第十二讲 非经典截面数据计量模型

一、离散被解释变量模型

主要内容

- 1-二值选择模型
- 2-多值选择模型

第1节 二值选择模型

- 一 实验基本原理

1.二值选择模型

假设研究人们买房的问题时，人们有两种选择；y=1(买房)或者 y=0(不买)，然而是否买房取决于人们的收入，对房价的基本预期，结婚与否等影响因素。假设把这些因素作为解释变量：$y_i = \mathbf{x_i}'\boldsymbol{\beta} + \varepsilon_i \quad (i = 1, \dots, n)$ （这里 y 是不可观测的潜变量）。

由此模型得到的 y 估计值可能会出现大于 1 或者小于 0 的不合实际的情况，为了使估计值在[0,1]范围内，考虑 y 的两点分布函数：

$$\begin{cases} P(y=1|x) = F(x,\beta) \\ P(y=0|x) = 1 - F(x,\beta) \end{cases}$$

通过选择合适的 F 函数形式（随机变量的累积分布函数）来保证$0 \le \hat{y} \le 1$。由于

$E(y|x) = 1 * P(y=1|x) + 0 * P(y=0|x) = P(y=1|x)$，所以$\hat{y}$可以理解为“y=1”的概率。

若 F 为标准正态分布函数，那么有下面的等式成立，此模型就是 probit 模型：

$$P(y=1|x) = F(x,\beta) = \varphi(x'\beta) \equiv \int_{-\infty}^{x'\beta} \emptyset(t)dt$$

若 F 为“逻辑分布”分布函数，那么下面的等式就是 logit 模型：

$$P(y=1|x) = F(x,\beta) = \Lambda(x'\beta) \equiv \frac{e^{x'\beta}}{1+e^{x'\beta}}$$

由于逻辑分布的累积分布函数有解析表达式，而标准正态分布没有，所以 logit 模型的计算相比 probit 模型简单。

以 logit 为例，通过下面的变形整理 logit 模型可以整理得到以下等式，

$$\ln f(y_i|x_i,\beta) = y_i \ln[\Lambda(x'\beta)] + (1-y_i)\ln\left[1-\Lambda\left(x'\beta\right)\right]$$

则 logit 模型的样本对数似然函数为，

$$\ln L(\beta|y,x) = \sum_{i=1}^{n} y_i \ln[\Lambda(x'\beta)] + \sum_{i=1}^{n} (1-y_i)\ln\left[1-\Lambda\left(x'\beta\right)\right]$$

使用 Stata 来最大化此非线性函数就可以求的模型的相关估计量。

相同的样本使用 logit 模型和 probit 模型估计出的参数估计值由于两模型假定的分布函数不同，两个参数估计值是不可比的。对此 Amemiya 提出，同一个样本的 logit 模型与 probit 模型的参数估计值大概有如下关系：$\beta_{logit} \approx 1.6 * \beta_{probit}$

另外可以使用 Stata 计算两个模型的边际效应，然后进行比较。注意，在这个非线性模型估计中，$\hat{\beta}_{MLE}$并不表示边际效应，只是表示解释变量影响的方向。

2.二值选择模型的异方差问题

由于标准的 probit 模型或者 logit 模型的扰动项是服从同方差假设的，但是很多回归模型是存在异方差的。这时可以使用 Stata 进行“似然比检验”（LR）来检测异方差的存在。

以 probit 模型为例，“同方差”原假设 H_0 为 $P(y_i = 1|x_i) = \varphi(x'\beta/\sigma)$，此时 $\sigma = 1$，而异方差的替代假设 H_1 为 $P(y_i = 1|x_i) = \varphi(x'\beta/\sigma_i)$ ，其中 $\sigma_i \equiv VAR(\varepsilon_i)$

