观测数据的处理方法

陈久宇 编著
林 见

上海交通大学出版社

观测数据的处理方法
上海交通大学出版社出版
（淮海中路1984弄19号）
新华书店上海发行所发行
浙江上虞汤浦印刷厂排版
常熟市大义印刷厂印装

开本850×1168毫米 1/32 印张13.375 字数306,000
1987年5月第1版 1987年6月第1次印刷
印数：1—2800
统一书号：ISBN7—313—00007—3/021 科技书目：150—317

定价：2.50元

目　录

第一章　绪论……（1）
§1-1　回归分析在观测资料整理中的意义 ……（1）
§1-2　回归分析和相关分析 ……（2）
§1-3　回归方程式 ……（3）
§1-4　回归参数估计的最小二乘法与最大似然估计法 ……（5）
第二章　一元线性回归及其应用……（7）
§2-1　一元线性回归的基本原理和方法 ……（7）
§2-2　一元线性回归方程的有效性与精度等问题 …（16）
§2-3　电子计算机在一元线性回归方程中的应用 ……（37）
§2-4　一元线性回归方程的稳定性 ……（42）
§2-5　根据回归方程预报 y 的取值的补充说明 ……（49）
§2-6　对观测数据的评价和对观测工作的要求 ……（55）
§2-7　回归分析中的统计检验 ……（57）
§2-8　两条回归直线的比较 ……（76）
§2-9　一元回归曲线——线性化变换分析法 ……（92）
§2-10　配曲线回归方程的有效性和精度问题 ……（105）
第三章　多元线性回归分析和应用……（112）
§3-1　二元线性回归、回归平面……（113）
§3-2　k 元线性回归——回归超平面 ……（122）
§3-3　多元线性回归方程的有效性和精度问题 ……（126）
§3-4　两个多元线性回归方程的比较 ……（167）
§3-5　自变量在多元回归方程中的重要性考察 ……（171）

§3-6 多元线性回归分析的步骤总结 ……(184)
第四章 多项式回归及用正交多项式配回归 ……(189)
§4-1 多项式回归 ……(189)
§4-2 加权多项式回归 ……(195)
§4-3 正交多项式配回归 ……(198)
§4-4 在电子计算机上实现正交多项式配回归 ……(217)
第五章 最佳回归方程和逐步回归分析 ……(231)
§5-1 观测数据中具体问题的多因子性质 ……(231)
§5-2 最佳回归方程的概念 ……(239)
§5-3 选择最佳回归方程的方法 ……(239)
§5-4 逐步回归方程中挑选和剔除因子的概念 ……(246)
§5-5 逐步回归分析法的具体步骤 ……(257)
§5-6 逐步回归分析的计算实例 ……(300)
§5-7 在电子计算机上实现逐步回归分析 ……(333)
第六章 插值函数 ……(339)
§6-1 概述 ……(339)
§6-2 拉格朗日(Lagrange)插值 ……(340)
§6-3 在电子计算机上实现拉格朗日一元 n 点插值 ……(357)
§6-4 分段插值 ……(358)
§6-5 样条(Spline)插值函数 ……(361)
§6-6 在电子计算机上实现样条插值函数的计算 ……(382)
§6-7 数值积分 ……(386)
§6-8 在电子计算机上实现数值积分 ……(394)
附 录 ……(397)
附表 1 u 检验的 u_α 值表 ……(397)
附表 2 小子样 t_α 分布数值表 ……(398)
附表 3 χ^2 检验的 χ^2 分布表 ……(399)

附表 4　F 检验的 F 分布表 ……（400）
附表 5　正交多项式表……（406）

第一章 绪 论

§1-1 回归分析在观测资料整理中的意义

在数理统计中，回归分析是处理变量之间关系的一种常用方法。近十多年来，由于测量仪器的发展及电子计算机的广泛应用，它的作用就更大了，不仅能运算大量的观测数据（很大的子样），而且还能处理多因子（很多变量）之间的关系，并从这些因子中抽选重要的因子，建立最优的回归方程。在已确定若干个因子的条件下，可通过手算（对较简单的情况）及计算机计算来求得回归方程。观测资料的整理分析工作就是处理变量之间的关系，与求得这种关系的规律性。从以后的论述可以看出，回归分析在这一工作中是一个甚为有用的数学工具。

