3.3异质性与非线性

梁思靖

2019.10.10

回归与匹配

- 回归和匹配都是用来控制协变量的研究策略
- 相同的核心假设：条件独立假设
- 回归是一种匹配估计量
- 一个例子：志愿服兵役对之后收入水平的影响
  D_i 表示个体i是否参军，协变量 X_i 是离散的

$$E(Y_i|D_i=1)-E(Y_i|D_i=0)$$
$$=E(Y_{1i}-Y_{0i}|D_i=1)+\{E(Y_{0i}|D_i=1)-E(Y_{0i}|D_i=0)\}$$

- 条件独立假设：$\{Y_{0i},Y_{1i}\}\perp D_i|X_i$

匹配的数学表达

- 定义处理效应： $\delta_{TOT} \equiv E(Y_{1i} - Y_{0i}|D_i = 1)$
$= E\{E(Y_{1i} - Y_{0ii}|D_i = 1)|D_i = 1\}$
$= E\{E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 1)|D_i = 1\}$
$= E\{E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 0)|D_i = 1\}$
$= E[\delta_X|D_i = 1]$

  其中 $\delta_X \equiv E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 0)$

- 离散形式： $\delta_{TOT} = \sum_x \delta_x P\,(X_i = x|D_i = 1)$

- 无条件平均处理效应：
$\delta_{ATE} = \{E(Y_{1i}|X_i, D_i = 1) - E(Y_{0i}|X_i, D_i = 0)\}$
$= \sum_x \delta_x P(X_i = x) = E[Y_{1i} - Y_{0i}]$

回归的数学表达

- $Y_i = \sum_x d_{ix}\alpha_x + \delta_R D_i + e_i$

其中d_{ix}是虚拟变量，当$X_i = x$时，$d_{ix} = 1$
α_x估计$X_i = x$对收入的影响，δ_R是回归估计量

- $\delta_R = \frac{cov(Y_i, \tilde{D}_i)}{V(\tilde{D}_i)}$

$$= \frac{E\left[\left(D_i - E\left[D_i|X_i\right]\right)Y_i\right]}{E\left[\left(D_i - E\left[D_i|X_i\right]\right)^2\right]} = \frac{E\left\{\left(D_i - E\left[D_i|X_i\right]\right)E[Y_i|D_i, X_i]\right\}}{E\left[\left(D_i - E\left[D_i|X_i\right]\right)^2\right]}$$

- 展开条件期望：$E[Y_i|D_i, X_i] = E[Y_i|D_i = 0, X_i] + \delta_X D_i$
- 分子化简为：$E\left\{\left(D_i - E\left[D_i|X_i\right]\right)E[Y_i|D_i, X_i]\right\}$

$$= E\left\{\left(D_i - E\left[D_i|X_i\right]\right)E[Y_i|D_i = 0, X_i]\right\}$$
$$+ E\left\{\left(D_i - E\left[D_i|X_i\right]\right)\delta_x D_i\right\}$$
$$= E\left\{\left(D_i - E\left[D_i|X_i\right]\right)^2 \delta_X\right\}$$

回归与匹配的联系

- $\delta_R = \frac{E\left\{\left(D_i - E[D_i|X_i]\right)^2 \delta_X\right\}}{E\left[\left(D_i - E[D_i|X_i]\right)^2\right]}$

$$= \frac{E\left\{E\left[\left(D_i - E[D_i|X_i]\right)^2|X_i\right]\delta_X\right\}}{E\left\{E\left[\left(D_i - E[D_i|X_i]\right)^2|X_i\right]\right\}} = \frac{E[\sigma_D^2(X_i)\delta_X]}{E[\sigma_D^2(X_i)]}$$

