第 11 章　二值选择模型

11.1　离散被解释变量的例子

二值选择(binary choices)：考研或不考研；就业或待业；买房或不买房；买保险或不买保险；贷款申请被批准或拒绝；出国或不出国；回国或不回国；战争或和平；生或死。

多值选择(multiple choices)：对不同交通方式的选择(走路、骑车、坐车上班)；对不同职业的选择。

这类模型被称为“离散选择模型”(discrete choice model)或“定

性反应模型”(qualitative response model)。

有时被解释变量只能取非负整数：

企业在某段时间内获得的专利数；某人在一定时间内去医院看病的次数；某省在一年内发生煤矿事故的次数。

这类数据称为“计数数据”(count data)，被解释变量也是离散的。

考虑到离散被解释变量的特点，通常不宜用 OLS 进行回归。

11.2　二值选择模型

假设个体只有两种选择，比如$y=1$(考研)或$y=0$(不考研)。

所有解释变量都包括在向量$\boldsymbol{x}$中。

“线性概率模型”(Linear Probability Model，简记 LPM)：

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i \quad (i=1,\cdots,n)$$

优点：计算方便，容易得到边际效应。

缺点：（1）由于$\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$，故$\varepsilon_i = 1-\boldsymbol{x}_i'\boldsymbol{\beta}$或$\varepsilon_i = -\boldsymbol{x}_i'\boldsymbol{\beta}$，因此$\varepsilon_i$必然与$\boldsymbol{x}_i$相关，导致估计不一致。

（2）ε_i服从两点分布，而非正态分布。

（3）由于$\mathrm{Var}(\varepsilon_i)=\mathrm{Var}(\boldsymbol{x}_i'\boldsymbol{\beta})$，故扰动项$\varepsilon_i$的方差依赖于$\boldsymbol{x}_i$，存在异方差(故应使用稳健标准误)。