借助回归分析所求得的回归方程即测值变量之间关系的数学表达式，在必要的时候，它可使我们不必直接测定某些数据，而只需把与该数据有关的另一些数据测出来，通过它们之间存在着的数学关系式，推算出所要测定的量，即所谓预报；或根据专业知识和工程原理，对已求得的规律作分析和判断，以取得结构是否处于正常运行状态的信息；或经过分析判断，认定这个规律是正常的情况下，可用它同以后的观测量相对照，以判断结构的动向是否失常，后两者即所谓监视结构安全。本书中，我们把回归分析作为重点内容，并认为读者已有一定的概率论、数理统计与线性代数等数理基础。在介绍回归分析的原理和方法时结合一些大坝观测实例，以帮助大家理解和运用。

§1--2 回归分析和相关分析

变量之间的关系一般有两种.

一种是变量间存在着完全确定的关系. 这可用最简单的例子静水压强来说明，$p=\gamma h$，其中 γ 是水的重度，通常看作常量，h 是水深，p 是水深为 h 处的静水压强，若已知水深 h，则静水压强 p 就可以按照上式求得. 在科学技术中，有大量问题可以用理论方法建立变量间的确定关系，这种关系叫做函数关系.

另一种是变量之间存在着不确定性的关系. 许多变量间的函数关系往往一时还不能从理论上建立起来，在观测中，大量的问题属于这一类. 例如经过若干假定之后，可以用力学理论推求混凝土坝顶位移量同库水位、坝体温度、材料徐变等之间的确定函数关系，但是其理论计算结果，一般很难与实测结果相符合，这种与实测结果不相符合的程度，各坝之间肯定也是不相同的. 所以坝顶位移量与库水位、坝体温度、材料徐变等之间的关系可认为是一种不存在确定性关系的例子. 对一个具体的混凝土坝来讲，影响位移的因素除了上述三个之外，可能还有其他很多因素，其中有的因素是人们一时还没有认识到，有的虽然已经被人们认识了，但暂时还无法测量. 由于影响因素多，再加上在观测各种变量时产生的误差，以及上述因素有时又会偶然地综合在一起，这就造成了变量之间关系的不确定性. 然而，这种不存在确定性关系的随机变量之间的关系也不是没有规律可循的，大量偶然性中蕴含着必然性的规律. 只要人们经过长期多次观测，就不难发现这种随机变量之间确实也存在着某种客观的规律性，我们把这一类变量之间的关系称为统计相关.

当然，变量之间的确定性关系(即函数关系)和不存在确定性关系(即统计相关)，二者之间没有一条不可逾越的鸿沟. 确定性关系往往是通过大量的统计相关关系表现出来的，也就是当对事

物内部规律了解得更加深刻的时候，统计相关关系就可能转化为确定性关系。但未被彻底认识的确定性关系，或尚未发展上升到确定性关系时，往往在某一个认识阶段上，仅仅表现为统计相关的关系。

相关分析和回归分析研究的都是随机变量之间的相关关系。但是在数理统计中，二者的意义是有区别的。相关分析是把变量都看做随机变量，其目的是确定变量之间的相互联系的程度如何，分析中假定所有随机变量的误差必须都呈正态分布。回归分析则是应用数学方法对大量观测数据加以处理，从而确定上述不存在确定性关系的变量之间的关系规律性，并用数学关系式表达出来。在回归分析中，假定因变量的误差呈正态分布，而对自变量的误差分布并无要求，也就是只考虑在自变量保持一系列定值时，因变量这个随机变量是如何变化的。总之，相关分析的前提是把全部变量视为随机变量，而回归分析则只把因变量视为随机变量，把自变量视为非随机变量。

§1-3 回归方程式

$E\{Y|X_1\}, E\{Y|X_2\}, \cdots E\{Y|X_n\}$表示在$X=X_1, X=X_2 \cdots X=X_n$时$Y$的条件数学期望。我们假定在大坝观测中的变量都是连续型正态分布的。

