

§ 4.6 受限被解释变量数据模型 ——选择性样本

Model with Limited Dependent Variable
——Selective Samples Model

- 一、经济生活中的受限被解释变量问题
- 二、“截断”问题的计量经济学模型
- 三、“归并”问题的计量经济学模型

**The Bank of Sweden Prize in Economic
Sciences in Memory of Alfred Nobel 2000**
**"for his development of theory and
methods for analyzing selective samples"**



James J Heckman

USA

- **“Shadow Prices, Market Wages and Labour Supply”, *Econometrica* 42 (4), 1974, P679-694**
发现并提出“选择性样本”问题。
- **“Sample Selection Bias as a Specification Error”, *Econometrica* 47(1), 1979, P153-161**
证明了偏误的存在并提出了**Heckman**两步修正法。

一、经济生活中的受限被解释变量问题

1、“截断”（truncation）问题

- 由于条件限制，样本不能随机抽取，即不能从全部个体，而只能从一部分个体中随机抽取被解释变量的样本观测值，而这部分个体的观测值都大于或者小于某个确定值。“掐头”或者“去尾”。
- 消费函数例题：被解释变量最底**200**元、最高**10000**元。原因：抽样。
- 离散选择模型的例题：银行贷款，实际上是选择性样本，通常表现为“截断样本”。原因：问题的局限。

类似的实际问题很多

能够获得贷款的企业是全部有贷款需求的企业中表现良好的一部分

2、“归并”(censoring)问题

- 将被解释变量的处于某一范围的样本观测值都用一个相同的值代替。
- 经常出现在“检查”、“调查”活动中，因此也称为“检查”(censoring)问题。
- 需求函数模型中用实际消费量作为需求量的观测值，如果存在供给限制，就出现“归并”问题。
- 被解释变量观测值存在最高和最低的限制。例如考试成绩，最高**100**，最低**0**，出现“归并”问题。

二、“截断”问题的计量经济学模型

1、思路

- 如果一个单方程计量经济学模型，只能从“掐头”或者“去尾”的连续区间随机抽取被解释变量的样本观测值，那么很显然，抽取每一个样本观测值的概率以及抽取一组样本观测值的联合概率，与被解释变量的样本观测值不受限制的情况是不同的。
- 如果能够知道在这种情况下抽取一组样本观测值的联合概率函数，那么就可以通过该函数极大化求得模型的参数估计量。

2、截断分布

$$f(\xi|\xi > a) = \frac{f(\xi)}{P(\xi > a)}$$

a 为随机变量 ξ 分布范围内的一个常数

$$f(\xi|\xi > c) = \frac{f(\xi)}{P(\xi > c)} = \frac{1/(b-a)}{\int_c^b \frac{1}{b-a} d\xi} = \frac{1}{b-c}$$

如果 ξ 服从均匀分布 $U(a, b)$ ，但是它只能在 (c, b) 内取得样本观测值，那么取得每一个样本观测值的概率

$$\begin{aligned}
 f(\xi|\xi > a) &= \frac{f(\xi)}{P(\xi > a)} \\
 &= \frac{(2\pi\sigma^2)^{-1/2} e^{-(\xi-\mu)^2/(2\sigma^2)}}{1 - \Phi(\alpha)} \\
 &= \frac{\frac{1}{\sigma} \phi(\frac{\xi - \mu}{\sigma})}{1 - \Phi(\alpha)}
 \end{aligned}$$

ξ 服从正态分布

$$P(\xi > a) = 1 - \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi(\cdot)$$

