

# 因果推断在观察性研究中的应用 II：分析 | 因果科学读书会第4期

## 【读书会相关信息】

“因果”并不是一个新概念，而是一个已经在多个学科中使用了数十年的分析技术。通过前两季的分  
享，我们主要梳理了因果科学在计算机领域的前沿进展。如要融会贯通，我们需要回顾数十年来在社会学、经济学、医学、生物学等多个领域中，都是使用了什么样的因果模型、以什么样的范式、解决了什么样的问题。我们还要尝试进行对比和创新，看能否以现在的眼光，用其他的模型，为这些研究提供新的解决思路。

如果大家对这个读书会感兴趣，欢迎报名：

[https://pattern.swarma.org/mobile/study\\_group/10?from=wechat](https://pattern.swarma.org/mobile/study_group/10?from=wechat)

【时间】2021年11月21日 9:00--11:00

## 【主讲人】

邓宇昊，北京大学数学科学学院统计学2018级博士生，导师为周晓华教授，主要研究方向为生物统计、因果推断、临床试验研究中的统计学方法，已在Biometrics、Statistics in Medicine等杂志发表多篇论文。

## 【笔记小分队】

段月然 中国地质大学（北京）

陈虹宇 西安财经大学

## 【讲座笔记】

记录人：陈虹宇 西安财经大学

## 逆概率加权估计

例子：非正式收入对经济行为有何影响？（Imbens、Rubin 和 Sacerdote 的一项研究）

- 实验组（成功者）：在马萨诸塞州彩票中玩过并赢得大笔奖金的个人。
- 对照组（失败者）：买彩票但只中了少量奖品的个人。
- 协变量：协变量包括彩票购买数量、教育程度、是否工作、买彩票前六年的收入、买彩票六年后的平均收入、年龄、中奖年份、性别等。

## 倾向得分（propensity score）

- 观察性研究一般先做非混淆性假设：给定协变量后，处理和潜在结果独立

1. 定义:给定协变量之后, 一个人属于处理组的概率

$$e(X_i) = P(W_i = 1 | X_i)$$

2. 两条性质:

- 均衡性质: 给定倾向得分后, 协变量与处理是独立的, 处理组与对照组协变量分布近似相等

$$W_i \perp X_i | e(X_i)$$

- 给定倾向得分后处理与潜在结果独立, 可把多维的协变量转化为一维的倾向得分, 降低了所需控制的变量的维度

$$W_i \perp (Y_i(1), Y_i(0)) | e(X_i)$$

- 倾向得分一般用logistic回归来估计

## 逆概率加权估计 (Horvitz-Thompson估计)

$$\mathbb{E} \left[ \frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} \right] = \mathbb{E}_{\text{sp}}[Y_i(1)], \quad \mathbb{E} \left[ \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)} \right] = \mathbb{E}_{\text{sp}}[Y_i(0)]$$

用这两个经验期望之差来估计平均因果作用  $\hat{\tau}$  :

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \left( \frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)} \right) = \frac{1}{N} \sum_{i=1}^N \left( \frac{(W_i - e(X_i)) \cdot Y_i^{\text{obs}}}{e(X_i) \cdot (1-e(X_i))} \right)$$

- 实际中常用估计的倾向得分来代替真实的倾向得分

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \left( \frac{W_i \cdot Y_i^{\text{obs}}}{\hat{e}(X_i)} - \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-\hat{e}(X_i)} \right) = \frac{1}{N} \sum_{i=1}^N \left( \frac{(W_i - \hat{e}(X_i)) \cdot Y_i^{\text{obs}}}{\hat{e}(X_i) \cdot (1-\hat{e}(X_i))} \right)$$

- 估计的倾向得分  $\hat{e}(X)$  往往比真实的倾向得分 $e(X)$  更好,原因是估计的倾向得分在协变量  $\mathbf{X}$ 上略有过度拟合, 产生比随机实验更好的均衡效果
- 估计的倾向得分可能比真实的倾向得分有更好的性质, 具体体现在均方误差会更小

## 标准化的倾向得分

- 对权重标准化可以改进估计量的 (均方误差) 性质。

1. 加权估计与回归的联系:加权估计量可通过加权回归得到

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i$$

其中, 每个个体的权重为

$$\hat{\lambda}_i = \begin{cases} \frac{1}{1-\hat{e}(X_i)} & \text{如果 } W_i = 0 \\ \frac{1}{\hat{e}(X_i)} & \text{如果 } W_i = 1 \end{cases}$$

$W_i = 0$  表示个体处于对照组,  $W_i = 1$  表示个体处于处理组

- 该模型不代表一个科学问题, 它是人为设置的一个工作模型
- 不管这个线性模型是否设定正确, 由这个模型来估的  $\hat{\tau}$  都是相合估计

## 2. 在回归中纳入协变量:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i \beta + \varepsilon_i$$

权重同上

- 纳入协变量的方程拥有一个好的性质: 得到的估计量是双稳健的, 即只要倾向得分模型或回归模型正确, 因果作用的估计量就是相合的
- 倾向得分模型正确指的是  $e(X)$  是一个 logistic 回归, 用 logistic 回归来估计会得到倾向得分的相合估计
- 若倾向得分模型并不符合 logistic 模型, 错误的使用 logistic 回归来估,  $\hat{e}(X)$  不是  $e(X)$  的相合估计, 但真实的数据满足这个线性方程, 那么也能保证得到的  $\hat{\tau}$  相合
- 相合: 当样本量趋无穷时,  $\hat{\tau}$  估计的因果作用离真实的因果作用很接近

## 子分类估计

### 估计的倾向得分的检验

#### 1. 简介

- 真实情况下往往需要用估计的倾向得分来代替真实的倾向得分, 估计的倾向得分是否拥有类似的均衡性质有待检验:
  - 首先, 划分几个层  $0 = b_0 < b_1 < \dots < b_j = 1$ , 定义层指标

$$B_i(j) = \begin{cases} 1 & \text{如果 } b_{j-1} \leq \hat{e}(X_i) < b_j \\ 0 & \text{否则} \end{cases}$$

- 然后, 检验  $W_i \perp X_i \mid B_i(1), \dots, B_i(j)$ , 即检验划分层后是否依旧有均衡性质
- 为保证倾向得分在处理组与对照组之间具有可比性 (即在两组中有重叠), 取倾向得分的上界与下界

$$\underline{e}_t = \min_{i: W_i=1} \hat{e}(X_i), \quad \bar{e}_c = \max_{i: W_i=0} \hat{e}(X_i)$$

$\underline{e}_t$  是处理组中倾向得分的下界,  $\bar{e}_c$  是对照组倾向得分的上界

- 最后采取逐步检验的形式，先看只有一个层够不够，若只有一个层能实现均衡性则不用对倾向得分进行细分，若一层不够则需对倾向得分进行分层
- 检验取线性化的倾向得分

$$\hat{\ell}(x) = \log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$$

意义：衡量个体处于处理组和处于对照组的odds ratio的对数

## 2. 构造倾向得分分层

$$N_c(j) = \sum_{i=1}^N (1 - W_i) \cdot B_i(j)$$

$$N_t(j) = \sum_{i=1}^N W_i \cdot B_i(j)$$

$$\bar{\ell}_c(j) = \frac{1}{N_c(j)} \sum_{i=1}^N (1 - W_i) \cdot B_i(j) \cdot \hat{\ell}(X_i)$$

$$\bar{\ell}_t(j) = \frac{1}{N_t(j)} \sum_{i=1}^N W_i \cdot B_i(j) \cdot \hat{\ell}(X_i)$$

$N_c(j)$  表示第j层内对照组的个体数， $N_t(j)$  表示第j层内处理组的个体数， $\bar{\ell}_c(j)$  表示第j层中对照组的平均线性化倾向得分， $\bar{\ell}_t(j)$  表示第j层中处理组的平均线性化倾向得分

- 用t统计量检验先  $\bar{\ell}_c(j)$  和  $\bar{\ell}_t(j)$  是否接近
- 若t统计量超过所设的阈值，则说明在这一层内倾向得分波动太大，需要再细分，若t统计量小于所设的阈值，则说明在这一层内倾向得分趋于稳定，不需再分，可直接检验均衡性
- 线性化倾向得分优势：线性化的倾向得分的表现更对称，更接近正态分布
- 若一层内倾向得分波动较大，则需对该层进一步细分：在  $N_c(j) + N_t(j)$  个个体的倾向得分的中位数处分为两层
- 新两层需满足的条件：
  - 每一新层内的对照组个体数、处理组个体数都不小于 3
  - 每一新层内的对照组个体数、处理组个体数都不小于  $\boxtimes+2$ （其中  $\boxtimes$  是协变量的数量）

