内容简介

本书为北京大学数学科学学院概率统计系“应用多元统计分析”课程使用多年的教材，它主要介绍一些实用的多元统计分析方法的理论及其应用，并列举了各方面的应用实例，同时还以国际上著名的统计分析软件SAS系统作为典型工具，通过实例介绍如何处理数据分析中的各种实际问题。

本书共分十一章。第一章为绪论；第二、第三章介绍多元统计分析的理论基础——多元正态分布及其参数的估计和检验问题；第四章重点介绍多因变量的多元线性回归的有关问题，包括模型、参数的估计及其性质、假设检验、变量筛选，以及双重筛选逐步回归问题；第五、第六章介绍分类问题（判别与聚类）；第七到第九章介绍降维的多变量方法（主成分分析、因子分析和对应分析方法）；第十章讨论两组相关变量的典型相关分析；第十一章介绍近年来发展的偏最小二乘回归分析方法；并且在每一章内都配有适量的习题。“附录”中介绍了本课程所需的矩阵代数的有关内容；书末附有“部分习题参考解答或提示”，这些都将更便于读者自学。

本书可作为综合大学、工科大学或高等师范学院数学系、应用数学系、经济学等相关专业的本科生或研究生教材或教学参考书；对于其他领域中从事应用统计的工作人员也是一本极好的学习参考书。

作者简介

高惠璇 北京大学数学科学学院教授。1965年毕业于北京大学数学力学系。长期从事概率论与数理统计的教学、科研工作，主要研究方向是统计计算、统计软件与应用多元统计方法，曾参加过国家教委《数学软件的研究与开发》项目和统计软件的开发及推广普及工作。

北京大学数学教学系列丛书

应用多元统计分析

高惠璇 编著

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/高惠璇编著. —北京：北京大学出版社，2005.1

(北京大学数学教学系列丛书)

ISBN 7-301-07858-7

Ⅰ.应… Ⅱ.高… Ⅲ.多元分析：统计分析-高等学校-教材 Ⅳ.O212.4

中国版本图书馆 CIP 数据核字(2004)第 124211 号

书　　名：应用多元统计分析

著作责任者：高惠璇　编著

责 任 编 辑：邱淑清

标 准 书 号：ISBN 7-301-07858-7/O・0613

出 版 发 行：北京大学出版社

地　　址：北京市海淀区中关村北京大学校内　100871

网　　址：http://cbs.pku.edu.cn　电子信箱：zpup@pup.pku.edu.cn

电　　话：邮购部 62752015　发行部 62750672　理科编辑部 62752021

印　刷　者：北京大学印刷厂

经　销　者：新华书店

890 mm×1240 mm　A5　13.625 印张　392 千字

2005 年 1 月第 1 版　　2005 年 1 月第 1 次印刷

印　　数：0001—4000 册

定　　价：21.00 元

序　言

自 1995 年以来，在姜伯驹院士的主持下，北京大学数学科学学院根据国际数学发展的要求和北京大学数学教育的实际，创造性地贯彻教育部“加强基础，淡化专业，因材施教，分流培养”的办学方针，全面发挥我院学科门类齐全和师资力量雄厚的综合优势，在培养模式的转变、教学计划的修订、教学内容与方法的革新，以及教材建设等方面进行了全方位、大力度的改革，取得了显著的成效。2001 年，北京大学数学科学学院的这项改革成果荣获全国教学成果特等奖，在国内外产生很大反响。

在本科教育改革方面，我们按照加强基础、淡化专业的要求，对教学各主要环节进行了调整，使数学科学学院的全体学生在数学分析、高等代数、几何学、计算机等主干基础课程上，接受学时充分、强度足够的严格训练；在对学生分流培养阶段，我们在课程内容上坚决贯彻“少而精”的原则，大力压缩后续课程中多年逐步形成的过窄、过深和过繁的教学内容，为新的培养方向、实践性教学环节，以及为培养学生的创新能力所进行的基础科研训练争取到了必要的学时和空间。这样既使学生打下宽广、坚实的基础，又充分照顾到每个人的不同特长、爱好和发展取向。与上述改革相适应，积极而慎重地进行教学计划的修订，适当压缩常微、复变、偏微、实变、微分几何、抽象代数、泛函分析等后续课程的周学时。并增加了数学模型和计算机的相关课程，使学生有更大的选课余地。

在研究生教育中，在注重专题课程的同时，我们制定了 30 多门研究生普选基础课程（其中数学系 18 门），重点拓宽学生的专业基础和加强学生对数学整体发展及最新进展的了解。

