第四节　带虚变量的回归预测技术

- 在回归模型中，因变量受两类变量的影响
  - 数量变量：产量、销售量、收入、价格、成本、身高、温度
  - 品质变量：性别、文化程度、宗教、战争、地震、季节、历史、社会、文化背景、婚否
  - 品质变量不像数量变量那样可以用不同的数值表现，他只能以品质、属性、种类等具体形式来表现，出现为1，不出现为0。
  - 例：
    - 某冷饮的销售量=f（人均收入、季节）
    - 储蓄总额=f（人均收入、物价指数、政策）

一、虚变量的概念

- 水平：虚变量的出现形式

  - 例：政策　政策有变化为1水平
    政策无变化为2水平

    季节　　春季为1水平
    夏季为2水平
    秋季为3水平
    冬季为4水平

- 反应：

用$\delta_i(j,k)$表示第i个样本第j个虚变量取第k个水平的反应。

其中 $\delta_i(j,k)=\begin{cases}1 & \text{当第}i\text{个样本第}j\text{个虚变量取第}k\text{个水平}\\ 0 & \text{否则}\end{cases}$

- 例：若 y=f(季节、销售政策、价格)

第一次观测是在夏季、政策有改变时得到的

第二次观测是在春季、政策无改变时得到的

$$\delta_1(1,1)=0 \quad \delta_2(1,1)=1$$
$$\delta_1(1,2)=1 \quad \delta_2(1,2)=0$$
$$\delta_1(1,3)=0 \quad \delta_2(1,3)=0$$
$$\delta_1(1,4)=0 \quad \delta_2(1,4)=0$$
$$\delta_1(2,1)=1 \quad \delta_2(2,1)=0$$
$$\delta_1(2,2)=0 \quad \delta_2(2,2)=1$$

反应表，反应矩阵：

例：某服装销售量 $y=f(x_1,x_2)$,其中x_1表示季节,有春、夏、秋、冬四个水平；x_2表示销售策略，分为保留原销售策略和改变原销售策略两个水平。构造反应表：

		季节x_1				政策 x_2	
样本编号	y_i	春	夏	秋	冬	有	无
1	30	1	0	0	0	0	1
2	28	0	1	0	0	0	1
3	35	0	0	1	0	0	1
4	25	0	0	0	1	1	0
5	23	1	0	0	0	1	0

反应矩阵：
$$X=\begin{pmatrix}1&0&0&0&0&1\\0&1&0&0&0&1\\0&0&1&0&0&1\\0&0&0&1&1&0\\1&0&0&0&1&0\end{pmatrix}$$

若有m个虚变量 $x_1, x_2, \cdots, x_m$，第j个虚变量x_j有r_j个水平，则一般的反应表见书**104**页。

表 8-2

样本编号	自变量 / 反应 / 水平 / 因变量	x_1	…	x_m
		$C_{11}C_{12}\cdots C_{1r_1}$	…	$C_{m1}\cdots C_{mr_m}$
1	y_1	$\delta_1(1,1)\delta_1(1,2)\cdots\delta_1(1,r_1)$	……	$\delta_1(m,1)\cdots\delta_1(m,r_m)$
2	y_2	$\delta_2(1,1)\delta_2(1,2)\cdots\delta_2(1,r_1)$	……	$\delta_2(m,1)\cdots\delta_2(m,r_m)$
⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮	⋮ ⋮ ⋮
n	y_n	$\delta_n(1,1)\delta_n(1,2)\cdots\delta_n(1,r_1)$	……	$\delta_n(m,1)\cdots\delta_n(m,r_m)$

二、建立预测模型

- 只含虚变量

假设有m个虚变量 $x_1, x_2, \cdots, x_m$，其中x_j有r_j个水平，y与x有线性关系

$$y_i = \sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j,k)b_{jk} + e_i \qquad i = 1,2,\cdots,n.$$

