生物数学模型的统计学基础

唐守正　李　勇　著

科　学　出　版　社

2002

前　言

利用数学模型描述客观世界是科学研究和技术开发的一种重要手段.这种方法在力学、物理学和工程技术中早已获得巨大的成功,在化学中也取得了重大进展.数学,曾经被认为在生物学中的作用等于零.可是到了 20 世纪,数学在生物学中也起到越来越大的作用.数量遗传学和数量生态学等新兴学科相继诞生,标志着数学这个古老学科已经成为生物学研究的一种重要工具.数学在生物学中应用的一种主要形式就是生物数学模型.在第二次世界大战以后,随着计算机技术的高速发展,数据库、数值模拟等技术已经逐渐成为许多行业不可缺少的基础.在这种环境中,到 20 世纪 70 年代以后,用数学模型方法研究生物现象,已形成一种浪潮,冲击着生物学的许多领域.

生物数学模型研究如何利用数学语言和数学工具,描述生物学现象和规律,通过数学或逻辑推理得到一些结论,然后再将这些结论用来解释、预测生物学现象和发现新的规律.

生物数学模型与其他领域的数学模型可以起到类似的作用.但是,生物学现象与力学现象、物理现象甚至化学现象有着一些非常明显的不同特点.首先,生物的特点在于它的多样性,即差异性."没有完全相同的两个个体",这句话的含义不同于"我们不能两次进入同一条河流",因为有时两个同种个体之间在某个性质上的差异(例如大小)甚至可以超出种间差异.最近克隆的多莉羊出现的早衰现象,说明即使克隆生物可能也存在差异.其次,生物与众多的环境因素相关,决定了生物学现象的复杂性,这种复杂性是力学、物理和化学无法比拟的.第三,在生物与环境的协同进化中所发生的变异,导致"同一物种"与"同一环境"在不同时间可能有不同的关系.这三个特点决定了在生物数学模型中几乎不存在普遍

适用的“常数”和所谓的“确定性模型”，甚至难以定义什么是测量对象的“真值”. 因此，参数估计就成为建立生物数学模型的首要问题之一. 近年来，发展出许多统计学方法来处理模型的参数估计问题. 这些统计学方法有可能成为生物数学模型建模的有用工具.

这里举两个例子说明正确使用统计方法对建立生物数学模型的重要性.

第一，在系统生态学中，对于一个大系统，往往把它分解成一系列相互联系的子系统. 一个子系统的输出，可能是另外一些系统的输入，每个子系统还可能受到外部的控制，整个系统的状态方程可以写成一个联立方程组

$$\begin{cases} \dfrac{\mathrm{d}y_1}{\mathrm{d}t} = f_1(y_1, \cdots, y_p, x_1, c_1), \\ \cdots\cdots\cdots\cdots \\ \dfrac{\mathrm{d}y_p}{\mathrm{d}t} = f_p(y_1, \cdots, y_p, x_p, c_p), \end{cases} \tag{0.1}$$

其中 y 是系统的状态变量，x 是控制变量，c 是系统参数. 经济数学家早已认识到，对于线性方程组，上述模型中的变量 y 是内生变量，变量 x 是外生变量，它们在模型中起到完全不同的作用，经典的回归方法给出不相容的参数估计. 对于线性模型，现在已有了比较完善的统计方法处理这个问题，对于非线性模型，也必须建立相应的统计方法.

第二，数量遗传学中，遗传学家已经建立了远交群体子代表现型 y_2 与亲本表现型“平均值”y_1 之间的模型原型

$$y_2 = m + by_1 + e,$$

其中 e 是误差. 如果我们无法测得亲本表现型的“平均值”，只能得到个别亲本的“实现值”，如何估计模型的参数？现代统计理论已经证明，采用经典回归方法将给出不正确的结果. 这也是建模时必须解决的问题.

