体育经济分析：理论与应用

附加实证专题：匹配方法

周正卿

02 May 2022

引言

大纲

上次实证方法

- 条件独立假设: $(Y_{0i}, Y_{1i}) \perp D_i | X_i$
- CEF、LPF
- 遗漏变量偏误公式

今天

- project proposals.
- 匹配估计 (*MHE* 3.2 和 C&T 25.4).

匹配方法

主旨

- 回忆 **条件独立假设CIA** [†]： 在给定**可观测的控制变量**条件下，处置变量的分配就像RCT一样好。
- **匹配估计量**的直观想法
- 假若承认条件独立假设 $\Longleftrightarrow (Y_{1i}, Y_{0i}) \perp D_i | X_i$ ， 那么可以计算出以 $X_i = x$ 为条件的处置效应：

$$\tau(x) = E\left[Y_{1i} - Y_{0i} \mid X_i = x\right]$$

- **匹配的想法是:** 通过"控制"可观测变量（几乎相同），即将干预组和控制组的个体按可观测特征匹配，解决基于可观测变量自选择造成的偏差

† 通过增加控制变量的方式解决了选择偏误的问题，但当选择偏误由于不可观测的因素造成时，就该方法就失去作用。**基于可观测特征的选择（selection on observables）：只根据可观测特征而选择是否接受处置**

目标

- 回到**因果推断的根本问题**：

1. 要估计 $\tau_i = Y_{1i} - Y_{0i}$
2. 无法同时观察到 Y_{1i} 和 Y_{0i}

- 大多数实证"武器库"都指向为干预组个体去寻找 Y_{0i}，即干预组的不可观察的反事实。

目标

- 回到**因果推断的根本问题**：

1. 要估计 $\tau_i = Y_{1i} - Y_{0i}$
2. 无法同时观察到 Y_{1i} 和 Y_{0i}

- 大多数实证"武器库"都指向为干预组个体去寻找 Y_{0i}，即干预组的不可观察的反事实。

- **匹配方法的思路不同**

- 将控制组和干预组的个体用可观测的特征 X_i 进行配对，以 "匹配成功的"控制组个体的结果变量为参照，将其作为干预组个体结果变量 Y_{1i} 的反事实结果估计值 $\widehat{Y_{0i}}$。

正式的

- 我们想为每一个**干预组的个体（$D_i = 1$）**构建一个反事实
- i 个体的反事实只使用与其可观测特征 X_i 的个体。
- 假设干预组有 N_T 个体，控制组有 N_C 个体，预期有：
  - 干预组有 N_T 个权重集，
  - 控制组 N_C 个体中有与之相匹配的权重

正式的

- 我们想为每一个**干预组的个体（$D_i = 1$）**构建一个反事实
- i 个体的反事实只使用与其可观测特征 X_i 的个体。
- 假设干预组有 N_T 个体，控制组有 N_C 个体，预期有：
  - 干预组有 N_T 个权重集，
  - 控制组 N_C 个体中有与之相匹配的权重: $w_i(j)$ $(i = 1, \ldots, N_T;\ j = 1, \ldots, N_C)$

正式的

- 我们想为每一个**干预组的个体（$D_i = 1$）**构建一个反事实
- i 个体的反事实只使用与其可观测特征 X_i 的个体。
- 假设干预组有 N_T 个体，控制组有 N_C 个体，预期有：
  - 干预组有 N_T 个权重集，
  - 控制组 N_C 个体中有与之相匹配的权重: $w_i(j)$ $(i = 1, \ldots, N_T;\ j = 1, \ldots, N_C)$
- 假定 $\sum_j w_i(j) = 1$ 。干预组 i 个体的反事实结果估计为：

$$\widehat{Y_{0i}} = \sum_{j \in (D=0)} w_i(j) Y_j$$

正式的

- 如果对干预组个体 i 的反事实估计值为：

$$\widehat{Y_{0i}} = \sum_j w_i(j) Y_j$$

- 那么每一个个体 i 的处置效应估计值为：

$$\hat{\tau}_i = Y_{1i} - \widehat{Y_{0i}} = Y_{1i} - \sum_j w_i(j) Y_j$$

正式的

- 如果对干预组个体 i 的反事实估计值为：

$$\widehat{Y_{0i}} = \sum_j w_i(j) Y_j$$

- 那么每一个个体 i 的处置效应估计值为：

$$\hat{\tau}_i = Y_{1i} - \widehat{Y_{0i}} = Y_{1i} - \sum_j w_i(j) Y_j$$

所以对干预组的处置效应的**一般化的匹配的处置效应**是

$$\hat{\tau}_M = \frac{1}{N_T} \sum_{i \in (D=1)} \left(Y_{1i} - \widehat{Y_{0i}} \right) = \frac{1}{N_T} \sum_{i \in (D=1)} \left(Y_{1i} - \sum_{j \in (D=0)} w_i(j) Y_j \right)$$

为匹配对象加权†

- 所以匹配方法的核心是如何进行加权。††

† 🤦 †† 除此之外，再加上有趣的、政策相关的、具有可信的条件独立性假设。还有数据。

为匹配对象加权†

- 所以匹配方法的核心是如何进行加权。††

Q 哪里可以找到这些方便的权重?

