第 28 章　处理效应

28.1　处理效应与选择难题

经济学中常希望评估某项目或政策实施后的效应，比如政府推出的就业培训项目(job training program)。

此类研究称为“项目效应评估”(program evaluation)，而项目效应也称为“处理效应”(treatment effect)。

项目参与者的全体构成“实验组”或“处理组”(treatment group，或 the treated)，而未参与项目者则构成“控制组”(control group)或“对照组”(comparison group)。

考虑就业培训的处理效应评估。一个天真的做法是直接对比实验组与控制组的未来收入或就业状况。

但参加就业培训者的未来收入比未参加者通常更低。难道就业培训反而有害？

是否参加培训是参加者自我选择(self selection)的结果，岗位好收入高的人群不需要参加培训，而参加者多为失业或低收入者。

由于实验组与对照组成员初始条件不相同，故存在“选择偏差”(selection bias)。

即使实验组的未来收入低于对照组，我们真正感兴趣的问题是，实验组的未来收入是否会比这些人如果未参加培训项目的(假想)未来收入更高。

Rubin(1974)提出了以下“反事实框架”(a counterfactual framework)，称为“鲁宾因果模型”(Rubin Causal Model)。

以虚拟变量$D_i = \{0,1\}$表示个体 i 是否参与此项目，即 1 为参与，而 0 为未参与。称D_i为“处理变量”(treatment variable)。

记其未来收入或感兴趣的结果(outcome of interest)为y_i。

对于个体 i，未来收入y_i可能有两种状态，取决于是否参加项目：

$$y_i=\begin{cases} y_{1i} & \text{若 } D_i=1 \\ y_{0i} & \text{若 } D_i=0 \end{cases}$$

y_{0i}表示个体 i 未参加项目的未来收入，

y_{1i}表示个体 i 参加项目的未来收入。

想知道$(y_{1i}-y_{0i})$，即个体 i 参加项目的因果效应。

如果个体 i 参加项目，可观测到 y_{1i}，但看不到 y_{0i}；

反之，如果个体 i 未参加项目，可观测到 y_{0i}，但看不到 y_{1i}。

个体只能处于一种状态，故只能观测到 y_{0i} 或 y_{1i}，而无法同时观测到 y_{0i} 与 y_{1i}，是一种“数据缺失”(missing data)问题。

可将 y_i 写为

$$y_i = (1 - D_i) y_{0i} + D_i y_{1i} = y_{0i} + \underbrace{(y_{1i} - y_{0i})}_{\text{处理效应}} D_i$$

其中，$(y_{1i} - y_{0i})$ 为个体 i 参加项目的处理效应。

不同个体的处理效应不同，故将(y_{0i}, y_{1i}, D_i)视为来自三维随机向量(y_0, y_1, D)总体的一个随机抽样。

假设样本为 iid，即对于任何$i \neq j$，(y_{0i}, y_{1i}, D_i)的概率分布与(y_{0j}, y_{1j}, D_j)相同，且相互独立。

这意味着不存在溢出效应，此假定称为“个体处理效应稳定假设”(Stable Unit Treatment Value Assumption，简记 SUTVA)。

由于处理效应$(y_{1i}-y_{0i})$为随机变量，称其期望值为“平均处理效应”(Average Treatment Effect，简记 ATE)：

$$\mathrm{ATE} \equiv \mathrm{E}(y_{1i}-y_{0i})$$

ATE 表示从总体中随机抽取某个体的期望处理效应，无论该个体是否参与项目。

如果仅考虑项目参加者的平均处理效应，称为“参与者平均处理效应”(Average Treatment Effect on the Treated，简记 ATT 或 ATET)或“参与者处理效应”(Treatment Effect on the Treated，简记 TOT)：

$$\mathrm{ATT} \equiv \mathrm{E}(y_{1i}-y_{0i} \mid D_i=1)$$

对于政策制定者，ATT 可能更为重要。ATE 与 ATT 一般不相等。

不能同时观测y_{0i}与y_{1i}，应如何估计 ATE 或 ATT？

简单地比较项目参与者与未参与者的收入，将导致选择偏差：

$$\underbrace{\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)}_{\text{参与者与未参与者的平均差异}} = \underbrace{\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 1)}_{ATT} + \underbrace{\mathrm{E}(y_{0i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)}_{\text{选择偏差}}$$

上式第一项为 ATT，而第二项为参与者的平均y_{0i}与未参与者的平均y_{0i}之差，即选择偏差。

由于低收入者通常更倾向于选择参加培训项目，故选择偏差一般为负，导致实验组与控制组的收入之差(即$\mathrm{E}(y_{1i}\mid D_i=1)-\mathrm{E}(y_{0i}\mid D_i=0)$)低估参与者平均处理效应(ATT)。

如果选择偏差的绝对值足够大，则可能导致$\mathrm{E}(y_{1i}\mid D_i=1)-\mathrm{E}(y_{0i}\mid D_i=0)<0$，出现参加培训者的收入反而低于未参加者的情形。

定义“非参与者平均处理效应”(Average Treatment Effect on the Untreated，简记 ATU)为

$$\mathrm{ATU}\equiv\mathrm{E}(y_{1i}-y_{0i}\mid D_i=0)$$

由于个体根据参加项目的预期收益$\mathrm{E}(y_{1i}-y_{0i})$而自我选择是否参加项目，导致对处理效应的估计困难，称为“选择难题”(the selection problem)。

28.2 通过随机分组解决选择难题

解决选择难题的方法之一是随机分组，使得个体 i 的D_i(是否参加项目)通过抛硬币或电脑随机数而决定，则D_i独立于(y_{0i}, y_{1i})。

此时，$\mathrm{ATE}=\mathrm{ATT}$，因为$\mathrm{E}(y_{1i}-y_{0i}\mid D_i=1)=\mathrm{E}(y_{1i}-y_{0i})$(由于$(y_{1i}-y_{0i})$独立于$D_i$)。

对于 ATE 的估计，只要比较实验组与控制组的平均收入即可：

$$\mathrm{E}(y_{1i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0) = \mathrm{E}(y_{1i}) - \mathrm{E}(y_{0i}) = \mathrm{ATE} = \mathrm{ATT}$$

因为D_i独立于(y_{0i}, y_{1i})。在随机分组的情况下，只需要计算样本中实验组与控制组的平均收入之差，即可一致地估计平均处理效应，即“差额估计量”(differences estimator)。

上述结果在更弱的均值独立(mean independence)条件下也成立，即只要y_{0i}, y_{1i}都均值独立于D_i。

如果只关心 ATT，则只需要y_{0i}均值独立于D_i即可，因为选择偏差$\mathrm{E}(y_{0i} \mid D_i = 1) - \mathrm{E}(y_{0i} \mid D_i = 0)$为 0。

如果只有观测数据，很可能不满足“y_{0i}均值独立于D_i”的假设。

可使用以下两类方法。

第一类方法假设个体依可测变量选择是否参加项目(第 3-7 节)。

第二类方法假设个体依不可测变量选择(第 8 节)。

28.3 依可测变量选择

除(y_i, D_i)外，通常还可观测到个体 i 的一些特征，比如年龄、性别、培训前收入，记为向量$\boldsymbol{x}_i$，也称为“协变量”(covariates)。总体可由$(y_0, y_1, D, \boldsymbol{x})$来表示。

