第十章　回归分析

回归分析方法是数理统计中的常用方法之一，是处理多个变量之间相关关系的一种数学方法.

第一节　回归分析的概述

在客观世界中变量之间的关系有两类，一类是确定性关系，例如欧姆定律中电压 U 与电阻 R、电流 I 之间的关系为 U=IR，如果已知这三个变量中的任意两个，则另一个就可精确地求出.另一类是非确定性关系即所谓相关关系.例如，正常人的血压与年龄有一定的关系，一般来讲年龄大的人血压相对地高一些，但是年龄大小与血压高低之间的关系不能用一个确定的函数关系表达出来.又如施肥量与农作物产量之间的关系，树的高度与径粗之间的关系也是这样.另一方面，即便是具有确定关系的变量，由于试验误差的影响，其表现形式也具有某种程度的不确定性.

具有相关关系的变量之间虽然具有某种不确定性，但通过对它们的不断观察，可以探索出它们之间的统计规律，回归分析就是研究这种统计规律的一种数学方法.它主要解决以下几方面问题.

（1）从一组观察数据出发，确定这些变量之间的回归方程.

（2）对回归方程进行假设检验.

(3) 利用回归方程进行预测和控制.

回归方程最简单的也是最完善的一种情况，就是线性回归方程.许多实际问题，当自变量局限于一定范围时，可以满意地取这种模型作为真实模型的近似，其误差从实用的观点看无关紧要.因此，本章重点讨论有关线性回归的问题.现在有许多数学软件如 Matlab,SAS 等都有非常有效的线性回归方面的计算程序，使用者只要把数据按程序要求输入到计算机，就可很快得到所要的各种计算结果和相应的图形，用起来十分方便.

我们先考虑两个变量的情形.设随机变量 y 与普通变量 x 之间存在着某种相关关系.这里 x 是可以控制或可精确观察的变量，如在施肥量与产量的关系中，施肥量是能控制的，可以随意指定几个值 $x_1,x_2,\cdots,x_n$,故可将它看成普通变量，称为自变量，而产量 y 是随机变量，无法预先作出产量是多少的准确判断，称为因变量.本章只讨论这种情况.

由 x 可以在一定程度上决定 y，但由 x 的值不能准确地确定 y 的值.为了研究它们的这种关系，我们对(x,y)进行一系列观测，得到一个容量为 n 的样本(x 取一组不完全相同的值)：$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,其中 y_i 是 $x=x_i$ 处对随机变量 y 观察的结果.每对(x_i,y_i)在直角坐标系中对应一个点，把它们都标在平面直角坐标系中，称所得到的图为散点图.如图 10-1 所示.

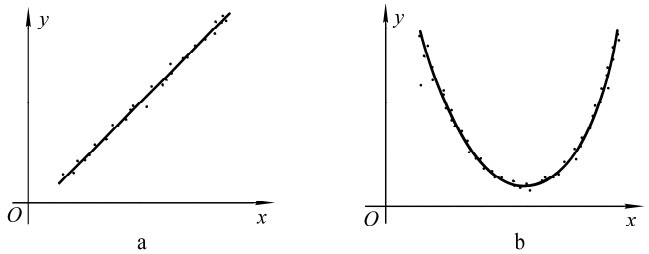


图 10-1

由图 10-1（a）可看出散点大致地围绕一条直线散布，而图 10-1（b）中的散点大致围

绕一条抛物线散布，这就是变量间统计规律性的一种表现.

如果图中的点像图 10-1（a）中那样呈直线状，则表明 y 与 x 之间有线性相关关系，我们可建立数学模型

$$y=a+bx+\varepsilon \qquad (10\text{-}1)$$

来描述它们之间的关系.因为 x 不能严格地确定 y，故带有一误差项 ε，假设 $\varepsilon\sim N(0,\sigma^2)$,相当于对 y 作这样的正态假设，对于 x 的每一个值有 $y\sim N(a+bx,\sigma^2)$,其中未知数 a,b,σ^2 不依赖于 x,(10.1)式称为一元线性回归模型(Univariable linear regression model).

在(10.1)式中，a,b,σ^2 是待估计参数.估计它们的最基本方法是最小二乘法，这将在下节讨论.记 $\hat{a}$ 和 $\hat{b}$ 是用最小二乘法获得的估计，则对于给定的 x，方程

$$\hat{y}=\hat{a}+\hat{b}x \qquad (10\text{-}2)$$

称为 y 关于 x 的线性回归方程或回归方程，其图形称为回归直线.(10-2)式是否真正描述了变量 y 与 x 客观存在的关系，还需进一步检验.

