Lecture 8

在观察性研究中，**匹配（Matching）** 是一种通过选择与处理组个体在协变量上相似的对照组个体，以减少混杂偏差的技术。其核心目标是使处理组和对照组在观察到的协变量上尽可能相似，从而模拟随机实验的条件，提高因果效应估计的准确性。

为什么要用匹配?

在观察性数据中，处理组的分配通常不是随机的，可能导致协变量分布不均衡。例如，接受职业培训的人可能更年轻或教育水平更高。如果不调整这些差异，直接比较处理组和对照组的结果会产生偏差。匹配通过"平衡"协变量分布，降低混杂因素的影响。

常见的匹配方法

方法	原理	适用场景
精确匹配	要求协变量完全匹配（如性别必须相同）	关键协变量必须严格一致（如性别、地区）
最近邻匹配	为每个处理组个体选择协变量最接近的对照组个体（如欧氏距离最小）	协变量维度低、样本量较大时
卡尺匹配	在最近邻匹配基础上，限制匹配的最大距离（卡尺），避免匹配到差异过大的个体	防止匹配到不相似的个体，提高质量
倾向得分匹配	基于倾向得分（propensity score）进行匹配，平衡协变量的整体分布	高维协变量时更高效

倾向得分匹配（Propensity Score Matching）的步骤

1. **估计倾向得分**：
   使用逻辑回归、随机森林等模型预测每个个体接受处理的概率 (propensity score)：
   $$p(X_i) = P(W_i = 1 \mid X_i) \tag{1}$$
   - 协变量 X_i 包括所有可能影响处理分配的变量（如年龄、教育、收入）。
2. **选择匹配方法**：
   - **最近邻匹配**：为每个处理组个体选择倾向得分最接近的对照组个体。
   - **卡尺匹配**：在倾向得分的一定范围内（如±0.1）寻找匹配。
   - **核匹配**：使用加权平均，权重随倾向得分距离增加而衰减。
3. **执行匹配**：
   根据匹配方法筛选对照组样本，形成匹配后的数据集。
4. **评估匹配质量**：
   - **标准化均值差（SMD）**：检查匹配后各协变量的均值差异是否缩小（通常要求SMD < 0.1）。
   - **可视化**：绘制匹配前后倾向得分的分布图（如直方图、QQ图）。
5. **估计因果效应**：
   在匹配后的样本中，直接比较处理组和对照组的平均结果差异（如ATE）。
   公式：
   $$\hat{\tau} = \frac{1}{N_T} \sum_{i \in T} Y_i - \frac{1}{N_C} \sum_{j \in C} Y_j \tag{2}$$

实例：职业培训对收入的影响

目标：估计参加职业培训（处理组）对年收入（结果）的因果效应。
步骤：

1. **数据准备**：收集参与者的年龄、教育、工作经验、性别等协变量，以及是否参加培训和年收入。
2. **估计倾向得分**：用逻辑回归建模 $P(\text{参加培训} \mid \text{年龄、教育、经验等})$。
3. **卡尺匹配**：为每个处理组个体匹配倾向得分相差不超过0.05的对照组个体。
4. **评估平衡性**：
   - 匹配前：处理组的平均教育水平显著高于对照组。
   - 匹配后：两组的年龄、教育、经验等协变量的SMD均<0.1。
5. **计算ATE**：匹配后处理组的平均收入为8万元，对照组为6.5万元，ATE=1.5万元。

§1 Matching

9 Matching in observational studies

Matching has a long history in empirical research, and W. Cochran and D. Rubin popularized it in statistical causal inference. In this section, we consider matching techniques in analyzing average treatment effect, and see what unique problems this technique brings to practitioners.

1. Matching 的情境设定

Let us first consider a hypothetical situation, i.e., ① the number of units in the control group is significantly larger than the treatment group. Specifically, ② for each unit i inside the treatment group($W_i = 1$), there exists a $h(i)$ in the control group($W_{h(i)} = 0$) satisfying

$$X_i = X_{h(i)}.$$

In that case, we have

$$Prob(W_i = 1, W_{h(i)} = 0|X_i, X_{h(i)}) = Prob(W_i = 1|X_i) \times Prob(W_{h(i)} = 0|X_{h(i)}) = p(X_i) \times (1 - p(X_i)).$$

Similarly,

$$Prob(W_i = 0, W_{h(i)} = 1|X_i, X_{h(i)}) = Prob(W_i = 0|X_i) \times Prob(W_{h(i)} = 1|X_{h(i)}) = (1 - p(X_i)) \times p(X_i).$$

In particular, ⇓

$$Prob(W_i = 1, W_{h(i)} = 0|W_i + W_{h(i)} = 1, X_i, X_{h(i)}) = \frac{p(X_i) \times (1 - p(X_i))}{p(X_i) \times (1 - p(X_i)) + (1 - p(X_i)) \times p(X_i)} = 1/2.$$

2. 使用 exact matching 进行 test

Therefore, if we ① pick data (X_i, W_i, Y_i) such that $W_i = 1$ or $i = h(j)$ for some $W_j = 1$, then we can do a randomization trick to establish the CI of the estimator. That is, calculating (挑选符合 matching 的 n_1 组 pairs 作为 sample)

