第六章　回归分析

前几章所讨论的内容，其目的在于寻求被测量的最佳值及其精度。在生产和科学实验中，还有另一类问题，即测量与数据处理的目的并不在于获得被测量的估计值，而是为了寻求两个变量或多个变量之间的内在关系，这就是本章所要解决的主要问题。

表达变量之间关系的方法有散点图、表格、曲线、数学表达式等，其中数学表达式能较客观地反映事物的内在规律性，形式紧凑，且便于从理论上作进一步分析研究，对认识自然界量与量之间关系有着重要意义。而数学表达式的获得是通过回归分析方法完成的。

第一节　回归分析的基本概念

一、函数与相关

在生产和科学实验中，人们常遇到各种变量。从辩证唯物主义观点来看，这些变量之间是相互联系、相互依存的，它们之间存在着一定的关系。人们通过实践，发现变量之间的关系可分为两种类型：

1. 函数关系（即确定性关系）

数学分析和物理学中的大多数公式属于这种类型。例如以速度 v 作匀速运动的物体，走过的距离 s 与时间 t 之间，有如下确定的函数关系：

$$s = vt$$

若上式中的变量有两个已知，则另一个就可由函数关系精确地求出。

2. 相关关系

在实际问题中，绝大多数情况下变量之间的关系不那么简单。例如，在车床上加工零件，零件的加工误差与零件的直径之间有一定的关系，知道了零件直径可大致估计其加工误差，但又不能精确地预知加工误差。这是由于零件在加工过程中影响加工误差的因素很多，如毛坯的裕量、材料性能、背吃刀量、进给量、切削速度、零件长度等，相互构成一个很复杂的关系，加工误差并不由零件直径这一因素所确定。像这种关系，在实践中是大量存在的，如材料的抗拉强度与其硬度之间；螺纹零件中螺纹的作用中径与螺纹中径之间；齿轮各种综合误差与有关单项误差之间；某些光学仪器、电子仪器等开机后仪器的读数变化与时间之间；材料的性能与其化学成分之间等。这些变量之间既存在着密切的关系，又不能由一个（或几个）变量（自变量）的数值精确地求出另一个变量（因变量）的数值，而是要通过实验和调查研究，才能确定它们之间的关系，我们称这类变量之间的关系为相关关系。一般讲，多考虑一些变量会减少所考察的因变量的不确定性，但不是绝对的。

应该指出，函数和相关关系虽然是两种不同类型的变量关系，但是它们之间并无严格的界限。一方面由于测量误差等原因，确定性的关系在实际中往往通过相关关系表现出来。例如尽管从理论上物体运动的速度、时间和运动距离之间存在着函数关系，但如果我们做多次反复地实测，每次测得的数值并不一定满足 $s = vt$ 的关系。在实践中，为确定某种函数关系

中的常数，往往也是通过实验。另一方面，当对事物内部的规律性了解得更加深刻的时候，相关关系又能转化为确定性关系。事实上，实验科学（包括物理学）中的许多确定性的定理正是通过对大量实验数据的分析和处理，经过总结和提高，从感性到理性，最后才能得到更能深刻地反映变量之间关系的客观规律。

二、回归分析的主要内容

回归分析（Regression Analysis）是英国生物学家兼统计学家高尔顿（Galton）在1889年出版的《自然遗传》一书中首先提出的，是处理变量之间相关关系的一种数理统计方法。上面已经提到，由于相关变量之间不存在确定性关系，因此，在生产实践和科学实验所记录的这些变量的数据中，存在着不同程度的差异。回归分析就是应用数学的方法，对大量的观测数据进行处理，从而得出比较符合事物内部规律的数学表达式。概括地说，本章主要解决以下几方面的问题：

1）从一组数据出发，确定这些变量之间的数学表达式——回归方程或经验公式。

2）对回归方程的可信程度进行统计检验。

3）进行因素分析，例如从对共同影响一个变量的许多变量（因素）中，找出哪些是重要因素，哪些是次要因素。

回归分析是数理统计中的一个重要分支，在工农业生产和科学研究中有着广泛的应用。当今在实验数据处理、经验公式的求得、因素分析、仪器的精度分析、产品质量的控制、某些新标准的制定、气象及地震预报、自动控制中的数学模型的制定及其他许多场合中，回归分析往往是一种很有用的工具。

三、回归分析与最小二乘的关系

回归分析是基于最小二乘原理的，回归方程系数的求解，特别是一元线性回归方程的求解与最小二乘法有一定的相似性，两者的主要不同是，最小二乘法对研究事物内部规律的数学表达式——经验公式，在得到该公式待求参数估计量后，只进行精度评价，而不研究所拟合的经验公式的整体质量。回归分析求解回归方程系数后，还需进一步对所得的回归方程——经验公式的整体精度进行分析和检验，以确定回归方程的质量水平，并定量地评价回归方程与实际研究的事物规律的符合程度，即进行回归方程的方差分析与显著性检验等。由此表明，最小二乘原理是回归分析的主要理论基础，而回归分析则是最小二乘原理的实际应用与扩展，它不仅研究一元回归分析，还有多元回归分析等内容。