## 3. 例子：如何根据估计的倾向得分来构造样本的分层

步	层	下界	上界	宽度	对照数	处理数	t统计量
1	1	0.00	0.94	0.94	4462	742	36.3
2	1	0.00	0.06	0.06	2540	61	3.2
	2	0.06	0.94	0.88	1922	681	23.7
3	1	0.00	0.02	0.01	1280	20	2.2
	2	0.02	0.06	0.05	1260	41	0.5
	3	0.06	0.20	0.14	1163	138	3.9
	4	0.20	0.94	0.74	759	543	10.9

第一步中，该层的t统计量为36.3，说明在完整样本中倾向得分的波动大，因此要分层；第二步，在中位数处分为两层，这两层的t统计量都超过了阈值，因此，再分别对这两层进行细分为四层；第三步，第二层的t值已小于1，则只需对第一、三、四层再进行细分

步	层	下界	上界	宽度	对照数	处理数	t统计量
4	1	0.00	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.1	0.05	604	46	0.5
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.94	0.57	301	351	5.6

第四步，得到七层，在这七层中只有第二、六、七层需要进一步细分，但由于第二层、第六层的样本量不够，不能继续细分，因此只将第七层做进一步细分

步	层	下界	上界	宽度	对照数	处理数	t统计量
5	1	0.01	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	-0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.50	0.13	181	144	2.5
	8	0.50	0.94	0.44	120	207	2.3

第五步，将第七、八层继续进行细分；最终得到10层

步	层	下界	上界	宽度	对照数	处理数	t统计量
6	1	0.01	0.01	0.00	644	6	-0.0
	2	0.01	0.02	0.01	636	14	1.7
	3	0.02	0.06	0.05	1260	41	0.5
	4	0.06	0.11	0.05	604	46	-0.3
	5	0.11	0.20	0.09	559	92	1.0
	6	0.20	0.37	0.17	458	192	1.2
	7	0.37	0.42	0.05	101	61	0.3
	8	0.42	0.50	0.08	80	83	0.7
	9	0.50	0.61	0.11	73	90	0.8
	10	0.61	0.94	0.34	47	117	-0.3

#### 4. 检验倾向得分的整体均衡性

- 需检验命题：给定已分层指标后，协变量是否独立于处理组

$$W_i \perp X_i \mid B_i(1), \dots, B_i(J)$$

- 对第k个变量分别考虑其在对照组和处理组每一层中的平均：

$$\bar{X}_{c,k}(j) = \frac{1}{N_c(j)} \sum_{i:W_i=0} B_i(j) \cdot X_{ik}$$

$$\bar{X}_{t,k}(j) = \frac{1}{N_t(j)} \sum_{i:W_i=1} B_i(j) \cdot X_{ik}$$

- 把  $X_k$  当做结局，若结局无因果作用，则处理组和对照组中X的均值近似相等
- 记第k个协变量在第j层的平均因果作用为  $\hat{\tau}_k(j)$

$$\hat{\tau}_k(j) = \bar{X}_{t,k}(j) - \bar{X}_{c,k}(j)$$

- 用奈曼或费希尔方法检验每层的  $\tau$  的加权平均是否等于0

$$\hat{\tau}_k = \sum_{j=1}^J \frac{N_c(j)+N_t(j)}{N} \cdot \hat{\tau}_k(j)$$

#### 5. 检验倾向得分在所有层内的均衡性

- 在整体上均衡时可能在某些层内不均衡
- 检验方法：方差分析

##### a. 饱和模型

$$SSR_k^{\text{ur}} = \sum_{i=1}^N \left( X_{ik} - \sum_{j=1}^J \left( \bar{X}_{c,k} W_i + \bar{X}_{t,k} (1 - W_i) \right) \cdot B_i(j) \right)^2$$

允许在每一层内，对照组和处理组都有一个不同的X的系数

## b. 简化模型

$$SSR_k^r = \sum_{i=1}^N \left( X_{ik} - \sum_{j=1}^J \bar{X}_k(j) \cdot B_i(j) \right)^2$$

在每一层内  $\bar{X}_{c,k}$  和  $\bar{X}_{t,k}$  都相等

- 检验  $t_k(j) = 0$  :

$$F = \frac{(SSR_k^r - SSR_k^{ur})/J}{SSR_k^{ur}/(N-2J)}$$

## 截断

- 倾向得分接近0，个体可能更多来自对照组，对照者很难找到与之匹配的处理者；倾向得分接近1，个体可能更多来自处理组，很难找到与之匹配的对照者
- 截断定义：丢弃倾向得分接近0或1的个体
- 注意：改变了待估量，因为其删除了一些样本，使得协变量分布改变
- 缺点：牺牲了外部有效性，它是有偏估计，不能代表原始样本中的平均因果作用
- 优点：提高了内部有效性，将总体限制成截断后的样本，估计会变得更精确、可信

## 实例分析：子分类估计

- 基本思想：先按照倾向得分估计对样本进行分层，在每一层内估计平均因果作用，之后再加权平均组合起来
- 具体例子：彩票数据

	低	中	高	总计
	$\hat{e}(X_i) < 0.0891$	$0.0891 \leq \hat{e}(X_i) \leq 0.9109$	$\hat{e}(X_i) > 0.9109$	
中小奖	82	172	5	259
中大奖	4	151	82	237
总计	86	323	87	496

根据截断准则得到阈值，将倾向得分分为低、中、高，将过低和过高的样本扔掉，只保留中间323个样本做分析

子类	最小倾向得分	最大倾向得分	对照数	处理数	t统计量
1	0.03	0.24	67	13	-0.1
2	0.24	0.32	32	8	0.9
3	0.32	0.44	24	17	1.7
4	0.44	0.69	34	47	2.0
5	0.69	0.99	15	66	1.6

用截断后的样本重新拟合倾向得分模型，最终分为5层，3、4、5层因样本数问题不能再细分（再细分协变量数量会大于样本数）；在每一层内估计平均因果作用，再按各层样本数加权平均，即得到总体的平均因果作用估计

协变量	全部样本		截断的样本		截断的样本，分5层	
	估计值	s.e.	估计值	s.e.	估计值	s.e.
无	-6.2	(1.4)	-6.6	(1.7)	-5.7	(2.0)
部分	-2.8	(0.9)	-4.0	(1.2)	-5.1	(1.2)
所有	-5.1	(1.0)	-5.3	(1.1)	-5.7	(1.1)

全样本、截断样本、分5层的截断样本（子分类估计）的彩票结果分析：全样本中，纳入协变量会严重影响因果作用的估计值，估计值对协变量很敏感，不稳健；用截断后样本，与用全样本相比，其因果作用较集中；分5层的截断样本（即使用子分类估计），因果作用很接近

### 子分类估计的奈曼推断

- 第j层内平均作用差异估计  $\hat{\tau}^{dif}(j)$  :处理组的平均结局-对照组的平均结局
- 子分类的平均因果作用估计为

$$\hat{\tau}^{strat} = \sum_{j=1}^J q(j)\hat{\tau}^{dif}(j)$$

其中，权重  $q(j) = N(j)/N$

- 奈曼推断能给出一个保守的方差
- 每一层内对照组结局方差

$$s_c(j)^2 = \frac{1}{N_c(j)-1} \sum_{i=1}^N (1 - W_i) \cdot B_i(j) \cdot \left(Y_i^{obs} - \overline{Y}_c^{obs}(j)\right)^2$$

- 每一层内处理组结局方差

$$s_t(j)^2 = \frac{1}{N_t(j)-1} \sum_{i=1}^N W_i \cdot B_i(j) \cdot \left(Y_i^{obs} - \overline{Y}_t^{obs}(j)\right)^2$$

- 根据奈曼准则，计算  $\hat{\tau}^{dif}(j)$  的方差

$$\begin{aligned} \widehat{V}\left(\hat{\tau}^{dif}(j)\right) &= \frac{s_c(j)^2}{N_c(j)} + \frac{s_t(j)^2}{N_t(j)} = \\ &\frac{1}{N_c(j) \cdot (N_c(j)-1)} \sum_{i:W_i=0} B_i(j) \cdot \left(Y_i^{obs} - \overline{Y}_c^{obs}(j)\right)^2 + \\ &\frac{1}{N_t(j) \cdot (N_t(j)-1)} \sum_{i:W_i=1} B_i(j) \cdot \left(Y_i^{obs} - \overline{Y}_t^{obs}(j)\right)^2 \end{aligned}$$

