生物多元分析

张 焱 编著

SHENGWU
DUOYUAN
FENXI

西南师范大学出版社

生化、农林、医卫类专业用书

生物多元分析

张　焱　编著

西南师范大学出版社
1999 年

特约编辑：赵宏量
责任编辑：米加德
封面设计：王　煤

生物多元分析

张　焱　编著

西南师范大学出版社出版、发行
（重庆　北碚）
重庆北碚培萃印刷厂印刷
开本：850×1168 1/32 印张：8.875 字数：250千
1999年4月第1版　1999年4月第1次印刷
印数：1～1000
ISBN 7-5621-2042-0/Q・16

定价：11.80元

序

《生物多元分析》出版了，作为长期从事多元分析及其应用研究的科研工作者，看到多元分析在生物、农林、医卫等方面扩展应用的专著出版，倍感亲切，浮想联翩，被邀作序，也深感快慰与荣幸.

多元分析又称多元统计分析，是国内近二十年来最活跃也最重要的学科分支之一，应用面极为广泛. 因为各种事物都要受很多因素条件(多元)的制约，并影响其现状和发展. 要研究它们就需要多元分析这样的理论、方法和技术来处理有关的多维总体. 但现代的信息十分丰富，多维数据庞杂，多元分析计算繁难，如果不是微型计算机的出现和迅速发展，多元分析的计算和应用简直是寸步难行. 这也是为什么在计算机未大量使用前，多元分析与应用不能得到很好发展的原因. 以石油勘探为例，一般人以为这只与打钻找油有关，殊不知它需要极为复杂和庞大的计算，需要多元分析等多种方法技术的支持，故 1973 年中国当时的最大型计算机(每秒百万次)首先是为石油勘探部门所建造和使用的. 其计算量的巨大，使人难以想象，由于笔者 1973 年至 1974 年亲身参加过这个工作的应用研究，所以才有以上的体验. 实际在气象数值预报的问题上，也有类似的情况.

多元分析作为应用性极强的学科分支，促进了其它边缘学科的发展. 以数学地质为例，多元分析等数学方法在地质科学中的深入应用，使数学地质这一新兴的边缘学科分支迅速发展和成长. 中国知名学者、教授赵鹏大、李裕伟、王学仁、王世称和笔者等的大量应用研究工作，在国际数学地质杂志上曾有专文介绍. 因此，国际

数学地质协会(IAMG)在1981年主动吸收了这些人成为该会最早的中国会员. 笔者在美国 Wisconsin 大学统计系和地质系做访问研究时,他们也对中国数学地质方面的研究工作印象深刻. 中国数学地质在多元分析等数学方法的支持下,得到了充分迅速的发展. 在武汉、成都和长春都建立了数学地质专业的博士点.

多元分析本身二十多年来,在中国蓬勃壮大与发展,也是笔者亲眼目睹和身历其过程的. 1977年,中国科学院应用数学研究所方开泰同志所倡导和组织的多元分析全国学术研讨会在美丽的黄山召开,这是文革后召开的首届有关多元分析的会议,我国很多知名学者如陈希孺、成平、赵鹏大、张尧庭等都参加了. 会上报告内容丰富,讨论热烈,气氛活泼自由,其情其景,令人难忘,至今犹历历在目. 会上决定多元分析学术会今后每二年定期在全国各地举行,由中科院应用数学研究所方开泰等负责总的筹划;随后中国科学院计算中心研究员杨自强、张建中等也组织发起了有关多元分析计算的全国会议,同样也定期不断地召开和延续. 在主任委员方开泰的影响和倡导下,中国现场统计研究会多元分析委员会也决定召开全国多元分析应用方面的会议. 因方开泰经常在国外讲学,国内便由我们组织召开了五届全国会议,分别在成都、青岛、烟台、福州等地先后举行. 一时间包括多元分析的理论、计算、应用各个方面的学术讨论会相继举行,此起彼伏,十分活跃. 在此基础上,方开泰教授又与国际知名的多元分析权威学者共同发起和组织了在香港召开的国际多元分析及应用的学术会议(1992),会议规模盛大,国际知名的多元分析专家纷纷参加,国内学者出席的也有近20人. 随着学术活动的蓬勃开展,多元分析为中国的社会主义建设作出了相应的贡献,如在我国资源预测,石油开发,矿产勘探,地球化学找矿等有关数学地质的丰富和卓有成效的应用成果中,多元分析功不可没. 我们只从这一个小小侧面,充分说明和反映了邓小平同志所提出的"科学技术是第一生产力"的论断是非常正确的. 改

