计量经济学

自然实验和反事实估计框架

张晨峰

2016年6月15日

华东理工大学商学院

8. 自然实验和反事实估计框架

主要内容

- 理想的实验
- 回归与匹配
- 断点回归设计

8.1 理想的实验

医院能够使人变得更健康吗？

利用NHIS的数据，下面的表格给出了最近去过医院和没有去过医院的人的平均健康状况。

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

8.1 理想的实验

鲁宾因果框架(Rubin Causal Model)

个体健康状况的潜在结果

$$Y_i = \begin{cases} Y_{1i} & D_i = 1 \\ Y_{0i} & D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

对于个体来说，只能观测到Y_{1i}或Y_{0i}，所以可以理解为一个缺失数据问题。 平均处理效应(average treatment effect,ATE)为

$$\tau_{ATE} = E(Y_{1i} - Y_{0i})$$

处理的平均处理效应(average treatment effect on the treated,ATT)

$$\tau_{ATT} = E(Y_{1i} - Y_{0i}|D_i = 1)$$

鲁宾因果框架

把是否去医院接受治疗带来的不同结果进行简单比较

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

前半部分是处理的平均因果效应，后半部分是选择性偏误。 给定随机分配下D_i的独立性，我们可以对因果效应继续简化

$$E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1] = E[Y_{1i}-Y_{0i}|D_i=1] = E[Y_{1i}-Y_{0i}]$$

8.1 理想的实验

田纳西的师生比例改进计划(STAR)

这项实验将学生分配至三个处理组：小班、普通班及普通/助理班。对随机实验的第一个问题就是随机化是否成功地平滑了不同处理组间的各种特征。

Students who entered STAR in kindergarten

	Variable	Small	Regular	Regular/Aide	Joint P-value
1.	Free lunch	.47	.48	.50	.09
2.	White/Asian	.68	.67	.66	.26
3.	Age in 1985	5.44	5.43	5.42	.32
4.	Attrition rate	.49	.52	.53	.02
5.	Class size in kindergarten	15.10	22.40	22.80	.00
6.	Percentile score in kindergarten	54.70	48.90	50.00	.00

8.1 理想的实验

Table 2.2.2: Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

教育与收入的研究

经验研究中，教育水平和收入之间的因果联系告诉我们

- 如果人们在一个完美的受控实验中改变他的受教育水平，那么平均而言他们会赚到多少钱
- 或者如果人们随机选择受教育水平，从而使得他们在各方面都可比时，受教育水平的差异带来的收入水平的不同

条件独立假设(CIA)

实验保证了我们感兴趣的变量与潜在结果无关，从而使得被比较的组别之间是真正可比的。我们可以将这一概念推广到因果变量取值超过两个并且有一系列控制变量需要给定的更复杂的情况，来使得因果推断得以成立。这就带来了条件独立假设。

教育与收入的研究

$$E[Y_i|C_i=1]-E[Y_i|C_i=0]=E[Y_{1i}|C_i=1]-E[Y_{0i}|C_i=1] \\ +E[Y_{0i}|C_i=1]-E[Y_{0i}|C_i=0] \tag{2}$$

条件独立假设(CIA)

条件独立假设指的是给定观测到的特点X_i，选择性偏误消失。正式地说，也就是

$$\{Y_{0i},Y_{1i}\}\perp C_i|X_i$$

换而言之，即

$$E[Y_i|X_i,C_i=1]-E[Y_i|X_i,C_i=0]=E[Y_{1i}-Y_{0i}|X_i]$$

教育与收入的研究

假设个体的收入函数为$Y_{si} \equiv f_i(s_i)$，它表示个体i接受s年教育后会获得的收入。在随机实验中，由于S_i是在给定X_i下随机分配的，所以条件独立假设自然成立。则给定X_i，不同教育水平下平均收入的差异就可解释为教育的因果效应。换而言之

$$E[Y_i|X_i, S_i = s] - E[Y_i|X_i, S_i = s-1] = E[f_i(s) - f_i(s-1)|X_i]$$

这意味着，我们对X_i可取的每个值都构造了一个因果效应。经验研究者往往发现用一个综合指标来汇总一系列估计值会显得十分有用。可以用X_i的边际分布做权重，通过对X_i每个可能值对应的因果效应进行加权平均来计算无条件因果效应。**回归为我们提供了一个简单易用的经验研究策略，它可以自动地将条件独立假设转化为我们需要估计的因果效应。**

