无混淆分配下的因果推断

慧航

2025年11月

匹配的假设

在Unconfoundedness假设条件下，可以方便的得到处理效应的识别。关键假设：

1. Unconfoundedness假设（CIA假设）：

$$w_i \perp\!\!\!\perp (y_i(1), y_i(0)) | x$$

2. 共同支撑假设（Common support assumption, CSA）：

$$0 < P(w_i = 1 | x_i) < 1$$

匹配的原理

CIA意味着均值独立，即：

$$\mathbb{E}\left(y_i\left(1\right)|x_i, w_i\right) = \mathbb{E}\left(y_i\left(1\right)|x_i\right)$$

$$\mathbb{E}\left(y_i\left(0\right)|x_i, w_i\right) = \mathbb{E}\left(y_i\left(0\right)|x_i\right)$$

而CSA需要对于相同的x_i，都有处理组和非处理组：经常需要trimming。

匹配的原理

由于:

$$\begin{aligned}&\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)\\&=\mathbb{E}\left(y_i\left(1\right)|x_i,w_i=1\right)-\mathbb{E}\left(y_i\left(0\right)|x_i,w_i=0\right)\\&=\mathbb{E}\left(y_i\left(1\right)|x_i\right)-\mathbb{E}\left(y_i\left(0\right)|x_i\right)\\&=\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)|x_i\right)\end{aligned}$$

因而

$$\mathbb{E}\left[\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)\right]=\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)\right)$$

同理:

$$\begin{aligned}&\mathbb{E}\left[\mathbb{E}\left(y_i|w_i=1,x_i\right)-\mathbb{E}\left(y_i|w_i=0,x_i\right)|w_i=1\right]=\\&\qquad\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)|w_i=1\right)\end{aligned}$$

无混淆分配下的推断：回归方法

方法一：回归

设定

$$\mathbb{E}\left(y_i|w_i=1,x_i\right)=\mathbb{E}\left(y_i\left(1\right)|x_i\right)=\mu_1\left(x_i\right)$$

$$\mathbb{E}\left(y_i|w_i=0,x_i\right)=\mathbb{E}\left(y_i\left(0\right)|x_i\right)=\mu_0\left(x_i\right)$$

平均处理效应：

$$\hat{\tau}_{reg}=\frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1\left(x_i\right)-\hat{\mu}_0\left(x_i\right)\right]$$

1. 简单线性回归：

$$y_i=\alpha+x_i'\beta+\tau\cdot w_i+\epsilon_i$$

或者：

$$y_i=\alpha+\dot{x}_i'\beta+\tau\cdot w_i+w_i\dot{x}_i'\delta+\epsilon_i$$

2. 非参数回归

无混淆分配下的推断：匹配

或者：

- Nearest Neighbor Matching：
  1. 给定一个正的常数M，比如$M=1$
  2. 令$d(\cdot,\cdot)$为一个距离函数，比如欧几里得距离：

$$d(x_i,x_j)=(x_i-x_j)'(x_i-x_j)$$

  或者Mahalanobis距离：

$$d(x_i,x_j)=(x_i-x_j)'\Sigma_x^{-1}(x_i-x_j)$$

- caliper：距离小于一个临界值（caliper）即匹配成功

临近匹配

Nearest-neighbor matching: 对于任意处理组的i，从控制组中找到最近的M个控制组个体，记：

$$J_M(i) = \{l_1(i), ..., l_M(i)\}$$

定义

$$\hat{y}_i(0) = \frac{1}{M} \sum_{m \in J_M(i)} y_m$$

可以使用：

$$\frac{1}{N_1} \sum_{i|w_i=1} [y_i(1) - \hat{y}_i(0)]$$

匹配的细节

实践中，有不同的匹配方案：

1. 选择M，一般而言如果控制组数量远远大于实验组数量，可以使用较多的M
2. 序贯/非序贯
   - 序贯：按顺序一个一个配对
   - 非序贯：所有的实验组和控制组放在一起考虑
3. 贪婪/非贪婪
   - 贪婪：有放回
   - 非贪婪：无放回
4. 先进行分组或者分层（stratification），组内进行匹配
5. 使用propensity score进行排序，进而匹配
6. Stata:
   - psmatch2
   - teffects nnmatch（推荐）

匹配的步骤

倾向得分匹配的步骤：

1. 定义距离
2. 匹配
3. 评估匹配结果：
   1. balancing: t-test、Standardised Bias
   2. unconfoundedness：使用明显无效应的其他的Y
4. 评估政策效应
5. 敏感性分析

