定量社会科学的因果推断

Causal Inference in Quantitative Social Sciences

江　艇
中国人民大学经济学院

Last updated: March 7, 2021

第二类识别假设

- 第二类识别假设：分配机制不取决于潜在结果。

$$\Pr(D=1|X,Y^0,Y^1)=\Pr(D=1|X)$$

- 第二类识别假设（重新表述）：给定可观测变量，潜在结果均值独立于处理状态。

Assumption ID.2: $\mathbb{E}\left(Y^d|D,X\right)=\mathbb{E}\left(Y^d|X\right),\ d=0,1$

等价地，

$$\mathbb{E}(Y^0|D=1,X)=\mathbb{E}(Y^0|D=0,X)$$
$$\mathbb{E}(Y^1|D=1,X)=\mathbb{E}(Y^1|D=0,X)$$

- 此时**可观测变量相同条件下的组间均值差异**能够识别条件平均处理效应 $\tau(x)$, $\tau_1(x)$ 和 $\tau_0(x)$.

$$
\begin{aligned}
&\mathbb{E}(Y|X=x, D=1) - \mathbb{E}(Y|X=x, D=0) \\
=& \mathbb{E}(Y^1|X=x, D=1) - \mathbb{E}(Y^0|X=x, D=0) \\
=& \mathbb{E}(Y^1|X=x) - \mathbb{E}(Y^0|X=x) \\
=& \mathbb{E}(Y^1 - Y^0|X=x) \\
=& \tau(x)
\end{aligned}
$$

要想识别 $\tau_1(x)$，需要用到 $\mathbb{E}(Y^0|D=1, X) = \mathbb{E}(Y^0|D=0, X)$.
要想识别 $\tau_0(x)$，需要用到 $\mathbb{E}(Y^1|D=1, X) = \mathbb{E}(Y^1|D=0, X)$.
证明过程类似。因此

$$\tau(x) = \tau_1(x) = \tau_0(x)$$

- 进而能够识别平均处理效应

$$\tau = \mathbb{E}_X \left[\tau(x)\right]$$
$$\tau_1 = \mathbb{E}_X \left[\tau_1(x)|D=1\right]$$
$$\tau_0 = \mathbb{E}_X \left[\tau_0(x)|D=0\right]$$

请注意，$\tau \neq \tau_1 \neq \tau_0$，因为 $F_X(x) \neq F_{X|D=1}(x) \neq F_{X|D=0}(x)$.

$$\begin{aligned}\tau_1 &= \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=1\right) - \mathbb{E}\left(Y|X, D=0\right)|D=1\right] \\ &= \mathbb{E}\left(Y|D=1\right) - \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=0\right)|D=1\right]\end{aligned}$$

$$\begin{aligned}\tau_0 &= \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=1\right) - \mathbb{E}\left(Y|X, D=0\right)|D=0\right] \\ &= \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=1\right)|D=0\right] - \mathbb{E}\left(Y|D=0\right)\end{aligned}$$

$$\begin{aligned}\tau &= \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=1\right) - \mathbb{E}\left(Y|X, D=0\right)\right] \\ &= \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=1\right)\right] - \mathbb{E}_X \left[\mathbb{E}\left(Y|X, D=0\right)\right]\end{aligned}$$

而简单的组间均值比较的表达式却是

$$\begin{aligned}\tau &= \mathbb{E}\left(Y|D=1\right) - \mathbb{E}\left(Y|D=0\right) \\ &= \mathbb{E}_X\left[\mathbb{E}\left(Y|X,D=1\right)|D=1\right] - \mathbb{E}_X\left[\mathbb{E}\left(Y|X,D=0\right)|D=0\right]\end{aligned}$$

