第十一章
样本自选择模型

杨宇彤

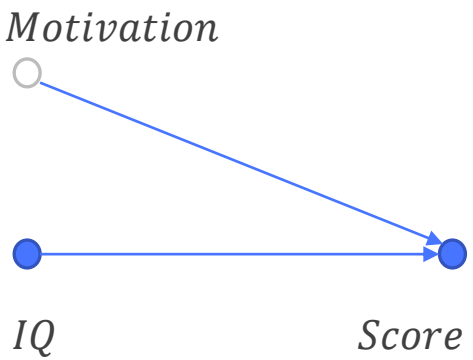
CIRG，2020/12/10

本章框架

- 11.1：样本自选择偏差产生原因的直观理解
- 11.2：样本自选择偏差解决办法的直观理解
- 11.3：传统Heckman样本选择模型

1、样本自选择偏差产生原因的直观理解

- 假设我们要研究智商对大学学习成绩的影响
- 随机分配样本：对总体随机抽样
  - $Score_i = \alpha + \beta IQ_i + e_i$
  - 假设其中不可观测特征e_i是个体的学习动力：$e_i = Motivation$，并且$Motivation_i$和IQ_i不相关
  - 变量路径图随机：



1.1 一个例子

表1 随机分配上大学的个体情况

Score（成绩）（可观测到）	IQ（智商）（可观测到）	Motivation（学习动力）（不可观测到）
73	90	-5
78	90	0
83	90	5
75	100	-5
80	100	0
85	100	5
77	110	-5
82	110	0
87	110	5

1.1 一个例子

- 回归分析：

由于$\mathbb{E}(e_i|IQ_i) = \mathbb{E}(Motivation_i|IQ_i) = 0$，系数$\beta$不会受到干扰项的影响，它反映了$IQ$对$Score$的因果影响

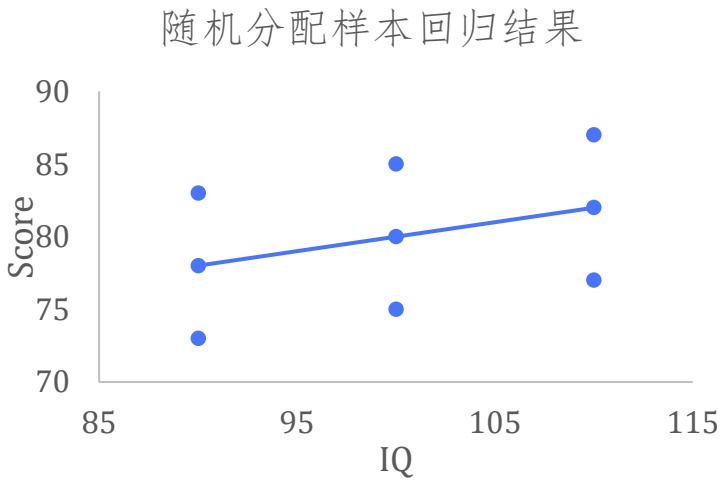
- 回归结果：$Score_i = 60 + 0.2IQ_i + e_i$

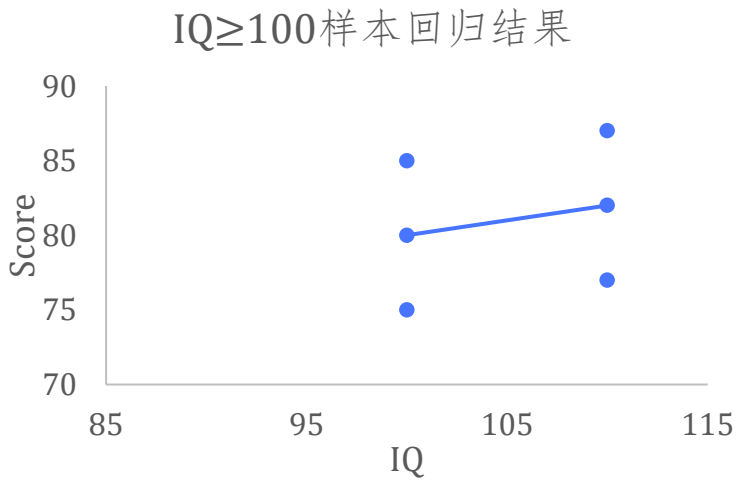
1.2 自选择样本

- 如果上大学与否并非随机分配，而是由个体自己选择决定，那么上大学样本里的个体特征（IQ和$Score$）的分布就会与总体不一样。这种情况下的样本称为自选择样本
- 样本的自选择（sample self-selection）有两种情况:
  - 基于可观测变量选择（selection on observables）
  - 基于不可观测变量选择（ selection on unobservables）