森林的生长和演化是一种生物现象，它是林学和生态学研究的重要内容之一. 早在 20 世纪 60 年代已经出现了现代意义下的

数学模型对森林的生长规律的研究，进而指导森林经营工作. 世界林联从 70 年代开始召开独立的有关森林生长和收获模型的国际会议，标志着“森林生长和收获模型”已经被国际承认为一个独立的研究方向. 我国在这方面起步较晚，直到 1989 年，我们研究小组的课题“林分生长动态模型与模拟技术”获得了国家自然科学基金面上项目的资助(符伍儒，唐守正分别第一、第二负责)，这是我国第一个国家级的有关森林生长和收获模型的课题. 在自然科学基金委员会的关心下，我们研究小组又连续获得自然科学基金重点项目“我国主要人工用材林生长模型、经营模型和优化控制”(唐守正负责)和农林倾斜项目“林分生长的地理和种源变异及其模型的研究”(李希菲负责)，从而较全面地开展了森林生长与收获模型方面的研究.

注意到在以往的“森林生长和收获模型”研究中，应用统计学工具的经验和缺陷，从课题的立项开始，我们就把总结和发展统计学工具，作为研究的一个重要内容，以便使模型建立在更坚实的基础上. 课题组经过近十年的努力，在怎样用数学语言建立森林生长与收获模型、怎样解释数学模型、需要一些什么新的数学方法来处理森林生长与收获模型等方面积累了一些经验和得到一些新结果，并且，进一步认识到必须重视生物数学模型的统计学基础研究.

就生物数学模型的统计学基础来说，唐守正首先注意到联立方程组和度量误差模型在生态模型的解释、参数估计、模型耦合及模型整体化等方面的重要性，并猜测到联立方程组、混合模型和度量误差模型之间可能存在某种比 Anderson 曾经发现的更深入的联系，并且认为这一点可能成为生态模型的建模的更好的基础或工具. 刘秀芳教授和李勇副教授带领一些数学研究生完成了非线性度量误差模型极限定理的基础工作和联立方程组与度量误差模型关系研究的基础工作. 李希菲研究员等带领一些林学研究生研究了这些统计方法在森林生长与收获模型方面应用的优缺点. 本书是上述这些课题在统计学基础方面研究的一个总结.

为了内容的系统完整性，本书从一元线性模型开始，逐步引入联立方程组，混合误差模型，度量误差模型以及向非线性模型的推广，讨论了这些统计模型之间的关系以及它们和一些林学数学模型的应用和局限．希望通过这些总结与讨论，有助于理解应用统计方法的“生物数学模型”和“统计模型”的关系与差异，有助于进一步寻找适用于“生物数学模型”的更好的数学工具．

本书包含了一些国内统计专业书籍中没有或很少介绍的内容以及我们在上述研究中得到的一些结果．例如，因子分析的方差类型；组合误差结构模型；带限制的似乎不相关模型；联立方程组模型中含有方程间限制的模型的可识别性；混合模型的参数估计方法；度量误差模型中的广义函数关系、广义结构关系、广义超结构关系模型；度量误差模型和联立方程组的关系；非线性度量误差模型和生物数学模型等等．

本书的例子和数据大多来自上述三个自然科学基金课题，参加上述课题的主要人员还有彭世揆教授、郎奎健教授、宋铁英教授、孟献宇教授、佘光辉、李凤日、杜纪山、王明亮、王雪峰、张春梅等．老一代的数学家严士健教授和林学家关毓秀教授始终关心和指导上述课题的研究和进展．本书的出版得到“华夏英才基金”的资助．对上述单位与个人一并表示感谢．