（4）可能出现$\hat{y}>1$或$\hat{y}<0$的不现实情形，参见图 11.1。

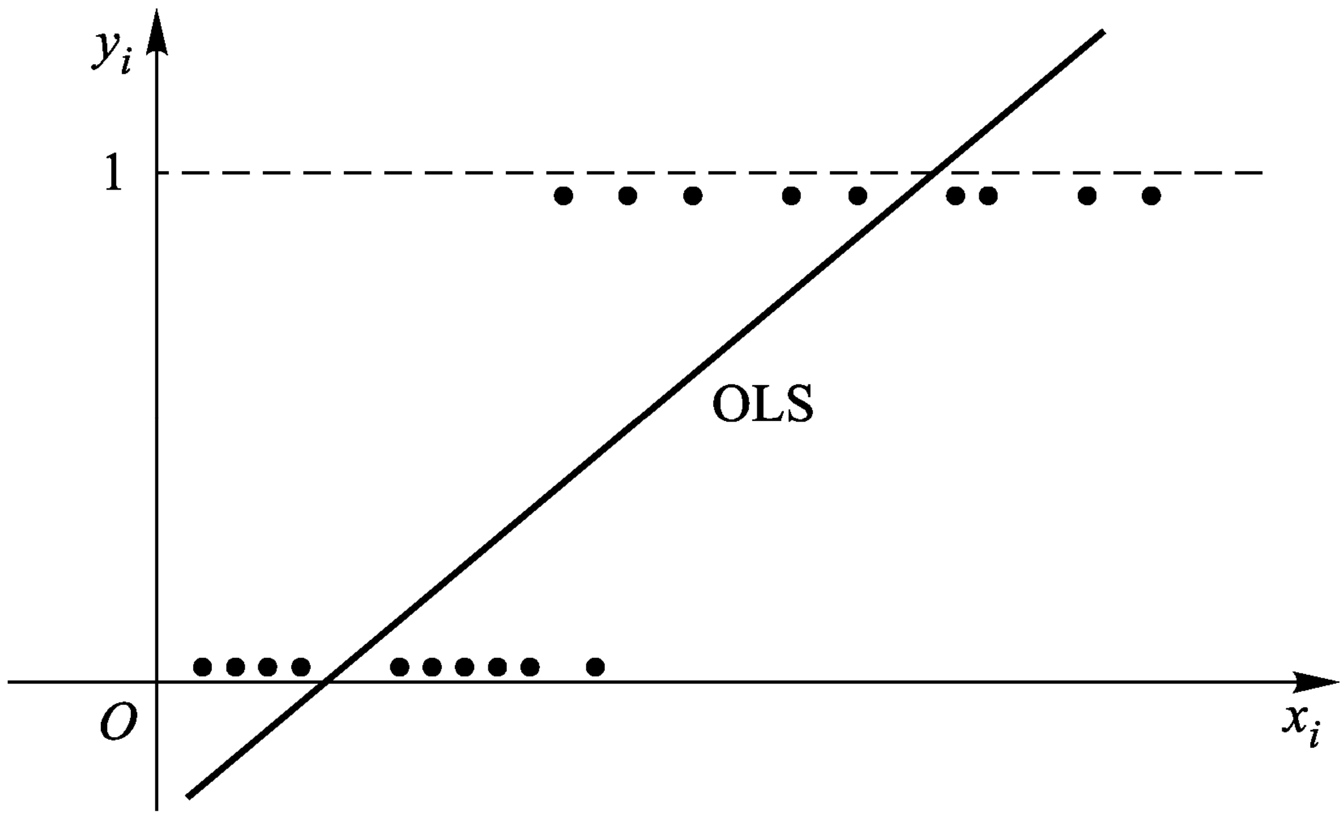

图 11.1　OLS 与二值选择模型

为使 y 的预测值总是介于$[0,1]$之间，给定$\boldsymbol{x}$，考虑 y 的两点分布概率：

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

函数$F(\boldsymbol{x},\boldsymbol{\beta})$也称“连接函数”(link function)。

通过选择合适的$F(\boldsymbol{x},\boldsymbol{\beta})$(比如，某随机变量的 cdf)，可保证$0\le\hat{y}\le 1$，并将$\hat{y}$理解为“$y=1$”发生的概率，因为：

$$\mathrm{E}(y\,|\,\boldsymbol{x}) = 1\cdot\mathrm{P}(y=1\,|\,\boldsymbol{x}) + 0\cdot\mathrm{P}(y=0\,|\,\boldsymbol{x}) = \mathrm{P}(y=1\,|\,\boldsymbol{x})$$

如果$F(\boldsymbol{x}, \boldsymbol{\beta})$为标准正态的 cdf：

$$\mathrm{P}(y=1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = \Phi(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \int_{-\infty}^{\boldsymbol{x}'\boldsymbol{\beta}} \phi(t)\, dt$$

该模型称为“Probit”。

如果$F(\boldsymbol{x}, \boldsymbol{\beta})$为“逻辑分布”(logistic distribution)的 cdf：

$$\mathrm{P}(y=1 \mid \boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta}) = \Lambda(\boldsymbol{x}'\boldsymbol{\beta}) \equiv \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} \quad (11.1)$$

该模型称为“Logit”。逻辑分布的密度函数关于原点对称，期望为 0，方差为$\pi^2/3$(大于标准正态的方差)，具有厚尾(fat tails)。

由于逻辑分布的 cdf 有解析表达式(而标准正态没有)，故计算 Logit 比 Probit 更为方便。

对于此非线性模型，进行 MLE 估计。

以 Logit 模型为例。第 i 个观测数据的概率密度为

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \begin{cases} \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), \ \text{若} y_i = 1 \\ 1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}), \ \text{若} y_i = 0 \end{cases}$$

将其更紧凑地写为

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = \left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{y_i} \left[1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{1-y_i}$$

取对数可得

$$\ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}) = y_i \ln\left[\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right] + (1 - y_i)\ln\left[1 - \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})\right]$$

假设样本中的个体相互独立，则整个样本的对数似然函数为

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

在此非线性模型中，估计量 $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ 并非边际效应(marginal effects)。以 Probit 为例，

$$\frac{\partial \mathrm{P}(y = 1 \mid \boldsymbol{x})}{\partial x_k} = \frac{\partial \mathrm{P}(y = 1 \mid \boldsymbol{x})}{\partial (\boldsymbol{x}'\boldsymbol{\beta})} \cdot \frac{\partial (\boldsymbol{x}'\boldsymbol{\beta})}{\partial x_k} = \phi(\boldsymbol{x}'\boldsymbol{\beta}) \cdot \beta_k$$

由于 Probit 与 Logit 使用的分布函数不同，其参数估计值并不直接可比。须计算边际效应，然后进行比较。

但对于非线性模型，边际效应不是常数，随着解释变量而变。常用的边际效应概念：

(1) **平均边际效应** (average marginal effect)，即分别计算在每个样本观测值上的边际效应，然后进行简单算术平均。

(2) **样本均值处的边际效应** (marginal effect at mean)，即在 $\boldsymbol{x} = \bar{\boldsymbol{x}}$ 处的边际效应。