如果个体 i 对D_i的选择完全取决于可观测的$\boldsymbol{x}_i$，称为“依可测变量选择”(selection on observables)，则可以找到估计处理效应的合适方法(即使没有合适的工具变量)。

如果个体对D_i的选择完全取决于$\boldsymbol{x}_i$，则在给定$\boldsymbol{x}_i$的情况下，潜在结果(y_{0i}, y_{1i})将独立于D_i。

Rosenbaum and Rubin (1983)提出“可忽略性”假设：

假定 28.1 可忽略性 (Ignorability)。

给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})独立于D_i，记为$(y_{0i}, y_{1i}) \perp D_i \mid \boldsymbol{x}_i$，其中“$\perp$”表示相互独立。

“可忽略性”的含义是，给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})对于D_i的影响可以忽略。

可忽略性也称为“无混淆性”(unconfoundedness)，“条件独立假定”(Conditional Independence Assumption，简记 CIA)，或“依可测变量选择”(selection on observables)。

此假定意味着，给定$\boldsymbol{x}_i$，则(y_{0i}, y_{1i})在处理组与控制组的分布完全一样，即

$$F(y_{0i}, y_{1i} \mid \boldsymbol{x}_i, D_i = 1) = F(y_{0i}, y_{1i} \mid \boldsymbol{x}_i, D_i = 0)$$

$F(\cdot)$为分布函数。在很多情况下，只需更弱的均值独立假定。

假定 28.2 均值可忽略性 (Ignorability in Mean)。

$\mathrm{E}(y_{0i} \mid \boldsymbol{x}_i, D_i) = \mathrm{E}(y_{0i} \mid \boldsymbol{x}_i)$，而且$\mathrm{E}(y_{1i} \mid \boldsymbol{x}_i, D_i) = \mathrm{E}(y_{1i} \mid \boldsymbol{x}_i)$。这意味着，在给定$\boldsymbol{x}_i$的情况下，$y_{0i}$与$y_{1i}$都均值独立于$D_i$。

如果可忽略性假定成立，则原则上可将$\boldsymbol{x}_i$直接作为控制变量引入以下回归方程，以解决遗漏变量问题：

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma D_i + \varepsilon_i$$

但不清楚$\boldsymbol{x}_i$是否应以线性形式进入上述方程。如果遗漏非线性项，仍可能存在遗漏变量偏差。

解决方法之一为基于鲁宾反事实框架的匹配估计量。

从此方程可看出，可忽略性是很强的假定；它意味着回归方程已包括了所有相关变量，不存在任何与解释变量相关的遗漏变量。

如果$\boldsymbol{x}_i$中已包含较丰富的协变量(a rich set of covariates)，可认为可忽略性假定基本得到满足，遗漏变量偏差较小。

28.4 匹配估计量的思想

假设个体 i 属于处理组，匹配估计量的基本思路是，找到属于控制组的某个体j,使得个体j与个体i的可测变量取值尽可能相似(匹配)，即$\boldsymbol{x}_i \approx \boldsymbol{x}_j$。

基于可忽略性假设，则个体 i 与个体 j 进入处理组的概率相近，具有可比性；故可将y_j作为y_{0i}的估计量，即$\hat{y}_{0i} = y_j$。

可将$(y_i - \hat{y}_{0i}) = y_i - y_j$作为对个体 i 处理效应的度量。

对处理组中的每位个体都如此进行匹配；类似地，对控制组每位个体也进行匹配，然后对每位个体的处理效应进行平均，即可得到“匹配估计量”(matching estimators)。

由于匹配的具体方法不同，故存在不同的匹配估计量。

首先，是否放回；如果不放回(no replacement)，则每次都将匹配成功的个体(i,j)从样本中去掉；如果有放回，则将匹配成功个体留在样本中，参与其余匹配。

其次，是否允许并列(ties)，比如控制组个体j与k的可测变量都与处理组个体i一样接近。

如果允许并列，则将y_j与y_k的平均值作为y_{0i}的估计量，即$\hat{y}_{0i}=(y_j+y_k)/2$。

如果不允许并列，则电脑程序将根据数据排序选择个体j或k；匹配结果可能与数据排序有关，故建议先将样本随机排序。

以上为一对一(one-to-one)匹配，也可以进行一对多匹配，比如一对四匹配，即针对每位个体寻找四位不同组的最近个体匹配。

匹配估计量一般存在偏差(bias)，除非在“精确匹配”(exact matching)的情况下，即对于所有匹配都有$\boldsymbol{x}_i = \boldsymbol{x}_j$。

更常见的为“非精确匹配”(inexact matching)，只能保证$\boldsymbol{x}_i \approx \boldsymbol{x}_j$。

在非精确匹配的情况下，如进行一对一匹配，则偏差较小，但方差较大；

进行一对多匹配可降低方差(使用了更多信息)，但偏差增大(使用了更远的信息)。

Abadie et al (2004)建议进行一对四匹配，以最小化均方误差。

例 假设样本容量为 7，其中包括 3 位控制组个体与 4 位处理组个体。同时假设$\boldsymbol{x}_i$仅包含一个变量x_i。进行有放回的一对一匹配，且允许并列。

表 28.1 匹配估计量的简单例子

i	D_i	x_i	y_i	匹配结果	$\hat{y}_{0i}$	$\hat{y}_{1i}$
1	0	2	7	{5}	7	8
2	0	4	8	{4, 6}	8	7.5
3	0	5	6	{4, 6}	6	7.5
4	1	3	9	{1, 2}	7.5	9
5	1	2	8	{1}	7	8
6	1	3	6	{1, 2}	7.5	6
7	1	1	5	{1}	7	5

Curse of dimension.

k=1, 体积. 1, 0.01 体积. $d = 0.01^{\frac{1}{k}}$

28.5 倾向得分匹配

$\boldsymbol{x}_i$可能包括多个变量。如直接用$\boldsymbol{x}_i$进行匹配，可能遇到数据稀疏的问题，很难找到与$\boldsymbol{x}_i$相近的$\boldsymbol{x}_j$。

可使用某函数$f(\boldsymbol{x}_i)$，将 K 维向量$\boldsymbol{x}_i$的信息压缩到一维，根据$f(\boldsymbol{x}_i)$进行匹配。

定义$\boldsymbol{x}_i$与$\boldsymbol{x}_j$之间的“马氏距离”(Mahalanobis distance)为

$$d(i,j)=(\boldsymbol{x}_i-\boldsymbol{x}_j)'\hat{\boldsymbol{\Sigma}}_X^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)$$

其中，二次型矩阵$\hat{\boldsymbol{\Sigma}}_X^{-1}$为$\boldsymbol{x}$的样本协方差矩阵之逆矩阵。

使用马氏距离进行匹配，称为“马氏匹配”(Mahalanobis matching)。

马氏匹配的缺点是，如果$\boldsymbol{x}$包括的变量较多或样本容量不够大，则不易找到好的匹配。

Rosenbaum and Rubin (1983)提出使用“倾向得分” (propensity score，简记 p-score)来度量距离。

定义 个体 i 的倾向得分为，在给定$\boldsymbol{x}_i$的情况下，个体 i 进入处理组的条件概率，即 $p(\boldsymbol{x}_i) \equiv \mathrm{P}(D_i = 1 \mid \boldsymbol{x} = \boldsymbol{x}_i)$，或简记 $p(\boldsymbol{x})$。