教材建设是教学成果的一个重要体现。与修订的教学计划相

配合，我们进行了有组织的教材建设，计划自 1999 年起用 8 年的时间修订、编写和出版 40 余种教材，这就是将陆续呈现在大家面前的《北京大学数学教学系列丛书》。这套丛书凝聚了我们近十年在人才培养方面的思考，记录了我们教学实践的足迹，体现了我们教学改革的成果，反映了我们对新世纪人才培养的理念，代表了我们新时期的数学教学水平。

经过 20 世纪的空前发展，数学的基本理论更加深入和完善，而计算机技术的发展使得数学的应用更加直接和广泛，而且活跃于生产第一线，促进着技术和经济的发展，所有这些都正在改变着人们对数学的传统认识。同时也促使数学研究的方式发生巨大变化。作为整个科学技术基础的数学，正突破传统的范围而向人类一切知识领域渗透。作为一种文化，数学科学已成为推动人类文明进化、知识创新的重要因素，将更深刻地改变着客观现实的面貌和人们对世界的认识。数学素质已成为今天培养高层次创新人才的重要基础。数学的理论和应用的巨大发展必然引起数学教育的深刻变革。我们现在的改革还是初步的。教学改革无禁区，但要十分稳重和积极；人才培养无止境，既要遵循基本规律，更要不断创新。我们现在推出这套丛书，目的是向大家学习。让我们大家携起手来，为提高中国数学教育水平和建设世界一流数学强国而共同努力。

张继平

2002 年 5 月 18 日

于北京大学蓝旗营

前　言

多元统计分析是数理统计学30多年来迅速发展起来的一个分支。特别在计算机非常普及、各种统计分析软件不断推出的今天，多元统计分析方法已广泛地应用到社会科学和自然科学的许多领域中。北京大学概率统计系自1985年成立以来，一直开设"应用多元统计分析"课程。编者在近20年来教学和科研的基础上，编写了《应用多元统计分析》一书。本书的目的是介绍一些实用的多元统计分析方法的理论及其应用，并以国际上著名的标准统计分析软件SAS系统作为典型工具，通过实例介绍如何用统计软件处理数据分析中的各种实际问题。

本书共有十一章及附录。第一章"绪论"介绍多元统计分析研究的对象，应用领域及多元数据的图表示法；第二章介绍多元正态分布及其参数的估计和性质；第三章首先介绍三个重要分布，即威沙特(Wishart)分布、霍特林(Hotelling) T^2 分布、威尔克斯(Wilks)分布及它们的性质，然后讨论多元正态总体中参数的假设检验问题；第四章"回归分析"重点介绍多因变量的多元线性回归的有关问题，包括模型、参数的估计及其性质、假设检验、变量筛选，以及双重筛选逐步回归问题。从第五章至第十章介绍另一些常用的统计方法，如判别分析、聚类分析、主成分分析、因子分析、对应分析方法以及典型相关分析。第十一章介绍近年来发展的偏最小二乘回归分析方法。在"附录"中介绍了本课程所需的矩阵代数的有关内容。书末还给出书中部分习题参考解答或提示。

"应用多元统计分析"是一门应用性很强的课程。本书不仅介绍了各种常用的多元统计分析方法的统计背景和实际意义，说明该方法的统计思想、数学原理及解题步骤，还列举出各方面的应用实例。本书将多元统计方法的介绍与在计算机上实现这些方法的统计软件(SAS系统)结合起来，使读者不仅学到统计方法的理论知识，还知

配合，我们进行了有组织的教材建设，计划自 1999 年起用 8 年的时间修订、编写和出版 40 余种教材，这就是将陆续呈现在大家面前的《北京大学数学教学系列丛书》。这套丛书凝聚了我们近十年在人才培养方面的思考，记录了我们教学实践的足迹，体现了我们教学改革的成果，反映了我们对新世纪人才培养的理念，代表了我们新时期的数学教学水平。

经过 20 世纪的空前发展，数学的基本理论更加深入和完善，而计算机技术的发展使得数学的应用更加直接和广泛，而且活跃于生产第一线，促进着技术和经济的发展，所有这些都正在改变着人们对数学的传统认识。同时也促使数学研究的方式发生巨大变化。作为整个科学技术基础的数学，正突破传统的范围而向人类一切知识领域渗透。作为一种文化，数学科学已成为推动人类文明进化、知识创新的重要因素，将更深刻地改变着客观现实的面貌和人们对世界的认识。数学素质已成为今天培养高层次创新人才的重要基础。数学的理论和应用的巨大发展必然引起数学教育的深刻变革。我们现在的改革还是初步的。教学改革无禁区，但要十分稳重和积极；人才培养无止境，既要遵循基本规律，更要不断创新。我们现在推出这套丛书，目的是向大家学习。让我们大家携起手来，为提高中国数学教育水平和建设世界一流数学强国而共同努力。