A 统计学家提供了选项，但需要谨慎/负责任地选择。

- 例如，如果对于所有 (i, j) 组合赋予相同权重 $w_i(j) = \frac{1}{N_C}$ ，那么相当于回到了均值差。这意味这些寻找的权重在赋权后，不能与条件独立假设冲突。
- **计划是：** 选择的权重能够表明干预组特征 X_i 与控制组特征 X_j 是 ***如何接近的***

† 🤦 †† 除此之外，再加上有趣的、政策相关的、具有可信的条件独立性假设。还有数据。

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量
- 如果 X 是 **离散的**，那么可以考虑平等性，如 $w_i(j) = I(X_i = X_j)$，根据需要进行缩放从而得到 $\sum_j w_i(j) = 1$ 。

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量
- 如果 X 是 **连续的**，那么可以考虑***相似性*** 而非***平等性***。
- *最近邻匹配* 使用的是 X_i 和 X_j 之间的欧几里得距离，为干预组个体选择一个最接近的控制组观察点

$$d_{i,j} = \left(X_i - X_j\right)' \left(X_i - X_j\right)$$

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量
- 如果 X 是 **连续的**，那么可以考虑***相似性*** 而非***平等性***。
- *最近邻匹配* 使用的是 X_i 和 X_j 之间的欧几里得距离，为干预组个体选择一个最接近的控制组观察点

$$d_{i,j} = \left(X_i - X_j\right)' \left(X_i - X_j\right)$$

- 每个个体的处置效应记作： $\hat{\tau}_i = Y_{1i} - Y_{0j}^i$ ，其中 Y_{0j}^i 是 i 个体在控制组中最近邻
- **ATE估计值:** $\hat{\tau}_M = \frac{1}{N_T} \sum_i \hat{\tau}_i$
- 如果CIA成立并且有足够的重叠区域，就能计算处置效应。
- 缺点是:欧式距离受量纲的影响，两点之间的欧氏距离与原始数据测量单位相关 → 原始数据要进行标准化处理

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量
- 如果 X 是 **连续的**，那么可以考虑***相似性*** 而非***平等性***。
- 使用马哈拉诺比斯距离作为*最近邻匹配*：在 X_i 与 X_j 之间使用最近的马氏距离 选择唯一匹配

$$d_{i,j} = \left(X_i - X_j\right)' \Sigma_X^{-1} \left(X_i - X_j\right)$$

其中 Σ_X^{-1} 是 X 的协方差矩阵

相似性Proximity

- 我们选择的权重 $w_i(j)$ 应该是反映干预组特征 X_i 与控制组特征 X_j ***如何接近的*** 代理变量
- 如果 X 是 **连续的**，那么可以考虑***相似性*** 而非***平等性***。
- 使用马哈拉诺比斯距离作为*最近邻匹配*： 在 X_i 与 X_j 之间使用最近的马氏距离 选择唯一匹配

$$d_{i,j} = \left(X_i - X_j\right)' \Sigma_X^{-1} \left(X_i - X_j\right)$$

其中 Σ_X^{-1} 是 X 的协方差矩阵

- **ATE估计值:** $\hat{\tau}_M = \frac{1}{N_T} \sum_i \hat{\tau}_i$ ，其中 $\left(\hat{\tau}_i = Y_{1i} - Y_{0j}^i\right)$
- 如果CIA成立并且有足够的重叠区域，就能计算处置效应。
- 优点:马氏距离不受量纲的影响，两点之间的马氏距离与原始数据测量单位无关；由标准化数据和中心化数据计算出的二点之间的马氏距离相同；排除变量之间的相关性的干扰。缺点:夸大了变化微小的变量的作用。

匹配法的假设与条件

1. **条件独立假设:** $(Y_{0i}, Y_{1i}) \perp D_i | X_i$
    - CIA在匹配法语境下的理解为：在给定可观测变量 X_i 后，进入干预组还是控制组是随机分配的，不会因为潜在结果的好坏而决定是否接受干预
2. **重叠(Overlap):** $0 < Pr\left(D_i = 1 \mid X_i\right) < 1$
    - 给定可观测特征 $X_i = x$, 个体接受干预的概率大于0并小于1

- 两者同时满足，给定观测特征 $X_i = x$ 的ATE为：
- 对特征为 x 的个体 $ATT(x) = ATU(x) = ATE(x) \rightarrow ATE = E_X[ATE(X)] = \sum_x ATE(x)P(x)$

单邻居 → 多邻居匹配 → 倾向得分匹配

- 单邻居匹配在 $N_C \gg N_T$ 时，就会丢掉**很多**控制组个体 → 信息浪费
  - 多邻居使用**核匹配Kernel matching**，需要选择：核密度函数+带宽
- 当使用多个可观测特征匹配时，面临"维数的诅咒"
  - 可观测特征维数增加时，为每个干预组个体找到一个很好的、近距离的控制组会很困难。
- 倾向得分法(Propensity Score Methods，PSM)
  - 原理：通过函数关系将多维特征变量 X 变换为一维的倾向得分 $ps(X_i)$ 之后，再根据倾向得分进行匹配。倾向得分是可观测特征为 $X_i = x$ 的个体进入干预组（接受干预）的概率