- 现在的问题是：**What is the best way to "condition" on X?**

非参数方法：匹配

- 一个准确匹配 (exact matching) 的假想例子

id	treat	x	y
1	1	1	y_1
2	1	1	y_2
3	1	2	y_3
4	1	3	y_4
5	0	1	y_5
6	0	2	y_6
7	0	2	y_7
8	0	2	y_8

	treat	control
$x = 1$	y_1, y_2	y_5
$x = 2$	y_3	y_6, y_7, y_8
$x = 3$	y_4	

- ATT 的估计

$$\begin{aligned}
\hat{\tau}_1(x=1) &= \left(\frac{y_1+y_2}{2}\right) - y_5 \\
\hat{\tau}_1(x=2) &= y_3 - \left(\frac{y_6+y_7+y_8}{3}\right) \\
\hat{\tau}_1(x=3) &\text{ 无法估计}
\end{aligned}$$

$$\begin{aligned}
\hat{\tau}_1 &= \hat{\tau}_1(x=1)\times\frac{2}{3} + \hat{\tau}_1(x=2)\times\frac{1}{3} \\
&= \left(\frac{y_1+y_2+y_3}{3}\right) - \left(y_5\times\frac{2}{3} + \left(\frac{y_6+y_7+y_8}{3}\right)\times\frac{1}{3}\right) \\
&= \frac{1}{3}\left((y_1-y_5)+(y_2-y_5)+\left(y_3-\frac{y_6+y_7+y_8}{3}\right)\right) \\
&= \frac{1}{n_T}\sum_{i\in T\cup C_T}(2D_i-1)w_i y_i
\end{aligned}$$

- ATU 的估计

$$\begin{aligned}
\hat{\tau}_0(x=1) &= \left(\frac{y_1+y_2}{2}\right) - y_5 \\
\hat{\tau}_0(x=2) &= y_3 - \left(\frac{y_6+y_7+y_8}{3}\right) \\
\hat{\tau}_0(x=3) &\text{ 无法估计}
\end{aligned}$$

$$\begin{aligned}
\hat{\tau}_0 &= \hat{\tau}_0(x=1)\times\frac{1}{4} + \hat{\tau}_0(x=2)\times\frac{3}{4} \\
&= \left(\left(\frac{y_1+y_2}{2}\right)\times\frac{1}{4} + y_3\times\frac{3}{4}\right) - \left(\frac{y_5+y_6+y_7+y_8}{4}\right) \\
&= \frac{1}{4}\left[\left(\frac{y_1+y_2}{2}-y_5\right) + (y_3-y_6) + (y_3-y_7) + (y_3-y_8)\right]
\end{aligned}$$

- ATE 的估计

$$\begin{aligned}\hat{\tau}(x=1) &= \left(\frac{y_1+y_2}{2}\right) - y_5 \\ \hat{\tau}(x=2) &= y_3 - \left(\frac{y_6+y_7+y_8}{3}\right) \\ \hat{\tau}(x=3) &\text{ 无法估计}\end{aligned}$$

$$\begin{aligned}\hat{\tau} &= \hat{\tau}(x=1)\times\frac{3}{7} + \hat{\tau}(x=2)\times\frac{4}{7} \\ &= \left(\left(\frac{y_1+y_2}{2}\right)\times\frac{3}{7} + y_3\times\frac{4}{7}\right) - \left(y_5\times\frac{3}{7} + \left(\frac{y_6+y_7+y_8}{3}\right)\times\frac{4}{7}\right) \\ &= \frac{1}{7}\left[(y_1-y_5) + (y_2-y_5) + \left(y_3 - \frac{y_6+y_7+y_8}{3}\right)\right. \\ &\quad \left. + \left(\frac{y_1+y_2}{2} - y_5\right) + (y_3-y_6) + (y_3-y_7) + (y_3-y_8)\right] \\ &= \hat{\tau}_1\times\frac{3}{7} + \hat{\tau}_0\times\frac{4}{7}\end{aligned}$$

参数方法：含控制变量的线性回归

- 假定 $\mathbb{E}(Y|D=1,X)$ 和 $\mathbb{E}(Y|D=0,X)$ 的函数形式，然后使用线性回归分别对处理组和控制组进行估计。

$$\hat{\mu}_1(X) = \mathbb{E}(Y|\widehat{D=1},X)$$
$$\hat{\mu}_0(X) = \mathbb{E}(Y|\widehat{D=0},X)$$
$$\hat{\tau}(X) = \hat{\tau}_1(X) = \hat{\tau}_0(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$$
$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$
$$\hat{\tau}_1 = \frac{1}{n_T}\sum_{i\in T}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$
$$\hat{\tau}_0 = \frac{1}{n_C}\sum_{i\in C}[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)]$$

