计量经济学

自然实验和反事实估计框架

张晨峰

2019 年 6 月 9 日

华东理工大学商学院

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

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

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)

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

1.理想的实验

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

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

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

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

2.回归与匹配

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。**

什么情况下条件独立假设为经验研究提供一个可信的基础？

一个例子就是由Black等（2003）中对失业工人强制再培训项目所进行的研究。在该项研究中，作者关注强制再培训项目是否成功地提高了失业工人的工资。他们发掘的事实是：强制再培训项目的入选资格取决于基本的个体特征、过去的失业记录和工作历史。当某些强制接受培训的组别中工人数目大于受培训的限额数目时，接受培训的机会是以抽签的方式决定的。因此，给定导致工人被分配至不同组别的协变量，培训状况是随机分配的。

不合格的控制变量

假设我们感兴趣于大学教育对收入的影响，我们是否要把职业选择作为一个控制变量？记W_i是表示个体i是否为白领工人的虚拟变量，该个体的收入水平为Y_i。每个个体接受或不接受大学教育都带来收入水平和职业选择的两种不同的潜在结果，分别记为$\{Y_{1i}, Y_{0i}\}$和$\{W_{1i}, W_{0i}\}$，于是收入水平Y_i和职业选择W_i为

$$Y_i = C_i Y_{1i} + (1 - C_i) Y_{0i}$$

$$W_i = C_i W_{1i} + (1 - C_i) W_{0i}$$

其中，$C_i = 1$表示大学毕业水平，$C_i = 0$为其他。我们假设C_i是随机分配的。由于独立性

$$\mathrm{E}[\mathrm{Y}_1|C_i = 1] - \mathrm{E}[Y_i|C_i = 0] = \mathrm{E}[Y_{\mathrm{u}} - Y_{0_1}]$$

$$\mathrm{E}[W_i|C_i = 1] - \mathrm{E}[W_i|C_i = 0] = \mathrm{E}[W_{1i} - W_{0i}]$$

即我们分别将Y_i和W_i关于C_i回归就可以得到平均因果效应。

不合格的控制变量

给定白领职业，考虑大学毕业生和非大学毕业生的收入差距

$$\begin{aligned}&\mathrm{E}\left[Y_i|W_i=1,C_i=1\right]-\mathrm{E}\left[Y_i|W_i=1,C_i=0\right]\\=&\mathrm{E}\left[Y_{1i}|W_{1i}=1,C_i=1\right]-\mathrm{E}\left[Y_{0i}|W_{0i}=1,C_i=0\right]\end{aligned}$$

由$\{Y_{1i},W_{1i},Y_{0i},W_{0i}\}$的联合分布与$C_i$相互独立可知

$$\begin{aligned}&\mathrm{E}\left[Y_{1i}|W_{1i}=1,C_i=1\right]-\mathrm{E}\left[Y_{0i}|W_{0i}=1,C_i=0\right]\\=&\mathrm{E}\left[Y_{1i}|W_{1i}=1\right]-\mathrm{E}\left[Y_{0i}|W_{0i}=1\right]\end{aligned}$$

这意味不合格控制变量带来的问题

$$\begin{aligned}&\mathrm{E}\left[Y_{1i}|W_{1i}=1\right]-\mathrm{E}\left[Y_{0i}|W_{0i}=1\right]\\=&\mathrm{E}\left[Y_{1i}-Y_{0i}|W_{1i}=1\right]+\left\{\mathrm{E}\left[Y_{0i}|W_{1\mathrm{i}}=1\right]-\mathrm{E}\left[Y_{0i}|W_{0i}=1\right]\right\}\end{aligned}$$

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，都是基于条件独立假设成立。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