1.2.1 基于可观测变量选择

- 不会导致估计结果偏差
- 假设个体i上大学的效用$Utility$取决于个体可观测特征智商IQ_i，即$Utility_i = IQ_i - 100$
- 假设当$Utility \geq 0$，个体会选择上大学



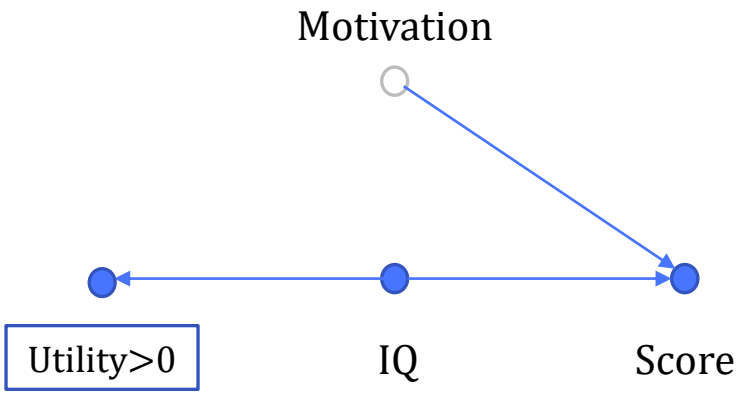


1.2.1 基于可观测变量选择

- 回归结果与随机分配上大学的结果一样

$$Score_i = 60 + 0.2IQ_i + e_i$$

- 虽然造成了样本与总体不一致，但由于是可观测变量造成的，在结果方程里通过控制可观测变量就避免了偏差
- 基于可观测变量自选择的样本的变量路径



1.2.2 基于不可观测变量选择

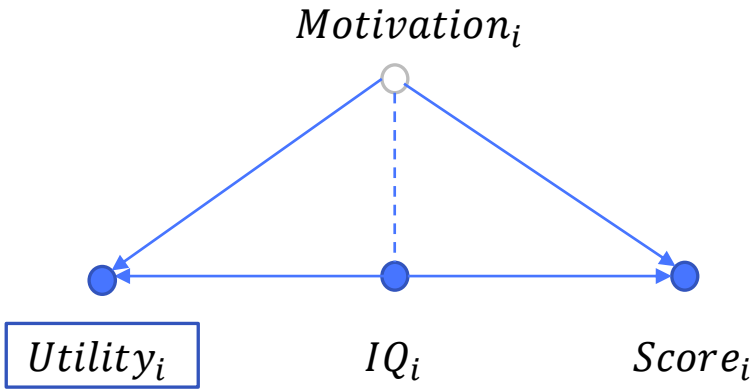
- 如果个体i上大学的效用不仅取决于个体可观测特征智商，还取决于个体不可观测特征学习动力，假设效用函数为：

$$Utility_i = -9 + 0.1IQ_i + Motivation_i$$

- 当$Utility > 0$，个体才会选择上大学
- 此时，$Score_i = \alpha + \beta IQ_i + e_i$中的$IQ_i$和干扰项$e_i$在样本里是相关的，虽然它们在总体里是不相关的

1.2.2 基于不可观测变量选择

- 基于不可观测变量自选择的样本的变量路径



- 变量$Utility_i$是一个对撞变量
- IQ_i和$Score_i$之间存在因果路径$IQ_i \rightarrow Score_i$和衍生路径$Motivation_i \cdots Score_i$