实际问题中，随机变量 y 有时与多个普通变量 $x_1,x_2,\cdots,x_p(p>1)$有关，可类似地建立数学模型

$$y=b_0+b_1x_1+\cdots+b_px_p+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2), \qquad (10\text{-}3)$$

其中 $b_0,b_1,\cdots,b_p,\sigma^2$ 都是与 $x_1,x_2,\cdots,x_p$ 无关的未知参数.(10-3)式称为多元线性回归模型，和前面一个自变量的情形一样，进行 n 次独立观测，得样本：

$$(x_{11},x_{12},\cdots,x_{1p},y_1),\cdots,(x_{n1},x_{n2},\cdots,x_{np},y_n)$$

有了这些数据之后，我们可用最小二乘法获得未知参数的最小二乘估计，记为 $\hat{b}_0,\hat{b}_1,\hat{b}_2,\cdots\hat{b}_n$，得多元线性回归方程

$$\hat{y}=\hat{b}_0+\hat{b}_1x_1+\cdots+\hat{b}_px_p \qquad (10\text{-}4)$$

同理，(10-4)式是否真正描述了变量 y 与 $x_1,x_2,\cdots,x_p$ 客观存在的关系，还需进一步检验.

第二节　参数估计

1. 一元线性回归

最小二乘法是估计未知参数的一种重要方法，现用它来求一元线性回归模型(10-1)式中 a 和 b 的估计.

最小二乘法的基本思想是：对一组观察值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,使误差 $\varepsilon_i=y_i-(a+bx_i)$的平方和

$$Q(a,b)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left[y_i-\left(a+bx_i\right)\right]^2 \qquad (10\text{-}5)$$

达到最小的 $\hat{a}$ 和 $\hat{b}$ 作为参数 a 和 b 的估计，称其为最小二乘估计(Least squares estimates).

直观地说，平面上直线很多，选取哪一条最佳呢？很自然的一个想法是，当点(x_i,y_i)，$i=1,2,\cdots,n$,与某条直线的偏差平方和比它们与任何其他直线的偏差平方和都要小时，这条直线便能最佳

地反映这些点的分布状况，并且可以证明，在某些假设下，$\hat{a}$ 和 $\hat{b}$ 是所有线性无偏估计中最好的.

根据微分学的极值原理，可将 $Q(a,b)$ 分别对 a,b 求偏导数，并令它们等于零，得到方程组：

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum\limits_{i=1}^{n}\left(y_i - a - bx_i\right) = 0, \\ \dfrac{\partial Q}{\partial b} = -2\sum\limits_{i=1}^{n}\left(y_i - a - bx_i\right)x_i = 0. \end{cases} \tag{10-6}$$

即

$$\begin{cases} na + \left(\sum\limits_{i=1}^{n} x_i\right)b = \sum\limits_{i=1}^{n} y_i, \\ \left(\sum\limits_{i=1}^{n} x_i\right)a + \left(\sum\limits_{i=1}^{n} x_i^{\,2}\right)b = \sum\limits_{i=1}^{n} x_i y_i. \end{cases} \tag{10−7}$$

(10-7)式称为正规方程组.

由于 x_i 不全相同，正规方程组的参数行列式

$$\begin{vmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^{\,2} \end{vmatrix} = n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n}(x_i - \overline{x})^2 \neq 0.$$

则(10-7)式有唯一解

$$\begin{cases} \hat{b} = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}, \\ \hat{a} = \hat{y} - \hat{b}\overline{x}. \end{cases} \tag{10-8}$$

于是，所求的线性回归方程为

$$\hat{y} = \hat{a} + \hat{b}x. \tag{10-9}$$

若将 $\hat{a} = \hat{y} - \hat{b}\overline{x}$ 代入上式，则线性回归方程亦可表示为

$$\hat{y} = \overline{y} + \hat{b}(x - \overline{x}). \tag{10-10}$$

(10-10)式表明，对于样本观察值$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,回归直线通过散点图的几何中心 $(\overline{x},\overline{y})$.回归直线是一条过点$(\overline{x},\overline{y})$,斜率为 $\hat{b}$ 的直线.

上述确定回归直线所依据的原则是使所有观测数据的偏差平方和达到最小值.按照这个原理确定回归直线的方法称为最小二乘法.“二乘”是指 Q 是二乘方（平方）的和.如果 y 是正态变量，也可用极大似然估计法得出相同的结果.

为了计算上的方便，引入下述记号：

$$\begin{cases} S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2, \\ S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^{\,2}-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2, \\ S_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right). \end{cases} \tag{10-11}$$

这样，*a*,*b* 的估计可写成：

$$\begin{cases} \hat{b}=\dfrac{S_{xy}}{S_{xx}}, \\ \hat{a}=\dfrac{1}{n}\sum_{i=1}^{n}y_i-\left(\dfrac{1}{n}\sum_{i=1}^{n}x_i\right)\hat{b}. \end{cases} \tag{10-12}$$

例 10. 1　某企业生产一种毛毯，1~10 月份的产量 x 与生产费用支出 *y* 的统计资料如表 10-1 所示.求 *y* 关于 *x* 的线性回归方程.

表 10-1

月份	1	2	3	4	5	6	7	8	9	10
x(千条)	12.0	8.0	11.5	13.0	15.0	14.0	8.5	10.5	11.5	13.3
y(万元)	11.6	8.5	11.4	12.2	13.0	13.2	8.9	10.5	11.3	12.0

解　为求线性回归方程，将有关计算结果列表如表 10-2 所示.