8.2 回归与匹配

匹配

匹配法对由每个协变量的特定值所决定的个体计算处理组和控制组之间的平均差异，然后用加权平均的方法将这些平均因果效应汇总到一个总的因果效应中。

回归与匹配

回归和匹配都是用来控制协变量的研究策略。而回归可以看做是一种特殊的匹配估计量，特定类型的一种加权后的匹配估计量(Angrist,2008)。

i	D_i	x_i	y_i	匹配结果	$\hat{y}_{0i}$	$\hat{y}_{1i}$
1	0	2	7	{5}	7	8
2	0	4	8	{4, 6}	8	7.5
3	0	5	6	{4, 6}	6	7.5
4	1	3	9	{1, 2}	7.5	9
5	1	2	8	{1}	7	8
6	1	3	6	{1, 2}	7.5	6
7	1	1	5	{1}	7	5

志愿兵服役对之后收入的影响(Angrist,1988)

被处理的平均处理效应

$$\delta_{TOT} \equiv E[Y_{1i} - Y_{0i}|D_i = 1] = E[\delta_X|D_i = 1]$$

其中$\delta_x \equiv E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0]$，对于$X_i$的每个特定值，例如$X_i = x$，我们将相应的收入的平均差距记为$\delta_x$。 在离散情况下，匹配估计量可以写为

$$E[Y_{1i} - Y_{0i}|D_i = 1] = \sum_x \delta_x P(X_i = x|D_i = 1)$$

同样，还可以方便地构造无条件平均处理效应

$$\delta_{ATE} = E[Y_{1i} - Y_{0i}] = \sum_x \delta_x P(X_i = x)$$

回归和匹配的差别只在于将处理效应δ_X加权平均到一个总体平均处理效应时使用的权重不同。

匹配策略可行性

如果解释变量(协变量)所决定的子集中的元素并非既有被处理的个体，也有作为控制的个体，匹配策略就未必可行。

倾向评分定理

若条件独立假设成立，也就是$\{Y_{0i}, Y_{1i}\} \perp D_i|X_i$，那么给定协变量向量的某个值函数$p(X_i)$(即倾向得分)，则潜在结果与处理状况仍然相互独立，即

$$\{Y_{0i}, Y_{1i}\} \perp D_i|p(X_i)$$

其中

$$p(X_i) \equiv E[D_i|X_i] = P[D_i = 1|X_i]$$

倾向评分匹配

- 用类似于logit或probit等参数模型来估计$p(X_i)$
- 用匹配法对处理效应进行估计

城市居民大学教育的收入回报

关注的问题

- 一个任意选取的大学生如果一开始没上大学的话会是什么收入水平
- 一个任意选取的非大学生如果上大学的话会是什么收入水平

8.2 回归与匹配

表 1　　预测倾向值的 Probit 回归结果

	回归系数	标准误	Z 值
城市户口	-1.35	.35	-3.83***
单位性质:党政机关	1.89	.28	6.75***
单位性质:国有企业	.11	.22	.48
单位性质:国有事业	1.40	.22	6.26***
单位性质:集体企事业	.32	.31	1.02
父亲单位性质:党政机关	-.03	.28	-.11
父亲单位性质:国有事业	.28	.19	1.47
父亲单位性质:集体企事业	-.26	.31	-.84
女性	.07	.15	.44
党员	-1.38	.18	-7.66***
年龄	-.22	.03	-8.19***
年龄平方	0	0	6.81***
截距	6.13	.78	7.90***
Log likelihood = -678.365			
Pseudo R^2 = 0.1746			

城市居民大学教育的收入回报

基于倾向值进行匹配

- 虽然每个个体都有倾向值得分，但有些人的倾向值太高或太低，因此无法找到相匹配的个体。
- “匹配样本”中倾向值的取值范围被称为共同区间(common support)。

