21世纪管理学研究生规划教材

Applied Multivariate Statistical Analysis

应用多元统计分析

李卫东 ◎编著

北京大学出版社
PEKING UNIVERSITY PRESS

21世纪管理学研究生规划教材

Applied Multivariate Statistical Analysis

应用多元统计分析

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

应用多元统计分析/李卫东编著. —北京:北京大学出版社,2008.11
(21世纪管理学研究生规划教材)
ISBN 978-7-301-14352-0

Ⅰ.应… Ⅱ.李… Ⅲ.多元分析:统计分析-研究生-教材 Ⅳ.O212.4

中国版本图书馆CIP数据核字(2008)第161040号

书　　名:应用多元统计分析
著作责任者:李卫东　编著
策划编辑:石会敏
责任编辑:石会敏
标准书号:ISBN 978-7-301-14352-0/F·2036
出版发行:北京大学出版社
地　　址:北京市海淀区成府路205号　100871
网　　址:http://www.pup.cn
电　　话:邮购部 62752015　发行部 62750672　编辑部 62752926　出版部 62754962
电子邮箱:em@pup.pku.edu.cn
印　刷　者:河北滦县鑫华书刊印刷厂
经　销　者:新华书店
787毫米×1092毫米　16开本　24.5印张　546千字
2008年11月第1版　2008年11月第1次印刷
印　　数:0001—4000册
定　　价:45.00元

序　言

在社会经济领域，许多现象是复杂的、多维的。面对多样的社会经济问题，仅从单方面、单指标考虑和分析，无法满足科学、客观地认识现实世界的需要。在当今的信息和知识经济时代，人类研究的科学和社会问题更加高深、复杂和庞大，有效地收集和分析数据以提取信息和获得知识变得更加须臾不可离。随着信息技术的发展和网络的逐渐普及，各类数据扑面而来，我们沉浸于数据的海洋之中。然而如何有效地从这些数据中挖掘出有效的信息，为管理、决策服务，是摆在我们面前的重要任务。多元统计分析提供了对多维数据进行分析的理论和方法，是解决此问题的有效工具和手段。

多元统计分析作为数理统计学的一个分支，在自然科学、社会经济、管理等领域得到了广泛应用。然而综观我国出版的多元统计分析的相关书籍，多数数理味较浓，适合财经类、管理类专业学生应用的较少。笔者在北京交通大学从事研究生多元统计分析课程的教学已有数年，期间积累了不少的资料和教学经验，多年来一直希望能编写一本满足财经类、管理类专业的研究生适用的教材，同时也为从事经济、管理实践和实际数据分析人员提供一本实用的参考书。

本书的主要特点为：

第一，系统性。多元统计分析的边界目前尚未有明确的界定，本书力求将相关的内容体现出来。在传统的推断统计原理、回归分析、聚类分析、主成分分析、因子分析、典型相关分析等基础上，将结合分析、路径分析、结构方程、多维标度法等内容也引入进来，力求内容完整。

第二，应用性。根据本人多年来从事本课程的教学经验和体会，对于财经类、管理类专业学生，不宜过多渲染数学原理，而应着重介绍方法的适用条件、基本思想、基本操作流程，并借助于计算机软件，实现方法的运作，并能够对分析结果进行说明。因此，本书侧重介绍各种多元统计分析方法的思想，尽可能简化相关数学原理，以讲

清楚为原则，避开过多的数学内容和工具，力求做到深入浅出，使多元统计分析的原理通俗易懂。同时，本书尽可能多地选择鲜活的案例，力求将方法的应用落到实处。其中有些案例就来自本人从事实际课题研究的成果。

第三，可操作性。方法工具的学习贵在应用。多元统计分析作为一个重要的工具，只有在应用中才能焕发生命力。本书将结合 SPSS、SAS、EViews 等相关软件，对相关案例做出详细说明，并给出相应的处理程序及输出结果，以促使读者能够进入应用的领域。

全书主要由两大部分组成：第一部分包括前四章，简要介绍多元统计分析的基础理论，包括多元描述统计、多元统计推断原理，这是多元统计分析的基础。第二部分从第五章至第十六章，依次介绍近年来广泛运用的、卓有成效的各种多元统计分析方法及其应用实例，包括回归分析、方差分析、判别分析、聚类分析、主成分分析、因子分析、多维标度法、对应分析、典型相关分析、路径分析、结构方程模型等。