革开放的20年来，在邓小平理论的指引下，多元分析工作者亲身感受到科教兴国的威力，也感受到科学春天的气息与温暖.

多元分析不仅在现代科技、工矿、地质、气象、工程等方面取得了不少卓有成效的成果，它还一步一步地扩展到农林、生物、医卫甚至社会经济、管理、人口、教育等各个研究领域，论文很多，有关的多元分析书籍也不少. 80年代初期出版的专著，如张尧庭、方开泰合写的《多元统计分析引论》(科学出版社，1982). 这本书作为多元分析数学理论基础为广大多元分析工作者所推崇和引用，拙著《多元分析》(北京地质出版社，1982)则作为高等学校试用教材供数学地质研究生学习和地质科学专业师生的参考书，但这两本书数学理论较多，偏深偏难. 对实际应用的工作者来说，王学仁编著的《地质数据的多变量统计分析》(科学出版社，1982)，则结合实际应用较好，易为实际应用工作者所接受，但稍偏于地质. 而与生物、农林、医卫相结合的多元分析书，却一直未能见到. 张焱编著的《生物多元分析》的出版填补了多元分析应用的一大空白，因为生物科学是与人类生存和发展密切相关的前沿学科，所以，本书出版的意义之重要也就不言自明了！张焱教授长期从事多元分析和软科学的应用研究，先后主研了"卧龙自然保护区大熊猫环境生态系统研究"等重大科研项目，撰写过《环境因子对冷箭竹开花影响的数学模型》等30多篇论文，获得过四川省及国家土地管理局科技进步二等奖2项、优秀成果一等奖1项，其应用研究成果是很显著的. 他这样的专家来编著生物多元分析，当然是驾轻就熟，众望所归！

本书着重阐明多元统计分析的基本思路，充分说明方法的实质，强调其在生物和生命科学中的实际应用，使应用者能易于了解和掌握多元分析方法的本质及其解法格式和典型范例，并易于应用到研究相关规律、数值分类、模式识别和预测预报的实际中去，本书还选编有多元分析计算程序，提倡使用相应的计算机软件，以减少实际工作者在运用多元分析复杂计算时遇到的困难.

本书行文流畅，脉络清晰，深入浅出，论述科学严谨，逻辑性强，是一本适用于研究生、本科生等不同层次读者的优秀著作；同时本书具有很强的实用性及应用面的广泛性，故对农林、医卫、化学、生物专业的实际工作者也是一本不可多得的工具书. 笔者在此不揣冒昧地欣然为之作序，并向读者推荐这本好书. 同时也希望广大读者多多指教，使多元分析在生物科学中的应用，能够进一步地丰富与发展，在教学和实际研究中，开好花，结美果！

王柏钧

1998 年 8 月 18 日于成都气象学院

编者的话

在生物科学的发展过程中，由于研究对象的复杂性，研究工作曾长期停留在观察和描述的阶段. 现代科学技术的发展，边缘学科、交叉学科的相继产生，特别是数学和电子计算机技术在生物科学研究中日益广泛的应用，大大地促进了生物科学研究的定量化、动态化和模型化. 人们比较熟悉的生物统计，就是出现最早、应用最广的一门生物学与数学之间的交叉学科. 80 年代以来，我国大学生物、农学和医学等专业已将它列为必修课程. 随着微型计算机的逐步普及，多元统计分析也在生物科学的研究中显示出特殊的威力，使一些用常规生物统计不能解决的问题，变得迎刃而解了. 实践证明，多元统计分析是生物科学研究中一个极为有效的工具.