② $$\widehat{\tau} = \frac{1}{n_1} \sum_{W_i=1} Y_i - \frac{1}{n_1} \sum_{W_i=0} Y_i, \text{ where } n_1 = \{W_i = 1\},$$ (用 n_1 组 pairs 计算 ATE estimator)

then ③ for each i, drawing $W_i^* = 1$ with probability $1/2$, $W_{h(i)}^* = 1 - W_i$. Then calculate the new estimator (在每个 pair 内重新抽取 W_i 和 $W_{h(i)}$)

④ $$\widehat{\tau}_b^* = \frac{1}{n_1} \sum_{W_i^*=1} Y_i - \frac{1}{n_1} \sum_{W_i^*=0} Y_i,$$ ($H_0: \mu_1 = \mu_0$ 下 τ 和 τ_b^* 同分布)

and repeat this step for $b = 1, \cdots, B$. This helps perform hypothesis testing (by choosing the $\alpha/2$ and $1 - \alpha/2$th quantile of $\widehat{\tau}_i^*$). However, in general, exact matching is hard since it is not simple to find $h(i)$ that satisfies $X_i = X_{h(i)}$.

3. Inexact matching 的判定

In practice, however, it is not easy to find exact matching. What we can achieve is find $h(i)$ such

that $X_i \approx X_{h(i)}$. For example, we consider

(对于 treatment group 中的 sample，在 control group 中寻找 X 最"接近"的 sample 组成 matching)

$$\widehat{h}(i) = \arg\min_{k:W_k=0} d(X_i, X_k) \quad \forall i: W_i = 1$$

where d is a distance, such as the Euclidean distance and the Mahalanobis distance

4. Mahalanobis distance $d(x,y) = |x-y|_2^2$ or $d(x,y) = (x-y)^\top \Omega^{-1}(x-y)$ (Mahalanobis distance),

Ω is the sample covariance matrix of X_i, i.e.,

$$\Omega = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^\top.$$

Remark [The idea of Mahalanobis distance] The idea of using Mahalanobis distance is simple. To illustrate, suppose $X_i = (X_{i1}, \cdots, X_{ip})^\top$ where X_{ij} is independent of X_{ik} for $j \neq k$. Moreover, assume $Var(X_{ij}) = \sigma_j^2$ and $\mathbf{E}X_{ij} = 0$. If that happens, then

$$\mathbf{E}X_i X_i^\top = diag(\sigma_1^2, \cdots, \sigma_p^2).$$

Based on this observations, we have

$$(X_i - X_j)^\top \Omega^{-1}(X_i - X_j) = \sum_{k=1}^{p} \frac{(X_{ik} - X_{jk})^2}{\sigma_k^2}.$$

(Mahalanobis distance 相当于进行了 normalization)

In other words, if an element X_{ik} has large variance, then statisticians may consider this dimension to be less informative, and thus give a small weight to this dimension. From this observation, we can see that the Mahalanobis is suitable for the data that have significant different scales among different dimensions.

Remark [What happens if we have large dimension covariates] To illustrate what happens, suppose $X_i \sim N(0, I_p)$. In that case

$$\mathbf{E}|X_i - X_k|_2^2 = 2p,$$ (高维情况下点之间距离很大)

that is, on average, the distance between X_i and X_k is $2p$, making it hard to find a close pair. Therefore, for a high dimension covariate set, sorts of dimension reduction techniques, such as PCA, needs to be resorted to before finding the matching pair.

5. Inexact matching 的 ATE estimator We may consider the matching ATE estimator to be generated in the following form:

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i(1) - \frac{1}{n}\sum_{i=1}^{n}\widehat{Y}_i(0),$$

(共 n 组"pairs"，每组"pair"中的 samples 的 X 很接近，其中有 1 个 sample 来自 treatment group，M 个 samples 来自 control group)

where

$$\widehat{Y}_i(1) = \begin{cases} Y_i \text{ if } W_i = 1, \\ \frac{1}{M}\sum_{k\in J_i} Y_k, \text{ where } J_i \text{ consists of } M \text{ minimizers of } d(X_i, X_k) \text{ with } W_k = 1 \text{ if } W_i = 0 \end{cases}$$

and

$$\widehat{Y}_i(0) = \begin{cases} Y_i \text{ if } W_i = 0, \\ \frac{1}{M}\sum_{k\in J_i} Y_k, \text{ where } J_i \text{ consists of } M \text{ minimizers of } d(X_i, X_k) \text{ with } W_k = 0 \text{ if } W_i = 1. \end{cases}$$

b. Inexact matching estimator 的问题

However, inevitably this kind of estimator incurs bias. Actually, Abadie and Imbens found that the estimator for bias is of the following form

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} \widehat{B}_i, \text{ where } \widehat{B}_i = (2W_i - 1)\frac{1}{M}\sum_{k\in J_i}(\widehat{\mu}_{1-Z_i}(X_i) - \widehat{\mu}_{1-Z_i}(X_k)).$$

Therefore, if we have a large dimension X_i, then $\widehat{\mu}_{1-Z_i}(X_i) - \widehat{\mu}_{1-Z_i}(X_k)$ becomes large in absolute value, and the bias becomes non-negligible.

