高级计量经济学

理论经济学博士课程 2023-2024

Lecture 14: Selection on Observables

黄嘉平

工学博士 经济学博士
深圳大学中国经济特区研究中心 讲师

办公室 粤海校区汇文楼1510
E-mail huangjp@szu.edu.cn
Website https://huangjp.com

用回归方法估计 ATE

Using Regression under Random Assignment

假设个体处理效应是恒定的（没有个体间差异），即 $Y_{1i} - Y_{0i} = \rho$。此时的估计目标就变成了参数 ρ。

考虑下面的回归模型

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

其中 $\alpha = E[Y_{0i}], \varepsilon_i = Y_{0i} - E[Y_{0i}]$。由此模型可得

$$E[Y_i^{\text{obs}} \mid W_i = 1] = \alpha + \rho + E[\varepsilon_i \mid W_i = 1]$$
$$E[Y_i^{\text{obs}} \mid W_i = 0] = \alpha + E[\varepsilon_i \mid W_i = 0]$$

取两式之差可得

$$\tau^{\text{diff}} = \rho + \text{选择偏差}$$

在随机分配机制下，选择偏差为零，$\tau^{\text{diff}} = \rho = \tau_{\text{ATE}} = \tau_{\text{ATET}}$。平均处理效应可以用回归系数的 OLS 估计量来估计，即 $\hat{\tau}_{\text{ATE}} = \hat{\tau}_{\text{ATET}} = \hat{\rho}_{\text{OLS}}$。

回归方法的优点是可以方便地加入其他控制变量，并可以很容易地拓展到多值处理变量。缺点是需要假设处理效应不存在个体差异。

干扰

Confounding

和回归分析类似，在潜在结果模型中，处理变量和潜在结果变量也可能受到其他因素的影响，我们称其为干扰（confounding）。

如果干扰以变量的形式存在，我们将其称为干扰变量（confounding variables, confounders）或者协变量（covariates），记为 $\boldsymbol{X}_i$。协变量的存在可能威胁到独立性假设 $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp W_i$ 的正当性，从而影响平均处理效应的估计。

当协变量可观测时，我们可以考虑下面的替代条件

- 条件独立假设（conditional independence assumption, CIA）：
$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp W_i \mid \boldsymbol{X}_i$$
- 条件均值独立假设（conditional mean independence, CMI）：
$$E[Y_{1i} \mid W_i, \boldsymbol{X}_i] = E[Y_{1i} \mid \boldsymbol{X}_i], \quad E[Y_{0i} \mid W_i, \boldsymbol{X}_i] = E[Y_{0i} \mid \boldsymbol{X}_i]$$

CMI 弱于 CIA，下面为了方便我们都假设 CIA。

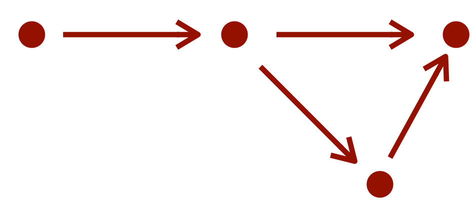
令 $\tau^{\text{diff}}(\boldsymbol{X}_i) = E[Y_i^{\text{obs}} \mid W_i = 1, \boldsymbol{X}_i] - E[Y_i^{\text{obs}} \mid W_i = 0, \boldsymbol{X}_i]$，则 $\tau^{\text{diff}}(\boldsymbol{X}_i)$ 是控制协变量后的平均处理效应。整体平均处理效应可以用 $\tau^{\text{diff}}(\boldsymbol{X}_i)$ 计算。

由于 CIA 和 CMI 基于协变量的可观测性，因此也被称为 selection on observables 条件。

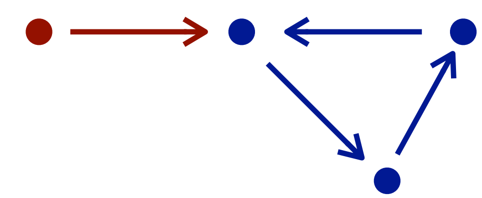
用有向无环图表达因果关系

Directed Acyclic Graph (DAG) Representation

Pearl (2000) 提倡用有向无环图（directed acyclic graph, DAG）表达潜在结果模型中的因果关系。



有向图（directed graph）



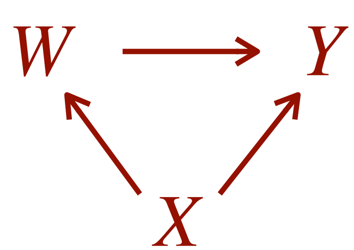
环（cycle）

我们可以用图中的节点代表变量，用边代表因果关系的方向。例如 $W \rightarrow Y$ 代表 W 引起了 Y 的变化（W causes Y），W 为因 Y 为果。

