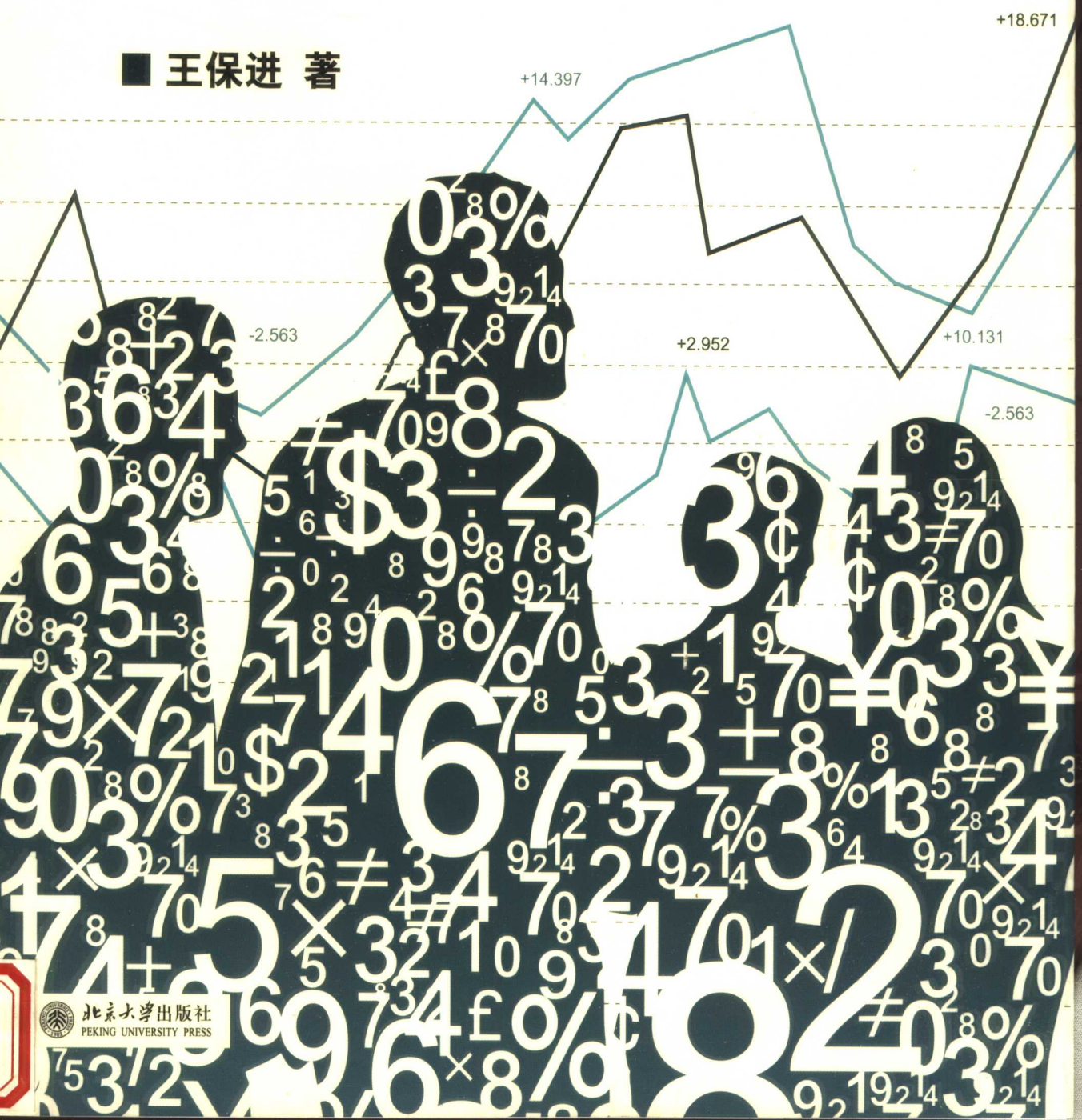
社会科学
研究方法丛书

H·EDU

多变量分析

统计软件与数据分析

■ 王保进 著



北京大学出版社
PEKING UNIVERSITY PRESS

社会科学研究方法丛书

多变量分析

统计软件与数据分析

■ 王保进 著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

多变量分析:统计软件与数据分析/王保进著. —北京:北京大学出版社,2007.8
(社会科学研究方法丛书)
ISBN 978-7-301-12473-4

Ⅰ. 多… Ⅱ. 王… Ⅲ. 统计分析-应用软件 Ⅳ. C819

中国版本图书馆 CIP 数据核字(2007)第 093105 号

高等教育文化事业有限公司(Taiwan)出版《多变量分析:套装程式与资料分析》一书的繁体版,并授权北京大学出版社在中国内地出版发行此书简体字版本,书名变更为《多变量分析:统计软件与数据分析》。
《多变量分析:套装程式与资料分析》,王保进著,2004 年 8 月第 1 版,ISBN:957-411-818-5。

