样本自选择模型

瞿博洋

CIGR，2020/12/17

本章内容

- 第四节 Heckman样本选择模型的应用例子
- 第五节 内生选择变量处置效应模型
- 第六节 样本自选择模型运用常见问题

4.1 模型设置

- 研究受教育程度对女性工资水平的影响
- 简化模型后，结果方程为：

$$Wage_i^* = \alpha + \beta_i Edu_i + \beta_2 Age_i + e_{1i}$$

- 其中，$Wage_i^*$是工资水平，Edu_i是受教育程度，Age_i是年龄，干扰项e_{1i}包含了不可观测但会影响工资水平的变量（如个体性格）。
- 对于总体或随机分配样本：

$$\mathbb{E}(e_{1I}|Edu_i, Age_i) = 0$$

4.1 模型设置

- 只有参加工作的人，才能观测到工资水平
- 是否参加工作是自我选择的，选择方程:

$$Utility_i^* = \gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i + e_{2i}$$

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

- 其中，$Children_i$是小孩的数量，干扰项e_{2i}包含了不可观测但会影响工资水平的变量（如个体性格）。

4.1 模型设置

- e_{1i}和e_{2i}都包含了一些相同的不可观测的变量，所以二者是相关的
- Heckman模型假设二者的分布是相关系数为ρ的二元正态分布，即：

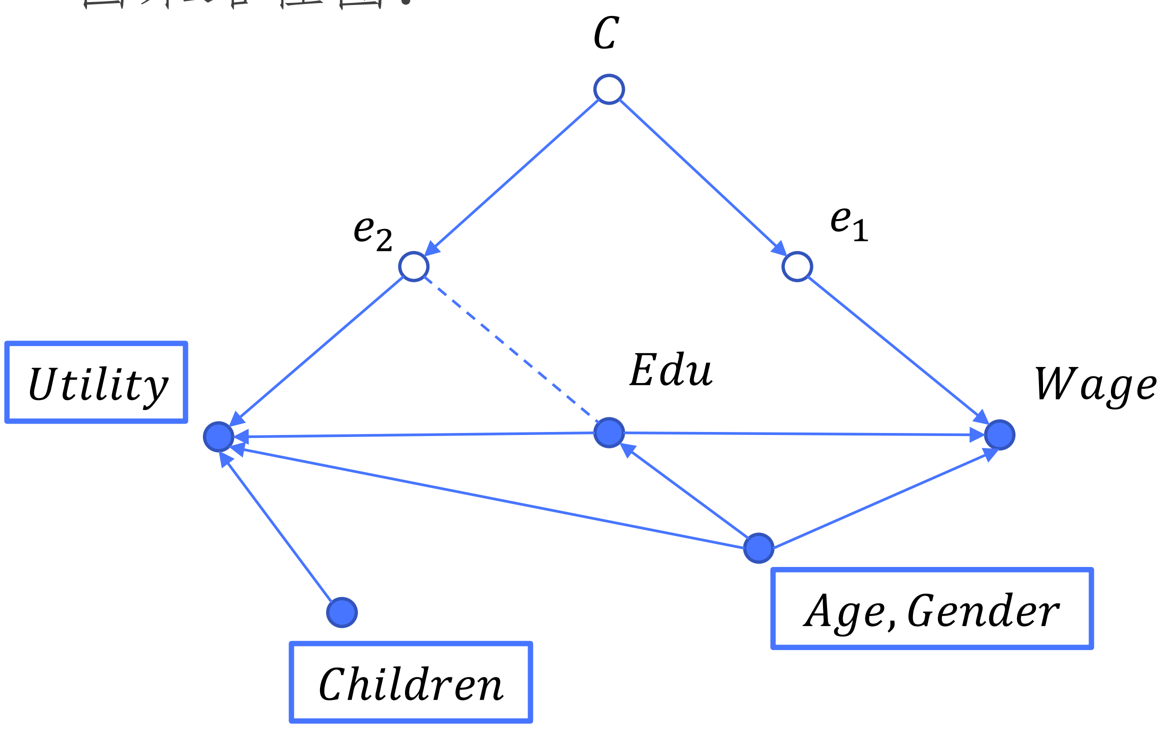
$$\begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right)$$

- 对于工资的观测结果为：

$$\begin{cases} Wage_i = Wage_i^*, & \text{如果} Work_i = 1 \\ Wage_i \text{缺失}, & \text{如果} Work_i = 0 \end{cases}$$