- 得到保守估计方差之后，可以构造保守置信区间用于待研究的科学问题的推断



## 子分类估计降低偏差

### 1. 偏差来源

假设潜在结果模型满足这样的线性关系：

$$\mathbb{E}_{\text{sp}}[Y_i(w) | X_i = x] = \alpha + \tau_{\text{sp}} \cdot w + \beta'x$$

将整个样本的  $\bar{Y}_1 - \bar{Y}_0$  得到的估计记为  $\hat{\tau}^{\text{dif}}$ ，它会有一个bias, 这个bias是协变量不平衡所造成的偏差, 若对照组与处理组的协变量是不平衡的，那么估计的因果作用是有偏差的，该偏差为

$$\mathbb{E}[\hat{\tau}^{\text{dif}} - \tau_{\text{fs}} | X, W] = (\bar{X}_t - \bar{X}_c)\beta$$

- 子分类估计量按倾向得分分层，由于倾向得分均衡性质，在每一层内  $\bar{X}_t(j)$  与  $\bar{X}_c(j)$  都比较接近，所以偏差项接近于0，然后对总体加权，然后偏差更接近于0，所以子分类估计降低了偏差

### 2. 进一步降低偏差

- 在每一层内分别拟合一个线性模型：

$$Y_i^{\text{obs}} = \alpha(j) + \tau(j) \cdot W_i + X_i\beta(j)$$

注意：不同层内的 $\beta$ 不同

- 得到协变量调整估计

$$\hat{\tau}^{\text{strat,adj}} = \sum_{j=1}^J \hat{\tau}^{\text{adj}}(j) \cdot q(j)$$

按层的比例做加权

- 若感兴趣的是处理组上的平均因果作用，则

$$\hat{\tau}_t^{\text{strat,adj}} = \sum_{j=1}^J \hat{\tau}^{\text{adj}}(j) \cdot \frac{N_t(j)}{N_t}$$

即权重相应变成处理组在每层的比例

## 子分类估计与加权估计的联系

- 联系：子分类估计也可以看成是加权估计，权重为

$$\hat{\lambda}_i = \begin{cases} \sum_{j=1}^J B_i(j) \cdot \frac{N(j)}{N_c(j)} & \text{如果 } W_i = 0 \\ \sum_{j=1}^J B_i(j) \cdot \frac{N(j)}{N_t(j)} & \text{如果 } W_i = 1 \end{cases}$$

- 子分类估计使用了“粗化”的倾向得分，认为倾向得分在每一层内是常数，即

$$\tilde{e}(X_i) = \sum_{j=1}^J B_i(j) \cdot \frac{N_t(j)}{N(j)}$$

- 子分类估计是对“粗化”的倾向得分做加权估计

## 逆概率加权估计与子分类估计

- 当层数足够多，每一层内的倾向得分波动足够小，逆概率加权估计与子分类估计是相近的
- 推荐使用子分类估计的3个理由
  - a. 如果倾向得分估计错误，逆概率加权估计受到的影响更大；估计量的倒数可能不稳定。子分类估计不依赖于个体的倾向得分估计，受到倾向得分估计错误影响小
  - b. 子分类估计的方差更小，因为倾向得分向层内均值集中了，因此子分类估计更稳健
  - c. 关于纳入协变量调整，子分类估计允许局部（层内分别）调整，而逆概率加权只能进行全局调整
- 实例：逆概率加权估计与子分类估计的结果比较

	全部样本		截断的样本	
	逆概率加权	子分类	逆概率加权	子分类
偏差	4.34	2.68	1.29	0.30
方差	2.59 <sup>2</sup>	0.83 <sup>2</sup>	1.29 <sup>2</sup>	1.15 <sup>2</sup>
均方误差	5.06 <sup>2</sup>	2.81 <sup>2</sup>	1.83 <sup>2</sup>	1.19 <sup>2</sup>

- 子分类估计较逆概率加权估计有更小的偏差和方差
- 基于截断样本的估计的偏差和方差都更小

## 匹配

- 适用情况：适用于少量的处理组个体但是有大量的对照组个体的情况
- 思想：对一个处理组个体，从对照组中找一个个体，让它们的协变量匹配（完全一致或相近）
- 例子：“他找到工作是因为他参加了就业培训项目。”
  - 结局：他找到了工作
  - 原因：他参加了就业培训项目
  - 要判断其是否正确需找一个未参加过就业培训项目的人，且一协变量与这个人相同（如同样是30岁未婚上过大学），若该对照组未找到工作，则该命题成立

## 匹配的分类

### 1. 根据匹配的精度划分

- 精确匹配：协变量完全一致的匹配
- 非精确匹配：允许匹配的时候存在一定的差异

### 2. 根据配对的选取方式划分

- 无放回：对照组的个体只能用一次

- 有放回：对照组的个体可以多次使用

### 3. 无放回的精确保配

- 处理组中取一个体*i*，设其协变量为  $X_i = x$  ,然后在对照组中找一个体  $i_m$  ,协变量为  $X_{mi} = x$  ,若能找到这样的对照者，则产生了“配对随机化”
- 对于配对，配对内的因果作用的估计：

$$\hat{\tau}_i = Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}}$$

- 处理组的平均因果作用：

$$\hat{\tau}_t = \frac{1}{N_t} \sum_{i:W_i=1} \hat{\tau}_i = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}})$$

- 可供匹配的对照个体较少，匹配过程会发生冲突（如给第一个处理者匹配了对照者，导致第二个处理者无人可匹配），导致精确保配在实际中几乎不可能
- 解决方案：
  - a. 允许对照者被重复匹配，即有放回的精确保配
  - b. 仍然使用无放回的精确保配，按事先指定的顺序匹配
  - c. 非精确保配

### 4. 无放回的非精确保配

选择个体对照者思路：最小化协变量的距离

$$m_i = \arg \min_{i' \in \mathbb{I}_c} \|X_i - X_{i'}\|$$

全局最优匹配：

$$\arg \min_{m_1, \dots, m_{N_t} \in \mathbb{I}_c} \sum_{i=1}^{N_t} \|X_i - X_{m_i}\|$$

- 局部最优不代表全局最优
- 准则：先匹配难匹配的，再匹配好匹配的

### 5. 距离测度

- 非精确保配涉及距离测度，需定义什么算近，什么算远
- 定义：

$$d_V(x, x') = (x'V^{-1}x)^{1/2}$$

- V的选取
  - a. 马氏距离

$$V_M = \frac{1}{2} \cdot \left( \frac{1}{N_c} \sum_{i:W_i=0} (X_i - \bar{X}_c)^{\otimes 2} + \frac{1}{N_t} \sum_{i:W_i=1} (X_i - \bar{X}_t)^{\otimes 2} \right)$$

b. 欧氏距离

$$V_E = \text{diag}(V_M)$$

## 6. 匹配的偏差

- 协变量匹配不完全一致时会产生偏差：

$$B_i = \mathbb{E}_{\text{sp}}[Y(0) \mid X_i = x] - \mathbb{E}_{\text{sp}}[Y(0) \mid X_i = X_{m_i}]$$

- 偏差可用线性模型校正，设对照组中模型满足

$$\mathbb{E}_{\text{sp}}[Y_i(0) \mid X = x] = \alpha + x\beta_c$$

偏差为

$$B_i = (X_i - X_{m_i})\beta_c$$

将偏差代入因果估计中，消除偏差影响

$$\hat{Y}_i(0) = Y_{m_i} + (X_i - X_{m_i})\hat{\beta}_c$$

得到调整的后果作用估计

$$\hat{\tau}_i^{\text{adj}} = Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}} - (X_i - X_{m_i})\hat{\beta}_c$$

## 7. 有放回匹配

- 选取对照者思路：最小化协变量的距离
- 得到因果作用估计：

$$\hat{\tau}_t = \frac{1}{N_t} \sum_{i:W_i=1} (Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}})$$

- 因对照者有重复匹配的可能性，所以记  $L(i)$  为对照组中个体*i*被使用的次数，有

$$L(i) = \sum_{j=1}^{N_t} 1_{i \in \mathcal{M}_j}, \quad i \in \mathbb{I}_c$$

其中，满足  $\sum_i L(i) = N_t$

## 8. 匹配的数量是否会影响因果作用的估计

- 给每个处理者都指定M个匹配，得ATT估计：

$$\hat{\tau}^M = \frac{1}{N_t} \sum_{i=1}^{N_t} \left( Y_i(1) - \frac{1}{M} \sum_{j \in \mathcal{M}_i} Y_j(0) \right)$$

也可写成

$$\hat{\tau}^M = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i(1) - \hat{Y}_i(0)) = \frac{1}{N} \sum_{i=1}^N \left( W_i - \frac{L(i)}{M} \right) \cdot Y_i^{\text{obs}}$$