唐守正　李　勇

1999，11，30

符　号　表

$\mathbf{0}$ 或 $\underset{n\times m}{\mathbf{0}}$	所有(分量)元素为0的 $n\times m$ 矩阵(向量)
$\mathbf{1}$ 或 $\underset{n\times m}{\mathbf{1}}$	所有元素(分量)为1的 $n\times m$ 矩阵(向量)
$\boldsymbol{I}_n$	n 阶单位矩阵
$\underset{n\times m}{\boldsymbol{A}}$ 或 $\boldsymbol{A}_{n\times m}$	$\boldsymbol{A}$ 为 $n\times m$ 矩阵
$\boldsymbol{A}_{i\cdot}$	矩阵 $\boldsymbol{A}$ 的第 i 行向量
$\boldsymbol{A}_{\cdot j}$	矩阵 $\boldsymbol{A}$ 的第 j 列向量
$\boldsymbol{A}>0$	$\boldsymbol{A}$ 为对称正定方阵
$\boldsymbol{A}\geqslant 0$	$\boldsymbol{A}$ 为对称半正定方阵
$\boldsymbol{A}\geqslant\boldsymbol{B}$	$\boldsymbol{A}\geqslant 0,\boldsymbol{B}\geqslant 0$ 且 $\boldsymbol{A}-\boldsymbol{B}\geqslant 0$
$\boldsymbol{A}\leqslant\boldsymbol{B}$	$\boldsymbol{A}\geqslant 0,\boldsymbol{B}\geqslant 0$ 且 $\boldsymbol{B}-\boldsymbol{A}\geqslant 0$
$\boldsymbol{A}'$	矩阵或向量 $\boldsymbol{A}$ 的转置
$\boldsymbol{A}^{+}$	矩阵 $\boldsymbol{A}$ 的加号逆
$\mathrm{diag}(\boldsymbol{a})$	以向量 $\boldsymbol{a}$ 为对角线元素的对角方阵
$\mathrm{diag}(A_1,A_2,\overset{\cdots}{A},A_n)$	主对角线上为 $A_1,A_2,\overset{\cdots}{A},A_n$ 的对角矩阵
$(a_{ij})_{m\times n}$ 或 $\underset{m\times n}{(a_{ij})}$	第 (i,j) 位置的元素为 a_{ij} 的 $m\times n$ 阶矩阵
$\underset{x}{\mathrm{argmin}}f(x)$ 或 $\mathrm{argmin}f(x)$	在函数定义域中使 $f(x)$ 达到极小值的点
$\underset{x\in D}{\mathrm{argmin}}f(x)$ 或 $\underset{D}{\mathrm{argmin}}f(x)$	在区域 D 中使函数 $f(x)$ 达到极小值的点
$\underset{x}{\mathrm{argmax}}f(x)$ 或 $\mathrm{argmax}f(x)$	在函数定义域中使 $f(x)$ 达到极大值的点
$\underset{x\in D}{\mathrm{argmax}}f(x)$ 或 $\underset{D}{\mathrm{argmax}}f(x)$	在区域 D 中使函数 $f(x)$ 达到极大值的点
$\det(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的行列式
$\mathrm{tr}(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的迹
$\mathrm{rk}(\boldsymbol{A})$ 或 $\mathrm{rank}(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的秩
$\vec{\boldsymbol{A}}$	矩阵 $\boldsymbol{A}$ 的按列拉直

$\boldsymbol{A}\otimes\boldsymbol{B}$	矩阵$\boldsymbol{A}$和$\boldsymbol{B}$的叉积
$\boldsymbol{E}(\boldsymbol{x})$	随机变量或随机向量$\boldsymbol{x}$的数学期望
$\boldsymbol{M}(\boldsymbol{x})$	$\boldsymbol{x}$的行向量张成的空间
$\boldsymbol{H}\in\boldsymbol{M}(\boldsymbol{x})$	矩阵$\boldsymbol{H}$的每一个行向量都属于$\boldsymbol{M}(\boldsymbol{x})$
$\mathrm{var}(x),\mathrm{var}(\boldsymbol{x})$	随机变量x的方差,$\boldsymbol{x}$的方差矩阵
$\mathrm{cov}(\boldsymbol{x})$	随机向量$\boldsymbol{x}$的方差矩阵
$\mathrm{cov}(\boldsymbol{x},\boldsymbol{y})$	随机向量$\boldsymbol{x}$与$\boldsymbol{y}$的协方差矩阵
$N(\mu,\boldsymbol{v})$	均值为μ,协方差矩阵为$\boldsymbol{v}$的正态分布
$x\sim N(\mu,\boldsymbol{v})$	x服从均值为μ,协方差矩阵为$\boldsymbol{v}$的正态分布
$F(n,m)$	自由度为(n,m)的F分布
$F_\alpha(n,m)$	自由度为(n,m)的F分布的上α分位点
$x\sim F(n,m)$	x服从自由度为(n,m)的F分布
$t(n)$	自由度为n的学生分布
$x\sim t(n)$	x服从自由度为n的学生分布
$\chi^2(n)$	自由度为n的χ^2分布
$x\sim\sigma^2\chi^2(n)$	$\frac{x}{\sigma^2}$服从自由度为n的χ^2分布
R^n	n维欧几里得空间

