

## 第 14 章 受限被解释变量

被解释变量的取值范围有时受限制, 称为“受限被解释变量”(Limited Dependent Variable)。

### 14.1 断尾回归

对线性模型  $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ , 假设只有满足  $y_i \geq c$  的数据才能观测到。

例:  $y_i$  为所有企业的销售收入, 而统计局只收集规模以上企业数据, 比如  $y_i \geq 100,000$ 。被解释变量在 100,000 处存在“左边断尾”。

## 断尾随机变量的概率分布

随机变量  $y$  断尾后，其概率密度随之变化。

记  $y$  的概率密度为  $f(y)$ ，在  $c$  处左边断尾后的条件密度函数为

$$f(y | y > c) = \begin{cases} \frac{f(y)}{P(y > c)}, & \text{若 } y > c \\ 0, & \text{若 } y \leq c \end{cases}$$

由于概率密度曲线下面积为 1，故断尾变量的密度函数乘以因子  $\frac{1}{P(y > c)}$ 。

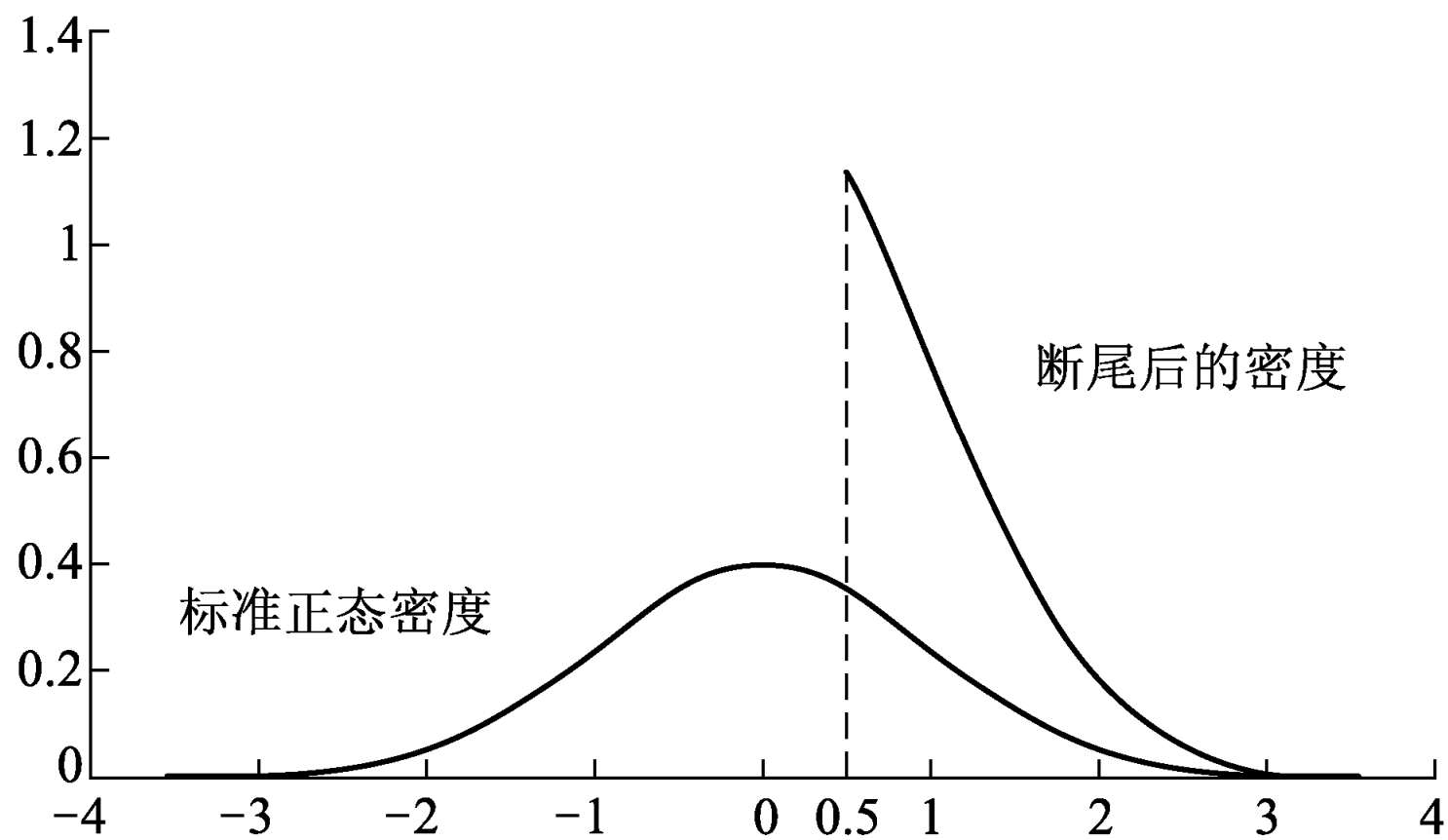


图 14.1 断尾的效果

断尾分布的期望也发生变化。以左边断尾为例。

对于最简单情形， $y \sim N(0, 1)$ ，可证明(参见附录)