LR 检验的结果若接受原假设，则使用同方差 probit 模型，若拒绝则使用异方差 probit 模型。

- 二 实验内容和数据来源
- 根据某统计资料，得到美国妇女就业的数据统计集，形成数据文件“womenwork.dta”，用来研究影响美国妇女就业的因素。被解释变量是work（就业work=1，不就业work=0），解释变量是age（年龄），married（婚否），children（子女数），education（教育年限）。完整的数据在本书附带光盘里的data文件夹的“womenwork.dta”工作文件中。

- 利用以上数据，建立合适模型对就业的影响因素进行计量分析，由于被解释变量取值有两个可以建立二值选择模型来分析问题。

- # 三 实验操作指导
- **1.建立logit模型分析**
- （1）使用logit模型回归
- Stata中使用logit模型回归的命令语句格式如下：
- logit y x1 x2 … [if] [in] [weight] [,options]
- 该命令中logit表示使用logit模型进行回归，相应y表示模型的被解释变量，x表示模型的解释变量，if表示logit的回归条件，in表示回归的范围，weight表示给观测值的加入权重，options的内容如下表所示：

Model	
noconstant	无常数项
offset（varname）	约束 varname 的系数为 1
asis	保留完全预测变量
SE/Robust	
vce(vcetype)	vcetype 可能包括 oim, robust, cluster clustvar, bootstrap，或者 jackknife
Reporting	
level(#)	设置置信度，默认值是 95
or	输出机会比
max options	
maximize_options	控制最大化过程；很少用到
nocoef	不输出系数表格栏；很少用

- 本实验中，在Stata命令窗口中输入如下命令。
- use womenwork, clear
- 输入此命令来打开需要的数据文件。
- logit work age education married children
- 输入此命令对被解释变量为work，解释变量为age、education、married、children的模型使用logit模型进行回归估计。

- 在这个回归结果图中log likelihood即对数似然值，不断的试错迭代是logit模型的估计方法，在逐步进行回归时，通过比较不同模型的-2LL判断模型的拟合优度，选择取值更小的模型。LR chi2(4)是卡方检验的统计量，也就是回归模型无效假设所对应的似然比检验量；其中4为自由度，Prob>chi2 是其对应的P值，在这个估计结果显示以p=0显著说明模型的有效性。其实这两个指标与线性回归结果中F统计量和P值的功能是大体一致的。另外结果中的Pseudo R2是准R2，虽然不等于R2,但可以用来检验模型对变量的解释力，因为二值选择模型是非线性模型，无法进行平方和分解，所以没有，但是准衡量的是对数似然函数的实际增加值占最大可能增加值的比重，所以也可以很好的衡量模型的拟合准确度。此logit模型中拟合优度为0.1882。
- coef是自变量对应的系数估计值，OLS通过t检验来检验估计量是否显著，logit模型通过z检验来判断其显著性；通过z检验结果可以看到此模型中系数均以p=0显著不为0。

- （2）由于估计系数不像线性模型能够表示解释变量的边际效应，所以Stata中有额外的命令语句来计算解释变量的边际效应：
- mfx [compute] [if] [in] [,options]
- 此命令语句中mfx表示对回归之后的模型计算解释变量的边际效应，其中options内容如下表所示：

predict（predict_option）	为 predict_option 计算边际效应
varlist(varlist)	为 varlist 计算边际效应
dydx	计算边际效应，是默认设置
eyex	以 dlny/dlnx 形式计算弹性
dyex	以 dy/dlnx 形式计算弹性
eydx	以 dlny/dx 形式计算弹性
nodiscrete	把虚拟变量视为连续变量
nose	不计算标准差
at(atlist)	在这些值处计算边际效应

- 本实验中，在进行logit模型回归估计后，在Stata命令窗口中输入如下命令：
- mfx
- 此命令计算模型回归之后，解释变量取值在样本均值处的边际效应。

- 此输出结果显示了每一个解释变量的平均边际影响，另外读者可以自己设定计算在边际影响的点，其原理就是命令语句options中的at(atlist)将其具体化，例如“mfx, at (x1=0)”表示计算x1取值为0，其他解释变量取值在样本均值处的边际效应；而“mfx”默认是在所有解释变量在样本平均值处的边际效应。