张 继 平

2002 年 5 月 18 日

于北京大学蓝旗营

前　言

多元统计分析是数理统计学30多年来迅速发展起来的一个分支。特别在计算机非常普及、各种统计分析软件不断推出的今天，多元统计分析方法已广泛地应用到社会科学和自然科学的许多领域中。北京大学概率统计系自1985年成立以来，一直开设“应用多元统计分析”课程。编者在近20年来教学和科研的基础上，编写了《应用多元统计分析》一书。本书的目的是介绍一些实用的多元统计分析方法的理论及其应用，并以国际上著名的标准统计分析软件SAS系统作为典型工具，通过实例介绍如何用统计软件处理数据分析中的各种实际问题。

本书共有十一章及附录。第一章“绪论”介绍多元统计分析研究的对象，应用领域及多元数据的图表示法；第二章介绍多元正态分布及其参数的估计和性质；第三章首先介绍三个重要分布，即威沙特(Wishart)分布、霍特林(Hotelling) T^2 分布、威尔克斯(Wilks)分布及它们的性质，然后讨论多元正态总体中参数的假设检验问题；第四章“回归分析”重点介绍多因变量的多元线性回归的有关问题，包括模型、参数的估计及其性质、假设检验、变量筛选，以及双重筛选逐步回归问题。从第五章至第十章介绍另一些常用的统计方法，如判别分析、聚类分析、主成分分析、因子分析、对应分析方法以及典型相关分析。第十一章介绍近年来发展的偏最小二乘回归分析方法。在“附录”中介绍了本课程所需的矩阵代数的有关内容。书末还给出书中部分习题参考解答或提示。

“应用多元统计分析”是一门应用性很强的课程。本书不仅介绍了各种常用的多元统计分析方法的统计背景和实际意义，说明该方法的统计思想、数学原理及解题步骤，还列举出各方面的应用实例。本书将多元统计方法的介绍与在计算机上实现这些方法的统计软件(SAS系统)结合起来，使读者不仅学到统计方法的理论知识，还知

道如何解决实际问题。书中全部实例都是用 SAS 系统完成分析计算，并且每一章都配有适量的习题，其中大部分习题都附有参考解答或提示，以便于读者自学。

本书是北京大学数学科学学院概率统计系为开设的限选专业课“应用多元统计分析”所编写的教材。国内目前虽有一些介绍多元统计方法的教材，因偏重的方面不相同，并不能很好地满足要求。国外这方面较好的教材目前虽已有中译本，但由于篇幅太大给学生增加了经济上的负担。为达到本课程所要求的目的，编者在已编写的讲义基础上，通过反复使用、多次修改后编写出版了此书。

本书的读者对象是理工科类、经济类，特别是统计学学科等各专业学习应用统计的本科生，以及其他各个领域中需要进行数据分析处理的实际工作者。本书适用于每周 3～4 学时、每学期约讲授 54～72 学时“应用多元统计分析”课程或相关课程的教材，其中有些内容可供任课教师酌情选用。

本书因篇幅关系，应用实例的 SAS 程序没有在正文中给出，正文中只列出主要计算结果。为方便读者学习与掌握本书内容，我们另准备了《应用多元统计分析》附盘（3 寸软盘）一张，其内容包括正文所有实例的 SAS 程序，各章所有练习题的原始数据及用编程方法解答的 SAS 程序，以供读者参考。需要此附盘的读者请从网站“ftp://162.105.69.120/gaohx”上下载附盘上的文件，或与北京大学数学科学学院（邮编：100871）作者联系。

高惠璇

2003 年 7 月于北京大学

目　录

第一章 绪 论

§1.1 引 言

多元统计分析(简称多元分析)是运用数理统计的方法来研究多变量(多指标)问题的理论和方法,它是一元统计学的推广.

在实际问题中,很多随机现象涉及到的变量不是一个,而经常是多个变量,并且这些变量间又存在一定的联系.我们常常需要处理多个变量的观测数据.例如考察学生的学习情况时,就需了解学生在几个主要科目的考试成绩.表 1.1 给出某年级随机抽取的 12 名学生 5 门主课期末考试的成绩.

