

§ 4.3 离散被解释变量数据计量经济学 模型（一）——二元选择模型

Models with Discrete Dependent Variables—Binary Choice Model

- 一、二元离散选择模型的经济背景
- 二、二元离散选择模型
- 三、二元Probit离散选择模型及其参数估计
- 四、二元Logit离散选择模型及其参数估计
- 五、二元离散选择模型的检验

说明

- 在经典计量经济学模型中，被解释变量通常被假定为连续变量。
- 离散被解释变量数据计量经济学模型 (Models with Discrete Dependent Variables) 和离散选择模型 (DCM, Discrete Choice Model)。
- 二元选择模型 (Binary Choice Model) 和多元选择模型 (Multiple Choice Model)。
- 本节只介绍二元选择模型。

- 离散选择模型起源于Fechner于1860年进行的动物条件二元反射研究。
- 1962年，Warner首次将它应用于经济研究领域，用以研究公共交通工具和私人交通工具的选择问题。
- 70、80年代，离散选择模型被普遍应用于经济布局、企业定点、交通问题、就业问题、购买决策等经济决策领域的研究。
- 模型的估计方法主要发展于80年代初期。

一、二元离散选择模型的经济背景

实际经济生活中的二元选择问题

- 研究选择结果与影响因素之间的关系。
- 影响因素包括两部分：**决策者的属性**和**备选方案的属性**。
- 对于单个方案的取舍。例如，购买者对某种商品的购买决策问题，求职者对某种职业的选择问题，投票人对某候选人的投票决策，银行对某客户的贷款决策。由**决策者的属性决定**。
- 对于两个方案的选择。例如，两种出行方式的选择，两种商品的选择。由**决策者的属性和备选方案的属性共同决定**。

二、二元离散选择模型

1、原始模型

- 对于二元选择问题，可以建立如下计量经济学模型。其中 Y 为观测值为1和0的决策被解释变量； X 为解释变量，包括选择对象所具有的属性和选择主体所具有的属性。

$$Y = XB + N \quad y_i = X_i B + \mu_i$$

$$E(\mu_i) = 0 \quad E(y_i) = X_i B$$

$$p_i = P(y_i = 1) \quad 1 - p_i = P(y_i = 0)$$

$$E(y_i) = 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0) = p_i$$

$$E(y_i) = P(y_i = 1) = X_i B$$

左右端矛盾

$$\mu_i = \begin{cases} 1 - X_i B & \text{当 } y_i = 1, \text{ 其概率为 } X_i B \\ -X_i B & \text{当 } y_i = 0, \text{ 其概率为 } 1 - X_i B \end{cases}$$

具有异
方差性

- 由于存在这两方面的问题，所以原始模型不能作为实际研究二元选择问题的模型。
- 需要将原始模型变换为效用模型。
- 这是离散选择模型的关键。

2、效用模型

$$U_i^1 = X_i B^1 + \varepsilon_i^1$$

第*i*个个体 选择**1**的效用

$$U_i^0 = X_i B^0 + \varepsilon_i^0$$

第*i*个个体 选择**0**的效用

$$U_i^1 - U_i^0 = X_i (B^1 - B^0) + (\varepsilon_i^1 - \varepsilon_i^0)$$



$$y_i^* = X_i B + \mu_i^*$$

作为研究对象的二元选择模型

$$P(y_i = 1) = P(y_i^* > 0) = P(\mu_i^* > -X_i B)$$

- 注意，在模型中，效用是不可观测的，人们能够得到的观测值仍然是选择结果，即**1**和**0**。
- 很显然，如果不可观测的 $U^1 > U^0$ ，即对应于观测值为**1**，因为该个体选择公共交通工具的效用大于选择私人交通工具的效用，他当然要选择公共交通工具；
- 相反，如果不可观测的 $U^1 \leq U^0$ ，即对应于观测值为**0**，因为该个体选择公共交通工具的效用小于选择私人交通工具的效用，他当然要选择私人交通工具。

