行选择进入断点两侧，导致在断点两侧的分布不均匀，即分配变量x的密度函数$f(x)$在断点$x=c$处不连续，出现左极限不等于右极限的情形。可检验如下原假设：

$$H_0: \lim_{x\to c^+} f(x) - \lim_{x\to c^-} f(x) = 0 \qquad (16.13)$$

McCrary (2008)提出通过“局部线性密度估计量”(local linear density estimator)的非参数方法进行检验。

Cattaneo, Jansson and Ma (2018)提出更有效率的检验方法，也称为**操纵检验**(manipulation test)，使用更高阶的局部多项式回归估计密度函数，且进行偏差校正，可通过 Stata 命令 `rddensity` 实现。

综上所述，由于断点回归在操作上存在不同选择，故在实践中，一般建议同时汇报以下各种情形，以保证结果的稳健性。

(1) 分别汇报三角核、均匀核与二次核的局部线性回归结果；

(2) 分别汇报使用不同带宽的结果；

(3) 分别汇报包含协变量与不含协变量的结果；

(4) 检验协变量的条件期望是否在断点处跳跃；

(5) 检验驱动变量的密度函数是否在断点处连续(操纵检验)。

16.7 断点回归的 Stata 案例

估计断点回归的 Stata 命令包括 `rd` 与 `rdrobust`。前者较早出现，安装方法为 `ssc install rd`。

命令 `rd` 通过最小化 MSE 来选择最优带宽(Imbens and Kalyanaraman, 2009)，并进行局部线性回归，详见陈强(2014, 第28 章)。

命令 `rdrobust` 提供 10 种不同方法计算带宽，还提供 Calonico, Cattaneo and Titiunik (2014) 的偏差校正估计量、稳健标准误及稳健置信区间。

命令 `rdrobust` 带有一个画图的子命令 `rdplot`。
下载安装 `rdrobust` 的方法：

打开 Github 网页 https://github.com/rdpackages/rdrobust

点击右上方绿色按钮“Code”，在下拉式菜单中点击“Download Zip”。

下载 zip 文件，解压后找到其中名为“Stata”的文件夹，并将其中全部文件存入 Stata 的 ado\plus\r 文件夹即可(可用命令 `sysdir` 查找此文件夹的位置)。

命令 rdrobust 的基本句型为

. rdrobust depvar runvar, c(#) p(#) kernel(uniform) bwselect(cerrd) covs(varlist) all

其中，depvar 为结果变量，runvar 为驱动变量。

选择项 c(#)指定断点位置，默认为 c(0)。

选择项 p(#)指定局部回归的阶数，默认为 p(1)，即局部线性回归。

选择项 kernel(uniform)指定使用均匀核，而默认为三角核 kernel(triangular)，另一备选核函数为二次核 kernel(epanechnikov)。

选择项 bwselect(cerrd)选择最小化 CER 的带宽，而默认选项 bwselect(mserd)选择最小化 MSE 的带宽。

选择项 covs(varlist)指定加入断点回归的协变量。

选择项 all 表示同时汇报三种估计结果，即(1)传统估计值与传统标准误，(2)偏差校正估计值与传统标准误，以及(3)偏差校正估计值与稳健标准误。

命令 rdplot 的基本句型为

```
. rdplot depvar runvar,c(#) p(#)
```

此命令将驱动变量 runvar 分成若干组(bins)，算出结果变量在每组的组内样本均值(sample average within bin)，然后画散点图，以及处理组与控制组的全局回归(global regression)拟合线。

选择项 c(#)指定断点位置，默认为 c(0)。选择项 p(#)指定全局回归的多项式阶数，默认为 p(4)，即四次多项式。

以命令 rdrobust 自带的美国参议院选举数据集 rdrobust_senate.dta 为例演示断点回归的操作。

该数据集以美国的州(*state*)为观测单位，变量 *year* 表示选举年份(election year)，用于研究在任参议员竞选连任时，是否具有“在位优势”(incumbency advantage)。

驱动变量 *margin*(取值介于-100 到 100)表示民主党的领先差额(margin of victory)，定义为在争夺某参议院议席时，民主党的得票百分数(vote share)，减去其最强对手的得票百分数。

结果变量 *vote*(取值介于 0 到 100)表示民主党在该席位下次选举(六年以后)的得票百分数。

协变量包括 *class*(取值 1-3，表示参议院席位的三个类别，对应于不同的选举周期)，*termshouse*(民主党候选人已累计当过几届众议员)，*termssenate*(民主党候选人已累计当过几届参议员)，以及 *population*(州人口)。