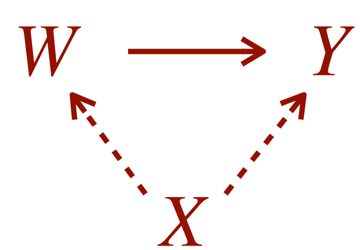
$$W \longrightarrow Y$$

注意：这里 W 的变化应是人为干预（intervention）引起的

如果存在同时影响 W 和 Y 的协变量 X，则 DAG 可以表达为



X 可观测



X 不可观测

注意：如果两个节点间没有边连接，则代表二者间**不存在**直接因果关系。

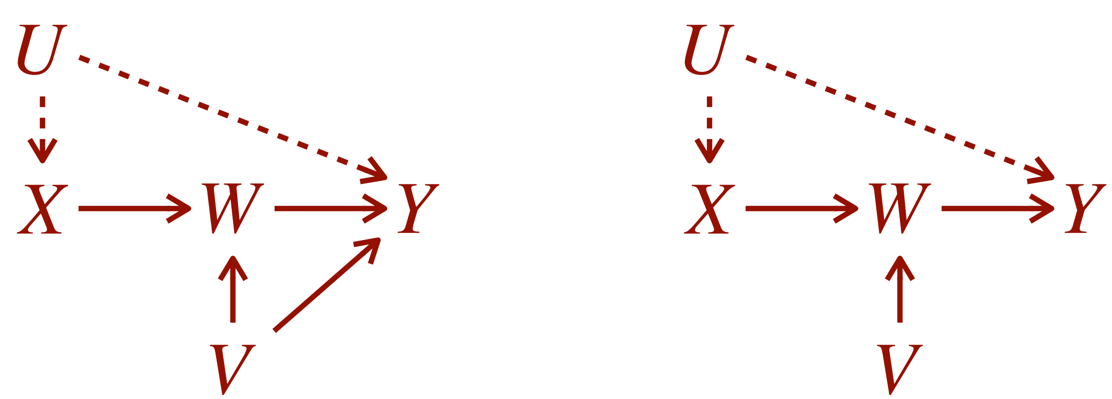
Pearl (2000, 2009). Causality: Models, Reasoning, and Inference, 1st & 2nd Edition. Cambridge University Press.

基于可观测协变量的识别

Identification under Selection on Observables

下面以奖学金和学习成绩间的因果效应问题为例。设处理变量 W 为奖学金获得情况，潜在结果 Y 为奖学金发放后的学生成绩，协变量包括 X（入学时的成绩），U（入学前的学术能力），V（学生的个人特征等）。

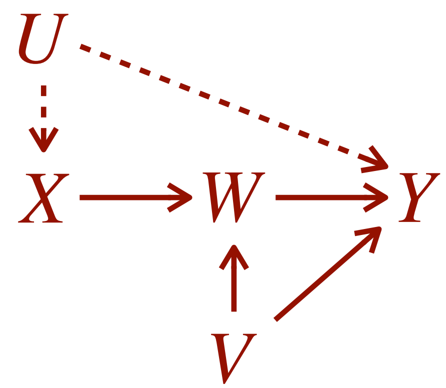
可以考虑右侧两种模型：

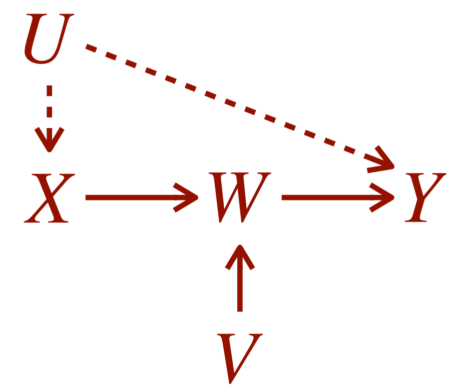


因果效应（直接或间接）可以用图中的路径（path）表达。例如 $X \to W \to Y$ 代表 X 通过 W 对 Y 产生因果效应。模型中是否存在干扰可以通过是否存在后门路径（backdoor path）判断