表 10-2

产量 *x*	费用支出 *y*	x^2	*xy*	y^2
12.0	11.6	144	139.2	134.56
8.0	8.5	64	68	72.25
11.5	11.4	132.25	131.1	129.96
13.0	12.2	169	158.6	148.84
15.0	13.0	225	195	169
14.0	13.2	196	184.8	174.24
8.5	8.9	72.25	75.65	79.21
10.5	10.5	110.25	110.25	110.25
11.5	11.3	132.25	129.95	127.69
13.3	12.0	176.89	159.6	144
Σ　117.3	112.6	1421.89	1352.15	1290

S_{xx}=1421.89-$\frac{1}{10}$ (117.3)^2=45.961,

S_{xy}=1352.15-$\frac{1}{10}$×117.3×112.6=31.352,

$\hat{b}=\frac{S_{xy}}{S_{xx}}$=0.6821,　$\hat{a}=\frac{112.6}{10}$-0.6821×$\frac{117.3}{10}$=3.2590,

故回归方程：$\hat{y}$ =3.2590+0.6821*x*.

2. 多元线性回归

多元线性回归(Multiple linear regression)分析原理与一元线性回归分析相同，但在计算上

要复杂些.

若$(x_{11},x_{12},\cdots,x_{1p},y_1),\cdots,(x_{n1},x_{n2},\cdots,x_{np},y_n)$为一样本，根据最小二乘法原理，多元线性回归中未知参数 $b_0,b_1,\cdots,b_p$ 应满足：

$$Q=\sum_{i=1}^{n}(y_i-b_0-b_1x_{i1}-\cdots-b_px_{ip})^2$$

达到最小.

对 Q 分别关于 $b_0,b_1,\cdots,b_p$ 求偏导数，并令它们等于零，得

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{i1}-\cdots-b_px_{ip})=0,\\ \dfrac{\partial Q}{\partial b_j}=-2\sum\limits_{i=1}^{n}(y_i-b_0-b_1x_{i1}-\cdots-b_px_{ip})x_{ij}=0,\ j=1,2,\cdots,p.\end{cases}$$

即

$$\begin{cases}b_0n+b_1\sum\limits_{i=1}^{n}x_{i1}+b_2\sum\limits_{i=1}^{n}x_{i2}+\cdots+b_p\sum\limits_{i=1}^{n}x_{ip}=\sum\limits_{i=1}^{n}y_i,\\ b_0\sum\limits_{i=1}^{n}x_{i1}+b_1\sum\limits_{i=1}^{n}x_{i1}^{\ 2}+b_2\sum\limits_{i=1}^{n}x_{i1}x_{i2}+\cdots+b_p\sum\limits_{i=1}^{n}x_{i1}x_{ip}=\sum\limits_{i=1}^{n}x_{i1}y_i,\\ \cdots\cdots\\ b_0\sum\limits_{i=1}^{n}x_{ip}+b_1\sum\limits_{i=1}^{n}x_{i1}x_{ip}+b_2\sum\limits_{i=1}^{n}x_{i2}x_{ip}+\cdots+b_p\sum\limits_{i=1}^{n}x_{ip}^{\ 2}=\sum\limits_{i=1}^{n}x_{ip}y_i.\end{cases}\qquad(10\text{-}13)$$

(10.13)式称为正规方程组，引入矩阵

$$\boldsymbol{X}=\begin{pmatrix}1&x_{11}&x_{12}&\cdots&x_{1p}\\1&x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\vdots&&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{np}\end{pmatrix},\quad \boldsymbol{Y}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad \boldsymbol{B}=\begin{pmatrix}b_0\\b_1\\\vdots\\b_p\end{pmatrix},$$

于是(10-13)式可写成

$$\boldsymbol{X}'\ \boldsymbol{X}\boldsymbol{B}=\boldsymbol{X}'\ \boldsymbol{Y}.\qquad(10\text{-}13)'$$

（10−13）′ 式为正规方程组的矩阵形式.若$(\boldsymbol{X}'\ \boldsymbol{X})^{-1}$ 存在,则

$$\hat{\boldsymbol{B}}=\begin{pmatrix}\hat{b}_0\\\hat{b}_1\\\vdots\\\hat{b}_p\end{pmatrix}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}.\qquad(10\text{-}14)$$

方程 $\hat{y}=\hat{b}_0+\hat{b}_1x_1+\cdots+\hat{b}_px_p$ 为 p 元线性回归方程.

例 10.2　如表 10-3 所示， x 和 z 表示某一种特定的合金中所含的 A 及 B 两种元素的百分数，现 x 及 z 各选 4 种，共有 4×4=16 种不同组合，y 表示各种不同成分的铸品数，根据表中资料求二元线性回归方程.

表 10-3

所含 Ax	5	5	5	5	10	10	10	10	15	15	15	15	20	20	20	20

所含 Bz	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
铸品数 y	28	30	48	74	29	50	57	42	20	24	31	47	9	18	22	31

解 由(10-13)式，根据表中数据，得正规方程组

$$\begin{cases} 16b_0+200b_1+40b_2=560, \\ 200b_0+3000b_1+500b_2=6110, \\ 40b_0+500b_1+120b_2=1580. \end{cases}$$

解 得：$\hat{b}_0$=34.75, $\hat{b}_1$=-1.78, $\hat{b}_2$=9. 于是所求回归方程为：

$$y=34.75-1.78x+9z.$$

第三节 假设检验

从上述求回归直线的过程看，用最小二乘法求回归直线并不需要对 y 与 x *一定具有线性相关关系，对*任何一组试验数据$(x_i,y_i)(i=1,2,\cdots,n)$都可用最小二乘法形式地求出一条 y 关于 x 的回归直线.若 y 与 x 间不存在某种线性相关关系，那么这种直线是没有意义的，这就需要对 y 与 x 的线性回归方程进行假设检验，即检验 x 的变化对变量 y 的影响是否显著.这个问题可利用线性相关的显著性检验来解决.