3、最大似然估计

- 欲使得效用模型可以估计，就必须为随机误差项选择一种特定的概率分布。
- 两种最常用的分布是标准正态分布和逻辑（**logistic**）分布，于是形成了两种最常用的二元选择模型—**Probit模型**和**Logit模型**。
- 最大似然函数及其估计过程如下：

$$F(-t) = 1 - F(t)$$


标准正态分布或逻辑分布的对称性


$$\begin{aligned} P(y_i = 1) &= P(y_i^* > 0) = P(\mu_i^* > -X_i B) \\ &= 1 - P(\mu_i^* \leq -X_i B) \\ &= 1 - F(-X_i B) = F(X_i B) \end{aligned}$$

$$P(y_1, y_2, \dots, y_n) = \prod_{y_i=0} (1 - F(X_i B)) \prod_{y_i=1} F(X_i B)$$

似然函数

$$L = \prod_{i=1}^n (F(X_i B))^{y_i} (1 - F(X_i B))^{1-y_i}$$


$$\ln L = \sum_{i=1}^n (y_i \ln F(X_i B) + (1 - y_i) \ln(1 - F(X_i B)))$$


$$\frac{\partial \ln L}{\partial B} = \sum_{i=1}^n \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] X_i = 0$$

1阶极值条件

- 在样本数据的支持下，如果知道概率分布函数和概率密度函数，求解该方程组，可以得到模型参数估计量。

三、二元Probit离散选择模型及其参数估计

1、标准正态分布的概率分布函数

$$F(t) = \int_{-\infty}^t (2\pi)^{-1/2} \exp(-x^2/2) dx$$

$$f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$$

2、重复观测值不可以得到情况下二元Probit 离散选择模型的参数估计

$$\begin{aligned}\frac{\partial \ln L}{\partial \mathbf{B}} &= \sum_{y_i=0} \frac{-f_i}{1-F_i} \mathbf{X}_i + \sum_{y_i=1} \frac{f_i}{F_i} \mathbf{X}_i \\ &= \sum_{i=1}^n \left(\frac{q_i f(q_i \mathbf{X}_i \mathbf{B})}{F(q_i \mathbf{X}_i \mathbf{B})} \right) \mathbf{X}_i \\ &= \sum_{i=1}^n \lambda_i \mathbf{X}_i \\ &= \mathbf{0}\end{aligned}$$

$$q_i = 2y_i - 1$$

- 关于参数的非线性函数，不能直接求解，需采用完全信息最大似然法中所采用的迭代方法。
- 应用计量经济学软件。
- 这里所谓“重复观测值不可以得到”，是指对每个决策者只有一个观测值。如果有多个观测值，也将其看成为多个不同的决策者。

例 贷款决策模型

- **分析与建模：**某商业银行从历史贷款客户中随机抽取**78**个样本，根据设计的指标体系分别计算它们的“商业信用支持度”（**CC**）和“市场竞争地位等级”（**CM**），对它们贷款的结果（**JG**）采用二元离散变量，**1**表示贷款成功，**0**表示贷款失败。目的是研究**JG**与**CC**、**CM**之间的关系，并为正确贷款决策提供支持。