现在来论述回归方程的概念。

研究随机变量之间的相互关系时，回归分析只把因变量视为随机变量，而把自变量视为非随机变量，即把自变量视为一系列定值。因此，回归分析的主要目的就是要确定因变量的条件数学期望是如何随着自变量的变动而变动，也就是说，要找出这种变动的规律。然后根据这些规律，从自变量求出和因变量相应的条件数学期望。例如，确定了坝顶水平位移的条件数学期望与库水位之间的变动规律，就可以从已知库水位求出相应的位移平均值。

如果只考虑两个变量间的相关关系，则对应于自变量的每一数值，因变量只有一个唯一的条件平均数与之对应。所以，因变量的条件数学期望与自变量之间存在着的关系，就变成函数关系。这种表示因变量条件数学期望随自变量的变动而变动的数学表达式，就称为回归方程，而回归方程的函数图形就称为回归曲线。

用上面的符号表示，理论的回归方程是

$$E\{Y|X\}=f(X,A,B)。\qquad(1\text{-}3\text{-}1)$$

当自变量不是一个而是 n 个时，理论回归方程为

$$E\{Y|X_1,X_2\cdots,X_n\}=f(X_1,X_2\cdots,X_n,A,B_1,B_2,\cdots B_n)。\qquad(1\text{-}3\text{-}2)$$

注意，上述的理论回归方程或称真正回归方程，一般是不能求得的，因为式中的参数 $B_1,B_2,\cdots,B_n$，常数项 A 等理论上都是从母体资料来的。数理统计学中所讨论的一切问题，归根结蒂，都是抽样估计问题。抽样估计就是从全部研究对象（母体）中随机地抽出一部分对象（子样）进行观察，并根据所获得的子样资料对于母体的数量特征和规律性进行估计。回归分析也是这样，如果要求出真正的坝顶水平位移条件数学期望 $\overline{Y}_X$ 与库水位 X 相关的回归方程，当然要长期连续不断取得极其大量的库水位测值 X 和坝顶水平位移 Y。然而，这既不可能，也无必要。我们只要定期地取得部分观测资料，并根据这部分观测资料，去找回归方程，对坝顶水平位移的条件数学期望 $\overline{Y}_X$ 随着库水位 X 的变化而变化的全部规律性作出估计，这就是一种抽样估计。所以回归方程参数的确定，是应用数理统计的方法，根据子样资料对上述理论回归方程的参数进行估计。如果用 $b_1,b_2,\cdots,b_n$ 和 a 等作为 $B_1,B_2,\cdots,B_n$ 和 A 的估计值，则所测得的回归方程叫做经验回归方程，用

$$\hat{y}=f(x),$$

$$\hat{y}=f(x_1,x_2,\cdots,x_n),$$

来代替式（1-3-1）和式（1-3-2）。因此，寻找回归方程就归结为根据子样资料，去找 $b_1,b_2,\cdots,b_n$ 及 a 等参数，得出经验回归方程。我

们所说的回归方程即指经验回归方程。

顺便指出，对某个问题作回归分析时，总是面临着选择自变量（即所谓选择因子）的任务，而这件事颇为复杂。在大坝原型观测的资料整理分析中，要涉及坝工的专业知识和数理统计的知识，我们把它留在本书第五章叙述。但一般的回归分析都是根据确定的自变量和确定的函数形式去找回归方程式，所以比较省事。本书在前四章，主要介绍在事先已确定一些自变量和已确定函数形式的条件下，如何去找出回归方程式。

§1-4 回归参数估计的最小二乘法与最大似然估计法

上述的理论回归方程式可能有多种类型。一个自变量时，理论回归方程 (1-3-1) $E\{Y|X\}=f(X)$ 的类型可能是一元线性的，如

$$E\{Y|X\}=\overline{Y}_X=A+BX, \qquad (1\text{-}4\text{-}1)$$

也可能是一元非线性，如二次抛物线方程

$$E\{Y|X\}=\overline{Y}_X=A+BX+CX^2, \qquad (1\text{-}4\text{-}2)$$

也可能是别的类型。k 个自变量时，理论回归方程 (1-3-2) $E\{Y|X_1,X_2,\cdots,X_k\}=f(X_1,X_2,\cdots X_k)$ 的类型可能是多元线性的，如