因为当且仅当 $b\neq 0$ 时，变量 y 与 x 之间存在线性相关关系.因此我们需要检验假设：

$$H_0:b=0;H_1:b\neq 0 \qquad (10.15)$$

若拒绝 H_0，则认为 y 与 x 之间存在线性关系，所求得的线性回归方程有意义；若接受 H_0，则认为 y 与 x 的关系不能用一元线性回归模型来表示，所求得的线性回归方程无意义.

关于上述假设的检验，我们介绍 3 种常用的检验法.

1. 方差分析法（*F* 检验法）

当 x 取值 $x_1,x_2,\cdots,x_n$ 时，得 y 的一组观测值 $y_1,y_2,\cdots,y_n$,记

$$Q_{总}=S_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

称为 $y_1,y_2,\cdots,y_n$ 的总偏差平方和(Total sum of squares)，它的大小反映了观测值 $y_1,y_2,\cdots,y_n$ 的分散程度.对 $Q_{总}$进行分析：

$$Q_{总}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2$$

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

$$=Q_{剩}+Q_{回}, \qquad (10\text{-}16)$$

其中

$$Q_{剩}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2,$$

$$Q_{回}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}\left[(\hat{a}+\hat{b}x_i)-(\hat{a}+\hat{b}\bar{x})\right]^2=\hat{b}^2\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

$Q_{剩}$称为剩余平方和(Residual sum of squares)，它反映了观测值 y_i 偏离回归直线的程度，这种偏离是由试验误差及其他未加控制的因素引起的.可证明 $\hat{\sigma}^2=\dfrac{Q_{剩}}{n-2}$ 是 σ^2 的无偏估计.

$Q_{回}$为回归平方和(Regression sum of squares)，它反映了回归值 $\hat{y}_i$ (i=1,2,⋯,n)的分散程度，它的分散性是因 x 的变化而引起的.并通过 x 对 y 的线性影响反映出来.因此 $\hat{y}_1,\hat{y}_2,\ldots,\hat{y}_n$ 的分散性来源于 $x_1,x_2,\cdots,x_n$ 的分散性.

通过对 $Q_{剩}$、$Q_{回}$的分析，$y_1,y_2,\cdots,y_n$ 的分散程度 $Q_{总}$的两种影响可以从数量上区分开来. $Q_{剩}$较小时，偏离回归直线的程度小.$Q_{回}$较大时，分散程度大.因而 $Q_{回}$与 $Q_{剩}$的比值反映了这种线性相关关系与随机因素对 y 的影响的大小；比值越大，线性相关性越强.

可证明统计量

$$F=\frac{Q_{回}}{1}\Big/\frac{Q_{剩}}{n-2}\overset{H_0真}{\sim}F(1,n\text{-}2)\qquad(10\text{-}17)$$

给定显著性水平 α,若 $F\geqslant F_{\alpha}$,则拒绝假设 H_0，即认为在显著性水平 α 下，y 对 x 的线性相关关系是显著的.反之，则认为 y 对 x 没有线性相关关系，即所求线性回归方程无实际意义.

检验时，可使用方差分析表 10-4.

表 10-4

方差来源	平方和	自由度	均方	F 比
回归 剩余	$Q_{回}$ $Q_{剩}$	1 n-2	$Q_{回}$/1 $Q_{剩}$/(n-2)	$F=\dfrac{Q_{回}}{Q_{剩}/(n-2)}$
总计	$Q_{总}$	n-1		

其中：

$$\begin{cases}Q=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{b}^2S_{xx}{}^2=S_{xy}{}^2/S_{xx},\\ Q_{剩}=Q_{总}-Q_{回}=S_{yy}-S_{xy}{}^2/S_{xx}.\end{cases}\qquad(10\text{-}18)$$

例 10.3　在显著性水平 α=0.05,检验例 10. 1 中的回归效果是否显著？

解　由例 10.1 知，

$$n=10，S_{xx}=45.961,\quad S_{xy}=31.352,$$
$$S_{yy}=22.124,\quad Q_{回}=S_{xy}{}^2/S_{xx}=21.3866,$$
$$Q_{剩}=Q_{总}-Q_{回}=22.124-21.3866=0.7374,$$

$$F=Q_{回}\Big/\frac{Q_{剩}}{n-2}=232.0217>F_{0.05}(1,8)=5.32.$$

故拒绝 H_0，即两变量的线性相关关系是显著的.