感谢多年来上过我的多元统计分析课程的同学，正是他们的需求和想法，促使我不断地思考和探求相关知识。还要特别感谢北京大学出版社的石会敏老师、朱启兵老师对我的期望和对书稿的认真编辑，正是他们的期望给了我很大的动力。我的两名研究生刘欣、曾年初协助我进行了部分案例的搜集和校对，妻子刘志英协助进行了书稿的校对工作，在此一并感谢。感谢全家人对我工作的理解和支持。同时谨向对本书的出版有过帮助的师长和朋友表示衷心的感谢。

本书可作为经济、管理、社会科学等有关专业的研究生教材或参考书，同时也可作为管理咨询、市场研究、数据分析等领域的实际工作者实用的参考书。根据教学实践，本书讲授 48 课时较为合适。本书的内容相对较多，教师在选用此书作为教材时可以根据课时灵活选讲。

由于水平有限，书中难免存在不足之处，恳请读者批评指正。

李卫东

2008 年 7 月

目　录

第一章　绪论 …… (1)
第一节　引言 …… (3)
第二节　多元统计分析的应用 …… (5)
第三节　多元统计分析的发展 …… (7)
第四节　多元统计分析方法的分析流程 …… (8)
第五节　相关统计软件的说明 …… (9)

第二章　多元描述统计分析 …… (15)
第一节　多元描述统计量 …… (17)
第二节　多元数据的图形表示 …… (20)

第三章　多元正态分布及参数估计 …… (31)
第一节　基本概念 …… (33)
第二节　多元正态分布 …… (36)
第三节　多元正态分布均值向量和协方差阵的估计 …… (39)
第四节　几种常用的抽样分布 …… (40)
第五节　实例分析 …… (42)

第四章　多元正态分布均值向量和协方差阵的检验 …… (47)
第一节　总体均值向量的检验 …… (49)
第二节　协方差阵的检验 …… (58)
第三节　多个正态总体参数的检验 …… (60)

第五章　回归分析 …… (71)
第一节　回归分析的基本思想 …… (73)
第二节　多元线性回归模型 …… (77)
第三节　回归分析专题 …… (94)

第六章　聚类分析 …… (109)
第一节　聚类分析的基本思想 …… (111)
第二节　相似性的度量 …… (113)
第三节　系统聚类法 …… (120)

第四节　动态聚类法 …… (138)
第五节　有序样品的聚类 …… (142)
第六节　聚类方法的新进展 …… (146)

第七章　判别分析 …… (153)
第一节　判别分析的基本思想 …… (155)
第二节　距离判别法 …… (156)
第三节　Fisher 判别法 …… (161)
第四节　贝叶斯判别法 …… (173)
第五节　逐步判别法 …… (180)

第八章　主成分分析 …… (191)
第一节　主成分的含义及其思想 …… (193)
第二节　主成分模型及其几何意义 …… (194)
第三节　主成分的推导及性质 …… (196)
第四节　主成分分析的应用 …… (200)

第九章　因子分析 …… (207)
第一节　因子分析的模型 …… (210)
第二节　因子载荷矩阵的估计 …… (213)
第三节　因子旋转 …… (216)
第四节　因子得分 …… (219)
第五节　因子分析的基本步骤 …… (222)

第十章　对应分析法 …… (239)
第一节　对应分析的基本思想 …… (241)
第二节　对应分析的基本原理 …… (243)
第三节　实例分析 …… (248)

第十一章　典型相关分析 …… (259)
第一节　典型相关分析的基本原理 …… (261)
第二节　典型相关分析的基本步骤 …… (268)
第三节　实例分析 …… (269)

第十二章　偏最小二乘回归分析 …… (275)
第一节　引言 …… (277)
第二节　偏最小二乘回归分析的基本原理 …… (279)
第三节　实例分析 …… (283)

第十三章　结合分析 …… (289)
第一节　引言 …… (291)
第二节　结合分析的基本原理 …… (292)
第三节　结合分析的分析步骤 …… (295)
第四节　结合分析的应用及进展 …… (302)
第五节　实例分析 …… (307)

