普通高等教育“十三五”规划教材／教辅

应用多元统计分析
——基于R的实验

韩明 编著

同济大学出版社
TONGJI UNIVERSITY PRESS

应用多元统计分析

——基于R的实验

韩 明 编著

同济大学出版社
TONGJI UNIVERSITY PRESS

内容提要

本书基于《应用多元统计分析》(第2版)(韩明,同济大学出版社)的内容,编写了基于R的实验.在每一章(从第2章开始)的前面,首先按照原教材简要介绍有关概念、理论和相关背景,然后是与本章内容对应的实验.全书由12章组成,通过40个实验,着重培养学生的动手能力、应用R软件分析和解决多元统计问题的能力.实验的内容与原教材的例题、应用案例不重复.本书既可以与原教材配套使用,也可以单独使用.

本书注重可读性,图文并茂,可供高等院校有关专业本科生和研究生作为"多元统计分析""多元统计实验"等课程的教材(或参考书),也可作为全国大学生(研究生)"数学建模竞赛"、全国大学生"统计建模大赛"的培训教材(或参考书),还可供相关专业的教师和科技人员、广大自学者参考.

图书在版编目(CIP)数据

应用多元统计分析:基于R的实验/韩明编著.—
上海:同济大学出版社,2019.8
ISBN 978-7-5608-8563-6

Ⅰ.①应… Ⅱ.①韩… Ⅲ.①多元分析—统计分析—高等学校—教材 Ⅳ.①O212.4

中国版本图书馆CIP数据核字(2019)第115682号

应用多元统计分析——基于R的实验
韩 明 **编著**
责任编辑 张 莉 **助理编辑** 任学敏 **责任校对** 徐春莲 **封面设计** 潘向蓁

出版发行 同济大学出版社 www.tongjipress.com.cn
(地址:上海市四平路1239号 邮编:200092 电话:021-65985622)
经 销 全国各地新华书店
印 刷 大丰科星印刷有限责任公司
排 版 南京月叶图文制作有限公司
开 本 710 mm×960 mm 1/16
印 张 16.5
字 数 330 000
版 次 2019年8月第1版 2019年8月第1次印刷
书 号 ISBN 978-7-5608-8563-6

定 价 42.00元

前　言

随着大数据、人工智能在日常生活中的渗透，学习多元统计分析的人越来越多.“多元统计分析”课程已经被越来越多高校列为相关专业的必修课或选修课.特别是随着相关软件的普及，人们不再只满足于学习一些理论知识，而是将多元统计分析作为工具，借助计算机和相关软件进行数据处理和分析.

多元统计分析是统计学中应用性很强的一个分支，它的应用范围十分广泛.在“多元统计分析”“多元统计实验”课程的教与学过程中，主要难点是涉及的理论比较抽象、计算比较复杂(需要借助有关软件在计算机上实现).

作者根据多年来的教学实践，深感内容简练但又实用的“多元统计分析”“多元统计实验”教材的重要性.作者认为，对于侧重于“应用”多元统计方法进行数据处理和分析的读者，重点不在于理解多元统计方法的理论证明和公式推导，而是要应用有关软件对数据进行分析，特别是要理解多元统计方法的目的、应用条件和结果的解释.本书注重可读性，图文并茂(配图 76 幅)；自第 2 章开始，每一章首先按照原教材简要介绍本章的有关概念、理论和相关背景，然后是与本章内容对应的实验.

考虑到作为一款免费软件，R 软件具有丰富的资源、良好的扩展性和完备的辅助系统，本书的实验采用 R 软件，并给出了相应的代码.通过 40 个实验，突出 R 软件的应用；着重培养学生的动手能力、应用 R 软件分析和解决多元统计问题的能力.书名确定为《应用多元统计分析 —— 基于 R 的实验》，主要是突出 R 软件在多元统计分析中的作用.

感谢王家宝教授在作者写作本书过程中的指导和鼓励.感谢使用《应用多元统计分析》第 1—2 版、《应用多元统计分析 —— 基于 R 的实验》的读者.愿本书的出版能对广大师生在“多元统计分析”“多元统计实验”课程的教与学的过程中有所帮助.

本书参考了一些国内外文献，在此向有关作者表示感谢.虽然作者努力使本书写成一本既有特色又便于教学(或自学)的教材(或参考书)，但由于水平所限，书中难免还存在一些疏漏甚至是错误，恳请专家和读者批评指正.