在估计 $p(\boldsymbol{x})$时，可使用参数估计(probit 或 logit)或非参数估计，最流行的方法为 logit。

使用倾向得分度量个体间距离，它不仅是一维变量，而且取值介于$[0,1]$之间。

即使$\boldsymbol{x}_i$与$\boldsymbol{x}_j$距离很远，仍可能$p(\boldsymbol{x}_i) \approx p(\boldsymbol{x}_j)$。

使用倾向得分作为距离函数进行匹配，称为“倾向得分匹配”(Propensity Score Matching，简记 PSM)。

PSM 的理论依据在于，如果可忽略性假定成立，则只须在给定$p(\boldsymbol{x})$的情况下，(y_{0i}, y_{1i})就独立于D_i。

命题(倾向得分定理) $(y_0, y_1) \perp D \mid \boldsymbol{x} \quad \Rightarrow \quad (y_0, y_1) \perp D \mid p(\boldsymbol{x})$

证明： 由于 D 为虚拟变量，故只须证明$\mathrm{P}\left[D=1 \mid y_0, y_1, p(\boldsymbol{x})\right]$与$y_0, y_1$无关即可。

$$
\begin{aligned}
&\mathrm{P}\left[D=1 \mid y_0, y_1, p(\boldsymbol{x})\right] \\
&=\mathrm{E}\left[D \mid y_0, y_1, p(\boldsymbol{x})\right] \\
&=\mathrm{E}_{y_0, y_1, \boldsymbol{x}}\left[\mathrm{E}(D \mid y_0, y_1, \boldsymbol{x}) \mid y_0, y_1, p(\boldsymbol{x})\right] \quad \text{(迭代期望定律)} \\
&=\mathrm{E}_{y_0, y_1, \boldsymbol{x}}\left[\mathrm{E}(D \mid \boldsymbol{x}) \mid y_0, y_1, p(\boldsymbol{x})\right] \quad \text{(可忽略性假定)} \\
&=\mathrm{E}_{y_0, y_1, \boldsymbol{x}}\left[p(\boldsymbol{x}) \mid y_0, y_1, p(\boldsymbol{x})\right] \\
&=p(\boldsymbol{x})
\end{aligned}
$$

为了能够进行匹配，需要在$\boldsymbol{x}$的每个可能取值上都同时存在处理组与控制组的个体，即“重叠假定”(overlap assumption)或“匹配假定”(matching assumption)。

假定 28.3 重叠假定。

对于$\boldsymbol{x}$的任何可能取值，都有$0 < p(\boldsymbol{x}) < 1$。

此假定意味着处理组与控制组这两个子样本存在重叠，故名“重叠假定”；它也是进行匹配的前提，故也称“匹配假定”。

它保证了处理组与控制组的倾向得分取值范围有相同的部分(common support)，参见图 28.1。

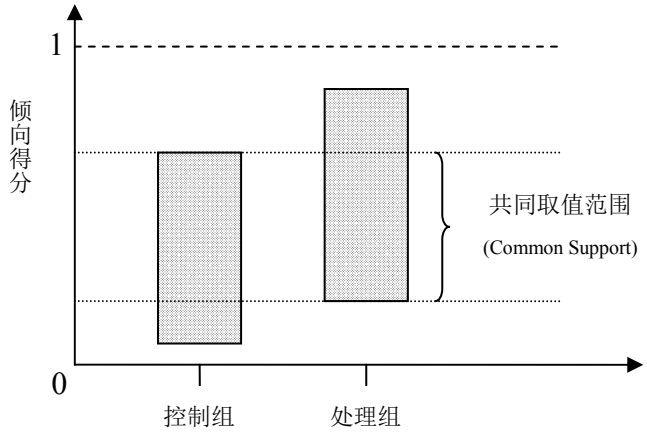


图 28.1 倾向得分的共同取值范围

在进行匹配时，为提高匹配质量，可仅保留倾向得分重叠部分的个体(但会损失样本容量)。

如果倾向得分的共同取值范围太小，则会导致偏差。

通过倾向得分匹配计算平均处理效应的一般步骤如下。

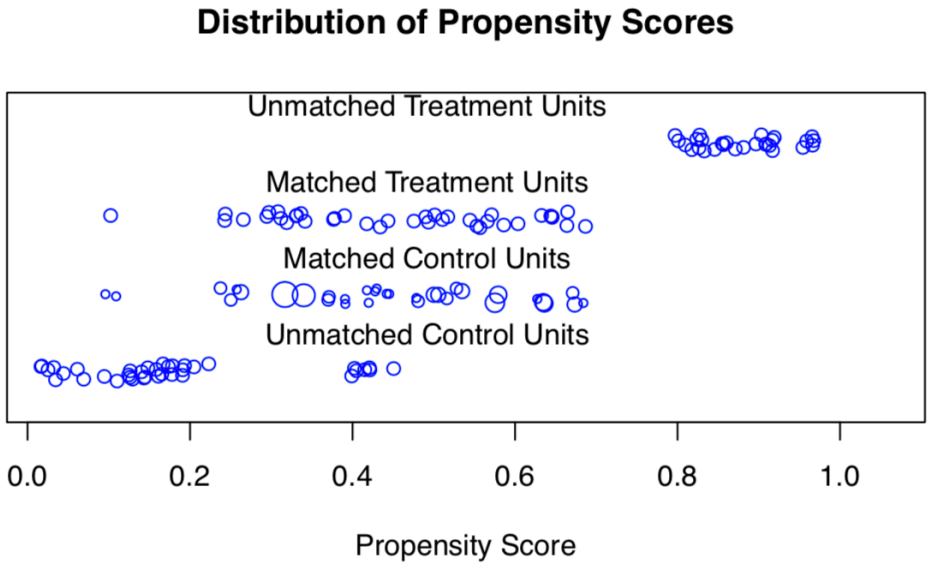
(1) 选择协变量$\boldsymbol{x}_i$。尽量将可能影响(y_{0i}, y_{1i})与D_i的相关变量包括进来，以满足可忽略性假设。

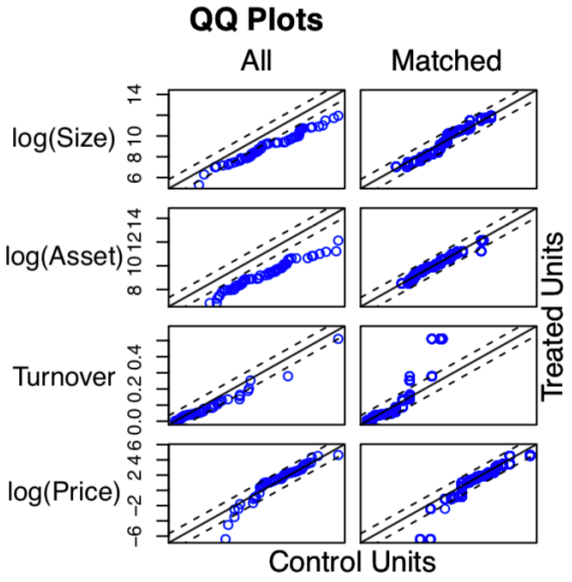
(2) 估计倾向得分。Rosenbaum and Rubin (1985)建议使用形式灵活的 logit 模型，比如包括$\boldsymbol{x}_i$的高次项与互动项。