- 该估计量方差为

$$\mathbb{V}(\hat{\tau}^M) = \frac{1}{N_t} \left( \sigma_t^2 + \frac{\sigma_c^2}{M} \right)$$

- 使用再多的匹配最多只能降低一半的方差，因为有一半方差来自处理者，一般配2-3个足够，匹配更多的个体对方差作用不大

## 全样本上的匹配估计量

考虑带协变量调整、有放回的匹配估计量

- 处理者上的匹配估计量

$$\hat{\tau}_t^{\text{adj}} = \frac{1}{N_t} \sum_{i: W_i=1} (Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}} - (X_i - X_{m_i}) \hat{\beta}_c)$$

- 对照者上的匹配估计量

$$\hat{\tau}_c^{\text{adj}} = \frac{1}{N_c} \sum_{i: W_i=0} (Y_i^{\text{obs}} - Y_{m_i}^{\text{obs}} - (X_i - X_{m_i}) \hat{\beta}_t)$$

- 整个样本上平均因果作用的调整匹配估计

$$\hat{\tau}^{\text{adj}} = \frac{N_c}{N_t + N_c} \cdot \hat{\tau}_c^{\text{adj}} + \frac{N_t}{N_t + N_c} \cdot \hat{\tau}_t^{\text{adj}}$$

## 匹配的实例：用该实例说明匹配的方法及几个变种

### 1. 简略介绍

- 处理组：21 位接触了铬元素和镍元素的铁路电焊工
- 对照组：26 位未接触铬元素和镍元素的对照人群
- 三个协变量：年龄、种族和当前吸烟行为
- 响应变量：DNA 蛋白质交联的测量
- 数据展示

电焊工组					对照组				
编号	年龄	种族	吸烟者	DPC	编号	年龄	种族	吸烟者	DPC
1	38	C	N	1.77	1	48	AA	N	1.08
2	44	C	N	1.02	2	63	C	N	1.09
3	39	C	Y	1.44	3	44	C	Y	1.10
4	33	AA	Y	0.65	4	40	C	N	1.10
5	35	C	Y	2.08	5	50	C	N	0.93
6	39	C	Y	0.61	6	52	C	N	1.11
7	27	C	N	2.86	7	56	C	N	0.98
8	43	C	Y	4.19	8	47	C	N	2.20
9	39	C	Y	4.88	9	38	C	N	0.88
10	43	AA	N	1.08	10	34	C	N	1.55
11	41	C	Y	2.03	11	42	C	N	0.55
12	36	C	N	2.81	12	36	C	Y	1.04
13	35	C	N	0.94	13	41	C	N	1.66
14	37	C	N	1.43	14	41	AA	Y	1.49
15	39	C	Y	1.25	15	31	AA	Y	1.36
16	34	C	N	2.97	16	56	AA	Y	1.02
17	35	C	Y	1.01	17	51	AA	N	0.99
18	53	C	N	2.07	18	36	C	Y	0.65
19	38	C	Y	1.15	19	44	C	N	0.42
20	37	C	N	1.07	20	35	C	N	2.33
21	38	C	Y	1.63	21	34	C	Y	0.97
					22	39	C	Y	0.62
					23	45	C	N	1.02
					24	42	C	N	1.78
					25	30	C	N	0.95
					26	35	C	Y	1.59
均值		AA	吸烟者		均值		AA	吸烟者	
38		10%	52%		43		19%	35%	

- 一开始协变量不均衡，对照组比电焊工组的年龄稍大

## 2. 基于倾向得分的匹配

- 倾向得分是对协变量的一个降维
- 匹配倾向得分会均衡协变量
- 注意：对照组中 21 个最大的倾向得分估计值  $\hat{e}(x_i)$  的均值为 0.46 略低于处理组倾向得分估计值  $\hat{e}(x_i)$  的均值的 0.51，因此配对匹配无法完全消除这种差距。年龄、倾向得分不均衡，精确匹配不可行
- 只能采取采用非精确匹配
- 取两个个体距离是倾向得分估计值  $\hat{e}(x_i)$  的差值的平方,得距离矩阵

电焊工	对照者1	对照者2	对照者3	对照者4	对照者5	对照者6
1	0.10	0.13	0.00	0.00	0.05	0.06
2	0.04	0.06	0.02	0.01	0.01	0.02
3	0.19	0.23	0.01	0.02	0.12	0.14
4	0.13	0.18	0.00	0.01	0.08	0.09
5	0.26	0.32	0.03	0.06	0.18	0.20
6	0.19	0.23	0.01	0.02	0.12	0.14
7	0.29	0.35	0.05	0.07	0.20	0.23
8	0.12	0.16	0.00	0.01	0.07	0.08
9	0.19	0.23	0.01	0.02	0.12	0.14
10	0.00	0.01	0.07	0.05	0.00	0.00
11	0.15	0.19	0.00	0.01	0.09	0.11
12	0.13	0.17	0.00	0.01	0.07	0.09
13	0.14	0.19	0.00	0.01	0.08	0.10
14	0.11	0.15	0.00	0.00	0.06	0.08
15	0.19	0.23	0.01	0.02	0.12	0.14
16	0.16	0.20	0.01	0.02	0.10	0.11
17	0.26	0.32	0.03	0.06	0.18	0.20
18	0.00	0.01	0.08	0.05	0.00	0.00
19	0.20	0.25	0.02	0.03	0.13	0.15
20	0.11	0.15	0.00	0.00	0.06	0.08
21	0.20	0.25	0.02	0.03	0.13	0.15

- 该矩阵看起来不直观，因此需做改进

## 3. 对距离矩阵的改进：卡尺

- 规则：引入宽度为  $\delta$  的卡尺，如果两位个体的倾向得分差异大于  $\delta$ ，则这个距离设置为  $\infty$ ，如果两位个体的倾向得分差异小于  $\delta$ ，则这个距离就是对  $x_k$  和  $x_l$  接近度的度量（距离一般采用马氏距离来测算）
- 卡尺的选择：经常取值为  $\varepsilon \delta(\delta)$  的标准差或标准差的倍数（50% 或 20%），该例选择50%（因样本量较小）
- 使用卡尺后的距离矩阵

电焊工	对照者1	对照者2	对照者3	对照者4	对照者5	对照者6
1	$\infty$	$\infty$	6.15	0.08	$\infty$	$\infty$
2	$\infty$	$\infty$	$\infty$	0.33	$\infty$	$\infty$
3	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
4	$\infty$	$\infty$	12.29	$\infty$	$\infty$	$\infty$
5	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
6	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
7	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
8	$\infty$	$\infty$	0.02	5.09	$\infty$	$\infty$
9	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
10	0.51	$\infty$	$\infty$	$\infty$	10.20	11.17
11	$\infty$	$\infty$	0.18	$\infty$	$\infty$	$\infty$
12	$\infty$	$\infty$	7.06	0.33	$\infty$	$\infty$
13	$\infty$	$\infty$	7.57	$\infty$	$\infty$	$\infty$
14	$\infty$	$\infty$	6.58	0.18	$\infty$	$\infty$
15	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
16	$\infty$	$\infty$	8.13	$\infty$	$\infty$	$\infty$
17	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
18	9.41	$\infty$	$\infty$	$\infty$	0.18	0.02
19	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$
20	$\infty$	$\infty$	6.58	0.18	$\infty$	$\infty$
21	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$	$\infty$

4. 最优配对匹配

- 定义：最优配对匹配是把每一位处理受试者与以为不同的对照者进行匹配，使得匹配对内的总距离最小化
- 最佳优先算法或贪心算法通常不会找到最优配对匹配
- 统计软件 R 的 optmatch 包中的配对匹配函数 pairmatch 可实现分配问题的解决
- 用最优匹配算法得到的结果：年龄、种族差距变小，很多个体的协变量没有配对成功

电焊工组					匹配对照组				
配对	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$	
1	38	C	N	0.46	45	C	N	0.32	
2	44	C	N	0.34	47	C	N	0.28	
3	39	C	Y	0.57	39	C	Y	0.57	
4	33	AA	Y	0.51	41	C	N	0.40	✗
5	35	C	Y	0.65	34	C	Y	0.67	
6	39	C	Y	0.57	31	AA	Y	0.55	✗
7	27	C	N	0.68	35	C	Y	0.65	✗
8	43	C	Y	0.49	41	AA	Y	0.35	✗
9	39	C	Y	0.57	34	C	N	0.54	✗
10	43	AA	N	0.20	50	C	N	0.23	✗
11	41	C	Y	0.53	44	C	Y	0.47	
12	36	C	N	0.50	42	C	N	0.38	
13	35	C	N	0.52	40	C	N	0.42	
14	37	C	N	0.48	44	C	N	0.34	
15	39	C	Y	0.57	35	C	N	0.52	✗
16	34	C	N	0.54	38	C	N	0.46	
17	35	C	Y	0.65	36	C	Y	0.64	
18	53	C	N	0.19	52	C	N	0.20	
19	38	C	Y	0.60	36	C	Y	0.64	
20	37	C	N	0.48	42	C	N	0.38	
21	38	C	Y	0.60	30	C	N	0.63	✗
	均值	%AA	%Y	均值	均值	%AA	%Y	均值	
	38	10	52	0.51	40	10	38	0.46	