韩　明

2019 年 3 月

目　录

1 绪 论

多元统计分析(Multivariate Statistical Analysis)是应用统计方法来研究多变量(多指标)问题的理论和方法,它是统计学的一个重要分支.

在实际问题中,受多个变量共同作用和影响的现象大量存在.当变量较多时,变量之间不可避免地存在相关性.我们常需要处理多个变量的观测数据,那么如何对多个变量的观测数据进行有效的分析和研究呢?如果把多个变量分开处理不仅会丢失一些信息,往往也不容易取得好的研究结论.多元统计分析,通过对多个变量的观测数据的分析,来研究这些变量之间的相互关系以及揭示这些变量内在的变化规律.

1.1 多元统计分析概述

早在19世纪就出现了处理二维正态总体的一些方法,但系统地处理多维概率分布总体的统计分析问题则开始于20世纪.多元统计分析起源于20世纪初,1928年Wishart发表的论文《多元正态总体样本协方差阵的精确分布》,可以说是多元统计分析的开端.之后Fisher, Hotelling, Roy,许宝禄等人作出了一系列奠基性的工作,使多元统计分析在理论上得到迅速的发展.

20世纪40年代,多元统计分析在心理、教育、生物等方面有不少的应用,但由于计算量大,其发展受到影响.20世纪50年代,随着计算机的出现和发展,多元统计分析在地质、医学、气象、社会学等方面得到了广泛的应用.20世纪60年代,通过应用和实践又完善和发展了理论,由于新理论和新方法的不断出现又促使它的应用范围更加扩大.20世纪70—80年代,在我国才受到各个领域的极大关注,近40年来,我国在多元统计分析的理论和应用上取得了许多显著的成绩.

进入21世纪后,人们获得的数据正以前所未有的速度迅速增加,产生了海量数据、大数据、超大型数据库等,遍及超级市场销售、银行存款、天文学、粒子物理、化学、医学、生物学以及政府统计等领域,多元统计分析与人工智能、数据库技术等

相结合，已经在经济、商业、金融、天文、地理、农业、工业等方面取得了成功的应用.

“多元统计分析”也称为“多元分析”(Multivariate Analysis).例如 Mardia et al.(1979)的书，书名为 *Multivariate Analysis*.英国著名的统计学家 Kendall 在《多元分析》一书中，把多元统计分析所研究的内容和方法概括为以下几个方面：

(1) 简化数据结构(降维问题)

简化数据结构就是将某些复杂的数据结构通过变量变换等方法，使相互依赖的变量变成互不相关的，或把高维空间的数据投影到低维空间，使问题得到简化而损失的信息又不太多.例如，主成分分析、因子分析、对应分析等就是这样的一类方法.

(2) 分类与判别(归类问题)

归类问题就是对所考察的观测点(或变量)按照相近程度进行分类(或归类).例如，聚类分析、判别分析等就是解决这类问题的统计方法.

(3) 变量间的相互联系

相互依赖关系：分析一个或几个变量的变化是否依赖于另外一些变量的变化？如果是，建立变量之间的定量关系式，并用于预测或控制——回归分析.

变量之间的相互关系：分析两组变量之间的相互关系——典型相关分析.

(4) 多元数据的统计推断

这是关于参数估计和假设检验的问题.特别是多元正态分布的均值向量和协方差矩阵的估计和假设检验等问题.

(5) 多元统计分析的理论基础

多元统计分析的理论基础包括多维随机向量(特别是多维正态随机向量)，以及由此定义的各种多元统计量，推导它们的分布并研究其性质，研究它们的抽样分布理论.

1.2 多元统计分析的应用

多元统计分析是统计学中应用性很强的一个分支，它的应用范围十分广泛.多元统计分析可以应用于几乎所有的领域，主要包括经济学、农业、地质学、医学、工业、气象学、金融、精算、物理学、地理学、军事科学、文学、法律、环境科学、考古学、体育科学、遗传学、教育学、生物学、管理科学、水文学等，还有一些交叉学科或方向等.多元统计分析的应用实在是难以一一罗列，以下简要地介绍一下多元统计分析在文学、数据挖掘(作为交叉学科或方向的代表)领域的应用.

在**文学**方面，自从 20 世纪 30 年代末，英国著名的统计学家 Yule 把统计方法

引入到文学词汇的研究以来，这个领域已经取得了不少进展，其中最有名的是Mosteller与Wallace在20世纪60年代初对美国立国三大文献之一的《联邦主义者文集》的研究.