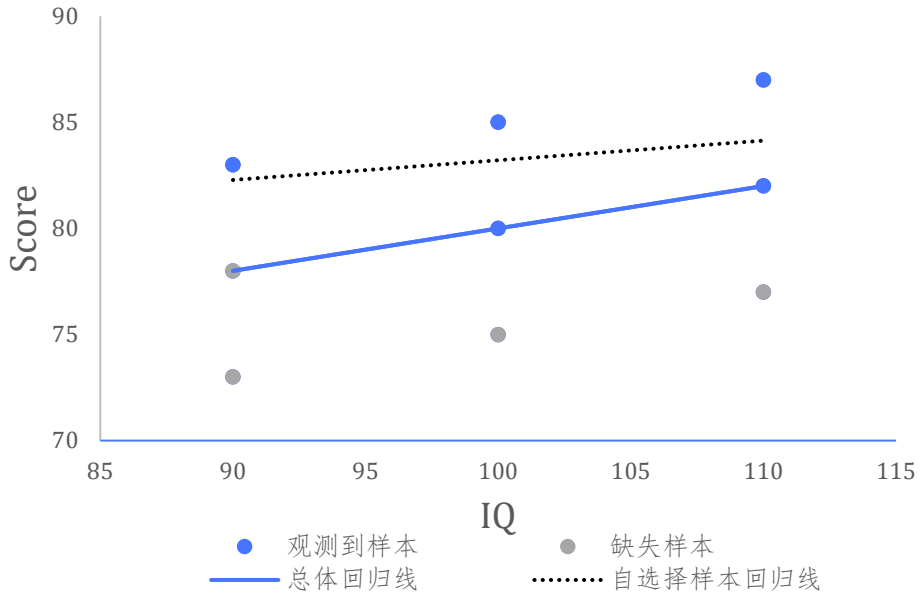
2、样本自选择偏差解决办法的直观理解

表2（基于不可观测变量）自选择上大学的个体情况

Score（成绩） （可观测到）	IQ（智商） （可观测到）
.	90
.	90
83	90
.	100
80	100
85	100
.	110
82	110
87	110

2、样本自选择偏差解决办法的直观理解

- 回归结果：$Score_i = 73 + 0.092IQ_i + e_i$



- 由于自选择缺失的样本点，造成自选择样本回归线的斜率比总体回归线的斜率小

2、样本自选择偏差解决办法的直观理解

- 要通过自选择样本去“倒推”总体的因果关系，必须知道选择方程，即个体如何自选择进入样本

$$Utility_i = -9 + 0.1IQ_i + Motivation_i$$

- 虽然我们观测不到每个个体的$Motivation_i$，但我们知道$Motivation$的分布，即其有三个可能值$(-5, 0, 5)$
- $Utility_i > 0 \Longrightarrow Motivation_i \geq 9 - 0.1 \times IQ_i$

2、样本自选择偏差解决办法的直观理解

表3 通过选择方程推断的样本个体智商和学习动力的分布

	成　　绩			人数分布		
	（1）	（2）	（3）	（4）	（5）	（6）
	Motivation（学习动力）			Motivation（学习动力）		
IQ（智商）	-5	0	5	-5	0	5
90	.	.	83	0	0	1
100	.	80	85	0	1	1
110	.	82	87	0	1	1

2、样本自选择偏差解决办法的直观理解

- 由表3的信息，我们可以计算样本中不同IQ的人的平均学习动力

$$\mathbb{E}(Motivation_i|IQ_i = 90, X) = 5$$
$$\mathbb{E}(Motivation_i|IQ_i = 100, X) = 2.5$$
$$\mathbb{E}(Motivation_i|IQ_i = 110, X) = 2.5$$

- 智商与学习动力的相关性方程：

$$\mathbb{E}(Motivation_i|IQ_i, X) = f(IQ_i)$$

2、样本自选择偏差解决办法的直观理解

- 在回归方程中将智商与学习动力的相关性“控制”掉，剩下的干扰项v_i与自变量不相关

$$Score_i = \alpha + \beta IQ_i + \mathbb{E}(Motivation_i|IQ_i, X) + v_i$$

- 把$\mathbb{E}(Motivation_i|IQ_i, X)$称为调整项$Adjust_i$

2、样本自选择偏差解决办法的直观理解

表4 增添调整项的自选择样本数据

Score（成绩）	IQ（智商）	$\mathbb{E}(Motivation_i \mid IQ_i)$
.	90	.
.	90	.
83	90	5.0
.	100	.
80	100	2.5
85	100	2.5
.	110	.
82	110	2.5
87	110	2.5

2、样本自选择偏差解决办法的直观理解

- 用添加调整项的表4中的样本数据回归调整的结果方程：

$$Score_i = \alpha + \beta IQ_i + Adjust_i + v_i$$

其中，$Adjust_i = \mathbb{E}(Motivation_i | IQ_i)$

- 新的方程回归得到的IQ_i系数等于0.2，与用原方程对随机分配回归得到的IQ系数是一样的

3、传统Heckman样本选择模型

- 模型设置
- Heckman模型如何解决样本选择偏差

3.1 模型设置

- 结果方程：

$$Y_i^* = \alpha + \boldsymbol{X_i'\beta} + e_{1i}$$

- 选择方程：

$$D_i^* = \boldsymbol{Z_i'\gamma} + e_{2i}$$

$$\begin{cases} D_i = 1，\ \text{如果} D_i^* > 0 \\ D_i = 0，\ \text{如果} D_i^* \leq 0 \end{cases}$$

- 样本中观测到的结果为：

$$\begin{cases} Y_i = Y_i^*，\ \text{如果} D_i = 1 \\ Y_i \text{缺失}，\ \text{如果} D_i = 0 \end{cases}$$