结果变量 *vote* 依赖于上次选举的 *margin*。如果 *margin* 大于 0，则民主党当选议员，可能享有在位优势，或可利用其当政六年的政治资源影响下次选举的结果。

以 *margin* = 0 为断点进行断点回归，因为 *margin* 取值刚好大于0(当选为参议员)与刚好小于 0(未选为参议员)的候选人具有可比性。

首先，在安装命令 `rdrobust` 时，已将数据集 rdrobust_senate.dta 存入 Stata 的系统路径，故可用命令 `sysuse` 打开此数据集，并考察变量的统计特征。

```
. sysuse rdrobust_senate.dta,clear

. sum
```

Variable	Obs	Mean	Std. dev.	Min	Max
state	1,390	40.01367	21.99304	1	82
year	1,390	1964.63	28.05466	1914	2010
vote	1,297	52.66627	18.12219	0	100
margin	1,390	7.171159	34.32488	-100	100
class	1,390	2.023022	.8231983	1	3
termshouse	1,108	1.436823	2.357133	0	16
termssenate	1,108	4.555957	3.720294	1	20
population	1,390	3827919	4436950	78000	3.73e+07

其次，使用命令 rdplot，画整个样本的断点回归图，直观考察是否存在断点。

```
. rdplot vote margin
```

命令 rdplot 默认在断点两侧分别画四次回归的拟合图。

RD Plot with evenly spaced mimicking variance number of bins using spacings estimators.

Cutoff c = 0	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	595	702
Order poly. fit (p)	4	4
BW poly. fit (h)	100.000	100.000
Number of bins scale	1.000	1.000

Number of obs = 1297
Kernel = Uniform

Outcome: vote. Running variable: margin.

	Left of c	Right of c
Bins selected	15	35
Average bin length	6.667	2.857
Median bin length	6.667	2.857
IMSE-optimal bins	8	9
Mimicking Var. bins	15	35
Rel. to IMSE-optimal:		
Implied scale	1.875	3.889
WIMSE var. weight	0.132	0.017
WIMSE bias weight	0.868	0.983

