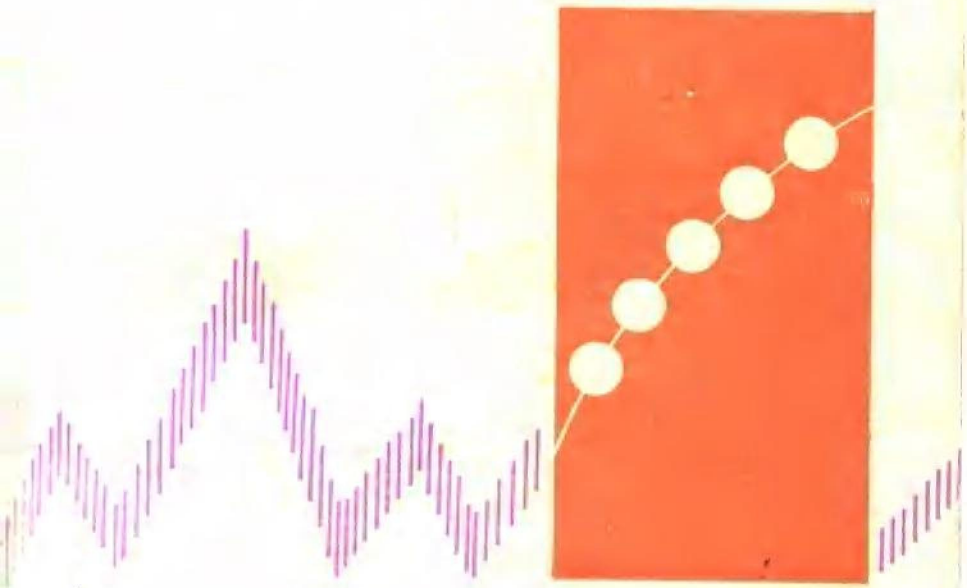
统计模型与预报方法

安鸿志　顾岚



气象出版社

统计模型与预报方法

安鸿志　　顾　岚

气象出版社

内 容 简 介

本书介绍各种回归和自回归模型，以及它们的各种混合模型的统计方法。其中包括许多新的方法和研究成果，如样条回归、门限回归和门限自回归分析，以及模型自变元的各种最佳准则选择方法并附有部分实例和FORTRAN语言程序。

本书可供气象、水文、地震、生物、医学、经济、人口统计、工业自动控制、国防与空间技术等领域的科研人员和工程技术人员阅读，亦可作为大专院校和部分中等技术学校的教学参考书。

时序模型与预报方法

安鸿志　　顾　岚

责任编辑　刘生长

气象出版社出版

（北京西郊白石桥路46号）

中国科学技术情报所印刷厂印刷

新华书店北京发行所发行　　全国各地新华书店经售

*　　*　　*

开本：787×1092 1/32　印张：9.5　字数：203　千字

1986年8月第一版　　1986年8月第一次印刷

印数：1—2500

统一书号：13194·0333　定价：2.25元

ISBN 7-5029-0008-X/P·0005

前　言

统计模型方法是数据分析的一种重要方法，主要用于处理受到偶然因素影响的各种量的记录值，其目的在于掌握这些量的变化规律。

客观世界中的各种量的变化规律可以分为两类。一类是确定性的规律，粗略地说，就是一数量能被若干它数量唯一确定，或者其自身按某种确定性的规律变化，这些关系或规律称为确定性的。例如，圆的面积被其半径唯一确定。又如，自由落体的位置被一时间的二次多项式所描述。另一类是非确定性的规律，它不同于前一类。例如，某地区的逐月降雨量，它不能被其它量按一准确关系唯一确定，其自身也无法用某种确定性的函数准确描述。可以说它不具有确定性的变化规律，但不能说它没有规律。基于对多年的实际记录的雨量资料的分析，人们会发现某些寓于大量偶然数据中的相对稳定性规律。这种规律被称为统计性规律。研究和统计分析这种规律，是概率论与数理统计学的重要内容。