$$E(y | y > c) = \frac{\phi(c)}{1 - \Phi(c)}$$

对于任意实数  $c$ ，定义“反米尔斯比率” (Inverse Mill's Ratio, 简记 **IMR**) 为

$$\lambda(c) \equiv \frac{\phi(c)}{1 - \Phi(c)}$$

则  $E(y | y > c) = \lambda(c)$ 。

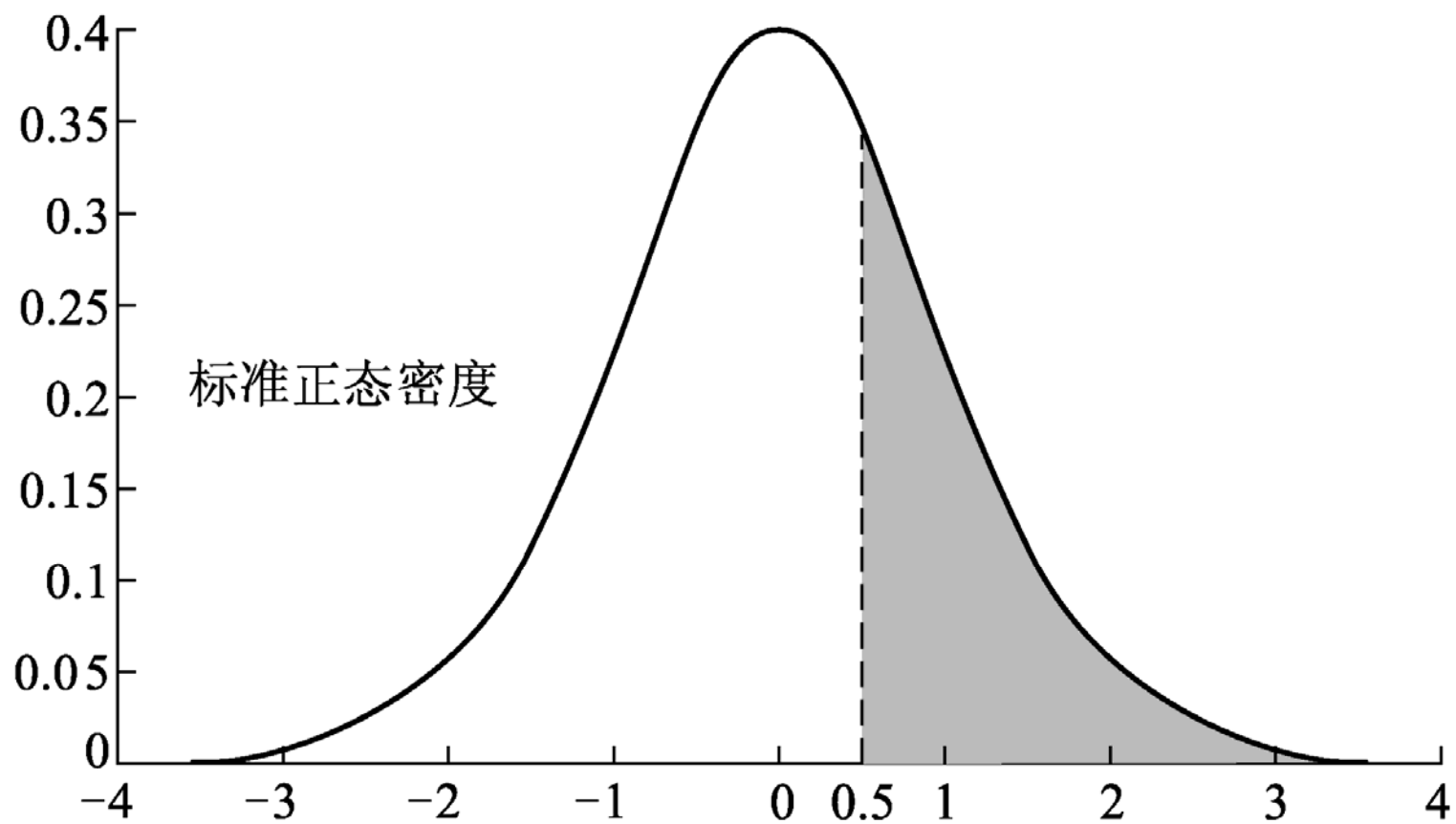


图 14.2 反米尔斯比率

对于正态分布  $y \sim N(\mu, \sigma^2)$ ，定义  $z \equiv \frac{y - \mu}{\sigma} \sim N(0, 1)$ ，则  $y = \mu + \sigma z$ 。故

$$\begin{aligned} E(y \mid y > c) &= E(\mu + \sigma z \mid \mu + \sigma z > c) = E\left[\mu + \sigma z \mid z > (c - \mu)/\sigma\right] \\ &= \mu + \sigma E\left[z \mid z > (c - \mu)/\sigma\right] = \mu + \sigma \cdot \lambda[(c - \mu)/\sigma] \end{aligned}$$

对于模型  $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ ， $\varepsilon_i \mid \mathbf{x}_i \sim N(0, \sigma^2)$ ，则  $y_i \mid \mathbf{x}_i \sim N(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2)$ ，故

$$E(y_i \mid y_i > c) = \mathbf{x}_i' \boldsymbol{\beta} + \sigma \cdot \lambda[(c - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]$$

如果用 OLS 估计  $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ ，则遗漏了非线性项  $\sigma \cdot \lambda[(c - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]$ ，与  $\mathbf{x}_i$  相关，导致 OLS 不一致。