国内介绍多元分析的书籍，虽然这十多年来已陆续出版了好些种，但这些书基本上还是偏重于理论，难度较大，实用性不强. 目前，还没有一本专门为生物科学研究人员编写的多元分析书籍. 为适应生物、农学和医学对多元统计分析的迫切需要，我们根据多年为生物系研究生讲授多元分析课所搜集的资料和积累的经验，编写了这本以实际应用为主，又兼顾理论脉络的《生物多元分析》一书，它既可作大学生物、农学和医学等专业研究生和高年级本科生的教材与教学参考书，也可作为生物、农学和医学科研工作者和实际工作者的工具书.

本书针对生物科学中研究相关规律、数值分类、模式识别和预测预报的需要，分章介绍了多元回归分析、判别分析、聚类分析、主成分分析、因子分析、典型相关分析和对应分析等多元统计分析方法. 着重阐明这些方法的数学模型、解决问题的基本思想方法和解

法程序，并尽可能地选编了一些富有启发性的典型生物学实例；对于多元分析的理论则只介绍其脉络和系统，一般不加以证明，对理论有兴趣的读者可以进一步阅读多元分析的专著.

虽然多元统计分析的计算有一套巧妙的数学方法，但毕竟较难，而且计算工作量很大，必须使用计算机方能顺利完成. 由于目前国内尚缺流行的汉化多元统计分析软件，因此，本书还选编了多元回归分析、主成分分析和典型相关分析等最常用的多元分析计算机软件程序，并配有相应的运行实例，供读者学习和使用多元分析软件时参考. 这无疑会给实际研究工作带来很大的方便.

本书在编写和试用过程中，得到了四川师范学院教务处、研究生处、数学系和生物系的热情关心和支持；特别是在这次出版过程中，得到了我的母校——西南师范大学出版社的亲切关怀和大力支持，在此，谨一并表示衷心感谢.

本书在编写中，编者学习、参考和使用了一些作者的宝贵资料，在此，谨向这些作者致以敬意、歉意和衷心感谢.

由于水平所限，本书谬误处一定有之，敬请读者批评指正.

编者

1998 年 8 月

目　录

第一章　多元回归分析

回归分析(Regression Analysis)是研究随机现象中变量之间相关关系的一种基本的统计分析方法，在生物、医学、气象、地质、地震、工农业生产以及经济领域的定量预测中，应用非常广泛.

多元回归，过去在《生物统计》中是指一个因变量(预报对象、响应变量)、多个自变量(预报因子)的(一对多)回归模型，本章所讲的多元回归则是指的多个因变量、多个自变量的(多对多)回归模型，为了加以区别，我们把前一类"一对多"回归称为一元(多自变量)回归，后一类"多对多"回归称为多元回归. 这种多元回归与一元回归有很多相似之处，因此，本章从一元回归模型的回顾入手，对应地介绍多元回归模型及其应用，同时，介绍近代回归分析中的一些新结果.

§1.1　一元多自变量回归分析

设随机变量 y 与 m 个自变量(非随机变量)$x_1, x_2, \cdots, x_m$ 有线性关系，于是有 y 的线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \qquad (1.1.1)$$

其中 ε 是随机误差，服从中心正态分布 $N(0, \sigma^2)$，$\beta_0, \beta_1, \cdots, \beta_m$ 为模型参数.

回归分析的主要任务是：

(1) 由观测值 $(x_{t1}, x_{t2}, \cdots, x_{tm}; y_t)(t = 1, 2, \cdots, n)$ 确定未知参数 $\beta_0, \beta_1, \cdots, \beta_m$ 的估计值 —— 回归系数 $b_0, b_1, \cdots, b_m$，从而得出 y 关于 $x_1, x_2, \cdots, x_m$ 的线性回归方程：

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_m x_m .$$

其中$\hat{y}$表示 y 的回归估计值.

(2) 对线性关系及各自变量的回归贡献进行显著性检验.

(3) 利用回归方程进行预报或控制.