- 常见情形一：

$$\mathbb{E}(Y|D=1,X)=\alpha_1+\gamma X$$
$$\mathbb{E}(Y|D=0,X)=\alpha_0+\gamma X$$

即

$$\mathbb{E}(Y|X)=\alpha_0+(\alpha_1-\alpha_0)D+\gamma X$$

可以通过 Y 对 D 和 X 的线性回归进行估计

$$Y_i=\hat{\alpha}_0+(\hat{\alpha}_1-\hat{\alpha}_0)\,D_i+\hat{\gamma}X_i+e_i$$

因此

$$\hat{\tau}(X)=(\hat{\alpha}_1+\hat{\gamma}X)-(\hat{\alpha}_0+\hat{\gamma}X)=\hat{\alpha}_1-\hat{\alpha}_0$$
$$\hat{\tau}=\hat{\tau}_1=\hat{\tau}_0=\hat{\alpha}_1-\hat{\alpha}_0$$

以 D 为核心解释变量，X 为控制变量的多元线性回归，D 的斜率系数估计即为对 ATE, ATT 和 ATU 的估计。此时隐含了同质处理效应的假定。

- 常见情形二：当 X 为离散变量时，假定其有 $x_1, \ldots, x_R$ 种不同取值，构造一组虚拟变量 $W_r = \mathbb{1}(X = x_r)$.

$$\mathbb{E}(Y|D=1,X) = \alpha_1 + \sum_{r=2}^{R} \gamma_r W_r$$

$$\mathbb{E}(Y|D=0,X) = \alpha_0 + \sum_{r=2}^{R} \gamma_r W_r$$

即

$$\mathbb{E}(Y|X) = \alpha_0 + (\alpha_1 - \alpha_0)D + \sum_{r=2}^{R} \gamma_r W_r$$

可以通过 Y 对 D 和 $W_r (r = 2, \ldots, R)$ 的线性回归进行估计

$$Y_i = \hat{\alpha}_0 + (\hat{\alpha}_1 - \hat{\alpha}_0) D_i + \sum_{r=2}^{R} \hat{\gamma}_r W_r + e_i$$

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

此时也隐含了同质处理效应的假定。

- 常见情形三：

$$\mathbb{E}(Y|D=1,X)=\alpha_1+\gamma_1 X$$
$$\mathbb{E}(Y|D=0,X)=\alpha_0+\gamma_0 X$$

即

$$\mathbb{E}(Y|X)=\alpha_0+(\alpha_1-\alpha_0)D+\gamma_0 X+(\gamma_1-\gamma_0)D\cdot X$$

可以通过 Y 对 D, X 及其交互项的线性回归进行估计

$$Y_i=\hat{\alpha}_0+(\hat{\alpha}_1-\hat{\alpha}_0)\,D_i+\hat{\gamma}_0 X_i+(\hat{\gamma}_1-\hat{\gamma}_0)\,D_i\cdot X_i+e_i$$

因此

$$\tau(X_i)=\tau_1(X_i)=\tau_0(X_i)=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\,X_i$$

$$\hat{\tau}=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\,\bar{X}$$
$$\hat{\tau}_1=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\,\bar{X}_{D=1}$$
$$\hat{\tau}_0=(\hat{\alpha}_1-\hat{\alpha}_0)+(\hat{\gamma}_1-\hat{\gamma}_0)\,\bar{X}_{D=0}$$