2. 相关系数法（t 检验法）

为了检验线性回归直线是否显著，还可用 x 与 y 之间的相关系数来检验.相关系数的定

义是：

$$r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}. \qquad (10\text{-}19)$$

由于

$$Q_{回}/Q_{总}=\frac{S_{xy}^{\ 2}}{S_{xx}S_{yy}}=r^2,\quad |r|\leqslant 1,\quad \hat{b}=\frac{S_{xy}}{S_{xx}},$$

因此

$$r=\frac{\hat{b}S_{xx}}{\sqrt{S_{xx}S_{yy}}}.$$

显然 r 和 $\hat{b}$ 的符号是一致的，它的值反映了 x 和 y 的内在联系.

提出检验假设：　　H_0: $r=0$;　　H_1: $r\neq 0$. 　　(10-20)

可以证明，当 H_0 为真时，

$$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\ \sim t(n\text{-}2). \qquad (10\text{-}21)$$

故 H_0 的拒绝域为

$$t\geqslant t_{\alpha/2}(n\text{-}2) \qquad (10\text{-}22)$$

由上例的数据可算出

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=0.9832,$$

$$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}=15.2352>t_{0.025}(8)=2.3060.$$

故拒绝 H_0，即两变量的线性相关性显著.

在一元线性回归预测中，相关系数检验与 F 检验法等价，在实际中只需作其中一种检验即可.

与一元线性回归显著性检验原理相同，为考察多元线性回归这一假定是否符合实际观察结果，还需进行以下假设检验：

$$H_0: b_1=b_2=\cdots=b_p=0;\quad H_1: b_i \text{ 不全为零}.$$

可以证明统计量

$$F=\frac{U}{p}\Big/\frac{Q}{n-p-1}\overset{H_0\text{真}}{\sim} F(p,n\text{-}p\text{-}1). \qquad (10\text{-}23)$$

其中　$U=\boldsymbol{Y}'\ \boldsymbol{X}(\boldsymbol{X}'\ \boldsymbol{X})^{-1}\boldsymbol{X}'\ \boldsymbol{Y}\text{-}n\,\hat{y}^2$,　　$Q=\boldsymbol{Y}'\ \boldsymbol{Y}\text{-}\boldsymbol{Y}'\ \boldsymbol{X}(\boldsymbol{X}'\ \boldsymbol{X})^{-1}\boldsymbol{X}'\ \boldsymbol{Y}$.

给定显著性水平 α，若 $F\geqslant F_\alpha$，则拒绝 H_0,即认为回归效果是显著的.

第四节　预测与控制

1. 预测

由于 x 与 y 并非确定性关系，因此对于任意给定的 $x=x_0$，无法精确知道相应的 y_0 值，但可由回归方程计算出一个回归值 $\hat{y}=\hat{a}+\hat{b}x_0$,可以以一定的置信度预测对应的 y 的观察值的取值范围，也即对 y_0 作区间估计，即对于给定的置信度 1- α,求出 y_0 的置信区间[称为预测区间(Prediction interval)]，这就是所谓的预测问题.

对于给定的置信度 1- α,可证明 y_0 的 1- α 预测区间为

$$\left(\hat{y}_0 \pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\right). \tag{10-24}$$

给定样本观察值，作出曲线

$$\begin{cases} y_1(x)=\hat{y}(x)-t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{xx}}}, \\ y_2(x)=\hat{y}(x)+t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{xx}}}. \end{cases} \tag{10-25}$$

这两条曲线形成包含回归直线 $\hat{y}=\hat{a}+\hat{b}$ x 的带形域，如图 10-2 所示，这一带形域在 $x=\overline{x}$ 处最窄，说明越靠近，预测就越精确.而当 x_0 远离 $\overline{x}$ 时，置信区域逐渐加宽，此时精度逐渐下降.

在实际的回归问题中，若样本容量 n 很大，在 $\overline{x}$ 附近的 x 可得到较短的预测区间，又可简化计算

$$\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}\approx 1,$$

$$t_{\frac{\alpha}{2}}(n-2)\approx z_{\frac{\alpha}{2}},$$

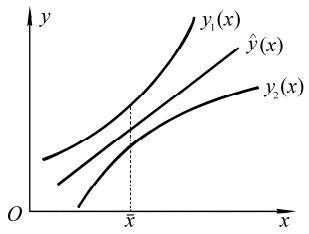


图 10-2

故 y_0 的置信度为 1- α 的预测区间近似地等于

$$\left(\hat{y}-\hat{\sigma}z_{\frac{\alpha}{2}},\hat{y}+\hat{\sigma}z_{\frac{\alpha}{2}}\right). \tag{10-26}$$

特别地，取 1- α=0.95,y_0 的置信度为 0.95 的预测区间为

$$\left(\hat{y}_0-1.96\hat{\sigma},\hat{y}_0+1.96\hat{\sigma}\right)$$

取 1- α=0.997,y_0 的置信度为 0.997 的预测区间为

$$\left(\hat{y}_0-2.97\hat{\sigma},\hat{y}_0+2.97\hat{\sigma}\right)$$

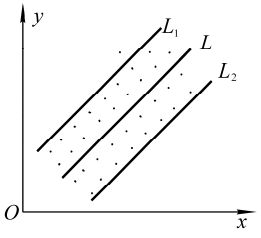


图 10-3

可以预料，在全部可能出现的 y 值中，大约有 99.7%的观测点落在直线 L_1:$y=\hat{a}-2.97\hat{\sigma}+\hat{b}x$ 与直线 L_2:$y=\hat{a}+2.97\hat{\sigma}+\hat{b}x$ 所夹的带形区域内.如图 10-3 所示.