$$ps(X_i = x) = P(D_i = 1 \mid X_i = x)$$

倾向得分匹配

原理

- 匹配法的假设与条件：**CIA + Overlap**
- **CIA** → **倾向得分定理**
  - 如果 $(Y_{0i}, Y_{1i}) \perp D_i | X_i$，等价于 $(Y_{0i}, Y_{1i}) \perp D_i | ps(X_i)$.
  - 通俗理解: 倾向得分 $ps\left(X_i\right)$ 总结了变量 X_i 中包含的所有相关信息
- **倾向得分是均衡得分** → 一维得分代替多维特征→ 若干预组和控制组的个体有相同的倾向得分，两组的**可观测特征分布在两组就是均衡的**†

$$X_i \perp D_i \mid ps\left(X_i\right)$$

  - 若只关心均值，则上式可简化为： $E\left(X_i \mid D_i=1, ps\left(X_i\right)\right)=E\left(X_i \mid D_i=0, ps\left(X_i\right)\right)$

† 不可观测的特征不一定是均衡的，因此前提要求不可观测的特征与干预变量无关。在RCT中，倾向得分是已知的。每个人被分到干预组概率是50%，倾向得分就是50%。在观测数据中，倾向得分是未知的。

倾向得分定理的证明

- 定理： $(Y_{0i},\, Y_{1i}) \perp D_i | X_i$，等价于 $(Y_{0i},\, Y_{1i}) \perp D_i | ps(X_i)$.

要想证明该定理，就要表明 $Pr\left(D_i = 1 \mid Y_{0i},\, Y_{1i},\, ps(X_i)\right) = ps(X_i)$：在给定 $ps(X_i)$ 后， D_i 与 $(Y_{0i},\, Y_{1i})$ 是独立的

$$
\begin{aligned}
&Pr\Big[D_i = 1 \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big] \\
&\quad = E\Big[D_i \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big] \\
&\quad = E\Big[E\Big(D_i \Big| Y_{0i},\, Y_{1i},\, ps(X_i),\, X_i\Big) \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big] \\
&\quad = E\Big[E\Big(D_i \Big| Y_{0i},\, Y_{1i},\, X_i\Big) \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big]
\end{aligned}
$$

倾向得分定理的证明

$$Pr\Big[D_i = 1 \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big] = \cdots = E\Big[E\Big(D_i \Big| Y_{0i},\, Y_{1i},\, X_i\Big) \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big]$$

$$= E\Big[E\Big(D_i \Big| X_i\Big) \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big]$$

$$= E\Big[ps(X_i) \Big| Y_{0i},\, Y_{1i},\, ps(X_i)\Big]$$

$$= ps(X_i)$$

$$\therefore (Y_{0i},\, Y_{1i}) \perp D_i | X_i \implies (Y_{0i},\, Y_{1i}) \perp D_i | ps(X_i) \quad ✔$$

直觉

Q 这到底是怎么回事?

- X_i 承载的信息肯定要比 $ps(X_i)$ 多， 那么如何才能通过给定 $ps(X_i)$ 情况来获得干预的条件独立性呢?
- A_1 干预的条件独立性并不是从 X_i 中提取所有可能的信息。实际上只关心给定 D_i 下其他特征变量与 (Y_{0i}, Y_{1i}) 无关的情况。
- A_2 回到主要关注点：**选择偏误**→人们会选择是否进入干预组。 如果控制住 X 后反映了两个人进入干预组的概率是相等的，并且如果 X_i 解释了所有的选择进入干预组的原因(CIA)，那么这两个人不可能会选择进入干预组。

估计

- 如何获得倾向得分?

我们对它们进行估计--有很多方法可以做到这一点：

1. 灵活性 (*如*, 相互项) Probit或Logit模型
2. 核函数估计
3. 其他方法：机器学习等

估计

MHE (p. 83)

问题

> 最大的问题是如何选择最优模型去估计 $ps(X_i)$...

回答

> 要视具体应用的情况而定。越来越多的实证文献表明，在连续协变量中带有几个多项式项的倾向得分的Logit模型在实践中运作良好.....

应用

- 假设已经获得了的倾向性分数估计值 $hatp(X_i)$ 。接下来呢?
- *选项 1：* 将其作为控制变量 + 回归
  - *选项 1a：* 使用 $ps(X_i)$ 进行 **条件回归**

$$Y_i = \alpha + \delta D_i + \beta ps(X_i) + u_i \tag{1a}$$

  - *选项 1b：* 如果想要得到异质性处置效应，意味着异质性处置效应与特征 X 共变，增加干预变量与倾向得分 $ps(X_i)$ 的**交互项**

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i ps(X_i) + \beta ps(X_i) + u_i \tag{1b}$$

回归的异质性

- 再思考一下这种情况下的异质性处置效应

$$
\begin{aligned}
Y_{0i} &= \alpha + \beta X_i + u_i \\
Y_{1i} &= Y_{0i} + \delta_1 + \delta_2 X_i
\end{aligned}
$$

例如, 估计处置效应依赖于 X_i

$$
\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
&= D_i \left(Y_{0i} + \delta_1 + \delta_2 X_i \right) + (1 - D_i) Y_{0i} \\
&= Y_{0i} + \delta_1 D_i + \delta_2 D_i X_i \\
&= \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i
\end{aligned}
$$

