

专题：定性响应模型

Models with Discrete Dependent Variables

- 6.1 定性响应模型的性质
- 6.2 线性概率模型(LPM)
- 6.3 Logit模型
- 6.4 Probit模型
- 6.5 LPM、Logit模型及Probit模型比较

➤ 什么是定性响应模型？

- 简单地说，被解释变量Y是定性变量的回归模型。
 - 定性变量包括：
 - 顺序型：产品等级、教育程度、满意程度；（顺序取值）
 - 分类型：性别、种族、肤色、宗教、国籍、地区、政治动乱和党派等。（0-1取值，没有次序）

➤ 定性响应模型的用途

- 研究某些决策选择（被解释变量Y的取值）受其他变量（解释变量X）影响的概率。
- 定制响应模型，也称概率模型。

➤ 定性响应模型分类

- 二值响应模型 (Binary Response)
 - 线性概率模型 (LPM)
 - Logit模型
 - Probit模型
 - Tobit模型
- 多值响应模型 (Multiple-response)
 - 顺序响应模型
 - 无序响应模型
 - 持续期限模型

线性概率模型 (Linear Probability Model, LPM)

➤ 线性概率模型 (LPM):

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (15.2.1)$$

其中 X = 家庭收入。 $Y=1$ ，如果该家庭拥有住房； $Y=0$ ，如果该家庭不拥有住房。

Y_i 服从贝努利概率分布：

Y_i	概率
0	$1 - P_i$
1	P_i
总和	1

Y_i 的期望（估计值）： $E(Y_i | X_i) = \beta_1 + \beta_2 X_i = P_i$

Y_i 的方差： $\text{var}(u_i) = P_i(1 - P_i)$

LPM 的估计方法：OLS

线性概率模型 (Linear Probability Model, LPM)

- 线性概率模型的不足：
- 随机扰动项 μ_i 非正态分布；
 - 随机扰动项 μ_i 的异方差；
 - 难以满足 $0 \leq E(Y_i | X_i) \leq 1$ 的约束；
 - 偏低的 R^2 。

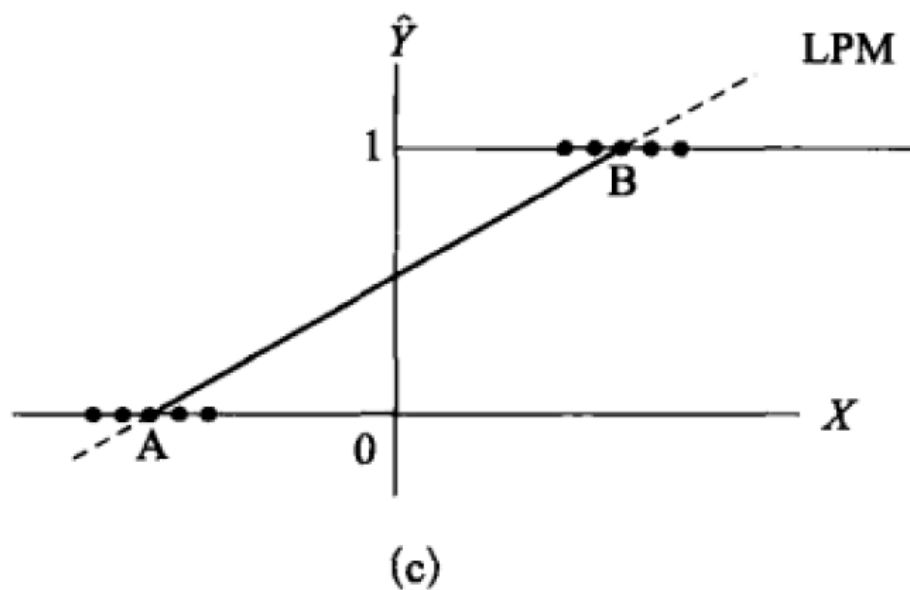


图 15—1 线性概率模型

线性概率模型 (Linear Probability Model, LPM)