- （3）计算模型预测的百分比来计算模型的拟合优度。

如果发生概率的预测值$\hat{y} \geq 0.5$，那么认为其$\hat{y}$=1；若$\hat{y} < 0.5$，那么$\hat{y}$=0；将预测发生值与实际值进行比较就可以得到准确预测的百分比。当然这里的门限值 0.5 在 Stata 中，读者可以根据自己的需要进行特别设定，但是 Stata 中默认的门限值是 0.5。

Stata 中执行该命令的语句为：

estat classification [if] [in] [weight] [,all] [cutoff(#)]

此命令语句表示根据预测概率进行分类，if 表示分类时观测值的条件，in 表示取值的范围等，weight 表示观测值的权重，all 表示忽略 if 和 in 的设定对所有观测值进行分类，cutoff 表示门限值（默认值为 0.5)

本实验中在命令窗口中输入以下命令语句，可以得到图 9.3 的运行结果：

estat clas

其中，结果图中 1177 和 296 所在位置是指正确预测所在类别的个数；分类依据也在结果中间显示 sensitivity（敏感性）=pr($\hat{y}_i = 1|y_i = 1$)，则 87.64%=1177/1343；specificity（特异性）= pr($\hat{y}_i = 0|y_i = 0$)，则 45.05%=296/657。

结果图的最后一行显示正确预测百分比为 73.65%，这个数字也刻画出了 logit 模型的拟合优度。

- 如果要检验这个分类的依据或者要获得每个预测值，可以利用此二值模型进行预测分析，Stata中二值选择模型的预测的命令语句如下所示：
- predict [type] newvar [if] [in] [,single_options]
- 其中predict是表示对模型进行预测的命令；newvar表示预测新变量的名称，type可以表明设定新变量的类型；if和in表示对此预测设定的条件和范围；single_options的内容以下表所示：

single_options	描述
Main	
xb	线性预测
stdp	计算预测的标准差
score	似然函数对 xb 的一阶导数
pr	概率预测，此为默认选项
Options	
nooffset	预测值不包括 offset 和 exposure 选项所设定的变量

- 本实验中，在Stata命令窗口中输入如下预测命令，可以得到预测结果图：
- predict p1, pr
- 此命令可以获得此模型的个体估计的值并记为新变量p1
- list work p1
- 此命令可以将实际值与估计值对应罗列，对比看到预测值和实际值的一致程度。

前面已经解释到，二值选择模型中，被解释变量的估计值是其取值 1 的概率。其中按照若实际值 work=1 且 p1≥ 0.5则说明预测是正确的，否则是错误的，读者可以手动从结果图中数一下，然后得到正确预测的百分比与上图的结果是相同的。

- （1）ROC曲线（受试者操控曲线）
- 此曲线是指图9.3提到的敏感性与（1-特异性）的散点图，即预测值等于1的准确率与错误率的散点图。Stata中绘画该ROC曲线命令语句为:
- lroc [x] [if] [in] [weight] [,options]
- 其中lroc表示绘图ROC曲线命令，if和in表示对绘制图时的条件和范围的设定，weight表示对观测值的权重设定，另外命令中的自变量x不能单独使用，必须与options中beta(matname)同时使用，而options的内容如下表所示：

All	对所有观测值作图
Nograph	不显示图形
beta(matname)	模型估计量保存在行矩阵 matname 中

- 本实验中，在以上工作后，在命令窗口中输入如下命令绘制ROC曲线图
- lroc

- 因为准确率就是曲线下面的面积，读此图可以看到ROC曲线是完全在45度直线上面，所以准确率高于错误率，即准确率大于0.5。此图曲线下方面积=0.7806，就是预测的准确率是0.7806。

