受限因变量模型

主要内容

- 断尾回归模型
- Tobit模型

实验1：断尾回归模型

- 实验基本原理

对一个随机变量 y 而言，当其断尾后，概率密度函数会发生变化。假如 y 原来的概率密度为 f(y)，则左端断尾后的条件密度函数为：

$$f(y|y>c)=\begin{cases}\dfrac{f(y)}{P(y>c)} & \text{如果 } y>c\\ 0 & \text{如果 } y\le c\end{cases}$$

可以证明，存在断尾的情况下，普通最小二乘是有偏的。

但 MLE 可以得到一致的估计。例如，当被解释变量左端断尾时，其条件密度函数为：

$$f(y_i|y_i>c,x_i)=\frac{\frac{1}{\sigma}\phi[(y_i-x_i'\beta)/\sigma]}{1-\Phi[(c-x_i'\beta)/\sigma]}$$

其中，ϕ是标准正态分布的概率密度函数，Φ是标准正态分布的累积分布函数。由此，可以计算出整个样本的似然函数，然后使用极大似然估计法进行估计。

注释：

$$f(\mathrm{y})=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\mathrm{y}-\mu)^2}{2\sigma^2}}=\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{(\frac{y-\mu}{\sigma})^2}{2}}=\frac{1}{\sigma}\Phi(\frac{y-\mu}{\sigma})$$

$$p(\mathrm{y}>\mathrm{c})=p(\mathrm{x}\beta+\mu>\mathrm{c})=\mathrm{p}(\mu>c-\mathrm{x}\beta)=1-\mathrm{p}(\mu\le c-\mathrm{x}\beta)$$

$$=1-\Phi(\frac{c-x\beta}{\sigma})$$

$$\ln L=-\frac{n}{2}(\ln(2\pi)+\ln\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{2}(\mathrm{Y}_i-\mathrm{X}_i\beta)^2-\sum_{i=1}^{n}(1-\Phi(\frac{c-X_i\beta}{\sigma}))$$

- **实验内容及数据来源**
- 文件名“laborsupply.dta”工作文件给出了1975年妇女劳动供给的一些数据，主要变量有：lfp=各妇女在1975年是否工作（该变量取1表示该妇女在1975年有工作），whrs=妇女的工作时间，kl6=年龄小于6岁的孩子个数，k618=年龄在6岁到18岁之间的孩子个数，wa=妇女的年龄，we=妇女的受教育年限。很显然，当某妇女在1975年没有工作时，我们观察到的该妇女的工作时间为0。
- 利用这些数据，我们要研究各个因素对妇女劳动时间的影响，并讲解断尾回归模型的拟合与预测。

- **实验操作指导**
- **1 利用普通最小二乘法进行回归**
- 我们首先利用这些数据进行普通最小二乘回归。键入以下命令：
- regress whrs kl6 k618 wa we if whrs > 0
- 其中，被解释变量为whrs，解释变量为kl6、k618、wa和we，条件语句if表明，我们对妇女工作时间大于0的数据进行回归。
- 这里，我们主要是为了和后面断尾回归的结果进行比较。

- **2 断尾回归的操作**
- 断尾回归的基本命令为：
- truncreg depvar [indepvar] [if] [in] [weight] [,options]
- 其中，truncreg代表“断尾回归”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。表11.2显示了各options选项及其含义。

表 11.2　断尾回归中 options 的内容表

noconstant	模型不包含常数项
ll(varname|#)	左端断尾的下限（lower limit）
ul(varname|#)	右端断尾的上限（upper limit）
offset(varname)	约束变量 varname 的系数为 1
constraints(constraints)	进行约束回归
collinear	保留多重共线性变量
level(#)	设置置信度，默认值 95%
vce(type)	设置估计量的标准差，常用的主要有：cluster，robust，bootstrap，oim，jackknife 等
noskip	进行模型整体显著性的似然比检验

- 对于“laborsupply.dta”的数据而言，1975年没有工作的妇女的劳动时间都被设定为0，事实上也就是其具体劳动时间的数据没有被统计到，这样，我们可以进行一个左端断尾的回归，命令如下：
- truncreg whrs kl6 k618 wa we, ll(0)
- 这里，选项ll(0)设定左端断尾的下限为0。

3 断尾回归的预测

对断尾回归模型进行预测的基本命令格式如下：

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

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

其中，第一种预测命令中，predict 代表预测的基本命令语句，newvar 代表生成的新变量的名称，type 代表新变量的类型，if 代表条件语句，in 代表范围语句，statistic 代表要预测的统计量。

第二种命令是对方程水平的得分变量的预测。stub 代表生成的新变量的前缀，而预测的第一个新变量为$\frac{\partial \ln L}{\partial x_i'\beta}$，第二个新变量为$\frac{\partial \ln L}{\partial \sigma}$。

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)	$E(y_i^*), y_i^* = \max\{a, \min(y_i, b)\}$

- 下面，我们结合本例对选项进行具体的说明。
- 1.拟合的标准误（stdp）也被称作standard error of the fitted value，可以将其看做观测值处于均值水平下的标准误。预测的标准误（stdf）也被称作the standard error of the future or forecast value，指的是每个观测值的点预测的标准误。根据两种标准误的计算公式可知，stdf预测的标准误总是比stdp预测的要大。
- 我们对上面的断尾回归进行默认预测以及stdp和stdf的预测，采用如下命令：
- predict y
- predict p, stdp
- predict f, stdf
- list whrs y p f in 1/10
- 其中，第一步为默认预测，并将预测值命名为y；第二步预测的是拟合的标准误，并将预测值命名为p；第三步预测的是预测的标准误，并将其命名为f；最后一步列出原序列值whrs和各预测值的前10个观测值。

2.pr(a, b)计算$y_i|x_i$在区间(a, b)被观测到的概率，也就是Pr $(a < x_i'\beta + \varepsilon_i < b)$。其中，a 和 b 可以是数字或变量名。我们用 lb 和 ub 来表示变量名。pr(20,50)计算的是Pr $(20 < x_i'\beta + \varepsilon_i < 50)$，pr(lb,ub)计算的是Pr $(lb < x_i'\beta + \varepsilon_i < ub)$。如果我们把 a 设定为缺失值“.”，则表示−∞；把 b 设定为缺失值“.”，则表示+∞。

3.e(a, b)计算的是$E(x_i'\beta + \varepsilon_i | a < x_i'\beta + \varepsilon_i < b)$，也就是说给定$y_i|x_i$在开区间(a, b)的条件下，$y_i|x_i$的期望值。a 和 b 的设定与在选项pr(a, b)处相同。

4.ystar(a,b)计算的是$E(y_i^*)$。当$x_i'\beta + \varepsilon_i \le a$时，$y_i^* = a$；当$x_i'\beta + \varepsilon_i \ge b$时，$y_i^* = b$；其余情况下，$y_i^* = x_i'\beta + \varepsilon_i$。a 和 b 的设定与在选项pr(a, b)处相同。

5. 选项 nooffset 只有在之前的断尾回归中设定了 offset()选项时才有意义。预测时加上 nooffset，则会忽略模型拟合时所设定的 offset()选项。从而，线性预测汇报的是$x_i'\beta$而非$x_i'\beta + offset_i$。

实验2：截取回归模型

- 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。

- **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，使用文件名“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)))