用上述模型描述前一节的例子

- 结果方程：

$$Score_i^* = \alpha + \beta IQ_i + e_{1i}$$

- 选择方程：

$$Utility_i^* = \gamma_0 + \gamma_1 IQ_i + \gamma_2 Parent\ Education_i + e_{2i}$$

$$\begin{cases} D_i = 1，\ 如果 Utility_i^* > 0 \\ D_i = 0，\ 如果 Utility_i^* \leq 0 \end{cases}$$

用上述模型描述前一节的例子

- 样本中观测到的结果为：

$$\begin{cases} Score_i = Score_i^*，\text{如果} D_i = 1 \\ Score_i \text{缺失}，\text{如果} D_i = 0 \end{cases}$$

- 假设e_{1i}和e_{2i}包含相同的不可观测变量（如$Motivation$变量），二者是相关的
- $Heckman$选择模型假设二者的相关关系如下：

$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)$$

3.2 Heckman模型如何解决样本选择偏差

- 自选择样本观测结果的条件期望函数：

$$
\begin{aligned}
&\mathbb{E}(Y_i \mid \text{样本}, \boldsymbol{X_i}) \\
=\ &\mathbb{E}(Y_i \mid D_i = 1, \boldsymbol{X_i}) \\
=\ &\mathbb{E}(Y_i^* \mid D_i^* > 0, \boldsymbol{X_i}) \\
=\ &\mathbb{E}(\alpha + \boldsymbol{X_i'\beta} + e_{1i} \mid \boldsymbol{Z_i'\gamma} + e_{2i} > 0, \boldsymbol{X_i}) \\
=\ &\alpha + \boldsymbol{X_i'\beta} + \mathbb{E}(e_{1i} \mid e_{2i} > -\boldsymbol{Z_i'\gamma}, \boldsymbol{X_i}) \\
=\ &\alpha + \boldsymbol{X_i'\beta} + \mathbb{E}(e_{1i} \mid e_{2i} > -\boldsymbol{Z_i'\gamma})
\end{aligned}
$$

3.2 Heckman模型如何解决样本选择偏差

- 增加控制变量后样本结果回归方程为：

$$Y_i = \alpha + \boldsymbol{X_i'\beta} + \mathbb{E}(e_{1i} \mid e_{2i} > -\boldsymbol{Z_i'\gamma}) + v_i$$

- 此时，$\mathbb{E}(v_i \mid \text{样本}, \boldsymbol{X_i'}) = 0$

$$\mathbb{E}(e_{1i} \mid e_{2i} > -\boldsymbol{Z_i'\gamma}) = \rho\sigma\lambda(\boldsymbol{Z_i'\gamma}) = \rho\sigma\frac{\phi\left(\frac{-\boldsymbol{Z_i'\gamma}}{\sigma}\right)}{1-\Phi\left(\frac{-\boldsymbol{Z_i'\gamma}}{\sigma}\right)}$$

逆米尔斯比例（IMR）

- 其中：

$$\lambda_i = \lambda(\boldsymbol{Z_i'\gamma}) = \frac{\phi\left(\frac{-\boldsymbol{Z_i'\gamma}}{\sigma}\right)}{1-\phi\left(\frac{-\boldsymbol{Z_i'\gamma}}{\sigma}\right)}$$

- 增加逆米尔斯比例项λ的样本结果回归方程变为：

$$Y_i = \alpha + \boldsymbol{X_i'\beta} + \rho\sigma\lambda_i + \upsilon_i$$

两阶段估计法

- 第一阶段，使用Probit模型估计样本选择方程：

$$\Pr(D_i = 1 \mid \boldsymbol{Z_i}) = \Pr(e_{2i} > -\boldsymbol{Z_i\gamma} \mid \boldsymbol{Z_i}) = \Phi(\boldsymbol{Z_i\gamma})$$

- 把$\boldsymbol{\gamma}$的估计值$\hat{\boldsymbol{\gamma}}$带入IMR公式：

$$\lambda_i = \lambda(\boldsymbol{Z_i'}\hat{\boldsymbol{\gamma}}) = \frac{\phi\left(\frac{-\boldsymbol{Z_i'}\hat{\boldsymbol{\gamma}}}{\sigma}\right)}{1 - \phi\left(\frac{-\boldsymbol{Z_i'}\hat{\boldsymbol{\gamma}}}{\sigma}\right)}$$

两阶段估计法

- 第二阶段，使用样本数据，将Y_i对$\boldsymbol{X_i'\beta}$和λ_i进行回归：

$$Y_i = \alpha + \boldsymbol{X_i'\beta} + \rho\sigma\lambda_i + \upsilon_i$$

- 两阶段法通过在回归方程中加入IMR，修正了样本选择偏差，得到的$\widehat{\boldsymbol{\beta}}$总体$\boldsymbol{\beta}$的一致估计值