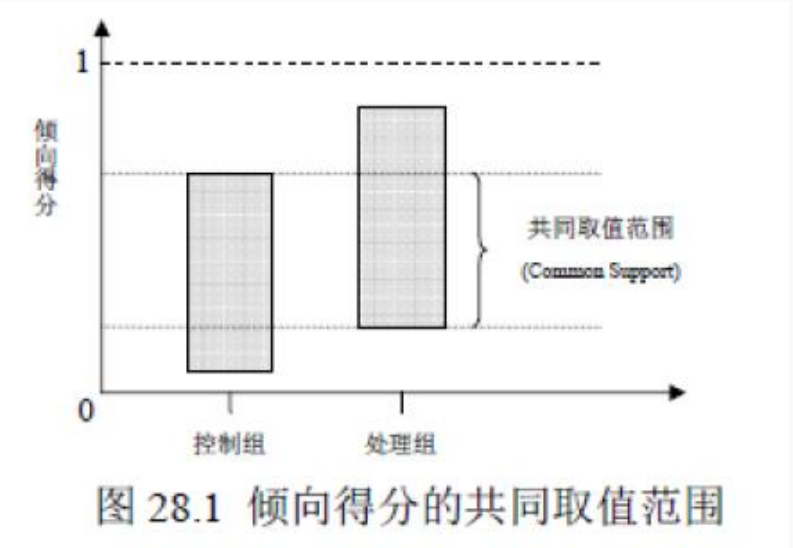


图 28.1 倾向得分的共同取值范围

城市居民大学教育的收入回报

常用匹配方法

- 邻近匹配(找与A的倾向值得分最接近的未上大学的个体B匹配)
- 半径匹配(以个体A的倾向值为中心，以某个数值为半径，在这个范围内的所有没上过大学的个体与A匹配)
- 核匹配(将没有受过大学教育的人的收入值加权平均，而权重则是核方程的取值)

城市居民大学教育的收入回报

基于匹配样本进行因果系数估计

表 2　　倾向值匹配的结果

	受过大学教育的人	没受过大学教育的人	因果关系系数	标准误	T 值
邻近匹配	233	2971	.730	.057	12.87 ***
半径匹配（半径 0.01）	230	2971	.770	.055	13.93 ***
核心匹配	232	2971	.764	.053	14.36 ***

注：（1）由于我们这里主要是用没有受过大学教育的人去匹配受过大学教育的人，我们的关注点是那些受过高等教育的人（即接受了某种“处理”的人），因此这里的因果关系系数即“受到处理的个体的平均处理效果”（average treatment effect of the treated），简写为 ATT；（2）* p <0.1，**p <0.05，***p <0.01（两端检验）。

倾向评分匹配的局限

- 通常要求**比较大的样本容量**以得到高质量的匹配。
- 要求处理组与控制组的倾向得分有**较大的共同取值范围**；否则，将丢失较多观测值，导致剩下的样本不具有代表性。
- **只控制可测变量的影响**，如存在依不可测变量选择的情况，仍有“隐性偏差”。

8.3 断点回归设计

断点回归法

- 清晰断点回归(sharp RD)
- 模糊断点回归(fuzzy RD)