4.1 模型设置

- 因果路径图：



4.1 模型设置

- 在结果方程里加入逆米尔斯比例作为调整项

$$Wage_i = \alpha + \beta_i Edu_i + \beta_2 Age_i + \rho\sigma\lambda_i + v_i$$

- 其中，逆米尔斯比例为：

$$\lambda_i = \frac{\phi(\gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i)}{1 - \Phi(\gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i)}$$

4.2 样本数据

- 样本里有2000个观测值，以下显示前20个：

```
. list wage work age education children if _n<=20
```

	wage	work	age	educat~n	children
1.	.	0	22	10	0
2.	20.31285	1	36	10	0
3.	.	0	28	10	0
4.	.	0	37	10	0
5.	16.14224	1	39	10	1
6.	14.95799	1	33	10	2
7.	18.44339	1	57	10	1
8.	17.57406	1	45	16	0
9.	.	0	39	12	0
10.	18.48312	1	25	10	3
11.	29.40447	1	26	16	0
12.	.	0	28	10	1
13.	.	0	52	10	1
14.	24.83475	1	38	16	3
15.	27.17002	1	36	16	4
16.	16.86481	1	32	12	3
17.	33.82108	1	36	16	5
18.	18.97637	1	46	12	5
19.	.	0	39	16	0
20.	.	0	34	10	0

4.3 手动估计模型

- 运用Probit模型估计选择模型：

$$\Pr(Work_i = 1|Z_i) = \Phi(\gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i)$$

```
. probit work education age children

Iteration 0:   log likelihood = -1266.2225
Iteration 1:   log likelihood = -1048.0634
Iteration 2:   log likelihood = -1044.0756
Iteration 3:   log likelihood = -1044.0621
Iteration 4:   log likelihood = -1044.0621

Probit regression                               Number of obs     =      2,000
                                                LR chi2(3)        =     444.32
                                                Prob > chi2       =     0.0000
Log likelihood = -1044.0621                     Pseudo R2         =     0.1755
```

work	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
education	.071238	.0107186	6.65	0.000	.0502299	.092246
age	.0412517	.0040689	10.14	0.000	.0332769	.0492265
children	.4084356	.0269242	15.17	0.000	.3556651	.461206
_cons	-2.535539	.1906351	-13.30	0.000	-2.909177	-2.161901

4.3 手动估计模型

- 用估计出来的系数去估计逆米尔斯系数

$$\lambda_i = \frac{\phi(\gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i)}{1 - \Phi(\gamma_0 + \gamma_1 Edu_i + \gamma_2 Age_i + \gamma_3 Children_i)}$$

- Stata代码如下：

```
. predict z if e(sanple),xb

. generate phi=normalden(z)

. generate PHI=normal(z)

. generate lambda=phi/PHI
```

4.3 手动估计模型

- 生成的新的变量如下所示

```
. list wage work age education children z phi PHI lambda if _n<=20
```

	wage	work	age	educat~n	children	z	phi	PHI	lambda
1.	.	0	22	10	0	-.9156223	.2623383	.1799325	1.457981
2.	20.31285	1	36	10	0	-.3380987	.37678	.3676444	1.024849
3.	.	0	28	10	0	-.6681122	.31914	.252031	1.266273
4.	.	0	37	10	0	-.296847	.3817468	.3832916	.9959697
5.	16.14224	1	39	10	1	.1940919	.3914982	.576948	.6785675
6.	14.95799	1	33	10	2	.3550174	.3745773	.6387117	.5864575
7.	18.44339	1	57	10	1	.9366222	.2572854	.8255236	.3116634
8.	17.57406	1	45	16	0	.4605941	.3587922	.6774551	.5296177
9.	.	0	39	12	0	-.0718678	.3979133	.4713536	.8441929
10.	18.48312	1	25	10	3	.4334395	.3631739	.6676522	.5439567
11.	29.40447	1	26	16	0	-.3231878	.3786421	.3732765	1.014374
12.	.	0	28	10	1	-.2596766	.3857158	.3975566	.9702159
13.	.	0	52	10	1	.7303638	.3055462	.7674161	.3981494
14.	24.83475	1	38	16	3	1.397139	.1503278	.9188141	.1636106
15.	27.17002	1	36	16	4	1.723071	.0904077	.9575621	.0944145
16.	16.86481	1	32	12	3	.8646771	.2745088	.806392	.3404161
17.	33.82108	1	36	16	5	2.131507	.0411472	.9834763	.0418385
18.	18.97637	1	46	12	5	2.259072	.0310971	.9880605	.0314729
19.	.	0	39	16	0	.2130841	.3899873	.5843693	.6673645
20.	.	0	34	10	0	-.4206021	.3651702	.3370228	1.083518