在1985至1986年复旦大学李贤平教授对我国名著《红楼梦》的原著者进行了研究.使用的统计方法主要是多元统计分析.先选定数十个与情节无关的虚词作为变量，把《红楼梦》一书中的120回作为120个样品，统计每一回(即每个样品)中选定的这些虚词(即变量)出现的频数.由此得到的数据矩阵作为分析的依据.

在《红楼梦》原著者的研究中使用较多的是聚类分析、主成分分析、典型相关分析等方法，由分析结果可以看出：

(1) 前80回和后40回截然地分为两类，证实了前80回和后40回不是出于一个人的手笔；

(2) 前80回是否为曹雪芹所写？通过曹雪芹的另一著作，做类似的分析，结果证实了用词手法完全相同，断定为曹雪芹一人手笔；

(3) 而后40回是否为高鹗写的？分析结果发现，后40回依回目的先后可分为几类，得出的结论推翻了后40回是高鹗一人所写.后40回的成书比较复杂，既有残稿也有外人笔墨，不是高鹗一人所续.

以上这些论证在红学界引起了轰动.他们用多元统计分析方法提出了关于《红楼梦》作者和成书过程的新学说.

李贤平教授等还把这类方法用于其他作家和作品，结果证明统计方法的分辨能力是很强的.

在**数据挖掘**方面，随着科学技术的发展，利用数据库技术来存储、管理数据，利用机器学习的方法来分析数据，从而挖掘出大量的隐藏在数据背后的知识，这种思想的结合形成了深受人们关注的非常热门的研究领域：数据库中的知识发现(knowledge discovery in databases)。数据挖掘(data mining)技术便是其中的一个最为关键的环节.数据挖掘、机器学习(machine learning)等为统计学(包括“多元统计分析”)提供了一个新的应用领域，同时也提出了很多挑战.多元统计分析中的聚类分析(cluster analysis)是按照某种相近程度，将用户数据分成一系列有意义的集合，例如在金融领域中，将贷款对象分为低风险和高风险等.数据挖掘是一个交叉学科，它涉及数据库、人工智能、统计学、并行计算等不同学科和领域，近年来受到各界的广泛关注.应该指出，Johnson & Wichern在*Applied Multivariate Statistical Analysis*(6th ed. 2007)中补充了“数据挖掘”部分，以及多元统计分析方法在数据挖掘中的应用.数据挖掘与统计学有着密切的关系，那么统计学如何为数据挖掘服务呢？这是在“数据挖掘”飞速发展的今天统计学必须回答的一个问题.令人高兴的是，现在可以从统计学在数据挖掘领域里的研究与应用情况看到对

这个问题的各种回答.数据挖掘对统计学带来的挑战,无疑将推动统计学的发展(韩明,2001).关于统计分析与数据挖掘,感兴趣的读者可参考相关文献(薛薇,2014)等.

1.3 本书的基本框架和内容安排

随着大数据、人工智能在我们日常生活的渗透,学习多元统计分析的人越来越多."多元统计分析"课程已经被越来越多高校列为相关专业的必修课或选修课.《多元统计分析》教材的特点各有不同,有的教材侧重理论的讲述,读者需要具备较深厚的数学基础;有的教材则注重模型的应用,理论和技术细节不是重点.作者认为,对于侧重"应用"多元统计方法进行数据处理和分析的读者,重点不在于理解多元统计方法的理论证明和公式推导,而是要应用有关软件对数据进行分析,特别是要理解多元统计方法的目的、应用条件和结果的解释.

多元统计分析通常涉及较为复杂的理论,计算繁琐.大多数多元统计方法几乎无法手工计算,必须借助计算机和有关软件来实现.相关软件的种类很多,有些功能齐全,有些价格便宜,有些容易操作,有些需要更多的实践才能掌握.这里就不一一罗列了.其实,读者只要学会使用一种软件,使用其他的软件也不会困难,看看帮助和说明即可.学习软件的最好方式是在使用中学.