- 改进：改进距离矩阵，用罚函数取代无穷大的惩罚，试图通过满足卡尺宽度来避免惩罚项，所有的处理者和对照者都有可能匹配，匹配的概率受惩罚项约束，带惩罚项的匹配，它牺牲了稍微超出卡尺宽度的几个匹配对，但允许配对者之间的竞争
- 用罚函数取代无穷大的惩罚的匹配结果：在边缘均值上差不多，但在个体层面上完全匹配的个体数增加

电焊工组					匹配对照组				
配对	年龄	种族	吸烟者	$\hat{\theta}(x)$	年龄	种族	吸烟者	$\hat{\theta}(x)$	
1	38	C	N	0.46	44	C	N	0.34	
2	44	C	N	0.34	47	C	N	0.28	
3	39	C	Y	0.57	36	C	Y	0.64	
4	33	AA	Y	0.51	41	AA	Y	0.35	
5	35	C	Y	0.65	35	C	Y	0.65	
6	39	C	Y	0.57	39	C	Y	0.57	
7	27	C	N	0.68	30	C	N	0.63	
8	43	C	Y	0.49	45	C	N	0.32	×
9	39	C	Y	0.57	36	C	Y	0.64	
10	43	AA	N	0.20	48	AA	N	0.14	
11	41	C	Y	0.53	44	C	Y	0.47	
12	36	C	N	0.50	41	C	N	0.40	
13	35	C	N	0.52	40	C	N	0.42	
14	37	C	N	0.48	42	C	N	0.38	
15	39	C	Y	0.57	35	C	N	0.52	
16	34	C	N	0.54	38	C	N	0.46	
17	35	C	Y	0.65	34	C	Y	0.67	
18	53	C	N	0.19	52	C	N	0.20	
19	38	C	Y	0.60	34	C	N	0.54	×
20	37	C	N	0.48	42	C	N	0.38	
21	38	C	Y	0.60	31	AA	Y	0.55	×
	均值	%AA	%Y	均值	均值	%AA	%Y	均值	
	38	10	52	0.51	40	14	38	0.45	

## 5. 多重对照最优匹配

- 在多重对照匹配中，每一位处理受试者至少与一位但可能不止一位对照者相匹配，使得匹配集中处理受试者和对照者之间的总距离最小化
- 分类
  - a. 固定比率匹配
    - 定义：就是给每位处理受试者匹配相同数量的对照者，如1:2、1:3
    - 优点：固定比率匹配的主要优点是，可以按通常的方式从处理组和对照组中计算汇总统计量，而不用基于观察值不等权重进行直接校正。当潜在对照者数量充足且必须消除的偏差不大时，这种优势将非常重要（降低方差的同时理论性质不难）
  - b. 数量可变匹配
    - 定义：一个受试者匹配不同数量的对照者，其优化算法不仅决定谁与谁匹配，还决定给每位处理受试者分配多少数量的对照者。需要规则对选择进行约束
    - 优点：
      - 匹配集将更紧密地匹配。相比于固定比率匹配而言，匹配集内的总距离通常相当小，如果试图均匀地分配对照者，则会在倾向得分上产生更大的错误匹配
      - 不要求对照者人数时处理受试者人数的整数倍
      - 配对匹配可以达到的效果具有明确的限制，而多重匹配也存在这些明确的限制，但这些限制在可变比率匹配时表现得效果更好（经验上的论证）
- 可变匹配的结果



电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$	
1	38	C	N	0.46	44	C	N	0.34	
2	44	C	N	0.34	47	C	N	0.28	
3	39	C	Y	0.57	36	C	Y	0.64	
4	33	AA	Y	0.51	41	AA	Y	0.35	
5	35	C	Y	0.65	35	C	Y	0.65	
6	39	C	Y	0.57	36	C	Y	0.64	
7	27	C	N	0.68	30	C	N	0.63	
8	43	C	Y	0.49	45	C	N	0.32	
9	39	C	Y	0.57	35	C	N	0.52	
10	43	AA	N	0.20	51	AA	N	0.12	
					56	AA	Y	0.13	
					48	AA	N	0.14	
					44	C	Y	0.47	
11	41	C	Y	0.53	41	C	N	0.40	
12	36	C	N	0.50	40	C	N	0.42	
13	35	C	N	0.52	42	C	N	0.38	
14	37	C	N	0.48	39	C	Y	0.57	
15	39	C	Y	0.57	38	C	N	0.46	
16	34	C	N	0.54	34	C	Y	0.67	
17	35	C	Y	0.65	63	C	N	0.09	
18	53	C	N	0.19	56	C	N	0.15	
					52	C	N	0.20	
					50	C	N	0.23	
					34	C	N	0.54	
19	38	C	Y	0.60	42	C	N	0.38	
20	37	C	N	0.48	31	AA	Y	0.55	
21	38	C	Y	0.60					
均值					均值	%AA	%Y	均值	
38					40	14	40	0.45	

- 最优完全匹配的结果：

电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$	
1	38	C	N	0.46	40	C	N	0.42	
2	44	C	N	0.34	47	C	N	0.28	
2					45	C	N	0.32	
2					44	C	N	0.34	
2					41	AA	Y	0.35	
2					42	C	N	0.38	
2					42	C	N	0.38	
2					41	C	N	0.40	
3	41	C	Y	0.53	39	C	Y	0.57	
3	39	C	Y	0.57					
3	39	C	Y	0.57					
3	39	C	Y	0.57					
3	39	C	Y	0.57					
3	38	C	Y	0.60					
3	38	C	Y	0.60					
4	33	AA	Y	0.51	31	AA	Y	0.55	
5	35	C	Y	0.65	35	C	Y	0.65	
5					34	C	Y	0.67	

电焊工组					匹配对照组							
匹配集	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$				
6	27	C	N	0.68	30	C	N	0.63				
7	43	C	Y	0.49	44	C	Y	0.47				
8	43	AA	N	0.20	51	AA	N	0.12				
8					56	AA	Y	0.13				
8					48	AA	N	0.14				
9	36	C	N	0.50	35	C	N	0.52				
9	35	C	N	0.52								
10	37	C	N	0.48	38	C	N	0.46				
10	37	C	N	0.48								
11	34	C	N	0.54	34	C	N	0.54				
12	35	C	Y	0.65	36	C	Y	0.64				
12					36	C	Y	0.64				
13					63	C	N	0.09				
13	53	C	N	0.19	56	C	N	0.15				
13					52	C	N	0.20				
13					50	C	N	0.23				
均值					均值	%AA	%Y	均值				
38					39	10	55	0.50				

其匹配度增高的同时由于匹配集的大小不相等会导致匹配效率低下的问题

- 匹配效率低下改进的两个变体：

a. 变体一：匹配集是配对或三元组，只使用 26 位对照者中的 21 位对照者

电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$	
1	38	C	N	0.46	42	C	N	0.38	
1					42	C	N	0.38	
2	44	C	N	0.34	45	C	N	0.32	
2					44	C	N	0.34	
3	39	C	Y	0.57	36	C	Y	0.64	
3	38	C	Y	0.60					
4	33	AA	Y	0.51	31	AA	Y	0.55	
5	35	C	Y	0.65	35	C	Y	0.65	
6	39	C	Y	0.57	39	C	Y	0.57	
6	39	C	Y	0.57					
7	27	C	N	0.68	30	C	N	0.63	
8	43	C	Y	0.49	44	C	Y	0.47	
8	41	C	Y	0.53					
9	43	AA	N	0.20	51	AA	N	0.12	
9					48	AA	N	0.14	

电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$	
10	36	C	N	0.50	35	C	N	0.52	
10	35	C	N	0.52					
11	37	C	N	0.48	41	C	N	0.40	
11					38	C	N	0.46	
12	39	C	Y	0.57	36	C	Y	0.64	
12	38	C	Y	0.60					
13	34	C	N	0.54	34	C	N	0.54	
14	35	C	Y	0.65	34	C	Y	0.67	
15	53	C	N	0.19	56	C	N	0.15	
15					52	C	N	0.20	
16	37	C	N	0.48	40	C	N	0.42	
均值					均值	%AA	%Y	均值	
38					39	10	52	0.50	

