第二节　一元线性回归模型

一、模型的建立

二、参数估计

三、回归方程的显著性检验

回归分析就是要研究具有相关关系的变量之间的统计规律性。回归分析在工农业生产及科学研究中有着广泛的应用。在试验数据的处理、经验公式的寻找、产品的统计质量管理、市场预测、某些新标准的制订、自动控制中数学模型的建立、气象预报、地质勘探、医药卫生等许多领域中都经常需要用到回归分析。

在回归分析中，把变量分成两类，一类是因变量，常是实际问题中所关心的一些指标，常用 y 表示；另一类是自变量，是影响因变量取值的一些变量，常用 $x_1, x_2, \cdots, x_p$ 表示。

在回归分析中要研究的主要问题是:

1、确定 y 与 $x_1, x_2, \cdots, x_p$ 之间的线性表达式，即估计表达式中的未知参数，这种表达式称为回归方程;

2、对求得的回归方程的可信度进行检验;

3、判断自变量 $x_i (i = 1, 2, \cdots, p)$ 对 y 有无影响，剔除影响不大的变量;

4、利用所得的回归方程进行预测和控制。

一、模型的建立

最简单的回归分析是一元线性回归分析。为了弄清它解决什么问题，我们先看一个例子。

例10.2.1 根据专业知识，合金钢的强度 $y(kg/mm^2)$ 与钢材中碳的含量 $x(\%)$ 之间有着密切关系。为了冶炼出符合所要求强度的钢，常通过控制钢水中碳的含量来达到目的。为此需要了解 y 与 x 的关系。下表是合金钢的强度和碳含量这两个变量之间的12组观测数据。

表10.2.1 碳含量$x(\%)$与钢的强度$y(kg/mm^2)$ 的12组数据

序号	1	2	3	4	5	6	7	8	9	10	11	12
x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20	0.21
y	42.0	43.5	45.0	45.5	45.0	47.5	49.0	53.0	50.0	55.0	55.0	60.0

由于各种因素的影响，即使钢水中碳的含量相同，合金钢的强度也不完全相同，因而它们间的关系是一种非确定性的关系。为了探求它们之间的关系，我们先把 (x_i, y_i) 看成是平面直角坐标系中的一个点，画出散点图：

由散点图可以发现，这些点散布在一直线附近，但又不全在这条直线上，因此我们可以认为变量 y 与变量 x 之间近似存在线性关系：

$$y=\beta_0+\beta_1 x$$

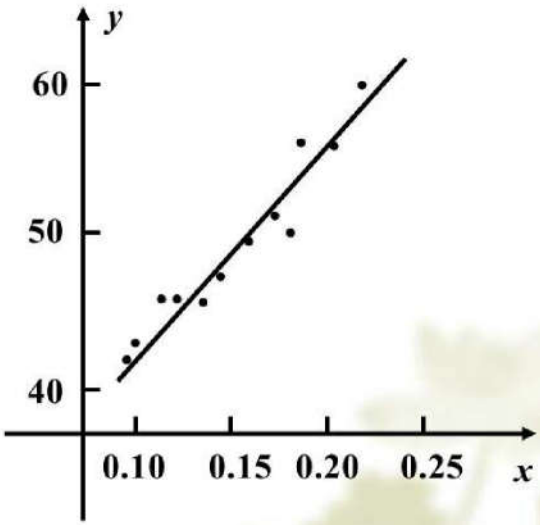

图10.2.1　　散点图

观察所画的散点图上点的分布规律，如果发现这些点散布在一直线附近，但又不全在一条直线上，那么我们可以认为变量 y 与变量 x 之间近似存在线性关系。这种关系由两部分组成，一部分是由于 x 的变化引起 y 的变化部分，记为：$\beta_0+\beta_1 x$，另一部分是由随机因素引起的误差，记为：ε，于是有 $y=\beta_0+\beta_1 x+\varepsilon$。现有数据对 (x_i,y_i)，$i=1,2,\cdots,n$，其中 $y_1,y_2,\cdots,y_n$ 相互独立，而诸 x_i 是一般变量，其值可以精确测量或严格控制，β_0，β_1 为未知参数，ε_i 是不可观测的随机误差。又设 $E(\varepsilon)=0$，$D(\varepsilon)=\sigma^2$，从而可得一元线性回归的数学模 型如下：

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ \ i=1,2,\cdots,n \\ E(\varepsilon_i)=0,\ \ D(\varepsilon_i)=\sigma^2,\ \ i=1,2,\cdots,n \end{cases} \tag{10.2.1}$$

若还有 $\varepsilon \sim N(0,\sigma^2)$，即有 $y \sim N(\beta_0+\beta_1 x,\sigma^2)$，则模型加强为：

$$\begin{cases} y_i=\beta_0+\beta_1 x_i+\varepsilon_i,\ i=1,2,\cdots,n \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n,\ i.i.d.\sim N(0,\sigma^2) \end{cases} \tag{10.2.2}$$