可见，预测区间意义与置信区间的意义相似，只是后者对未知参数而言，前者是对随机变量而言.

例 10.4 给定 $\alpha=0.05$,$x_0=13.5$,问例 10. 1 中生产费用将会在什么范围？

解 当 $x_0=13.5$,y_0 的预测值为：

$$\hat{y}_0=3.2590+0.6821\times 13.5=12.4674$$

给定 $\alpha=0.05$,$t_{0.025}(8)=2.306$，

$$\hat{\sigma}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{0.7374}{8}}=0.3036,$$

$$\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}=\sqrt{1+\frac{1}{10}+\frac{(13.5-11.73)^2}{45.961}}=1.0808,$$

故

$$t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}=2.306\times 0.3036\times 1.0808=0.7567.$$

即 y_0 将以 95%的概率落在(12.4674±0.7567)区间，即预报生产费用在(11.7107,13.2241)万元之间.

2. 控制

控制实际上是预测的反问题，即要求观察值 y 在一定范围内 $y_1<y<y_2$ 内取值，应考虑把自变量 x 控制在什么范围，即对于给定的置信度 1-α,求出相应的 x_1,x_2,使 $x_1<x<x_2$ 时，x 所对应的观察值 y 落在(y_1',y_2')之内的概率不小于 1-α.

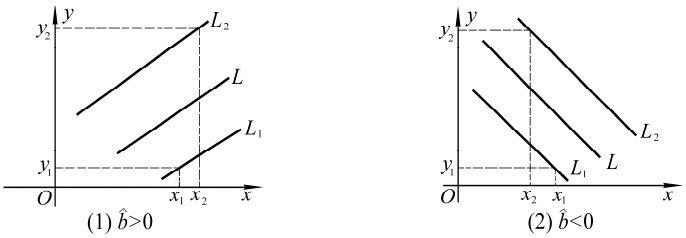
当 n 很大时，从方程

$$\begin{cases} y_1=\hat{y}-\hat{\sigma}z_{\frac{\alpha}{2}}=\hat{a}+\hat{b}x-\hat{\sigma}z_{\frac{\alpha}{2}}, \\ y_2=\hat{y}+\hat{\sigma}z_{\frac{\alpha}{2}}=\hat{a}+\hat{b}x+\hat{\sigma}z_{\frac{\alpha}{2}}. \end{cases} \tag{10-27}$$

分别解出 x 来作为控制 x 的上、下限：

$$\begin{cases} x_1=(y_1-\hat{a}+\hat{\sigma}z_{\frac{\alpha}{2}})\Big/\hat{b}, \\ x_2=(y_2-\hat{a}-\hat{\sigma}z_{\frac{\alpha}{2}})\Big/\hat{b}. \end{cases} \tag{10-28}$$

当 $\hat{b}>0$ 时，控制区间为(x_1,x_2)；如图 10-4（a）所示，当 $\hat{b}<0$ 时，控制区间为（x_2,x_1）.如图 10-4（b）所示,



（a） （b）

图 10-4

注意，为了实现控制，我们必须使区间(y_1,y_2)的长度不小于 $2z_{\frac{\alpha}{2}}\hat{\sigma}$,即：

$$y_2-y_1>2\hat{\sigma}\ z_{\frac{\alpha}{2}}.$$

第五节　非线性回归的线性化处理

前面讨论了线性回归问题，对线性情形我们有了一整套的理论与方法.在实际中常会遇见更为复杂的非线性回归问题，此时一般是采用变量代换法将非线性模型线性化，再按照线性回归方法进行处理.举例如下：

模型　　$y=a+b\sin t+\varepsilon$，　$\varepsilon\sim N(0,\sigma^2)$,　　(10-29)

其中 a,b,σ^2 为与 t 无关的未知参数，只要令 $x=\sin t$,即可将(10-29)化为(10-1).

模型　　$y=a+bt+ct^2+\varepsilon$，　$\varepsilon\sim N(0,\sigma^2)$,　　(10-30)

其中 a,b,c,σ^2 为与 t 无关的未知参数.令 $x_1=t,x_2=t^2$,得

$$y=a+bx_1+cx_2+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2),\qquad (10\text{-}31)$$

它为多元线性回归的情形.

模型　　$\dfrac{1}{y}=a+b/x+\varepsilon$，　$\varepsilon\sim N(0,\sigma^2)$,

令 $y'=\dfrac{1}{y}$，　$x'=\dfrac{1}{x}$,则有　　$y'=a+bx'+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$,

化为(10-1)式.

模型　　$y=a+b\ln x+\varepsilon$，　$\varepsilon\sim N(0,\sigma^2)$,

令 $x'=\ln x$,则有　　$y=a+bx'+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$,

可化(10-1)式.