我们可将清晰断点回归设计看作一类选择偏误来自可观测变量的经验研究方法。模糊断点回归则可被视为一种工具变量法。

8.3 断点回归设计

获得国家杰出奖学金的高中生是否更愿意读研究生？

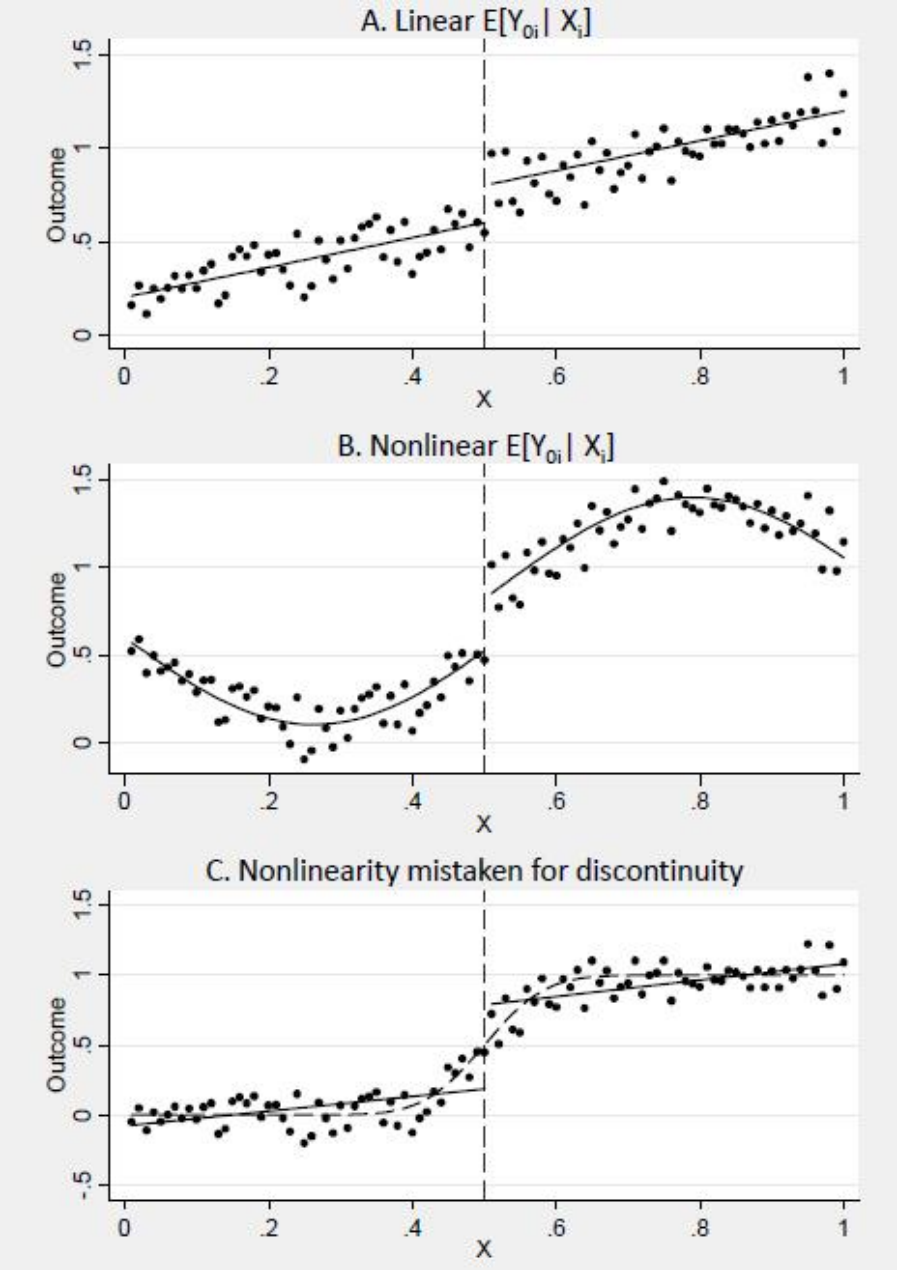
处理状态D_i可以写为如下函数

$$D_i = \begin{cases} 1 & x_i \geq x_0 \\ 0 & x_i \leq x_0 \end{cases}$$

清晰断点回归法通过比较PSAT分数刚好高于和低于国家杰出奖学金分数线的那些高中生的研究生入学率来回答这一问题。 回归模型

$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i$$

8.3 断点回归设计


A. Linear E[Y0i | Xi]
Outcome
0
.5
1
1.5
0
.2
.4
.6
.8
1
X
B. Nonlinear E[Y0i | Xi]
Outcome
0
.5
1
1.5
0
.2
.4
.6
.8
1
X
C. Nonlinearity mistaken for discontinuity
Outcome
-.5
0
.5
1
1.5
0
.2
.4
.6
.8
1
X

8.3 断点回归设计

非线性的挑战

- 多项式回归，例如$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho D_i + \eta_i$
- 邻域内的非参数方法

8.3 断点回归设计

模糊断点回归设计(fuzzy RD)

模糊断点回归设计要挖掘的是给定某个协变量时，处理状态的概率或者期望值所发生的不连续变化。清晰断点设计中，当协变量越过阈值，处理概率就从0变为1，而模糊断点设计中允许处理概率有小幅提升。

退休与城镇家庭消费(邹红等，2015)