其中$\sigma_D^2(X_i)$是给定X_i下D_i的条件方差

$$\sigma_D^2(X_i) \equiv E\left[\left(D_i - E[D_i|X_i]\right)^2|X_i\right]$$

- 回归模型的参数是匹配模型参数的加权平均

回归与匹配的联系

- $\sigma_D^2(X_i) = P(D_i = 1|X_i)\big(1 - P(D_i = 1|X_i)\big)$
- 回归估计量：$\delta_R = \frac{E\left[\sigma_D^2(X_i)\delta_x\right]}{E\left[\sigma_D^2(X_i)\right]} = \frac{\sum_x \delta_x\left[P\left(D_i=1\middle|X_i=x\right)\left(1-P\left(D_i=1\middle|X_i=x\right)\right)\right]P(X_i=x)}{\sum_x\left[P\left(D_i=1\middle|X_i=x\right)\left(1-P\left(D_i=1\middle|X_i=x\right)\right)\right]P(X_i=x)}$
- 匹配估计量：$E(Y_{1i} - Y_{0i}|D_i = 1) = \sum_x \delta_x P\left(X_i = x|D_i = 1\right) = \frac{\sum_x \delta_x P\left(D_i=1\middle|X_i=x\right)P(X_i=x)}{\sum_x P\left(D_i=1\middle|X_i=x\right)P(X_i=x)}$

其中$\mathrm{P}(X_i = x|D_i = 1) = \frac{P(D_i = 1|X_i = x)P(X_i=x)}{P(D_i=1)}$

- $P\left(D_i = 1\middle|X_i = x\right) = \frac{1}{2}$时条件方差最大

回归与匹配

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

Race	Average earnings in 1988-1991	Differences in means by veteran status	Matching estimates	Regression estimates	Regression minus matching
	(1)	(2)	(3)	(4)	(5)
Whites	14537	1233.4	-197.2	-88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

- 表3.3.1的结果显示，回归和匹配两种方法得到的结果类似
- 回归和匹配的区别：处理效应 δ_X 加权平均的权重不同。匹配权重——处理组中协变量的分布，回归权重——条件方差
- 匹配法最大权重在概率最大的的个体处理效应组，回归最大权重在条件方差最大的个体处理效应组

有序处理和连续处理

- 假设表示处理水平的变量 S_i 是连续分布的随机变量
  条件期望函数 $h(t) \equiv E[Y_i|S_i = t]$，其导数 $h'(t)$
  用 S_i 对 Y_i 做回归有：
  $$\frac{cov(Y_i, S_i)}{V(S_i)} = \frac{E[Y_i(S_i - E[S_i])]}{E[S_i(S_i - E[S_i])]} = \frac{E[h(S_i)(S_i - E[S_i])]}{E[S_i(S_i - E[S_i])]}$$
- 由微积分基本定理，分子

$$E[h(S_i)(S_i - E[S_i])] = \int_{-\infty}^{\infty}\int_{-\infty}^{u} h'(t)(u - E[S_i])g(u)dudt$$

$$= \int_{-\infty}^{\infty} h'(t)\int_{-\infty}^{u} (u - E[S_i])g(u)dudt$$

$g(u)$ 是在 u 处 S_i 的 密度函数

- 内层积分定义为：

$$\mu_t \equiv \{E[S_i|S_i \geq t] - E[S_i|S_i < t]\}\{P(|S_i \geq t)[1 - P(|S_i \geq t)]\}$$

$$\frac{E[Y_i(S_i - E[S_i])]}{E[S_i(S_i - E[S_i])]} = \frac{\int h'(t)\,\mu_t\,dt}{\int \mu_t\,dt}$$

- 加入协变量：$\frac{E[Y_i(S_i - E[S_i|X_i])]}{E[S_i(S_i - E[S_i|X_i])]} = \frac{E[\int h'_X(t)\mu_{tX}\,dt]}{E[\int \mu_{tX}\,dt]}$

- $h'_X(t) = \frac{\partial E[Y_i|X_i, S_i=1]}{\partial t}$

- $\mu_{tX} \equiv \{E[S_i|X_i, S_i \geq t] - E[S_i|X_i, S_i < t]\}\left\{P\left(\left|S_i \geq t\right|X_i\right)\left[1 - P\left(|S_i \geq t|X_i\right)\right]\right\}$

用倾向评分控制协变量

- 定义协变量向量的值函数为倾向得分
  $p(X_i) \equiv E[D_i|X_i] = P[D_i = 1|X_i]$
- 定理3.3.1（倾向评分定理）：若条件独立假设
  $\{Y_{0i}, Y_{1i}\} \perp D_i|X_i$ 成立，则有 $\{Y_{0i}, Y_{1i}\} \perp D_i|p(X_i)$
- 证明：$P[D_i = 1|Y_{ji}, p(X_i)] = E[D_i|Y_{ji}, p(X_i)]$

$$= E\{E[D_i|Y_{ji}, p(X_i), X_i]|Y_{ji}, p(X_i)\}$$
$$= E\{E[D_i|Y_{ji}, X_i]|Y_{ji}, p(X_i)\}$$
$$= E\{E[D_i|X_i]|Y_{ji}, p(X_i)\}$$
$$= E\{p(X_i)|Y_{ji}, p(X_i)\}$$

