Tobit 模型

一、实验基本原理

当被解释变量为截取数据时，我们虽然有全部的观测数据，但对于某些观测数据，被解释变量y_i被压缩在一个点上了。此时，y_i的概率分布就变成由一个离散点与一个连续分布所组成的“混合分布”（mixed distribution）。

假设真实情况为$y_i = x_i'\beta + \varepsilon_i$（$y_i$为不可观测的潜变量），$\varepsilon_i | x_i \sim N(0, \sigma^2)$。可以观测到的变量 $y_i^* = \begin{cases} y_i & \text{如果 } a < y_i < b \\ a & \text{如果 } y_i \le a \\ b & \text{如果 } y_i \ge b \end{cases}$

在这种情况下，可以证明，如果用 OLS 来估计，无论使用的是整个样本，还是去掉离散点后的子样本，都不能得到一致的估计。

下面，为了书写方便，我们用左端截取来说明实验原理。假定左端截取的截取点为 c，那么，$y_i > c$时的概率密度依然不变，为$\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma}), \forall y_i > c$。而$y_i \leq c$时的分布却被挤到一个点$y_i^* = c$上了，即$P(y_i^* = c|x) = 1 - P(y_i > c|x) = \Phi[(c - x_i'\beta)/\sigma]$。从而，该混合分布的概率密度函数可以写为：

$$f(y_i^*|x) = [\Phi\left(\frac{c - x_i'\beta}{\sigma}\right)]^{1(y_i^*=c)}[\frac{1}{\sigma}\phi(\frac{y_i - x_i'\beta}{\sigma})]^{1(y_i^*>c)}$$

其中，1(.)为“示性函数”（indicator function），即如果括号里的表达式为真，则取值为 1；否则，取值为 0。由此，可以写出整个样本的似然函数，然后使用 MLE 来估计。

- **二、案例分析：实验内容及数据来源**
- 我们要研究汽车重量对每加仑耗油下行驶的路程的影响，使用本书附带光盘的data文件夹下的“usaauto.dta”工作文件。主要变量有：mpg=每加仑汽油所行驶的英里数，weight=汽车的重量等。
- 利用“usaauto.dta”的数据，我们会讲解截取回归的操作及预测。
- 需要说明的是，这个数据本身不是截取数据，但为了展示tobit回归的相关操作，我们会对数据进行处理，然后讲解相关命令的操作。

- **实验操作指导**
- **（1） 普通最小二乘回归**
- 为了与数据处理后的tobit回归进行比较，我们这里先进行OLS回归。
- 键入命令：
- generate wgt=weight/1000
- regress mpg wgt
- 其中，第一步为生成一个新变量wgt，其值为变量weight的1/1000。第二步为mpg对wgt的回归。

- **（2） 截取回归的操作**
- 截取回归的基本命令为：
- tobit depvar [indepvar] [if] [in] [weight], ll[(#)] ul[(#)] [options]
- 其中，tobit代表“截取回归”的基本命令语句，depvar代表被解释变量的名称，indepvar代表解释变量的名称，if代表条件语句，in代表范围语句，weight代表权重语句，options代表其他选项。可用的options选项包括offset()、vce()、level()等，其含义和断尾回归处相同。此外，ll表示左截取点，ul表示右截取点，这两个选项至少需要设定一个，可以同时设定。对于ll和ul选项，可以设定截取点的值，也可以不设定。当只键入ll或ul选项而不设定截取点的值时，tobit命令会自动设定被解释变量的最小值为左截取点（当ll选项被设定时），被解释变量的最大值为右截取点（当ul选项被设定时）。

- 下面，我们通过例子来加深对命令的理解。
- 在“usaauto.dta”工作文件中，变量mpg的最小值为12，最大值为41。假定我们的数据为截取数据，当mpg的真实值小于或等于20时，我们只知道其不超过20，而不知道具体的取值。
- 我们先对数据进行变换，使用命令：
- replace mpg=20 if mpg<=20
- 即，将小于或等于20的mpg值设为20。然后，我们进行tobit回归：
- tobit mpg wgt, ll
- 这里，要注意选项是两个小写的字母el，而不是数字1。

- 事实上，我们没有必要先使用replace命令，直接使用选项ll(20)就可以得到图11.5的结果。前面之所以要对数据进行变换，主要是为了提醒读者，tobit命令是用于截取数据的。在实际的研究中，如果数据类型非截取，直接使用regress就可以了；只有在数据为截取数据时，才有必要使用tobit。

- 3 tobit回归的预测

对截取回归模型进行预测的基本命令格式和断尾回归相同，为：

predict [type] newvar [if] [in] [, statistic nooffset]

predict [type] {stub * | newvar$_{reg}$ newvar$_{lnsigma}$} [if] [in], scores

可用的选项及其解释亦与断尾回归处相同，在此不再赘述。

表 11.3 给出了主要的 statistic 统计量及其含义。

表 11.3　断尾回归预测中的 statistic 选项

xb	线性预测（默认选项）
stdp	拟合的标准误（standard error of the prediction）
stdf	预测的标准误（standard error of the forecast）
pr(a,b)	$\Pr(a < y_i < b)$
e(a,b)	$E(y_i \mid a < y_i < b)$
ystar(a,b)	$E(y_i^*), y_i^* = \max\{a, \min(y_i, b)\}$

- 小结
- (1)Tobit y x,ll(0)
- 表示取y>0的数据进行回归分析；
- (2)Tobit y x,ll(0) ul(100)
- 表示取0<y<100的数据进行回归分析。
- (3)predict yhat,xb(表示y的预测值)
- (4)predict p,stdp（表示拟合的标准误,即均值预测标准误）
- (5)predict f,stdf（表示预测的标准误，即个别值预测标准误）
- (6) predict pr, pr(20,40)（pr(20<y<40)）
- (7)predict yyhat,e(20,40)(E(y|20<y<40))
- (8)predict ystar (E(y*),y*=max(a,min(y,b)))