(3) **在某代表值处的边际效应** (marginal effect at a

representative value)，即给定$\boldsymbol{x}^*$，在$\boldsymbol{x} = \boldsymbol{x}^*$处的边际效应。

在非线性模型中，样本均值处的个体行为并不等于样本中个体的平均行为(average behavior of individuals differs from behavior of the average individual)。

对于政策分析而言，平均边际效应(Stata 的默认方法)，或在某代表值处的边际效应通常更有意义。

$\hat{\boldsymbol{\beta}}_{\text{MLE}}$并非边际效应，它究竟有什么含义？

对于 Logit 模型，记 $p \equiv \mathrm{P}(y=1|\boldsymbol{x})$，则$1-p=\mathrm{P}(y=0|\boldsymbol{x})$。

由于 $p=\dfrac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$， $1-p=\dfrac{1}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})}$， 故

$$\frac{p}{1-p}=\exp(\boldsymbol{x}'\boldsymbol{\beta})$$

$$\ln\left(\frac{p}{1-p}\right)=\boldsymbol{x}'\boldsymbol{\beta}$$

$p/(1-p)$称为“几率比”(odds ratio)或“相对风险”(relative risk)。

【例】假设在检验药物疗效的随机实验中，“$y=1$”表示“生”，“$y=0$”表示“死”；则几率比为 2 意味着存活的概率是死亡概率的两倍。

$\hat{\beta}_j$表示解释变量x_j增加一个微小量引起“对数几率比”(log-odds ratio)的边际变化。

也可视$\hat{\beta}_j$为半弹性，即x_j增加一单位引起几率比的变化百分比。比如，$\hat{\beta}_j = 0.12$，意味着x_j增加一单位引起几率比增加 12%。

另一解释：假设x_j增加一单位，从x_j变为x_j+1，记p的新值为p^*，则新几率比与原先几率比的比率可写为：

$$\frac{\dfrac{p^*}{1-p^*}}{\dfrac{p}{1-p}}=\frac{\exp[\beta_1+\beta_2x_2+\cdots+\beta_j(x_j+1)+\cdots+\beta_Kx_K]}{\exp(\beta_1+\beta_2x_2+\cdots+\beta_jx_j+\cdots+\beta_Kx_K)}=\exp(\beta_j)$$

有些研究者偏好计算$\exp(\hat{\beta}_j)$，它表示解释变量x_j增加一单位引起几率比的变化倍数。

比如，$\hat{\beta}_j=0.12$，则$\exp(\hat{\beta}_j)=e^{0.12}=1.13$，故当$x_j$增加一单位时，新几率比是原先几率比的 1.13 倍，或增加 13%，因为$\exp(\hat{\beta}_j)-1=1.13-1=0.13$。

基于此，Stata 称$\exp(\hat{\beta}_j)$为几率比(odds ratio)。

如果$\hat{\beta}_j$较小，则$\exp(\hat{\beta}_j)-1\approx\hat{\beta}_j$(将$\exp(\hat{\beta}_j)$泰勒展开)，此时以上两种方法是等价的。

如果x_j至少必须变化一个单位(比如性别、婚否等虚拟变量，年龄，子女个数)，则应使用$\exp(\hat{\beta}_j)$。

对于 Probit 模型，无法对其系数$\hat{\boldsymbol{\beta}}_{\text{MLE}}$进行类似的解释。

如何衡量二值模型的拟合优度呢？

由于不存在平方和分解公式，无法计算 R^2 。

Stata 仍然汇报一个"准 R^2" (Pseudo R^2)，由 McFadden (1974) 所提出：

$$准R^2 \equiv \frac{\ln L_0 - \ln L_1}{\ln L_0}$$

$\ln L_1$ 为原模型的对数似然函数之最大值，而 $\ln L_0$ 为以常数项为唯一解释变量的对数似然函数之最大值。

由于 y 为离散的两点分布，似然函数的最大可能值为 1，故对数似然函数的最大可能值为 0，记为 $\ln L_{\max}$ 。由于 $0 \ge \ln L_1 \ge \ln L_0$，故