此时隐含了异质处理效应的特定形式（条件平均处理效应随 X 线性变化）假定。

- 当 X 为离散变量时，可以把常见情形二和常见情形三结合起来，例如

$$\mathbb{E}(Y|D=1,X)=\beta_1\mathbb{1}(X=1)+\beta_2\mathbb{1}(X=2)+\beta_3\mathbb{1}(X=3)$$
$$\mathbb{E}(Y|D=1,X)=\gamma_1\mathbb{1}(X=1)+\gamma_2\mathbb{1}(X=2)+\gamma_3\mathbb{1}(X=3)$$

即

$$\begin{aligned}\mathbb{E}(Y|D,X)=&\gamma_1\mathbb{1}(X=1)+\gamma_2\mathbb{1}(X=2)+\gamma_3\mathbb{1}(X=3)\\&+(\beta_1-\gamma_1)D\cdot\mathbb{1}(X=1)\\&+(\beta_2-\gamma_2)D\cdot\mathbb{1}(X=2)\\&+(\beta_3-\gamma_3)D\cdot\mathbb{1}(X=3)\end{aligned}$$

这一模型被称作饱和模型 (saturated model)，此时模型的线性形式不再具有限制性，它等价于匹配方法。

$$\tau(X=k)=\beta_k-\gamma_k$$

$$\tau_1=\frac{1}{n_T}\sum_{i\in T}\left[(\beta_1-\gamma_1)\cdot\mathbb{1}(X_i=1)+(\beta_2-\gamma_2)\cdot\mathbb{1}(X_i=2)+(\beta_3-\gamma_3)\cdot\mathbb{1}(X_i=3)\right]$$

$$\tau_0=\frac{1}{n_C}\sum_{i\in C}\left[(\beta_1-\gamma_1)\cdot\mathbb{1}(X_i=1)+(\beta_2-\gamma_2)\cdot\mathbb{1}(X_i=2)+(\beta_3-\gamma_3)\cdot\mathbb{1}(X_i=3)\right]$$

$$\tau=\frac{1}{n}\sum_{i=1}^{n}\left[(\beta_1-\gamma_1)\cdot\mathbb{1}(X_i=1)+(\beta_2-\gamma_2)\cdot\mathbb{1}(X_i=2)+(\beta_3-\gamma_3)\cdot\mathbb{1}(X_i=3)\right]$$

这种等价性只有当 X 离散时才能实现；当 X 连续时，模糊匹配 (fuzzy matching) 和基于函数形式的外推 (functional extrapolation) 会得到不同的结果。

尤其重要的常见情形二：控制固定效应

- 考虑如下模型，

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$$

得到 $\hat{\beta}_1^{OLS}$ 的一种等价做法是，Y_i 和 D_i 分别对 X_i 进行回归，保留残差，

$$\begin{aligned} Y_i &= c_1 + c_2 X_i + \tilde{Y}_i \\ D_i &= d_1 + d_2 X_i + \tilde{D}_i \end{aligned}$$

则

$$\hat{\beta}_1^{OLS} = \frac{\widehat{\text{Cov}}\left(Y, \tilde{D}\right)}{\widehat{\text{Var}}\left(\tilde{D}\right)} = \frac{\widehat{\text{Cov}}\left(\tilde{Y}, \tilde{D}\right)}{\widehat{\text{Var}}\left(\tilde{D}\right)}$$

其中后者还能给出正确的标准误。

- 要想准确估计一个解释变量的系数，这个解释变量必须具有足够的变动性 (variation)。例如，在教育回报率的例子中，

$$\texttt{wage}_i = b_0 + b_1 \cdot \texttt{edu}_i + e_i$$

为了估计 b_1，要求样本中存在各种不同受教育水平的个体。

- 现在方程右边加入性别控制变量（作用何在 ?），

$$\texttt{wage}_i = b_0 + b_1 \cdot \texttt{edu}_i + c \cdot \texttt{male}_i + e_i$$

这等价于

$$\widetilde{\texttt{wage}}_i = b_1 \cdot \widetilde{\texttt{edu}}_i + e_i$$

其中 $\widetilde{\texttt{wage}}_i$ 和 $\widetilde{\texttt{edu}}_i$ 分别是 $\texttt{wage}_i$ 和 $\texttt{edu}_i$ 对 $\texttt{male}_i$ 回归得到的残差。