- （2）goodness-of-fit拟合优度检验
- 此检验是考察该模型对所用数据的拟合优度，在Stata中命令语句为：
- estat gof [if] [in] [weight] [,group(#) all outsample table]
- 其中，if和in表示对检测拟合优度时的条件和范围的设定，weight表示对观测值的权重设定，group（#）表示使用合理的#分位数进行检验；all表示对所有观测值进行检验，若无后面可选项则默认就是all;outsample表示对估计区间外的样本调整自由度，table表示各组列表。
- 本实验中在Stata命令窗口输入如下命令检验此模型的拟合优度，然后可以得到检验结果：
- estat gof

- 读此图的方法是P值越大，说明模型的拟合优度越好。

- **2.建立probit模型分析**
- 前面是使用logit模型对womenwork.dta进行分析，现在使用probit模型对此问题进行分析。两种方法在Stata中的操作是很一致的。
- 在Stata命令窗口中输入如下命令：
- use womenwork, clear
- 使用此命令打开所需要文件。
- probit work age education married children
- 此命令表示使用probit模型进行回归。
-
- 此图的解读方法与Logit模型结果图是完全一样的，probit模型估计结果显示系数估计值相比logit估计值发生了很大变化，且均显著通过了模型系数的显著性检验；另外模型的准R2是0.1889，相比logit模型稍有改进。

- 由于logit与probit模型得出的参数估计值不可直接比较，根据本节开始介绍的原理已了解到两模型的边际效应可以比较。Stata中probit模型的边际效应得出方法与logit是相同的。
- 在Stata命令窗口中输入如下命令计算probit模型回归后解释变量在样本均值处的边际效应:
- mfx

- 可以看到与前面的logit模型比较，两模型分析的边际效应是大致相同的。然后来计算probit模型的拟合优度，具体操作方法也与logit模型是一致的。

- 计算准确预测百分比，Stata命令窗口输入如下命令：
- estat clas
- 此图的解读方法与上面logit模型得到的是完全一样的，显然可以得到：sensitivity（敏感性）=87.64%，specificity（特异性）=45.05%，correctly classified（正确预测百分比）=73.65%。可以看到，这个结果与logit模型是完全一致的。
- 另外为了检验这个结果，可以同样输入如下命令：
- predict p2, pr
- 此命令可以获得此模型的个体估计的值并记为新变量p2
- list work p2

此命令可以将实际值与估计值对应罗列，对比预测值和实际值的一致程度。若 work=1 且$p2 \geq 0.5$则说明预测是正确的，否则是错误的。检测得到正确预测的百分比与上图的结果是否相同。

- 其次是使用ROC曲线来检测预测的准确度，在Stata命令窗口中输入如下命令，可以得到ROC曲线：
- lroc

- 此图的读法与logit的ROC图是一致的，由于logit模型与probit模型的sensitivity与specificity是相同的，那么ROC曲线一定是相同的，且曲线下方的面积同样是0.7806。

- 最后是godness-of-fit拟合优度检验，在Stata命令窗口中输入如下命令：
- estat gof

此检验显示 p 值是 0.8650，相比 logit 模型的此检验结果，$p_{logit} > p_{probit}$ ，即 logit 模型对样本数据的拟合优度更好。

- **3.二值选择模型的异方差问题**
- Stata中对probit二值选择模型进行异方差检验和回归的命令语句如下：
- hetprob y x1 x2 …[if] [in] [weight] , het（varlist [offset(varname)]) [,options]
- 其中hetprob表示对模型进行异方差probit模型估计和异方差检验，if和in表示对检测拟合优度时的条件和范围的设定，weight表示对观测值的权重设定，选择项 het(varilist)是影响扰动项的变量清单，在该命令语句的输出结果里，会汇报LR检验的结果，据此判断是否应该使用此异方差模型，options的内容如下表所示：

options	描述
Model	
noconstant	无常数项
offset（varname）	约束此变量的系数为 1
Asis	保留完全预测变量
constraints（constraints）	应用特定的线性约束
collinear	保留多重共线性预测变量
SE/Robust	
vce(vcetype)	vcetype 可能包括 oim, robust, cluster *clustvar* opg bootstrap, 或者 jackknife
Reporting	
level(#)	设置置信度，默认值 95
noskip	进行似然比检验
nolrtest	进行 wald 检验