异质性

- 最后一个方程为：

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i X_i + \beta X_i + u_i$$

如果想要得到 $ps(X_i)$ 和 $D_i ps(X_i)$ 的估计系数：

$$Y_i = \alpha + \delta_1 D_i + \delta_2 D_i ps(X_i) + \beta ps(X_i) + u_i \tag{1b}$$

这意味着要区分：

1. **特定组别的处置效应：** 对于每一个 X_i 的 $\delta_1 + \delta_2 ps(X_i)$
2. **ATE：** $\delta_1 + \delta_2 \overline{p}(X_i)$

更为灵活

- 使用倾向得分匹配的初衷是：降低维度，估计/选择/假设更少的参数
- 在线性回归中加入 $ps(X_i)$ 和 $D_i ps(X_i)$ 作为协变量并不能完全发挥灵活/非参数估计的潜力

分块

- *选择 2：* 对倾向得分进行分块（分层）

1. 将 $hatp(X_i)$ 的范围分成 K 区间（例如_，0.05宽的区间）。
2. 将样本放置到对应的 $\hat{p}(X_i)$ 区间内
3. 计算每个区块内的处置效应差异 $hattau_k$
4. 依据每个区块的样本比重，对 $\hat{\tau}_k$ 取加权平均

$$\hat{\tau}_{Block} = \sum_{k=1}^{K} \hat{\tau}_k \frac{N_{1k} + N_{0k}}{N}$$

注意： 使用 $ps(X_i)$ 分块与核函数计算的距离逻辑相似

选择分块

- 对倾向得分分区间，需要定义每个区间
- 实际操作中会有多次尝试

  1. **选择合理的区间范围**
  2. 保证**每个特征变量** 在每个分块中要均衡

     -如果特征变量在分块中**不平衡**，重新分块

     - 如果特征变量在分块中**平衡**，进行下一步
- 缺陷：多变量情形难以保证每个变量在每个块中都均衡，需要接受个别不重要变量的不均衡，因此需要一定的主观判断。

重叠

- 分块强调了**重叠假设**： $0 < Pr\left(D_i | X_i\right) < 1.$
- 如果某个区块包含了零个干预或者控制样本，就无法计算出 $\hat{\tau}_k$.
- 注意事项： Logit/Probit 可以隐藏违规行为，因为它迫使倾向得分估计值 $0 < \hat{ps}(X_i) < 1$
- *常见做法：* 经验性地执行重叠的检验
  - 在干预组中将 $\hat{ps}(X_i)$ 低于最低倾向得分的控制组样本删除
  - 在控制组中将 $\hat{ps}(X_i)$ 高于最高倾向得分的干预组样本删除

加权

- *选项 3：* 依据倾向得分为每个样本反向赋权
- Q 用 $1/\hat{ps}(X_i)$ 加权意义何在?
- A 考虑一下"老朋友"(带有偏误的)两组均值之差：

$$\hat{\tau}_{Diff} = \overline{Y}_T - \overline{Y}_C = \frac{\sum_i D_i Y_i}{\sum_i D_i} - \frac{\sum_i (1 - D_i) Y_i}{\sum_i (1 - D_i)}$$

- 由于样本对是否干预有选择，导致偏误：

$$E\left[Y_{0i} | D_i = 1\right] \neq E\left[Y_{0i}\right]$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_i Y_i}{ps(X_i)}\right]$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_i Y_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_i Y_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right]$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\Bigg(E\left[\frac{D_iY_{1i}}{ps(X_i)}\,\middle|\,X_i\right]\Bigg)$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\Bigg(E\left[\frac{D_iY_{1i}}{ps(X_i)}\,\middle|\,X_i\right]\Bigg) = E\Bigg(\frac{E\left[D_i \mid X_i\right]E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\Bigg)$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\left(E\left[\frac{D_iY_{1i}}{ps(X_i)} \,\middle|\, X_i\right]\right) = E\left(\frac{E\left[D_i \mid X_i\right]E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right)$$

$$= E\left(\frac{ps(X_i)E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right)$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\left(E\left[\frac{D_iY_{1i}}{ps(X_i)} \,\middle|\, X_i\right]\right) = E\left(\frac{E\left[D_i \mid X_i\right]E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right)$$

$$= E\left(\frac{ps(X_i)E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right) = E\left(E\left[Y_{1i} \mid X_i\right]\right)$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\left(E\left[\frac{D_iY_{1i}}{ps(X_i)}\,\middle|\,X_i\right]\right) = E\left(\frac{E\left[D_i \mid X_i\right]E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right)$$

$$= E\left(\frac{ps(X_i)E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right) = E\Big(E\left[Y_{1i} \mid X_i\right]\Big) = E\left[Y_{1i}\right]$$

加权的依据

- 假设已知 $ps(X_i)$，并且用 $1/ps(X_i)$ 对每个**干预组**个体进行加权。

$$E\left[\frac{D_iY_i}{ps(X_i)}\right] = E\left[\frac{D_i\left(D_iY_{1i} + (1-D_i)Y_{0i}\right)}{ps(X_i)}\right] = E\left[\frac{D_iY_{1i}}{ps(X_i)}\right]$$

$$= E\left(E\left[\frac{D_iY_{1i}}{ps(X_i)}\,\middle|\, X_i\right]\right) = E\left(\frac{E\left[D_i \mid X_i\right]E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right)$$

$$= E\left(\frac{ps(X_i)E\left[Y_{1i} \mid X_i\right]}{ps(X_i)}\right) = E\Big(E\left[Y_{1i} \mid X_i\right]\Big) = E\left[Y_{1i}\right]$$