书　　名:多变量分析——统计软件与数据分析
著作责任者:王保进　著
责任编辑:丁莉华
标准书号:ISBN 978-7-301-12473-4/C·0446
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 62117788　出版部 62754962
电子邮箱:law@pup.pku.edu.cn
印　刷　者:北京大学印刷厂
经　销　者:新华书店
787 毫米×1092 毫米　16 开本　29 印张　539 千字
2007 年 8 月第 1 版　2007 年 8 月第 1 次印刷
定　　价:48.00 元

Preface 序

近二十年来量化科学研究发展的趋势之一，就是多变量分析统计方法的应用日趋普遍，使研究者得以从资料中萃取更丰富的资讯，供作决策之用。特别是线性结构方程方法论的形成，使得量化科学研究得以更严谨的方法验证变量间的因果关系，从而确认理论体系的配适性。另一方面，电脑科技进步的神速，使得电脑的运算能力益加快速，所能处理资料的量也更加庞大，特别是统计软件的不断推陈出新，更为研究者添加了资料分析的利器。

在业界，多变量分析方法论的书籍犹如汗牛充栋，多得不胜枚举，而近年来出版此一领域书籍的共同特征，就是重视方法论在资料分析上的应用。目前“多变量分析”课程在各学术领域的研究院所中，多已列为选修课程，然而相关的方法论书籍并不多见，更缺乏资料分析取向的参考书籍。本书的出版，恰可弥补现有多变量分析领域的不足，提供给研究者在进行量化科学研究时，一本便利的食谱式参考书籍。

本书内容共分十章，涵盖了多因子方差分析、主成分分析、因子分析、判别分析、聚类分析、典型相关、多维标度、线性结构方程、逻辑斯谛回归分析及对数线性方程等常见的多变量分析方法。除第一章的简介与方法选择外，各章的撰写，第一节都是相关基本原理的介绍，为减轻初入门读者的负担，本书尽可能不加入公式的推导过程；第二节则是 SPSS 与 SAS 在该统计方法应用的程序介绍；从第三节起，则举一些实例，说明如何利用 SPSS 与 SAS 进行资料分析，并解释报表结果。至于对报表解释，为节省篇幅，主要以国内较普及的 SPSS 为主，对 SAS 部分的报表，雷同的部分则采用精简的解释方式，因此对习惯使用 SAS 的读者而言，建议您在报表解释部分可对照 SPSS 相同内容的解释。

由于多变量分析属于推论统计上进阶的方法论，因此本书假设读者对 SPSS

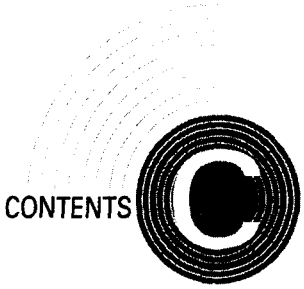
与 SAS 的基本操作，都已具备基本素养，因此书中并无有关两大统计软件基本操作的说明。有需要的读者，可参阅国内外相关的著作。若您是一位初入门的多变量分析方法使用者，建议您从第一章逐章阅读，应该可以提升对方法论的认识；至于对方法论已相当熟悉的人，则可依研究需求，在各节中找到您所需要的信息。

笔者进入大学服务已经十余年，一路走来要特别感谢授业恩师政治大学教育研究所马信行博士的启迪与教诲。*Educational Administration Quarterly* 在 2000 年的教育行政学术研究人力资源专刊中，登载了一篇“The Write Stuff: A Study of Productive Scholars in Educational Administration”的论文，论文中明确地指出，“明师”对学术研究与知识生产的重要性。如果笔者在教育行政与研究方法论的领域中有丝毫的成就，那都要感谢马教授的启迪，并扮演 mentoring 的角色及一路的支持与鼓励！