- 易知

$$\widetilde{\texttt{wage}}_i = \begin{cases} \texttt{wage}_i - \overline{\texttt{wage}}_m & \text{if } \texttt{male}_i = 1 \\ \texttt{wage}_i - \overline{\texttt{wage}}_f & \text{if } \texttt{male}_i = 0 \end{cases}$$

$$\widetilde{\texttt{edu}}_i = \begin{cases} \texttt{edu}_i - \overline{\texttt{edu}}_m & \text{if } \texttt{male}_i = 1 \\ \texttt{edu}_i - \overline{\texttt{edu}}_f & \text{if } \texttt{male}_i = 0 \end{cases}$$

因此对于估计 b_1 而言真正有用的变动性不是受教育水平的整体变动性，而是其在性别组内的变动性。我们把控制 $\texttt{male}_i$ 的操作称为控制性别固定效应。

- 类似地，如果在方程右边加入地区控制变量，

$$\texttt{wage}_i = b_0 + b_1 \cdot \texttt{edu}_i + \sum_{r=2}^{R} c^r \cdot \texttt{region}_i^r + e_i$$

其中

$$\texttt{region}_i^r = \begin{cases} 1 & \text{if } i \text{ is from region } r \\ 0 & \text{otherwise} \end{cases}$$

等价于

$$\widetilde{\texttt{wage}}_i = b_1 \widetilde{\texttt{edu}}_i + e_i$$

其中 $\widetilde{\texttt{wage}}_i$ 和 $\widetilde{\texttt{edu}}_i$ 分别是 $\texttt{wage}_i$ 和 $\texttt{edu}_i$ 对所有地区虚拟变量回归得到的残差，也即去地区均值以后工资和受教育水平。控制了地区固定效应以后，真正有用的变动性是地区内部受教育水平的变动性。

插曲：非参数回归

- 含控制变量的线性回归是回归调整法 (regression adjustment) 的特例。我们既可以用参数方法也可以用非参数方法估计 $\hat{\mu}_1(X)$ 和 $\hat{\mu}_0(X)$.
- 核回归 (kernel regression).
  - 用 $X = x$ 附近的 Y 的均值近似 $X = x$ 点处 Y 的均值。
  - $K(\cdot)$ 是核函数，本质就是权重函数，多表现为某个随机变量的密度函数形式。常见的核函数包括：
    - ▷ Uniform (rectangular) kernel:

$$K(u) = \frac{1}{2}\mathbb{1}\big(|u| < 1\big)$$

    - ▷ Triangular kernel:

$$K(u) = (1 - |u|)\mathbb{1}\big(|u| < 1\big)$$

▷ Epanechnikov kernel:

$$K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\big(|u|<1\big)$$

▷ Gaussian kernel:

$$K(u) = \phi(u)$$

– 核估计量

$$\mathbb{E}(\widehat{Y|X=x}) = \frac{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i-x}{h}\right)}$$

当 $n \to \infty$ 时，$h \to 0$，称为带宽 (bandwidth)。

– 核估计量是 Y 的加权平均，距离 $X = x$ 越近的观测值被赋予越高权重。

- 局部线性回归 (local linear regression).
  - 考察如下 $X = x$ 附近处的回归

$$\min_{a,b} \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right)$$

$$\widehat{\mathbb{E}(Y|X = x)} = \hat{a}$$

  - 基本想法是，如果条件期望函数是平滑的，那么在 $X = x$ 附近就可以用线性函数近似。从这个视角，可以把核回归看作是局部常数项回归 (local constant regression)。局部线性回归通常比局部常数项回归偏误更小，尤其当 x 取边界值时。[1]

[1] 在回归断点设计中这种情形是重要的。

示例 5. 私立学校的经济回报 (Dale and Krueger, 2002, *QJE*).

The college matching matrix

Applicant group	Student	Private			Public			1996 earnings
		Ivy	Leafy	Smart	All State	Tall State	Altered State	
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.