参见图 14.3。总体回归线为 $\alpha + \beta x_i$ ，而样本回归线为 $\hat{\alpha} + \hat{\beta} x_i$ 。

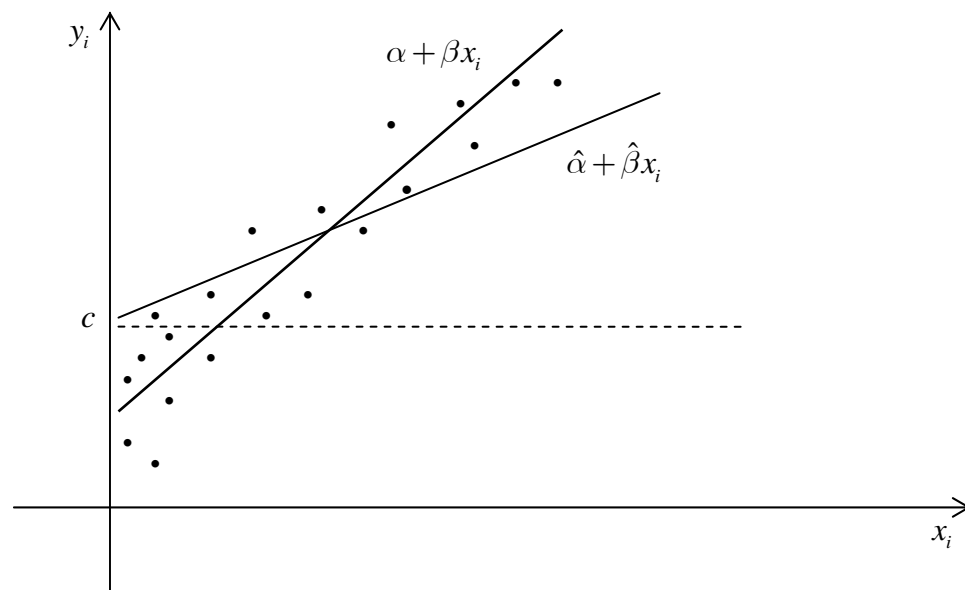


图 14.3 断尾回归示意图

使用 MLE 可得到一致估计。断尾前的概率密度：

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)^2\right\} = \frac{1}{\sigma} \phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

样本被观测到的概率：



$$\begin{aligned}
P(y_i > c \mid \mathbf{x}_i) &= 1 - P(y_i \leq c \mid \mathbf{x}_i) \\
&= 1 - P\left(\frac{y_i - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma} \leq \frac{c - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}_i\right) \\
&= 1 - P\left(\frac{\varepsilon_i}{\sigma} \leq \frac{c - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}_i\right) \\
&= 1 - \Phi\left(\frac{c - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma}\right)
\end{aligned}$$

断尾后的条件密度:

$$f(y_i \mid y_i > c, \mathbf{x}_i) = \frac{\frac{1}{\sigma} \phi[(y_i - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]}{1 - \Phi[(c - \mathbf{x}_i' \boldsymbol{\beta})/\sigma]}$$

## 14.2 零断尾泊松回归与负二项回归

计数数据有时仅包括正整数，不包括取值为 0 的观测值，称为“零断尾” (zero-truncated)。

例：在商场发放问卷调查，研究消费者每周去商场的次数。

例：在公交车上发放问卷调查，研究乘车者每周坐公交的次数。

如果不对似然函数进行调整，将得不到一致估计。