$P[D_i = 1|Y_{ji}, p(X_i)]$不依赖于 $Y_{ji}, j = 0,1$

- 由定理3.3.1和条件独立假设有：
  $E(Y_{1i} - Y_{0i}|D_i = 1)$
  $= E\{E(Y_i|p(X_i), D_i = 1) - E(Y_i|p(X_i), D_i = 0)|D_i = 1\}$
- 倾向评分定理使用：
  1.用logit或probit等参数模型来估计 $p(X_i)$
  2.用匹配法估计处理效应
- 估计处理效应的方法：
  1.根据 $p(X_i)$ 分层，用每层的样本条件期望计算
  2.将具有相似 $p(X_i)$ 的处理个体和控制个体匹配

计算处理效应

- 计算处理效应的方法：

条件独立假设意味着：

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

- 无条件平均处理效应：

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

- 样本估计量：

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

- 回归方程的估计量：

$$\delta_R = \frac{cov(Y_i,\tilde{D}_i)}{V(\tilde{D}_i)} = \frac{E\left[\left(D_i - E[D_i|X_i]\right)Y_i\right]}{E\left[\left(D_i - E[D_i|X_i]\right)^2\right]} = \frac{E\left[\left(D_i - p(X_i)\right)Y_i\right]}{E\left[p(X_i)\left(1-p(X_i)\right)\right]}$$

- 加权平均估计类：$E\left\{g(X_i)\left[\frac{Y_iD_i}{p(X_i)} - \frac{Y_i(1-D_i)}{\left(1-p(X_i)\right)}\right]\right\}$

$g(X_i)$ 是一个权重函数，

令 $g(X_i) = 1$，平均处理

令 $g(X_i) = \frac{p(X_i)}{P(D_i=1)}$，匹配法估计量

令 $g(X_i) = \frac{p(X_i)\left(1-p(X_i)\right)}{E\left[p(X_i)\left(1-p(X_i)\right)\right]}$，回归估计量

倾向评分模型 VS. 回归

- 针对收入偏低人群的政府资助培训项目，低收入是项目参与者和其他人群唯一的区别
- 控制过去收入变化来估计该项目对参与者收入的因果效应

  被解释变量：1978年收入

  控制变量：人口统计学变量、1974、1975年收入

Table 3.3.2: Covariate means in the NSW and observational control samples

Variable	NSW		Full Samples		P-score Screened Samples	
	Treated	Control	CPS-1	CPS-3	CPS-1	CPS-3
	(1)	(2)	(3)	(4)	(5)	(6)
Age	25.82	25.05	33.23	28.03	25.63	25.97
Years of schooling	10.35	10.09	12.03	10.24	10.49	10.42
Black	0.84	0.83	0.07	0.20	0.96	0.52
Hispanic	0.06	0.11	0.07	0.14	0.03	0.20
Dropout	0.71	0.83	0.30	0.60	0.60	0.63
Married	0.19	0.15	0.71	0.51	0.26	0.29
1974 earninigs	2,096	2,107	14,017	5,619	2,821	2,969
1975 earnings	1,532	1,267	13,651	2,466	1,950	1,859
Number of Obs.	185	260	15,992	429	352	157

不同控制变量下回归估计值

Table 3.3.3: Regression estimates of NSW training effects using alternate controls

Specification	Full Samples			P-Score Screened Samples	
	NSW	CPS-1	CPS-3	CPS-1	CPS-3
	(1)	(2)	(3)	(4)	(5)
Raw Difference	1,794 (633)	-8,498 (712)	-635 (657)		
Demographic controls	1,670 (639)	-3,437 (710)	771 (837)	-3,361 (811) [139/497]	890 (884) [154/154]
1975 Earnings	1,750 (632)	-78 (537)	-91 (641)	no obs. [0/0]	166 (644) [183/427]
Demographics, 1975 Earnings	1,636 (638)	623 (558)	1,010 (822)	1,201 (722) [149/357]	1,050 (861) [157/162]
Demographics, 1974 and 1975 Earnings	1,676 (639)	794 (548)	1,369 (809)	1,362 (708) [151/352]	649 (853) [147/157]