评估匹配结果：协变量

评估匹配结果：

1. normalized difference:

$$\hat{\Delta}_{ct} = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_c^2 + s_t^2}{2}}}$$

2. t-stat:

$$T_{ct} = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}}}$$

Stata: tebalance

Balancing

Table 16: Summary Statistics for Non-experimental Lalonde Data

	CPS controls (N_c=15,992)		trainees (N_t=185)			
Covariate	mean	(s.d.)	mean	(s.d.)	t-stat	nor-dif
Black	0.07	0.26	0.84	0.36	28.6	2.43
Hisp	0.07	0.26	0.06	0.24	-0.7	-0.05
Age	33.23	11.05	25.82	7.16	-13.9	-0.80
Married	0.71	0.45	0.19	0.39	-18.0	-1.23
Nodegree	0.30	0.46	0.71	0.46	12.2	0.90
Education	12.03	2.87	10.35	2.01	-11.2	-0.68
E'74	14.02	9.57	2.10	4.89	-32.5	-1.57
U'74	0.12	0.32	0.71	0.46	17.5	1.49
E'75	13.65	9.27	1.53	3.22	-48.9	-1.75
U'75	0.11	0.31	0.60	0.49	13.6	1.19

	Full Sample nor-dif	Matched Sample nor-dif	ratio of nor-dif
Black	2.43	0.00	0.00
Hispanic	-0.05	0.00	-0.00
Age	-0.80	-0.15	0.19
Married	-1.23	-0.28	0.22
Nodegree	0.90	0.25	0.28
Education	-0.68	-0.18	0.26
E'74	-1.57	-0.03	0.02
U'74	1.49	0.02	0.02
E'75	-1.75	-0.07	0.04

Matching

其他匹配方法:

1. Kernel matching
2. Radius matching
3. Stratification or interval matching
4. Propensity score matching

倾向得分匹配

倾向得分（Propensity score）方法：Rosenbaum and Rubin(1983)证明：

$$w_i \perp\!\!\!\perp (y_i(1), y_i(0)) | x_i \Longleftrightarrow w_i \perp\!\!\!\perp (y_i(1), y_i(0)) | P(x_i)$$

因而控制倾向得分匹配就足够了。

倾向得分匹配

Rosenbaum and Rubin(1983)提出了三阶段的方法：

1. 估计倾向得分：
$$P\left(w_i|x_i\right)$$
2. 用y_i对w_i和p_i做回归，得到$\hat{\mathbb{E}}\left(y_i|w_i,p_i\right)$
3. 估计ATT：
$$\frac{1}{N_1}\sum_{i|w_i=1}\left[y_i\left(1\right)-\hat{y}_i\left(0\right)\right]$$

注意在使用Propensity Score时，一定要注意Common support假设：trimming!

逆概率加权法

注意到，由于：

$$\begin{aligned}
\mathbb{E}\left(\frac{w_i y_i}{P(x_i)}\right) &= \mathbb{E}\left(\frac{w_i y_i(1)}{P(x_i)}\right) \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{w_i y_i(1)}{P(x_i)}|x_i\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(w_i y_i(1)|x_i)}{P(x_i)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}(w_i|x_i)\,\mathbb{E}(y_i(1)|x_i)}{P(x_i)}\right] \\
&= \mathbb{E}\left[\mathbb{E}(y_i(1)|x_i)\right] = \mathbb{E}(y_i(1))
\end{aligned}$$

同理：

$$\mathbb{E}\left(\frac{(1-w_i)\,y_i}{1-P(x_i)}\right) = \mathbb{E}(y_i(0))$$

逆概率加权法

因而平均处理效应：

$$\tau_{ATE} = \mathbb{E}\left[\frac{w_i y_i}{P(x_i)} - \frac{(1-w_i)\, y_i}{1-P(x_i)}\right]$$

可以使用：

$$\hat{\tau}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_i y_i}{P(x_i)} - \frac{(1-w_i)\, y_i}{1-P(x_i)}\right]$$

进行估计，称为Inverse Propensity Weighting(IPW)。其中$P(x_i)$可以使用Logistic sieve估计量（Hirano, Imbense and Ridder, 2003）。
Stata: teffects ipw