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

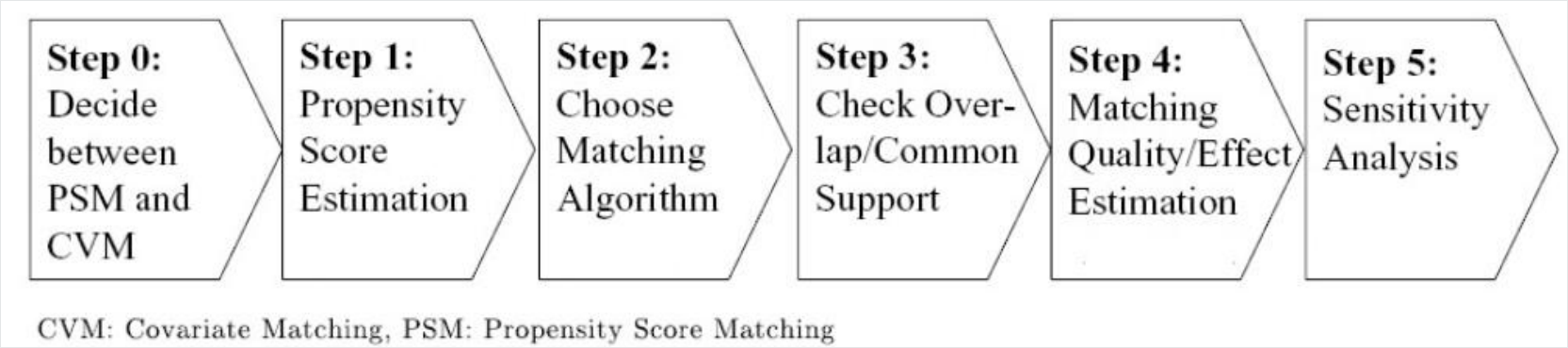
$$p(X_i) \equiv E[D_i|X_i] = P[D_i = 1|X_i]$$

2.回归与匹配

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

PSM的步骤



CVM: Covariate Matching, PSM: Propensity Score Matching

2.回归与匹配

变量的选择

- 包含同时影响处理状况和结果的变量
- 使得匹配满足CIA条件的变量
- 考虑共同区间
- 遗漏变量问题会导致结果有偏

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

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			

2.回归与匹配

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)。

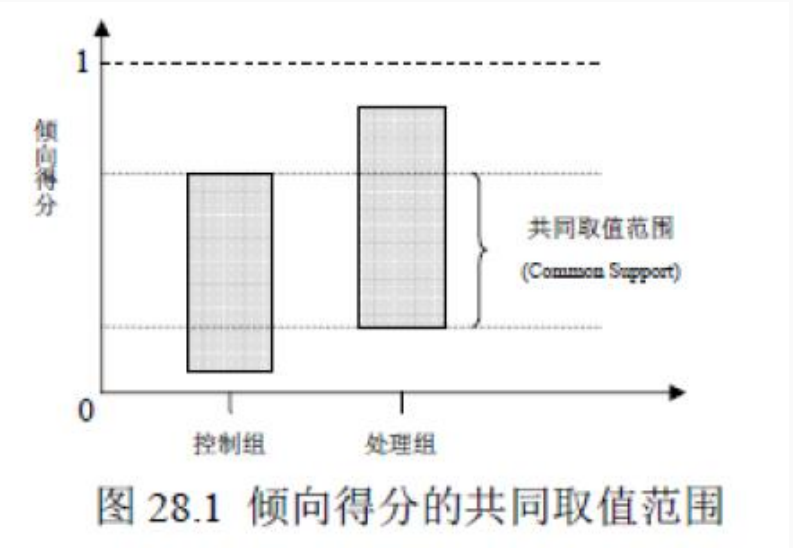


图 28.1 倾向得分的共同取值范围

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)