- 后门路径：连接处理变量 W 和结果变量 Y 的路径中，存在指向 W 的边。

图中的 $W \leftarrow X \leftarrow U \to Y$ 和 $W \leftarrow V \to Y$ 是后门路径。如果模型中存在后门路径，则可能存在干扰，使 τ^{diff} 包含选择偏差。





在 CIA 条件 $(Y_1, Y_0) \perp\!\!\!\perp W \mid X$ 成立时，我们可以通过控制协变量 X 阻断后门路径 $W \leftarrow X \leftarrow U \rightarrow Y$ 的发生。

路径中如果存在**不变的**变量，则该路径被阻断

另外，我们假设下面的条件成立：

- 共同支撑假设（common support assumption）或重叠假设（overlap）：$0 < \Pr(W = 1 \mid X) < 1$

此时，

$$
\begin{aligned}
E[Y_{1i} - Y_{0i} \mid X_i] &= E[Y_i^{\text{obs}} \mid W_i = 1, X_i] - E[Y_i^{\text{obs}} \mid W_i = 0, X_i] \\
&= \tau^{\text{diff}}(X_i)
\end{aligned}
$$

$$
\Rightarrow \quad
\begin{aligned}
\tau_{\text{ATE}} &= E_X\big[E[Y_{1i} - Y_{0i} \mid X_i]\big] = E[\tau^{\text{diff}}(X_i)] \\
\tau_{\text{ATET}} &= E_X\big[E[Y_{1i} - Y_{0i} \mid X_i] \mid W_i = 1\big] = E[\tau^{\text{diff}}(X_i) \mid W_i = 1]
\end{aligned}
$$

在上面的两个模型中，后门路径 $W \leftarrow X \leftarrow U \rightarrow Y$ 可以通过控制 X 或 U 阻断，但是 $W \leftarrow V \rightarrow Y$ 只能通过控制 V 阻断。

如果 V 不可观测，则上图左侧的模型无法对因果效应进行识别，右侧的则可以。

以上例子取自 Abadie & Catteneo (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10, 465-503.

选择偏差的分解*

Decomposition of Selection Bias*

Heckman et al. (1998) 对选择偏差 $E[Y_{0i} \mid W_i = 1] - E[Y_{0i} \mid W_i = 0]$ 进行了分解。他们得出

$$\text{选择偏差} = B_1 + B_2 + B_3$$

- B_1：来源于弱重叠（weak overlap）的偏差。
  如果存在 $\Pr(W = 1 \mid X = x) = 0, \Pr(W = 0 \mid X = x) > 0$ 或 $\Pr(W = 0 \mid X = x) = 0, \Pr(W = 1 \mid X = x) > 0$ 的 x，此时针对这个 x 无法进行处理组与对照组间的比较。

- B_2：来源于弱平衡（weak balance）的偏差。
  如果存在 $\Pr(W = 1 \mid X = x) \neq \Pr(W = 0 \mid X = x)$ 的 x，则关于这个 x，处理组与对照组中的观测值数量不同，会造成估计偏差。

- B_3：来源于不可观测协变量的偏差（selection on unobservables）。
  如果存在不可观测的协变量，则 CIA 或 CMI 不成立。（类似于回归分析中的遗漏变量偏差）

Heckman, Ichimura, Smith, & Todd (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica*, 66:5, 1017-1098.