● 样本观测值

CC=XY
CM=SC

JG	XY	SC	JGF	JG	XY	SC	JGF	JG	XY	SC	JGF
0	125.0	-2	0.0000	0	1500	-2	0.0000	0	54.00	-1	0.0000
0	599.0	-2	0.0000	0	96.00	0	0.0000	1	42.00	2	1.0000
0	100.0	-2	0.0000	1	-8.000	0	1.0000	0	42.00	0	0.0209
0	160.0	-2	0.0000	0	375.0	-2	0.0000	1	18.00	2	1.0000
0	46.00	-2	0.0000	0	42.00	-1	6.5E-13	0	80.00	1	6.4E-12
0	80.00	-2	0.0000	1	5.000	2	1.0000	1	-5.000	0	1.0000
0	133.0	-2	0.0000	0	172.0	-2	0.0000	0	326.0	2	0.0000
0	350.0	-1	0.0000	1	-8.000	0	1.0000	0	261.0	1	0.0000
1	23.00	0	0.9979	0	89.00	-2	0.0000	1	-2.000	-1	0.9999
0	60.00	-2	0.0000	0	128.0	-2	0.0000	0	14.00	-2	3.9E-07
0	70.00	-1	0.0000	1	6.000	0	1.0000	1	22.00	0	0.9991
1	-8.000	0	1.0000	0	150.0	-1	0.0000	0	113.0	1	0.0000
0	400.0	-2	0.0000	1	54.00	2	1.0000	1	42.00	1	0.9987
0	72.00	0	0.0000	0	28.00	-2	0.0000	1	57.00	2	0.9999
0	120.0	-1	0.0000	1	25.00	0	0.9906	0	146.0	0	0.0000
1	40.00	1	0.9998	1	23.00	0	0.9979	1	15.00	0	1.0000
1	35.00	1	0.9999	1	14.00	0	1.0000	0	26.00	-2	4.4E-16
1	26.00	1	1.0000	0	49.00	-1	0.0000	0	89.00	-2	0.0000
1	15.00	-1	0.4472	0	14.00	-1	0.5498	1	5.000	1	1.0000
0	69.00	-1	0.0000	0	61.00	0	2.1E-12	1	-9.000	-1	1.0000
0	107.0	1	0.0000	1	40.00	2	1.0000	1	4.000	1	1.0000
1	29.00	1	1.0000	0	30.00	-2	0.0000	0	54.00	-2	0.0000
1	2.000	1	1.0000	0	112.0	-1	0.0000	1	32.00	1	1.0000
1	37.00	1	0.9999	0	78.00	-2	0.0000	0	54.00	0	1.4E-07
0	53.00	-1	0.0000	1	0.000	0	1.0000	0	131.0	-2	0.0000
0	194.0	0	0.0000	0	131.0	-2	0.0000	1	15.00	0	1.0000

obs	JG	CC	CM			
1	0.000000	125.0000	-2.000000			
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19	1.000000	15.00000	-1.000000			
--	--	--	--			

Equation Specification

Equation specification

Binary dependent variable followed by list of regressors.

lg c cc cm

Binary estimation method: ☒ Probit ☐ Logit ☐ Extreme value

Estimation settings

Method: BINARY - Binary choice (logit, probit, extreme value)

Sample: 1 78

OK

Cancel

Options

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
------	-------	---------	-------	------	--------	----------	----------	-------	--------

Dependent Variable: JG

Method: ML - Binary Probit (Quadratic hill climbing)

Date: 11/10/05 Time: 11:04

Sample: 1 78

Included observations: 78

Convergence achieved after 13 iterations

Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	8.797358	7.544067	1.166129	0.2436
CC	-0.257882	0.228894	-1.126642	0.2599
CM	5.061789	4.458482	1.135317	0.2562
Mean dependent var	0.410256	S.D. dependent var		0.495064
S.E. of regression	0.090067	Akaike info criterion		0.118973
Sum squared resid	0.608402	Schwarz criterion		0.209616
Log likelihood	-1.639954	Hannan-Quinn criter.		0.155259
Restr. log likelihood	-52.80224	Avg. log likelihood		-0.021025
LR statistic (2 df)	102.3246	McFadden R-squared		0.968942
Probability(LR stat)	0.000000			
Obs with Dep=0	46	Total obs		78
Obs with Dep=1	32			

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
------	-------	---------	-------	------	--------	----------	----------	-------	--------

Estimation Command:

=====

BINARY(D=N) JG C CC CM

Estimation Equation:

=====

$JG = 1 - @CNORM(-(C(1) + C(2)*CC + C(3)*CM))$

Substituted Coefficients:

=====

$JG = 1 - @CNORM(-(8.797358365 - 0.2578816621*CC + 5.061788659*CM))$

输出的估计结果

- **该方程表示**，当**CC**和**CM**已知时，代入方程，可以计算贷款成功的概率**JGF**。例如，将表中第**19**个样本观测值**CC=15**、**CM=-1**代入方程右边，计算括号内的值为**0.1326552**；查标准正态分布表，对应于**0.1326552**的累积正态分布为**0.5517**；于是，**JG**的预测值**JGF=1-0.5517=0.4483**，即对应于该客户，贷款成功的概率为**0.4483**。

模拟预测

obs	JG	CC	CM	JGF	
1	0.000000	125.0000	-2.000000	0.000000	
2	0.000000	599.0000	-2.000000	0.000000	
3	0.000000	100.0000	-2.000000	0.000000	
4	0.000000	160.0000	-2.000000	0.000000	
5	0.000000	46.00000	-2.000000	0.000000	
6	0.000000	80.00000	-2.000000	0.000000	
7	0.000000	133.0000	-2.000000	0.000000	
8	0.000000	350.0000	-1.000000	0.000000	
9	1.000000	23.00000	0.000000	0.997922	
10	0.000000	60.00000	-2.000000	0.000000	
11	0.000000	70.00000	-1.000000	0.000000	
12	1.000000	-8.000000	0.000000	1.000000	
13	0.000000	400.0000	-2.000000	0.000000	
14	0.000000	72.00000	0.000000	0.000000	
15	0.000000	120.0000	-1.000000	0.000000	
16	1.000000	40.00000	1.000000	0.999803	
17	1.000000	35.00000	1.000000	0.999999	
18	1.000000	26.00000	1.000000	1.000000	
19	1.000000	15.00000	-1.000000	0.447233	
20	0.000000	69.00000	-1.000000	0.000000	

- **预测：**如果有一个新客户，根据客户资料，计算的“商业信用支持度”（**XY**）和“市场竞争地位等级”（**SC**），代入模型，就可以得到贷款成功的概率，以此决定是否给予贷款。

3、重复观测值可以得到情况下二元Probit离散选择模型的参数估计

- 思路

- 对每个决策者有多个重复（例如**10**次左右）观测值。
- 对第*i*个决策者重复观测 n_i 次，选择 $y_i=1$ 的次数比例为 p_i ，那么可以将 p_i 作为真实概率 P_i 的一个估计量。
- 建立“概率单位模型”，采用广义最小二乘法估计。
- 实际中并不常用。

- 对第*i*个决策者重复观测*n*次，选择 $y_i=1$ 的次数比例为 p_i ，那么可以将 p_i 作为真实概率 P_i 的一个估计量。

$$p_i = P_i + e_i = F(X_i B) + e_i$$

定义“观测到的”概率单位

$$E(e_i) = 0$$

$$Var(e_i) = p_i(1 - p_i)/n_i$$

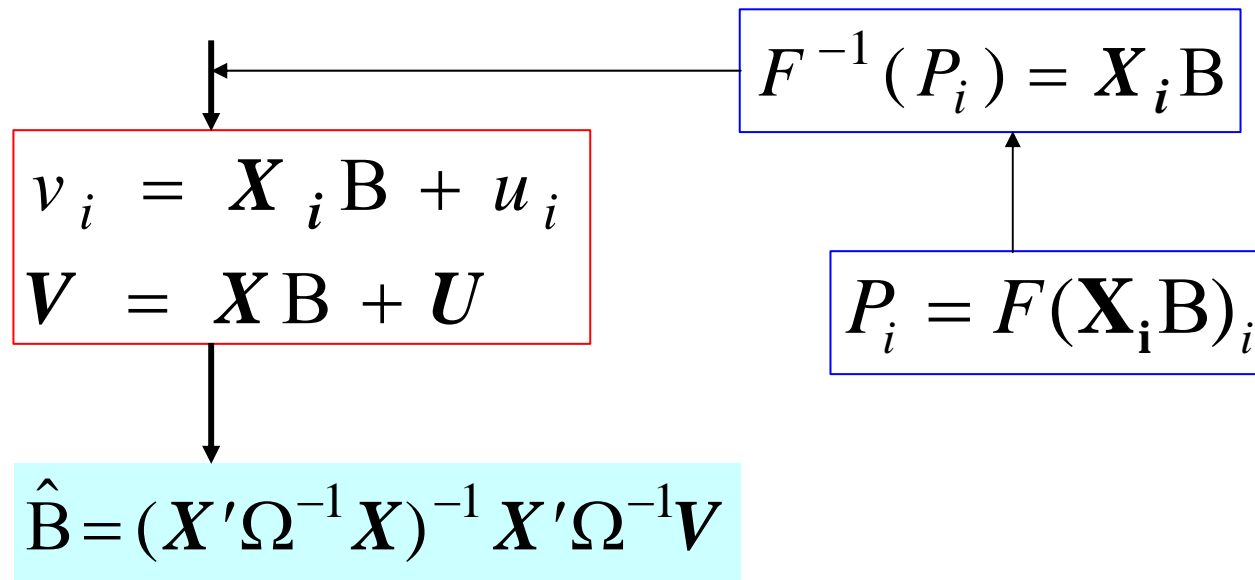
$$v_i = F^{-1}(p_i) = F^{-1}(P_i + e_i)$$