2.回归与匹配

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”。

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。

获得国家杰出奖学金的高中生是否更愿意读研究生？

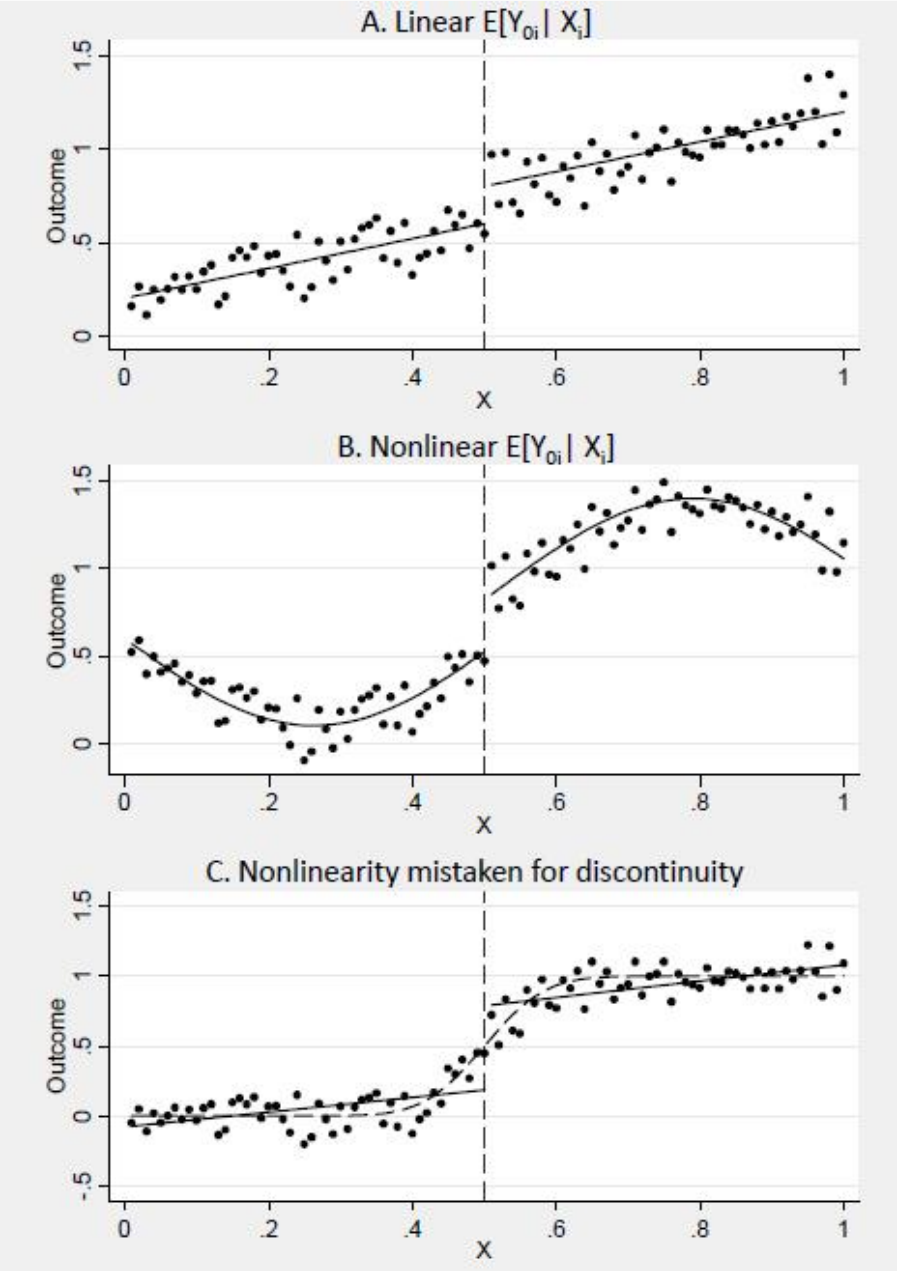
处理状态D_i可以写为如下函数

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

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

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

3.断点回归设计


A. Linear E[Y_{0i}| X_i]
Outcome
X
B. Nonlinear E[Y_{0i}| X_i]
Outcome
X
C. Nonlinearity mistaken for discontinuity
Outcome
X

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

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。

$$P\left[\mathrm{D}_i = 1|X_i\right] = \begin{cases} g_1\left(X_i\right) & \text{if } x_i \geq x_0 \\ g_0\left(X_i\right) & \text{if } x_i < x_0 \end{cases}$$

其中$g_1\left(X_i\right) \neq g_0\left(X_i\right)$。因此，处理的概率为

$$P\left[\mathrm{D}_i = 1|X_i\right] = g_0\left(X_i\right) + \left[g_1\left(X_i\right) - g_0\left(X_i\right)\right] T_i$$

其中$\mathrm{T}_i = 1\left(X_i \geq X_0\right)$，因此上式可以写为

$$\begin{aligned} E\left[\mathrm{D}_i|X_i\right] &= \gamma_{oo} + \gamma_{o1}X_i + \gamma_{o2}X_i^2 + \ldots + \gamma_{op}X_i^p \\ &+\pi\mathrm{T}_i + \gamma_1^*X_i\mathrm{T}_i + \gamma_2^*X_i^2\mathrm{T}_i + \ldots + \gamma_p^*X_i^p\mathrm{T}_i \end{aligned}$$

意味着可以用T_i及其交互项作为D_i的工具变量。

退休与城镇家庭消费(邹红等，2015)