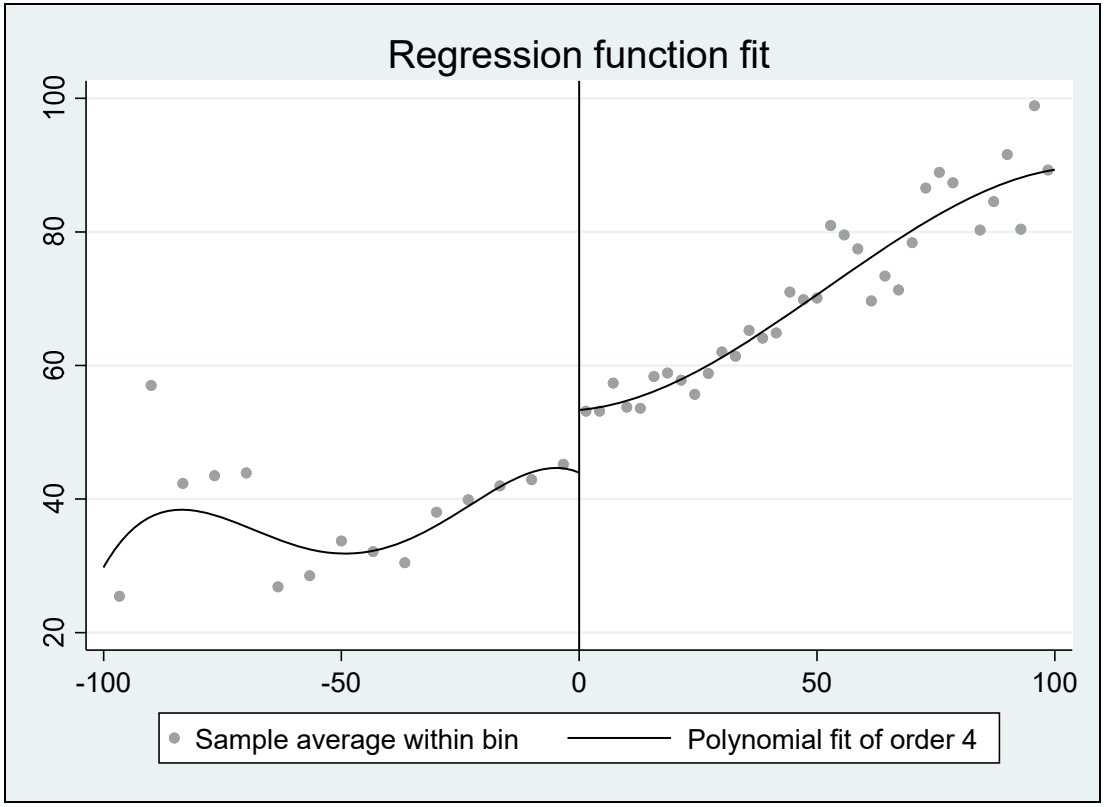

图 16.7 整个样本的断点回归图

结果变量在断点 $margin = 0$ 向上跳跃，或有在位优势。但这只是全局回归(未限制带宽)，且高次回归并不稳定，故仅为示意图。

使用最小化 MSE 的默认带宽与默认的三角核进行断点回归。

```
. rdrobust vote margin
```

```
Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 |  Left of c  Right of c           Number of obs =       1297
                   |                                 BW type       =      mserd
     Number of obs |        595         702           Kernel        = Triangular
Eff. Number of obs |        360         323           VCE method    =         NN
    Order est. (p) |          1           1
    Order bias (q) |          2           2
       BW est. (h) |     17.754      17.754
       BW bias (b) |     28.028      28.028
         rho (h/b) |      0.633       0.633

Outcome: vote. Running variable: margin.
```

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	7.4141	1.4587	5.0826	0.000	4.5551	10.2732
Robust	-	-	4.3110	0.000	4.0937	10.9193

平均处理效应的点估计为 7.41，且在 1%的水平上显著。

使用选择项 all，同时汇报三种结果。

```
. rdrobust vote margin,all
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	360	323
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	17.754	17.754
BW bias (b)	28.028	28.028
rho (h/b)	0.633	0.633

Number of obs	= 1297
BW type	= mserd
Kernel	= Triangular
VCE method	= NN

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	7.4141	1.4587	5.0826	0.000	4.5551	10.2732
Bias-corrected	7.5065	1.4587	5.1460	0.000	4.64747	10.3655
Robust	7.5065	1.7413	4.3110	0.000	4.0937	10.9193

偏差校正的点估计为 7.51，略有增大；相应的稳健标准误也有所增大，但仍在 1%的水平上显著。对比均匀核的估计结果：

```
. rdrobust vote margin,kernel(uniform) all
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	271	235
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	11.597	11.597
BW bias (b)	22.944	22.944
rho (h/b)	0.505	0.505

Number of obs =	1297
BW type =	mserd
Kernel =	Uniform
VCE method =	NN

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	7.2025	1.6129	4.4656	0.000	4.04127	10.3637
Bias-corrected	7.5935	1.6129	4.7080	0.000	4.43226	10.7547
Robust	7.5935	1.8521	4.0999	0.000	3.96341	11.2235

使用选择项 bwselect(msetwo)，允许断点两侧的带宽不同。

```
. rdrobust vote margin,bwselect(msetwo) all
```

```
Sharp RD estimates using local polynomial regression.

      Cutoff c = 0 |  Left of c  Right of c           Number of obs =       1297
                   |                                 BW type       =     msetwo
     Number of obs |       595         702           Kernel        = Triangular
Eff. Number of obs |       336         326           VCE method    =         NN
    Order est. (p) |         1           1
    Order bias (q) |         2           2
       BW est. (h) |    16.170      18.126
       BW bias (b) |    27.104      29.344
         rho (h/b) |     0.597       0.618

Outcome: vote. Running variable: margin.

------------------------------------------------------------------------------
           Method |     Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
     Conventional |     7.4536     1.4972     4.9785   0.000     4.51924       10.388
Bias-corrected    |     7.5335     1.4972     5.0319   0.000     4.59916      10.4679
           Robust |     7.5335     1.7595     4.2817   0.000       4.085       10.982
------------------------------------------------------------------------------
```

使用最小化 CER 的方法选择最优带宽。

```
. rdrobust vote margin,bwselect(cerrd) all
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	284	248
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	12.407	12.407
BW bias (b)	28.028	28.028
rho (h/b)	0.443	0.443

Number of obs	1297
BW type	cerrd
Kernel	Triangular
VCE method	NN

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	7.6316	1.6801	4.5424	0.000	4.3387	10.9244
Bias-corrected	7.6817	1.6801	4.5723	0.000	4.38884	10.9746
Robust	7.6817	1.8406	4.1735	0.000	4.07422	11.2892

使用选择项 bwselect(certwo)，通过最小化 CER 选择最优带宽，但允许断点两侧的带宽不同。