- 同样地, 用 $1/(1-ps(X_i))$ 对 **控制组** 个体可以得到：

$$E\left[\frac{(1-D_i)Y_i}{1-ps(X_i)}\right] = E\left[Y_{0i}\right]$$

加权：倾向得分匹配估计值

- 因此，我们可以通过以下方式估计无偏的处置效应

$$\hat{\tau}_{psWeight} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{ps(X_i)} - \frac{(1-D_i Y_i)}{1-ps(X_i)}\right]$$

- *直观上：* 以 X_i 为函数为干预组个体接受干预赋予更高的 $ps(X_i)$，来试图克服选择偏差

加权：倾向得分匹配估计值

- 因此，我们可以通过以下方式估计无偏的处置效应

$$\hat{\tau}_{psWeight} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{ps(X_i)} - \frac{(1-D_i Y_i)}{1-ps(X_i)}\right]$$

- *直观上：* 以 X_i 为函数为干预组个体接受干预赋予更高的 $ps(X_i)$，来试图克服选择偏差
- 试图让干预组个体回到**像是随机分配一样**
  - (1) 为**低分值 $ps(X_i)$ 干预组个体**提高权重，
    (2) 为**高分值 $ps(X_i)$ 控制组样本**提高权重。

例子

- 假设对某些个体 i ， $ps(X_i) = 0.80.$
- 这个ps得分说明某些带有 X_i 特征的个体分配到**干预组**概率是分配到**控制组**概率的4倍
- 反向加权可以解决每一个具有 X_i 特征个体的不平衡问题
  - 如果 i 是 **干预组**，那么他权重是 $1/ps(X_i) = 1/0.80 = 1.25$
  - 如果 i 是 **控制组**，那么他权重是 $1/(1 - ps(X_i)) = 1/(1 - 0.80) = 5$
  - 那么， $5/1.25$ 是 4 ！

例子

- 假设对某些个体 i， $ps(X_i) = 0.80.$
- 这个ps得分说明某些带有 X_i 特征的个体分配到**干预组**概率是分配到**控制组**概率的4倍
- 反向加权可以解决每一个具有 X_i 特征个体的不平衡问题
  - 如果 i 是 **干预组**，那么他权重是 $1/ps(X_i) = 1/0.80 = 1.25$
  - 如果 i 是 **控制组**，那么他权重是 $1/(1 - ps(X_i)) = 1/(1 - 0.80) = 5$
  - 那么， $5/1.25$ 是 4 ！
- 该加权方法使我们回到：给定每一组 X_i 时， 进入干预组和控制组的平衡

加权：最后问题

- *实践中问题：* 不一定能够保证 $\sum_i \hat{ps}(X_i) = 1$.
- *解决方法：* 通过加总将权重进行归一化处理。
- 应用归一化加权倾向得分估计值：

$$\hat{\tau}_{pWeight} = \sum_{i=1}^{N} \frac{\frac{D_i Y_i}{\hat{p}(X_i)}}{\sum_i \frac{D_i}{\hat{p}(X_i)}} - \sum_{i=1}^{N} \frac{\frac{(1-D_i) Y_i}{1-\hat{p}(X_i)}}{\sum_i \frac{(1-D_i)}{1-\hat{p}(X_i)}}$$

- Hirano, Imbens和Ridder (2003) 认为该估计值是有效的

为什么只选择其一就可以?

- 加权平均数没有什么特别之处，而且回归可以加权
- 因此，一个**基于回归的估计**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

为什么只选择其一就可以?

- 加权平均数没有什么特别之处，而且回归可以加权
- 因此，一个**基于回归的估计**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

其**权重**

$$w_i = \sqrt{\frac{D_i}{\hat{ps}(X_i)} + \frac{(1 - D_i)}{1 - \hat{ps}(X_i)}}$$

为什么只选择其一就可以?

- 加权平均数没有什么特别之处，而且回归可以加权
- 因此，一个**基于回归的估计**

$$Y_i = \alpha + X_i\beta + \tau D_i + u_i$$

其**权重**

$$w_i = \sqrt{\frac{D_i}{\hat{ps}(X_i)} + \frac{(1 - D_i)}{1 - \hat{ps}(X_i)}}$$

- 提供了一个**双重稳健**的特性，你有两个机会是正确的：一个是 $ps(X_i)$ 计算正确，另一个是回归模型设定正确。只要保证其一就可以。

为什么只选择其一就可以?

- 替代方案是，双重稳健的意思是将倾向得分区块与回归结合

为什么只选择其一就可以?

- 替代方案是，双重稳健的意思是将倾向得分区块与回归结合
- *Step 1* 对每个区块 k，进行回归

$$Y_i = \alpha_k + X_i\beta_k + \tau_k D_i + u_i$$

为什么只选择其一就可以?