表 1.1 12 名学生 5 门课程的考试成绩

序号	政治(X_1)	语文(X_2)	外语(X_3)	数学(X_4)	物理(X_5)
1	99	94	93	100	100
2	99	88	96	99	97
3	100	98	81	96	100
4	93	88	88	99	96
5	100	91	72	96	78
6	90	78	82	75	97
7	75	73	88	97	89
8	93	84	83	68	88
9	87	73	60	76	84
10	95	82	90	62	39
11	76	72	43	67	78
12	85	75	50	34	37

表 1.1 提供的数据,如果用一元统计方法,势必要对多门课程分别分析,每次分析处理一门课程的成绩.这样处理,由于忽视了课程之间可能存在的相关性,因此,一般说来,丢失信息太多,分析的结果

不能客观全面地反映某年级学生的学习情况. 本书将要讨论的多元统计方法,它同时对多门课程的成绩进行分析. 这样的分析对诸课程间的关系、相依性和相对重要性等都能提供有用的信息. 如果说一元统计分析是研究一个随机变量统计规律性的学科,那么多元统计分析则是研究多个随机变量之间相互依赖关系以及内在统计规律性的一门统计学科.

由于大量实际问题都涉及到多个变量,这些变量又是随机变量,如学生的学习成绩随着被抽取学生的不同,成绩也有变化(我们往往需要依据它们来推断全年级的学习情况). 所以要讨论多元随机变量的统计规律性. 多元统计分析就是讨论多元随机变量的理论和统计方法的总称. 其内容既包括一元统计学中某些方法的直接推广,也包括多元随机变量特有的一些问题. 多元统计分析是一类范围很广的理论和方法.

就以学生成绩为例,我们可以研究很多问题:用各科成绩的总和作为综合指标,来比较学生学习成绩的好坏;根据各科成绩相近程度对学生进行分类(如成绩好的与成绩差的,又如文科成绩好的与理科成绩好的);研究各科成绩之间的关系(如物理与数学成绩的关系,文科成绩与理科成绩的关系);等等. 所有这些都属于多元统计分析的研究内容.

综上所述,多元统计分析是以 p 个变量的 n 次观测数据所组成的数据矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

为依据的. 根据实际问题的需要,给出种种方法. 英国著名统计学家肯德尔(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)

简化数据结构即是将某些较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的.例如主成分分析、因子分析,以及对应分析等多元统计方法就是这样的一类方法.

2. 分类与判别(归类问题)

归类问题即是对所考察的观测点(或变量)按相似程度进行分类(或归类).例如聚类分析和判别分析等方法就是解决这类问题的统计方法.

3. 变量间的相互联系

(1) 相互依赖关系:分析一个或几个变量的变化是否依赖于另一些变量的变化?如果是,建立变量间的定量关系式,并用于预测或控制——回归分析.

(2) 变量间的相互关系:分析两组变量间的相互关系——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及协方差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及多维正态随机向量,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.这些不仅是统计估计和假设检验的基础,也是多元统计分析的理论基础.

多元统计分析起源于20世纪初,1928年威沙特(Wishart)发表的论文《多元正态总体样本协方差阵的精确分布》,可以说是多元分析的开端.之后费希尔(Fisher)、霍特林(Hotelling)、罗伊(Roy)、许宝騄等人作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,在许多领域中也有了实际应用.由于用统计方法解决实际问题时需要的计算量很大,使其发展受到影响,甚至停滞了相当长的时间.20世纪50年代中期,随着电子计算机的出现和发展,使得多元统计分析在地质、气象、医学、社会学等方面得到广泛的应用.60

年代通过应用和实践又完善和发展了理论，由于新理论、新方法的不断出现又促使它的应用范围更加扩大. 70 年代初期在我国才受到各个领域的极大关注，近 30 年来我国在多元统计分析的理论研究和应用上也取得了很多显著成绩，有些研究工作已达到国际水平，并已形成一支科技队伍，活跃在各条战线上.

§1.2　多元统计分析的应用

多元统计分析是解决实际问题的有效的数据处理方法. 随着电子计算机使用的日益普及，多元统计方法已广泛地应用于自然科学、社会科学的各个方面. 以下我们列举多元统计分析的一些应用领域.