(3) 进行倾向得分匹配。

如果倾向得分估计得较准确，应使$\boldsymbol{x}_i$在匹配后的处理组与控制组之间分布较均匀，比如，匹配后的处理组均值$\bar{\boldsymbol{x}}_{treat}$与控制组均值$\bar{\boldsymbol{x}}_{control}$较接近；称为“数据平衡”(data balancing)。

$\bar{\boldsymbol{x}}_{treat}$与$\bar{\boldsymbol{x}}_{control}$的差距与计量单位有关，故一般针对$\boldsymbol{x}$的每个分量 x


Distribution of Propensity Scores
Unmatched Treatment Units
Matched Treatment Units
Matched Control Units
Unmatched Control Units
0.0
0.2
0.4
0.6
0.8
1.0
Propensity Score


QQ Plots
All
Matched
log(Size)
log(Asset)
Turnover
log(Price)
Treated Units
Control Units

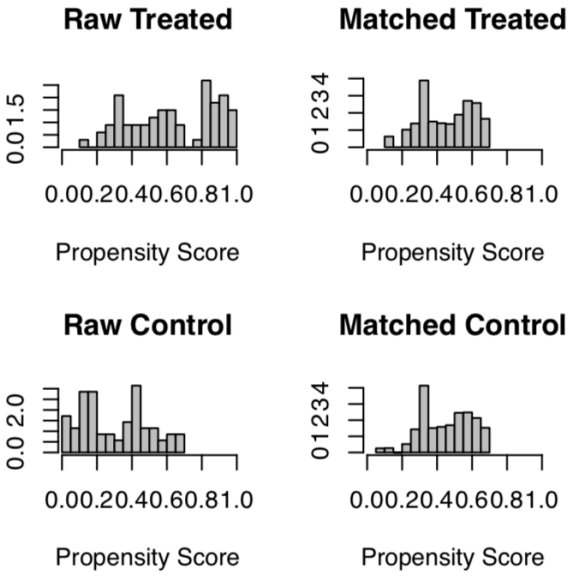

Raw Treated
Matched Treated
Raw Control
Matched Control
Propensity Score

Table 11: Diagnostics of Propensity Score Matching: Less vs. More

The first four columns of panel A and B report the weighted mean and standard deviation of the treatment group (Less) and control group (More). In panel A, the weight is 1 for all observations because there is no matching yet; in panel B, the weight for units in the treatment group is 1 and the weight for a unit in the control group is proportional to the number of treatment units to which it was matched. The last two columns in panel A present the mean difference and the p-value of the heterogeneity robust t-test that the difference is zero. The last two columns in panel B report the weighted version of the same statistics. Panel C reports the number of firms matched, unmatched and discarded.

	Mean Treated	Std. Treated	Mean Control	Std. Control	Mean Diff.	p-value
Panel A: Summary of balance of all data						
Prop. Score	0.632	0.242	0.304	0.192	0.328	0.000
log(Size)	8.888	1.379	10.255	1.755	−1.367	0.000
log(Asset)	9.191	1.135	11.008	1.493	−1.817	0.000
Turnover	0.068	0.080	0.095	0.105	−0.027	0.098
log(Price)	1.368	2.048	1.340	1.716	0.028	0.932
Panel B: Summary of balance of matched data						
Prop. Score	0.454	0.147	0.454	0.148	0.000	0.992
log(Size)	9.280	1.369	9.357	1.175	−0.078	0.817
log(Asset)	9.910	0.838	9.925	0.847	−0.016	0.944
Turnover	0.077	0.094	0.078	0.060	−0.001	0.957
log(Price)	1.448	1.959	1.802	1.566	−0.354	0.428

Panel C: Number of firms		
	Control	Treated
All	70	67
Matched	33	39
Unmatched	37	0
Discarded	0	28

考察“标准化差距”(standardized differences)或“标准化偏差”(standardized bias)：

$$\frac{\left|\bar{x}_{treat}-\bar{x}_{contrl}\right|}{\sqrt{(s^2_{x,treat}+s^2_{x,control})/2}}$$

其中，$s^2_{x,treat}$与$s^2_{x,control}$分别为处理组与控制组变量 x 的样本方差。一般要求此标准化差距不超过 10%；如果超过，则应回到第(2)步、甚至第(1)步，重新估计倾向得分；或者改变具体的匹配方法。

(4) 根据匹配后样本(matched sample)计算平均处理效应。参加者平均处理效应(ATT)估计量的一般表达式为

$$\widehat{\mathrm{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} (y_i - \hat{y}_{0i})$$

其中，$N_1 = \sum_i D_i$ 为处理组个体数。

未参加者平均处理效应(ATU)估计量的一般表达式为：

$$\widehat{\mathrm{ATU}} = \frac{1}{N_0} \sum_{j:D_j=0} (\hat{y}_{1j} - y_j)$$

其中，$N_0 = \sum_j (1 - D_j)$ 为控制组个体数。

平均处理效应(ATE)估计量的一般表达式为

$$\widehat{\mathrm{ATE}}=\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{1i}-\hat{y}_{0i})$$

$N=N_0+N_1$；如果$D_i=1$，则$\hat{y}_{1i}=y_i$；如果$D_i=0$，则$\hat{y}_{0i}=y_i$。

在进行倾向得分匹配时，有不同的具体方法。

方法之一为“k近邻匹配”(k-nearest neighbor matching)，即寻找倾向得分最近的k个不同组个体。如果$k=1$，则为“一对一匹配”(one-to-one matching)。

但即使“最近邻居”也可能相去甚远。

方法之二限制倾向得分的绝对距离$|p_i - p_j| \le \varepsilon$，一般建议$\varepsilon \le 0.25\hat{\sigma}_{pscore}$，其中$\hat{\sigma}_{pscore}$为倾向得分的样本标准差；称为“卡尺匹配”(caliper matching)或“半径匹配”(radius matching)。

方法之三为“卡尺内最近邻匹配”(nearest-neighbor matching within caliper)，在给定的卡尺ε范围寻找最近匹配，此法较流行。

以上三种方法本质上都是近邻匹配法。

另一类匹配方式为整体匹配法，每位个体的匹配结果为不同组的全部个体(通常去掉在 common support 之外的个体)，只是根据个体距离不同给予不同的权重(近者权重大，远者权重小，超出一定范围权重可为 0)。

比如，在估计 ATT 时，$\hat{y}_{0i}$的估计量为

$$\hat{y}_{0i} = \sum_{j:D_j=0} w(i,j) y_j$$

其中，$w(i,j)$为适用于配对(i,j)的权重。

如果使用核函数来计算权重$w(i,j)$，则为**方法之四**“核匹配”(kernel matching)(Heckman et al, 1997, 1998)，其权重表达式为

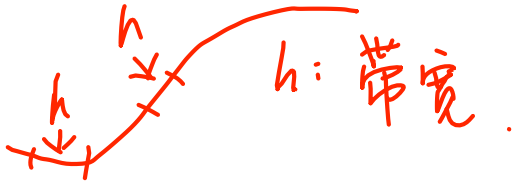
Heckit.

$$w(i,j) = \frac{K[(\boldsymbol{x}_j - \boldsymbol{x}_i)/h]}{\sum_{k:D_k=0} K[(\boldsymbol{x}_k - \boldsymbol{x}_i)/h]}$$

其中，h 为带宽，$K(\cdot)$为核函数。$\hat{y}_{0i}$可视为“核回归估计量”。

"Hyperparameter."