具有统计性规律的变量或者数据资料，在自然科学、社会科学以及生产、生活的各个领域中都会遇到。分析这类数据中存在的统计规律，有的是为了对某些量进行预报；有的是以控制和管理为目的；此外也还有其它各种目的；其中以预报为目的的最为普遍。例如，在气象领域中，人们可以记录各种气象要素的实际数值，分析这些数据的最主要的目的是预报

重要气象要素的未来值，如雨量、气温等。在经济领域中，类似的数据分析也多是为预报服务。比如各种商品的销售量的预报，各种产值、产量的预报等等。这些预报与企业的管理和经济发展有着密切的关系。在工农业生产、自动化技术以及军工和空间技术领域里，类似的数据分析问题也很多，目的也比较广泛。在其它领域中，也会遇到上述类型的数据分析问题，比如对心电图的数据分析，流行病的预报等。解决上述数据分析问题，在统计学中有各种不同的方法，其中比较重要的一种是建立统计模型的方法。

统计模型有各种不同的类型，比如有回归模型，还有自回归与混合回归模型等。它们属于概率统计学的不同的分支，所涉及的概率论与数理统计的专门知识也较多。这对应用者来说，掌握比较全面的统计模型方法具有一定的困难。在解决实际问题时，往往不是用一种统计模型分析就能奏效，或者事先也无法知道哪一种模型比较合适。因此，应用者的确须要了解较多的统计模型方法。为了适应广大应用者的需要，我们收集了较全和较新的实用统计模型方法，并且力求少用或者不用概率统计的专门知识，以介绍方法为主写成此书。

本书共分六章。在第一章里，我们用直观方法定义各种模型，并且叙述模型拟合的最小二乘原理及其有关预备公式。在第二至四章里，分别介绍回归、自回归和混合回归模型的具体拟合方法。其中包括对于线性与非线性模型的拟合方法，对于回归变元的最佳准则选择方法，以及拟合后的预报方法。在第五章里，给出各种拟合方法的数值例子，并附以主要方法的语言程序。以上各章不仅不涉及概率统计知

识，而且也不包括比较复杂的数学证明推导。对于暂时还不熟悉矩阵证明技巧和概率统计知识的读者，容易通过阅读以上各章，学会各种模型统计方法。对于有兴趣知道各种预备公式的证明的读者，可以阅读第六章§1，这也不需要概率统计知识。对于具有概率统计基础的读者，可以阅读第六章其余各节。在这些节中，我们给出了各类模型的严格统计定义和某些简单性质，并介绍了关于本书的模型拟合方法的理论研究成果。有兴趣了解它们的数学证明的读者，可以查阅书中提供的文献。

由于水平所限，书中可能有不足或错误之处，请读者批评指正。

符 号 说 明

$\equiv$ 定义为（或恒等于）。

$\approx$ 渐近相等（或近似相等）。

$\delta_{i,j}$ 克朗内克(Kronecker)符号，即当$i=j$时$\delta_{i,i}=1$，当$i\neq j$时$\delta_{i,j}=0$。

A' 矩阵A的转置矩阵。

$detA$ 方阵A的行列式。

A^{-1} 满秩方阵的逆矩阵。

$A>0$ A为对称正定方阵。

$(a_{i,j})$ 矩阵A的元素表示法。在容易发生误会时，用$a_{i,j}$表示，比如$a_{i,j-1}$。

$j\in J$ 正整数j属于正整数集合J。

J^c 正整数集合J的余集（相对某一完全集合而言）。

$J\subset M$ 正整数集合J是M的子集合。

$J\supset M$ 正整数集合M是J的子集合。

$J\backslash M$ 正整数集合J与M的差集。

$J\cup\{j\}$ 正整数集合J再添上单元集合$\{j\}$的元j后的合集。又称为J与$\{j\}$的并集。

$J\backslash\{j\}$ 从正整数集合中去掉$\{j\}$的元素j后的差集。

log 表示以e为底的自然对数。

Ez 随机变量（或矢量）的均值(矢量)。

$N(0,Q)$ 均值为0，方差阵为Q的正态分布。

$\tilde{\beta}\sim N(0,Q)$ 随机矢量$\tilde{\beta}$的分布收敛于正态分布$N(0,Q)$。

$E\{z|z_n, z_{n-1}, \cdots, z_1\}$ 在已知$z_n, z_{n-1}, \cdots, z_1$条件下，对随机变量(或矢量)$z$的条件均值。