一、教育学

n 个考生报考北京大学概率统计系. 每个考生参加 p 门课(语文、数学、政治、外语、物理、化学……)的考试，各门课的成绩记为 $y_{i1}, y_{i2}, \cdots, y_{ip} (i=1,2,\cdots,n)$. 又每个考生在高中学习期间，$m$ 门主要课程成绩为 $x_{i1}, x_{i2}, \cdots, x_{im} (i=1,2,\cdots,n)$. 经过对这些大量的资料作统计分析，我们能够得出：

(1) 高考成绩和高中学习期间成绩的关系，即给出两组变量线性组合间的关系，从而可由考生在高中学习期间的成绩来预测高考的综合成绩或某些科目的成绩.

(2) 给出考生成绩次序排队的最佳方案(最佳组合). 总分可以体现一个考生成绩好坏，但对报考概率统计系的学生，按总分从高到低的顺序录取并不是很合适的，如果按适当的权重加权求和，比如数学、物理、外语的权重相对高些，然后按加权和的顺序录取也许更合适些.

此外利用 n 个学生在高中学习期间 m 门主要课程的考试成绩，可对学生进行分类，如按文、理科成绩分类，按总成绩分类等. 若准备给优秀学生发奖，那么一等奖、二等奖的比例应该是多少？应用多元统计分析的方法可以给出公平合理地确定.

二、医学

医生对病人的诊断是靠对病人观测若干症状后来综合评定的.如一个人发高烧,医生根据他的体温高低、白血球数目及其他症状来判断他是患感冒、肺炎还是其他疾病.再比如某人发现其腹部有肿瘤,医生根据肿瘤的大小、生长的速度、边界是否清楚,以及质硬或软等症状来判断肿瘤是良性或恶性.

为了判断更为准确可靠,事先应有一批经专家确诊或手术后经病理化验确诊的病例资料,根据这批资料利用多元统计方法可建立诊断的准则(即专家系统).对来就诊的病人,按专家系统的要求,观测若干项指标后,根据诊断准则,即可作出诊断.

三、气象学

全国各地建立了很多气象站,在不同时间各气象站都记录了降雨量、气温、气压、湿度、风速、风向等气象指标资料.对这些资料作统计分析,可以得出:

(1) 指标间的关系,如降雨量与前一天的气温、气压、湿度等的关系,利用该关系可对降雨量作预报.

(2) 不同地点气象指标之间的关系,如某地有气象站,长期记录各项气象指标的资料.今计划在该站附近建一大型化工厂,厂区的气象条件是我们关心的,而在此处新建一气象站又不可能.最后采用的办法是在该厂区临时建一个观测站,与气象站同时测定气象指标;然后利用这些资料用多元统计分析方法建立两地气象指标的关系,以达到今后可由气象站的气象资料来预报该厂区的气象情况.

四、环境科学

(1) 为了了解某大型化工厂对环境的污染程度,在厂区建立很多监测点,每天定时测定各种污染气体的浓度.用统计分析方法分析处理这些资料,可对厂区按污染情况分为几类,如分为严重污染、一般污染和轻污染三类;并为今后监测点的布局提供既合理又经济的方案.

(2) 许多学者研究了洛杉矶地区大气中污染物质的浓度. 在较长的一段时间内,每天定时测定该地区与污染有关的几个指标值,利用多元统计检验的方法,首先判断洛杉矶地区空气污染程度在一周内是固定不变或周末与平时有显著差异;其次对这庞杂的观测数据用一种易解释的方法加以归纳化简.

五、地质学

随着电子计算机的普及及地质科学向定量化发展,地质学和数学(主要是多元统计方法)结合起来产生了边缘学科——数学地质,多元分析是其主要内容之一. 王学仁先生在《地质数据的多变量统计分析》一书中介绍了多元分析方法及其在地质学中的应用. 应用多元统计方法处理各种地质观测数据,对成矿规律的评价、矿产预测、构造解释推断、勘探工程部署等等都得出了一些定量的依据,并获得了一些找矿信息.

六、考古学

(1) 考古学家根据一群坟墓中的陪葬品(特别是陶瓷和珠宝),利用它们在式样和装饰上的差别,把它们按时间顺序排列起来.

(2) 考古学家在古代墓地上,挖掘出若干个头盖骨,它们可能都是来自同一种族,或两个对抗种族(战死的战友和敌人都可能被埋在同一个坑内). 对每个头盖骨可测得多种数据,利用头盖骨的数据来判断所属的种族或性别;并研究最佳的测量法及最少的测量项目.