种族与吸烟都完全均衡了，结果较之前要好

b. 变体二：任何匹配集中最多使用两位电焊工或最多使用三位对照者，使用全部 26 位对照者

电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{\theta}(x)$	年龄	种族	吸烟者	$\hat{\theta}(x)$	
1					44	C	N	0.34	
1	38	C	N	0.46	41	AA	Y	0.35	
1					42	C	N	0.38	
2					50	C	N	0.23	
2	44	C	N	0.34	47	C	N	0.28	
2					45	C	N	0.32	
3	39	C	Y	0.57	36	C	Y	0.64	
3	39	C	Y	0.57					
4	33	AA	Y	0.51	31	AA	Y	0.55	
5	35	C	Y	0.65	34	C	Y	0.67	
6	27	C	N	0.68	30	C	N	0.63	
7	43	C	Y	0.49	44	C	Y	0.47	
7	41	C	Y	0.53					
8	39	C	Y	0.57	39	C	Y	0.57	
8	39	C	Y	0.57					
9					51	AA	N	0.12	
9	43	AA	N	0.20	56	AA	Y	0.13	
9					48	AA	N	0.14	

电焊工组					匹配对照组				
匹配集	年龄	种族	吸烟者	$\hat{\theta}(x)$	年龄	种族	吸烟者	$\hat{\theta}(x)$	
10	36	C	N	0.50					
10	35	C	N	0.52	35	C	N	0.52	
11					42	C	N	0.38	
11	37	C	N	0.48	38	C	N	0.46	
12	34	C	N	0.54	34	C	N	0.54	
13	35	C	Y	0.65	35	C	Y	0.65	
14					63	C	N	0.09	
14	53	C	N	0.19	56	C	N	0.15	
14					52	C	N	0.20	
15	38	C	Y	0.60	36	C	Y	0.64	
15	38	C	Y	0.60					
16					41	C	N	0.40	
16	37	C	N	0.48	40	C	N	0.42	
	均值	%AA	%Y	均值	均值	%AA	%Y	均值	
	38	10	52	0.51	39	11	56	0.50	

- 总结：在特定意义上，最优完全匹配是用于观察性研究的最佳设计；其思路接近分层，其基于协变量取值将受试者划分为不同的层，但要求每层必须至少包含一位处理受试者和至少一位对照者
- 实现：R 的 optmatch 包中的 fullmatch 函数

## 精细均衡

- 定义：精细均衡是对最优匹配的约束，强制对一个名义变量进行均衡
- 精细均衡以均衡协变量为目的，但它不需要精确地匹配这个名义变量
- 3个应用场景：
  - a. 有多个取值水平的名义变量，很难用倾向得分来平衡时
  - b. 罕见的二分类变量，很难用距离开控制均衡时
  - c. 几个名义变量相互作用，难用倾向得分估计时
- 在精细均衡约束下，可用配对约束其他协变量
- 精细均衡和距离矩阵可以突出不同的协变量
- 例子：针对电焊工案例中种族没有被精确的均衡问题，利用精细均衡均衡种族变量
  - 第一步：构造种族与处理变量的交叉列表

	AA	C
电焊工	2	19
潜在对照者	5	21

- 第二步：为了达到均衡，从每个种族取值类别中确定必须移除的对照者的数量：删除 3 个 AA、2 个 C
- 第三步：决定删除谁，构造一个新的距离矩阵，构造一个 26×26 的矩阵，添加 5 行虚拟行，为每位必须移除的对照者添加一行，设置其与自身相同类别受试者之间的距离为零，与其他类别

受试者之间的距离为无穷大

电焊工	对照者1	对照者2	对照者3	对照者4	对照者5	对照者6
1	∞	∞	6.15	0.08	∞	∞
2	∞	∞	∞	0.33	∞	∞
3	∞	∞	∞	∞	∞	∞
4	∞	∞	12.29	∞	∞	∞
5	∞	∞	∞	∞	∞	∞
6	∞	∞	∞	∞	∞	∞
7	∞	∞	∞	∞	∞	∞
8	∞	∞	0.02	5.09	∞	∞
9	∞	∞	∞	∞	∞	∞
10	0.51	∞	∞	∞	10.2	11.17
11	∞	∞	0.18	∞	∞	∞
12	∞	∞	7.06	0.33	∞	∞
13	∞	∞	7.57	∞	∞	∞
14	∞	∞	6.58	0.18	∞	∞
15	∞	∞	∞	∞	∞	∞
16	∞	∞	8.13	∞	∞	∞
17	∞	∞	∞	∞	∞	∞
18	9.41	∞	∞	∞	0.18	0.02
19	∞	∞	∞	∞	∞	∞
20	∞	∞	6.58	0.18	∞	∞
21	∞	∞	∞	∞	∞	∞
E1	∞	0	0	0	0	0
E2	∞	0	0	0	0	0
E3	0	∞	∞	∞	∞	∞
E4	0	∞	∞	∞	∞	∞
E5	0	∞	∞	∞	∞	∞

- 第四步：为这个新的平方距离矩阵找到一个最优匹配
- 第五步：丢弃额外行及其匹配的对照者
- 最终输出结果：

电焊工组					匹配对照组			
配对	年龄	种族	吸烟者	$\hat{e}(x)$	年龄	种族	吸烟者	$\hat{e}(x)$
1	38	C	N	0.46	44	C	N	0.34
2	44	C	N	0.34	47	C	N	0.28
3	39	C	Y	0.57	36	C	Y	0.64
4	33	AA	Y	0.51	41	AA	Y	0.35
5	35	C	Y	0.65	35	C	Y	0.65
6	39	C	Y	0.57	36	C	Y	0.64
7	27	C	N	0.68	30	C	N	0.63
8	43	C	Y	0.49	45	C	N	0.32
9	39	C	Y	0.57	39	C	Y	0.57
10	43	AA	N	0.20	50	C	N	0.23
11	41	C	Y	0.53	44	C	Y	0.47
12	36	C	N	0.50	41	C	N	0.40
13	35	C	N	0.52	40	C	N	0.42
14	37	C	N	0.48	42	C	N	0.38
15	39	C	Y	0.57	35	C	N	0.52
16	34	C	N	0.54	38	C	N	0.46
17	35	C	Y	0.65	34	C	Y	0.67
18	53	C	N	0.19	52	C	N	0.20
19	38	C	Y	0.60	34	C	N	0.54
20	37	C	N	0.48	42	C	N	0.38
21	38	C	Y	0.60	31	AA	Y	0.55
	均值 38	%AA 10	%Y 52	均值 0.51	均值 40	%AA 10	%Y 38	均值 0.46

- 种族变量达到精细均衡，整体中的比例均衡，但不完全匹配，电焊工组的10号与21号都是黑人与白人的配对
- 某些情况下精细均衡不可行，如例中的吸烟变量，对照组中的吸烟者数量较少，不管去掉多少都无法达到均衡

## 无组别匹配（也称非二部图匹配）

- 二部图：一个无向图当中，所有的点可以分成两个子集。这两个子集当中的点各自互不相交，并且图当中的所有边关联的顶点都属于两个不同的集合

- 二部图匹配：在二分图当中，如果我们选择了一条边就会连通对应的两个点。这也就构成了一个匹配，我们规定一个顶点最多只能构成一个匹配，也就是说所有的匹配之间没有公共的点。把对照组当成一个图，把处理组当成一个图，二图之间类似一个映射关系
- 非二部图匹配：不限制两个受试者集合，如剂量匹配中吃一片药和吃二、三、四片药的受试者并没有明确的区分对照组与处理组，即不严格区分哪个受试者该属于哪个集合
- 应用：用于治疗剂量匹配，或多个对照组匹配
- 匹配思路：从一个平方对称距离矩阵开始，矩阵的一行和一系列对应每位受试者且记录任意两位受试者之间的距离，然后将受试者分成匹配，使得配对受试者之间的总距离最小

### 1. 最优非二部图匹配：

ID	1	2	3	4	5	6
1	0	106	119	231	110	101
2	106	0	207	126	192	68
3	119	207	0	156	247	25
4	231	126	156	0	34	67
5	110	192	247	34	0	212
6	101	68	25	67	212	0

将六位受试者配成三对，约束自己与自己不能匹配，

结果：1和2配对，3和6配对，4和5配对

### 2. 用非二部图匹配执行处理-对照匹配

ID	1	2	3	4	5	6
1	0	$\infty$	$\infty$	231	110	101
2	$\infty$	0	$\infty$	126	192	68
3	$\infty$	$\infty$	0	156	247	25
4	231	126	156	0	$\infty$	$\infty$
5	110	192	247	$\infty$	0	$\infty$
6	101	68	25	$\infty$	$\infty$	0