$\hat{\alpha}$ 对矢量α的某种估计值。

$\tilde{\alpha}$ 对矢量α的某种估计值的误差$\hat{\alpha}-\alpha$。

$\hat{\beta}_Y$ 对矢量β的Yule-Walker估计值。

$\|a\|^2$ 矢量a的分量的平方和。

(3.5) 在某章中§ 3 内的第5公式。余者相同。

(2.3.5) 指第二章§ 3 的公式(3.5)，在第二章以外的各章引用此(3.5)式时，使用此记号，在第二章内引用时，仍用(3.5)式。余者相同

图1、表1 指在同一章内的图1、表1；在同一章内，图和表连续编号。在其它章引用时，须加定语说明。

目　　录

第一章　预备知识

§1　统计规律与统计模型

一个量或几个量的取值受到偶然因素的影响时，无法用确定的数量关系或函数关系描述它们，在统计学中，称这些量为随机量或随机序列。在这些量之间，或其自身前后之间往往存在着某种统计依赖关系，也就是说，在大量的重复观察或丰富的数据资料中，存在着相对稳定性规律。这种规律称为统计性规律。当这种规律能用某一模型方式描述，或能近似描述时，称这种随机量或随机序列适合此类模型。这种模型，可以通过对相应的量的实测数据的计算分析而获得估计。所以，将这种模型称为统计模型。

统计模型的具体形式，在少数情况下能够依靠被考查的诸量的实际背景所决定，在大多数情况下并非都能如此。因此，目前所使用的各种统计模型，在绝大多数情况下都是对真实统计规律的近似描述。从另一方面来说，真实模型形式总是比较复杂的，而我们实际使用的模型又不可能太复杂了，所以，近似描述手段又是十分必要的。

现在，我们分析几个实例，以阐明统计模型产生的实际背景，以及对真实统计规律的近似描述方法。

例1，用雷达跟踪一飞行目标时，记录下目标的距离、方位角和高低角诸量，并换算成直角座标量。今考虑三个直角座标分量中的一个，记做z_k，k表示第k个采样时刻，采样

是等间隔的。由于雷达的量测带有误差，这些误差受到各种偶然因素的影响，所以，被记录到的飞行目标的轨道值也是随机序列量。从这些实际测量数据z_1，z_2，…，z_n估计目标的飞行轨道，就是一种数据处理问题。此时，如果将真实的轨道用$\varphi(k)$表示，雷达的误差量用u_k表示，那么有

$$z_k = \varphi(k) + u_k \quad k = 1, 2, \cdots, n \qquad (1.1)$$

$\varphi(k)$是非随机的客观飞行轨道的离散采样，而误差量u_k是在零值附近起伏的随机序列量。对u_k的一种简单的直观描述是：它的平均值$\overline{u}_n = (u_1 + u_2 + \cdots + u_n)/n \approx 0$，其标准偏差$\sigma_n^2 = (u_1^2 + u_2^2 + \cdots + u_n^2)/n \approx \sigma^2$，此处$\sigma^2$为一非零常数。(1.1) 式连同对$u_k$的描述，称为对序列$z_k$的模型描述。以后简称模型 (1.1) 式。由于$\varphi(k)$的函数形式没有被指定，所以我们说 (1.1) 式的模型形式没有确定。如果我们事先知道飞行目标在做匀速直线飞行，那么 (1.1) 式便可写成

$$z_k = \alpha_1 + \alpha_2 \cdot k + u_k \quad k = 1, 2, \cdots, n \qquad (1.2)$$