- 替代方案是，双重稳健的意思是将倾向得分区块与回归结合
- *Step 1* 对每个区块 k，进行回归

$$Y_i = \alpha_k + X_i\beta_k + \tau_k D_i + u_i$$

- *Step 2* 对所有区块的处置效应加权平均

$$\hat{\tau} = \sum_{k=1}^{K} \hat{\tau}_k \frac{N_{1k} + N_{0k}}{N}$$

分清主次

- 不要(过于)迷恋那些花哨的东西
- 任何评审都会以下两个**主要要求**提出质疑：

分清主次

- 不要(过于)迷恋那些花哨的东西
- 任何评审都会以下两个**主要要求**提出质疑：
    1. **条件独立假设**满足吗?
    2. 在干预组和控制组是否满足**重叠**?

分清主次

- 不要(过于)迷恋那些花哨的东西
- 任何评审都会以下两个**主要要求**提出质疑：
  1. **条件独立假设**满足吗?
  2. 在干预组和控制组是否满足**重叠**?
- 对于第**2**种质疑，使用倾向得分匹配[†]时可以在数据中寻找证据
- 可以绘制**T**和**C**的 $ps(X_i)$ 的分布。

† 随着 X 维度扩大，检查 X 空间中的重叠情况会很困难

- 情况一： $ps(X_i)$ 不完全重叠且均值不相等

- 情况二： $ps(X_i)$ 的估计值被强制执行完全重叠，但均值不相等

- 情况三： 基于Logit回归估值（0-1）的 $\hat{ps}(X_i)$ 隐藏了某些没有重叠的真实 $ps(X_i)$

回归与匹配相同点

例子

- 假设观测数据如下：ID是个体编号， Y是观测结果（健康情况）， D是处 置与否（服药与否），AGE是可观测特征（年龄）

ID	Y	D	AGE
1	10	1	30
2	5	1	30
3	12	0	30
4	6	0	30
5	5	1	40
6	12	1	40
7	7	0	40
8	4	0	40
9	6	0	40
10	5	0	40

例子

- 按年龄和接受处置与否分成四组，得到每组Y的平均值和人数（括号中数值）。

	D = 1	D = 0
AGE=30	7.5 (2)	9 (2)
AGE=40	8.5 (2)	5.5 (4)

- 倘若满足CIA，那么：
  - $\widehat{ATE}\left(AGE_i = age\right) = \bar{Y}_1\left(AGE_i = age\right) - \overline{Y_0}\left(AGE_i = age\right) = \widehat{ATT} = \widehat{ATU}$
  - $\overline{ATT} = \sum_{age} P\left(AGE_i = age \mid D = 1\right) \times \widehat{ATT}\left(AGE_i = age\right)$
  - $\widehat{ATE} = \sum_{age} P\left(AGE_i = age\right) \times \widehat{ATE}\left(AGE_i = age\right)$

精确匹配

- 对年龄为 30 岁的:
  - 干预组: $\overline{Y_1}\left(AGE_i = 30\right) = \frac{10+5}{2} = 7.5$
  - 控制组: $\overline{Y_0}\left(AGE_i = 30\right) = \frac{12+6}{2} = 9$
  - 平均处置效应: $\widehat{ATE}\left(AGE_i = 30\right) = \bar{Y}_1\left(AGE_i = 30\right) - \bar{Y}_0\left(AGE_i = 30\right) = -1.5$
- 对年龄为 40 岁的:
  - 干预组: $\overline{Y_1}\left(AGE_i = 40\right) = \frac{5+12}{2} = 8.5$
  - 控制组: $\overline{Y_0}\left(AGE_i = 40\right) = \frac{7+4+6+5}{4} = 5.5$
  - 平均处置效应: $\widehat{ATE}\left(AGE_i = 40\right) = \overline{Y_1}\left(AGE_i = 40\right) - \bar{Y}_0\left(AGE_i = 40\right) = 3$

精确匹配

- 匹配相对回归方法的优点:
  - 通过相减，年龄对健康的影响已经去除，不需要假设年龄与健康的函数关系。
- 计算 $\widehat{ATE}$，用不同年龄人数的比率 $P\left(AGE_i = age\right)$ 为权重:
  - $\begin{aligned}\widehat{ATE} &= P\left(AGE_i = 30\right) \times \widehat{ATE}\left(AGE_i = 30\right) + P\left(AGE_i = 40\right) \times \widehat{ATE}\left(AGE_i = 40\right) \\ &= \frac{4}{10}(-1.5) + \frac{6}{10} \times 3 = 1.2\end{aligned}$
- 精确匹配法允许不同年龄的平均处置效应是不同的， 即**允许异质处置效应**。要得到总体平均处置效应，使用相应人数比例进行加权平均。

回归方法：完全饱和模型

- **完全饱和模型:**对解释变量（控制变量+处置变量)的所有可能组合值(含交互项)都有对应系数的回归模型。
- 用4个虚拟变量D0AGE30、D1AGE30、D0AGE40、D1AGE40来代表四个可能组合。
- 将完全饱和回归模型设置如下:

$$Y_i = \beta_1 D0AGE30_i + \beta_2 D1AGE30_i + \beta_3 D0AGE40_i + \beta_4 D1AGE40_i + e_i$$