KNN，LLR，非参，半参。



如用局部线性回归来估计$w(i,j)$，则为**方法之五**“局部线性回归匹配”(local linear regression matching)。

方法之六使用更为光滑的“三次样条”(cubic spline)来估计$w(i,j)$，称为“样条匹配”(spline matching)。

究竟应使用以上哪种具体匹配方法或参数(比如，k 近邻匹配的 k 取值，是否放回，如何处理并列)，文献中尚无明确指南。

不存在适用于一切情形的绝对好方法，只能根据具体数据来选择匹配方法。

如果控制组个体并不多(N_0较小)，则应进行有放回的匹配。

如果存在较多具有可比性的控制组个体，则可考虑一对多匹配或核匹配，以提高匹配效率。

实践中建议尝试不同的匹配方法，考察其稳健性。

PSM 的局限性：

(1) PSM 通常要求比较大的样本容量以得到高质量的匹配。

(2) PSM 要求处理组与控制组的倾向得分有较大的共同取值范围(common support)；否则，将丢失较多观测值，导致剩下的样本不具有代表性。

(3) PSM 只控制可测变量的影响，如存在依不可测变量选择(selection on unobservable)，仍有“隐性偏差”(hidden bias)。

28.6 倾向得分匹配的 Stata 实例

28.7 偏差校正匹配估计量

在倾向得分匹配第一阶段估计倾向得分时，存在不确定性(可使用 probit，logit 或非参估计)。

Abadie and Imbens (2002, 2004, 2006, 2011)又重新回到更简单的马氏距离，进行有放回且允许并列(ties)的 k 近邻匹配。

A&I (2016)

errors-in-variable.

由于非精确匹配(inexact matching)一般存在偏差，Abadie and Imbens 提出了偏差校正的方法，通过回归的方法来估计偏差，然后得到“偏差校正匹配估计量”(bias-corrected matching estimator)。

Abadie and Imbens 还通过在处理组或控制组内部进行二次匹配，得到在异方差条件下也成立的稳健标准误。

PSM：X 可测

28.8 双重差分倾向得分匹配

对于观测数据，如存在依不可测变量选择，有几种处理方法：

(1) 尽量使用更多的相关可测变量，以满足可忽略性假定，然后使用匹配估计量。

(2) 如果影响处理变量D_i的不可观测变量不随时间而变，而且有面板数据，则可使用“双重差分倾向得分匹配估计量”(differences-in-differences PSM estimator)。

(3) 使用断点回归法，特别是模糊断点回归，参见第 9-12 节。

(4) 使用工具变量法，比如第 13 节的处理效应模型。

(5) 根据依可测变量选择的影响来估计依不可测变量选择的影响，参见 Altonji et al (2005)。

本节介绍即双重差分 PSM，由 Heckman et al (1997, 1998)提出。

假设有两期面板数据，记实验前的时期为t'，实验后的时期为t。

在时期t'，实验还未发生，故所有个体的潜在结果均可记为$y_{0t'}$。

在时期t，实验已经发生，故有两种潜在结果，记为y_{1t}(参与实验)与y_{0t}(未参与实验)。

双重差分 PSM 成立的前提为均值可忽略性假定：

$$\mathrm{E}(y_{0t} - y_{0t'} \mid \boldsymbol{x}, D = 1) = \mathrm{E}(y_{0t} - y_{0t'} \mid \boldsymbol{x}, D = 0)$$

在此假定下，可一致地估计 ATT：

$$\widehat{ATT}=\frac{1}{N_1}\sum_{i:\, i\in I_1\cap S_p}\left[(y_{1ti}-y_{0t'i})-\sum_{j:\, j\in I_0\cap S_p}w(i,j)(y_{0tj}-y_{0t'j})\right]$$

其中，S_p为共同取值范围的集合(common support)，$I_1=\{i: D_i=1\}$(处理组的集合)，$I_0=\{i: D_i=0\}$(控制组的集合)，N_1为集合$I_1\cap S_p$所包含的处理组个体数，而$w(i,j)$为对应于配对(i,j)的权重，可通过核匹配或局部线性回归匹配的方法来确定。

方括弧内第一项$(y_{1ti}-y_{0t'i})$为处理组个体 i 实验前后的变化，而第二项中的$(y_{0tj}-y_{0t'j})$则为控制组个体j的前后变化。

双重差分 PSM 法的步骤如下。

第一步、根据处理变量D_i与协变量$\boldsymbol{x}_i$估计倾向得分。

第二步、对于处理组的每位个体 i，确定与其匹配的全部控制组个体(即确定集合S_p)。

第三步、对于处理组的每位个体 i，计算其结果变量的前后变化$(y_{1ti}-y_{0t'i})$。

第四步、对于处理组的每位个体 i，计算与其匹配的全部控制组个体的前后变化$(y_{0tj}-y_{0t'j})$，其中$j\in I_0\cap S_p$。

第五步、针对$(y_{1ti}-y_{0t'i})$与$(y_{0tj}-y_{0t'j})$，根据公式进行倾向得分核匹配或局部线性回归匹配，即得到$\widehat{ATT}$。

Gary King.

双重差分 PSM 法的优点：可以控制不可观测(unobservable)但不随时间变化(time invariant)的组间差异，比如处理组与控制组分别来自两个不同的区域，或处理组与控制组使用了不同的调查问卷。

28.9 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是否超过某断点所决定。

据以进行分组的变量x_i被称为“分组变量”(assignment variable)。

例 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：

$$D_i = \begin{cases} 1 & \text{若} x_i \geq 500 \\ 0 & \text{若} x_i < 500 \end{cases}$$

记不上大学与上大学的两种潜在结果分别为(y_{0i}, y_{1i})。

由于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，故D_i独立于(y_{0i}, y_{1i})，满足可忽略性假定。

但不能使用 PSM，因为重叠假定不满足，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$x_i < 500$，完全没有交集！

处理变量D_i为x_i的函数，记为$D(x_i)$。由于函数$D(x_i)$在 $x = 500$ 处存在一个断点(discontinuity)，故可估计D_i对y_i的因果效应。

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。

他们高考成绩的细微差异只是由于“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，可视为准实验。

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

$$\begin{aligned}\text{LATE} &\equiv \mathrm{E}(y_{1i} - y_{0i} \mid x = 500) \\ &= \mathrm{E}(y_{1i} \mid x = 500) - \mathrm{E}(y_{0i} \mid x = 500) \\ &= \lim_{x \downarrow 500} \mathrm{E}(y_{1i} \mid x) - \lim_{x \uparrow 500} \mathrm{E}(y_{0i} \mid x)\end{aligned}$$

从上 从下

其中，$\lim_{x \downarrow 500}$ 与 $\lim_{x \uparrow 500}$ 分别表示从 500 的右侧与左侧取极限。假设 $\mathrm{E}(y_{1i} \mid x)$ 与 $\mathrm{E}(y_{0i} \mid x)$ 为连续函数，故其极限值等于函数取值。

一般地，假设断点为某常数 c，而分组规则为

$$D_i = \begin{cases} 1 & \text{若} x_i \geq c \\ 0 & \text{若} x_i < c \end{cases}$$

假设在实验前，结果变量 y_i 与 x_i 之间存在如下线性关系：

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

假设$D_i = \mathbf{1}(x_i \geq c)$的处理效应为正，则$y_i$与$x_i$之间的线性关系在$x = c$处就存在一个向上跳跃的断点。