Methods under Selection on Observables

匹配估计量（加权）

Matching Estimator (Weighting)

假设只存在一个可观测协变量 X_i，且 X_i 取离散值 $x_1, \ldots, x_m$。同时假设 CIA 和共同支撑假设成立。
前面定义了 $\tau^{\text{diff}}(X_i) = E[Y_i^{\text{obs}} \mid W_i = 1, X_i] - E[Y_i^{\text{obs}} \mid W_i = 0, X_i]$，因此 ATE 和 ATET 可以表达为

$$\begin{aligned}
\tau_{\text{ATE}} &= E[Y_{1i} - Y_{0i}] = E\big[E[Y_{1i} - Y_{0i} \mid X_i]\big] \\
&= E[\tau^{\text{diff}}(X_i)] = \sum_{k=1}^{m} \tau^{\text{diff}}(x_k) \Pr(X_i = x_k), \\
\tau_{\text{ATET}} &= E[Y_{1i} - Y_{0i} \mid W_i = 1] = E\big[E[Y_{1i} - Y_{0i} \mid X_i] \mid W_i = 1\big] \\
&= E[\tau^{\text{diff}}(X_i) \mid W_i = 1] = \sum_{k=1}^{m} \tau^{\text{diff}}(x_k) \Pr(X_i = x_k \mid W_i = 1)
\end{aligned}$$

根据等式最右侧的表达式可知，平均处理效应是条件均值之差 $\tau^{\text{diff}}(X_i)$ 的加权平均。

样本中对应 $\tau^{\text{diff}}(x_k)$ 的是 $\hat{\tau}^{\text{diff}}(x_k) = \frac{1}{n_{1k}} \sum_{j=1}^{n_{1k}} y_{1j}^{\text{obs}} - \frac{1}{n_{0k}} \sum_{j=1}^{n_{0k}} y_{0j}^{\text{obs}}$，而用作权重的概率也可以由样本提供的经验分布计算。因此平均处理效应的估计量是

$$\hat{\tau}_{\text{ATE}} = \textstyle\sum_{k=1}^{m} \hat{\tau}^{\text{diff}}(x_k)\, p_k, \quad \hat{\tau}_{\text{ATET}} = \textstyle\sum_{k=1}^{m} \hat{\tau}^{\text{diff}}(x_k)\, p_{k|W_i=1}$$

文献中称这种估计量为匹配估计量（matching estimator）或 subclassification。

回归调整

Regression Adjustment

令 D_{ik} 为 $X_i = x_k$ 时取值为 1 的虚拟变量。如果假设个体处理效应恒定，则可以考虑下面的回归模型

$$Y_i^{\text{obs}} = \rho W_i + \sum_{k=1}^{m} \beta_k D_{ik} + \varepsilon_i$$

Angrist & Pischke (2009) 应用 FWL 定理得出 ρ_{OLS} 可以表达为

$$\rho_{\text{OLS}} = \sum_{k=1}^{m} \tau^{\text{diff}}(x_k)\,\omega_k, \quad \omega_k = \frac{\text{Var}[W_i \mid X_i = x_k]\,\Pr(X_i = x_k)}{\sum_{r=1}^{m} \text{Var}[W_i \mid X_i = x_k]\,\Pr(X_i = x_r)}$$

因此，$\hat{\rho}_{\text{OLS}}$ 是用条件方差进行加权的匹配估计量。只有当 $\tau^{\text{diff}}(x_k) = \tau^{\text{diff}}$ （不随 x_k 的取值而改变），或 $\omega_k = p_k$ 时，$\hat{\rho}_{\text{OLS}}$ 才能正确估计 ATE 和 ATET。然而实践中，这两个条件都很难成立。

因为 $W_i \in \{0,1\}$，其条件方差是

$$\text{Var}[W_i \mid X_i = x_k] = \Pr(W_i = 1 \mid X_i = x_k)\big(1 - \Pr(W_i = 1 \mid X_i = x_k)\big)$$

因此当 $\Pr(W_i = 1 \mid X_i = x_k) = \frac{1}{2}$ 时取最大值，此时回归赋予 $\tau^{\text{diff}}(x_k)$ 的权重最大。

Angrist & Pischke (2009). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.

匹配是相似个体间的比较

Dale, S. B. & Krueger, A. (2002). Estimating the payoff to attending a more selective college: an application of selecting on observables and unobservables. *QJE*, 117(4): 1491—1527.

TABLE 2.1
The college matching matrix

Applicant group	Student	Private: Ivy	Private: Leafy	Private: Smart	Public: All State	Public: Tall State	Public: Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

表取自 Angrist & Pischke (2015). *Mastering 'Metrics*. PUP.

读私立大学比读公立大学更能带来高收入吗?