目　录

第一章　一元线性模型

建立数学模型有三个基本问题,第一是如何将实际问题转化为数学模型,第二是如何估计模型中的未知参数,第三是如何根据观测数据对已建立的模型做出一些判断,例如某些参数是否相等,某些关系是否成立等.线性模型是各种模型中最基本的一种,由线性模型发展出来的各种方法和技术,可以在一定程度和推广到非线性模型.大量的实际问题也可以由线性模型来描述,或作为初级近似由线性模型近似地描述.模型参数估计方法有多种,例如,实验数据直接估计法,矩法,回归估计法等.这些方法大多可以用线性模型的理论来解决.在一般统计学教科书中的多数经典判断问题,也可以转化为一元线性模型问题.

本章主要介绍独立等方差线性模型(即一元线性模型)的基础理论和如何把统计上的一些经典问题(如回归、方差分析、协方差分析、数量化方法Ⅰ等)转化为一元线性模型问题的方法,并通过例子演示如何利用一元线性模型的理论解决这些经典问题.本章是以后各章的基础.希望这一章能够帮助读者掌握如何应用一元线性模型的基本理论解决实际问题.

1.1　一元线性模型的基本理论

线性模型是 Rao 在 20 世纪 70 年代提出的一种统计模型(Rao,1973),它概括了许多线性常规统计问题,例如,平均数估计,线性回归,方差分析和协方差分析,数量化方法Ⅰ等(董文泉等,1978),因而在各方面获得广泛应用.

一元线性模型是所有线性模型的基础.多元统计分析(唐守正,1986)从应用角度讲述了设计矩阵列满秩时的一元线性模型的

基本理论和应用.为了便于理解,本章第一节介绍一元线性模型的基本概念和一些必要的数学内容,其余各节叙述一元线性模型在一些典型特例下的应用,这些典型特例是在构造模型时经常碰到的情况.

在本书的公式中通常用大写字母表示含有误差的变量观测值(这个误差可能来自于测量误差和抽样误差等),用小写字母表示不含观测误差的变量观测值;用黑体或大写希腊字母表示矩阵,根据需要用黑体或斜体表示向量.

一元线性模型的一般形式为

$$\begin{cases} Y_1 = x_{11}\beta_1 + x_{12}\beta_2 + \cdots + x_{1p}\beta_p + e_1 \\ Y_2 = x_{21}\beta_1 + x_{22}\beta_2 + \cdots + x_{2p}\beta_p + e_2 \\ \cdots\cdots\cdots\cdots \\ Y_n = x_{n1}\beta_1 + x_{n2}\beta_2 + \cdots + x_{np}\beta_p + e_n \end{cases}$$

其中,Y_t 是第 t 个样品因变量的观测值,它是包含误差的观测值;x_{ti}是第 t 个样品第 i 个自变量的值,它不包含观测误差;β_i 是第 i 个未知参数,e_t 是 Y_t 的观测误差,$1\leqslant t\leqslant n$,$1\leqslant i\leqslant p$.上式可写成矩阵形式

$$Y_t = \boldsymbol{x}^{(t)}\boldsymbol{\beta} + e_t, \qquad 1\leqslant t\leqslant n$$

或

$$\boldsymbol{Y} = \boldsymbol{x\beta} + \boldsymbol{e},$$

其中 $\boldsymbol{x}^{(t)} = (x_{t1} \quad x_{t2} \quad \cdots \quad x_{tp})$,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{x} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

为了叙述方便,约定一些在本书中使用的符号.设 $\boldsymbol{A}$ 为一个 $m\times n$ 阶矩阵,当需要强调阶数时,我们把其阶数用小号字写在它的下面,即用 $\underset{m\times n}{\boldsymbol{A}}$ 来表示.例如 $\underset{1\times 5}{\boldsymbol{x}}$ 表示一个 1×5 阶矩阵,或等价