$$F^{-1}(P_i + e_i) = F^{-1}(P_i) + \frac{e_i}{f(F^{-1}(P_i))}$$

$$v_i = F^{-1}(P_i) + u_i$$

$$E(u_i) = 0$$

$$Var(u_i) = \frac{P_i(1 - P_i)}{n_i (f(F^{-1}(P_i)))^2}$$



V 的观测值通过求解标准正态分布的概率分布函数的反函数得到

$$p_i = \int_{-\infty}^{v_i} (2\pi)^{-1/2} \exp(-t^2/2) dt$$

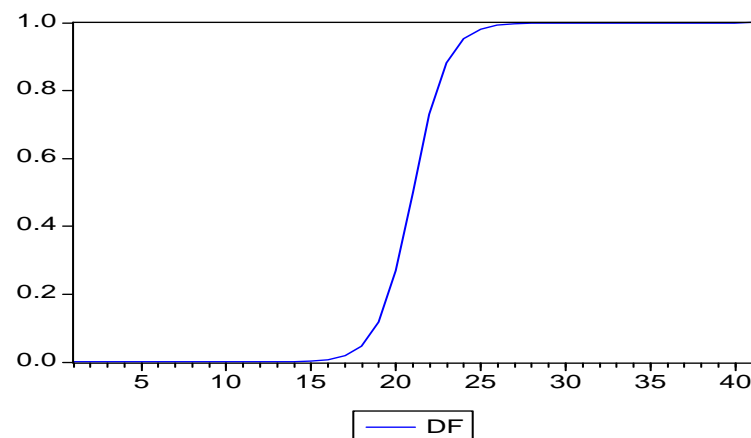
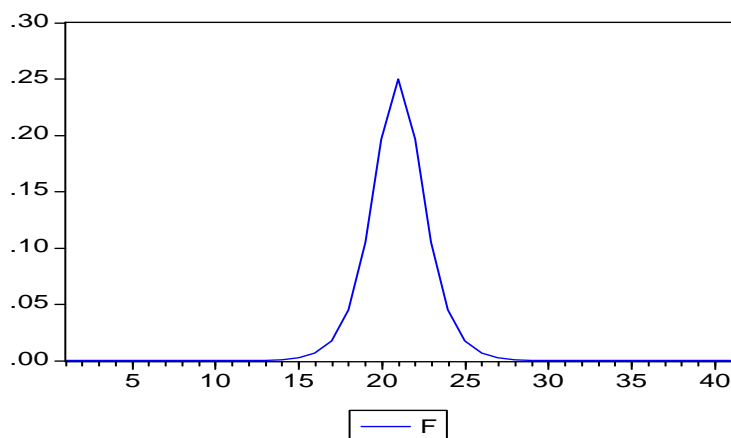
实际观测得到的