用矩阵形式，令

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = [\delta_i(j,k)] \qquad B = \begin{pmatrix} b_{11} \\ \vdots \\ b_{1r_1} \\ \vdots \\ b_{m1} \\ \vdots \\ b_{mr_m} \end{pmatrix} \qquad E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$Y = XB + E \qquad \hat{Y} = X\hat{B}$$

由最小二乘法，可知 $\hat{B}$ 满足正规方程

$$X'X\hat{B} = X'Y \qquad (*)$$

$$\because r(X) \le \sum_{j=1}^{m} r_j - (m-1)$$

$\therefore X'X$是奇异矩阵

又$\because$可以证明 $r(X'X \mid X'Y) = r(X) = r(X'X)$

$\therefore$(*)式有无穷多解.

可以证明，用任意一组解作预测，预测结果相同。

$\hat{B}$的一个特解：书106页。

由于假定 X^TX 的秩为 $\sum_{j=1}^{m} r_j - (m-1)$，故在求解时，先删去第 j 个自变量第一个水平所对应的方程（$j=2, 3, \cdots, m$），总共删去 $m-1$ 个，然后令 $\hat{b}_{j1}=0$，$j=2, 3, \cdots, m$。此时，剩下的方程组其系数矩阵是满秩的，故可惟一解出其余的 $\hat{b}_{jk}$。这种特定的解，称为方程组（8.2—4）的特解。记为 $\hat{B}^{0r}$

$$\hat{B}^{0r} = (\hat{b}^0_{11}, \hat{b}^0_{12}, \cdots, \hat{b}^0_{1r_1}; 0, \hat{b}^0_{22}, \cdots, \hat{b}^0_{2r_2}; \cdots; 0, \hat{b}^0_{mr_2}, \cdots, \hat{b}^0_{mr_m})$$

由此得到预测方程为

$$\hat{y}_i = \sum_{j=1}^{m} \sum_{k=1}^{r_j} \delta_i(j,k) \hat{b}^0_{jk} \qquad (8.2-5)$$

例：

$$X'XB=\left(\begin{array}{cccccc}2&0&0&0&1&1\\0&1&0&0&0&1\\0&0&1&0&0&1\\0&0&0&1&1&0\\1&0&0&1&2&0\\1&1&1&0&0&3\end{array}\right)\left(\begin{array}{c}b_{11}\\b_{12}\\b_{13}\\b_{14}\\b_{21}\\b_{22}\end{array}\right)=X'Y=\left(\begin{array}{c}y_1+y_5\\y_2\\y_3\\y_4\\y_4+y_5\\y_1+y_2+y_3\end{array}\right)$$

令 $\hat{b}_{21}=0$, 因此得到

$$\left(\begin{array}{ccccc}2&0&0&0&1\\0&1&0&0&1\\0&0&1&0&1\\0&0&0&1&0\\1&1&1&0&3\end{array}\right)\left(\begin{array}{c}b_{11}\\b_{12}\\b_{13}\\b_{14}\\b_{22}\end{array}\right)=\left(\begin{array}{c}y_1+y_5\\y_2\\y_3\\y_4\\y_1+y_2+y_3\end{array}\right)=\left(\begin{array}{c}53\\28\\35\\25\\81\end{array}\right)$$

$$\therefore \hat{B} = (y_5, y_2 - y_1 + y_5, y_3 - y_1 + y_5, y_4, 0, y_1 - y_5)'$$

$$= (23, 21, 28, 25, 0, 7)'$$

$$\hat{y} = 23 \cdot \delta(1,1) + 21 \cdot \delta(1,2) + 28 \cdot \delta(1,3) + 25 \cdot \delta(1,4) + 7 \cdot \delta(2,2)$$

若冬季、保留原政策:

$$\hat{y} = 23 \cdot 0 + 21 \cdot 0 + 28 \cdot 0 + 25 \cdot 1 + 7 \cdot 1 = 32$$

- 一般情况

自变量中既有虚变量，也有普通变量

$$y : x_1, x_2, \cdots, x_m, \qquad x_{m+1}, \cdots, x_r$$

$$y_i = \sum_{j=1}^{m} \sum_{k=1}^{r_j} \delta_i(j,k) b_{jk} + \beta_{m+1} x_{m+1,i} + \cdots + \beta_r x_{r,i} + \varepsilon_i$$