均值的比较:

$$\overline{w}_{\text{Pri}} - \overline{w}_{\text{Pup}}$$
$$= \$92{,}000 - \$72{,}500$$
$$= \$19{,}500$$

$A \sim D$ 组分别代表志愿和录取结果相似的学生，从表中可以看出，C 组和 D 组只包含处理组和对照组其中之一。

基于 A 组和 B 组中的样本可得

$$\Delta\overline{w}_A = -\$5{,}000$$
$$\Delta\overline{w}_B = \$30{,}000$$

由此计算的匹配估计值为

$$\hat{\tau}_{\text{ATE}} = \frac{3}{5}\Delta\overline{w}_A + \frac{2}{5}\Delta\overline{w}_B$$
$$= \$9{,}000$$

匹配与回归

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000

如果用回归方法，可以考虑包含私立学校虚拟变量和 A 组虚拟变量的模型

$$W_i = \alpha + \beta P_i + \gamma A_i + u_i$$

此时，$\alpha = E[W_i \mid \text{Public}, B]$，$\alpha + \beta = E[W_i \mid \text{Private}, B]$，$\alpha + \gamma = E[W_i \mid \text{Public}, A]$，$\alpha + \beta + \gamma = E[W_i \mid \text{Private}, A]$。

平均处理效应为 $\tau_{\text{ATE}} = \beta$，其估计值是 $\hat{\tau}_{\text{ATE}} = \$10{,}000$。与匹配估计值的差异是因为回归用了不同的权重进行加权。

参军经历如何影响收入

Angrist (1998). Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants. Econometrica, 66(2): 249—288.

TABLE 3.3.1
Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

Race	Average Earnings in 1988–1991 (1)	Differences in Means by Veteran Status (2)	Matching Estimates (3)	Regression Estimates (4)	Regression Minus Matching (5)
Whites	14,537	1,233.4 (60.3)	−197.2 (70.5)	−88.8 (62.5)	108.4 (28.5)
Non-whites	11,664	2,449.1 (47.4)	839.7 (62.7)	1,074.4 (50.7)	234.7 (32.5)

Notes: Adapted from Angrist (1998, tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 nonwhites in the sample.

表取自 Angrist & Pischke (2009). *Mostly Harmless Econometrics*. PUP.

参军作为处理变量**不是随机分配**的！

个人意愿、身体素质、家庭背景等要素都能影响一个人是否志愿参军以及是否被录取。

控制变量包括：

- 种族
- 申请参军年份（1979-1982）
- 申请时的教育水平
- 适应性测验结果
- 出生年份

结果变量为1988-1991年间的税前收入

近邻匹配

Nearest-Neighbor Matching

前面介绍的匹配估计量是组间的匹配（控制协变量 $X_i = x_k$ 时，处理组与对照组间的均值比较）。匹配也可以在个体层面实现。

个体处理效应可以表达为

$$\tau_i = Y_{1i} - Y_{0i} = \begin{cases} Y_i^{\text{obs}} - Y_i^{\text{mis}} \text{ if } W_i = 1 \\ Y_i^{\text{mis}} - Y_i^{\text{obs}} \text{ if } W_i = 0 \end{cases}$$
$$= W_i(Y_i^{\text{obs}} - Y_i^{\text{mis}}) + (1 - W_i)(Y_i^{\text{mis}} - Y_i^{\text{obs}})$$

我们可以用下面的办法对 Y_i^{mis} 进行估计：

1. 针对每个处理组中的 i，从对照组中找到使 $\|X_i - X_j\|$ 最小的 j，并令 $\hat{Y}_i^{\text{mis}} = Y_{j(i)}^{\text{obs}}$；
2. 针对每个对照组中的 i，从处理组中找到使 $\|X_i - X_j\|$ 最小的 j，并令 $\hat{Y}_i^{\text{mis}} = Y_{j(i)}^{\text{obs}}$；
3. 计算个体处理效应的估计量 $\hat{\tau}_i = W_i(Y_i^{\text{obs}} - \hat{Y}_i^{\text{mis}}) + (1 - W_i)(\hat{Y}_i^{\text{mis}} - Y_i^{\text{obs}})$；
4. $\hat{\tau}_{\text{ATET}} = \frac{1}{N_1}\sum_{i:W_i=1}\hat{\tau}_i$，$\hat{\tau}_{\text{ATEC}} = \frac{1}{N_0}\sum_{i:W_i=0}\hat{\tau}_i$，$\hat{\tau}_{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}(N_1\hat{\tau}_{\text{ATET}} + N_0\hat{\tau}_{\text{ATEC}})$。