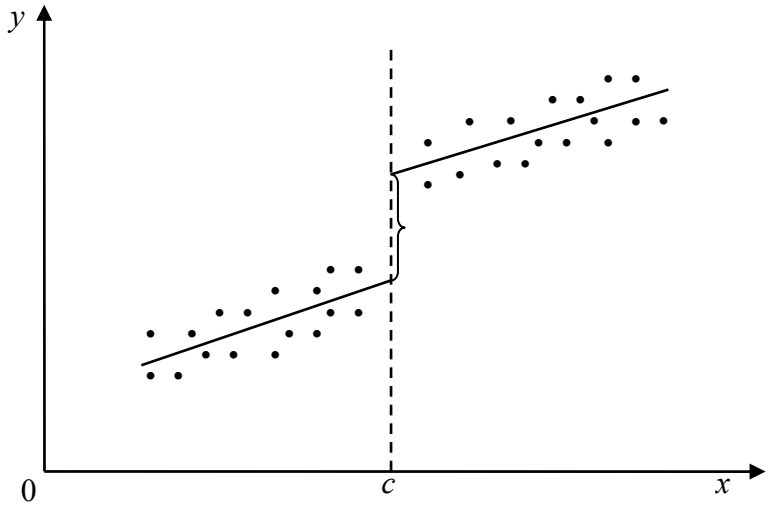
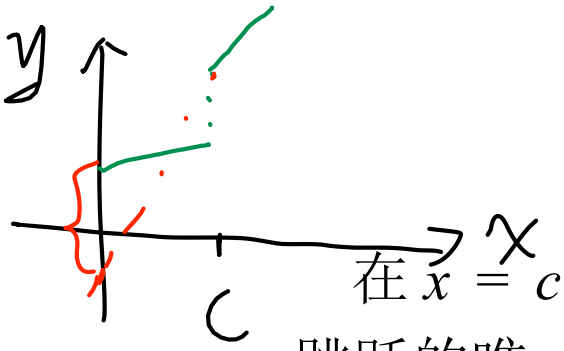


图 28.4 断点回归示意图



在 $x = c$ 附近，个体在各方面均无系统差别，造成 $\mathrm{E}(y_i \mid x)$ 在此跳跃的唯一原因只能是 D_i 的处理效应，故可将此跳跃视为在 $x = c$ 处 D_i 对 y_i 的因果效应。

$D_i = 1$，$\beta + \gamma$

为了估计此跳跃，将方程改写为：$D_i = 0$，β

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

① 线性　　② 全样本

变量 $(x_i - c)$ 为 x_i 的标准化，使得 $(x_i - c)$ 的断点为 0。

引入互动项 $\gamma(x_i - c)D_i$ 允许断点两侧的回归线斜率可以不同。

对此方程进行 OLS 回归，所得 $\hat{\delta}$ 就是在 $x = c$ 处的 LATE 估计量。

由于此回归线存在断点，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)比较强的一种准实验。

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差异来检验随机性。

但断点回归仅推断在断点处的因果关系，并不一定能推广到其他样本值，故外部有效性(external validity)受局限。

例 Thistlewaite and Campbell (1960)使用断点回归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，故成绩刚好达到获奖标准与差一点达到的学生具有可比性。

例 Angrist and Lavy (1999)在研究班级规模对成绩的影响时，利用了色列教育系统的一项制度进行断点回归；该制度限定班级规模的上限为 40 名学生，一旦超过 40 名学生(比如 41 名学生)，则该班级将被一分为二。

28.10 精确断点回归

断点回归可分为两种类型。一种类型是“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 $0 < a < b < 1$。

使用上述方程估计精确断点回归，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。

其次，既然断点回归是局部的随机实验，则原则上只应使用断点附近的观测值，却使用了整个样本。

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

$$y_i = \alpha + \beta_1(x_i - c) + \delta D_i + \gamma_1(x_i - c)D_i + \beta_2(x_i - c)^2 + \gamma_2(x_i - c)^2 D_i + \varepsilon_i \quad (c-h < x < c+h)$$

其中，$\hat{\delta}$为对 LATE 的估计量，可用稳健标准误来控制异方差。

但上式未确定 h 的取值，且仍依赖于具体的函数形式。

研究者开始转向非参数回归，不依赖于具体的函数形式，且可以通过最小化均方误差来选择最优带宽 h。

一般推荐使用局部线性回归，即最小化如下目标函数：

$$\min_{\{\alpha,\beta,\delta,\gamma\}} \sum_{i=1}^{n} K\left[(x_i - c)/h\right]\left[y_i - \alpha - \beta(x_i - c) - \delta D_i - \gamma(x_i - c)D_i\right]^2$$

其中，$K(\cdot)$ 为核函数。针对断点回归，较常用的核函数为三角核(triangular kernel)与矩形核(rectangular kernel，即均匀核)。

如使用矩形核，则为 OLS 回归，等价于上文的参数回归。此估计量也称为“局部沃尔德估计量”(local Wald estimator)。

考察最优带宽的选择。记$m_1(x) \equiv \mathrm{E}(y_1 \mid x)$，$m_0(x) \equiv \mathrm{E}(y_0 \mid x)$，则$\delta = m_1(c) - m_0(c)$，$\hat{\delta} = \hat{m}_1(c) - \hat{m}_0(c)$。

Imbens and Kalyanaraman (2009)提出通过最小化两个回归函数在断点处的均方误差来选择最优带宽：

cross-validation.

$$\min_h \ \mathrm{E}\left\{\left[\hat{m}_1(c) - m_1(c)\right]^2 + \left[\hat{m}_0(c) - m_0(c)\right]^2\right\}$$

也可在方程中加入影响结果变量y_i的其他协变量$\boldsymbol{w}_i$。

是否包括协变量$\boldsymbol{w}_i$不影响断点回归的一致性，但加入协变量可减少扰动项方差，使得估计更为准确。

如果协变量$\boldsymbol{w}_i$在$x=c$处的条件密度函数也存在跳跃，则不宜将$\hat{\delta}$全部归功于该项目的处理效应。

断点回归的隐含假设是，协变量$\boldsymbol{w}_i$的条件密度在$x=c$处连续。

为了检验此假设，可将$\boldsymbol{w}_i$中每个变量作为被解释变量，进行断点回归，考察其分布是否在$x=c$处有跳跃。

“内生分组”(endogenous sorting)：

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组变量(complete manipulation)，自行选择进入处理组或控制组，导致非随机分组，引起断点回归失效。

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。

对于内在分组，可从理论上讨论，也可根据数据进行检验。

x本身在c附近是否连续

假设存在内生分组，个体自行选择进入断点两侧，导致分组变量 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。

kernel density.