- 非参数估计结果：
  - $\tau(A) = -5,\ \tau(B) = 30.$
  - $\tau = (-5) \times (3/5) + 30 \times (2/5) = 9.$
  - $\tau_1 = (-5) \times (2/3) + 30 \times (1/3) = 20/3.$
  - $\tau_0 = (-5) \times (1/2) + 30 \times (1/2) = 25/2.$
- 参数估计结果：
  - 全样本的组间均值差异：$\tau_{\mathsf{MD}} = 19.5.$
  - 匹配样本的组间均值差异：$\tau_{\mathsf{MDM}} = 20.$
  - 全样本的控制回归：$\tau_{\mathsf{R}} = 10.$
  - 匹配样本的控制回归：$\tau_{\mathsf{RM}} = 10.$
  - 饱和模型
  - 匹配样本的加权简单回归

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

匹配与线性回归的比较

- 匹配作为控制 X 的非参数方法，无需假定同质处理效应或处理效应关于 X 的函数形式。
- 但是，即使 $\tau(x)$ 各不相同并且高度非线性，是否有可能线性回归一致地估计了 $\tau = \mathbb{E}_X[\tau(x)]$ 呢？
- 更重要的是,如前所述,匹配只是一种条件策略 (conditioning strategy),而识别假设不会因为选择了一种特定的条件策略而变得更可信；换句话说,
- **启示一**：**匹配的识别假设和线性回归的识别假设是一样的**——都是假设**ID.2**。事实上，当关于 $\mathbb{E}(Y^1|X)$ 和 $\mathbb{E}(Y^0|X)$ 的函数形式假设——即关于 $\mathbb{E}(Y|D=1,X)$ 和 $\mathbb{E}(Y|D=0,X)$ 的函数形式假设——正确时，可以证明，假设**LS.2**和假设**ID.2**是等价的。因此这两种方法的识别力度 (identification power) 是一样的，**认为匹配方法能够解决线性回归所无法解决的内生性问题，是大错特错**。

- 所以我们需要深入探究**在异质处理效应前提下**，参数方法何时产生偏误。但必须牢记，
- **启示二**：**对研究情境的充分认识和对协变量的妥善选择，以及在此基础上的因果识别工作，要优先于条件策略的选择。**
- 要保证参数方法能够得到平均处理效应的一致估计，下面两个条件中必须至少满足一个 (Imbens, 2015, *JHR*)：
  - 关于条件期望的函数形式假设是正确的。
  - 处理组和控制组的 X 分布相同。

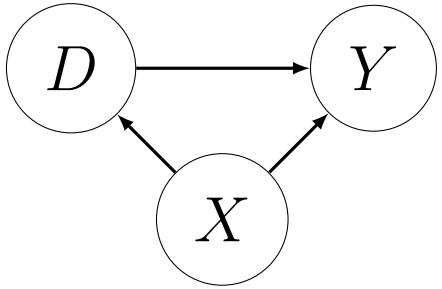
  否则，存在**由模型误设 (misspecification) 引起的外推偏误 (extrapolation bias)**。
- 通常而言，当处理组和控制组规模很不均衡时，其协变量分布往往也差异很大，此时可以考虑使用匹配方法。

示例 6. 评估职业培训项目的效果 (Lalonde, 1986, *AER*; Dehejia and Wahba, 1999, *JASA*, 2002, *REStat*).

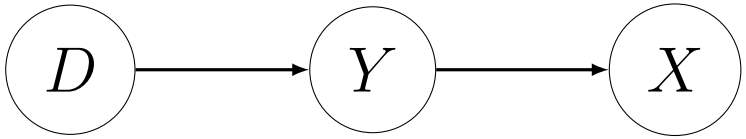
- 评估美国 1970 年代中期的某个职业培训项目。该项目是一项实地实验，招募一批劳动力市场上的弱势群体（戒毒瘾者、有犯案前科者、辍学者等），将其分为处理组和控制组，处理组个体可以获得 9-18 个月的工作机会。74、75 年为干预前，78 年为干预后。
- Lalonde (1986) 最先使用了这组数据，他先用处理组和控制组数据估计了这一职业培训项目的效果，然后用 PSID（收入动态面板调查）中的抽样数据代替控制组数据再次进行了估计，说明两者的差异。
- Dehejia and Wahba (1999, 2002) 发现，如果采用倾向得分匹配方法，那么即使用调查数据而非实际的控制组数据作为控制组，也能得到合意的结果。
- 实验组的干预前结果和调查数据控制组的干预前结果有很大差异。