  - 这里不包括截距常数项, 避免共线性

- 完全饱和回归模型对应的条件期望函数:

$$\begin{aligned} &E\left(Y_i \mid D0AGE30i, D1AGE30i, D0AGE40i, D1AGE40i\right) \\ &= \beta_1 D0AGE30_i + \beta_2 D1AGE30_i + \beta_3 D0AGE40_i + \beta_4 D1AGE40_i \end{aligned}$$

- 系数含义： β_1 为当 $D_i = 0, AGE_i = 30$ 时, Y_i 的均值；其他同理。

回归方法：完全饱和模型

- 完全饱和模型的系数对应了不同年龄干预组和控制组观测结果的均值
- 相同年龄组别的系数之差就是代表该年龄组别的处置效应 → 精确匹配结果

完全饱和回归模型和精确匹配得到结果完全一致

回归与匹配差异

控制变量饱和模型

- **控制变量饱和模型:** 对控制变量的所有可能组合值(仅含控制变量间交互)都有对应系数。
- 用2个虚拟变量AGE30和AGE40来代表两个可能组合:
  - 当 $AGE_i = 30$ 时， $AGE30_i = 1$； 否则， $AGE30_i = 0$ 。
  - 当 $AGE_i = 40$ 时， $AGE40_i = 1$； 否则， $AGE40_i = 0$ 。
- 将控制变量饱和模型设置如下:

$$Y_i = \beta D_i + \phi_1 AGE30_i + \phi_2 AGE40_i + e_i$$

  - 避免共线性，这里同样不包含截距项
  - D和AGE不再有交叉项，意味着对于不同年龄只有一个相同的处置效应估计值。
- 从条件期望值来看系数的含义： ϕ_1 为当 $D_i = 0, AGE_i = 30$ 时, Y_i 的均值 $\beta + \phi_1$ 为当 $D_i = 1, AGE_i = 30$ 时, Y_i 的均值; ϕ_2 为当 $D_i = 0, AGE_i = 40$ 时, Y_i 的均值; $\beta + \phi_2$ 为当 $D_i = 1, AGE_i = 40$ 时, Y_i 的均值

控制变量饱和模型

- 处置效应对于30和40两个组别的处置效应时同质的
- 与精确匹配法和完全饱和模型得到的处置效应估计略有出入

$\widehat{ATE}$ 差异：估计权重差异

- 在精确匹配模型或完全饱和回归模型中, $\widehat{ATE}$ 是每个 $\widehat{ATE}\left(AGE_i\right)$ 的加权平均值, 其权重为干预组中不同年龄个体的比率, 即:

$$\widehat{ATE} = P\left(AGE_i = 30\right) \times \widehat{ATE}\left(AGE_i = 30\right) + P\left(AGE_i = 40\right) \times \widehat{ATE}\left(AGE_i = 40\right)$$

- 在控制变量饱和回归模型中的处置效应是基于回归方法的估计系数，是建立在最小化估计方差基础上。对方差最小的年龄组别的处置效应赋予更大的权重。

- 要使精确匹配/完全饱和模型与控制变量饱和模型得到结果完全一致, **处置效应需不存在差异化**, 即 $\widehat{ATE}\left(X_i = x\right)$ 为常数。

- 由于控制变量饱和模型的处置变量系数 $\widehat{\beta}$ 使用的权重并不符合ATE对权重的定义, 因此并非ATE的一致或无偏估计量。

- 实际运用中还会涉及到**非饱和模型:** $Y_i = \beta D_i + \phi AGE_i + e_i$, 不同年龄对收入的影响一样，都等于 ϕ。

$\widehat{ATE}$ 差异：缺乏重叠以及控制变量不均衡

- 重叠：指干预组和控制组控制变量的分布范围是否重叠。
- 均衡性：指干预组和控制组控制变量的均值是否接近
- 回归方法中，未进行重叠和均衡检验。
    - 在不重叠、不均衡的情况下，回归估计的处置效应依赖于回归模型形式设定是否正确，即使正确，稳健性较低。

$\widehat{ATE}$ 差异：例子

- 假设收入只受受教育程度和智商的影响

ID	INC1	College	IQ	ID	INC1	College	IQ
1	15283	0	58	11	25281	1	101
2	15539	0	61	12	25467	1	102
3	15986	0	64	13	26622	1	116
4	16687	0	66	14	27781	1	120
5	16841	0	68	15	27675	1	127
6	17087	0	76	16	27347	1	129
7	18117	0	82	17	28334	1	133
8	20260	0	90	18	28671	1	135
9	19472	0	90	19	29705	1	146
10	18811	0	93	20	29809	1	147