Φ 是标准正态分布条件概率函数

3、截断被解释变量数据模型的最大似然估计

$$y_i = \mathbf{B}'\mathbf{X}_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$y_i | \mathbf{X}_i \sim N(\mathbf{B}'\mathbf{X}_i, \sigma^2)$$

$$f(y_i) = \frac{\frac{1}{\sigma} \phi((y_i - \mathbf{B}'\mathbf{X}_i) / \sigma)}{1 - \Phi((a - \mathbf{B}'\mathbf{X}_i) / \sigma)}$$

$$\ln L = -\frac{n}{2}(\ln(2\pi) + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{B}'\mathbf{X}_i)^2 - \sum_{i=1}^n \ln \left(1 - \Phi \left(\frac{a - \mathbf{B}'\mathbf{X}_i}{\sigma} \right) \right)$$

$$\frac{\partial \ln L}{\partial \begin{pmatrix} \mathbf{B} \\ \sigma^2 \end{pmatrix}} = \sum_{i=1}^n \begin{pmatrix} \left(\frac{y_i - \mathbf{B}'\mathbf{X}_i}{\sigma^2} - \frac{\lambda_i}{\sigma} \right) \mathbf{X}_i \\ -\frac{1}{2\sigma^2} + \frac{(y_i - \mathbf{B}'\mathbf{X}_i)^2}{2\sigma^4} - \frac{\alpha_i \lambda_i}{2\sigma^2} \end{pmatrix} = \sum_{i=1}^n \mathbf{g}_i = \mathbf{0}$$

$$\alpha_i = (a - \mathbf{B}'\mathbf{X}_i) / \sigma$$

$$\lambda_i = \phi(\alpha_i) / (1 - \Phi(\alpha_i))$$

- 求解该1阶极值条件，即可以得到模型的参数估计量。
- 由于这是一个复杂的非线性问题，需要采用迭代方法求解，例如牛顿法。

4、例题—城镇居民消费模型

—截断样本数据

cons	incom	cons	incom	cons	incom
11123.84	13882.62	5064.340	6778.03	5759.210	7041.87
7867.530	10312.91	7356.260	9999.54	4948.980	6569.23
5439.770	7239.06	4914.550	6901.42	6023.560	7643.57
5105.380	7005.03	6069.350	8399.91	8045.340	8765.45
5419.140	7012.9	4941.600	6926.12	5666.540	6806.35
6077.920	7240.58	5963.250	7321.98	5298.910	6657.24
5492.100	7005.17	6082.620	7674.2	5400.240	6745.32
5015.190	6678.9	9636.270	12380.43	5330.340	6530.48
11040.34	14867.49	5763.500	7785.04	5540.610	7173.54
6708.580	9262.46	5502.430	7259.25		
9712.890	13179.53	7118.060	8093.67		

将这组样本看成是在 ≥ 4500 的条件下随机抽取得到

Equation Specification

Equation Specification:
Dependent variable followed by list of regressors.
cons c incom


Dependent variable censoring points:
A number, a series, a series expression, or blank for no censoring
Left: 4500
Right:
☒ Field is actual censoring value
☐ Field is zero/one indicator of censoring
☒ Truncated sample

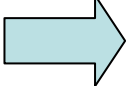

Distribution:
☒ Normal
☐ Logistic
☐ Extreme Value

Estimation Settings:
Method: CENSORED - Censored data (tobit)
Sample: 1 31

OK
Cancel
Options

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
------	-------	---------	-------	------	--------	----------	----------	-------	--------


 Dependent Variable: CONS
 Method: ML - Censored Normal (TOBIT)
 Date: 11/18/05 Time: 23:15
 Sample: 1 31


 Included observations: 31
 Truncated sample
 Left censoring (value) series: 4500
 Convergence achieved after 7 iterations 
 Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	200.7795	281.7348	0.712654	0.4761
INCOM	0.750072	0.032369	23.17269	0.0000

Error Distribution				
SCALE: C(3)	401.1625	54.96058	7.299096	0.0000

R-squared	0.948849	Mean dependent var	6433.182
Adjusted R-squared	0.945195	S.D. dependent var	1761.376
S.E. of regression	412.3449	Akaike info criterion	14.94657
Sum squared resid	4760793.	Schwarz criterion	15.08534
Log likelihood	-228.6718	Hannan-Quinn criter.	14.99181
Avg. log likelihood	-7.376511		