此时$\varphi(k) = \alpha_1 + \alpha_2 \cdot k$，有了明确的形式，其中$\alpha_1$表示目标的初始位置，$\alpha_2$表示飞行速度。但是$\alpha_1$和$\alpha_2$的具体值是未知的。这时，我们说 (1.2) 式的模型形式是已知的。由于此模型中有两个未知参数 α_1 和 α_2，连同σ^2就完全决定了 (1.2) 式的模型了，所以称 (1.2) 式的模型为有限参数模型。

匀速直线运动是最简单的飞行状态。在很多情况下，并非匀速直线飞行，而且我们也无法事先知道其飞行状态，尽管它是客观存在的。或者有时知道其飞行状态，但是由于太复杂而无法将它用有限参数形式写出。这时就须用有限参数模型做近似描述。由于轨道总是时间的连续函数，所以$\varphi(k)$

可以表示成无穷级数的形式，即

$$\varphi(k)=\sum_{j=1}^{\infty}\alpha_j k^{j-1}$$

从而（1.1）式可以写成

$$z_k=\sum_{j=1}^{\infty}\alpha_j k^{j-1}+u_k \quad k=1,\ 2,\ \cdots,\ n$$

与（1.2）式相比，此式可称为无限参数模型。我们将不去讨论这种模型。但是，从这种模型却可以引出一种近似描述方法，即取上式求和中的前p项做为近似模型：

$$z_k=\sum_{j=1}^{p}\alpha_j k^{j-1}+u_k \quad k=1,2,\cdots,\ n \qquad (1.3)$$

（1.3）式是熟知的多项式回归模型，是属于我们讨论的有限参数模型的一种。当然，还有不同于（1.3）式的其它有限参数模型，也能用于对（1.1）式的近似描述。根据具体数据z_1，z_2，…，z_n，选用哪种近似描述模型，以及怎样估计模型中的未知参数，这些正是本书要介绍的主要内容。根据实际数据，利用书中的方法获得的统计模型，称为拟合模型。拟合的基本原理为最小二乘法，这将在下一节介绍。这里的拟合一词也兼有“逼近”之义。

在例1中$\varphi(k)$仅是时刻k的函数，有时φ也可以是其它已知量的函数。试看下边一例。

例2，为了确定施某种化肥的增产效益，在相同条件下，在n个不同的单位面积的地块上，分别施用x_1，x_2，…，

x_n斤同种化肥，收获后，产量分别为z_1，z_2，…，z_n斤。由于产量受到各种偶然因素影响，不能表现出与施肥的确定关系。当我们把记录值（x_k，z_k），$k=1$，2，…，n，描绘在平面直角坐标图上时（见图1），发现这些数据点集中在某一曲线$z=\varphi(x)$的附近。虽然在此例中无法从直观背景确定$\varphi(x)$的形式，但是，仿照例1中引入近似模型描述的论述，

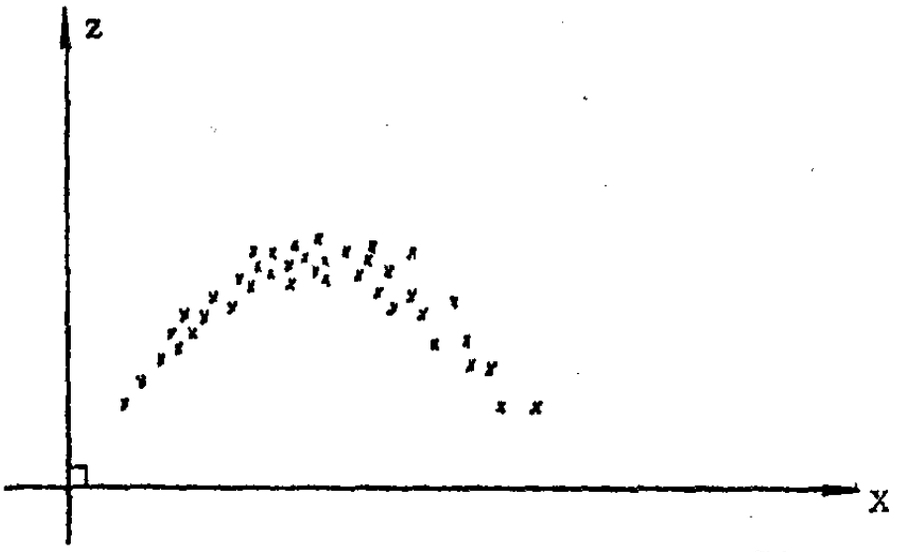


图1　单位面积的施肥量与产量的数据图

自然会引用类似（1.3）式的模型做为近似模型。比如用以下形式描述：

$$z_k=\alpha_1+\alpha_2x_k+\alpha_3x_k^2+\varepsilon_k \quad k=1,\ 2,\ \cdots,\ n \tag{1.4}$$