$$E\{Y|X_1,X_2\cdots X_k\}=\overline{Y}_X=A+B_1X_1+B_2X_2+\cdots+B_kX_k,$$

也可能是别的类型等等。

如把观测数据 $X_{1i},X_{2i},\cdots,X_{ki},Y_i(i=1,2,\cdots,n)$ 代入多元线性理论回归方程，可得

$$Y_i=A+B_1X_{1i}+B_2X_{2i}+\cdots+B_kX_{ki}+\varepsilon_i, \qquad (1\text{-}4\text{-}3a)$$

其中 ε_i 表示各次观测值的误差，通常有如下三个假定：

（1）误差 ε_i 没有系统性，ε_i 的数学期望全为零，$E(\varepsilon_i)=0$ $(i=1,2,\cdots,n)$；

（2）各次观测互相独立，并有相同的精度，即 ε_i 之间的协方差

可表为

$$\operatorname{cov}(\varepsilon_t,\varepsilon_h)=\begin{cases}0,\text{当 } t\neq h\\ \sigma^2,\text{当} t=h\end{cases}\quad\text{（即方差）；}$$

（3）观测误差服从正态分布。

这三条假定可概括为：误差 ε_i 相互独立地服从$N(0,\sigma^2)$分布，表示为 $\varepsilon_i\sim N(0,\sigma^2)$。

回归分析在大坝观测资料的数据处理、曲线拟合、建立经验公式以及在各类预报问题中均有广泛应用。但作不同应用时，计算重点常有差异。例如，在建立经验公式与曲线拟合中，关心的是对回归系数的估计；在预报问题中，关心的是回归方程的预报精度，这就需要给出误差 ε 的公共方差 σ^2 的估计值。在分析影响 Y 的主要因素时，就需要对各回归系数作显著性的检验，判定各个 X_i 的重要程度，为变量的取舍提供依据。

通常用最小二乘法根据子样资料去估计回归方程的参数。除了最小二乘法之外，还可采用最大似然估计法来估计回归方程中的参数。

本书主要采用最小二乘法，但在适当的地方也引入最大似然估计法。当 Y 是正态分布时，则最小二乘估计法与最大似然估计法给出相同的结果。

第二章　一元线性回归及其应用

§2-1　一元线性回归的基本原理和方法

一、回归直线

因变量 y 与单个自变量 x 之间存在着某种相关关系，这自然是一种最简单的情况。通过观测，我们得到了关于 x、y 两个变量的一组数据(子样)。若将这组数据的 x、y 值一一对应地绘在 x-y 坐标上，就得到散点图；从这个散点图若能直观地看出两个变量之间大致呈线性关系，则利用这个子样的资料找出能描述 x，y 之间线性关系的回归方程，称为一元线性回归方程，其函数图形称为回归直线。

例 2-1　某水库的水位与该坝 32* 坝段的坝基扬压力的测值记录如表 2-1。水库水位用 x 表示，扬压力用 y 表示，找出它的回归线性方程。

表2-1

编号	观测日期			水库水位	扬压力
	年	月	日	x (m)	y (kN/m)
1	1973	3	1	247.9	3950
2		4	25	251.1	4270
3		5	8	253.7	4580
4		7	6	255.8	4610
5			18	254.6	4400
6		8	15	259.1	4800
7		9	10	262.3	5070
8		10	12	261.5	5110
9		11	6	260.3	5220

我们按上表所列的 x,y 值，先绘出散点图，如图 2-1 所示。这些散点大致呈线性关系。那么，我们很自然地想到可以用一条直线来表示 x 与 y 的关系，并借助最小二乘法，可找到

$$y=a+bx. \qquad (2\text{-}1\text{-}1)$$

这就是扬压力 y 对水库水位 x 的回归方程，其图形称为回归直线。

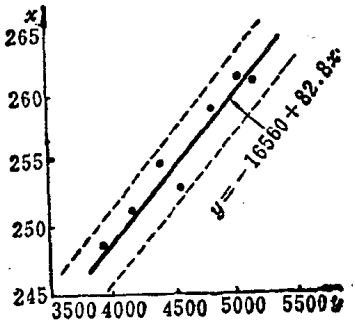

图 2-1

二、确定回归直线的方法——最小二乘法

回归直线应是在一切直线中最接近实际测点的直线，也就是用这条直线来代表 x、y 的关系与实际数据的误差比其他任何直线都要小。这就是在 §1-4 中所叙述的利用最小二乘法的原则来确定 a,b 值。