七、服装工业——服装的定型分类问题

一个服装公司希望生产足够多的成衣以适应大多数顾客的要求,而且使不合身的和卖不出去的服装尽量减少,这样不仅可满足社会需要且厂方也才可能多赚钱. 为此目的,首先在各地做抽样调查,对被调查人测量其身体的几十个部位的尺寸,然后对庞大的调查资料用多元统计方法分析处理,确定一种服装究竟需要有几种型号,每种型号服装的比例是多少,由身体的哪几个主要部位的尺寸决定.

八、经济学

(1) 构造中国国民收入的生产、分配与最终使用的计量经济模型. 例如根据我国几十年来财政收入与国民收入、工农业总产值、人口、就业人口、固定投资等因素相关,利用回归方法建立预测模型,以用于对今后的财政收入作预测.

(2) 在商业经济中,常常需要将很复杂的数据综合成商业指数形式,如物价指数、货币工资比、生活费用指数、商业活动指数等,用主成分分析可以从多个变量中构造出所需的商业指数.

(3) 为了研究不同地区农民收支的分布规律,抽样调查了全国28个省、市、自治区的农民生活消费支出情况,如食品、衣着、燃料、住房、生活用品、文化生活等的消费. 用聚类分析方法对28个地区分类,根据分类结果还可进一步研究各类地区农民的生活水平、富裕程度,以便进一步研究经济发展对策.

(4) 在经济学中,根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型.

九、农业

(1) 有 n 个不同地区,每个地区记录多种农作物的收获量,用多元统计方法对各个地区的总生产效率进行比较,并对不同的农业区域进行分类.

(2) 为了节省能源,对某地农用的手扶拖拉机的能源消耗进行抽样调查. 调查的内容为拖拉机在田间、运输、排灌、加工等作业时的燃油耗,以及在册月数、年平均更换零件数及平均燃油耗. 通过对调查资料作统计分析,达到对拖拉机的平均燃油耗作预测并对拖拉机进行分类,划分为淘汰类、大修类、小修类和继续使用类.

十、社会科学

青少年犯罪问题是一个很大的社会问题. 对待青少年犯罪,我们采取"以防为主、防重于治"的原则. 要预防犯罪,除了加强经常性的

教育外，还必然提出预测犯罪的问题. 如对青少年犯罪心理和行为倾向性在其犯罪行为发生之前便能进行预测，争取把它们消灭在萌芽状态，才能做到实际预防.

为此目的，1981 至 1982 年间中央教育科学研究所等几个单位协作进行了调查研究工作，调查对象为一般中学生，以及工读学校、少管所、劳教农场和劳改农场的青少年. 调查内容有两大方面：心理因素（如物质追求感、隔离感、无目的感、团伙义气感……）和外部因素（如性别、家庭平均收入、每月零花钱……）共 25 项指标，用多元统计方法分析处理这一批资料，找出青少年犯罪诸因素间的互相关系及其与犯罪行为的内在联系，并用反映数量关系的数学模型表示出来，用以描述青少年犯罪这一社会现象在个体身上的内在联系或变化规律，并借助这个模型对其他个体特征发展的趋向性进行比较科学的预测.

十一、文学

自从 30 年代末英国著名的统计学家尤尔（Yule）把统计方法引入到文学词汇的研究以来，这个领域已经取得不少进展，其中最有名的是 Mosteller 与 Wallace 在 60 年代初对美国立国三大历史文献之一的《联邦主义者》文集的研究.

在 1985 至 1986 年间复旦大学统计运筹系的李贤平教授对我国的名著《红楼梦》的著作权进行研究. 使用的统计方法主要是多元分析. 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）选定的这些虚词（即变量）出现的频数. 由此得到的数据阵作为分析的依据.

在《红楼梦》著作权的研究中使用较多的方法是聚类分析、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，这证实了前 80 回和后 40 回不是出于同一个人的手笔；

（2）前 80 回是否为曹雪芹所写？通过用曹雪芹的另一著作，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.

十二、其他

多元统计分析方法在其他很多领域中也有它的应用. 比如体育科研、军事科学、生物学、心理学、生态学、保险科学、火警预报、地震预报、中医阴阳学说研究等.

§1.3 多元统计数据的图表示法

图形有助于对所研究的数据的直观了解,一元或二元数据的一维或二维图形容易得到,三维图形虽也可以画出,但并不方便. 三维以上图形怎么表示?许多统计学家给出了多元数据的图示方法,但这方面的研究还处于不成熟状态,目前尚未有公认的方法. 这里介绍几种国际上近几十年来出现的方法,其中有一些依赖人工容易实现,但是有一些方法若是没有计算机的帮助,恐怕较难实现.