双向稳健的处理效应评估方法

然而IPW方法对倾向得分非常敏感。可以考虑使用Robins等人提出的双向稳健（Double robustness）方法：

1. 结合了回归方法和IPW
2. 只需要$P(x_i)$或者结果方程至少有一个设定正确（双向稳健）

最小化：

$$\min_{\alpha_0,\beta_0} \sum_{i|w_i=0} \frac{\left[y_i - \alpha_0 - \beta_0'(x_i - \bar{x}_i)\right]^2}{1 - P(x_i;\hat{\gamma})}$$

$$\min_{\alpha_1,\beta_1} \sum_{i|w_i=1} \frac{\left[y_i - \alpha_1 - \beta_1'(x_i - \bar{x}_i)\right]^2}{P(x_i;\hat{\gamma})}$$

平均处理效应为：

$$\hat{\tau}_{ATE} = \hat{\alpha}_1 - \hat{\alpha}_0$$

Stata: teffects aipw

双向稳健的处理效应评估方法

- 不妨考虑一阶条件（对α_1求导）：

$$-2\sum \frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}=0$$

- 考虑其总体形式

$$\begin{aligned}
\mathbb{E}\left(\frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}\right) &= \mathbb{E}\left[\mathbb{E}\left(\frac{w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]}{P\left(x_i;\hat{\gamma}\right)}|x_i\right)\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i\left[y_i-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i\left[y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)\right]|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}\left(w_i|x_i\right)\mathbb{E}\left(y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right]
\end{aligned}$$

双向稳健的处理效应评估方法

- 总体一阶条件：

$$\mathbb{E}\left[\frac{\mathbb{E}\left(w_i|x_i\right)\mathbb{E}\left(y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)}{P\left(x_i;\hat{\gamma}\right)}\right]=0$$

的两种情况：

- $P\left(x_i;\hat{\gamma}\right)$正确设定：那么$\alpha_1=\mathbb{E}\left[\mathbb{E}\left(y_i\left(1\right)\right)\right]=\mathbb{E}\left(y_i\left(1\right)\right)$
- $y_i\left(1\right)$函数形式设定正确：$\mathbb{E}\left(y_i\left(1\right)-\alpha_1-\beta_1'\left(x_i-\bar{x}\right)|x_i\right)=0$成立，$\alpha_1$得到一致估计。

Matching

其他方法：Imai and Ratkovic (2014)：解：

$$\frac{1}{N}\sum_{i=1}^{N} g_{\gamma}\left(w_i, x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{w_i}{P\left(x_i;\gamma\right)} - \frac{1-w_i}{1-P\left(x_i;\gamma\right)}\right) f\left(x_i\right)\right] = 0$$

仍然是双向稳健的。此外，Fan et al. (2016)做了推广。

DDML与因果推断

- 以上讨论的部分线性模型可以看做是对同质性处理效应的建模，即假设数据生成过程为：

$$y_i = \tau w_i + g(x_i) + u_i$$

其中$w = 0/1$为处理变量。根据该设定，其反事实为

$$y_i(0) = g(x_i) + u_i$$

$$y_i(1) = \tau + g(x_i) + u_i$$

那么处理效应$y_i(1) - y_i(0) = \tau$，从而平均处理效应与处理组平均处理效应相等，都为τ。

- 在无混淆分配的假设，即$w_i \perp\!\!\!\perp (y_i(0), y_i(1)) | x_i$的条件下，可以自然得到$u_i \perp\!\!\!\perp w_i | x_i$，从而我们可以使用双重机器学习的方法对以上问题进行估计。

异质性处理效应

- 然而以上模型对于处理效应同质性的假设仍然太强。为此我们不妨暂时放弃函数形式假设，先从识别层面探讨可能的异质性处理效应的识别问题。为此我们现引入如下假设：
  1. 假设1. **无混淆分配假设**：$w_i \perp\!\!\!\perp (y_i(0), y_i(1)) | x_i$
  2. 假设2. **共同支撑假设**：对于任意的$x_i = x$，倾向得分$p(x) \triangleq \mathrm{P}(w_i = w | x_i = x) > 0, w = 0/1$
  3. 假设3. **SUTVA假设**：$y_i = y_i(w_i)$

条件平均处理效应

- 如此，如果记$g(w,x)=\mathbb{E}(y_i|w_i=w,x_i=x)$，那么：

$$g(w,x)=\mathbb{E}(y_i|w_i=w,x_i=x)=\mathbb{E}(y_i(w)|w_i=w,x_i=x)=\mathbb{E}(y_i(w)|x_i=x)$$