第十四章　多维标度法 …… (313)
第一节　引言 …… (315)
第二节　多维标度法的基本原理 …… (316)
第三节　非度量方法 …… (320)
第四节　实例分析 …… (322)

第十五章　路径分析 …… (331)
第一节　引言 …… (333)
第二节　路径分析的基本原理 …… (336)
第三节　分解简单相关系数的路径分析 …… (343)
第四节　实例分析 …… (346)

第十六章　结构方程模型 …… (351)
第一节　引言 …… (353)
第二节　结构方程模型的基本原理 …… (356)
第三节　实例分析 …… (362)

附录　常用统计表 …… (371)

参考文献 …… (381)

第一章 绪 论

教学目的

本章是绪论部分，主要介绍社会经济现象的多维特征，多元统计分析的含义、应用及学科的发展简况。通过本章的学习，希望读者能够：

1. 理解社会经济现象的多维特征；
2. 掌握多元统计分析的含义；
3. 领会多元统计分析方法的广泛应用；
4. 掌握多元统计分析的基本流程；
5. 领会 SPSS、SAS 等统计软件的基本功能和应用。

第一章 绪 论

第一节　引言

一、多元数据的广泛存在

在现实工作、生活中，对各类社会、经济、技术等现象的系统认识，需要搜集和分析大量体现现象特征和状态的多维指标和数据。我们经常碰到需要用多个指标进行描述的现象，如对一个企业的科学认识，需要了解企业的规模、产量、产值、利润、税收、所属行业、所在地区、员工数、组织等多方面的信息；如考查了解一个中学生的学习情况，需要了解学生在几个主要科目的成绩，像语文、数学、英语、政治等的考试成绩。还有，国际竞争力是影响国家或地区发展的重要因素，根据瑞士洛桑国际管理与发展学院的研究，国际竞争力可分为企业管理、国内经济实力、国际化程度、政府作用、金融环境、基础设施、科学技术和国民素质八个方面，每个方面由若干个具体指标组成。企业竞争力是反映企业生存和发展的能力，可以从资源、能力、知识、环境等多个维度进行分析，也可从组织、管理、文化、产品、技术、品牌等多个维度进行分析。

企业文化是促进企业发展的重要因素，是多维度、多层次的。多名学者对之进行了定性、定量研究，提出了不同的文化测度模型。荷兰学者霍夫斯坦德(Hofstede)在对大量调查数据统计分析的基础上，总结出权力距离(Power Distance)、风险规避(Uncertainty Avoidance)、个人主义倾向(Individualism)和对抗性(Masculinity)四个文化维度，用来测度群体文化特征。在后来与东方学者的合作研究中，他又发现了第五个维度：时间维度。欧瑞利(O'Reilly)和柴特曼(Chatman)提出：组织文化测量维度既要能反映组织文化的特性，又要求能够反映组织成员对组织文化的偏好程度。通过采用 Q-Sorted 研究方法，他们提出了测量组织文化的八个维度，即创新、稳定性、相互尊重、结果导向、注重细节、团队导向、进取性、决策性这八个维度。美国学者奎因(Quinn)和卡迈隆(Cameron)等人通过大量的文献回顾和实证研究发现组织中的主导文化、领导风格、管理角色、人力资源管理、质量管理以及对成功的判断准则都对组织的绩效表现有显著影响。他们在前人的研究基础上提出竞争性文化价值模型，认为组织灵活性-稳定性、外部导向-内部导向这两个维度能够有效地衡量企业文化差异对企业效率的影响。美国学者德尼森(Denison)构建了一个能够描述有效组织的文化特质模型。该模型认为有四种文化特质即适应性、使命、一致性、投入和组织有效性显著相关，其中每个文化特质对应着三个子维度。国内一些专家学者也总结了具有东方文化特征的企业文化的特征维度，代表性的有清华大学张德教授提出的企业文化十四个特征维度：领导风格、能力绩效导向、人际和谐、科学求真、凝聚力、正直诚信、顾客导向、卓越创新、组织学习、使命与战略、团队精神、发展意识、社会责任、文化认同。