为使处理者与对照者相匹配，在处理受试者之间和对照者之间都设置无穷大距离，这里1、2、3为处理者，4、5、6为对照者

- 不建议使用非二部图匹配来执行处理-对照匹配：极大增大矩阵的大小，导致计算复杂度增加

### 3. 剂量匹配

ID	1	2	3	4	5	6
1	0	$\infty$	119	231	110	101
2	$\infty$	0	$\infty$	126	192	68
3	119	$\infty$	0	$\infty$	247	25
4	231	126	$\infty$	0	$\infty$	67
5	110	192	247	$\infty$	0	$\infty$
6	101	68	25	67	$\infty$	0

- 假设多吃了一片带来的收益是线性的，估计剂量效应
- 当服用的剂量有一定的差异时才能说明两人的结局差异来源于剂量不同
- 约束：当两位受试者的剂量相差小于 2 时，设置无穷大距离
- 结果：1与5匹配，2与4匹配，3与6匹配，在每个配对里估一个平均的剂量效应，再取平均即可

#### 4. 多个组匹配

ID	1	2	3	4	5	6
1	0	$\infty$	119	231	110	101
2	$\infty$	0	207	126	192	68
3	119	207	0	$\infty$	247	25
4	231	126	$\infty$	0	34	67
5	110	192	247	34	0	$\infty$
6	101	68	25	67	$\infty$	0

1与2一组，3与4一组，5与6一组，将同一组内的个体间设置为无穷大进行匹配，匹配结果类似正交设计

- 无组别匹配注意事项：
  - 奇数个受试者：给个槽

ID	1	2	3	4	5	槽
1	0	106	119	231	110	0
2	106	0	207	126	192	0
3	119	207	0	156	247	0
4	231	126	156	0	34	0
5	110	192	247	34	0	0
槽	0	0	0	0	0	0

- 最终将与槽相匹配的个体扔掉
- 想多扔掉一些受试者可以多加几个槽

例：最低工资会降低就业率吗（Card和Krueger的研究）

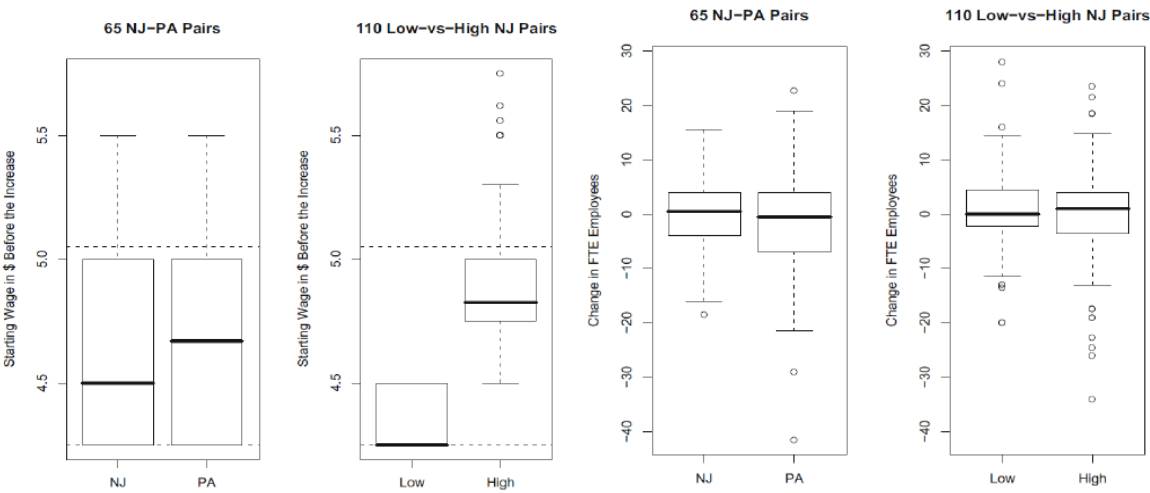
- 用匹配方法研究提高最低工资对就业率的影响
- 数据描述：

连锁餐厅	是否直营	每天营业时间	地点	起薪	编号	301	310	477	434	208	253
KFC	Yes	10.5	NJ	5	301	—	7.4	2067.6	2065.5	233.5	698.0
W	No	11.5	NJ	4.25	310	7.4	—	7.3	4.7	254.8	6.4
BK	No	16.5	PA	4.25	477	2067.6	7.3	—	749.5	641.2	1583.1
BK	No	16	PA	4.25	434	2065.5	4.7	749.5	—	642.5	1579.7
RR	Yes	17	NJ	4.62	208	233.5	254.8	641.2	642.5	—	475.7
RR	Yes	13	NJ	4.87	253	698.0	6.4	1583.1	1579.7	475.7	—

- 为能保证体现出最低工资政策的影响，保证可比性，对起薪差额小于 0.50 美元的两家餐厅的距离施加惩罚项，这样它们加薪都加到同一个水平时，他们的就业率差异能反应出因果作用
- 因营业时间对就业率的影响不大，这里对营业时间不需要完全匹配，只需精细均衡即可
- 将新泽西内部110个样本 进行施加了惩罚项的配对，最终分为受影响较大（即起薪低的店）和受影响较小（即起薪较高的店）两个组：

组别	类型	均值	最小值	下四分位数	中位数	上四分位数	最大值
受影响较大	NJ-vs.-NJ	4.33	4.25	4.25	4.25	4.50	4.50
受影响较小	NJ-vs.-NJ	4.91	4.50	4.75	4.83	5.00	5.75
差值	NJ-vs.-NJ	0.58	0.24	0.50	0.58	0.50	1.25

- 结果：



新泽西州提高最低工资之前的起薪                      全职等效就业人数的变化

- 虚线表示新泽西州政策要求的最低工资标准
- 用全职等效就业人数（FTE）衡量就业水平
- 可看出新泽西州和宾夕法尼亚州在政策实施前后就业率基本无变化，新泽西州内部配对在政策实施前后就业率也基本无变化，否定了提高最低工资会降低就业率的论断

- 该结论也受到了质疑：如选择样本有问题，宾夕法尼亚州和新泽西州关于最低工资的标准不同，关于这些质疑下回再叙

## 【精彩问答】

记录人：段月然 中国地质大学（北京）

1. Q: (1/倾向得分) 含义是什么？为什么权重是倾向得分分之一呢？

因果推断在观察性研究中的应用：分析

一、逆概率加权

### 加权估计量

- 因此可以得到平均因果作用的估计
- $$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \left( \frac{W_i \cdot Y_i^{\text{obs}}}{e(X_i)} - \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-e(X_i)} \right) = \frac{1}{N} \sum_{i=1}^N \left( \frac{(W_i - e(X_i)) \cdot Y_i^{\text{obs}}}{e(X_i) \cdot (1-e(X_i))} \right)$$
- 在实际中，倾向得分往往未知，需要估计。
- $$\hat{\tau} = \frac{1}{N} \sum_{i=1}^N \left( \frac{W_i \cdot Y_i^{\text{obs}}}{\hat{e}(X_i)} - \frac{(1-W_i) \cdot Y_i^{\text{obs}}}{1-\hat{e}(X_i)} \right) = \frac{1}{N} \sum_{i=1}^N \left( \frac{(W_i - \hat{e}(X_i)) \cdot Y_i^{\text{obs}}}{\hat{e}(X_i) \cdot (1-\hat{e}(X_i))} \right)$$

10

A: 在逆概率加权中（上图公式1），公式简化得到的最右侧公式中  $1/e(X_i)$  可以看作加权估计，它的权重就是  $1/e(X_i)$ ，就是倾向得分的倒数。

权重是倾向得分的倒数，这是逆概率加权的结果。

2. Q: 相合这个词，是什么概念？

A: 样本量大的时候，估计的准确。

吴鹏：consistent，指的是依概率收敛。某种意义上来说，可理解为asymptotically unbiased（渐近无偏）。

3. Q: 双稳健性估计部分，“倾向得分模型”指的是估计倾向性得分e(x)时设定的模型吗？

A: 对，一般是用Logistic模型来设定，但Logistic有可能设定错误导致  $\hat{e}(X_i)$  并不能趋向于  $e(X_i)$ 。这时候就是倾向得分模型设定错误。

## 双稳健估计

- 在回归中纳入协变量：
- $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i \beta + \varepsilon_i$
- 每个个体的权重为
- $$\hat{\lambda}_i = \begin{cases} \frac{1}{1-\hat{e}(X_i)} & \text{如果 } W_i = 0 \\ \frac{1}{\hat{e}(X_i)} & \text{如果 } W_i = 1 \end{cases}$$
- 这样得到的估计量是“双稳健”的：只要倾向得分模型或回归模型正确，因果作用的估计量就是相合的。