McCrary(2008)提出检验以下原假设：

$$H_0:\ \theta \equiv \ln \lim_{x \downarrow c} f(x) - \ln \lim_{x \uparrow c} f(x) \equiv \ln f^{+} - \ln f^{-} = 0$$

通过计算$\hat{\theta}$及其标准误，可检验密度函数$f(x)$是否$x=c$处连续。

内生分组也可能使得协变量$\boldsymbol{w}_i$在$x=c$两侧分布不均匀；故须检验协变量$\boldsymbol{w}_i$的条件密度在$x=c$处是否连续。

由于断点回归在操作上存在不同选择，实践中一般建议同时汇报以下各种情形，以保证稳健性：

(1) 分别汇报三角核与矩形核的局部线性回归结果(后者等价于线性参数回归)；

(2) 分别汇报使用不同带宽的结果(比如，最优带宽及其二分之一或两倍带宽)；

(3) 分别汇报包含协变量与不包含协变量的情形。

(4) 进行模型设定检验，包括检验分组变量与协变量的条件密度是否在断点处连续。

28.11 模糊断点回归

模糊断点回归的特征是，在断点 $x=c$ 处，个体得到处理的概率从 a 跳跃为 b，其中 $0<a<b<1$。即使 $x>c$，也不一定得到处理，但得到处理的概率在 $x=c$ 处有不连续的跳跃。

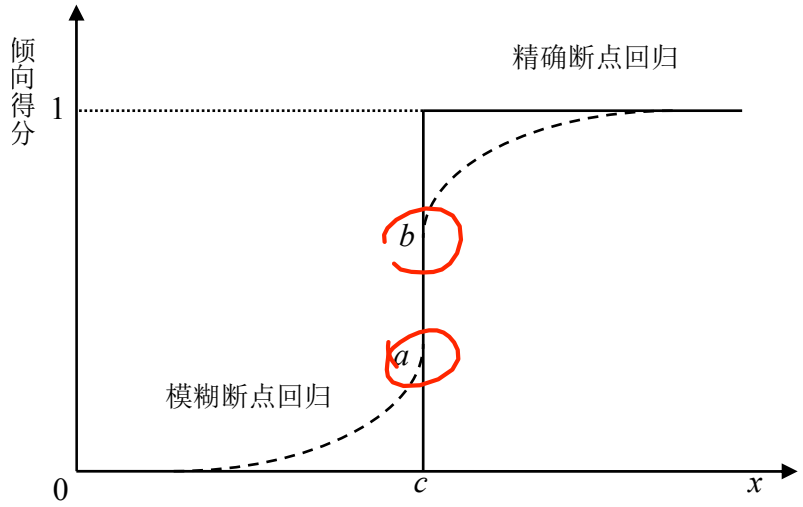


图 28.5 精确断点回归与模糊断点回归

例 高考成绩上线并不能完全保证上大学，能否上大学还取决于填报志愿，甚至有些上线考生放弃上大学的机会；而即使成绩未上线，但也可能因某种特长而得到加分，从而得到上大学的机会。上大学的概率确实在分数线的位置上有一个不连续的跳跃。

在模糊断点的情况下，处理变量 D 不完全由分组变量 x 所决定。

一般来说，影响处理变量 x 的其他因素也会影响结果变量 y，导致在回归方程中处理变量 D 与扰动项 ε 相关，故 OLS 不一致。

例 虽然成绩上线却因志愿不妥而落榜者多有较深实力，而这种不可观测的实力可以影响结果变量 y。

在模糊断点的情况下识别平均处理效应，需引入条件独立假定。

假定 给定 x，则 $(y_1 - y_0)$ 独立于 D，即 $(y_{1i} - y_{0i}) \perp D_i \mid x_i$。

此假定意味着，在给定分组变量 x 的情况下，D 可以与 y_0 相关，但不能与参加项目的收益 $(y_{1i}-y_{0i})$ 相关。

由于 $y=y_0+D(y_1-y_0)$，故

$$\begin{aligned}\mathrm{E}(y\mid x)&=\mathrm{E}(y_0\mid x)+\mathrm{E}\left[D(y_1-y_0)\mid x\right]\\&=\mathrm{E}(y_0\mid x)+\mathrm{E}(D\mid x)\cdot\mathrm{E}\left[(y_1-y_0)\mid x\right]\end{aligned}$$

其中，$\mathrm{E}\left[(y_1-y_0)\mid x\right]$ 是平均处理效应，而 $\mathrm{E}(D\mid x)$ 为倾向得分。对上式两边从 c 的右边取极限可得

$$\lim_{x\downarrow c}\mathrm{E}(y\mid x)=\lim_{x\downarrow c}\mathrm{E}(y_0\mid x)+\lim_{x\downarrow c}\mathrm{E}(D\mid x)\cdot\lim_{x\downarrow c}\mathrm{E}\left[(y_1-y_0)\mid x\right]$$

对上式两边从 c 的左边取极限可得

$$\lim_{x\uparrow c}\mathrm{E}(y\mid x)=\lim_{x\uparrow c}\mathrm{E}(y_0\mid x)+\lim_{x\uparrow c}\mathrm{E}(D\mid x)\cdot\lim_{x\uparrow c}\mathrm{E}\left[(y_1-y_0)\mid x\right]$$

假设$\mathrm{E}(y_0\mid x)$与$\mathrm{E}(y_1\mid x)$在 $x=c$ 处连续，则其左极限等于右极限，也等于其函数值，故$\lim_{x\downarrow c}\mathrm{E}(y_0\mid x)=\lim_{x\uparrow c}\mathrm{E}(y_0\mid x)$，而且

$$\lim_{x\downarrow c}\mathrm{E}\left[(y_1-y_0)\mid x\right]=\lim_{x\uparrow c}\mathrm{E}\left[(y_1-y_0)\mid x\right]=\mathrm{E}\left[(y_1-y_0)\mid x=c\right]。$$

两方程相减可得：

$$\lim_{x\downarrow c}\mathrm{E}(y\mid x)-\lim_{x\uparrow c}\mathrm{E}(y\mid x)=\left[\lim_{x\downarrow c}\mathrm{E}(D\mid x)-\lim_{x\uparrow c}\mathrm{E}(D\mid x)\right]\cdot\mathrm{E}\left[(y_1-y_0)\mid x=c\right]$$

$E(D|x) = P(D=1|x)\,1 + P(D=0|x)\,0 = P(D=1|x)$

b a Treatment

根据模糊断点回归的定义可知，

$\lim\limits_{x\downarrow c}\mathrm{E}(D\,|\,x)-\lim\limits_{x\uparrow c}\mathrm{E}(D\,|\,x)=b-a\neq 0$，故可将其作为分母：

$$\mathrm{LATE}\equiv\mathrm{E}\left[(y_1-y_0)\,|\,x=c\right]=\frac{\lim\limits_{x\downarrow c}\mathrm{E}(y\,|\,x)-\lim\limits_{x\uparrow c}\mathrm{E}(y\,|\,x)}{\lim\limits_{x\downarrow c}\mathrm{E}(D\,|\,x)-\lim\limits_{x\uparrow c}\mathrm{E}(D\,|\,x)}$$

b − a

上式的分子就是精确断点回归的 LATE，而分母为得到处理的概率 (即倾向得分)在断点 c 处的跳跃$(b-a)$。

此式是精确断点回归公式的推广（精确断点情况下，$b-a=1$）。

b、a 的估计：x=c 时，D=1 的比例

D=0

或者 probit. logit.