在人格心理学中，对人格特质分为以下五个方面：外向性、接纳性、责任感、情绪稳定

性、开放性。自我价值感，是一个多维度、多层次的心理结构。自我价值感的内涵十分丰富，包括自我评价、自我感受、自我价值判断、自我体验、人格倾向等成分。这些又可进行细分，区分为不同类型的价值感。

从以上列举的情况看，对于大多数社会经济现象，都具有多维特性，都需要用多个指标进行测量和分析。

二、多元统计分析的含义及其研究目标

随着网络的普及和迅猛发展，网络上的各种资源异常丰富，人们面临着数据量过大的问题。面对着数据的海洋，应当如何分析这些数据间的关系，如何把这些数据中的重要信息挖掘出来，进而把握系统的本质属性呢？以多维数据集合为对象，进行统计数据的收集、整理、显示、分析，以揭示各类现象内在数量规律性的理论和方法，就是多元统计分析。多元统计分析是一元统计学的推广。

对多维数据的处理，用一元统计分析方法，一方面会导致计算量过大，同时会忽略多个变量之间存在的相关性，导致部分信息的缺失；另一方面，在一些情况下，仅依靠一元统计分析的结论，有可能会误导我们，其结论是不可靠的。

然而对多元统计分析的界定，大家看法并不一致。有的书籍中避而不谈，有的认为是对多变量的处理分析。实际上，多元统计分析包括诸多内容，随着实践的发展，不断有新的分析方法和技术出现，这不断丰富着本学科的内容。但将各领域中用到的方法进行全面系统的梳理不太现实，因而笔者选择了较为常用和实用的方法和技术作为本书的主要内容。

多元统计分析主要可用于以下目标的研究：

第一，数据结构简化或数据压缩。在不损失有价值信息的前提下，尽可能用简单的形式表示所研究的现象，同时又能作出很好的解释。

第二，分类和组合。在存在先验信息或不存在任何先验信息的情况下，对所考察的对象或指标按照相似程度进行分类或归类，同时给出一定的分组规则。

第三，变量间的关系。变量之间存在着怎样的互动关系？它们是如何相互影响的？这些通常是人们所关注的焦点问题。

第四，预测。如何根据变量间的变化关系对其他变量或对未来进行预测？这也是我们经常碰到的问题。

第五，假设的构建与检验。为验证某些观点或假设，对以多元正态总体参数形式陈述的假设进行检验。

第六，信息的提取。从数据到知识需要三个层次，数据—信息—知识。面对海量、复杂的数据，如何从中提取有效的信息和知识，为管理决策服务，这是我们面临的重要问题。多维数据的图示法、主成分分析、因子分析等提供了强有力的工具和帮助。

从多元统计分析方法的理论和本质看，笔者认为，它既是一门科学，同时又是一门艺术。其原因在于，在多元统计分析中，变量的选择、方法工具的选择、检验标准的选取，往

往是见仁见智的过程，很多方法的应用成功与否并没有绝对的标准，而在于人们在适合的情景下应用。因而，它包含有艺术的成分。这一点在一些具体方法的应用中可以得到具体体现。

第二节　多元统计分析的应用

多元统计分析方法被广泛应用于自然科学、社会科学、经济、管理等多个领域中，实践表明，多元统计分析方法在处理包括大量实验单元，多个指标的海量、复杂数据方面，是一种很有实用价值的方法，特别是随着实验单元、指标个数的增加，其价值和重要性愈能体现出来。试想一下我们面对一个包括5 000个个体单位、60个指标的情形，传统的统计学方法显然较难处理此类情形，而多元统计分析则可以帮助我们很好地面对和处理此类情形。

近年来，关于多元统计分析方法应用的出版物增长很快，为此很难对多元统计分析方法的广泛应用作出全面系统的概括。简明起见，我们结合多元统计分析的目标和研究内容分别进行说明。

1. 数据结构简化或数据压缩

- 用少数几个因子代表影响消费者购买行为的因素。
- 多元统计分析用于体育运动项目的研究。如，对田径运动成绩的分析有助于确定各种运动的基本功。林德于1977年对八届奥林匹克运动会十项全能成绩用多元统计分析方法确定了四个基本体力因子：短跑速度、臂力、长跑耐力、腿力。
- 选择区域主导产业时，在产业的多个指标数据中可用因子分析方法确定若干个主要因子，为主导产业的选择提供参考。