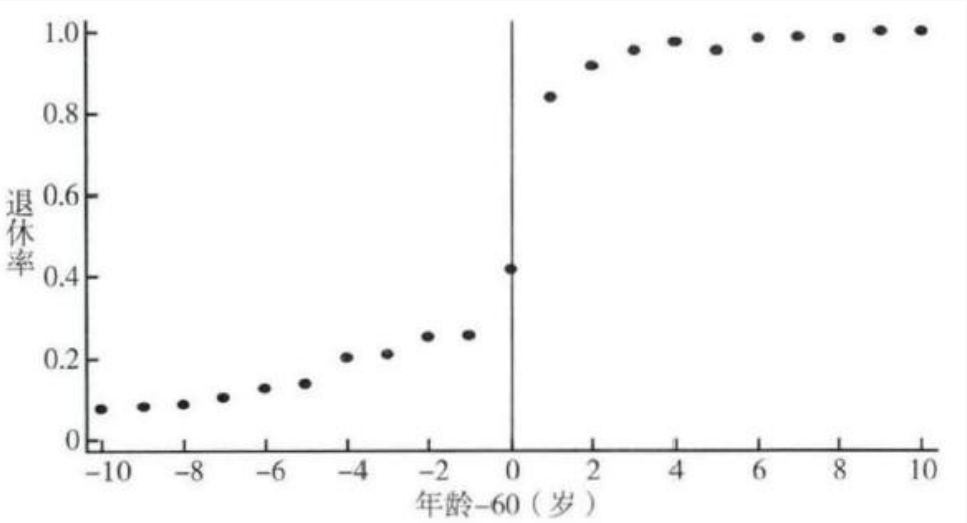


图 1　退休率与年龄

8.3 断点回归设计

$$Y_{st} = \beta_0 + \beta_1 R_{st} + \beta_2 S + \beta_3 S^2 + \varepsilon_{st} \quad (1)$$

$$R_{st} = a_0 + a_1 D_{st}(S > 0, D = 1) + a_2 S + a_3 S^2 + u_{st} \quad (2)$$

其中，下标 t 为时间，s 为户主年龄。(1)式中的 Y_{st} 为每个时期不同户主年龄上的家庭平均消费支出；R 为退休虚拟变量，如果男性户主的就业状态为退休时取值为 1，否则为 0，R_{st} 为每个年份上不同年龄的退休率。S 为年龄断点差(户主年龄 - 法定退休年龄 60)，即户主真实年龄减去退休断点(60)的差，S^2 是年龄断点差的平方，我们加入 S 的多阶项来构造非线性关系进行 RD 估计，多阶项的阶次选择通过 AIC 准则判断。(2)式中的实验变量 D_{st} 用来反应个体所处的年龄与断点之间的关系，当户主年龄断点差大于 0(即大于断点)，D_{st} 取值为 1，这些家庭为实验组；户主年龄断点差小于 0，D_{st} 取值为 0，这些家庭为控制组。

8.3 断点回归设计

表 3　　退休对家庭可支配收入、非耐用消费支出与食物支出的影响

	家庭可支配收入	非耐用消费支出	服务性消费支出	食物支出	在家食物支出
	(1)	(2)	(3)	(4)	(5)
退休 (IV = 年龄虚拟变量)	-0.388*** (0.128)	-0.090*** (0.029)	-0.254*** (0.071)	-0.201*** (0.042)	-0.074* (0.019)
(年龄 -60)	-0.005 (0.008)	-0.004* (0.002)	-0.009** (0.005)	0.023*** (0.004)	0.013*** (0.001)
$(年龄-60)^2$	-0.000 (0.000)	-0.001*** (0.000)	-0.000*** (0.000)	-0.002*** (0.000)	-0.001*** (0.000)
地区和年份虚拟变量	yes	yes	yes	yes	yes
常数项	9.351*** (0.075)	10.381*** (0.017)	9.173 (0.040)	9.637*** (0.024)	9.239*** (0.010)
样本数	3740	3740	2165	3740	3740
R^2	0.139	0.759	0.297	0.827	0.882

注：*、**、***分别表示在 10%、5%、1% 的水平下显著，括号内的标准差均为稳健标准差。