另外，还有下述模型　　$Q(y)=a+bx+\varepsilon,\ \varepsilon\sim N(0,\sigma^2)$,

其中 Q 为已知函数，且设 $Q(y)$存在单值的反函数，a,b,σ^2 为与 x 无关的未知参数.这时，令 $z=Q(y)$,得

$$z=a+bx+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$$

在求得 z 的回归方程和预测区间后，再按 $z=Q(y)$的逆变换，变回原变量 y.我们就分别称它们为关于 y 的回归方程和预测区间.此时 y 的回归方程的图形是曲线，故又称为曲线回归方程.

例 10.5　某钢厂出钢时所用的盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断扩大.通过试验，得到了使用次数 x 和钢包增大的容积 y 之间的 17 组数据如表 10-5，求使用

次数 x 与增大容积 y 的回归方程.

表 10-5

x	y	x	y
2	6.42	11	10.59
3	8.20	12	10.60
4	9.58	13	10.80
5	9.50	14	10.60
6	9.70	15	10.90
7	10.00	16	10.76
8	9.93	18	11.00
9	9.99	19	11.20
10	10.49		

解 散点图如图 10-5 所示.

看起来 y 与 x 呈倒指数关系 $\ln y=a+b\frac{1}{x}+\varepsilon$，记 $y'=\ln y, x'=\frac{1}{x}$,求出 x' ,y' 的值（表 10-6）.

表 10-6

x'	y'	x'	y'
0.5000	1.8594	0.0909	2.3599
0.3333	2.1041	0.0833	2.3609
0.2500	2.2597	0.0769	2.3795
0.2000	2.2513	0.0714	2.3609
0.1667	2.2721	0.0667	2.3888
0.1429	2.3026	0.0625	2.3758
0.1250	2.2956	0.0556	2.3979
0.1111	2.3016	0.0526	2.4159
0.1000	2.3504		

作(x' ,y')的散点图，如图 10-6 所示.

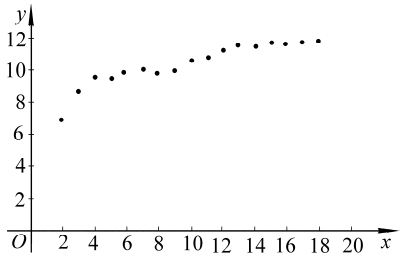


图 10-5

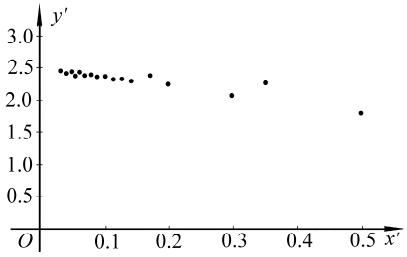


图 10-6

可见各点基本上在一直线上，故可设

$$y'=a+bx'+\varepsilon,\ \varepsilon\sim(0,\sigma^2),$$

经计算，得

$$\overline{x'}=0.1464,\ \overline{y'}=2.2963,$$

$$\sum_{i=1}^{n}(x_i')^2=0.5902,$$

$$\sum_{i=1}^{n}(y_i')^2 = 89.9311,$$

$$\sum_{i=1}^{n}x_i'y_i' = 5.4627.$$

$$\hat{b} = -1.1183, \quad \hat{a} = 2.4600.$$

于是 x' 对于 y' 的线性回归方程为

$$y' = -1.1183x' + 2.4600,$$

换回原变量得

$$\hat{y} = 11.7046\mathbf{e}^{-\frac{1.1183}{x}}.$$

现对 x' 与 y' 的线性相关关系的显著性用 F 检验法进行检验，得

$$F(1，16) = 379.3115 > F_{0.01}(1，16) = 8.53.$$

检验结论表明，此线性回归方程的效果是显著的.

小　结

本章介绍了在实际中应用非常广泛的数理统计方法之一——回归分析，并对线性回归作了参数估计、相关性检验、预测与控制及非线性回归的线性化处理.

1. 一元线性回归模型 $y=a+bx+\varepsilon$ 的最小二乘估计为

$$\hat{b} = \frac{S_{xy}}{S_{xx}}, \hat{a} = \bar{y} - \bar{x}\hat{b}.$$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i$，$S_{xx} = \sum_{i=1}^{n}x_i^2 - n\bar{x}^2$，

$$S_{xy} = \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}, \quad S_{yy} = \sum_{i=1}^{n}y_i^2 - n\bar{y}^2.$$

2. 变量 y 与 x 的线性相关性假设检验有下列所述：

(1) 方差分析法（*F* 检验法）

$$H_0: b=0; \quad H_1: b\neq 0.$$

$$F = Q_{回}\Big/\frac{Q_{剩}}{n-2} \overset{H_0真}{\sim} F_{\alpha}(1,n-2).$$

其中

$$Q_{回} = S_{xy}^2/S_{xx}, \quad Q_{剩} = Q_{总} - Q_{回} = S_{yy} - S_{xy}^2/S_{xx}.$$

给定显著性水平 α，若 $F \geqslant F_{\alpha}$，则拒绝 H_0,即认为 y 对 x 具有线性相关关系.

(2) 相关系数法（*t* 检验法）

$$H_0: r=0; \quad H_1: r\neq 0.$$

其中

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},\qquad t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\overset{H_0\text{真}}{\sim}t_{\frac{\alpha}{2}}(n-2).$$

若 $t\geqslant t_{\frac{\alpha}{2}}(n-2)$ 则拒绝 H_0.即认为两变量的线性相关性显著.