R软件是完全免费的、由志愿者管理的软件,其编程语言与S-plus所基于的S语言一样,很方便.在网站(http://cran.r-project.org/bin/windows/base)上可免费下载R软件的Windows版(当然也可以免费下载R软件的其他版本,如UNIX、LINUX、MacOS),点击"Download R3.5.1 for Windows"下载(注:作者在写作本书后期时的最新版本为R3.5.1),按照提示安装即可.还有不断加入的从事各个方向研究者编写的软件包和程序.在这个意义上可以说,其函数的数量和更新远远超过其他软件.它的所有计算过程和代码都是公开的,它的函数还可以被用户按需要改写.它的语言结构和C++、Fortran、MATLAB、Pascal、Basic等很相似,容易举一反三.对于一般非统计工作者来说,主要问题是它没有"傻瓜化".

考虑到作为一款免费软件,R软件具有丰富的资源(涵盖了多种行业数据分析中几乎所有的方法),良好的扩展性(方便的编写函数和程序包,可以胜任复杂数据的分析、精美图形的绘制),完备的帮助系统(每个函数都有统一格式的帮助).本书的实验均采用R软件,并给出了相应的代码.

近几年来有关R语言/软件与统计分析相结合的书越来越多,代表性的有:Clark(2007),薛毅(2007),汤银才(2008),Kabacoff(2013),Tsay(2013),James

et al.(2013),吴喜之(2013),薛薇(2014),陈景祥(2014),韩明(2017)等.

本书按照《应用多元统计分析》(第 2 版)(韩明,同济大学出版社)的内容(有修改),编写了基于 R 的实验.全书由 12 章组成,在每一章(从第 2 章开始)的前面,首先按照原教材简要介绍本章的有关概念、理论和相关背景,然后是与本章内容对应的实验.本书注重可读性,图文并茂(配图 76 幅);通过 40 个实验,突出 R 软件的应用,着重培养学生的动手能力、应用 R 软件分析和解决多元统计问题的能力.书名为《应用多元统计分析——基于 R 的实验》,主要是突出 R 软件在多元统计分析中的应用.

1.4 用于实验的数据集

本书中用于实验的数据集(按照在本书中首次出现的先后顺序)见表 1-1.

表 1-1 用于实验的数据集

数据集名称	所在的章节(或实验编号)	数据集名称	所在的章节(或实验编号)
mtcars	2.3.1, 2.3.3, 3.3.4	cholesterol	6.4.1
iris	2.3.2, 2.3.4, 7.4.1, 8.4.1	UScereal	6.4.4
women	3.3.1	USJudgeRatings	9.4.2
Boston	3.3.2	ability.cov	10.6.1
state.x77	3.3.3, 4.4.3	Harman74	10.6.2
stackloss	4.4.1, 4.4.2, 4.4.4	caith	11.3.4
USPop	5.3.5	smoke	11.3.5

在表 1-1 中所列数据集中,一部分包含在 R 的基础包(成功启动 R 意味着基础包的默认加载包已经成功加载到 R 的工作空间,用户可以直接调用),用函数“data( )”可以查询基础包中的数据集(Data sets in package ‘datasets’)名称(列表).除基础包外,其他包的数据集需要加载后才能调用.另外,本书用于实验的数据,除表 1-1 所列数据集外,还有一些数据需要导入(详见后面各章中的实验).

2 多元数据的表示及可视化

翻开报纸，打开电视或上网络浏览，就可以看到各种数据.比如高速公路通车里程、物价指数、股票行情、外汇牌价、犯罪率、房价、流行病的有关数据；当然还有国家统计局定期发布的各种国家经济数据、海关发布的进出口贸易数据等.从这些数据中，各有关方面可以提取对自己有用的信息.

某些企业每年都要花数目可观的经费来收集和分析数据.他们调查其产品目前在市场中的状况和地位并确定其竞争对手的态势；他们调查不同地区、不同阶层的民众对其产品的认知程度和购买意愿，以改进产品或推出新品种争取新顾客；他们还收集各地方的经济交通等信息，以决定如何保住现有市场和开发新市场.市场信息数据对企业是至关重要的.面对着一堆数据，我们该如何简洁明了地反映出其中规律性的东西或所谓的信息呢？一般首先对收集来的数据进行描述性分析，以初步发现其内在的规律性，然后再选择进一步分析的方法.

数据作为信息的载体，当然要分析数据中包含的主要信息，也就是分析数据的主要特征——数字特征.对一元数据，即样本数据(或观测值) $x_1, x_2, \cdots, x_n$ 是从一元总体中抽取的.一元数据的数字特征主要有：均值 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，方差 $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$，标准差 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$，等等.对于多元数据，除分析各分量的取值特征外，还要分析各分量之间的相关关系.