记  $f(y)$  为  $y$  的概率函数，而  $F(y) \equiv P(Y \leq y)$  为 cdf。如果存在零断尾，则断尾后的概率函数为

$$f(y | y \geq 1) = \frac{f(y)}{1 - F(0)}, \quad y = 1, 2, \dots$$

如果  $y$  服从泊松分布, 则

$$f(y | y \geq 1) = \frac{e^{-\lambda} \lambda^y}{y!(1 - e^{-\lambda})}, \quad y = 1, 2, \dots$$

进行 MLE 估计, 得到“零断尾泊松回归”(zero-truncated Poisson regression)。如果  $y$  服从负二项分布(NB1 或 NB2), 可进行“零断尾负二项回归”(zero-truncated negative binomial regression)。

### 14.3 随机前沿模型(选读)

## 14.4 偶然断尾与样本选择

被解释变量  $y_i$  的断尾有时与另一变量  $z_i$  有关，称为“偶然断尾”(incidental truncation)或“样本选择”(sample selection)。

称  $z_i$  为选择变量。

**例** 在美国的亚裔移民给人的整体印象是聪明能干。但在美国的亚裔并非亚洲人口的代表性样本。通常只有受过高等教育或具有吃苦冒险精神的亚裔才会“自我选择”(self selection)移民。

决定移民与否的变量便对被解释变量产生了断尾作用，故“样本选择”将导致“选择性偏差”(selection bias)。

例 妇女劳动力供给模型：

劳动时间方程  $\text{hours} = \alpha_0 + \alpha_1 \text{wage} + \alpha_2 \text{children} + \alpha_3 \text{marriage} + u$

工资方程  $w^o - w^r = \beta_0 + \beta_1 \text{age} + \beta_2 \text{education} + \beta_3 \text{children} + \beta_0 \text{location} + v$

$w^o$  表示 offered wage,  $w^r$  表示 reservation wage。

如果  $w^o - w^r < 0$ , 则选择不工作, 无法观测到劳动时间(hours), 造成劳动时间方程的偶然断尾与样本选择问题。

考虑二维正态随机向量 $(y, z)$ ，记期望为 $(\mu_y, \mu_z)$ ，标准差为 $(\sigma_y, \sigma_z)$ ，相关系数为 $\rho$ ，联合密度函数为 $f(y, z)$ 。