- 以上等式意味着为了获得$g(w,x)$的估计，我们可以使用根据w分组，使用y对x做回归分别得到$g(0,x)$以及$g(1,x)$，将两者相减就可以得到

$$g(1,x)-g(0,x)=\mathbb{E}(y_i(1)-y_i(0)|x_i=x)=\text{CATE}(x)$$

其中$\text{CATE}(x)$为给定特征x的平均处理效应，或者**条件平均处理效应**（conditional average treatment effects）。

交互模型

- Chernozhukov等（2017）即考虑了这种数据生成过程：

$$y = g(w, x) + u$$

- $$w = m(x) + v$$

  其中无混淆分配假设即$\mathbb{E}(u|w, x) = 0$，该设定实际上允许处理变量w与x之间任意的交互，或者说处理效应可以是x的任意函数，从而以上模型也被称为“**交互模型**”（**interactive model**）。
- 如果需要进一步得到ATE和ATT，可以计算

$$\text{ATE} = \mathbb{E}[g(1, x) - g(0, x)]$$

$$\text{ATT} = \mathbb{E}[g(1, x) - g(0, x)|w = 1]$$

逆概率加权

- 进一步，我们也可以将以上的识别方法写成逆概率加权的形式：

$$\begin{aligned}\mathbb{E}\left(\frac{w_i y_i}{p\left(x_i\right)}\middle| x_i=x\right) &= \frac{\mathbb{E}\left(w_i y_i|x_i=x\right)}{p\left(x\right)} = \frac{\mathbb{E}\left(w_i y_i\left(1\right)|x_i=x\right)}{p\left(x\right)} \\ &= \frac{\mathbb{E}\left(w_i|x_i=x\right)\mathbb{E}\left(y_i\left(1\right)|x_i=x\right)}{p\left(x\right)} \\ &= \mathbb{E}\left(y_i\left(1\right)|x_i=x\right)\end{aligned}$$

类似的，

$$\mathbb{E}\left[\frac{\left(1-w_i\right)y_i}{1-p\left(x_i\right)}\middle| x_i=x\right] = \mathbb{E}\left(y_i\left(0\right)|x_i=x\right)$$

- 从而条件平均处理效应可以使用

$$\mathrm{CATE}\left(x\right) = \mathbb{E}\left[\frac{w_i y_i}{p\left(x_i\right)} - \frac{\left(1-w_i\right)y_i}{1-p\left(x_i\right)}\middle| x_i=x\right]$$

进行识别。

逆概率加权

- 而更进一步，我们可以证明

$$\mathbb{E}\left(y_i\left(1\right)|x_i=x\right)=\mathbb{E}\left[g\left(1,x_i\right)+\frac{w_i\left(y_i-g\left(1,x_i\right)\right)}{p\left(x_i\right)}\middle|x_i=x\right]$$

$$\mathbb{E}\left(y_i\left(0\right)|x_i=x\right)=\mathbb{E}\left[g\left(0,x_i\right)+\frac{\left(1-w_i\right)\left(y_i-g\left(0,x_i\right)\right)}{1-p\left(x_i\right)}\middle|x_i=x\right]$$

- 好处：在实证中，$g\left(\cdot,\cdot\right)$函数和$p\left(x_i\right)$都需要通过模型进行估计，从而都可能会存在设定错误或者偏误。而这一方法的好处在于其是“**双向稳健**”（**doubly robust**）的，即只要$g\left(\cdot,\cdot\right)$函数和$p\left(x_i\right)$只要有一个估计准确，那么以上等式即成立。

DDML方法

- 基于此，Chernozhukov（2017）提出可以通过定义

$$\psi(\tau,\eta;y_i,w_i,x_i)=g(1,x_i)-g(0,x_i)+\frac{w_i(y-g(1,x_i))}{p(x_i)}-\frac{(1-w_i)(y-g(0,x_i))}{1-p(x_i)}-\tau$$

并使用矩条件$\mathbb{E}[\psi(\tau,\eta;y_i,w_i,x_i)]=0$求解出平均处理效应$\tau$

- 其中$\eta=(g(1,x),g(0,x),p(x))$为冗余参数。可以证明以上的矩条件是满足Neyman正交化条件的，即

$$\partial_\eta\mathbb{E}\psi(\tau,\eta;y,w,x)|_{\eta=\eta_0}=0$$

从而我们可以使用双重机器学习的方法对τ进行估计。

DDML方法

- 或者也可以对处理组平均处理效应进行评估。根据Farrell（2015），有

$$\mathbb{E}\left(y_i\left(1\right)|w_i=1\right)=\mathbb{E}\left[\frac{w_i y_i}{p_1}\right]$$

$$\mathbb{E}\left(y_i\left(0\right)|w_i=1\right)=\mathbb{E}\left[\frac{w_i g\left(0,x_i\right)}{p_1}+\frac{p\left(x_i\right)}{p_1}\frac{\left(1-w_i\right)\left(y_i-g\left(0,x_i\right)\right)}{1-p\left(x_i\right)}\right]$$