三、预测精度

$$\hat{y}_i = \sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta_i(j,k)\hat{b}_{jk} + \sum_{j=m+1}^{r}\beta_j x_{j,i} \qquad (i = 1,2,\cdots,n)$$

复相关系数 $$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

标准偏差 $$S = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - r}}$$

四、举例

某省农业生产资料购买力和农民货币收入统计数据如下：

年份	1975	1976	1977	1978	1979	1980
农资购买力y	1.3	1.3	1.4	1.5	1.8	2.1
货币收入x	4.7	5.4	5.5	6.9	9.0	10.0
年份	1981	1982	1983	1984	1985	
农资购买力y	2.3	2.6	2.7	3.0	3.2	
货币收入x	11.3	13.4	15.2	19.3	27.8	

试建立预测模型。若1986年农民货币收入为30，试求农资购买力？

解：（1）建立一元线性回归模型

$$\hat{y} = 1.0161 + 0.09357x \qquad R^2 = 0.8821 \quad F = 67.3266$$

$$\hat{y}_{1986} = 1.0161 + 0.09357 \times 30 = 3.8232$$

(2)引入虚变量来反应经济政策的影响。有两个水平，政策无变化为1水平，政策有变化为2水平，所以反应：

$$\delta_i(2,1) = \begin{cases} 1 & i < 1979 \\ 0 & i \geq 1979 \end{cases} \qquad \delta_i(2,2) = \begin{cases} 0 & i < 1979 \\ 1 & i \geq 1979 \end{cases}$$

反应表为

编号	y	x_1	无	有
1	1.3	4.7	1	0
2	1.3	5.4	1	0
3	1.4	5.5	1	0
4	1.5	6.9	1	0
5	1.8	9.0	0	1
6	2.1	10.0	0	1
7	2.3	11.3	0	1
8	2.6	13.4	0	1
9	2.7	15.2	0	1
10	3.0	19.3	0	1
11	3.2	27.8	0	1

$$\hat{b}_1 = 0.0692 \quad \hat{b}_{21} = 0.9855 \quad \hat{b}_{22} = 1.48$$

$$\therefore \hat{y} = 0.0692x_1 + 0.9855 \times \delta(2,1) + 1.48 \times \delta(2,2) \qquad R^2 = 0.9498 > 0.8821$$