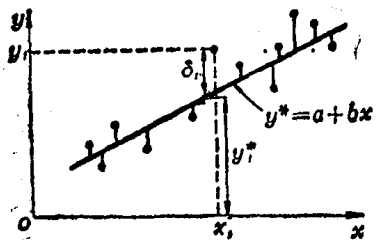

图 2-2

如果一共有 n 个测点数据，分别用 $x_t,y_t(t=1,2,\cdots,n)$ 代表，则在任一条直线

$$\hat{y}=y^*=a+bx \qquad (2\text{-}1\text{-}2)$$

上用 x_t 代入，$y^*_t(\hat{y}_t)$ 一般不正好等于 y_t，如图 2-2 所示，其误差为

$$\delta_t=y_t-y^*_t,$$

或

$$\delta_t = y_t - a - bx_t。 \qquad (2\text{-}1\text{-}3)$$

n 个测点所引起的总误差为

$$Q^* = \sum_{t=1}^{n} \delta_t^2 = \sum_{t=1}^{n} (y_t - a - bx_t)^2。 \qquad (2\text{-}1\text{-}4)$$

回归直线是应用最小二乘法原则来确定的，换句话说，回归系数 b 及常数项 a 应使 Q^* 达到极小值。

要使 Q^* 达到极小值，只要将上式分别对 a、b 求偏导数，并令它们等于零，

即

$$\frac{\partial Q^*}{\partial a} = -2\sum_{t=1}^{n} (y_t - a - bx_t) = 0, \qquad (2\text{-}1\text{-}5a)$$

$$\frac{\partial Q^*}{\partial b} = 2\sum_{t=1}^{n} (y_t - a - bx_t)x_t = 0。 \qquad (2\text{-}1\text{-}5b)$$

以上两方程称为正规方程，从中可求出 a，b 值。

从(2-1-5a)式得，

$$na = \sum_{t=1}^{n} y_t - b\sum_{t=1}^{n} x_t,$$

$$a = \sum_{t=1}^{n} y_t / n - b\sum_{t=1}^{n} x_t/n。$$

上式中，若以 $\sum_{t=1}^{n} y_t/n = \overline{y}$ 和 $\sum_{t=1}^{n} x_t/n = \overline{x}$ 分别代表 x_t、y_t 的算术平均数，则上式可变为如下简单的形式：

$$a = \overline{y} - b\overline{x}。 \qquad (2\text{-}1\text{-}6)$$

式中 a 称为回归方程的常数项。

从(2-1-5b)式又得

$$\sum_{t=1}^{n} x_t y_t - a \sum_{t=1}^{n} x_t - b \sum_{t=1}^{n} x_t^2 = 0.$$

以(2-1-6)式的 a 代入，整理后得

$$b = \frac{\sum_{t=1}^{n} x_t y_t - \frac{1}{n}\left(\sum_{t=1}^{n} x_t\right)\left(\sum_{t=1}^{n} y_t\right)}{\sum_{t=1}^{n} x_t^2 - \frac{1}{n}\left(\sum_{t=1}^{n} x_t\right)^2}, \qquad (2\text{-}1\text{-}7)$$

或

$$b = \frac{\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} \bar{x} y_t}{\sum_{t=1}^{n} x_t^2 - \sum_{t=1}^{n} \bar{x} x_t}$$

$$= \frac{\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} \bar{x} y_t - \sum_{t=1}^{n} x_t \bar{y} + \sum_{t=1}^{n} \bar{x}\bar{y}}{\sum_{t=1}^{n} x^2{}_t - \sum_{t=1}^{n} \bar{x} x_t - \sum_{t=1}^{n} \bar{x} x_t + \sum_{t=1}^{n} \bar{x}^2}$$

$$= \frac{\sum_{t=1}^{n} (x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}. \qquad (2\text{-}1\text{-}7a)$$