假设个体进入样本的“选择机制”(selection mechanism)为“选择变量 $z$ 大于某常数 $c$ ”。

比如，在妇女劳动力供给例子中， $z = w^o - w^r$ ，而 $c = 0$ 。

断尾后的联合分布：

$$f(y, z | z > c) = \frac{f(y, z)}{P(z > c)}$$

偶然断尾 $y$ 的条件期望：

$$E(y | z > c) = \mu_y + \rho\sigma_y\lambda[(c - \mu_z)/\sigma_z]$$

$\lambda(\cdot)$ 为反米尔斯比率(IMR)函数。

如果 $\rho = 0$ ( $y$ 与 $z$ 相互独立),则 $z$ 的选择过程并不对 $y$ 产生影响。

如果 $\rho > 0$ (即 $y$ 与 $z$ 正相关),则“ $z > c$ ”偶然断尾的结果是把 $y$ 的整个分布推向右边(因为 $\lambda(\cdot) > 0$ ),从而使得条件期望 $E(y|z > c)$ 大于无条件期望 $E(y)$ 。

在“ $z < c$ ”条件下,偶然断尾 $y$ 的条件期望为

$$E(y|z < c) = \mu_y - \rho\sigma_y\lambda[(\mu_z - c)/\sigma_z]$$

假设回归模型为 $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$ 。

$y_i$  是否可观测取决于选择变量  $z_i$  (取值为 0 或 1)

$$y_i = \begin{cases} \text{可观测} & z_i=1 \\ \text{不可观测} & z_i=0 \end{cases}$$

决定二值变量  $z_i$  的方程为

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

$$z_i^* = \mathbf{w}_i' \boldsymbol{\gamma} + u_i$$

$z_i^*$  为不可观测的潜变量。



假设 $u_i$ 服从正态分布，则 $z_i$ 为 Probit 模型，故 $P(z_i = 1 | \mathbf{w}_i) = \Phi(\mathbf{w}_i' \boldsymbol{\gamma})$ 。

可观测样本的条件期望：

$$\begin{aligned} E(y_i | y_i \text{ 可观测}) &= E(y_i | z_i^* > 0) = E(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i | \mathbf{w}_i' \boldsymbol{\gamma} + u_i > 0) \\ &= E(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i | u_i > -\mathbf{w}_i' \boldsymbol{\gamma}) = \mathbf{x}_i' \boldsymbol{\beta} + E(\varepsilon_i | u_i > -\mathbf{w}_i' \boldsymbol{\gamma}) \\ &= \mathbf{x}_i' \boldsymbol{\beta} + \rho \sigma_\varepsilon \lambda(-\mathbf{w}_i' \boldsymbol{\gamma}) \end{aligned}$$

其中， $E(\varepsilon_i) = E(u_i) = 0$ ，并将 Probit 扰动项的标准差 $\sigma_u$ 标准化为 1。

OLS 估计，将遗漏非线性项 $\rho \sigma_\varepsilon \lambda(-\mathbf{w}_i' \boldsymbol{\gamma})$ 。

如 $\mathbf{w}_i$ 与 $\mathbf{x}_i$ 相关，则 OLS 不一致，除非“ $\rho = 0$ ”（即  $y$  与  $z$  不相关）。

解释变量  $x_{ik}$  的边际效应:

$$\frac{\partial E(y_i | z_i^* > 0)}{\partial x_{ik}} = \beta_k + \rho\sigma_\varepsilon \frac{\partial \lambda(-\mathbf{w}_i'\boldsymbol{\gamma})}{\partial x_{ik}}$$

右边第一项为直接影响，第二项是通过改变个体进入样本可能性而产生的间接影响(即选择性偏差)。

如知道  $\boldsymbol{\gamma}$ ，就知道  $\lambda(-\mathbf{w}_i'\boldsymbol{\gamma})$ ，可把它作为解释变量引入回归方程。