这种匹配方法被称为近邻匹配（nearest-neighbor matching）。其优点是可以直接对应协变量为连续值或多变量的情况。近邻匹配有很多变种，比如考虑可重复或不可重复抽样，用多个匹配值计算 $\hat{Y}_i^{\text{mis}}$（取其均值），用不同的距离函数等。

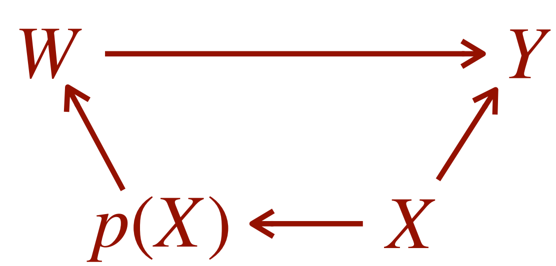
倾向得分匹配

Matching using Propensity Scores

倾向得分（propensity score）是个体接受处理的条件概率，即

$$p(X_i) = \Pr(W_i = 1 \mid X_i)$$

Rosenbaum & Rubin (1983) 证明，在 CIA 和共同支撑假设成立的情况下，我们可以将假设中基于协变量 X_i 的部分置换成倾向得分 $p(X_i)$，置换后的假设仍然成立。这说明，控制倾向得分和控制协变量都能够起到阻断后门路径的作用。



因此，我们可以将倾向得分运用到匹配中：

1. 针对 X_i 的每个取值 x_k，估计倾向得分 $\hat{p}(x_k)$；
2. 用 $\hat{p}(x_k)$ 替代 x_k 进行匹配。

注意：也有观点认为将倾向得分用于匹配会造成数据失衡、减小估计效率、增加估计偏差等问题。
King & Nielsen (2019). Why Propensity Scores Should Not Be Used For Matching. *Political Analysis*, 27:435-454.

Rosenbaum & Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:1, 41-55.

逆概率加权法

Inverse Probability Weighting

逆概率加权法（inverse probability weighting）是另一种利用倾向得分估计平均处理效应的方法。

Hirano et al. (2003) 分析了下面的加权估计量

$$\hat{\tau}^{\text{ipw}}_{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\frac{W_i Y_i^{\text{obs}}}{\hat{p}(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{(1-W_i)Y_i^{\text{obs}}}{1-\hat{p}(X_i)}$$

当某个 $\hat{p}(x_k)$ 的取值不等于 1/2 时，说明该 x_k 在处理组和对照组中存在弱平衡，可能导致估计偏差。加权估计量的目的就是调整这种弱平衡带来的偏差。

用 $1/\hat{p}(X_i)$ 做权重的一个缺点是权重之和不等于 1。如果将权重之和调整为 1，则对应的 $\hat{\tau}^{\text{ipw}}_{\text{ATE}}$ 可以通过加权回归求得。

Hirano, Imbens, & Ridder (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica, 71:4, 1161-1189.

更多参考文献

1. Morgan, S. L. & Winship, C. (2015).
   Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition. Cambridge University Press.

2. Pearl, J. (2009).
   Causality: Models, Reasonings, and Inference, 2nd Edition. Cambridge University Press.

3. Pearl, J., Glymour, M., & Jewell, N. P. (2016).
   Causal Inference in Statistics: A Primer. Wiley.

4. Cunningham, S. (2021).
   Causal Inference: The Mixtape. Yale University Press.
   https://mixtape.scunning.com/

5. Hernán, M. A. & Robins, J. M. (2020).
   Causal Inference: What If. CRC Press.
   https://www.hsph.harvard.edu/miguel-hernan/wp-content/uploads/sites/1268/2024/04/hernanrobins_WhatIf_26apr24.pdf

6. Huntington-Klein, N. (2022).
   The Effect: An Introduction to Research Design and Causality. CRC Press.
   https://theeffectbook.net/