- 用P得分值筛选的CPS-1样本估计结果和未筛选的CPS-3样本结果一样好
- 回归中使用正确控制变量可以很好剔除选择偏误

3.4 回 归 细 节

加权回归

- 当加权使得估计的数值更接近总体参数时，使用加权回归
- 异方差下使用加权回归的理由：对线性条件期望函数，存在异方差即条件方差函数 $E\left[e_i^2 \middle| X_i\right]$ 未必是常数，加权最小二乘法更精确

有限被解释变量和边际效应

- 被解释变量取值有限，线性回归模型不适宜，probit和Tobit等非线性模型更适合
- 例子：兰德健康保险实验（HIE），考察人们医疗是否考虑成本
  被解释变量：
  虚拟变量：个体在给定年份中是否有过医疗支出或是否住院
  非负变量：和医生面对面交流的次数、医疗总费用
- 控制组 $D_i = 0$：完全免费的医疗服务
  处理组 $D_i = 1$：支付一定费用的医疗服务（接受门诊治疗时，个人每年150美元，家庭每年450美元）

回归结果

Table 3.4.1: Average outcomes in two of the HIE treatment groups

Plan	Face-to-face visits	Outpatient Expenses (1984$)	Admissions	Prob. Any Medical (%)	Prob. Any Inpatient (%)	Total Expenses (1984$)
Free	4.55	340	.128	86.8	10.3	749
	(.168)	(10.9)	(.0070)	(.817)	(.45)	(39)
Individual Deductible	3.02	235	.115	72.3	9.6	608
	(.171)	(11.9)	(.0076)	(1.54)	(.55)	(46)
Deductible minus free	-1.53	-105	-0.013	-14.5	-0.7	-141
	(.240)	(16.1)	(.0103)	(1.74)	(.71)	(60)

- 伯努利实验：$E[Y_{1i} - Y_{0i}] = E[Y_{1i}] - E[Y_{0i}]$
  $= P[Y_{1i} = 1] - P[Y_{0i} = 1]$
- 是否发生医疗支出的变量实际表示为概率

- probit模型：假设个体是否参与由潜变量 Y_i^* 决定

$$Y_i^* = \beta_0^* + \beta_1^* D_i - v_i \ , v_i \sim N(0, \sigma_v^2)$$

潜在得分模型：$Y_i = 1[Y_i^* > 0]$

条件期望函数：

$$E(Y_i|D_i) = \Phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]$$

$$= \Phi\left[\frac{\beta_0^*}{\sigma_v}\right] + \left\{\Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_v}\right] - \Phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\} D_i$$

正数效应

- 医疗总费用：非负随机变量

$$E(Y_i|D_i) = E[Y_i|Y_i > 0, D_i]P\left[Y_i > 0\middle|D_i\right]$$

- 医疗费用差异：

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$
$$= E[Y_i|Y_i > 0, D_i = 1]P\left[Y_i > 0\middle|D_i = 1\right]$$
$$- E[Y_i|Y_i > 0, D_i = 0]P\left[Y_i > 0\middle|D_i = 0\right]$$
$$= \left\{P\left[Y_i > 0\middle|D_i = 1\right] - P\left[Y_i > 0\middle|D_i = 0\right]\right\}E[Y_i|Y_i > 0, D_i = 1]$$

参与效应

$$+ \{E[Y_i|Y_i > 0, D_i = 1] - E[Y_i|Y_i > 0, D_i = 0]\}P\left[Y_i > 0\middle|D_i = 0\right]$$

正数效应

正数效应

- 参与效应：医疗消费为正的个体在不同处理组时接受治疗的概率之差

  正数效应：给定个体接受医疗治疗，不同处理组之间平均医疗花费之差

- 分解正数效应：

$$
\begin{aligned}
&E[Y_i|Y_i>0,D_i=1]-E[Y_i|Y_i>0,D_i=0]\\
&=E[Y_{1i}|Y_{1i}>0]-E[Y_{0i}|Y_{0i}>0]\\
&=E[Y_{1i}-Y_{0i}|Y_{1i}>0]+E[Y_{0i}|Y_{1i}>0]\\
&-E[Y_{0i}|Y_{0i}>0]
\end{aligned}
$$