Heckman (1979)提出“两步估计法”，也称“Heckit”。

第一步：用 Probit 估计方程  $P(z_i = 1 | \mathbf{w}) = \Phi(\mathbf{w}_i'\boldsymbol{\gamma})$ ，得到估计值  $\hat{\boldsymbol{\gamma}}$ ，计算  $\hat{\lambda}(-\mathbf{w}_i'\hat{\boldsymbol{\gamma}})$ 。

第二步：用 OLS 回归  $y_i \xrightarrow{\text{OLS}} \mathbf{x}_i, \hat{\lambda}_i$ ，得到估计值  $\hat{\boldsymbol{\beta}}, \hat{\rho}, \hat{\sigma}_\varepsilon$ 。

更有效率的方法是 MLE。

在两步法中，第一步误差被带入第二步，效率不如 MLE 的整体估计。

两步法的优点在于，操作简便；对于分布的假设也更弱(即使不假设二维正态分布，也可能成立)。

为检验是否存在样本选择偏差(sample selection bias)，可使用似然比检验来检验原假设 “ $H_0 : \rho = 0$ ”。

如使用 Heckit，无法进行此 LR 检验。

## 14.5 归并回归

对于线性模型  $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ ，当  $y_i \geq c$  (或  $y_i \leq c$ ) 时，所有  $y_i$  都被归并为  $c$ ，称为“归并数据” (censored data)。

例 (上不封顶的数据, top coding) 在问卷调查中，常有诸如“收入在¥50,000 及以上”这样的选项。

例 (边角解) 考虑买车的决定，并考察“买车开支”这个变量。如果不买车，则“买车开支”的最优解为边角解，即买车开支为 0；反之，如果买车，则买车开支一定为正数。

例 (边角解) 考察“劳动时间”这个变量。对于失业或待业者而言，“劳动时间”的最优解为边角解，即劳动时间为 0；而就业者

的劳动时间一定为正数。

归并回归(censored regression)与断尾回归不同的是, 虽有全部观测数据, 但某些数据的  $y_i$  被压缩在一个点上。

$y_i$  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  $y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$  ( $y_i^*$  不可观测),  $\varepsilon_i | \mathbf{x}_i \sim N(0, \sigma^2)$ , 归并点为  $c = 0$ 。

假设可观测到:

$$y_i = \begin{cases} y_i^*, & \text{若 } y_i^* > 0 \\ 0, & \text{若 } y_i^* \leq 0 \end{cases}$$

如使用满足条件 “ $y_i > 0$ ” 的子样本，将导致断尾，出现偏差，  
因为

$$\begin{aligned} E(y_i \mid \mathbf{x}_i; y_i > 0) &= E(y_i^* \mid \mathbf{x}_i; y_i > 0) \quad (\text{给定 } y_i > 0, \text{ 必然 } y_i = y_i^*) \\ &= E(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \mid \mathbf{x}_i; y_i^* > 0) \\ &= \mathbf{x}_i' \boldsymbol{\beta} + E(\varepsilon_i \mid \mathbf{x}_i; \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i > 0) \\ &= \mathbf{x}_i' \boldsymbol{\beta} + E(\varepsilon_i \mid \mathbf{x}_i; \varepsilon_i > -\mathbf{x}_i' \boldsymbol{\beta}) \\ &= \mathbf{x}_i' \boldsymbol{\beta} + \sigma \cdot \lambda(-\mathbf{x}_i' \boldsymbol{\beta} / \sigma) \end{aligned}$$

由于忽略非线性项  $\sigma \cdot \lambda(-\mathbf{x}_i' \boldsymbol{\beta} / \sigma)$ ，导致扰动项与  $\mathbf{x}_i$  相关，故 OLS 不一致。

对于整个样本,

$$\begin{aligned} E(y_i | \mathbf{x}_i) &= 0 \cdot P(y_i = 0 | \mathbf{x}_i) + E(y_i | \mathbf{x}_i; y_i > 0) \cdot P(y_i > 0 | \mathbf{x}_i) \\ &= E(y_i | \mathbf{x}_i; y_i > 0) \cdot P(y_i > 0 | \mathbf{x}_i) \end{aligned}$$