- 其中p_1为处理组样本量占总样本量的比例，从而处理组平均处理效应：

$$\mathbb{E}\left(y_i\left(1\right)-y_i\left(0\right)|w_i=1\right)=\mathbb{E}\left[\frac{w_i\left(y_i-g\left(0,x_i\right)\right)}{p_1}-\frac{p\left(x_i\right)}{p_1}\frac{\left(1-w_i\right)\left(y_i-g\left(0,x_i\right)\right)}{1-p\left(x_i\right)}\right]$$

Chernozhukov（2017）提出可以根据以上条件，定义

$$\psi\left(\tau,\eta;y_i,w_i,x_i\right)=\frac{w_i\left(y_i-g\left(0,x_i\right)\right)}{p_1}-\frac{p\left(x_i\right)}{p_1}\frac{\left(1-w_i\right)\left(y_i-g\left(0,x_i\right)\right)}{1-p\left(x_i\right)}-\tau\frac{w_i}{p_1}$$

其中$\eta=\left(g\left(0,x\right),p\left(x\right),p_1\right)$为冗余参数。

DDML方法

NSW中的DDML

- 如上方法可以使用ddml命令直接进行估计，比如对于NSW项目的政策评估
- ddml_nsw.do

因果树

- 回归树通过不断划分特征空间的方法可以实现对y的预测或者对w的分类，这是一种“非参数”的方法：我们无需假设y或者w的数据生成过程，甚至也无需设置特征对y的影响的异质性，分类和回归树本身就可以发现这些“交互”作用。
- 那么是否可以将分类和回归树的方法直接用于处理效应的估计呢，从而解决异质性处理效应的估计问题呢?
- 如，我们可以构造一颗预测y的回归树，假设叶子结点的集合为$\Pi = \{\ell_1, ..., \ell_L\}$，我们可以仿照以上逆概率加权的方法，使用

$$\hat{\tau}_\ell = \frac{1}{\#\ell} \sum_{i \in \ell} \left[\frac{w_i y_i}{p(x_i)} - \frac{(1 - w_i) y_i}{1 - p(x_i)} \right]$$

对CATE进行估计。
- 然而以上方法存在很严重的问题。为了构建回归树，我们使用了逐步增加节点的纯度的方法，而纯度的增加就是节点内y的相似程度增加，而以上的估计本质上是在节点内对处理组和对照组的y进行加权比较，按照这样的算法，以上

因果树

- Athey和Imbens（2016）提出了适应因果推断的“**因果树**”（**causal tree**），使用了一种“**诚实方法**”（**honest approach**）构建树，从而达到对异质性处理效应的识别。
- 该方法的一个重要改进是将样本分为了两部分：训练集I^{tr}和估计集I^{est}，当然，为了选择超参数还可以多分出一个验证集。
- 重要的在于，在构建树的过程中仅仅使用训练集，而估计政策效应则使用估计集数据，因而这种树的构造方法也被称为“**双样本树**”（**double-sample tree**）。
- 此外，为了适应政策效应的评估，Athey和Imbens（2016）还改进了构造树时的准则，使用期望的MSE作为准则估计，从而可以更好的估计政策效应。

因果森林

- 随后，Wager和Athey（2019）进一步将因果树推广为**因果森林（causal forest）**。他们提出了两种不同的方法：
  - 双样本树，即使用Athey和Imbens（2016）的方法，每次从样本中进行有放回抽样，在每个bootstrap样本中都进行双样本树的建模，最后将每颗树得到的CATE平均得到最终的估计；
  - 倾向得分树：不使用双样本的方法，而是直接以w_{i}作为训练目标构建树。
- Wager和Athey（2019）证明以上因果森林的估计结果是渐近正态的，且其渐近方差可以方便的估计出。
- Athey等（2019）又对以上方法做了更多拓展，提出了“**广义随机森林**”（**generalized random forest**）以解决更多的参数估计问题，这里不再赘述。广义随机森林在R语言中可以使用grf包进行估计，在Python中也可以使用EconML包进行分析。