2. 分类和组合

- 根据产业发展的不同情况，将不同产业进行投资可行性等分类。
- 根据城市发展的情况，对城市进行分类。
- 对不同客户，根据其消费信贷情况，对其进行分类。
- 根据不同企业的经营、生产情况，对其进行分类。
- 对作品的著作权的归属进行分析。
- 利用多元统计分析方法进行税务识别，在发达国家和地区早已实行，如美国90%以上的税务稽查案件，都是通过计算机分析筛选出来的。美国国内收入局装有一套“货币—银行—企业”的检查系统，它的数据库里储存着来自银行、企业和货币使用者的流动信息。无论是经济实况稽查还是分行业专业稽查，一般都采用电脑计分、选样抽查的方法确定稽查对象。纳税人的纳税申报表根据国内收入局制定的标准，由电脑进行客观打分。一般而言，收入越高、减项越大、错误越多者，分数越高，被选定查税的可能性就越大。美国国内收入局通常每3年实行一次“衡量纳税人遵法稽查计划”，每次通过电脑选择数万件，对每件都彻底稽查，再根据结果改进电脑打分及选样抽查等的标准。

• 对不同地质条件的地区是否产油进行监测，获得样本，建立模型，以对新地区是否产油进行判定。

• 市场细分。市场细分是指营销者通过市场调研，依据消费者的需要、购买行为和购买习惯等方面的特征差异，把某一产品的市场整体划分为若干消费者群的市场分类过程。每一个消费者群就是一个细分市场，每一个细分市场都是具有类似需求倾向的消费者构成的群体。市场细分是市场定位的基础，是企业制定营销策略的基本依据之一。

3. 变量间的关系

• 企业绩效与战略之间的关系。

• 企业文化与企业绩效的关系。

• 创新与企业环境的关系。

• 儿童成绩与其家庭环境、身体素质等因素之间的关系。

• 个人所受教育程度、工作岗位、工作能力对其薪酬水平的影响。

4. 预测

• 通过对学生情况的连续跟踪，利用高考成绩得分及几个高中成绩变量与几个大学成绩之间的联系，构造用来预测学生在大学里成功与否的指标。

• 利用公司会计数据信息，构造识别具有潜在财务危机上市公司的方法。

• 根据产品销售状况与企业投放广告情况、价格水平、促销力度、产品质量、竞争产品等因素的关系，对某产品的销售情况进行预测。

5. 假设的构建与检验

• 利用多个变量数据来确定在新兴工业化国家中不同类型的公司是否呈现出不同的改革模式。

• 特定城市空气污染的程度在一周时间内是否固定不变，或周末与平时有无显著差异。

• 股市中是否存在周日效应。

• 不同广告方式等因素对产品的销量是否有显著差异。

• 对个人客户信用状况是否良好进行检验。

• 对企业文化的现状是否适应企业战略发展的需要进行检验。

• 不同国家或企业的竞争力是否有显著差异。

6. 信息的提取

• 对超市中不同顾客购买日用品、消费品等数据的整理分析，为超市的货物调配、摆放布局、进货品种等管理决策提供基本依据。

• 客户流失分析是电信运营商用来获取利润最直接最有效的手段之一。在目前竞争激烈的电信市场中，企业和客户之间的关系是经常变动的，一旦成为电信企业的客户，电信企业就要尽力保持这种客户关系。客户关系的最佳境界体现在三个方面：最长时间地保持这种关系；最多次数地和客户交易；保证每次交易的利润最大化。因此，电信公司对已有的客户进行流失分析是一项重要的工作。通过对大量客户原始记录数据的分类分析，结合以前拥有的客户流失数据，建立客户属性、服务属性、客户消费数据与客户流失可能性关联的多元统计模型，找出客户属性、服务属性、客户消费数据与客户流失的最