第二节　一元线性回归

一元回归是处理两个变量之间的关系的，即两个变量 x 和 y 之间若存在一定的关系，则可通过实验，分析所得数据，找出两者之间关系的经验公式。假如两个变量之间的关系是线性的就称为一元线性回归，这就是工程上和科研中常遇到的直线拟合问题。

一、一元线性回归方程

（一）回归方程的求法

下面通过具体例子来讨论这个问题。

例 6-1　测量某导线在一定温度 x 下的电阻值 y 得如下结果：

x/℃	19.1	25.0	30.1	36.0	40.0	46.5	50.0
y/Ω	76.30	77.80	79.75	80.80	82.35	83.90	85.10

试找出它们之间的内在关系。

为了研究电阻 y 与温度 x 之间的关系，把数据点在坐标纸上（见图 6-1），这种图叫散点图。从散点图可以看出，电阻 y 与温度 x 大致成线性关系。因此，人们假设 x 与 y 之间的内在关系是一条直线，这些点与直线的偏离是实验过程中其他一些随机因素的影响而造成的。这样就可以假设这组测量数据有如下结构型式：

$$y_t = \beta_0 + \beta x_t + \varepsilon_t, \quad t = 1, 2, \cdots, N \tag{6-1}$$

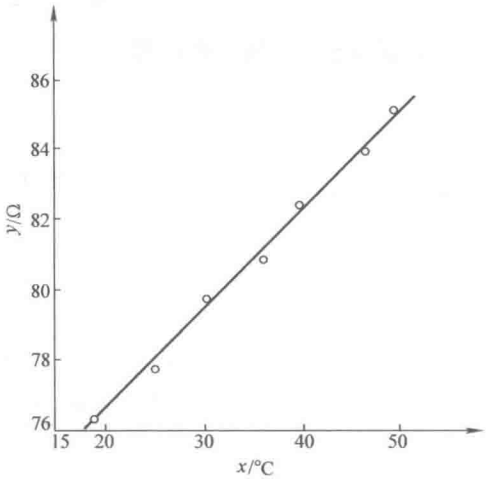

图 6-1

式中的 ε_1，ε_2，…，ε_N 分别表示其他随机因素对电阻 y_1，y_2，…，y_N 影响的总和，一般假设它们是一组相互独立并服从同一正态分布 $N(0, \sigma)$ 的随机变量（本章对 ε_t，$t = 1, 2, \cdots, N$ 均作这样的假设）。变量 x 可以是随机变量，也可是一般变量，不特别指出时，都作一般变量处理，即它是可以精确测量或严格控制的变量。这样，变量 y 是服从 $N(\beta_0 + \beta x_t, \sigma)$ 的随机变量。式（6-1）就是一元线性回归的数学模型。在例 6-1 中 $N = 7$，将表中的数据代入式（6-1），就可以得到一组测量方程，该方程组与式（5-7）完全相似，只是方程组中每个方程形式都相同，即都为式（6-1）的形式，但比式（5-7）中的方程形式更规范。由式（6-1）组成的方程组中有两个未知数，且方程个数大于未知数的个数，适合用最小二乘法求解。由此可见，回归分析只是最小二乘法的一个应用特例。

所以，人们用最小二乘法来估计式（6-1）中的参数 β_0、β。

设 b_0 和 b 分别是参数 β_0 和 β 的最小二乘估计，于是得到一元线性回归的回归方程

$$\hat{y} = b_0 + bx \tag{6-2}$$

式中，b_0、b 为回归方程的回归系数。

对每一个 x_t 由式（6-2）可以确定一个回归值 $\hat{y}_t = b_0 + bx_t$。实际测得值 y_t 与这个回归值 $\hat{y}_t$ 之差就是残余误差 v_t

$$v_t = y_t - \hat{y}_t = y_t - b_0 - bx_t \quad t = 1, 2, \cdots, N \tag{6-3}$$

应用最小二乘法求解回归系数，就是在使残余误差平方和为最小的条件下求解回归系数 b_0 和 b。这种方法我们在第五章中已经熟悉了。用矩阵形式，令

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad \boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \quad \boldsymbol{b} = \begin{pmatrix} b_0 \\ b \end{pmatrix} \quad \boldsymbol{V} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{pmatrix}$$

则式（6-3）的矩阵形式为

$$\boldsymbol{Y} - \boldsymbol{Xb} = \boldsymbol{V} \tag{6-4}$$

假定测得值 y_t 的精度相等，根据最小二乘原理，回归系数的矩阵解为

$$\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \boldsymbol{CB} \tag{6-5}$$