4.3 手动估计模型

- 估计回归方程

```
. reg wage education age lambda

      Source |       SS           df       MS      Number of obs   =     1,343
-------------+----------------------------------   F(3, 1339)      =    171.27
       Model |   14796.046         3  4932.01533   Prob > F        =    0.0000
    Residual |  38558.8486     1,339  28.7967503   R-squared       =    0.2773
-------------+----------------------------------   Adj R-squared   =    0.2757
       Total |  53354.8946     1,342  39.7577456   Root MSE        =    5.3663

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |     .98588   .0508305    19.40   0.000     .8861639    1.085596
         age |   .2123369   .0209086    10.16   0.000     .1713197    .2533542
      lambda |   3.973879   .5979168     6.65   0.000     2.800924    5.146835
       _cons |   .6543349   1.197644     0.55   0.585    -1.695129    3.003799
------------------------------------------------------------------------------
```

4.4 使用Stata的Heckman命令估计模型

- Stata中也有自带的Heckman命令可以用来直接估计模型的回归结果

```
. heckman wage education age, select(education age children) twostep

Heckman selection model -- two-step estimates   Number of obs     =      2,000
(regression model with sample selection)              Selected    =      1,343
                                                      Nonselected =        657

                                                Wald chi2(2)      =     432.15
                                                Prob > chi2       =     0.0000
```

wage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wage						
education	.98588	.0542582	18.17	0.000	.8795358	1.092224
age	.2123369	.0223243	9.51	0.000	.1685821	.2560917
_cons	.6543347	1.282338	0.51	0.610	-1.859001	3.16767
select						
education	.071238	.0107186	6.65	0.000	.0502299	.092246
age	.0412517	.0040689	10.14	0.000	.0332769	.0492265
children	.4084356	.0269242	15.17	0.000	.3556651	.461206
_cons	-2.535539	.1906351	-13.30	0.000	-2.909177	-2.161901
/mills						
lambda	3.973879	.6296416	6.31	0.000	2.739805	5.207954
rho	0.66708					
sigma	5.9570896					

5.1 模型设置

- 自选择样本偏差的原理及其处理方法的另一个应用，是估计内生二元选择变量(endogenous binary-treatment variable)的处置效应
- 一个常见的内生二元自选择变量模型：

$$Y_i = \alpha_0 + \alpha_1 D_i + \boldsymbol{X}_i'\boldsymbol{\beta} + \mathrm{e}_{1i}$$

- 其中，D_i是一个二元选择变量，$\boldsymbol{X}_i'$是控制变量
- 选择公式为：

$$Utility_i = \boldsymbol{Z}_i'\boldsymbol{\gamma} + e_{2i}$$

- 其中，只有当$Utility_i > 0$时才接受处置：

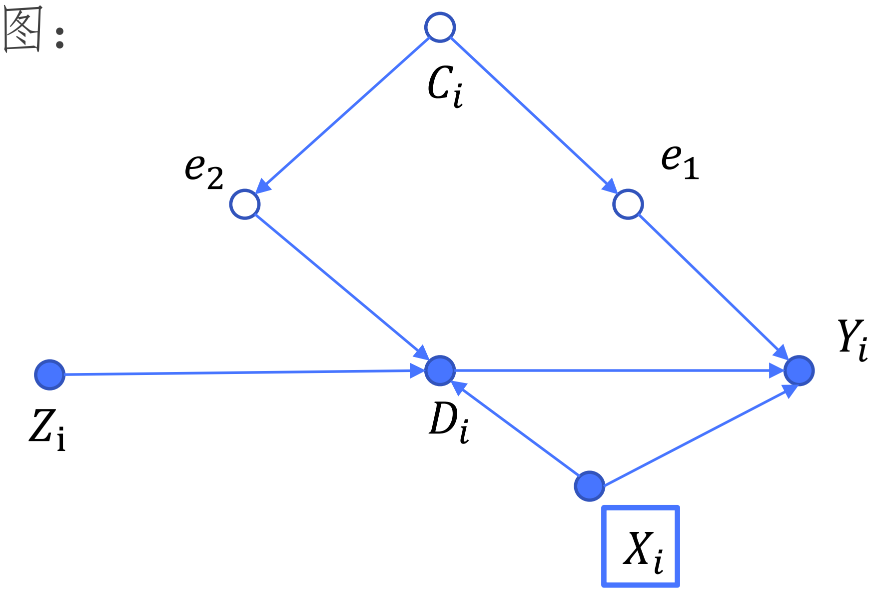
$$\begin{cases} D_i = 1, & \text{如果} Utility_i > 0 \\ D_i = 0, & \text{如果} Utility_i \leq 0 \end{cases}$$