Left censored obs	0	Right censored obs	0
Uncensored obs	31	Total obs	31

将这组样本看成是在 ≥ 4000 的条件下随机抽取得到

View Procs Objects Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: CONS
Method: ML - Censored Normal (TOBIT)
Date: 11/19/05 Time: 08:52
Sample: 1 31
Included observations: 31
Truncated sample
Left censoring (value) series: 4000
Convergence achieved after 5 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	237.2539	267.6126	0.886557	0.3753
INCOM	0.746924	0.031111	24.00847	0.0000

Error Distribution				
SCALE:C(3)	391.2665	50.05756	7.816331	0.0000

R-squared	0.949144	Mean dependent var	6433.182
Adjusted R-squared	0.945511	S.D. dependent var	1761.376
S.E. of regression	411.1556	Akaike info criterion	14.96703
Sum squared resid	4733371.	Schwarz criterion	15.18381
Log likelihood	-228.9890	Hannan-Quinn criter.	15.01227
Avg. log likelihood	-7.386742		

Left censored obs	0	Right censored obs	0
Uncensored obs	31	Total obs	31

参数由 0.750072
变化为

似然函数值为
什么变小?

似然函数值由一
228.6718减小为

将这组样本看成是在 ≤ 11500 、 ≥ 4500 条件下随机抽取得到

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: CONS									
Method: ML - Censored Normal (TOBIT)									
Date: 11/19/05 Time: 09:05									
Sample: 1 31									
Included observations: 31									
Truncated sample									
Left censoring (value) series: 4500									
Right censoring (value) series: 11500									
Convergence achieved after 7 iterations									
Covariance matrix computed using second derivatives									
		Coefficient	Std. Error	Z-Statistic	Prob.				
C		99.92608	213.2677	0.318980	0.7497				
INCOM		0.763322	0.036832	20.72443	0.0000				
Error Distribution									
SCALE:C(3)		404.9458	56.52227	7.164359	0.0000				
R-squared	0.949923	Mean dependent var	6433.182						
Adjusted R-squared	0.946346	S.D. dependent var	1761.376						
S.E. of regression	407.9931	Akaike info criterion	14.91250						
Sum squared resid	4660834.	Schwarz criterion	15.05127						
Log likelihood	-228.1437	Hannan-Quinn criter.	14.95773						
Avg. log likelihood	-7.359475								
Left censored obs	0	Right censored obs	0						
Uncensored obs	31	Total obs	31						

参数由 0.750072 变化为

似然函数值为
什么增大?

似然函数值由 -
228.6718 增大为

将这组样本看成是在 ≥ 0 条件下随机抽取得到

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
<p>Dependent Variable: CONS <u>Method: ML - Censored Normal (TOBIT)</u> Date: 11/19/05 Time: 09:11 Sample: 1 31 Included observations: 31 Truncated sample Left censoring (value) at zero Convergence achieved after 4 iterations Covariance matrix computed using second order approximations</p>									
				Coefficient	Std. Error				
	C			238.4742	275.9763				
	INCOM			0.746817	0.032100				
				Error Distribution					
	SCALE: C(3)			390.7451	49.00000				
	R-squared			0.949146	Mean dependent var				
	Adjusted R-squared			0.945514	S.D. dependent var				
	S.E. of regression			411.1454	Akaike info criterion				
	Sum squared resid			4733134.	Schwarz criterion				
	Log likelihood			-228.9968	F-statistic				
	Avg. log likelihood			-7.386994	Prob(F-statistic)				
	Left censored obs			0					
	Uncensored obs			31					