Remark For simplicity, here we assume $M = 1$, and define $\mu(x) = \mathbf{E}Y_i(1)|X_i = x$, $\nu(x) = \mathbf{E}Y_i(0)|X_i = x$. Then

$$\widehat{Y}_i(1) = W_iY_i(1) + (1 - W_i)Y_k(1), \quad \text{where } k = k(X_1, \cdots, X_n).$$

Taking conditional expectations on $X_1, \cdots, X_n$, we can roughly derive this result.

§2 Propensity score matching

10 Propensity score matching

In practice, however, it is unrealistic to assume that abundant overlap exists, especially when the covariates/confounders are of high–dimensions. In other words, when we try to use distance–based matching, it is possible that all X_j has a large distance to X_i. Therefore, we may need some techniques that compress the dimension of confounders and extract essential information. Propensity score is definitely one of the useful property to describe the covariates. Especially, the propensity score has a unique propety, as described below.

1. propensity score summarizes 了 X 的信息 (同 Lecture 5/§3/3)

Theorem 10.1. *Suppose W_i is independent of $Y_i(0), Y_i(1)$ conditional on X_i, then W_i is independent of $Y_i(0), Y_i(1)$ conditional on $p(X_i) = (\mathbf{E}W_i = 1|X_i)$.*

Proof Based on definition, it suffices to show that

$$\mathbf{E}h(W_i)g(Y_i(0),Y_i(1))|p(X_i) = \mathbf{E}h(W_i)|p(X_i) \times \mathbf{E}g(Y_i(0),Y_i(1))|p(X_i).$$

Notice that

$$\begin{aligned}\mathbf{E}h(W_i)g(Y_i(0),Y_i(1))|X_i &= \mathbf{E}h(W_i)|X_i \times \mathbf{E}g(Y_i(0),Y_i(1))|X_i \\ &= (h(1)p(X_i) + h(0)(1-p(X_i))) \times \mathbf{E}g(Y_i(0),Y_i(1))|X_i,\end{aligned}$$

therefore, from tower property,

$$\begin{aligned}\mathbf{E}h(W_i)g(Y_i(0),Y_i(1))|p(X_i) &= \mathbf{E}(h(W_i)g(Y_i(0),Y_i(1))|X_i)|p(X_i) \\ &= (h(1)p(X_i) + h(0)(1-p(X_i))) \times \mathbf{E}g(Y_i(0),Y_i(1))|p(X_i) \\ &= \mathbf{E}h(W_i)|p(X_i) \times \mathbf{E}g(Y_i(0),Y_i(1))|p(X_i),\end{aligned}$$

and we prove the result. □

2. Propensity score matching 的 test (p(X) discrete)

From this result, we know that the observational data behave like randomization experiment when conditional on $p(X_i) = x$. To use this property, we consider a simple situation. Suppose the propensity score only takes K values, i.e., $p(x) \in \{c_1, \cdots, c_K\}$. If we choose all X_i having propensity e_k, since W_i is independent of $Y_i(0), Y_i(1)$ conditional on $p(X_i) = e_k$, when we choose $p(X_i) = e_k$ and calculate the two–sample test statistics

$$\widehat{\tau}_k = \frac{1}{v_1}\sum_{i=1}^{v} W_i Y_i - \frac{1}{v_0}\sum_{i=1}^{v}(1-W_i)Y_i \text{ where } v = \#\{p(X_i) = e_k\}.$$

For

$$\mathbf{E}W_iY_i(1)|e(X_i) = \mathbf{E}W_i|e(X_i) \times \mathbf{E}Y_i(1)|e(X_i),$$

we have

$$\mathbf{E}\widehat{\tau}_k \approx \mathbf{E}Y_i(1) - \mathbf{E}Y_i(0).$$

Remark Generally speaking, if W_i is not independent of $Y_i(1)$ and $Y_i(0)$, e.g., when

$$\mathbf{E}W_iY_i(1)|X_i = p(X_i)\mathbf{E}Y_i(1)|X_i,$$

we have

$$\mathbf{E}W_iY_i(1) = \mathbf{E}p(X_i)\mathbf{E}Y_i(1)|X_i \neq \mathbf{E}Y_i(1),$$

so the two sample test may not work.

3. Propensity score matching 的 test (P(X) continuous)

In practice, the propensity score may be a continuous function. In that case, we artifically separate the region $[0,1)$ into several pieces $[\frac{k-1}{K}, \frac{k}{K})$; and calculate the two sample test statistics

$$\widehat{\tau}_k^* = \widehat{\tau}_k = \frac{1}{v_1}\sum_{i=1}^{v} W_iY_i - \frac{1}{v_0}\sum_{i=1}^{v}(1-W_i)Y_i \text{ where } v = \#\{p(X_i) \in [\frac{k-1}{K}, \frac{k}{K})\}.$$

If K is sufficiently large, then we may consider the propensity score to be the same, and the corresponding two sample test statistics should be consistent to the real ATE.