准R^2可写为$\dfrac{\ln L_1-\ln L_0}{\ln L_{\max}-\ln L_0}$。

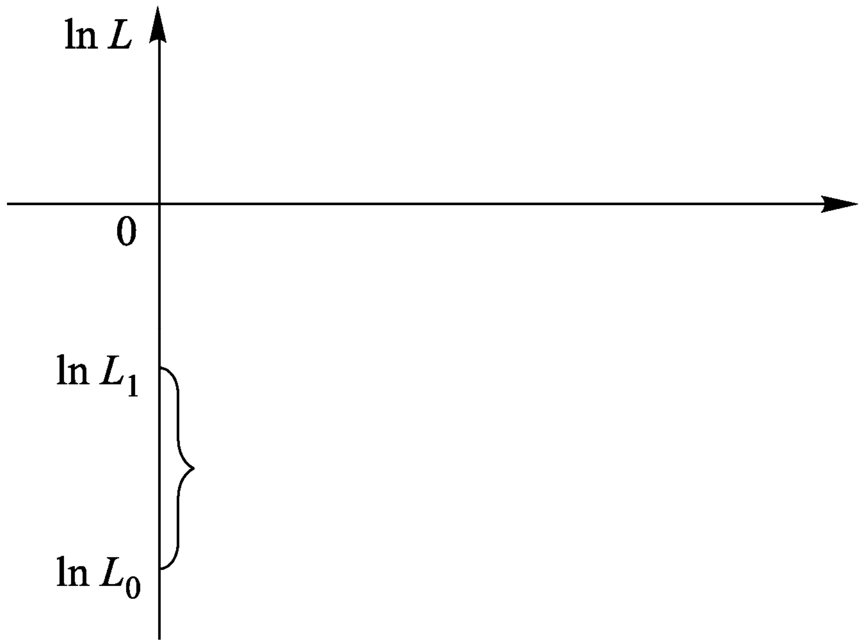

图 11.2　准R^2的计算

判断拟合优度的另一方法是计算“正确预测的百分比”(percent correctly predicted)。

如果发生概率的预测值$\hat{y} \geq 0.5$，则认为其预测$y = 1$；反之，则认为其预测$y = 0$。

将预测值与实际值(样本数据)进行比较，就能计算正确预测的百分比。

对于 Probit 与 Logit 模型，如果分布函数设定不正确，则为准最大似然估计(QMLE)。

由于二值选择模型的分布必然为两点分布(属于线性指数分布族)，故只要条件期望函数$\mathrm{E}(y\,|\,\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$正确，则 MLE 一致。

由于两点分布的特殊性，在 iid 的情况下，只要$\mathrm{E}(y\,|\,\boldsymbol{x}) = F(\boldsymbol{x}, \boldsymbol{\beta})$，稳健标准误就等于 MLE 的普通标准误。如果认为模型设定正确，就没有必要使用稳健标准误(但使用稳健标准误也没有错)。

如果模型设定不正确($\mathrm{E}(y\,|\,\boldsymbol{x}) \neq F(\boldsymbol{x}, \boldsymbol{\beta})$)，则 Probit 与 Logit 模型不能得到一致估计，使用稳健标准误也没有太大意义(只是更精确地估计错误参数的标准误)；首先应解决参数估计的一致性问题。

如果数据并非 iid，比如可将样本分为若干组(聚类)，而每组内的个体存在组内自相关，则应使用聚类稳健的标准误。

11.3 二值选择模型的微观基础

对于二值选择行为，可通过“潜变量”(latent variable)概括该行为的净收益(收益减去成本)。

如果净收益大于 0，则选择做；否则，选择不做。

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$$

其中，净收益 y^* 为潜变量，不可观测。选择规则为

$$y=\begin{cases}1, & 若\ y^*>0\\ 0, & 若\ y^*\le 0\end{cases}$$

因此，

$$\mathrm{P}(y=1\,|\,\boldsymbol{x})=\mathrm{P}(y^*>0\,|\,\boldsymbol{x})=\mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta}+\varepsilon>0\,|\,\boldsymbol{x})=\mathrm{P}(\varepsilon>-\boldsymbol{x}'\boldsymbol{\beta}\,|\,\boldsymbol{x})$$

假设$\varepsilon \sim N(0,\sigma^2)$或服从逻辑分布，则

$$\mathrm{P}(y=1\,|\,\boldsymbol{x})=\mathrm{P}(\varepsilon>-\boldsymbol{x}'\boldsymbol{\beta}\,|\,\boldsymbol{x})=\mathrm{P}(\varepsilon<\boldsymbol{x}'\boldsymbol{\beta})=F_{\varepsilon}(\boldsymbol{x}'\boldsymbol{\beta})$$