一、建立回归方程

设对 y 及 $x_1, x_2, \cdots, x_m$ 作 n 次独立观测,得

$$(x_{t1}, x_{t2}, \cdots, x_{tm}; y_t) \quad (t = 1, 2, \cdots, n),$$

将观测值代入(1.1.1)式,得

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_m x_{tm} + \varepsilon_t \quad (t = 1, 2, \cdots, n) \tag{1.1.2}$$

其中诸 ε_t 相互独立且均服从中心正态分布 $N(0, \sigma^2)$.

记

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}_{n\times(m+1)},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}_{(m+1)\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1},$$

则(1.1.2)有矩阵形式

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.1.3}$$

其中 $\boldsymbol{X}$ 称为设计矩阵,是非随机的.一般总是假设 $\operatorname{rank}(\boldsymbol{X}) = m + 1 < n$,使得 $\boldsymbol{X}'\boldsymbol{X}$ 为满秩矩阵.

1. $\boldsymbol{\beta}$ 的最小二乘估计

设
$$Q = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \tag{1.1.4}$$

称 Q 为误差平方和,使 Q 达到最小值的$\hat{\boldsymbol{\beta}} = (b_0, b_1, \cdots, b_m)'$ 称为 $\boldsymbol{\beta}$ 的最小二乘估计.

由于 $\partial Q / \partial \boldsymbol{\beta} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$

所以 $\hat{\beta}$ 应满足 $X'X\hat{\beta}=X'y$ (1.1.5)

(1.1.5) 称为正规方程.

在 $S=X'X$ 满秩的假定下,(1.1.5) 有唯一解:

$$\hat{\beta}=(X'X)^{-1}X'y \tag{1.1.6}$$

它就是 β 的最小二乘估计. Q 的最小值记为

$$Q_e=(y-X\hat{\beta})'(y-X\hat{\beta}) \tag{1.1.7}$$

称为剩余(或残差)平方和,记 $y-X\hat{\beta}=e$,称为剩余(或残差)向量.则

$$Q_e=e'e.$$

方程 $\hat{y}=X\hat{\beta}$ 称为 y 关于 $x_1,x_2,\cdots,x_m$ 的线性回归方程,则 $e=y-\hat{y}$. 关于 $\hat{\beta}$ 的性质,有下述定理.

定理 1.1.1 (高斯—马尔科夫定理)设线性模型:

$y=X\beta+\varepsilon$

$E(\varepsilon)=0,\quad \operatorname{cov}(\varepsilon,\varepsilon)=\sigma^2 I_n$

则 $\hat{\beta}=(XX)^{-1}X'y$ 具有性质:

(1) $E(\hat{\beta})=\beta,\quad \operatorname{cov}(\hat{\beta},\hat{\beta})=\sigma^2(X'X)^{-1}$

(2) 设 $T=cy$ 为 β 的任意线性无偏估计,则

$$\operatorname{Cov}(T,T)=\operatorname{Cov}(\hat{\beta},\hat{\beta})+E[(T-\hat{\beta})(T-\hat{\beta})'],$$

即 $\hat{\beta}$ 为 β 的最优线性无偏估计,简记为 BLUE.

2. 回归方程的建立

在实际问题中,求回归系数 β 的最小二乘估计很少直接使用正规方程(1.1.5),而是从中先消去

$$b_0=\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_1-\cdots-\hat{\beta}_m\bar{x}_m \tag{1.1.8}$$

得到仅有 m 个未知数的线性方程组 $\quad Lb=B$ (1.1.9)

其中 $\boldsymbol{L}=(l_{ij})_{m\times m}$,

$$\boldsymbol{b}=(\hat{\beta}_1,\cdots,\hat{\beta}_m)'=(b_1,b_2,\cdots,b_m)',$$

$$\boldsymbol{B}=(l_{1y},l_{2y},\cdots l_{my})',$$

$$l_{ij}=\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)=\sum_{t=1}^{n}x_{ti}x_{tj}-n\bar{x}_i\bar{x}_j,$$

$$\bar{x}_i=\frac{1}{n}\sum_{t=1}^{n}x_{ti},\bar{y}=\frac{1}{n}\sum_{t=1}^{n}y_t,$$