终状态的关系，并给出模型，寻找流失客户的主要特征，建立评分模型，按照流失程度，对已有客户进行等级评价。

通过上述例子，读者可以看到，虽然分析的具体问题各不相同，但都会用到多元统计分析方法，这表明了多元统计分析应用的极大广泛性。

第三节　多元统计分析的发展

多元统计分析最初发端于20世纪30年代。1928年维希特发表论文《多元正态总体样本协差阵的精确分布》，表明了多元统计分析的开端。其后，费希尔(R. A. Fisher)、霍特林(H. Hotelling)、罗伊(S. N. Roy)、许宝騄等人做了一系列奠基性工作，使多元统计分析在理论上得到了迅速发展。20世纪40年代，多元统计分析在心理、教育、生物学等领域得到了不少的应用。但由于该类方法计算量太大，使其发展受到影响，甚至停滞了相当长时间。50年代中期，随着电子计算机的出现和发展，使多元统计分析方法在地质、气象、医学、经济、管理、社会学及图像处理等方面得到了广泛应用。60年代通过应用和实践又进一步发展和完善了理论，同时由于理论的发展和完善又进一步扩大了它的应用范围。我国20世纪70年代初才关注到多元统计分析方法，改革开放以后，多元统计分析的理论和应用取得了较多进展，并在经济、管理实践中进行了广泛应用。

例如，中国科学院研究生院陈希孺和中国科学技术大学赵林城比较系统地研究了多元线性回归的LS和M估计的相合性、渐近正态性和线性表示等大样本性质，在一些情况下得到了或几乎得到了充要条件，有的问题得到了精确的阶估计和理想的界限。中国科学院应用数学所方开泰和上海财经大学张尧庭等在椭球总体的多元分析方面，中国科学院系统科学所吴启光和北京理工大学徐兴忠等在多种线性模型估计的容许性和其他统计决策问题方面，北京工业大学王松桂在线性回归的估计方面，以及东北师范大学史宁中在有约束的线性模型方面，中国人民大学何晓群教授在质量管理(六西格玛)方面，都取得了不少成果。

比线性模型复杂的多元模型是非线性参数模型、半参数和非参数模型。在这些模型的理论方面我国统计学者也做了许多工作。例如，中国科学院系统科学所成平等在研究半参数模型的渐近有效估计方面，陈希孺、赵林城和安徽大学陈桂景等在研究非参数回归、密度估计和非参数判别方面，东南大学韦博成等在用微分几何方法研究非线性(参数)回归方面，以及南京大学王金德在非线性回归估计的渐近性质方面均取得了一系列成果。在非参数理论的成果中，陈希孺和赵林城彻底解决了关于U统计量分布的非一致收敛速度问题，有关结果被美国《统计科学百科全书》以及美国和前苏联等出版的多本专著引述。东南大学韦博成、中国人民大学吴喜之以及云南大学王学仁和石磊等在模型和数据的统计诊断方面取得了许多成果。云南大学的学者还把他们的成果用于地质探矿的数据分析等实际问题，取得成功。解决数据与模型这一对矛盾的另一种途径是使用对模型不敏感的统计方法，即当模型与数据吻合或不太吻合时都能给出比较正确的结论，

这就是稳健统计方法。中国科学院系统科学所李国英和张健等在多元位置和散布阵的稳健估计及其性质,位置 M 估计的崩溃性质等方面也取得了一些成果。在多维试验设计方面,中国科学院数学所王元和应用数学所方开泰引进数论方法提出了均匀设计,能用于缺乏使用正交设计条件的情况。该设计方法已在国内的多个实际部门应用,效果良好。这一工作在国际上也受到重视。南开大学张润楚等在研究计算机试验设计方面也有一些成果。西北工业大学张恒喜等对小样本多元统计分析方法进行了分析,并出版了《小样本多元数据分析方法及应用》(2002)。

可以预计,随着计算机技术的发展,多元统计分析将逐步取代一元统计学,成为人们日常生活和工作中的必要工具。当然,随着数据量的不断扩大,这也给多元统计分析带来了极大挑战。

由于网络的发展和数据采集技术的进步,时间维度与横截面维度相结合而形成的综列数据越来越多,可以预计,在不久的将来,时间序列分析将和多元统计分析日趋融合,成为人们进行数据处理分析的重要工具。

第四节　多元统计分析方法的分析流程

利用多元统计分析方法解决实际问题时,通常有以下步骤:

第一,确定问题和目标。在实际问题中,由于涉及变量数量多,变量间的关系复杂,由于各种条件的限制我们不可能面面俱到,因而需要集中精力,抓主要矛盾,从错综复杂的数量关系中确定研究问题和目标。有目标才有动力,才有前进的方向。