```
. rdrobust vote margin,bwselect(certwo) all
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	595	702
Eff. Number of obs	266	252
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	11.299	12.667
BW bias (b)	27.104	29.344
rho (h/b)	0.417	0.432

Number of obs	=	1297
BW type	=	certwo
Kernel	=	Triangular
VCE method	=	NN

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	8.0175	1.7188	4.6645	0.000	4.64868	11.3864
Bias-corrected	8.0665	1.7188	4.6930	0.000	4.69767	11.4354
Robust	8.0665	1.8683	4.3176	0.000	4.40472	11.7283

在断点回归中引入协变量。

```
. rdrobust vote margin, all covs(class termshouse termssenate)
```

Covariate-adjusted Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	491	617
Eff. Number of obs	315	283
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	18.033	18.033
BW bias (b)	28.988	28.988
rho (h/b)	0.622	0.622

Number of obs	= 1108
BW type	= mserd
Kernel	= Triangular
VCE method	= NN

Outcome: vote. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	6.8499	1.4067	4.8694	0.000	4.09275	9.607
Bias-corrected	6.9884	1.4067	4.9679	0.000	4.2313	9.74556
Robust	6.9884	1.6636	4.2009	0.000	3.7279	10.249

Covariate-adjusted estimates. Additional covariates included: 3

作为稳健性检验，检验协变量 *population* 是否在 *margin* = 0 的断点处存在显著的跳跃：

```
. rdrobust population margin,all
```

Sharp RD estimates using local polynomial regression.

Cutoff c = 0	Left of c	Right of c
Number of obs	640	750
Eff. Number of obs	412	378
Order est. (p)	1	1
Order bias (q)	2	2
BW est. (h)	20.763	20.763
BW bias (b)	33.202	33.202
rho (h/b)	0.625	0.625

Number of obs	= 1390
BW type	= mserd
Kernel	= Triangular
VCE method	= NN

Outcome: population. Running variable: margin.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	-3.2e+05	6.5e+05	-0.4873	0.626	-1.6e+06	962492
Bias-corrected	-3.7e+05	6.5e+05	-0.5681	0.570	-1.7e+06	909721
Robust	-3.7e+05	7.8e+05	-0.4761	0.634	-1.9e+06	1.2e+06

进行操纵检验，考察驱动变量在断点处的密度函数是否连续，以排除内生分组的可能性。

由于命令 `rddensity` 依赖于命令 `lpdensity`(表示 local polynomial density)，故须分别从 Github 下载这两个命令。

安装 `lpdensity` 的方法：打开网页 https://github.com/nppackages/lpdensity，点击右上方绿色按钮“Code”，从下拉式菜单点击“Download Zip”。

下载 ZIP 文件并解压，然后找到其中名为“Stata”的文件夹，以及相应的路径名。比如，若压缩文件放在 F 盘的根目录，则路径名应为“F:\lpdensity-master\lpdensity-master\stata”。

使用该路径名，在 Stata 中输入如下命令即可完成安装：

```
net install lpdensity, from(F:\lpdensity-master\lpdensity-master\stata) replace。
```

安装 rddensity 的方法：打开网页 https://github.com/rdpackages/rddensity，点击右上方绿色按钮“Code”，从下拉式菜单点击“Download Zip”。

下载 ZIP 文件并解压，然后找到其中名为“Stata”的文件夹，以及相应的路径名。

比如，若压缩文件放在 F 盘的根目录，则路径名应为“F:\rddensity-master\rddensity-master\stata”。

使用该路径名，在 Stata 中输入如下命令即可完成安装：net install rddensity, from(F:\rddensity-master\rddensity-master\stata) replace。

下面，使用命令 rddensity 进行操纵检验。

```
. rddensity margin,plot
```

其中，选择项“plot”表示画图。

默认为无约束(unrestricted)的估计，即将断点两边的观测值视为两个样本，分别估计密度函数的左极限与右极限。

RD Manipulation test using local polynomial density estimation.

c = 0.000	Left of c	Right of c
Number of obs	640	750
Eff. Number of obs	408	460
Order est. (p)	2	2
Order bias (q)	3	3
BW est. (h)	19.841	27.119

Number of obs =	1390
Model =	unrestricted
BW method =	comb
Kernel =	triangular
VCE method =	jackknife

Running variable: margin.