设变量个数为 p,观测次数为 n,第 k 次观测值记为

$$X_{(k)} = (x_{k1}, x_{k2}, \cdots, x_{kp}) \quad (k = 1, 2, \cdots, n),$$

n 次观测数据组成的矩阵记为 $X = (x_{ij})_{n\times p}$.

一、轮廓图

轮廓图的作图步骤为:

(1) 作直角坐标系,横坐标取 p 个点,以表示 p 个变量;

(2) 对给定的一次观测值,在 p 个点上的纵坐标(即高度)与对

应的变量取值成正比；

(3) 连结此 p 个点得一折线，即为该次观测值的一条轮廓线；

(4) 对于 n 次观测值，每次都重复上述步骤，可画出 n 条折线，构成 n 次观测值的轮廓图.

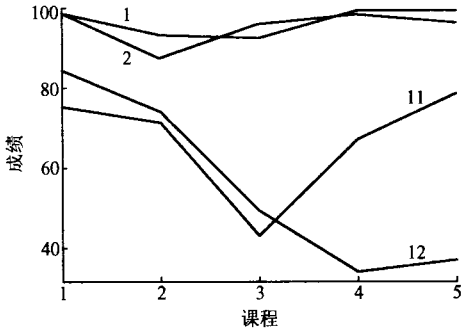

图 1.1　4 名学生学习成绩的轮廓图

如图 1.1 中 4 条折线为由表 1.1 给出的序号为 1、2 和 11、12 这 4 个学生学习成绩的轮廓线. 由该轮廓图可直观看出，哪几个学生成绩相似、哪些属优秀、哪些中等、哪些较差；对各门课程而言，也可直观地看出各课程成绩的好坏和分散情况等等. 这种图形在聚类分析中颇有帮助.

二、雷达图

雷达图的作图步骤是：

(1) 作一圆，并把圆周分为 p 等分；

(2) 连结圆心和各分点，把这 p 条半径依次定义为各变量的坐标轴，并标以适当的刻度；

(3) 对给定的一次观测值，把 p 个变量值分别取在相应的坐标轴上，然后将它们连结成一个 p 边形；

(4) n 次观测值可画出 n 个 p 边形.

这种图形既像雷达荧光屏上看到的图像，也像一个蜘蛛网. 因此有人称为雷达图，也有人称为蜘蛛图. 图 1.2 为表 1.1 中序号为 1 和 12 的学生学习成绩的雷达图. 各科都达到 100 分的学生对应着一个

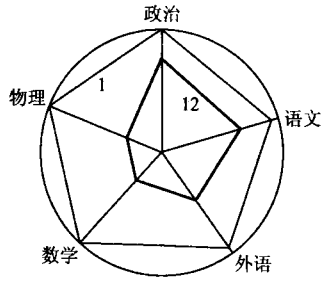

图 1.2 两名学生学习成绩的雷达图

面积最大的正五边形，如学生序号为 1 的图形接近正五边形，因此是学习成绩优秀的学生；另外，学习成绩差的学生，其图形面积也小，如学生序号为 12 的就是如此，而且其图形明显偏右上方，这意味着该学生的数学、物理和外语成绩极差，而语文和政治还算过得去.

当观测次数 n 较大时，为了获得较好的效果，每张图可以只画少数几次观测值，甚至只画一次观测值；为使图形效果更好，在雷达图中适当分配变量的坐标轴并选取合适的尺度是十分重要的. 如在学生成绩的雷达图中，有意识地把理科成绩分配在左边坐标轴上，文科在右边，则可根据图形偏左或偏右看出该学生是偏理还是偏文.

三、调和曲线图

从数学上看，较为完美的多元数据图表示方法可能是 Andcews 在 1972 年提出的三角多项式表示法，其思想是把多维空间中的一个点对应于二维平面上的一条曲线.

设 p 维数据 $X=(x_1,x_2,\cdots,x_p)'$（注：上角“′”表示转置，即行(列)转换为列(行)），则对应的曲线是

$$f_X(t)=\frac{x_1}{\sqrt{2}}+x_2\sin t+x_3\cos t+x_4\sin 2t+x_5\cos 2t+\cdots$$

$$(-\pi\leqslant t\leqslant\pi).$$

上式当 t 在区间$(-\pi,\pi)$上变化时，其轨迹是一条曲线.