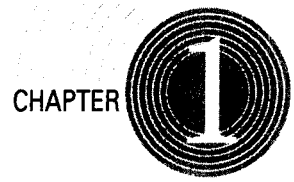
最后，笔者才疏学浅，不揣浅陋，虽校对再三，疏漏之处，尚祈各方先进不吝指正。

王保进

2004 年 6 月于台北



目 录



第一章

多变量分析简介

第一节　绪　　论

多变量分析(multivariate analysis)的方法论,近十年来因为个人电脑处理能力的剧增,统计软件的窗口化,加上复杂的新方法不断出现,以及验证性研究日趋受到重视,因此有了相当蓬勃的发展。目前多变量分析的各种方法,已经在企业、政府机构及大学各学科的研究中普遍使用,同时各学术研究领域中,属于量化(quantitative)研究者,不使用多变量分析作为案例分析方法的,更是凤毛麟角。表 1-1是教育心理学领域知名的学术季刊 *Journal of Educational Psychology* 在 1990、2000 及 2001 年三个年度所出版论文使用的案例分析方法。由表中可知,除了各种方差分析和多元回归分析,仍是研究者进行资料分析常用方法外,包括多变量分析、LISREL 及 HLM 等方法,近来已成为研究者经常使用的资料分析方法。

表 1-1　*Journal of Educational Psychology* 论文使用的案例分析方法

年代 / 分析方式	1990(62 篇)	2000(66 篇)	2001(49 篇)
卡方检验	8	8	7
相关系数	30	24	14
方差分析	41	39	31
多元回归	11	13	8
多变量分析[a]	5	22	14
LISREL	3	9	9
多层线性模型(HLM)	0	8	5
整合分析	1	2	3

注:a 包括区别分析、聚类分析、多维变量、典型相关、因素分析等多变量方法。

Hair、Anderson、Tatham 与 Black(1998)即指出,一个仅能进行双变量间关系的显著性检验,而不使用多变量分析的研究者,很显然地表现出其漠视可以提供有用信息的鲁棒性资料分析工具。而 Gatty(1966)也指出,在任何应用科学领域的研究目的中,研究问题若非以多变量分析处理,则结果会是相当肤浅的!由此可知,多变量分析确实可以让学术研究工作者除了能从测量资料中获取更多的信息外,同时可以让显著性检验的结果更具鲁棒性。

第二节　常用多变量分析方法

要对多变量分析作一明确的定义并非易事。就广义来看,当研究者同时进行多个变量的测量,假设这些变量间具多变量正态分布(multivariate normal distribu-

tion)的特性,进而同时分析至少两个变量间的关系,都可以视为是"多变量分析"(Dillon & Goldstein, 1984),此一定义,恰与分析单一变量的平均数或标准差的"单变量分析"(univariate analysis)或探讨两变量间配对相关的"双变量分析"(bivariate analysis)相对应。而 Stevens(1992)则认为所谓多变量分析是指同时分析多个因变量间关系的统计方法。甚至有更简易的说法,认为只要同时包括两个以上因变量的资料分析方法,都可以称为多变量分析。

在社会科学研究中,主要的多变量分析方法包括多变量方差分析(Multivariate analysis of variance, MANOVA)、主成分分析(Principal component analysis)、因子分析(Factor analysis)、典型相关(Canonical correlation analysis)、聚类分析(Cluster analysis)、判别分析(Discriminant analysis)、多维标度(Multidimensional scaling, MDS),以及近年来颇受瞩目的验证性因子分析(Confirmatory factor analysis)或线性结构模型(LISREL)与逻辑斯谛回归分析等,以下简单说明这些方法的观念与适用时机。

一、多变量方差分析

MANOVA 适用于同时探讨一个或多个自变量与两个以上因变量间因果关系的统计方法,依照研究者所操作自变量的个数,可以分为单因素(一个自变量)或多因素(两个以上自变量)MANOVA。进行 MANOVA 时,自变量必须是离散的称名或顺序变量,而因变量则必须是等距以上的变量。

二、主成分分析

主成分分析的主要功能在分析多个变量间的相关,以建构变量间的总体性指标(overall indicators)。当研究者测量一群彼此间具有高相关的变量,则在进行显著性检验前,为避免变量数过多,造成解释上的复杂与困扰,常会先进行主成分分析,在尽量不丧失原有信息的前提下,抽取出少数几个主成分,作为代表原来变量的总体性指标,达到资料减缩(data reduction)的功能。进行主成分分析时,并无自变量或因变量的区别,但所有的变量必须是等距以上的变量。

三、因子分析

因子分析与主成分分析常被研究者所混淆,因为二者的功能都是通过对变量间的相关分析,以达到简化数据功能。但不同的是,主成分分析是在找出变量间最佳线性组合(linear combination)的主成分,以说明变量间最多的变异量;至于因子分析,则在于找出变量间共同的潜在结构(latent structure)或因子(factor),以估计每一个变量在各因子上的负荷量(loading)。进行因子分析时,并无自变量或因变量的区分,但所有的变量必须是等距以上的变量。