地是一个 5 维行向量. 又如 $\underset{6\times 1}{\boldsymbol{x}}$ 表示一个 6×1 阶矩阵,或等价地是一个 6 维列向量. 对于随机变量(或随机向量)x,$\boldsymbol{E}(x)$表示 x 的数学期望. 对于任何随机变量 x,$\mathrm{var}(x)=\boldsymbol{E}(x-\boldsymbol{E}(x))^2$ 表示 x 的方差. 对于任何随机向量 $\boldsymbol{x}$,$\mathrm{cov}(\boldsymbol{x})$表示 $\boldsymbol{x}$ 的协方差矩阵,简称协方差矩阵. 当 $\boldsymbol{x}$ 为行向量时

$$\mathrm{cov}(\boldsymbol{x})=\boldsymbol{E}((\boldsymbol{x}-\boldsymbol{E}(\boldsymbol{x}))'(\boldsymbol{x}-\boldsymbol{E}(\boldsymbol{x}))),$$

当 $\boldsymbol{x}$ 为列向量时,

$$\mathrm{cov}(\boldsymbol{x})=\boldsymbol{E}((\boldsymbol{x}-\boldsymbol{E}(\boldsymbol{x}))(\boldsymbol{x}-\boldsymbol{E}(\boldsymbol{x}))').$$

假设观测误差 $\boldsymbol{e}$ 的均值为零,即 $\boldsymbol{E}(\boldsymbol{e})=\boldsymbol{0}$;观测误差 $\boldsymbol{e}$ 的协方差矩阵为 $\boldsymbol{\Psi}$,即 $\mathrm{cov}(\boldsymbol{e})=\boldsymbol{\Psi}$. 这样就有下面的模型:

$$\begin{cases}\underset{n\times 1}{\boldsymbol{Y}}=\underset{n\times p}{\boldsymbol{x}}\underset{p\times 1}{\boldsymbol{\beta}}+\underset{n\times 1}{\boldsymbol{e}}\\ \boldsymbol{E}(\boldsymbol{e})=\boldsymbol{0}\\ \mathrm{cov}(\boldsymbol{e})=\boldsymbol{\Psi}\end{cases}. \tag{1.1.1}$$

这里并不假定观测误差 $\boldsymbol{e}$ 的各个分量是互不相关的,因而 $\boldsymbol{\Psi}$ 不一定是对角矩阵. 我们称 $n\times(p+1)$阶矩阵$(\boldsymbol{Y}\quad \boldsymbol{x})$为**资料矩阵**;称 $\boldsymbol{Y}$ 为**观测向量**;称 $\boldsymbol{x}$ 为**设计矩阵**;称 $\boldsymbol{e}$ 为**观测误差**;称 $\boldsymbol{\beta}$ 为**模型参数向量**,简称为**参数向量**或**参数**;称 $\boldsymbol{\Psi}$ 为误差 - 协方差矩阵(variance - covariance of error 或 dispersion),简称为**误差矩阵**.

对于误差矩阵 $\boldsymbol{\Psi}$ 的不同假设,可以得到各种不同的线性模型. 本章所要讨论的是最基本的情形 $\boldsymbol{\Psi}=\sigma^2\boldsymbol{I}_n$(本书中 $\boldsymbol{I}_n$ 表示 n 阶单位矩阵),即

$$\begin{cases}\underset{n\times 1}{\boldsymbol{Y}}=\underset{n\times p}{\boldsymbol{x}}\underset{p\times 1}{\boldsymbol{\beta}}+\underset{n\times 1}{\boldsymbol{e}}\\ \boldsymbol{E}(\boldsymbol{e})=\boldsymbol{0}\\ \mathrm{cov}(\boldsymbol{e})=\sigma^2\boldsymbol{I}_n\end{cases}. \tag{1.1.2}$$

此时模型(1.1.2)称为**一元线性模型**. 如果再假定 $\boldsymbol{e}$ 服从正态分布,则称模型(1.1.2)为**一元正态线性模型**. 本章中参数估计部分的结果对一元线性模型都是正确的,但假设检验和统计推断只对一元正态线性模型成立. 对于非正态一元线性模型,当样本观测的数目 n 比较大时,本节中这些假设检验和统计推断的结果也近似