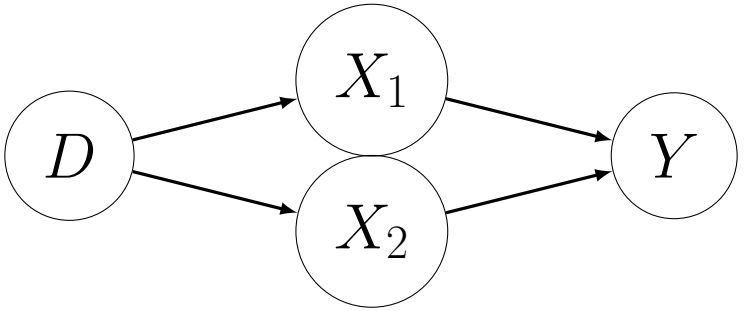
匹配方法实操

- 选择协变量 X
  - 必须控制

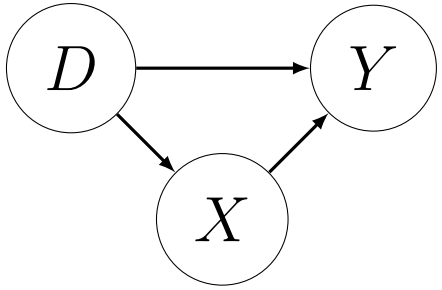


  - 必须不能控制





  - 只有当关心直接效应时才控制（常常是“坏控制 (bad control)”）



- 匹配方法的分类
  - 以协变量 X 本身作为匹配对象的方法称为协变量匹配 (covariate matching)。当协变量为离散变量时，可以进行准确匹配。当用于匹配的协变量过多时，会遇到维度的诅咒。当匹配协变量为连续变量时，准确匹配无法实现。
  - 近似匹配。一种想法是构造处理组个体协变量向量和控制组个体协变量向量的距离指标，例如马氏距离 (Mahalanobis distance)，为处理组个体匹配到这一距离指标最小的控制组个体。

$$\min_{c \in C} (\mathbf{X}_c - \mathbf{X}_t)' \Omega(\mathbf{X}_t)^{-1} (\mathbf{X}_c - \mathbf{X}_t)$$

    其中 $\Omega(\cdot)$ 是样本方差-协方差矩阵。
  - 另一种想法是去估计倾向得分（注意，倾向得分可以模型化为匹配协变量的非线性函数），然后直接用倾向得分的估计值进行近似匹配，这就是第二种主要的匹配方法：倾向得分匹配 (propensity score matching, PSM)。

- 是否放回 (replacement)：匹配过程中个体可否被重复使用，涉及 bias 和 variance 的权衡。
- 一配多时可以选择固定数目的匹配个体——最近邻 k 匹配，也可以选择距离指标在一定范围内的匹配个体——半径 (radius)/卡钳 (caliper) 匹配。
- 核 (kernel) 匹配：一配多计算匹配个体均值时，其权重是否以及如何随距离衰减。

$$\hat{E}(Y|X_t, D=0) = \sum_{c \in C} \frac{K\left(\frac{X_c - X_t}{h}\right) Y_c}{K\left(\frac{X_c - X_t}{h}\right)}$$

- 在大多数研究情境中，处理组规模往往小于控制组规模，能够为处理组个体找到匹配质量良好的控制组个体，因此 ATT 是主要的研究目标。

- 协变量匹配时，因为匹配是不精确的，需要对估计量进行偏误修正。以 ATE 为例，

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i^1 - \hat{Y}_i^0\right)$$

$$\hat{Y}_i^1 = D_i Y_i + (1 - D_i)\frac{1}{\|T_i\|}\sum_{t\in T_i}\left\{Y_t + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_t)\right\}$$

$$\hat{Y}_i^0 = (1 - D_i) Y_i + D_i\frac{1}{\|C_i\|}\sum_{c\in C_i}\left\{Y_c + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_c)\right\}$$