14

### 4. Q: 分数diff太大不匹配不行吗？为什么要截断？

**A:** 暂时没有涉及到匹配的问题。截断的意思就是把倾向得分太大或太小的部分扔掉。如果你不匹配的意思是扔掉，那就是截断的意思。

### 5. Q: 如何来确定分层的边界？

## 构造倾向得分分层

- 为了保证处理组和对照组重叠（否则处理组或对照组将没有数据进行比较），取
- $\underline{e}_t = \min_{i:W_i=1} \hat{e}(X_i), \bar{e}_c = \max_{i:W_i=0} \hat{e}(X_i)$
- 从一个层开始： $J = 1$ ，边界为  $b_0 = \underline{e}_t, b_1 = \bar{e}_c$ 。
- 这么多层是否足够？如果不够，需要细分。
- 取线性化的倾向得分（对数比率）
- $\hat{\ell}(x) = \log\left(\frac{\hat{e}(x)}{1-\hat{e}(x)}\right)$

17

**A:** 确定分层边界需要先给一个全局的上界和下界，上界就是处理组的倾向得分最小值，下界是对照组倾向得分的最大值。先设定全局的上界和下界后，再分层。分层就需要在一层内做t检验，如果检验没有通过需要在这一层中位数处截成两层。

**Q:** 截断把数据本身的版图 data self- pattern 给重构了，这种处理的出发点，考虑是数据的内聚吗，还是观察性任务的需要？

**A:** 可以当成内聚，也可以不用截断后的样本，但如果不用截断后的样本得到的估计量的方差可能会很大。比如实际情况要检验吃药有没有效使用了全部样本后发现方差太大，就会导致你不知道这个药是否有效。为了增加稳健性，可能会对倾向得分做一些处理，比如抛弃一些不合理的样本，就像是抛弃outliers一样。近两年也有一些关于trimming（截断）的前沿研究证明了当样本趋于无穷大的时候trimming，trimming的界限也会趋于1和0，这样仍然可以得到一个渐进无偏的估计。





6. Q: 35页的假设，为什么认为回归估计得到的是因果作用的真值呢？隐含的意思是回归估计得到的因果作用更准确吗？

## 逆概率加权与子分类

- 当层数足够多，每一层内的倾向得分波动足够小，逆概率加权估计与子分类估计是相近的。
- 但在实践中，推荐使用子分类估计。
  1. 如果倾向得分估计错误，逆概率加权估计受到的影响更大；估计量的倒数可能不稳定。
  2. 子分类估计的方差更小，因为倾向得分向层内均值集中了，因此子分类估计更稳健。
  3. 关于纳入协变量调整，子分类估计允许局部（层内分别）调整，而逆概率加权只能进行全局调整。

A: 这只是一个假设，没有隐含回归估计更准确的含义。这是观察性数据，我们并不知道它的真实值，只能假设一个ground truth，我们假设回归估计是一个真实值。因为加权估计和子分类估计都涉及到对倾向得分分层。用到倾向得分时，就会不稳定，这是因为它的方差很大。而回归估计的方差比较小，更稳健。所以就假设回归估计是一个真实值。

Q: 14页双稳健那里，我理解说：第一节课 认为非混杂性一定成立时，使用回归不管形式是否正确 得到的因果估计都是相合的；然后考虑到非混杂性不存在，使用你概率加权或者其他方式时，如果想要因果的估计是相合的，就要倾向性得分模型或者回归的形式正确了。请问这样理解对吗？

A:

7. Q: 这里讲的子分类是依据倾向性得分来分的，这里的倾向性得分是一开始就用比如LR估计出了每条样本的倾向性得分，然后再去分层吗？之后构造分层时还会在每个层内再重新估计倾向性得分吗？

A:

8. Q: 协变量的数量是如何决定的？是人为确定的吗？

A:

9. Q: 为什么在倾向性得分子分类方法中，需要检验每个协变量在每一层均衡？倾向性得分的目的是降维，使得我们不必在高维空间中做均衡，即不必保证每一层每一维都均衡。

p33指出降低了对回归协变量的敏感性，可不可以理解从bias-variance tradeoff的角度理解为variance降低？

A:

10. Q: 垂直符号能简单解释一下吗?

A:

11. Q: 构造分层时不会在每个层内再重新估计倾向性得分了。

A:

12. Q: 子分类估计，每个层内，都会做一个线性回归，协变量在每层的回归系数不一样，跟整体回归相比，会存在过拟合的问题吗？感觉不太符合直觉，协变量系数，在每一层不同，说明对y具有不同的影响？

A:

13. Q: 请问如果无法获得对照组信息时如何从观察数据中进行估计呢？比如存在selection bias时，只能观察到对撞节点一种取值的情况，别的取值无法获取。

A:

14. Q: 估计的倾向性得分，实际的倾向性得分，两者有什么区别呢，实际应用中，分别怎么计算？

A:

15. Q: 如果会过拟合，结论的泛化性会不会很差，换个数据集，结论会不会就相反了？

A:

16. Q: 子分类方式有没有R的软件包？

A: 子分类倾向得分的估计目前是没有，匹配是有软件包的。

17. Q: 匹配方法，能处理的最大样本数量，大概多大？样本量太大，是不是计算比较慢？

A: 在生物统计中一般是几千到几万。样本量太大，肯定会比较慢，本身匹配就比较慢。

19. Q: 估计倾向得分必须用logit模型吗

A: 还可以用probit

20. Q: 为什么说估计的倾向得分比真实的倾向得分可能要好？

A:

## 估计的倾向得分

- 估计的倾向得分 $\hat{e}(X)$ 往往比真实的倾向得分 $e(X)$ 更好。
- 估计的倾向得分在协变量 $X$ 上略有过度拟合，产生比随机实验更好的均衡效果。
- 因为倾向得分是用来平衡协变量的，所有过度拟合也没什么问题。
- 理论研究还表明，估计的倾向得分可能比真实的倾向得分有更好的性质。

☞

Q: 那是不是说明observational studies比RCT可能要好点了么？

A:

21. Q: 对于差额小于0.5的两家餐厅的距离施加惩罚，但是右边的矩阵的数值哪里体现了惩罚呢？（还是说右边的矩阵只是计算距离，计算距离之后再施加惩罚呢？）

因果推断在观察性研究中的应用：分析

三、匹配

实例：最低工资会降低就业率吗

- 如果对新泽西州的两家餐厅匹配，我们希望这两家匹配的餐厅在加薪前的起薪有很大不同。
- 对起薪差额小于0.50美元的两家餐厅的距离施加惩罚项。

连锁餐厅	是否直营	每天营业时间	地点	起薪	编号	301	310	477	434	208	253
KFC	Yes	10.5	NJ	5	301	—	7.4	2067.6	2065.5	233.5	698.0
W	No	11.5	NJ	4.25	310	7.4	—	7.3	4.7	254.8	6.4
BK	No	16.5	PA	4.25	477	2067.6	7.3	—	749.5	641.2	1583.1
BK	No	16	PA	4.25	434	2065.5	4.7	749.5	—	642.5	1579.7
RR	Yes	17	NJ	4.62	208	233.5	254.8	641.2	642.5	—	475.7
RR	Yes	13	NJ	4.87	253	698.0	6.4	1583.1	1579.7	475.7	—

82

A: 表格中出现的2000以上的数值都是惩罚得到的，还有600多和700的数据也都是惩罚得到的。

22. Q: 倾向性得分如果估计的不准，其实最终ATE也是有偏的？这儿是不是还有个trade-off啊？就是估计的倾向性得分和真实的倾向性得分之间？

A:

23. Q: 倾向得分相等 $p(t|X_1=x_{1i}, X_2=x_{2i})=p(t|X_1=x_{1j}, X_2=x_{2j})$ ，但 $x_i \neq x_j$ 。匹配个体i和j有问题吗？用倾向得分匹配后，处理组与对照组中协变量X就接近平衡吧？还要进一步检验处理组与对照组中协变量X的平衡吗？如果不平衡，可能是倾向得分匹配的不好吧？

A:

24. Q: 有评估倾向性得分模型的方法吗？

A: 一般是用logit回归


前面有一个地方提到了检验倾向得分的均衡性， 给定一个倾向得分的模型怎么知道这个模型好不好

25. Q: 目前的方法对我来说最难的地方就是假设，那些非混淆性假设怎么满足。

A:

26.

【读书会PPT】



20211121 【邓宇昊】观察性研究：分析1.pdf

1.16MB