此式的分子就是精确断点回归的 LATE，故可用精确断点回归(比如，局部线性回归)来估计此分子。

分母在形式上与分子完全一样，故也可用精确断点回归来估计，只要将结果变量 y 换为处理变量 D 即可。

进行模糊断点回归的另一方法为工具变量法。

定义 $Z_i = \mathbf{1}(x_i \geq c)$，则 Z_i 与处理变量 D_i 相关，满足相关性。

而 $Z_i = \mathbf{1}(x_i \geq c)$ 在断点 c 附近相当于局部随机实验，故只通过 D_i 影响 y_i，与扰动项 ε_i 不相关，满足外生性。

$Z_i \to W$

$\to D_i \to y_i - y_0$

因此，Z_i为D_i的有效工具变量，可使用 2SLS 进行估计。

如果使用相同的带宽 h，则此 2SLS 估计量在数值上正好等于使用矩形核的局部线性回归估计量。

以上的断点回归均假设在断点附近仿佛存在局部随机分组。

如果分组变量为年龄(时间)或地理区域，则这种解释一般行不通，称为“非随机断点设计”(Nonrandomized discontinuity design)。

例 以年龄 65 岁为分界线，年满 65 岁即可获得退休金。此时，分组变量为时间，是个确定性过程，个体无法控制。此时，须考虑以下三种可能性。

首先，年满 65 岁是否使得个体有资格参加其他项目，从而通过其他渠道影响结果变量。

其次，虽然年满 65 岁即可获得退休金，但退休金的效应可能需要几年后才能体现(即可能存在动态效应)。

最后，由于个体可以预见 65 岁以后将得到退休金，故可能在 65 岁之前就调整其经济行为。

对于这些可能性，应进行具体分析，才能得到令人信服的结论。

另一种非随机断点设计使用地理区域作为分组变量，以某种区域分界线作为断点，进行“地理断点回归”(geographic RD)。

例 Black (1999)通过比较在学区分界线两侧的房价来测算对居民对高质量小学教育的支付意愿(willingness to pay)。

由于个体一般可以选择住在学区分界线的哪一侧，故很难视为局部随机分组。需要说明，在分界线两侧，除了处理变量不同外，在其他方面均几乎没有差别。

为了保证分界线两侧的可比性，Black (1999)剔除了分界线为主要街道或高速公路的部分分界线(主要街道或高速公路两侧的社区可能有较大差别，尽管距离很近)。

例(强制矿工制度的长期经济影响) 在1573-1812年期间,西班牙殖民者在秘鲁与玻利维亚实行了一种称为“mining mita”的强制矿工征用制度。该制度规定，在离矿山较近的一定区域内，每个土著社区须提供其成年男性人口的七分之一作为强制矿工。

为研究此制度的长期经济影响，Dell (2010)使用断点回归来比较此区域分界线两侧的当代家庭消费与儿童发育不良比例。

为了保证分界线两侧具有可比性，该研究剔除了分界线一侧为平原而另一侧为安第斯山脉的部分分界线。

28.12 断点回归的 Stata 实例

28.13 处理效应模型

解决依不可测变量选择问题的另一方法：直接对处理变量D_i进行结构建模。Maddala (1983)提出“处理效应模型”(treatment effects model)：

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma D_i + \varepsilon_i$$

假设处理变量由以下“处理方程”(treatment equation)所决定：

$$D_i = \boldsymbol{1}(\boldsymbol{z}_i'\boldsymbol{\delta} + u_i)$$

其中，$\boldsymbol{z}_i$可以与$\boldsymbol{x}_i$有重叠的变量，但$\boldsymbol{z}_i$中至少有一个变量，比如z_{1i}，不在$\boldsymbol{x}_i$中。

并假设$\mathrm{Cov}(z_{1i}, \varepsilon_i) = 0$，即虽然$z_{1i}$影响个体是否参与项目$D_i$，但不直接影响结果变量$y_i$(只通过$D_i$间接影响$y_i$)；故可将$z_{1i}$视为 IV。

假设扰动项(ε_i, u_i)服从二维正态分布：

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & 1 \end{pmatrix} \right]$$

其中，ρ为(ε_i, u_i)的相关系数，而u_i的方差标准化为 1。

允许$\rho \neq 0$，这正是内生性的来源。

如果$\rho = 0$，则不存在内生性，可直接用 OLS 得到一致估计。

对于参加者而言，y_i的条件期望为

$$\begin{aligned}
\mathrm{E}(y_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid \boldsymbol{z}_i'\boldsymbol{\delta} + u_i > 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \mathrm{E}(\varepsilon_i \mid u_i > -\boldsymbol{z}_i'\boldsymbol{\delta}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \gamma + \rho\sigma_\varepsilon \lambda(-\boldsymbol{z}_i'\boldsymbol{\delta})
\end{aligned}$$

其中，$\lambda(\cdot)$为反米尔斯函数，即$\lambda(c) \equiv \dfrac{\phi(c)}{1-\Phi(c)}$（使用偶然断尾公式）。

未参加者的条件期望为

$$
\begin{aligned}
\mathrm{E}(y_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) &= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid \boldsymbol{z}_i'\boldsymbol{\delta} + u_i \le 0, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} + \mathrm{E}(\varepsilon_i \mid u_i \le -\boldsymbol{z}_i'\boldsymbol{\delta}, \boldsymbol{x}_i, \boldsymbol{z}_i) \\
&= \boldsymbol{x}_i'\boldsymbol{\beta} - \rho\sigma_\varepsilon \lambda(\boldsymbol{z}_i'\boldsymbol{\delta})
\end{aligned}
$$

将两方程相减，可得参加者与未参加者的条件期望之差：

$$
\mathrm{E}(y_i \mid D_i = 1, \boldsymbol{x}_i, \boldsymbol{z}_i) - \mathrm{E}(y_i \mid D_i = 0, \boldsymbol{x}_i, \boldsymbol{z}_i) = \gamma + \rho\sigma_\varepsilon \left[\lambda(-\boldsymbol{z}_i'\boldsymbol{\delta}) + \lambda(\boldsymbol{z}_i'\boldsymbol{\delta})\right]
$$

如果直接比较处理组与控制组的平均收益y_i，将遗漏上式右边第二项$\rho\sigma_\varepsilon\left[\lambda(-\boldsymbol{z}_i'\boldsymbol{\delta})+\lambda(\boldsymbol{z}_i'\boldsymbol{\delta})\right]$，导致不一致的估计(除非$\rho=0$)。

定义个体 i 的风险(hazard)为

$$\lambda_i=\begin{cases}\lambda(-\boldsymbol{z}_i'\boldsymbol{\delta}) & \text{若}D_i=1\\ -\lambda(\boldsymbol{z}_i'\boldsymbol{\delta}) & \text{若}D_i=0\end{cases}$$

可得到统一的方程(对于参加者与未参加者都适用)：

$$\mathrm{E}(y_i\mid\boldsymbol{x}_i,\boldsymbol{z}_i)=\boldsymbol{x}_i'\boldsymbol{\beta}+\gamma D_i+\rho\sigma_\varepsilon\lambda_i$$

可进行类似于 Heckit 的两步法估计。

第一步：用 Probit 估计方程 $\mathrm{P}(D_i = 1 \mid \boldsymbol{z}_i) = \Phi(\boldsymbol{z}_i'\boldsymbol{\delta})$，得到估计值 $\hat{\boldsymbol{\delta}}$，计算 $\hat{\lambda}_i$。

第二步：用 OLS 回归 $y_i \xrightarrow{\text{OLS}} \boldsymbol{x}_i, D_i, \hat{\lambda}_i$，得到估计值 $\hat{\boldsymbol{\beta}}, \hat{\gamma}, \widehat{\rho\sigma_\varepsilon}$。

两步法的优点是计算方便；其缺点在于，第一步的估计误差被带入第二步中，导致效率损失。

更有效率的做法是，使用 MLE，同时估计所有模型参数。

处理效应模型依赖于对结构方程的正确设定，如模型设定有误，z_{1i} 不是有效 IV，或扰动项不服从正态，都会导致不一致估计。