- 本实验中，在Stata命令窗口中输入如下命令进行异方差模型估计和检验，可以得到图9.12的运行结果：
- hetprob work age education married children, het (age education married children)

- 结果显示LR检验的结果是接受原假设，即模型不存在异方差问题。所以回归不应使用异方差回归模型，可以直接应用probit模型进行估计。

第2节 多值选择模型

- 一 实验基本原理
- 1.多值选择模型
- 有时候人们面临的选择是多个的，比如交通选择，入读大学的选择等等。假设个体可以选择的y=1,2,3,…,J ,其中J是正整数。当研究的被解释变量是这样多值离散的，建立的模型就是多值选择模型，而当J=2时，就是上节所说的probit或者logit模型。
-

$$P(y_i = j|x) = \begin{cases} \dfrac{e^{x_i'\beta_j}}{1+\sum_{j=1}^{J} e^{x_i'\beta_j}} & (j = 2,\dots,J) \\ \dfrac{1}{1+\sum_{j=1}^{J} e^{x_i'\beta_j}} & (j = 1) \end{cases}$$

其中“j=1”所对应的一组为参照组，且各项选择概率之和为 1，这个模型就是多值选择 logit 模型。

为估计多值 logit 模型，得到该模型第 i 个个体的对数似然函数为：

$$\ln L_i(\beta_1, \ldots, \beta_J) = \sum_{j=0}^{J} \mathbf{1}(y_i = j) * lnP(y_i = j|\mathbf{x})$$

其中，$\mathbf{1}(\cdot)$表示示性函数，若括号内的条件成立则该函数取值 1，否则取值 0。将所有个体的对数似然函数加总即得到整个样本的对数似然函数，然后最大化此函数值得到$\hat{\beta}_{MLE}$。

在多值选择模型下，因为 Probit 模型需要对多元正态分布进行评价，所以应用受到限制，所以应用最多的是多值 logit 模型，所以这里仅介绍多值 logit 模型。

2、相对风险（相对机会比）

介绍这个概念是因为$\hat{\beta}_{MLE}$代表了解释变量单位的增加引起的是相对风险的边际变化。

若 j=1 为参照组，相对风险 $= \ln[\frac{P(y=j)}{P(y=1)}] = \boldsymbol{x_i}'\boldsymbol{\beta_j}$，在多值选择模型中，参照组的选择很重要，因为对估计量的解释是以参照组为转移的。

另外多项选择模型必须满足“无关选择的独立性”，即将多值选择模型中任意选择两个，就会是二值 logit 模型。这在实践中较难满足，是其重大缺点。

- 二 实验内容和数据来源
- 本实验来自某统计资料，统计在购物时所选品牌与性别、年龄的关系。变量主要有brand（品牌），female（性别），age（年龄）。完整的数据在本书附带光盘data文件夹下“brand.dta”中。
- 本实验用此数据来以female和age为解释变量，brand为被解释变量，brand的取值是离散的，且有三个取值，应建立多值选择模型进行相关分析。

- 二实验操作指导
- **1.选择合理模型**
- 在Stata中将数据按照某个或某几个变量进行分类并按这个变量获得其频数分布的命令如下：
- tab varlist
- 其中varlist表示按照其分类的变量或者变量组合。
- 在本实验中，打开数据文件并将数据按brand取值分类，在Stata命令窗口中输入如下命令
- use brand ,clear
- tab brand
- 读图可知brand取值有三个，分别是1，2，3。由于所要探究的问题female和age对brand的影响，且假定了选择各个品牌之间是相互独立的，那么建立多值选择模型来分析问题是合理的。