➤ 线性概率模型的改进:

- 针对不满足 $0 \leq E(Y_i | X_i) \leq 1$ 情形
 - 思路1 (Censored data): **Tobit模型** (Probit模型的扩展)
当 Y_i 的估计值落在 $[0,1]$ 区间之外时,
 Y_i 的估计值 < 0 , 取值为 0
 Y_i 的估计值 > 1 , 取值为 1
 - 思路2 (函数变换): **Logit模型/Probit模型**
设计一种方法, 使得 $0 \leq E(Y_i | X_i) \leq 1$ 成立。
实现方式: CDF函数

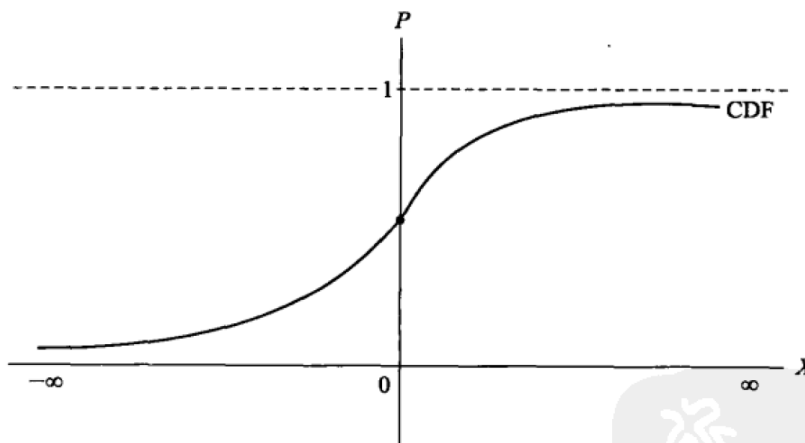


图 15—2 一个累积分布函数 (CDF)

➤ Logit模型

- 线性模型

$$P_i = \beta_1 + \beta_2 X_i$$

- Logistic分布函数（CDF）的转换：

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^{Z_i}}{1 + e^{Z_i}}$$

其中 $Z_i = \beta_1 + \beta_2 X_i$ 。

Logit模型： $L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \beta_1 + \beta_2 X_i + u_i$

其中： $\frac{P_i}{1 - P_i}$ 称为机会比率（odd ratio）

Logit模型的估计方法：MLE(极大似然法)

表 15—7 个性化教学体系 (PSI) 对课程成绩影响的数据

观测	GPA 分数	TUCE 分数	PSI	期末 成绩 Y	成绩 等级	观测	GPA 分数	TUCE 分数	PSI	期末 成绩 Y	成绩等级
1	2.66	20	0	0	C	17	2.75	25	0	0	C
2	2.89	22	0	0	B	18	2.83	19	0	0	C
3	3.28	24	0	0	B	19	3.12	23	1	0	B
4	2.92	12	0	0	B	20	3.16	25	1	1	A
5	4.00	21	0	1	A	21	2.06	22	1	0	C
6	2.86	17	0	0	B	22	3.62	28	1	1	A
7	2.76	17	0	0	B	23	2.89	14	1	0	C
8	2.87	21	0	0	B	24	3.51	26	1	0	B
9	3.03	25	0	0	C	25	3.54	24	1	1	A
10	3.92	29	0	1	A	26	2.83	27	1	1	A
11	2.63	20	0	0	C	27	3.39	17	1	1	A
12	3.32	23	0	0	B	28	2.67	24	1	0	B
13	3.57	23	0	0	B	29	3.65	21	1	1	A
14	3.26	25	0	1	A	30	4.00	23	1	1	A
15	3.53	26	0	0	B	31	3.10	21	1	0	C
16	2.74	19	0	0	B	32	2.39	19	1	1	A

注：成绩 $Y = 1$ ，如果期末成绩等级为 A；

$= 0$ ，如果期末成绩等级为 B 或 C。

TUCE=学期初为测试学生的宏观经济学知识而进行的一项考试的成绩（期初测试成绩）。

PSI = 1，如果采用新的教学方法；

$= 0$ ，如果不采用新的教学方法。

GPA=开始学习中级微观经济学时的平均成绩。

资料来源：L. Spector and M. Mazzeo, "Probit Analysis and Economic Education," *Journal of Economic Education*, vol. 11, 1980, pp. 37-44.