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
<p>Dependent Variable: CONS <u>Method: Least Squares</u> Date: 11/19/05 Time: 09:13 Sample: 1 31 Included observations: 31</p>									
	Variable			Coefficient	Std. Error	t-Statistic		Prob.	
	C			238.4742	275.9763	0.864111		0.3946	
	INCOM			0.746817	0.032100	23.26504		0.0000	
	R-squared			0.949146	Mean dependent var			6433.182	
	Adjusted R-squared			0.947393	S.D. dependent var			1761.376	
	S.E. of regression			403.9945	Akaike info criterion			14.90302	
	Sum squared resid			4733134.	Schwarz criterion			14.99554	
	Log likelihood			-228.9968	F-statistic			541.2621	
	Durbin-Watson stat			1.220819	Prob(F-statistic)			0.000000	

结果与OLS相同
似然函数值减小

似然函数
值最小

5、为什么截断被解释变量数据模型不能采用普通最小二乘估计

- 对于截断被解释变量数据计量经济学模型，如果仍然把它看作为经典的线性模型，采用**OLS**估计，会产生什么样的结果？
- 因为 y_i 只能在大于 a 的范围内取得观测值，那么 y_i 的条件均值为：

$$\begin{aligned} E(y_i | y_i > a) &= \int_a^{\infty} y_i \phi(y_i | y_i > a) dy_i \\ &= B'X_i + \sigma \frac{\phi((a - B'X_i) / \sigma)}{1 - \Phi((a - B'X_i) / \sigma)} \end{aligned}$$

$$E(y_i | y_i > a) = \mathbf{B}' \mathbf{X}_i + \sigma \lambda(\alpha_i)$$

$$\alpha_i = \frac{\alpha - \mathbf{B}' \mathbf{X}_i}{\sigma}$$

$$\begin{aligned} \frac{\partial E(y_i | y_i > a)}{\partial \mathbf{X}_i} &= \mathbf{B} + \sigma \left(\frac{d\lambda_i}{d\alpha_i} \right) \frac{\partial \alpha_i}{\partial \mathbf{X}_i} \\ &= \mathbf{B} + \sigma (\lambda_i^2 - \alpha_i \lambda_i) \left(\frac{-\mathbf{B}}{\sigma} \right) \\ &= \mathbf{B} (1 - \lambda_i^2 + \alpha_i \lambda_i) \\ &= \mathbf{B} (1 - \delta(\alpha_i)) \end{aligned}$$

$$y_i | y_i > a = E(y_i | y_i > a) + u_i = \mathbf{B}' \mathbf{X}_i + \sigma \lambda(\alpha_i) + u_i$$

$$\text{Var}(u_i) = \sigma^2 (1 - \lambda_i^2 + \lambda_i \alpha_i) = \sigma^2 (1 - \delta_i)$$

- 由于被解释变量数据的截断问题，使得原模型转换为包含一个非线性项模型。
- 如果采用**OLS**直接估计原模型：
 - 实际上忽略了一个非线性项；
 - 忽略了随机误差项实际上的异方差性。
 - 这就造成参数估计量的偏误，而且如果不了解解释变量的分布，要估计该偏误的严重性也是很困难的。

6、Heckman两步修正法

- Sample Selection Bias as a Specification Error, *Econometrica* 47(1), 1979, P153-161

$$w_i = x_{1i}\beta_1 + \varepsilon_{1i} \quad (1)$$

市场工
资方程

$$e_i^* = x_{2i}\beta_2 + \varepsilon_{2i} \quad (2)$$

工作倾
向方程

$$E(\varepsilon_1) = 0, \quad E(\varepsilon_2) = 0, \quad \varepsilon_1, \varepsilon_2 \text{正相关}$$

$$E(\varepsilon_{1i} | e_i^* \geq 0) = E(\varepsilon_{1i} | \varepsilon_{2i} \geq -x_{2i}\beta_2)$$

$$E(w_i | x_{1i}, e_i^* \geq 0) = x_{1i}\beta_1 + E(\varepsilon_{1i} | \varepsilon_{2i} \geq -x_{2i}\beta_2)$$