式中ε_k的直观描述与（1.3）式中的u_k相同，但是其实际背景与例1不同，它反映了不同地块上对施肥作用的偶然性差异。

通过例1和例2不难想象，（1.3）（1.4）式的更一般形式应当是：

$$z_k=\varphi(x_{1,k},\ x_{2,k},\ \cdots,\ x_{s,k},\ \alpha)+\varepsilon_k$$

$$k=1,\ 2,\ \cdots \qquad (1.5)$$

其中e_k的含义如前所述，φ是自变量$x_{1,k}$，$x_{2,k}$，…，$x_{s,k}$的s元已知函数，它被一个p维参数矢量$\alpha=(\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_p)^T$所唯一确定，$x_{1,k}$，$x_{2,k}$，…，$x_{s,k}$是被量测的其它量的记录值。例如，在（1.4）式中，$x_{1,k}\equiv 1$，以后称取常数1的变量为常值变量，$x_{2,k}=x_k$，$x_{3,k}=x^2_k$，$\alpha=(\alpha_1,\ \alpha_2,\ \alpha_3)^T$。

对于以上所引进的统计模型的进一步推广，是将（1.5）式中的自变量考虑为包含有随机量，或者随机序列量的模型。为此分析以下实例。

例3，某河流经某地的旬流量为z_k，其上游的旬降雨量为x_k，它们之间有着统计性的依赖关系。仿前两例可以用以下模型描述：

$$z_k=\varphi(x_k)+e_k \qquad k=1,\ 2,\ \cdots \qquad (1.6)$$

从直观背景和统计经验知道，z_k不仅与x_k有关，而且与x_{k-1}，x_{k-2}，…，x_{k-s}也有关系。所以应当考虑比（1.6）式更广泛的模型

$$z_k=\varphi(x_k,x_{k-1},\cdots,x_{k-s})+e_k \qquad k=1,\ 2,\ \cdots \qquad (1.7)$$

（1.6），（1.7）式中的e_k仍如前所述，φ是以随机序列x_k及其历史值x_{k-1}，x_{k-2}，…，x_{k-s}为其自变元的确定函数。至于φ的具体形式，与前述相同，在多数情况下是难于确定的，只能用近似手段描述。在（1.7）式中，φ的近似描述的形式是非常多的，比如，

$$\varphi(x_k,\ x_{k-1},\ \cdots,\ x_{k-s})=\alpha_1+\alpha_2x_k+\alpha_3x_{k-1}+\\ +\cdots+\alpha_{s+1}x_{k-s}$$

$$\varphi(x_k, x_{k-1}, \cdots, x_{k-s}) = \alpha_1 + \alpha_2 e^{\alpha_3 x_k} + \alpha_4 x_{k-1} x_{k-2} + \cdots + \alpha_{s+2} x_{k-s-1} x_{k-s}$$

……

都是关于φ的有限参数形式的逼近，将它们代入（1.7）式后，便得到新的有限参数模型，这些模型不同于(1.3)(1.4)式的模型。当然，如果能从气象和水文理论对z_k和x_k的依赖关系有所推敲，将帮助人们确定近似描述模型的范围。无论如何，由于不能准确给出φ的形式，最后不得不在众多的近似形式中挑选一个合适者。所以，挑选模型将是模型拟合方法中的重要一环。

除以上的统计模型外，还有其它类型的模型。其中一种是关于随机序列量自身的相互统计依赖关系的模型描述。在上述例子中，对于u_k和e_k都是用均值为零和方差为σ^2来描述的。但是，在很多情况下这种描述还不够精细。在统计学中，对于纯随机序列的自身统计规律，也有各种有限参数模型描述方法。