选择性偏误

- 正数效应的非因果性解决办法：删失回归（censored regression）
- Tobit模型：假设非参与者的潜在支出结果

$$Y_i = 1[Y_i^* > 0]Y_i^*$$

Y_i^* 是正态分布的潜在支出变量，可取负值。

$$E(Y_i|D_i) = \Phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma_v}\right][\beta_0^* + \beta_1^* D_i] + \sigma_v \phi\left[\frac{\beta_0^* + \beta_1^* D_i}{\sigma_v}\right]$$

- $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$

$$= \left\{\Phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_v}\right][\beta_0^* + \beta_1^*] + \sigma_v \phi\left[\frac{\beta_0^* + \beta_1^*}{\sigma_v}\right]\right\}$$

$$- \left\{\Phi\left[\frac{\beta_0^*}{\sigma_v}[\beta_0^*]\right] + \sigma_v \phi\left[\frac{\beta_0^*}{\sigma_v}\right]\right\}$$

协变量导致的非线性

- 非线性模型结果必须转化为边际效应才有用
- 边际效应：非线性模型中条件期望函数的变化

$$E\{E[Y_i|X_i,D_i=1]-E[Y_i|X_i,D_i=0]\}\ or\ E\left\{\frac{\partial E[Y_i|X_i,D_i]}{\partial D_i}\right\}$$

- 带有协变量的probit模型的条件期望函数：

$$E(Y_i|D_i)=\Phi\left[\frac{X_i'\beta_0^*+\beta_1^*D_i}{\sigma_v}\right]$$

- 平均处理效应：$E\left\{\Phi\left[\frac{X_i'\beta_0^*+\beta_1^*}{\sigma_v}\right]-\Phi\left[\frac{X_i'\beta_0^*}{\sigma_v}\right]\right\}$

  平均导数近似表达 $E\left\{\phi\left[\frac{X_i'\beta_0^*+\beta_1^*D_i}{\sigma_v}\right]\right\}\left(\frac{\beta_1^*}{\sigma_v}\right)$

- 非负有限解释变量的条件期望函数：$E(Y_i|X_i,D_i)=\Phi\left[\frac{X_i'\beta_0^*+\beta_1^*D_i}{\sigma_v}\right][X_i'\beta_0^*+\beta_1^*D_i]+\sigma_v\phi\left[\frac{X_i'\beta_0^*+\beta_1^*D_i}{\sigma_v}\right]$
- Tobit模型的边际效应的简单形式：

$$E\left\{\Phi\left[\frac{X_i'\beta_0^*+\beta_1^*D_i}{\sigma_v}\right]\right\}(\beta_1^*)$$

- 例子：抚养孩子对工作的影响
  被解释变量:工作时间、就业率
  表示抚养孩子的变量：是否有两个以上孩子的虚拟变量或孩子总数
  协变量:母亲年龄、第一胎生育时的年龄、种族虚拟变量以及母亲教育水平

各种估计方法结果比较

Table 3.4.2: Comparison of alternative estimates of the effect of childbearing on LDVs

		Right-hand side variable								
		More than two children					Number of children			
	Mean	OLS	Probit		Tobit		OLS	Probit MFX	Tobit MFX	
Dependent variable			Avg effect, full sample	Avg effect on treated	Avg effect, full sample	Avg effect on treated		Avg effect, full sample	Avg effect, full sample	Avg effect on treated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Full Sample										
Employment	.528	-.162	-.163	-.162	-	-	-.113	-.114	-	-
	(.499)	(.002)	(.002)	(.002)			(.001)	(.001)		
Hours worked	16.7	-5.92	-	-	-6.56	-5.87	-4.07	-	-4.66	-4.23
	(18.3)	(.074)			(.081)	(.073)	(.047)		(.054)	(.049)
Panel B: Non-white College Attendees over 30, first birth before age 20										
Employment	.832	-.061	-.064	-.070	-	-	-.054	-.048	-	-
	(.374)	(.028)	(.028)	(.031)			(.016)	(.013)		
Hours worked	30.8	-4.69	-	-	-4.97	-4.90	-2.83	-	-3.20	-3.15
	(16.0)	(1.18)			(1.33)	(1.31)	(.645)		(.670)	(.659)

各种估计方法结果比较

- 在被解释变量有限情况下使用非线性模型更好接近条件期望函数，但当考虑边际效应时，线性模型和非线性模型下的结论差别很小