$$E(w_i | x_{1i}, e_i^* \geq 0) = x_{1i}\beta_1 + \rho\sigma_1\lambda_i$$

$$w_i = x_{1i}\beta_1 + \rho\sigma_1\lambda_i + \mu_i$$

其中 ρ 为 $\varepsilon_1, \varepsilon_2$ 的相关系数,

σ_1 为 ε_{1i} 的标准差,

σ_2 为 ε_{2i} 的标准差。

$$\lambda_i = \frac{\phi\left(\frac{x_{2i}\beta_2}{\sigma_2}\right)}{\Phi\left(\frac{x_{2i}\beta_2}{\sigma_2}\right)}$$

$$w_i = x_{1i}\beta_1 + \rho\sigma_1\lambda_i + \mu_i$$

如何估计该模型？

- 第一步，用**probit**模型估计(2)，利用全部样本；利用估计结果，计算 λ_i 。
- 第二步，利用选择性样本，将 $(\rho \sigma_1)$ 作为一个待估计参数，估计模型，得到 β_1 的估计。

三、“归并”问题的计量经济学模型

1、思路

- 以一种简单的情况为例，讨论“归并”问题的计量经济学模型。即假设被解释变量服从正态分布，其样本观测值以**0**为界，凡小于**0**的都归并为**0**，大于**0**的则取实际值。如果 **y^*** 以表示原始被解释变量， **y** 以表示归并后的被解释变量，那么则有：

$$y = 0 \quad \text{当 } y^* \leq 0$$

$$y = y^* \quad \text{当 } y^* > 0$$

$$y^* \sim N(\mu, \sigma^2)$$

- 单方程线性“归并”问题的计量经济学模型为：

$$y_i = \mathbf{B}' \mathbf{X}_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

- 如果能够得到 \mathbf{y}_i 的概率密度函数，那么就可以方便地采用最大似然法估计模型，这就是研究这类问题的思路。
- 由于该模型是由**Tobin**于**1958**年最早提出的，所以也称为**Tobin**模型。

2、“归并”变量的正态分布


- 由于原始被解释变量 y^* 服从正态分布，有

$$P(y=0) = P(y^* \leq 0) = \Phi\left(-\frac{\mu}{\sigma}\right) = 1 - \Phi\left(\frac{\mu}{\sigma}\right)$$

$$P(y) = P(y^*) \quad \text{当 } y^* > 0$$

3、归并被解释变量数据模型的最大似然估计

$$\ln L = \sum_{y_i > 0} -\frac{1}{2} \left(\ln(2\pi) + \ln \sigma^2 + \frac{(y_i - B'X_i)^2}{\sigma^2} \right) + \sum_{y_i = 0} \ln \left(1 - \Phi \left(\frac{B'X_i}{\sigma} \right) \right)$$

- 该似然函数由两部分组成，一部分对应于没有限制的观测值，是经典回归部分；一部分对应于受到限制的观测值。
- 这是一个非标准的似然函数，它实际上是离散分布与连续分布的混合。
- 如何理解后一部分？为什么要求和？

- 如果样本观测值不是以**0**为界，而是以某一个数值**a**为界，则有

$$\begin{array}{ll} y = a & \text{当 } y^* \leq a \\ y = y^* & \text{当 } y^* > a \end{array}$$

$$y^* \sim N(\mu, \sigma^2)$$

估计原理与方法相同。

4、例题—城镇居民消费模型

—归并样本数据

11123.84

cons	incom	cons	incom	cons	incom
11000.00	13882.62	5064.340	6778.03	5759.210	7041.87
7867.530	10312.91	7356.260	9999.54	4948.980	6569.23
5439.770	7239.06	4914.550	6901.42	6023.560	7643.57
5105.380	7005.03	6069.350	8399.91	8045.340	8765.45
5419.140	7012.9	4941.600	6926.12	5666.540	6806.35
6077.920	7240.58	5963.250	7321.98	5298.910	6657.24
5492.100	7005.17	6082.620	7674.2	5400.240	6745.32
5015.190	6678.9	9636.270	12380.43	5330.340	6530.48
11000.00	14867.49	5763.500	7785.04	5540.610	7173.54
6708.580	9262.46	5502.430	7259.25		
9712.890	13179.53	7118.060	8093.67		