表 15—8

方程 (15.8.1) 的回归结果

Dependent Variable: Grade				
Method: ML-Binary Logit				
Convergence achieved after 5 iterations				
Variable	Coefficient	Std. Error	Z Statistic	Probability
C	-13.0213	4.931	-2.6405	0.0082
GPA	2.8261	1.2629	2.2377	0.0252
TUCE	0.0951	0.1415	0.67223	0.5014
PSI	2.3786	1.0645	2.2345	0.0255
McFadden $R^2 = 0.3740$ LR statistic (3 df) = 15.40419				

回归结果估计系数的解释：

解释1：解释变量 i 的1单位变动导致机会比率的对数 β_i 单位变化

解释2：解释变量 i 的1单位变动导致机会比率 e^{β_i} 单位变动；

比如：上述结果中，PSI的系数为2.3786，这表明：给定其他条件不变，接受新教学方法的学生获得A的概率比没有接受新教学方法的学生高出近10倍（ $e^{2.3786} - e^0 \approx 10.7897 - 1 = 9.7897 \approx 10$ ）。

➤ Probit模型

- 线性模型

$$P_i = \beta_1 + \beta_2 X_i$$

- Probit的转换:

第1步: $I_i = \beta_1 + \beta_2 X_i$
其中 X_i 表示第 i 个家庭的收入。

第2步:

给定正态性假定, $I_i^* \leq I_i$ 的概率可由标准化正态 CDF 算出^①:

$$P_i = P(Y = 1 | X) = P(I_i^* \leq I_i) = P(Z_i \leq \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i)$$

第3步: Probit模型

$$\begin{aligned} F(I_i) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_1 + \beta_2 X_i} e^{-z^2/2} dz \end{aligned}$$

Proit模型的估计方法: MLE(极大似然法)

表 15—13

Dependent Variable: grade				
Method: ML-Binary probit				
Convergence achieved after 5 iterations				
Variable	Coefficient	Std. Error	Z-Statistic	Probability
C	-7.4523	2.5424	-2.9311	0.0033
GPA	1.6258	0.6938	2.3430	0.0191
TUCE	0.0517	0.0838	0.6166	0.5374
PSI	1.4263	5950	2.3970	0.0165
LR statistic (3 df) = 15.5458			McFadden R^2 = 0.3774	
Probability (LR stat) = 0.0014				

回归结果估计系数的解释：

解释变量 i 的1单位变动导致被解释变量选择概率的 $f(\beta_1 + \beta_2 X_i)\beta_2$ 单位变动；

Logit模型与Probit模型比较

- Logit模型与Probit模型的分布比较:

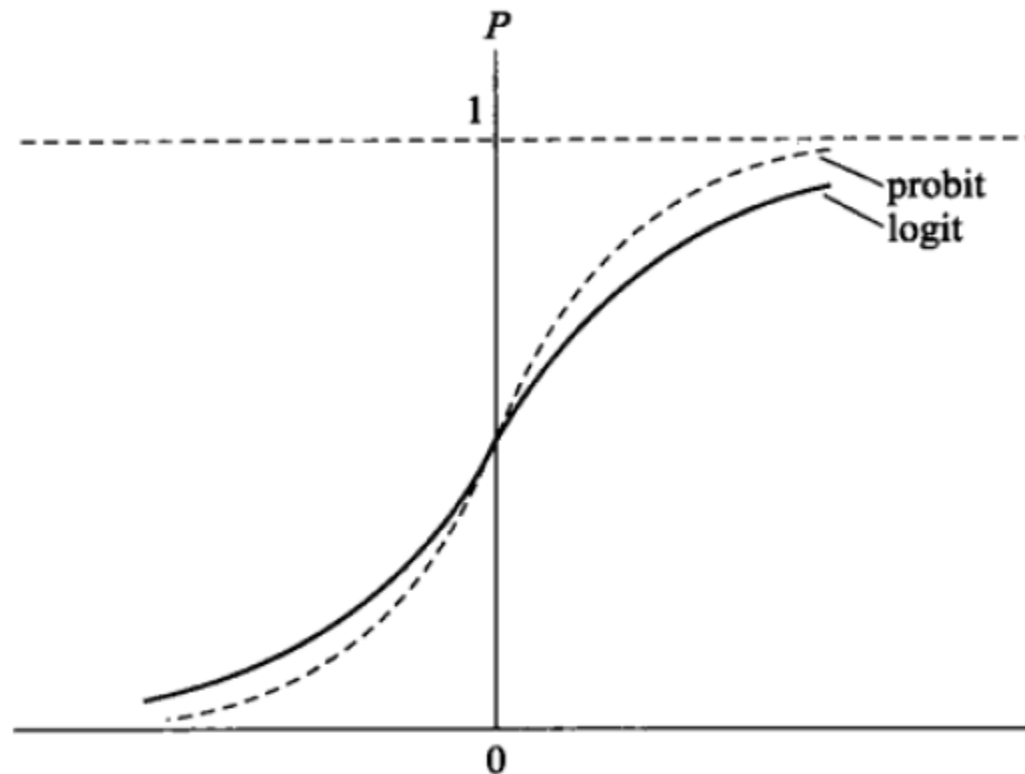


图 15—6 logit 和 probit 模型的累积分布

- 实际上，Logit模型与Probit模型分布比较接近，估计结果也比较接近。

Logit模型与Probit模型比较

➤ Logit模型与Probit模型的估计结果比较:

表 15—8

方程 (15.8.1) 的回归结果

Dependent Variable: Grade				
Method: ML-Binary Logit				
Convergence achieved after 5 iterations				
Variable	Coefficient	Std. Error	Z Statistic	Probability
C	-13.0213	4.931	-2.6405	0.0082
GPA	2.8261	1.2629	2.2377	0.0252
TUCE	0.0951	0.1415	0.67223	0.5014
PSI	2.3786	1.0645	2.2345	0.0255
McFadden $R^2 = 0.3740$ LR statistic (3 df) = 15.40419				

表 15—13

Dependent Variable: grade				
Method: ML-Binary probit				
Convergence achieved after 5 iterations				
Variable	Coefficient	Std. Error	Z-Statistic	Probability
C	-7.4523	2.5424	-2.9311	0.0033
GPA	1.6258	0.6938	2.3430	0.0191
TUCE	0.0517	0.0838	0.6166	0.5374
PSI	1.4263	5950	2.3970	0.0165
LR statistic (3 df) = 15.5458 McFadden $R^2 = 0.3774$				
Probability (LR stat) = 0.0014				

Logit模型与Probit模型比较

- 线性概率模型(OLS模型)与logit/probit模型估计结果比较:

比较项目	OLS 模型	Logit 模型/Probit 模型
估计方法	OLS (最小二乘法)	ML(极大似然法)
单个系数显著性评价	t 统计量 (t 分布)	Z 统计量 (标准正态)
拟合优度	R^2 , 或者调整的 R^2	McFadden R^2 或计数 R^2
所有系数联合检验	Wald 检验	LR(似然比)检验

- Eviews回归结果: 线性概率模型、Logit模型和Probit模型
- 系数比较:

$$\beta_{probit} * 1.81 \approx \beta_{logit}$$

$$\beta_{LPM} = 0.25\beta_{logit} \quad \text{除截距之外的其他系数}$$

$$\beta_{LPM} = 0.25\beta_{logit} + 0.5 \quad \text{截距}$$