四、典型相关

典型相关可视为积差相关或多元回归分析的扩展，主要功能在分析两个变量间的相关。进行多元回归分析的目的，是在分析一个或多个自变量与一个因变量间的关系，而典型相关中因变量也可以是多个；也就是说，典型相关的目的在于通过计算得到两个变量线性组合的加权系数，以使（maximum）两个变量间的相关达到最大化。进行典型相关时，两个变量间并无严格的自变量与因变量的区分，但两个变量都必须是等距以上的变量。

五、聚类分析

聚类分析的主要功能在进行分类（classification），当研究者有观测值时，常会根据观测值的相似性或差异性进行分类，以形成几个性质不同的类别，简化解释的工作。也就是说，聚类分析根据对变量进行测量的观测值进行分类，以达到组内同质、组间异质的目的。其次，聚类分析完成后，通常可以进行区别分析，以识别分类的效度。当然，在某些时候也可以对变量进行分类（此功能类似因子分析，因此多采用因子分析解决问题）。进行聚类分析时，并无自变量或因变量的区分，但原则上所有的变量必须是等距以上的变量。

六、区别分析

区别分析是多变量分析中应用相当广泛的统计方法，它可以用来对样本进行分类的工作；也可以用来了解不同类别样本在某些变量上的差异情形；同时也可以根据不同类别的样本在某些变量的实际表现，用来预测新的样本属于某一类别的概率。因此在行为科学中，常见研究者单独使用区别分析，建立区别函数（discriminant function），以对新样本进行预测；或是多变量方差分析的检验值达到显著水平后，比较不同组别样本在因变量平均数的差异情形；或是在聚类分析后，检验聚类的正确性。进行区别分析时，自变量是等距以上的变量，至于因变量通常是离散变量。

七、多维量表分析

多维量表分析基本上也是一种分类的统计方法，它在市场研究上普遍被应用。当研究者想要解释一群受试者（例如消费者）对一组客体（例如商品）在某些变量上相似性的测量值中所包含的信息，此时多维量表分析就是一个相当适用的方法。研究者只要将这一组客体在变量上的测量值转换成多维度的几何表征，就能够将这些客体有效地显示在这个几何空间中，达到分类的目的，同时也可以进一步解释这些几何表征所代表的潜在结构或意义。进行多维量表分析时，并无自变量或因变量的区分，同时变量可以是等距以上变量，也可以是称名或顺序变量。

八、线性结构方程

线性结构方程是一个相当具有变通与弹性的统计方法，随着研究者对变量间关系界定的差异，LISREL 的常见名称包括协方差结构分析（Covariance structure analysis）、潜变量分析（Latent variable analysis）、线性结构模型或验证性因子分析。LISREL 可视为多元回归分析与因子分析两个方法论的整合模型，让研究者可以探讨变量间的线性关系（回归分析），并对可测量显（manifest）变量与不可测量的潜（latent）变量间（因子分析）的因果模型作假设检验。

九、逻辑斯谛回归分析

逻辑斯谛回归可视为传统多元回归分析的一个特例。它和多元回归分析一样，都具有解释自变量与因变量间的关系，并可进行预测。所不同的是在进行多元回归分析时，包括自变量与因变量都必须是等距以上变量；但在进行逻辑斯谛回归分析时，自变量仍是等距以上变量，但因变量则是二分的称名变量。

十、对数线性方程

在基本统计学中，当研究者面对探讨两个称名或顺序变量间关系的研究问题时，都是以卡方检验来进行假设检验。当问题的性质是探讨两个称名变量间是否独立（independence）或是关联强度（strength of association）时，是以卡方独立性检验来进行假设检验。进行卡方独立性检验时，研究者必须将样本在两个称名变量上的反应，建立二维列联表（contingency table），以进一步根据列联表中各细格（cell）的次数反应，进行显著性检验。但当研究者面对三个或三个以上的称名变量时，所建立的多元列联表（multiway contingency table）间变量关联的分析，卡方独立性检验将无法解决这样的问题，此时适合的方法就是对数线性模型。利用对数线性模型来解决多元列联表的问题的目的，主要就在探讨构成列联表的多个名义变量间的关系，进而在精简的（parsimonious）原则下建构拟合的解释模型，并根据所建立的模型估计细格参数值，以了解各变量效果对细格次数的影响。

十一、Logit 对数线性模型

在对数线性模型中，多个名义变量间是互为因果的关系，并无自变量与因变量的区分，研究目的在探讨变量间的关联强度与性质。但有时研究者会面临变量间有自变量与因变量区分的情境。在基本统计学中，当研究者面对的问题性质是两个称名变量间有自变量与因变量的区别，目的在探讨两个变量间的因果关系时，多是以卡方齐性检验来进行假设检验。但自变量个数在两个以上时，卡方齐性检验就不再适用，而必须改用 Logit 对数线性模型（logit log-linear model）方法来对数据进行分析。Logit 对数线性模型的功能与多元回归分析相当类似，都可以用来探讨与解释因变量与自变量间的关系，但不同的是，多元回归分析的变量都是