检验我国是否存在退休消费骤降现象。

我国实行的是法定退休制度，退休应该存在明显的年龄断点。

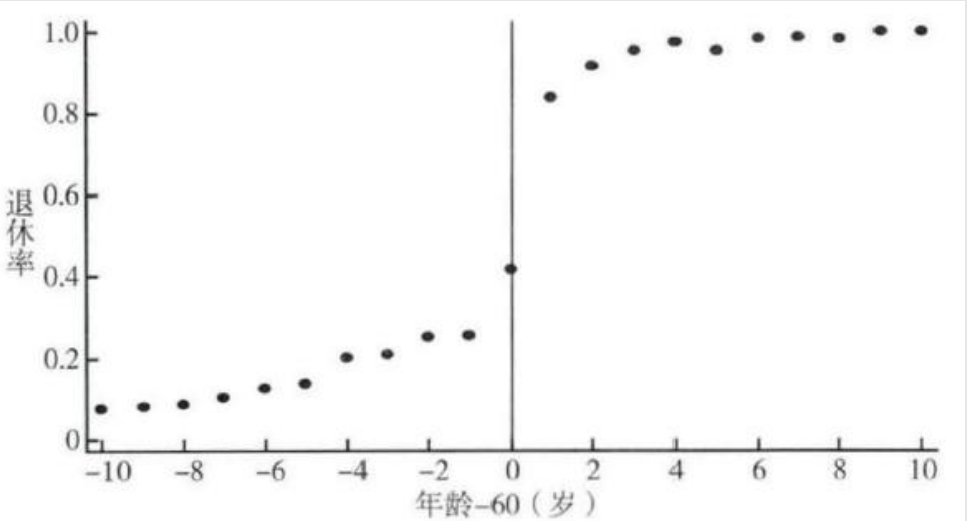


图 1　退休率与年龄

3.断点回归设计

退休与城镇家庭消费(邹红等，2015)

在中国现行退休制度安排下，并非所有人都是在规定退休年龄处停止工作，因为还有其他因素也会影响到退休决定，比如有的人会因为健康状况而更早一些停止工作，再如一些人可能会在办理了法律上的退休手续后返聘或者找到另外的工作等。所以，退休制度仅仅使得退休的可能性在政策规定的退休年龄处发生一个外生的跳跃，但不一定是完全从0直接变动到1的改变，具有这种特征的RD被称为模糊断点FRD。

3.断点回归设计

退休与城镇家庭消费(邹红等，2015)

利用外生的退休制度作为工具变量来识别退休对于消费的影响。对于个人来讲，退休制度的影响反映在个人的年龄是否达到退休规定的年龄，我们可以用个人是否达到退休年龄作为工具变量。同时把样本限制在政策规定的退休年龄附近的人群，因为在这个小的区域内比较好地控制年龄效应，再利用工具变量的思想，把政策规定的退休年龄之前和之后的人作为控制组和实验组，就可以利用退休制度对人们退休决定的外生冲击去估计退休对消费的影响。

退休与城镇家庭消费(邹红等，2015)

实证模型

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

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

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（户主年龄一法定退休年龄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，我们加人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。为了得到退休效应的无偏估计，我们可以用D_{st}作为R_{st}的工具变量。

3.断点回归设计

表 2　　退休制度对退休率的影响

	被解释变量:退休			
	(1)	(2)	(3)	(4)
年龄虚拟变量	0.459***	0.366***	0.339***	0.303***
	(0.013)	(0.018)	(0.020)	(0.023)
(年龄-60)	0.025***	0.046***	0.051***	0.064***
	(0.001)	(0.002)	(0.003)	(0.005)
$(年龄-60)^2$	-0.000	-0.000***	0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
$(年龄-60)^3$		-0.000***	-0.000***	-0.000***
		(0.000)	(0.000)	(0.000)
$(年龄-60)^4$			-0.000***	-0.000***
			(0.000)	(0.000)
$(年龄-60)^5$				0.000***
				(0.000)
地区和年份虚拟变量	yes	yes	yes	yes
常数项	0.387***	0.445***	0.445***	0.470***
	(0.007)	(0.010)	(0.009)	(0.012)
样本数	3740	3740	3740	3740
R^2	0.944	0.950	0.951	0.952
F 检验值	1295.82	436.83	303.89	170.83

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。

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	-0.388***	-0.090***	-0.254***	-0.201***	-0.074*
(IV = 年龄虚拟变量)	(0.128)	(0.029)	(0.071)	(0.042)	(0.019)
(年龄 - 60)	-0.005	-0.004*	-0.009**	0.023***	0.013***
	(0.008)	(0.002)	(0.005)	(0.004)	(0.001)
$(年龄-60)^2$	-0.000	-0.001***	-0.000***	-0.002***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
地区和年份虚拟变量	yes	yes	yes	yes	yes
常数项	9.351***	10.381***	9.173	9.637***	9.239***
	(0.075)	(0.017)	(0.040)	(0.024)	(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。