1986年时： $\delta(2,1) = 0 \quad \delta(2,2) = 1$

$$\hat{y}_{1986} = 0.0692 \times 30 + 1.48 = 3.556$$

五、虚变量的引入方式

- 1。加法方式
- 2。乘法方式
- 3。临界指标的虚拟变量的引入

1。加法方式——影响截距

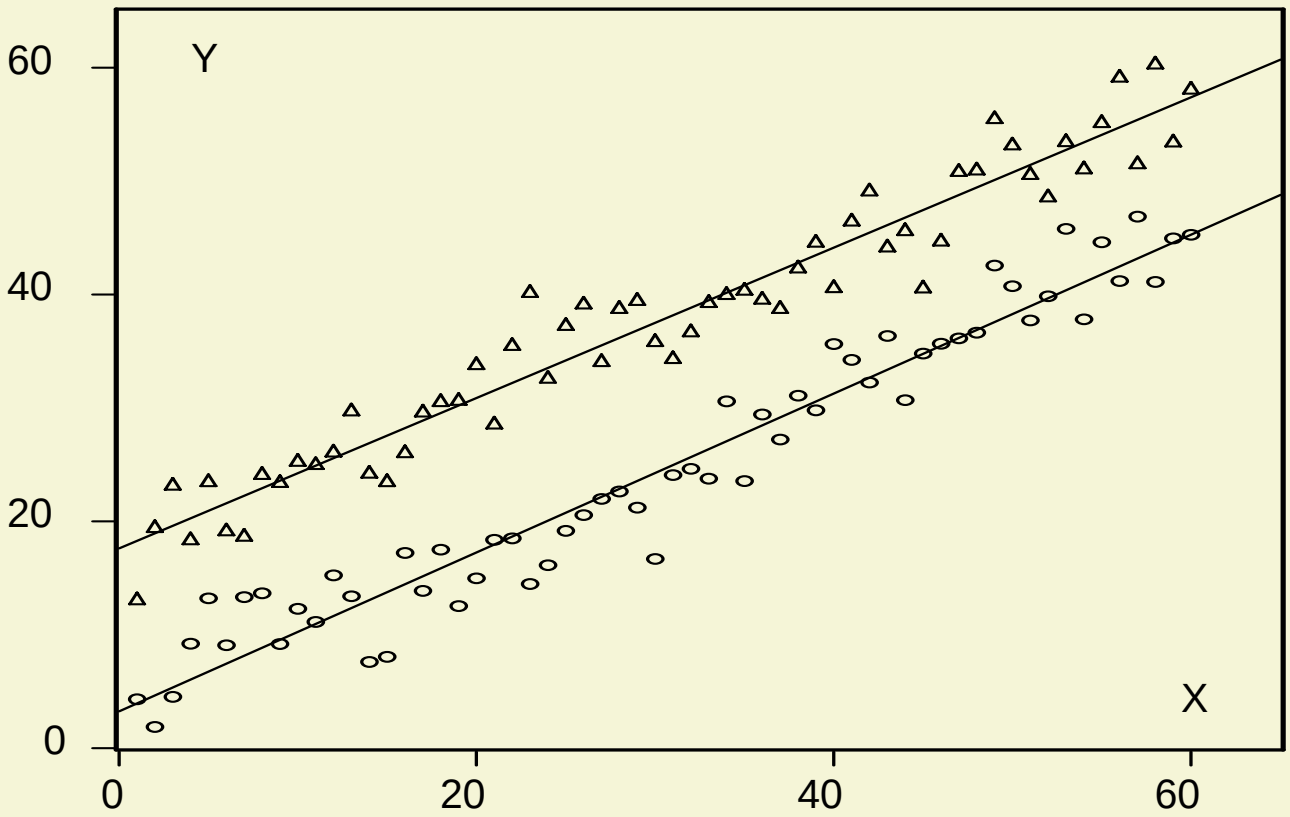
设有模型，

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 D + u_t ,$$

其中 y_t，x_t 为定量变量；D 为定性变量。

当 $D = 0$ 或 1 时，上述模型可表达为，

$$y_t = \begin{cases} \beta_0 + \beta_1 x_t + u_t , & (D = 0) \\ (\beta_0 + \beta_2) + \beta_1 x_t + u_t , & (D = 1) \end{cases}$$



例如：中国成年人体重 y（kg）与身高 x（cm）的回归关系如下：

$$y = -100 + x - 5D = \begin{cases} -105 + x & D = 1 \text{ (男)} \\ -100 + x & D = 0 \text{ (女)} \end{cases}$$

2。乘法方式——影响斜率

设有模型

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t D + u_t ,$$

其中 x_t 为定量变量；D 为定性变量。当 $D = 0$ 或 1 时，上述模型可表达为，

$$y_t = \begin{cases} \beta_0 + (\beta_1 + \beta_2)x_t + u_t , & (D = 1) \\ \beta_0 + \beta_1 x_t + u_t , & (D = 0) \end{cases}$$