四、二元Logit离散选择模型及其参数估计

1、逻辑分布的概率分布函数

$$F(t) = \frac{1}{1 + e^{-t}} \longrightarrow F(t) = \frac{e^t}{1 + e^t} = \Lambda(t)$$

$$f(t) = \frac{e^{-t}}{(1 + e^{-t})^2} \longrightarrow f(t) = \frac{e^t}{(1 + e^t)^2} = \Lambda(t)(1 - \Lambda(t))$$



Börsch-Supan于1987年指出:

- 如果选择是按照效用最大化而进行的, 具有极限值的逻辑分布是较好的选择, 这种情况下的二元选择模型应该采用**Logit**模型。

2、重复观测值不可以得到情况下二元logit 离散选择模型的参数估计

$$\begin{aligned}\frac{\partial \ln L}{\partial B} &= \sum_{i=1}^n \left[\frac{y_i f_i}{F_i} + (1 - y_i) \frac{-f_i}{(1 - F_i)} \right] X_i \\ &= \sum_{i=1}^n (y_i - \Lambda(X_i B)) X_i = 0\end{aligned}$$

- 关于参数的非线性函数，不能直接求解，需采用完全信息最大似然法中所采用的迭代方法。
- 应用计量经济学软件。

obs	JG	CC	CM			
1	0.000000	125.0000	-2.000000			
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19	1.000000	15.00000	-1.000000			

Equation Specification

Equation specification

Binary dependent variable followed by list of regressors.

lg c cc cm

Binary estimation method: ☐ Probit ☒ Logit ☐ Extreme value

Estimation settings

Method: BINARY - Binary choice (logit, probit, extreme value)

Sample: 1 78

OK

Cancel

Options

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
------	-------	---------	-------	------	--------	----------	----------	-------	--------

Dependent Variable: JG

Method: ML - Binary Logit (Newton-Raphson)

Date: 11/10/05 Time: 16:53

Sample: 1 78

Included observations: 78

Convergence achieved after 13 iterations

Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	16.11426	14.56353	1.106481	0.2685
CC	-0.465035	0.431760	-1.077068	0.2814
CM	9.379903	8.712437	1.076611	0.2817
Mean dependent var	0.410256	S.D. dependent var		0.495064
S.E. of regression	0.091187	Akaike info criterion		0.120325
Sum squared resid	0.623629	Schwarz criterion		0.210968
Log likelihood	-1.692674	Hannan-Quinn criter.		0.156611
Restr. log likelihood	-52.80224	Avg. log likelihood		-0.021701
LR statistic (2 df)	102.2191	McFadden R-squared		0.967943
Probability(LR stat)	0.000000			
Obs with Dep=0	46	Total obs		78
Obs with Dep=1	32			

View	Procs	Objects	Print	Name	Freeze	Estimate	Forecast	Stats	Resids
------	-------	---------	-------	------	--------	----------	----------	-------	--------

Estimation Command:

=====

BINARY(D=L,R) JG C CC CM

Estimation Equation:

=====

JG = 1-@LOGIT(-(C(1) + C(2)*CC + C(3)*CM))

Substituted Coefficients:

=====

JG = 1-@LOGIT(-(16.11426399 - 0.4650347429*CC + 9.379903458*CM))

obs	JG	CC	CM	JGFF	
1	0.000000	125.0000	-2.000000	0.000000	
2	0.000000	599.0000	-2.000000	0.000000	
3	0.000000	100.0000	-2.000000	0.000000	
4	0.000000	160.0000	-2.000000	0.000000	
5	0.000000	46.00000	-2.000000	3.64E-11	
6	0.000000	80.00000	-2.000000	0.000000	
7	0.000000	133.0000	-2.000000	0.000000	
8	0.000000	350.0000	-1.000000	0.000000	
9	1.000000	23.00000	-2.000000	0.995586	
10	0.000000	60.00000	-2.000000	5.41E-14	
11	0.000000	70.00000	-2.000000	6.13E-12	
12	1.000000	-8.00000	-2.000000	1.000000	
13	0.000000	400.0000	-2.000000	0.000000	
14	0.000000	72.00000	-2.000000	2.86E-08	
15	0.000000	120.0000	-1.000000	0.000000	
16	1.000000	40.00000	1.000000	0.998986	
17	1.000000	35.00000	1.000000	0.999901	
18	1.000000	26.00000	1.000000	0.999998	
19	1.000000	15.00000	-1.000000	0.440000	
20	0.000000	69.00000	-1.000000	9.76E-12	