其中,  $P(y_i > 0 | \mathbf{x}_i) = P(y_i^* > 0 | \mathbf{x}_i) = P(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i > 0 | \mathbf{x}_i)$

$$\begin{aligned} &= P(\varepsilon_i > -\mathbf{x}_i' \boldsymbol{\beta} | \mathbf{x}_i) = P\left(\frac{\varepsilon_i}{\sigma} > \frac{-\mathbf{x}_i' \boldsymbol{\beta}}{\sigma} \middle| \mathbf{x}_i\right) \\ &= 1 - \Phi(-\mathbf{x}_i' \boldsymbol{\beta} / \sigma) = \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma) \end{aligned}$$

$$E(y_i | \mathbf{x}_i) = E(y_i | \mathbf{x}_i, y_i > 0) \cdot P(y_i > 0 | \mathbf{x}_i) = \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma) [\mathbf{x}_i' \boldsymbol{\beta} + \sigma \cdot \lambda(-\mathbf{x}_i' \boldsymbol{\beta} / \sigma)]$$

是解释变量  $\mathbf{x}_i$  的非线性函数。如果使用 OLS 对整个样本进行回归, 非线性项将被纳入扰动项中, 导致不一致估计。

Tobin (1958)提出用 MLE 估计这个模型，称为“Tobit”。

在归并数据情况下， $y_i > 0$ 时的概率密度依然不变，仍为

$$\frac{1}{\sigma} \phi[(y_i - \mathbf{x}_i' \boldsymbol{\beta}) / \sigma]$$

$y_i \leq 0$ 时的分布被挤到“ $y_i = 0$ ”上，即

$$P(y_i = 0 | \mathbf{x}) = 1 - P(y_i > 0 | \mathbf{x}) = 1 - \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma)$$

该混合分布的概率密度为

$$f(y_i | \mathbf{x}) = [1 - \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma)]^{I(y_i=0)} \left[ \frac{1}{\sigma} \phi((y_i - \mathbf{x}_i' \boldsymbol{\beta}) / \sigma) \right]^{I(y_i>0)}$$



Tobit 模型的缺陷是对分布的依赖性强，不够稳健。

如果似然函数不正确（扰动项不服从正态分布或存在异方差），则 QMLE 估计不一致。

使用 Tobit 模型时，需要检验正态性与同方差性。

为了检验正态性，可将 Tobit 模型的 MLE 一阶条件视为某种矩条件，并基于此进行“条件矩检验”(conditional moment test)。

但条件矩统计量的真实分布与渐近分布有相当差距，存在较严重的显著性水平扭曲，故使用“参数自助法”来获得校正的临界值。

为了检验同方差的原假设“ $H_0: \sigma_i^2 = \sigma^2$ ”，考虑替代假设

“ $H_1 : \sigma_i^2 = \exp(\mathbf{z}_i' \boldsymbol{\alpha})$ ”，其中  $\mathbf{z}_i$  通常等于解释变量  $\mathbf{x}_i$  (也可不同)。

然后通过辅助回归，构建 LM 统计量来检验  $\boldsymbol{\alpha} = \mathbf{0}$ ，参见 Cameron and Trivedi (2010, p.550)。

如果发现扰动项不服从正态分布或存在异方差，解决方法之一为使用更稳健的“归并最小绝对离差法” (Censored Least Absolute Deviations，简记 CLAD)。

CLAD 法仅要求扰动项为 iid，即使在非正态与异方差情况下也一致，且在一定正则条件下，服从渐近正态分布。

将归并数据模型简洁地写为

$$y_i = \max(0, \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i)$$

如果  $\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i \geq 0$ ，则  $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ ；反之， $y_i = 0$ 。

CLAD 法的目标函数为离差绝对值之和：

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n |y_i - \max(0, \mathbf{x}_i' \boldsymbol{\beta})|$$

选择  $\boldsymbol{\beta}$  使得离差绝对值之和最小化，即可得到 CLAD 估计量。

## 14.6 归并数据的两部分模型（选读）

## 14.7 含内生解释变量的 Tobit 模型（选读）