称 $E(y)=\beta_0+\beta_1 x$ 为变量 y 关于变量 x 的回归函数。它在平均意义上表明 y 与 x 之间存在一种统计规律性。

在这里要研究的问题是：

（1）如何根据样本 $(x_i,y_i),\ i=1,2,\cdots,n$ 求出 β_0，β_1 的点估计 $\hat{\beta}_0$，$\hat{\beta}_1$，则称 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$ 为 y 关于 x 的一元线性回归方程，这表明 $\hat{y}$ 是 x 的函数，有时也记 $\hat{y}$ 为 $\hat{y}(x)$，称为拟合值或预测值。

（2）如何检验回归方程的可信度？

二. 参数 β_0，β_1 的最小二乘估计

要求模型（9.2.1）中参数 β_0，β_1 的估计，就是求这样的估计 $\hat{\beta}_0$，$\hat{\beta}_1$，它使观测值 y 与拟合值 $\hat{y}$ 之间的偏差平方和达到最小。记：

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \qquad (10.2.3)$$

则所求的 $\hat{\beta}_0$，$\hat{\beta}_1$ 应满足下列要求：

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\min_{\beta_0,\beta_1} Q(\beta_0,\beta_1) \qquad (10.2.4)$$

满足(10.2.4)的 $\hat{\beta}_0$，$\hat{\beta}_1$ 称为 β_0，β_1 的最小二乘估计，简记为 *LSE*

对 $Q(\beta_0,\beta_1)$ 求关于 β_0，β_1 的偏导数并令其等于零，得

$$\begin{cases}\dfrac{\partial Q}{\partial \beta_0}\triangleq 0\\[2ex] \dfrac{\partial Q}{\partial \beta_1}\triangleq 0\end{cases}$$

整理得正规方程组：

$$\begin{cases}\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0\\[2ex] \sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)x_i=0\end{cases}$$

于是得

$$\begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ \bar{x}\beta_0 + \dfrac{1}{n}\sum_{i=1}^{n} x^2 \beta_1 = \dfrac{1}{n}\sum_{i=1}^{n} x_i y_i \end{cases}$$

整理得:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \triangleq \dfrac{l_{xy}}{l_{xx}} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

其中 $$l_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y},$$

$$l_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

由此可知，要求 β_0，β_1 的最小二乘估计，只要先计算

$$\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} y_i,\ \sum_{i=1}^{n} x_i^2,\ \sum_{i=1}^{n} x_i y_i,\ \sum_{i=1}^{n} y_i^2$$

然后可算出 $\bar{x}$，$\bar{y}$，l_{xy}，l_{xy}，从而可以算出 $\hat{\beta}_0$，$\hat{\beta}_1$

$$\hat{\beta}_1 \equiv \frac{l_{xy}}{l_{xx}} \equiv \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 \equiv \bar{y} - \hat{\beta}_1 \bar{x}$$

下面求例10.2.1中合金刚强度关于碳含量的一元线性回归方程组。

由表10.2.1 可得 $\sum_{i=1}^{n} x_i = 1.90,\ \sum_{i=1}^{n} y_i = 590.5,$

$$\sum_{i=1}^{n} x_i^2 = 0.3194,\ \sum_{i=1}^{n} x_i y_i = 95.9250,$$

据此得 $\bar{x} = 0.1583,\ \bar{y} = 49.2083,$

$$l_{xx} = 0.0186,\ l_{xy} = 2.4292,$$

于是求得估计值如下：$\hat{\beta}_1 = l_{xy} / l_{xx} = 130.6022,$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 28.5340,$$

从而求得一元线性回归方程组为：

$$\hat{y} = 28.5340 + 130.6022x \qquad (10.2.6)$$

为了后面讨论的需要，下面研究最小二乘估计的性质

性质1 在模型(10.2.1)下，有 $E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$

性质2 在模型（10.2.1）下，有 $D(\hat{\beta}_1) = \sigma^2 / l_{xx},$

$$D(\hat{\beta}_0) = (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2,\ Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}}{l_{xx}}\sigma^2$$

性质3 在模型（10.2.2）下，$\hat{\beta}_0 \sim N(\beta_0, (\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}})\sigma^2),$

$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{l_{xx}})$，且 $\bar{y}$ 与 $\hat{\beta}_1$ 相互独立。

性质4 在模型（10.2.2）下，$E(\hat{y}) = \beta_0 + \beta_1 x = E(y),$

$$D(\hat{y}) = [\frac{1}{n} + \frac{(x-\bar{x})^2}{l_{xx}}]\sigma^2,\quad \hat{y} \sim N(\beta_0 + \beta_1 x, [\frac{1}{n} + \frac{(x-\bar{x})^2}{l_{xx}}]\sigma^2)$$