11040.34

Censored (11000) 估计

Dependent Variable: CONS

Method: **ML** - Censored Normal (TOBIT)

Date: 11/29/04 Time: 17:25

Sample: 1 31

Included observations: 31

Right censoring (value) series: **11000**

Convergence achieved after 8 iterations

Covariance matrix computed using second order method

	Coefficient	Standard Error
C	25.62933	30.00000
INCOM	0.775212	0.00000
Error Degrees of Freedom: 29		
SCALE:C(3)	396.7539	5.00000
R-squared	0.949968	
Adjusted R-squared	0.946394	
S.E. of regression	404.4725	
Sum squared resid	4580745.	
Log likelihood	-215.7708	
Avg. log likelihood	-6.960348	

Left censored obs 0

Uncensored obs 29

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: CONS									
Method: Least Squares									
Date: 11/19/05 Time: 09:36									
Sample: 1 31									
Included observations: 31									
<hr/>									
<hr/>									
Variable		Coefficient	Std. Error	t-Statistic	Prob.				
<hr/>									
C		283.3025	273.2348	1.036847	0.3084				
INCOM		0.740774	0.031782	23.30833	0.0000				
<hr/>									
R-squared		0.949325	Mean dependent var	6427.886					
Adjusted R-squared		0.947578	S.D. dependent var	1746.959					
S.E. of regression		399.9813	Akaike info criterion	14.88305					
Sum squared resid		4639566.	Schwarz criterion	14.97557					
Log likelihood		-228.6873	F-statistic	543.2782					
Durbin-Watson stat		1.241862	Prob(F-statistic)	0.000000					
<hr/>									
<hr/>									

参数估计结果、似然函数值都与OLS估计差异较大。为什么似然函数值大于OLS估计？

Right censored obs 2

Total obs 31

Censored (12000) 估计—与OLS相同

Dependent Variable: CONS

Method: **ML** - Censored Normal (TOBIT)

Date: 11/30/04 Time: 09:05

Sample: 1 31

Included observations: 31

Right censoring (value) series: **12000**

Convergence achieved after 4 iterations

Covariance matrix computed using second order

	Coefficient	Standard Error
C	283.3025	273.2348
INCOM	0.740774	0.031782
Error term		
SCALE:C(3)	386.8636	4639566.
R-squared	0.949359	
Adjusted R-squared	0.945742	
S.E. of regression	406.9253	
Sum squared resid	4636469.	
Log likelihood	-228.6873	
Avg. log likelihood	-7.377011	
Left censored obs	0	Right censored obs
Uncensored obs	31	Total obs

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
Dependent Variable: CONS									
Method: Least Squares									
Date: 11/19/05 Time: 09:36									
Sample: 1 31									
Included observations: 31									
Variable		Coefficient	Std. Error	t-Statistic	Prob.				
C		283.3025	273.2348	1.036847	0.3084				
INCOM		0.740774	0.031782	23.30833	0.0000				
R-squared		0.949325	Mean dependent var	6427.886					
Adjusted R-squared		0.947578	S.D. dependent var	1746.959					
S.E. of regression		399.9813	Akaike info criterion	14.88305					
Sum squared resid		4639566.	Schwarz criterion	14.97557					
Log likelihood		-228.6873	F-statistic	543.2782					
Durbin-Watson stat		1.241862	Prob(F-statistic)	0.000000					

5、实际模型中的Truncation与Censored

- 时间序列样本，不考虑。
- 截面上的全部个体作为样本，不考虑**Truncation**。
- 按照抽样理论选取截面上的部分个体作为样本，不考虑**Truncation**。
- 按照特定的规则选取截面上的部分个体作为样本，必须考虑**Truncation**。
- 截面数据作样本，根据样本观测值的经济背景，决定是否考虑**Censored**。