例4，重新讨论例1。为了研究雷达的测量误差u_k的统计特征，采用特殊手段获得准确的$\varphi(k)$之值，从而也就得到了$u_k = z_k - \varphi(k)$的各时刻取值。现在要分析u_k具有怎样的统计规律。比较简单的方法是计算u_k的平均值以及标准偏差值。这本书中，我们将要分析u_k的进一步的统计规律。比如，将(u_{k-1}, u_k)，$k=1, 2, \cdots, n$，描绘在平面直角坐标图上（见图2），假如这些点集聚在过零点的一条直线附近，仿照前面产生模型的讨论，可以引进以下模型：

$$u_k = \beta u_{k-1} + e_k \qquad k = 2, 3, \cdots, n \qquad (1.8)$$

其中e_k如前所述。(1.8) 式表现了u_k的自身之间的相依关

系。这是关于u_k的比较简单的模型描述中的一种。上面曾假设（u_{k-1}，u_k）在图2中围绕在一条过零点的直线附近，在

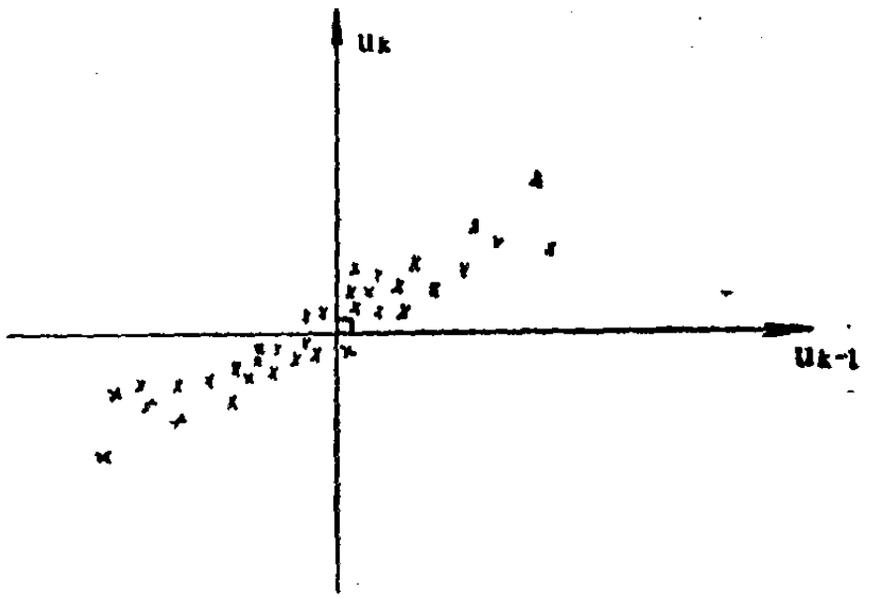


图2　（u_{k-1}，u_k）的数据平面图

多数情况下并非如此。由此不难想象，描写u_k自身统计规律的一般有限参数模型应当写成：

$$u_k = \varphi(u_{k-1},\ u_{k-2},\ \cdots,\ u_{k-p},\ \alpha) + e_k$$
$$k = p+1,\ p+2,\ \cdots,\ n \qquad (1.9)$$

而上式的一种常用简单形式为：

$$u_k = \alpha_1 u_{k-1} + \alpha_2 u_{k-2} + \cdots + \alpha_p u_{k-p} + e_k$$
$$k = p+1,\ p+2,\ \cdots,\ n \qquad (1.10)$$

模型（1.9）和（1.10）式不仅用于描述例1中的误差量u_k，对于其它随机性序列也适用。比如对例3中的河流旬流量z_k，以及降雨量x_k，都可以用（1.9）或（1.10)式近似描述它们各自的自身统计规律。但是，从直观上可以想象，在流量z_k和雨量x_k的数据资料都具备的情况下，仅仅建立z_k和x_k的自身的统计模型，对掌握z_k的变化规律是不充分的。