第二,根据相关理论,设置指标变量。对于实际研究问题,确定了目标后,面临的问题就是:在该项研究中应涉及哪些指标变量?这需要相关领域理论的支持,对面临的问题选择指标体系。指标的选择不宜过多过细,但也不宜遗漏重要变量。

第三,收集、整理统计数据。多元统计分析模型的建立是基于指标变量的样本统计数据,样本数据的质量决定了模型的作用大小。样本数据的取得可采用不同的随机抽样方式,如简单随机抽样、分层抽样、整群抽样等。考虑推断的可行性,在本书中,若不作特殊说明,样本均是按照简单随机抽样方式抽取。通常取得的样本数据可分为三类:一类是横截面数据,一类是时间序列数据,还有一类是综列数据。在整理数据时,应剔除异常值,对变量数据进行规格化、标准化等变换处理,以便于后续数据的分析。

第四,根据研究目标和数据,选择多元统计分析方法,构造理论模型。按照研究目标的要求,结合数据特性,选择合适的方法工具。如研究多变量之间的关系,可采用回归分析、路径分析、因子分析等工具;若要研究事物的分类,则可选用聚类分析、判别分析等方法。

第五,模型的估计。结合 SPSS、SAS、EViews 等计算机软件,进行统计计算,估计模型。

第六,模型的检验与调试。建立模型后,模型的效果如何,还需要进行统计检验和模

型实际应用的检验。如回归分析模型，需要考虑模型的显著性检验（F 检验）、回归系数的显著性检验（t 检验）等；若是因子分析模型，则需要检查其因子的信息提取率是否达到一定标准、因子的含义是否容易理解等。如果模型没有通过检验，则需要对模型进行调试，以得到相对满意的模型。

第七，模型的应用。在模型通过各种检验后，就可以运用统计模型做进一步的分析研究。在应用时，必须注意定量分析与定性分析的有机结合。

第五节　相关统计软件的说明

正是计算机软件技术的发展，才使得多元统计分析获得了新生。对于多元统计分析模型，手工计算显然是不可行的，必须借助于计算机软件。适用于多元统计分析的软件很多，代表性的有 SPSS、SAS、EViews、R、MATLAB、S-PLUS 等软件。本书中主要以 SPSS 和 SAS 软件为蓝本进行介绍。下面对这两个软件进行简要说明。

一、SPSS 软件简介

SPSS 是“社会科学统计软件包”（Statistical Package for the Social Science）的简称，是一种集成化的计算机数据处理应用软件。1968 年，美国斯坦福大学 H. Nie 等三位大学生开发了最早的 SPSS 统计软件，并于 1975 年在芝加哥成立了 SPSS 公司，至今已有三十余年的成长历史，全球约有 25 万家产品用户，广泛分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研、教育等多个领域和行业。SPSS 是世界上公认的三大数据分析软件之一（SAS、SPSS 和 SYSTAT）。1994—1998 年间，SPSS 公司陆续购并了 SYSTAT 公司、BMDP 软件公司、ISL 公司等，并将各公司的主打产品收纳 SPSS 旗下，从而使 SPSS 公司由原来的单一统计产品开发与销售转向为企业、教育科研及政府机构提供全面信息统计决策支持服务，成为走在最新流行的“数据仓库”和“数据挖掘”领域前沿的一家综合统计软件公司。伴随 SPSS 服务领域的扩大和深度的增加，SPSS 公司已决定将其全称更改为 Statistical Product and Service Solutions（统计产品与服务解决方案）。

SPSS 软件具有如下特点：

第一，集数据录入、资料编辑、数据管理、统计分析、报表制作、图形绘制为一体。从理论上说，只要计算机硬盘和内存足够大，SPSS 可以处理任意大小的数据文件，无论文件中包含多少个变量，也不论数据中包含多少个个案。

第二，统计功能系统全面，涵盖了各种的统计方法。包括常规的集中趋势统计量和离散程度统计量、相关分析、回归分析、方差分析、卡方检验、t 检验和非参数检验；也包括近期发展的多元统计技术，如多元回归分析、聚类分析、判别分析、主成分分析和因子分析等方法，并能在屏幕（或打印机）上显示（打印）如正态分布图、直方图、散点图等各种