- **2.模型回归**
- 多值选择模型有logit和probit多值选择模型，Stata中使用多值logit和probit模型的命令语句是：
- mlogit y x1 x2 … [if] [in] [weight] [,options]（multinomial logit 模型)
- mprobit y x1 x2 …[if] [in] [weight] [,options]（multinomial probit 模型)
- 此命令中if和in表示对检测拟合优度时的条件和范围的设定，weight表示对观测值的权重设定，options的内容如下表所示：

options	描述
Main	
noconstant	无常数项
baseoutcome(#)	设定基础类别
constraints（clist）	应用特定的线性约束
collinear	保留多重共线性预测变量
SE/Robust	
vce(vcetype)	vcetype 可能包括 oim, robust, cluster *clustvar*, opg bootstrap, 或者 jackknife
Reporting	
level(#)	设置置信度，默认值是 95
rrr	输出相对风险比率
max options	
maximize_options	控制最优化过程；很少使用

- 经常使用的命令语句是“mlogit y x1 x2 …，base(#)”或者“mprobit y x1 x2 …，base(#)”，其中#是指被解释变量的某个取值，其可以根据需要变动此参照组。本实验中，由于logit模型与probit模型操作相似，以多值logit为例进行操作。
- 在Stata命令窗口中输入如下命：
- mlogit brand age female, base(1)
- 此命令表示以age和female为解释变量，brand为被解释变量，以brand=1为参照组的多值logit模型回归。
- 根据前面原理部分的介绍，该题的多值logit模型是由三个方程组成的。Stata回归结果图显示出了j=2和j=3时对应的模型估计结果，自然由三种选择概率之和为1可得到j=1时模型结果。

前面介绍了$\hat{\beta}_{MLE}$代表了解释变量单位的增加引起的是相对风险比的边际变化，就可以对此结果进行解读了。例如 brand=2 时，female 的系数是 0.52，说明若 female 由 0 增加到 1，样本中个体平均选择 2 的概率相对选择 1 的概率的对数（即相对风险）增加 0.52；此时 age 的样本估计系数是 0.37，说明 age 增加 1，样本中个体平均选 2 的概率相对选择 1 的概率的对数增加 0.37。

- Stata中得出多值选择模型个体选择被解释变量每个取值的概率的命令语句格式（1）：
- predict [type] {stub*|newvars} [if] [in] [,statistic outcome(#,#,…) nooffset]
- 该预测命令语句中，type表示预测设定新变量的类型，{stub*|newvars}表示预测的新变量名称，if和in表示对检测拟合优度时的条件和范围的设定，outcome表示需要对其指定的类别进行概率预测。如果不设定outcome选项，则需设定k个新变量。如果是预测指数或者指数的标准差，则需设定1个新变量。outcome()中，outcome可以直接用类别的取值，也可以用#1 #2等表示类别的序号，当然也可用数值标签来表示。nooffset表示预测时的约束，statistic的内容主要包括:

pr	概率预测，此为默认值
xb	线性预测
stdq	计算预测的标准差

- 预测命令格式（2）：
- predict [type] {stub*|newvarlist} [if] [in], scores
- 此命令中type表示预测设定新变量的类型，{stub*|newvarlist}表示预测的新变量名称，if和in表示对检测拟合优度时的条件和范围的设定，score表示对数似然函数对每个方程的一阶导数，第1、2、…、k个变量为对数似然函数对地1、2、3、…、k个方程的一阶导数。

- 在本实验中，在Stata命令窗口中输入如下命令语句预测brand三个取值的概率然后列出如图9.15的预测结果：
- predict p1 p2 p3
- List
- 此图可以看出很多时候根据模型预测选择某个品牌的概率最大，但是实际上此个体未选择此品牌，就是预测失败了。若读整个个体选择的概率图，会有一个很明显的结论，年轻的人倾向于选择brand1（选择brand的概率较大），随着年龄增加选择brand2和brand3的概率增加，年龄越大的人倾向选择brand3。

二、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 * | $\text{newvar}_{\text{reg}}$ $\text{newvar}_{\text{lnsigma}}$} [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)))