5.1 模型设置

- 该模型有两个和Heckman相同的假设：
  - $\boldsymbol{Z}_i'$和$\boldsymbol{X}_i'$为外生变量，他们与干扰项无关
  - e_{1i}和e_{2i}服从二元正态分布：

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

  - 路径图：



5.1 模型设置

- 和Heckman模型类似的，我们计算条件期望
- 当$D_i = 1$时：

$$
\begin{aligned}
&\mathbb{E}(Y_i|\boldsymbol{X_i}, D_i = 1) \\
&= \alpha_0 + \alpha_1 + \boldsymbol{X}_i'\boldsymbol{\beta} + \mathbb{E}(\mathrm{e}_{1i}|D_i = 1, \boldsymbol{X_i}) \\
&= \alpha_0 + \alpha_1 + \boldsymbol{X}_i'\boldsymbol{\beta} + \mathbb{E}(\mathrm{e}_{1i}|e_{2i} > -\boldsymbol{Z}_i'\boldsymbol{\gamma}) \\
&= \alpha_0 + \alpha_1 + \boldsymbol{X}_i'\boldsymbol{\beta} + \rho\sigma\frac{\phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}{1 - \Phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}
\end{aligned}
$$

5.1 模型设置

- 当$D_i = 0$时：

$$
\begin{aligned}
&\mathbb{E}(Y_i|\boldsymbol{X_i}, D_i = 0) \\
&= \alpha_0 + \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\beta} + \mathbb{E}(\mathrm{e}_{1i}|D_i = 0, \boldsymbol{X_i}) \\
&= \alpha_0 + \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\beta} + \mathbb{E}(\mathrm{e}_{1i}|e_{2i} < -\boldsymbol{Z}_{\boldsymbol{i}}'\boldsymbol{\gamma}) \\
&= \alpha_0 + \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\beta} + \rho\sigma\frac{-\phi(-\boldsymbol{Z}_{\boldsymbol{i}}'\boldsymbol{\gamma})}{\Phi(-\boldsymbol{Z}_{\boldsymbol{i}}'\boldsymbol{\gamma})}
\end{aligned}
$$

5.2 估计方法

- 要获得α_1，有两种估计方法
  - 分别估计

  分别对$D_i = 1$和$D_i = 0$时的样本使用Heckman样本选择模型进行估计，然后将两个回归结果的截距相减，就能得到系数α_1

  - 整合估计

  将两个公式和在一起，表示为：

$$Y_i = \alpha_0 + \alpha_1 D_i + \boldsymbol{X}_i'\boldsymbol{\beta} + \rho\sigma\left[\frac{\phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}{1-\Phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}D_i + \frac{-\phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}{\Phi(-\boldsymbol{Z}_i'\boldsymbol{\gamma})}(1-D_i)\right] + u_i$$

  然后对上述方程使用Heckman样本选择模型模拟

5.3 实例

- 估计女性工会成员身份对工资水平的影响，结果方程如下：

$$Wage_i = \alpha_0 + \alpha_1 Age_i + \alpha_2 Grade_i + \alpha_3 Smsa_i + \alpha_4 Black_i + \alpha_5 Tenure_i + e_{1i}$$

- 其中，$Wage_i$是工资水平，Age_i是年龄，$Grade_i$是学历,$Tenure_i$是工作时限

- 进入工会的选择方程为：

$$Utility_i = \gamma_0 + \gamma_1 South_i + \gamma_2 Black_i + \gamma_3 Tenure_i + e_{2i}$$

$$\begin{cases} Union_i = 1, & \text{如果}Utility_i > 0 \\ Union_i = 0, & \text{如果}Utility_i \leq 0 \end{cases}$$