其中 $\hat{\mu}_d(X)$ 是 $\mu_d(X) = E(Y|X, D = d)$ 的 OLS 估计。

- 估计倾向得分
  - $h(\cdot)$ 形式的选择不是为了提供因果解释，而只是为了更好地近似条件期望函数。以 logit 为例，

$$P(D=1|X) \triangleq \pi(X) = \frac{\exp(h(X)'\gamma)}{1+\exp(h(X)'\gamma)}$$

  - 基于逐步回归的方程设定搜索：
    1. 先根据经验决定有哪些变量必须加入（如果没有先验信息，就只加入截距项）。
    2. 然后逐一加入其余所有一次项，对新增项的系数显著性进行 likelihood ratio test，统计量数值最大的一次项加入 $h(\cdot)$.
    3. 对余下的一次项重复步骤 2，直到本轮新增项系数的检验统计量最大值低于临界值 $C_{lin} = 1$.
    4. 然后逐一加入所有二次项（包括平方项和交互项），进行类似于步骤 2-3 的操作，统计量临界值 $C_{qua} = 2.71$.
  - 根据最终确定的 $h(\cdot)$ 估计倾向得分。

- 检查匹配样本的平衡性。
  - 计算组间的标准化差异 (normalized difference)。

$$\Delta = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\left(s_C^2 + s_T^2\right)/2}}$$

$$s_C^2 = \frac{1}{N_C - 1}\sum_{c \in C}\left(X_c - \bar{X}_C\right)^2,\ s_T^2 = \frac{1}{N_T - 1}\sum_{t \in T}\left(X_t - \bar{X}_T\right)^2$$

  - 这个统计量和检验两个样本均值是否相等的 t 统计量长得很像，但不建议使用后者。(为什么 ?)

$$t_{stat} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{s_C^2/N_C + s_T^2/N_T}}$$

  - 如果标准化差异很大，则要考虑删截样本，可以根据倾向得分进行删节。

- 协变量匹配：`teffects nnmatch` 和 `nnmatch` 都可以实现，并且都提供了详尽的匹配细节，但后者报告的标准误估计是错误的。
- 倾向得分匹配：`teffects psmatch` 和 `psmatch2` 都可以实现，并且都提供了详尽的匹配细节，但后者报告的标准误估计是错误的。
- 不论是协变量匹配还是倾向得分匹配，可能都有必要预先进行一步准确匹配，例如性别、行业等。
- 尽管有研究认为协变量匹配优于倾向得分匹配 (King and Nielsen, 2019, *Political Analysis*)，但倾向得分匹配仍然被广泛使用。当进行无放回的 $1:1$ 匹配时，一般不使用倾向得分匹配。

- **启示三：匹配的重点是找到“相像”的处理组和控制组个体，具体实现手段并不重要。**（言必称 PSM 只能反映对匹配方法的一知半解。）

示例 7. 信息不对称与融资决策 (Derrien and Kecskes, 2013, *JF*)

these restrictions on both groups of firms. We require that candidate control firms have the same two-digit SIC code as our treatment firms. We also require candidate control firms to be in the same total assets quintile, Q quintile, and cash flow quintile as our treatment firms.[6] We then retain candidate control firms that have the smallest difference in number of analysts compared to the corresponding treatment firms. We break any remaining ties based on the smallest differences in total assets, Q, and cash flow. To this end, we compute the difference between treatment firms and control firms for each of total assets, Q, and cash flow. We compute the rank of the difference for each of these three variables, and we compute the total rank across all three variables. We retain candidate control firms that have the lowest total rank.

- **启示四：要从两个层次理解匹配，作为数据预处理手段的匹配和作为非参数估计方法的匹配，前者远比后者重要。**