第九章 含定性变量的回归模型

§9.1 自变量中含有定性变量的回归模型

一、简单情况

首先讨论定性变量只取两类可能值的情况，例如研究粮食产量问题，y为粮食产量，x为施肥量，另外再考虑气候问题，分为正常年份和干旱年份两种情况，对这个问题的数量化方法是引入一个0-1型变量D，令：

$D_i=1$　　表示正常年份

$D_i=0$　　表示干旱年份

§9.1 自变量中含有定性变量的回归模型

粮食产量的回归模型为：

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \varepsilon_i$$

其中干旱年份的粮食平均产量为：

$$E(y_i \mid D_i = 0) = \beta_0 + \beta_1 x_i$$

正常年份的粮食平均产量为：

$$E(y_i \mid D_i = 1) = (\beta_0 + \beta_2) + \beta_1 x_i$$

§9.1 自变量中含有定性变量的回归模型

例9.1 某经济学家想调查文化程度对家庭储蓄的影响，在一个中等收入的样本框中，随机调查了13户高学历家庭与14户中低学历的家庭，

因变量y为上一年家庭储蓄增加额，

自变量x_1为上一年家庭总收入，

自变量x_2表示家庭学历，

高学历家庭$x_2=1$,低学历家庭$x_2=0$，

调查数据见表9.1：

§9.1 自变量中含有定性变量的回归模型

表9.1

序号	y（元）	x_1（万元）	x_2
1	235	2.3	0
2	346	3.2	1
3	365	2.8	0
4	468	3.5	1
5	658	2.6	0
6	867	3.2	1
7	1085	2.6	0
23	8950	3.9	0
24	9865	4.8	0
25	9866	4.6	0
26	10235	4.8	0
27	10140	4.2	0

§9.1 自变量中含有定性变量的回归模型

建立y对x1、x2的线性回归

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.938[a]	.879	.869	1288.68

a. Predictors: (Constant), X2, X1

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	290372875.924	2	145186437.962	87.425	.000
	Residual	39856639.705	24	1660693.321		
	Total	330229515.630	26			

§9.1 自变量中含有定性变量的回归模型

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-7976.809	1093.445		-7.295	.000
X1	3826.129	304.591	.921	12.562	.000
X2	-3700.330	513.445	-.529	-7.207	.000

两个自变量x_1与x_2的系数都是显著的，判定系数$R^2=0.879$，回归方程为：

$$\hat{y}=-7976+3826x_1-3700x_2$$

§9.1 自变量中含有定性变量的回归模型

这个结果表明，中等收入的家庭每增加1万元收入，平均拿出3826元作为储蓄。高学历家庭每年的平均储蓄额少于低学历的家庭，平均少3700元。

如果不引入家庭学历定性变量x_2，仅用y对家庭年收入x_1做一元线性回归，得判定系数$R^2=0.618$，拟合效果不好。

§9.1 自变量中含有定性变量的回归模型

家庭年收入x_1是连续型变量，它对回归的贡献也是不可缺少的。如果不考虑家庭年收入这个自变量，13户高学历家庭的平均年储蓄增加额为3009.31元，14户低学历家庭的平均年储蓄增加额为5059.36元，这样会认为高学历家庭每年的储蓄额比低学历的家庭平均少5059.36-3009.31=2050.05元，而用回归法算出的数值是3824元，两者并不相等。

§9.1 自变量中含有定性变量的回归模型

用回归法算出的高学历家庭每年的平均储蓄额比低学历的家庭平均少3824元，这是在假设两者的家庭年收入相等的基础上的储蓄差值，或者说是消除了家庭年收入的影响后的差值，因而反映了两者储蓄额的真实差异。而直接由样本计算的差值2050.05元是包含有家庭年收入影响在内的差值，是虚假的差值。所调查的13户高学历家庭的平均年收入额为3.8385万元，14户低学历家庭的平均年收入额为3.4071万元，两者并不相等。

§9.1 自变量中含有定性变量的回归模型

二、复杂情况

某些场合定性自变量可能取多类值，例如某商厦策划营销方案，需要考虑销售额的季节性影响，季节因素分为春、夏、秋、冬4种情况。为了用定性自变量反应春、夏、秋、冬四季，我们初步设想引入如下4个0-1自变量：

$$\begin{cases} x_1 = 1, & \text{春季} \\ x_1 = 0, & \text{其它} \end{cases} \qquad \begin{cases} x_2 = 1, & \text{夏 季} \\ x_2 = 0, & \text{其它} \end{cases}$$

$$\begin{cases} x_3 = 1, & \text{秋 季} \\ x_3 = 0, & \text{其它} \end{cases} \qquad \begin{cases} x_4 = 1, & \text{冬 季} \\ x_4 = 0, & \text{其它} \end{cases}$$

§9.1 自变量中含有定性变量的回归模型

可是这样做却产生了一个新的问题，即$x_1+x_2+x_3+x_4=1$，构成完全多重共线性。

解决这个问题的方法很简单，我们只需去掉一个0-1型变量，只保留3个0-1型自变量即可。例如去掉x_4，只保留x_1、x_2、x_3。

对一般情况，一个定性变量有k类可能的取值时，需要引入k-1个0-1型自变量。当k=2时，只需要引入一个0-1型自变量即可。