5.3 实例

- 展示1693个观测值中的前20个

```
. list wage age grade smsa black tenure if _n<=20
```

	wage	age	grade	smsa	black	tenure
1.	4.903638	20	12	1	1	.9166667
2.	3.3407572	20	12	1	1	1
3.	4.9892929	26	12	1	1	2.416667
4.	11.177726	26	17	1	0	3.416667
5.	7.2376854	26	12	1	0	.6666667
6.	4.9892929	25	12	1	0	1.416667
7.	4.282655	23	12	1	0	4.75
8.	5.7387546	20	12	1	0	2.5
9.	3.6748322	20	10	1	0	3.25
10.	7.4732333	23	15	1	0	1.666667
11.	8.0299786	23	15	1	0	2.333333
12.	5.888651	23	15	1	0	2.416667
13.	8.3083492	23	15	1	0	.3333333
14.	9.1006418	23	15	1	0	1.75
15.	10.192716	24	15	1	0	.4166667
16.	8.1584569	25	14	1	0	.75
17.	5.3319055	23	13	1	0	2
18.	4.8393968	21	8	1	0	.5833333
19.	3.6748322	20	12	1	0	1.166667
20.	4.9464658	27	12	1	0	3.083333

5.3 实例

- 使用etregress命令

```
. etregress wage age grade smsa black tenure, treat(union = south black tenure)

Iteration 0:   log likelihood =  -3140.811
Iteration 1:   log likelihood = -3053.6629
Iteration 2:   log likelihood = -3051.5847
Iteration 3:   log likelihood =  -3051.575
Iteration 4:   log likelihood =  -3051.575

Linear regression with endogenous treatment     Number of obs     =      1,210
Estimator: maximum likelihood                   Wald chi2(6)      =     681.89
Log likelihood =  -3051.575                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wage         |
         age |   .1487409   .0193291     7.70   0.000     .1108566    .1866252
       grade |   .4205658   .0293577    14.33   0.000     .3630258    .4781058
        smsa |   .9117044   .1249041     7.30   0.000     .6668969    1.156512
       black |  -.7882471   .1367078    -5.77   0.000     -1.05619   -.5203048
      tenure |   .1524015   .0369596     4.12   0.000     .0799621    .2248409
     1.union |   2.945815   .2749621    10.71   0.000       2.4069    3.484731
       _cons |  -4.351572   .5283952    -8.24   0.000    -5.387208   -3.315936
-------------+----------------------------------------------------------------
union        |
       south |  -.5807419   .0851111    -6.82   0.000    -.7475566   -.4139271
       black |   .4557499   .0958042     4.76   0.000     .2679771    .6435226
      tenure |   .0871536   .0232483     3.75   0.000     .0415878    .1327195
       _cons |  -.8855758   .0724506   -12.22   0.000    -1.027576   -.7435753
-------------+----------------------------------------------------------------
     /athrho |  -.6544347   .0910314    -7.19   0.000     -.832853   -.4760164
    /lnsigma |   .7026769   .0293372    23.95   0.000      .645177    .7601767
-------------+----------------------------------------------------------------
         rho |  -.5746478    .060971                      -.682005   -.4430476
       sigma |   2.019151   .0592362                      1.906325    2.138654
      lambda |    -1.1603   .1495097                     -1.453334   -.8672668
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =    19.84   Prob > chi2 = 0.0000
```

6.1 解释变量的选择

- 在实际运用中，要求$\boldsymbol{Z_i}$至少包含一个与$\boldsymbol{X_i}$不同的变量
- 假设$\boldsymbol{Z_i}=\boldsymbol{X_i}$，即$Y_i=\alpha_0+\boldsymbol{X}_i'\boldsymbol{\beta}+\lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\gamma}})+\mathrm{e}_i$。由于$\lambda(\boldsymbol{X}_i'\hat{\boldsymbol{\gamma}})$在定义域的大部分范围内是近线性的，所以会导致严重的共线性问题
- 需要一个工具变量，影响选择但不影响结果，称为排他性约束条件(exclusion constraints)

6.2 二元正态分布假设

- 如果干扰项不符合二元正态分布假设，调整项的计算就有可能是错误的
- 一种替代方案是，假定干扰项服从一些其他的特定的非正态分布
- 但现有的理论很少指出应该用何种分布来代替

6.3 选择模型必须为Probit模型

- 在Heckman模型中，一阶段的估计选择方程不能使用Logit模型，因为Logit模型不具有干扰项正态分布的假设，与Heckman模型不符合
- Probit模型：
  - 一个二元0/1变量的模型，取值取决于如下方程：

$$D_i^* = \boldsymbol{Z}_i'\boldsymbol{\gamma} + e_i$$

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

  - 假设e_i符合标准正态分布，则：

$$\Pr(D_i = 1|\boldsymbol{Z}_i) = \Phi(\boldsymbol{Z}_i'\boldsymbol{\gamma})$$

6.4 检查相关系数ρ

- 当e_{1i}和e_{2i}不相关时，样本自选择并不会造成估计偏差，这种情况也被称为外生样本选择
- 这种情况下不需要加入调整项