Probit
0.999999
1.000000
0.447233
0.000000

3、重复观测值可以得到情况下二元logit离散选择模型的参数估计

- 思路

- 对每个决策者有多个重复（例如**10**次左右）观测值。
- 对第*i*个决策者重复观测 n_i 次，选择 $y_i=1$ 的次数比例为 p_i ，那么可以将 p_i 作为真实概率 P_i 的一个估计量。
- 建立“对数成败比例模型”，采用广义最小二乘法估计。
- 实际中并不常用。

- 用样本重复观测得到的 p_i 构成“成败比例”，取对数并进行台劳展开，有

$$\ln\left(\frac{p_i}{1-p_i}\right) \approx \ln\left(\frac{P_i}{1-P_i}\right) + \frac{e_i}{P_i(1-P_i)}$$

$$\frac{P_i}{1-P_i} = e^{X_i B}$$

$$F(t) = \frac{1}{1+e^{-t}}$$

$$\frac{F(t)}{1-F(t)} = e^t$$

$$\ln\left(\frac{p_i}{1-p_i}\right) \approx \ln(e^{X_i B}) + u_i = X_i B + u_i$$

$$v_i = X_i B + u_i$$

$$V = XB + U$$

$$\hat{B} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} V$$

逻辑分布的概率
分布函数

五、二元离散选择模型的检验

1、检验假设举例—省略变量检验

- 经典模型中采用的变量显著性t检验仍然是有效的。
- 如果省略的变量与保留的变量不是正交的，那么对参数估计量将产生影响，需要进一步检验这种省略是否恰当。

零假设为: $H_0 : \mathbf{Y}^* = \mathbf{X}_1 \mathbf{B}_1 + \mu^*$

备择假设为: $H_1 : \mathbf{Y}^* = \mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2 + \mu^*$

2、统计量

用于检验的统计量为 Wald 统计量、LR 统计量和 LM 统计量，具体计算方法如下：

$$W = \hat{B}_2' V_2^{-1} \hat{B}_2$$

其中 $V_2 = \text{AsyVar}(\hat{B}_2)$ 。

$$LR = -2(\ln \hat{L}_0 - \ln \hat{L}_1)$$

如果X2中的变量省略后对参数估计量没有影响，那么H1和H0情况下的对数最大似然函数值应该相差不大，此时LR统计量的值很小，自然会小于临界值，不拒绝 H0。

其中 \hat{L}_0 、 \hat{L}_1 分别为 H_0 情形和 H_1 情形下的似然函数值的估计量。

$$LM = g_0' V_0^{-1} g_0$$

其中 g_0 是 H_1 情形下的对数似然函数对参数估计量的一阶导数向量，

用 H_0 情形下的最大似然参数估计量代入计算； V_0 是 H_1 情形下参数最大似然估计量的方差矩阵估计量。

3. 异方差性检验

- 截面数据样本，容易存在异方差性。
- 假定异方差结构为：

$$\text{Var}(\varepsilon|\mathbf{X}, \mathbf{Z}) = (\exp(\mathbf{Z}'\boldsymbol{\gamma}))^2$$

$$H_0 : \boldsymbol{\gamma} = \mathbf{0}$$

- 采用**LM**检验

将解释变量分为两类， \mathbf{Z} 为只与个体特征有关的变量。显然异方差与这些变量相关。

将异方差检验问题变为一个约束检验问题

4. 分布检验

- 检验关于分布的假设（**probit**、**logit**）。
- 一般不进行该项检验。
- 具体见相关教科书。