由于多元统计分析中的符号多而杂，因此需要**说明**：在一元统计学中一般用大写和小写字母分别来区分随机变量及其观测值，在本书后面的章节里，由于其他复杂的符号，我们可能不再遵守此约定(Anderson 在 *An Introduction to Multivariate Statistical Analysis*(3rd ed., 2003)中也采用了类似的作法)，请读者注意一个符号在每一章中的意义.

2.1 多元数据的表示

2.1.1 多元数据的一般格式

当人们要研究一个社会现象或自然现象时，通常要选择一些变量的特征来进行记录，从而形成多元数据.对于每个项目，这些变量的值被记录下来.

我们用 x_{ij} 表示第 j 个变量 $X_j(j=1, 2, \cdots, p)$ 在第 i 项或第 i 次($i=1, 2, \cdots, n$) 试验中的观测值，因此 p 个变量的 n 个观测值可以表示如下：

	变量 X_1	变量 X_2	…	变量 X_p
记录 1	x_{11}	x_{12}	…	x_{1p}
记录 2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮
记录 n	x_{n1}	x_{n2}	…	x_{np}

可以用一个有 n 行 p 列的矩阵来表示这些数据，称为**数据矩阵**，记为

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = (x_{ij})_{n\times p}.$$

于是以上数据矩阵包含了全部变量的所有观测值.

当这些变量处于同等地位时，就是聚类分析、主成分分析、因子分析、对应分析等模型的数据格式；当其中一个变量是因变量，而其他变量为自变量时，就是回归分析等模型的数据格式；若此时因变量还是分类变量，则为方差分析、判别分析等模型的数据格式.

2.1.2 多元数据的数字特征

把 p 个一维随机变量放在一起，就构成一个 p 维随机向量 $(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p)^{\mathrm{T}}$，如果同时对 p 个变量作一次观测，得到观测值 $(x_{11}, x_{12}, \cdots, x_{1p}) = \boldsymbol{X}_{(1)}^{\mathrm{T}}$，它是一个样品.观测 n 次就得到 n 个样品 $\boldsymbol{X}_{(i)}^{\mathrm{T}} = (x_{i1}, x_{i2}, \cdots, x_{ip})$，$i=1, 2, \cdots, n$，而 n 个样品就构成一个样本.

常把 n 个样品排成一个 $n \times p$ 矩阵(数据矩阵),记为

$$\boldsymbol{X}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}=\begin{pmatrix} \boldsymbol{X}_{(1)}^{\mathrm{T}} \\ \boldsymbol{X}_{(2)}^{\mathrm{T}} \\ \vdots \\ \boldsymbol{X}_{(n)}^{\mathrm{T}} \end{pmatrix}=(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p).$$

矩阵 $\boldsymbol{X}$ 的第 i 行 $\boldsymbol{X}_{(i)}^{\mathrm{T}}=(x_{i1}, x_{i2}, \cdots, x_{ip})(i=1, 2, \cdots, n)$ 是一个 p 维向量,矩阵 $\boldsymbol{X}$ 的第 j 列 $\boldsymbol{X}_j=\begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$ $(j=1, 2, \cdots, p)$ 表示对第 j 个变量的 n 次观测.

以下是多元数据的一些数字特征.

(1) 样本均值向量

$$\bar{\boldsymbol{X}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{(i)}=(\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)^{\mathrm{T}},$$

其中,$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}(j=1, 2, \cdots, p)$ 称为**样本均值**.

(2) 样本离差矩阵(又称交叉乘积矩阵)

$$\boldsymbol{A}=\sum_{k=1}^{n}(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})(\boldsymbol{X}_{(k)}-\bar{\boldsymbol{X}})^{\mathrm{T}}=(a_{ij})_{p\times p},$$

其中,$a_{ij}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)(i, j=1, 2, \cdots, p)$.

(3) 样本协方差矩阵

$$\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{A}=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}=(s_{ij})_{p\times p}$$

或 $\boldsymbol{S}^{*}=\frac{1}{n}\boldsymbol{A}$,其中,$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)$ $(i, j=1, 2, \cdots, p)$ 称为**样本协方差**,$s_{ii}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2(i=1, 2, \cdots, p)$ 称为**样本方差**,$\sqrt{s_{ii}}$ 称为**样本标准差**.