上式中 b 称为回归系数，它表示 $\hat{y}$ 随着 x 的增率。

由此可见，a,b 完全取决于给定的观测资料。有了 a,b 值，即可写出回归方程

$$\hat{y} = a + bx.$$

这里，顺便要求大家记住以后常用的几个符号：

$$S_{xx} = \Sigma x^2{}_t - \frac{1}{n}(\Sigma x_t)^2 = \Sigma (x_t - \bar{x})^2, \qquad (2\text{-}1\text{-}8)$$

$$S_{yy} = \Sigma y^2{}_t - \frac{1}{n}(\Sigma y_t)^2 = \Sigma (y_t - \bar{y})^2, \qquad (2\text{-}1\text{-}9)$$

$$S_{xy}=S_{yx}=\Sigma x_t y_t-\frac{1}{n}(\Sigma x_t)(\Sigma y_t)$$

$$=\Sigma(x_t-\bar{x})(y_t-\bar{y}). \qquad (2\text{-}1\text{-}10)$$

其中

$$\begin{cases}\bar{x}=\Sigma x_t/n,\\ \bar{y}=\Sigma y_t/n.\end{cases} \qquad (2\text{-}1\text{-}11)$$

注意，为了简单起见，今后一般均以总和号 Σ 代表 $\sum_{t=1}^{n}$.这样，式(2-1-7a)又可写成

$$b=S_{xy}/S_{xx}. \qquad (2\text{-}1\text{-}7b)$$

最后，小结一下：一元线性回归方程的建立，首先根据子样中各测值散点图的大致线性关系，决定采用线性模型；然后利用子样资料，根据最小二乘法原则，求出常数项 a 和回归系数 b，这样就完全确定了一元线性回归方程式

$$\hat{y}=a+bx.$$

对于这个子样资料而言，这个回归方程式，是误差平方和 $\Sigma\delta_i^2$ 最小的一个.

三、一元线性回归方程的特点

从上面的推导，可以看出一元线性回归方程式有两个特点：

（1）从式(2-1-6)可知，$\bar{x}$，$\bar{y}$ 满足回归直线方程式，即回归直线必通过 $\bar{x}$，$\bar{y}$ 点。我们如果把分布在 x-y 平面上的 n 个观测点(散点图)看作是相同的几个质点，那么 $\bar{x}$，$\bar{y}$ 便是这个质点系的质心位置，而回归直线必须通过这些散点的质心．记住这个直观的结论，对作回归直线很有帮助.

（2）从式(2-1-7a)的回归系数公式

$$b=\frac{\sum_{t=1}^{n}(x_t-\bar{x})(y_t-\bar{y})}{\sum_{t=1}^{n}(x_t-\bar{x})^2}$$

看出，其分母是所有测值 x_t 距其算术平均数 $\bar{x}$ 的离差平方和，一般 x 的测值总不完全相等，所以分母必然是正数，因此回归系数 b 的符号取决于式(2-1-7a)的分子，即离差 $(x_t-\bar{x})$ 与离差 $(y_t-\bar{y})$ 乘积之总和。当 $b>0$ 时，即回归直线的斜率是正数，$\hat{y}$ 值随 x 的增加而增加，而当 $b<0$ 时，则相反。

四、具体计算步骤和格式

按式(2-1-7b)、(2-1-8)式(2-1-10)计算回归系数 b。

S_{yy} 在计算 b 的时候虽然不用，但以后作方差分析时要用到，因此就顺便用(2-1-9)式计算出来。

就 S_{xy}，S_{xx}，S_{yy} 各式看来，还要计算出 x_t^2，y_t^2，x_ty_t 各量，因此，可以把计算的总步骤和格式列述于下：

(1)列表

见表 2-2，编号即测值序号，x_t，y_t 是测值，由此计算得 x_t^2，y_t^2，x_ty_t。

表2-2

编　号	x_t	y_t	x_t^2	y_t^2	x_ty_t
1					
2					
3					
⋮					
n					
Σ					

(2)将表 2-2 中各栏的数相加，将和数放在最后一行Σ中。