$\widehat{ATE}$ 差异：例子

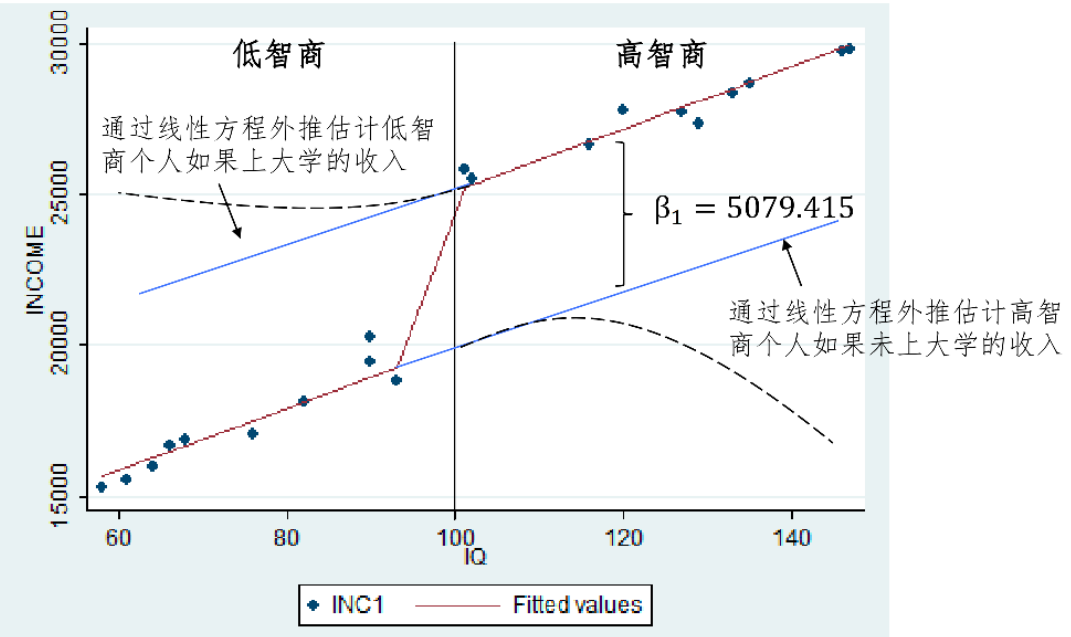
- 匹配方法：完全饱和模型和控制变量饱和模型
  - 精确匹配法: 数据不满足重叠要求, 无法匹配
    - 智商高于100的个体，都上大学,低于100，都没上大学 → 干预组和控制组个体无法匹配 → 无法估计各智商下的ATE，无法计算ATE
  - 完全饱和模型：先通过回归系数估计得到不同智商干预组（上 大学个体) 和控制组 (没上大学个体), 无干预组和控制组都 存在的观测结果, 无法估计。
    - 先获得解释变量的所有组合 → 给定任意智商水平，没有同时在干预组合控制组的数据 → 干预组和控制组个体无法匹配 → 无法估计各智商水平的ATE
  - 控制变量饱和模型: 对于只有干预组或控制组的数据, 权重为 0 , 无法估计处置效应 $\widehat{ATE}$

$\widehat{ATE}$ 差异：例子

- 唯一的出路就是使用：非饱和回归模型

$$INC_i = \alpha + \beta_1 \text{ College } e_i + \beta_2 IQ_i + e_i$$

- 假设IQ与INC是线性关系：IQ每增加1，收入均值增加 β_2



$\widehat{ATE}$ 差异：例子

- 严重取决于模型设定。如果真实的反事实的结果如虚线(曲线)所示，而非简单的线性关系，这种情况下，通过假设的线性函数外推得到的反事实结果就与真实情况相去甚远。
  - 在缺乏重叠情况下，回归得到的处置效应取决于模型设定正确与否
- 结果不稳健。即使关系假设正确，在缺乏重叠并不均衡的情况下，处置变量系数 β_1 的估计值也容易受到控制变量 β_2 的影响。
  - 用一组新数据INC2（小幅度改动）再次回归

$\widehat{ATE}$ 差异：例子

- 若将数据改为具备重叠和均衡情况时，用INC1和INC2再进行上述
  - 在中等智商的人，有些上了大学，有些没有上大学 → 重叠
  - 智商在上大学和没上大学的组别中均值相等 → 平衡
- 结果显示： 处置效应在INC1和INC2的估计结果比较稳健；尽管IQ在两个样本中系数差别较大。

总结

相同点

1.回归方法和匹配方法都是用于处理在估计处置效应中由于**可观测变量自选择**造成的偏差，它们都不能处理在估计处置效应中由于**不可观测变量自选择**造成的偏差。

2.**精确匹配**与**完全饱和回归模型**估计的处置效应是相同的。

总结

不同点

1.匹配法和控制变量饱和回归模型在计算平均处置效应(非条件的) $\widehat{ATE}$ 上采用不同的权重,对 $\widehat{ATE}\,(X_i = x)$ 进行加权平均。

2.匹配方法是先将样本根据可观测特征（控制变量）进行匹配，在观测特征达到均衡的基础上求解处置效应。**匹配方法是两步估计**，它允许我们先对可观测特征进行均衡性检验。**因此，使用匹配方法时应对数据是否满足重叠和是否均衡有明确的认识。**

总结

不同点

3.非饱和回归模型是将控制变量和处置变量一起纳入模型，**一步估计**出处置效应，对是否有重叠条件没有要求。**非饱和回归模型**可以通过**假设控制变量和观测结果的线性关系**去外推重叠区域外的“反事实结果”以达到估计处置效应的目的，其结果的合理性取决于**模型函数形式是否正确。**

4.匹配方法通过**同样特征**的干预组和控制组的观测结果均值相减得到处置效应。**可观测特征X对结果Y的影响通过均值相减消除了，其不需要假设特征是如何影响结果Y的**，因此**匹配是非参数方法**。回归方法则通过**线性函数LFP具体形式设定**假设特征变量X是如何影响结果Y的，因此**回归是参数估计方法**。