$F_{\varepsilon}(\cdot)$为ε的 cdf，此处用到密度函数关于原点对称的性质。

如果ε为正态分布，则为 Probit；如果ε为逻辑分布，则为 Logit。

对于任意常数$k > 0$，$\mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0) = \mathrm{P}(k\boldsymbol{x}'\boldsymbol{\beta} + k\varepsilon > 0)$。

记扰动项方差$\sigma^2 \equiv \mathrm{Var}(\varepsilon)$，则$\mathrm{Var}(k\varepsilon) = k^2\sigma^2$。

故$(k\boldsymbol{\beta}, k^2\sigma^2)$对模型的拟合与$(\boldsymbol{\beta}, \sigma^2)$完全一样，无法同时“识别”(identify)$\boldsymbol{\beta}$与$\sigma^2$。

对于 Probit 模型，令扰动项之方差σ^2为 1，即$\varepsilon \sim N(0,1)$；对于 Logit 模型，令扰动项之方差为$\pi^2/3$。

另一微观基础为“随机效用最大化”模型(Random Utility Maximization，简记 RUM)。

假设选择 a，可带来效用U_a；选择 b，可带来效用U_b。

如果$U_a > U_b$，则选 a，记$y=1$；如果$U_a \le U_b$，则选 b，记$y=0$。

假定 $U_a = \boldsymbol{x}'\boldsymbol{\beta}_a + \varepsilon_a$，$U_b = \boldsymbol{x}'\boldsymbol{\beta}_b + \varepsilon_b$。

由于效用方程中包含一个扰动项，故名“随机效用”。

$$
\begin{aligned}
\mathrm{P}(y=1\,|\,\boldsymbol{x}) &= \mathrm{P}(U_a > U_b\,|\,\boldsymbol{x}) \\
&= \mathrm{P}(\boldsymbol{x}'\boldsymbol{\beta}_a + \varepsilon_a > \boldsymbol{x}'\boldsymbol{\beta}_b + \varepsilon_b\,|\,\boldsymbol{x}) \\
&= \mathrm{P}\left[\boldsymbol{x}'(\boldsymbol{\beta}_a - \boldsymbol{\beta}_b) + (\varepsilon_a - \varepsilon_b) > 0\,|\,\boldsymbol{x}\right]
\end{aligned}
$$

定义$\boldsymbol{\beta} \equiv \boldsymbol{\beta}_a - \boldsymbol{\beta}_b$，$\varepsilon \equiv \varepsilon_a - \varepsilon_b$，又回到前面潜变量法的表达式。

如果ε_a与ε_b均为正态且相互独立，则$(\varepsilon_a - \varepsilon_b)$也服从正态分布。将$\mathrm{Var}(\varepsilon_a - \varepsilon_b)$标准化为 1，即得到 Probit 模型。

如果ε_a与ε_b相互独立，且均服从从非对称的“I 型极值分布”(Type I extreme value distribution)，cdf 为$F(\varepsilon) = \exp\left\{-\mathrm{e}^{-\varepsilon}\right\}$，则$(\varepsilon_a - \varepsilon_b)$服从逻辑分布。

随机效用法的优点是，容易推广到多值选择的情形。

11.4　二值选择模型中的异方差问题

标准的 Probit 或 Logit 模型假设扰动项为同方差，据此写出似然函数。对于这个同方差假设，可进行似然比检验(LR)。

对于 Probit 模型，同方差的原假设H_0为

$$\mathrm{P}(y_i = 1 | \boldsymbol{x}_i) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma\right)$$

其中，扰动项的标准差$\sigma = 1$。而“异方差”的替代假设H_1为

$$\mathrm{P}(y_i = 1 | \boldsymbol{x}_i) = \Phi\left(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma_i\right)$$

其中，$\sigma_i^2 \equiv \mathrm{Var}(\varepsilon_i)$。

假设σ_i^2依赖于外生变量$\boldsymbol{z} \equiv (z_1, \cdots, z_m)$:

$$\sigma_i^2 = \exp(\boldsymbol{z}_i'\boldsymbol{\delta})$$

$\boldsymbol{z}$可与解释变量$\boldsymbol{x}$有重叠部分，或包括$\boldsymbol{x}$，但不包括常数项。

两边取对数可得，

$$\ln \sigma_i^2 = \boldsymbol{z}_i'\boldsymbol{\delta}$$

在异方差的替代假设下，同样可写出似然函数，同时估计原方程与条件方差方程。

11.5 稀有事件偏差(选读)

11.6 含内生变量的 Probit 模型(选读)

11.7 双变量 Probit 模型(选读)

11.8 部分可观测的双变量 Probit 模型(选读)