三、回归方程的显著性检验

由一元线性回归方程系数的最小二乘估计公式可知，不管 y与 x之间是否有线性关系，只要给出 n对数据

$$(x_i, y_i),\ i = 1, 2, \cdots, n,$$

我们总能求出 $\hat{\beta}_0$，$\hat{\beta}_1$，从而得到回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$。

问题：这样得到的方程是否有实际意义？

在一元线性回归中，β_1反映 $E(y)$随 x变化程度的大小。

若 $\beta_1 \neq 0$，说明 $E(y)$随 x 的变化而变化，这时所给的方程才有实际意义。

若 $\beta_1 = 0$，说明 $E(y)$不随 x 的变化而变化，这时所给的方程没有实际意义。因此对回归方程要作显著性检验是：

$$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0$$

引起随机变量观测值 $y_1, y_2, \cdots, y_p$变化的原因有：

其一是由于 H_0 不真，从而 x 的变化引起 $E(y)$ 的变化；

其二是由于其它因素造成的随机误差所致。

为了分清究竟是什么原因引起随机变量观测值变化，因此需要把引起波动的上述两个原因从总的波动中分离出来。

考虑总偏差平方和 $SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=l_{yy}$,

把总偏差平方和分解:

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$=SS_e+SS_R$$

其中 $SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 称为回归平方和

$SS_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 称为残差平方和

下面求 SS_R, SS_e 的期望。

(1) 利用 $\hat{\beta}_1=l_{xy}/l_{xx}$, $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}$,

$$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\hat{\beta}_1^2 l_{xx}=l_{xy}^2/l_{xx}$$

由性质2知

$$E(SS_R)=E(\hat{\beta}_1^2 l_{xx})=\{D\hat{\beta}_1+[E\hat{\beta}_1]^2\}l_{xx}=\sigma^2+\beta_1^2 l_{xx}$$

上式表明，若 $\beta_1=0$, $E(SS_R)=\sigma^2$,

这时从平均意义上看， SS_R仅反映了随机误差引起差异;

$\beta_1\neq 0$, $E(SS_R)=\sigma^2+\beta_1^2 l_{xx}$,

这时 SS_R 还反映 $E(y)$ 随 x 变化所引起的差异。因此称之为回归平方和。

当 $\beta_1 = 0$ 时，由性质3，$\hat{\beta}_1 \sim N(\beta_1,\ \sigma^2 / l_{xx})$,

所以 $SS_R / \sigma^2 = \hat{\beta}_1^2 l_{xx} / \sigma^2 = [\hat{\beta}_1 / (\sigma / \sqrt{l_{xx}})]^2 \sim \chi^2(1)$

其自由度 $f_e = 1$。

由于 $y_1, y_2, \cdots, y_n \quad iid \sim N(\beta_0, \sigma^2)$,

$$SS_T / \sigma^2 = \sum (y_i - \bar{y})^2 / \sigma^2 \sim \chi^2(n-1)$$

其自由度为 $f_T = n - 1$。

又 $SS_e = SS_T - SS_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

其自由度为 $f_e = n - 2$。

$$E(SS_e) = E(SS_T) - E(SS_R) = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

由Cochran定理知，$SS_e / \sigma^2 \sim \chi^2(n-2)$,

且 SS_e 与 SS_R 相互独立。因此当 $\beta_1 = 0$,

$$E(SS_R) = \sigma^2, \qquad E(\frac{SS_e}{n-2}) = \sigma^2,$$

SS_e 只反映随机误差所引起的差异，故称为残差平方和。当

$$\beta_1 \neq 0, \qquad E(SS_R) > \sigma^2,$$

取检验统计量为：$F = \dfrac{SS_R}{SS_e / (n-2)} \sim F(1,\ n-2)$

对给定的 $\alpha > 0$，查表的上 α 分位数 $C = F_{1-\alpha}(1,\ n-2)$,

拒绝域是：$D = \{F_{值} > F_{1-\alpha}(1,\ n-2)\}$

对例10.2.1中的回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$，作显著性检验，得如下的方差分析表。

表10.2.2 方差分析表

来源	平　方　和	自由度	均方和MS	F比值
回归	$SS_R=l_{xy}^2/l_{xx}$ $=317.2589$	$f_R=1$	$MS_R=SS_R/f_R$ $=317.2589$	$F=\dfrac{SS_R/f_R}{SS_e/f_E}$ $=176.55$
残差	$SS_e=SS_T-SS_R$ $=17.9703$	$f_e=10$	$MS_e=SS_e/f_e$ $=1.7970$	
总和	$SS_T=335.2292$	$f_T=11$	$F_{0.05}(1,10)$ $=4.96<176.55$	

这表明在$\alpha=0.05$的水平上，方程（10.2.6）是有意义的。