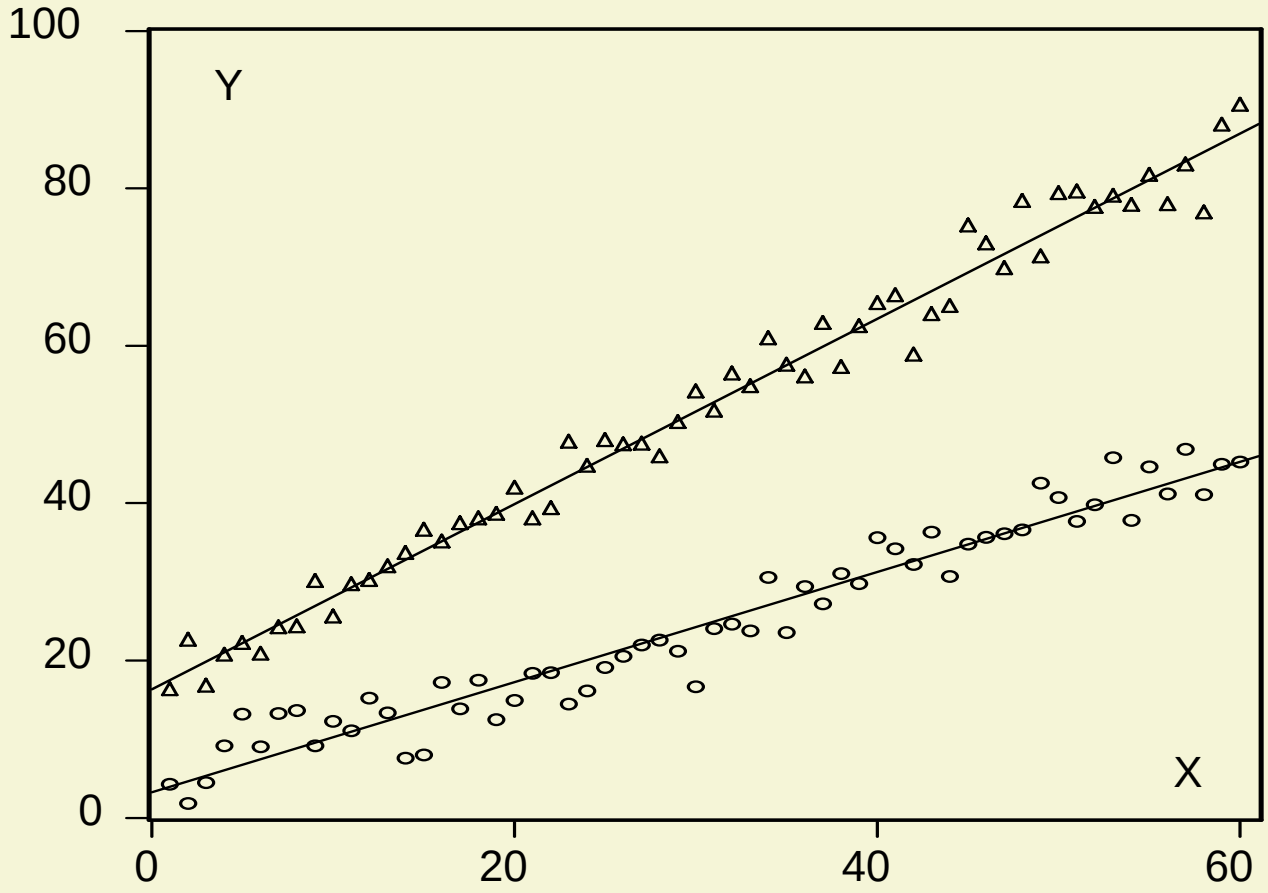
加法与乘法组合引入——
截距与斜率均不同

设有模型

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 D + \beta_3 x_t D + u_t ,$$