Method	T	P>|T|
Robust	-0.8753	0.3814

P-values of binomial tests. (H0: prob = .5)

Window Length / 2	<c	>=c	P>|T|
0.430	8	12	0.5034
0.861	17	25	0.2800
1.291	25	34	0.2976
1.722	45	47	0.9170
2.152	51	55	0.7709
2.583	66	65	1.0000
3.013	79	71	0.5678
3.444	94	86	0.6020
3.874	105	94	0.4785
4.305	115	107	0.6386

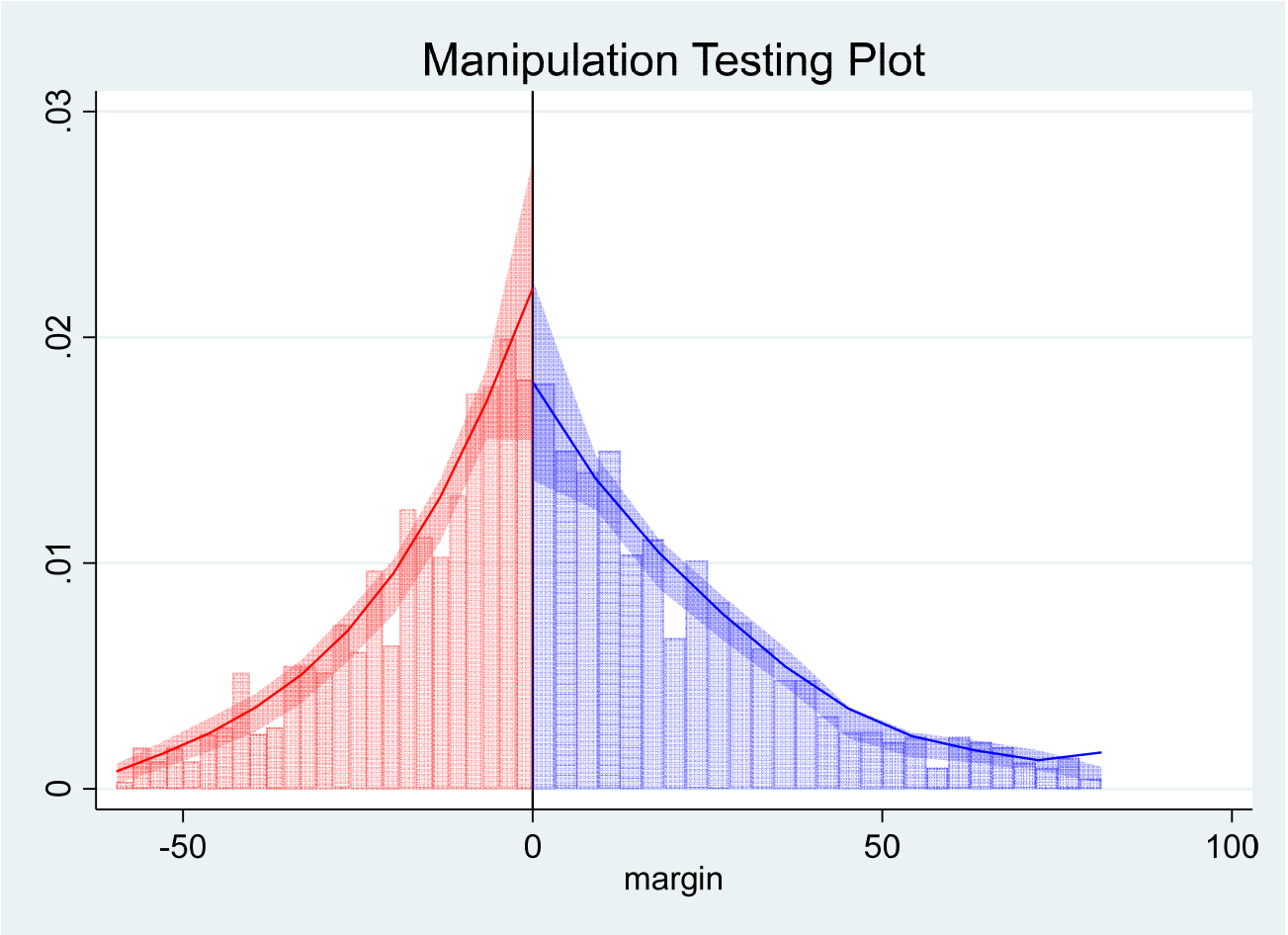

图 16.8 操纵检验的示意图

结果显示，该检验的 p 值高达 0.3814，故接受密度函数在断点处连续(左右极限相等)的原假设，不存在内生分组。

图 16.8 显示，断点左侧密度函数的置信区间与断点右侧的置信区间有明显的重叠，故二者的差异不显著。