计算式（6-5）的下列矩阵

$$\boldsymbol{A} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{pmatrix} N & \sum_{t=1}^{N} x_t \\ \sum_{t=1}^{N} x_t & \sum_{t=1}^{N} x_t^2 \end{pmatrix}$$

$$\boldsymbol{C} = \boldsymbol{A}^{-1} = \frac{1}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2} \begin{pmatrix} \sum_{t=1}^{N} x_t^2 & -\sum_{t=1}^{N} x_t \\ -\sum_{t=1}^{N} x_t & N \end{pmatrix} \tag{6-6}$$

$$\boldsymbol{B} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} = \begin{pmatrix} \sum_{t=1}^{N} y_t \\ \sum_{t=1}^{N} x_t y_t \end{pmatrix}$$

将 $\boldsymbol{C}$、$\boldsymbol{B}$ 代入式（6-5），解得 b_0、b

$$b = \frac{N\sum_{t=1}^{N} x_t y_t - \left(\sum_{t=1}^{N} x_t\right)\left(\sum_{t=1}^{N} y_t\right)}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2} = \frac{l_{xy}}{l_{xx}} \tag{6-7}$$

$$b_0 = \frac{\left(\sum_{t=1}^{N} x_t^2\right)\left(\sum_{t=1}^{N} y_t\right) - \left(\sum_{t=1}^{N} x_t\right)\left(\sum_{t=1}^{N} x_t y_t\right)}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2} = \bar{y} - b\bar{x} \tag{6-8}$$

式中

$$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x_t \tag{6-9}$$

$$\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y_t \tag{6-10}$$

$$l_{xx} = \sum_{t=1}^{N} (x_t - \bar{x})^2 = \sum_{t=1}^{N} x_t^2 - \frac{1}{N}\left(\sum_{t=1}^{N} x_t\right)^2 \tag{6-11}$$

$$l_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right) \tag{6-12}$$

$$l_{yy} = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N} y_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i\right)^2 \tag{6-13}$$

式中，l_{yy}是为了以后做进一步分析的需要而在这里一并写出。

将式（6-8）代入回归方程式（6-2），可得回归方程的另一种形式

$$\hat{y} - \bar{y} = b(x - \bar{x}) \tag{6-14}$$

由此可见，回归方程式（6-2）通过点（$\bar{x}$，$\bar{y}$），明确这一点对回归方程的作图是有帮助的。

由式（6-7）、式（6-8）求回归方程的具体计算通常是通过列表进行的。例 6-1 的计算见表 6-1 和表 6-2，由此可得回归方程

$$\hat{y} = 70.90\Omega + (0.2824\Omega/℃)x \tag{6-15}$$

这条回归直线一定通过（$\bar{x}$，$\bar{y}$）这一点，再令 x 取某一 x_0，代入回归方程式（6-15）求出相应的 $\hat{y}_0$，连接（$\bar{x}$，$\bar{y}$）和（x_0，$\hat{y}_0$）就是回归直线，并把它画在图 6-1 上。在本例中回归系数 b 的物理意义是温度上升 1℃，电阻平均增加 0.2824Ω。

表 6-1

序　号	x/℃	y/Ω	x^2/℃^{2}	y^2/Ω^2	xy/Ω·℃
1	19.1	76.30	364.81	5821.690	1457.330
2	25.0	77.80	625.00	6052.840	1945.000
3	30.1	79.75	906.01	6360.062	2400.475
4	36.0	80.80	1296.00	6528.840	2908.800
5	40.0	82.35	1600.00	6781.522	3294.000
6	46.5	83.90	2162.25	7039.210	3901.350
7	50.0	85.10	2500.00	7242.010	4255.000
Σ	246.7	566.00	9454.07	45825.974	20161.955

表 6-2

$\sum_{i=1}^{N} x_i = 246.7℃$	$\sum_{i=1}^{N} y_i = 566.00\Omega$	$N = 7$
$\bar{x} = 35.243℃$	$\bar{y} = 80.857\Omega$	
$\sum_{i=1}^{N} x_i^2 = 9454.07℃^2$	$\sum_{i=1}^{N} y_i^2 = 45825.974\Omega^2$	$\sum_{i=1}^{N} x_i y_i = 20161.955\Omega \cdot ℃$
$\left(\sum_{i=1}^{N} x_i\right)^2 \Big/ N = 8694.413℃^2$	$\left(\sum_{i=1}^{N} y_i\right)^2 \Big/ N = 45765.143\Omega^2$	$\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right) \Big/ N = 19947.457\Omega \cdot ℃$
$l_{xx} = \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2 / N$	$l_{yy} = \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2 / N$	$l_{xy} = \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right) / N$
$= 759.657℃^2$	$= 60.831\Omega^2$	$= 214.498\Omega \cdot ℃$
	$b = \frac{l_{xy}}{l_{xx}} = 0.2824\Omega/℃$	
	$b_0 = \bar{y} - b\bar{x} = 70.90\Omega$	
	$\hat{y} = 70.90\Omega + (0.2824\Omega/℃)x$	