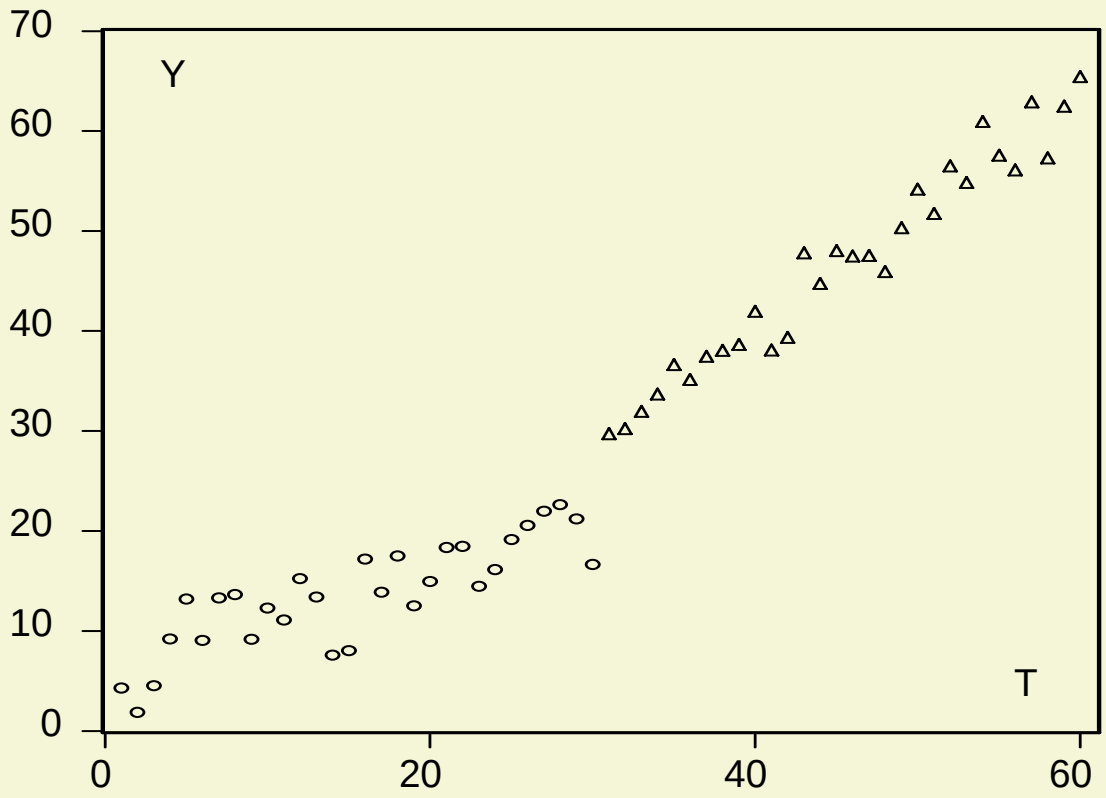
其中 x_t 为定量变量；D 为定性变量。当 $D = 0$ 或 1 时，上述模型可表达为，

$$y_t = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_t + u_t , & (D = 1) \\ \beta_0 + \beta_1 x_t + u_t , & (D = 0) \end{cases}$$


100
80
60
40
20
0
Y
X
0
20
40
60

3。临界指标的虚拟变量的引入

- 在经济转折时期，可以建立临界值指标的虚拟变量模型来反映
- 设转折时期 t* 转折时期的指标值= x*
- 虚拟变量 D=1（ t >= t*） D=0（ t < t*）
- 模型 $y = b_0 + b_1 x + b_2 (x-x^*) D + e$
- $t < t^*$ 时 $y = b_0 + b_1 x + e$
- $t >= t^*$ 时 $y = b_0 - b_2 x^* + (b_1 + b_2) x + e$
- 当$t = t^*$时， $x=x^*$ 两式计算的y 相等，两条直线在转折期连接成一条折线


70
60
50
40
30
20
10
0
Y
T
0
20
40
60

六、虚变量的作用

- 1、分离异常因素的影响.

  例如分析我国GDP的时间序列，必须考虑“文革”因素对国民经济的破坏性影响，剔除不可比的“文革”因素。

- 2、检验不同属性类型对因变量的作用.

  例如工资模型中的文化程度、季节对销售额的影响。

- 3、提高模型的精度.

  相当与将不同属性的样本合并，扩大了样本容量（增加了误差自由度，从而降低了误差方差）。