$$l_{jy}=\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)(y_t-\bar{y})$$

$$=\sum_{t=1}^{n}x_{tj}y_t-n\bar{x}_j\bar{y}\quad(i,j=1,2,\cdots,m).$$

从(1.1.9)可解得

$$\boldsymbol{b}=L^{-1}\boldsymbol{B}=(b_1,b_2,\cdots,b_m)',\tag{1.1.10}$$

而 $b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2-\cdots-b_m\bar{x}_m$,

于是得回归方程

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$$

和回归估计值

$$\hat{y}_t=\bar{y}+b_1(x_{t1}-\bar{x}_1)+b_2(x_{t2}-\bar{x}_2)+\cdots+b_m(x_{tm}-\bar{x}_m)$$

$$(t=1,2,\cdots,n,\cdots)\tag{1.1.11}$$

在逐步回归中,常将离差积和 l_{ij},l_{jy} 改为相关系数

$$r_{ij}=l_{ij}/(l_{ii}l_{jj})^{1/2}$$

$$r_{jy}=l_{jy}/(l_{jj}l_{yy})^{1/2}\quad(i,j=1,2,\cdots,m)\tag{1.1.12}$$

$$l_{yy}=\sum_{i=1}^{n}(y_t-\bar{y})^2=\sum_{t=1}^{n}y_t^2-n\bar{y}^2,$$

则方程组 $\boldsymbol{Lb}=\boldsymbol{B}$ 变为 $\boldsymbol{Rb}^*=\boldsymbol{r}$ (1.1.13)

其中 $\boldsymbol{R}=(r_{ij})_{m\times m}$ 称为样本相关矩阵

$$\boldsymbol{b}^*=(b_1^*,b_2^*,\cdots,b_m^*)'\tag{1.1.14}$$

$$\boldsymbol{r}=(r_{1y},\cdots,r_{my})'$$

解方程组(1.1.13)得　$\boldsymbol{b}^* = \boldsymbol{R}^{-1}\boldsymbol{r}$　　(1.1.15)

即

$$b_j^* = b_j(l_{jj}/l_{yy})^{1/2}$$

或

$$b_j = b_j^*(l_{yy}/l_{jj})^{1/2} \quad (j = 1,2,\cdots,m) \qquad (1.1.16)$$

$b_0^* = 0, b_1^*, b_2^*, \cdots, b_m^*$ 称为标准化回归系数，

$\hat{y}' = b_1^* x'_1 + b_2^* x'_2 + \cdots + b_m^* x'_m$ 称为 y 关于 $x_1, x_2, \cdots, x_m$ 的标准化回归方程(其中 $x'_1, x'_2, \cdots, x'_m, y'$ 均为相应变量的标准化). r_{jy} 反映了 y 同 x_j 之间的线性相关程度. 因此用(1.1.15)式求解 b^*，对于考察各自变量 x_j 同因变量 y 之间的线性相关程度比较清楚，但比用(1.1.10)式的计算量要大，通常是用计算机来进行计算.

例 1.1.1　下表 1.1.1 列出的是 1978 年～1985 年 8 年间我国的粮食总产量(y)、粮食单位面积产量(x_1)、农业机械总动力(x_2)、化肥施用量(x_3)等的统计数据，试建立我国的粮食总产量 y 关于粮食单位面积产量 x_1、农业机械总动力 x_2 和化肥施用量 x_3 的线性回归方程.

表 1.1.1

指标 / 年份(t)	粮食单位面积产量 x_1(公斤/公顷)	农机总动力 x_2(万焦耳/秒)	化肥施用量 x_3(万吨)	粮食总产量 y(万吨)
1978(1)	2 535	11 917 350	884	30 477
1979(2)	2 835	13 570 486	1 086.3	33 212
1980(3)	2 745	14 956 554	1 269.4	32 056
1981(4)	3 835	15 903 974	1 334.9	32 502
1982(5)	3 135	16 851 394	1 513.4	35 450
1983(6)	3 405	18 279 238	1 659.8	38 728
1984(7)	3 615	19 775 714	1 739.8	40 731
1985(8)	3 480	21 211 018	1 775.8	37 911