3. 给定 $x=x_0$ 时，y_0 的置信水平为 1- α 的预测区间为

$$\left(\hat{a}+\hat{b}x_0\pm t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right).$$

重要术语及主题

线性回归，最小二乘估计，预测与控制，非线性回归.

习 题 十

1.在硝酸钠($NaNO_3$)的溶解度试验中，测得在不同温度 x(℃)下，溶解于 100 份水中的硝酸钠份数 y 的数据如表 10-7 所示，试求 y 关于 x 的线性回归方程.

表 10-7

x_i	0	4	10	15	21	29	36	51	68
y_i	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

2.测量了 9 对父子的身高（单位：in，1 in=2.54 cm），所得数据如表 10-8 所示.

表 10-8

父亲身高 x_i	60	62	64	66	67	68	70	72	74
儿子身高 y_i	63.6	65.2	66	66.9	67.1	67.4	68.3	70.1	70

（1） 求儿子身高 y 关于父亲身高 x 的回归方程；

（2） 取 α=0.05，检验儿子的身高 y 与父亲身高 x 之间的线性相关关系是否显著；

（3） 若父亲身高 70 in，求其儿子的身高的置信度为 95%的预测区间.

3. 随机抽取了 10 个家庭，调查了他们的家庭月收入 x（$\times10^2$ 元）和月支出 y($\times10^2$ 元)，记录于下表 10-9 中：

表 10-9

x	20	15	20	25	16	20	18	19	22	16
y	18	14	17	20	14	19	17	18	20	13

求：（1） 在直角坐标系下作 x 与 y 的散点图，判断 y 与 x 是否存在线性关系；

（2） 求 y 关于 x 的一元线性回归方程；

（3） 对所得的回归方程作显著性检验.（取 α=0.025）

4.设 y 为树干的体积，x_1 为离地面一定高度的树干直径，x_2 为树干高度，一共测量了 31 棵树，数据列于表 10-10，作出 y 对 x_1,x_2 的二元线性回归方程，以便能用简单分法从 x_1 和 x_2 估计一棵树的体积，进而估计一片森林的木材储量.

表 10-10

x_1(直径)	x_2(高)	y(体积)	x_1(直径)	x_2(高)	y(体积)
8.3	70	10.3	12.9	85	33.8

8.6	65	10.3	13.3	86	27.4
8.8	63	10.2	13.7	71	25.7
10.5	72	10.4	13.8	64	24.9
10.7	81	16.8	14.0	78	34.5
10.8	83	18.8	14.2	80	31.7
11.0	66	19.7	15.5	74	36.3
11.0	75	15.6	16.0	72	38.3
11.1	80	18.2	16.3	77	42.6
11.2	75	22.6	17.3	81	55.4
11.3	79	19.9	17.5	82	55.7
11.4	76	24.2	17.9	80	58.3
11.4	76	21.0	18.0	80	51.5
11.7	69	21.4	18.0	80	51.0
12.0	75	21.3	20.6	87	77.0
12.9	74	19.1			

5. 一家从事市场研究的公司，希望能预测每日出版的报纸在各种不同居民区内的周末发行量，两个独立变量，即总零售额和人口密度被选作自变量.由 n=25 个居民区组成的随机样本所给出的结果列于表 10-11 中，求日报周末发行量 y 关于总零售额 x_1 和人口密度 x_2 的线性回归方程.

居民区	日报周末发行量 y_i ($\times 10^4$ 份)	总零售额 x_{i1} (10^5 元)	人口密度 x_{i2} ($\times$0.001m2)
1	3.0	21.7	47.8
2	3.3	24.1	51.3
3	4.7	37.4	76.8
4	3.9	29.4	66.2
5	3.2	22.6	51.9
6	4.1	32.0	65.3
7	3.6	26.4	57.4
8	4.3	31.6	66.8
9	4.7	35.5	76.4
10	3.5	25.1	53.0
11	4.0	30.8	66.9
12	3.5	25.8	55.9
13	4.0	30.3	66.5
14	3.0	22.2	45.3
15	4.5	35.7	73.6
16	4.1	30.9	65.1
17	4.8	35.5	75.2
18	3.4	24.2	54.6
19	4.3	33.4	68.7
20	4.0	30.0	64.8
21	4.6	35.1	74.7
22	3.9	29.4	62.7

23	4.3	32.5	67.6
24	3.1	24.0	51.3
25	4.4	33.9	70.8

6. 一种合金在某种添加剂的不同浓度之下，各做 3 次试验，得数据如表 10-12 所示：

浓度 x	10.0	15.0	20.0	25.0	30.0
抗压强度 y	25.2	29.8	31.2	31.7	29.4
	27.3	31.1	32.6	30.1	30.8
	28.7	27.8	29.7	32.3	32.8

(1) 作散点图.

(2) 以模型 $y=b_0+b_1x_1+b_2x_2+\varepsilon$, $\varepsilon\sim N(0,\sigma^2)$拟合数据，其中 b_0,b_1,b_2,σ^2 与 x 无关，求回归方程 $\hat{y}=\hat{b}_0+\hat{b}_1x+\hat{b}_2x^2$.