等距以上变量，通常以最小平方法进行模型估计与检验；Logit 对数线性模型的变量都是称名变量，通常以最大似然法进行模型估计与检验。

第三节　多变量分析方法的选择

对研究者而言，在使用多变量分析时，首先最重要的一件事，是必须能够根据研究问题或研究假设的性质，选择正确的多变量统计方法。在一般的多变量分析著作中，多数学者都有一套个人的方法，将多变量分析的方法加以分类。例如 Sharma(1996)根据自变量与因变量的个数、测量尺度，将常用的多变量分析方法加以分类。而 Hair 等人(1998)则是根据变量间的关系、因变量的个数及变量的测量尺度等三个条件，对多变量分析的方法加以分类。至于 Tabachnick 与 Fidell (2001)则将研究问题分为五类，并据此将各种单变量与多变量的方法加以分类。他们的分类包括：

一、变量间的相关

包括积差相关、多元回归分析、典型相关及对数线性模型(Log-linear model)都是用来探讨变量间相关(correlation)的适当统计方法。

二、比较平均数差异

包括 t 检验、单(多)因素方差分析、协方差分析、Hotelling's T^2 检验及单(多)因素多变量方差分析等都是可以用来检验不同组别样本在因变量上平均数差异的适当方法。

三、预测

包括多元回归分析、区别分析、Logit 对数线性模型及逻辑斯谛回归分析等都是用来预测样本在变量上的表现的适当方法。

四、变量的结构(structure)

包括主成分分析、因子分析及线性结构模型等都是用来探讨变量间结构的适当方法。至于分析观测值间的结构的方法，在 Tabachnick 与 Fidell 的著作中并未提及，包括聚类分析和多维量表分析都是适当的方法。

五、事件的时间系列

包括生存分析(survival analysis)、时间系列分析(time series analysis)，是研究者常用来分析一个空间事件或时间事件的纵贯面时间系列趋势的统计方法。

根据 Tabachnick 和 Fidell 的分类设计，若再考虑自变量与因变量的个数和测

量水平，以及研究问题或假设中是否包括协变量或干扰变量，则有关多变量分析统计方法的选择可以表 1-2 表示如下。

表 1-2 多变量分析方法的分类研究问题性质

研究问题性质	因变量		自变量		有无协变量	适用统计方法
	个数	数据类型	个数	数据类型		
变量间的相关	1	metric[a]	1	metric	N[c]	积差相关系数
			多	metric	N	多元相关系数、偏相关
	多	metric	多	metric	N	典型相关
	N		多	nometric	N	对数线性模型
比较平均数差异	1	metric	1	nometric	N	t 检验、单因子方差分析
					Y	单因子协方差分析
			多	nometric	N	多因子方差分析
					Y	多因子协方差分析
	多	metric	1	nometric	N	单因子多变量方差分析
					Y	单因子多变量协方差分析
			多	nometric	N	多因子多变量方差分析
					Y	多因子多变量方差分析
预测	1	metric	多	metric	N,Y	多元回归分析
		nometric[b]	多	metric	N	区别分析、逻辑斯谛回归分析
				nometric	N	Logit 对数线性模型
				metric,nometric	N	逻辑斯谛回归分析
变量的结构	多	metric	N		N	因子分析(潜在结构)、主成分分析(总体性指标)
			多	metric	N	LISREL
观测值	多	metric	N		N	聚类分析、多维量表分析
		nometric	N		N	多维量表分析
事件的时间系列	1	metric	N		N	ARIMA[d]、生存分析
			1 or 多	metric	N	Vector ARIMA
				nometric	N	Intervention ARIMA

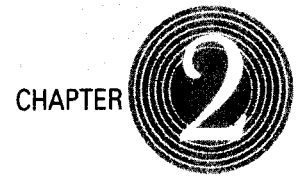
注：a. metric：表示用等距或等比量表所测量的连续数据。

b. nometric：表示用称名或顺序量表所测量的离散数据。

c. N 代表未具备，Y 代表具备。

d. ARIMA：Autoregressive Integrated Moving Average。

表 1-2 为社会科学中常用的多变量分析方法，研究者在确定研究问题或假设后，只要根据问题性质，然后考虑变量的个数和测量尺度，应该就可以正确地选择适当的多变量分析方法。



第二章

多变量方差分析