例如表 1.1 学生成绩数据中，学生 1 对应的曲线为

$$f_1(t)=\frac{99}{\sqrt{2}}+94\sin t+93\cos t+100\sin 2t+100\cos 2t,$$

学生 12 对应的曲线为

$$f_{12}(t)=\frac{85}{\sqrt{2}}+75\sin t+50\cos t+34\sin 2t+37\cos 2t,$$

它们的图形如图 1.3.

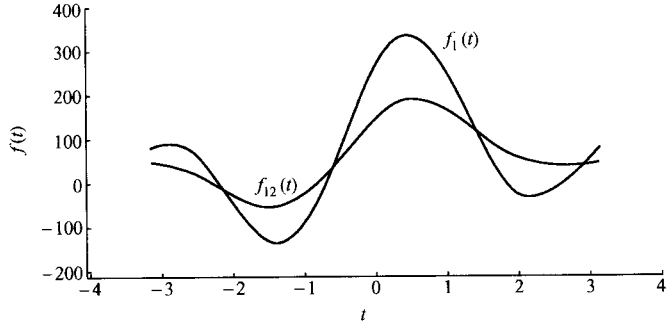

图 1.3　两名学生学习成绩的调和曲线图

n 次观测数据对应 n 条曲线，画在同一平面上就是一张调和曲线图. 在多项式的图表示中，当各变量的数值太悬殊时，最好先标准化——例如标准差标准化、极差标准化或极差正规化等后再作图. Andrews 证明了三角多项式图有许多很好的性质(见参考文献[1]).

作调和曲线图时一般要借助计算机作图，这种图对聚类分析帮助很大. 如果选择聚类统计量为距离，则同类的曲线拧在一起，不同类的曲线拧成不同的束，非常直观.

四、散布图矩阵

当 $p=2$ 时，常把 n 次二元观测数据点在平面上生成一张散布图，由散布图可以直观地看出变量 X 与 Y 之间的相关关系及相关的程度. 当 $p>2$ 时，我们也想借助散布图来直观给出变量之间，观测点之间的关系，可以对 p 个变量两两配对生成一张散布图矩阵. 通过这张图，不仅可以了解到每两个变量间的相关情况，在 SAS 系统中，还可通过“刷亮”方法来找出异常点(见参考文献[21]). 下面图 1.4 是 12 名学生 5 门课程成绩的散布图矩阵.

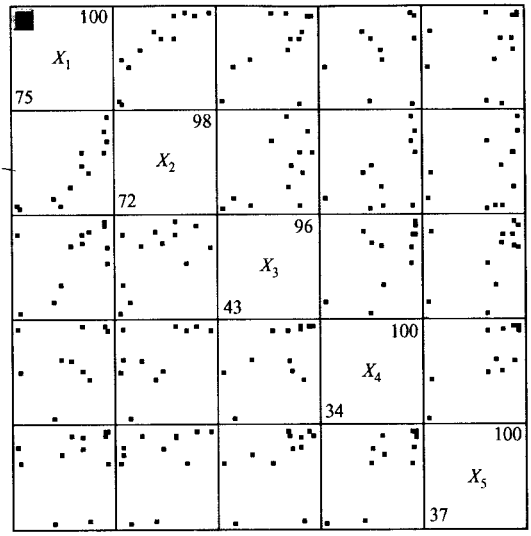

图 1.4 12 名学生学习成绩的散布图矩阵

五、其他方法

在多元数据的图表示法中，还有星座图、脸谱图、装饰图等表示法. 最为浪漫的可能是脸谱图，它把多元数据表示成一张张脸谱图. 脸的轮廓由上下两个椭圆构成，这些椭圆的长短轴及离心率等均由多元数据中某些变量来刻画；另一些变量决定鼻子长度，嘴的位置及圆弧的长度与向上还是向下，眼睛的大小，眼珠的位置，眉毛的角度等，如果变量很多，脸谱可以刻画得细致些，变量不多，则把一部分器官形态固定，只让另一部分器官变化. 在实际应用中，脸谱图也有发展，如在脸谱上加眼泪以表示很坏情况的出现；还可以在脸谱基础上加上体型，用一些变量来决定体型的胖瘦和高矮等.

最后我们指出，多元数据图表示法的难点在于变量过多. 如果有一种方法可以把高元数据投影到二维空间(平面)中去，并且在投影过程中不会过多地损失原有数据信息的话，就可以使用通常方法在平面上画出这些本来是高维数据的图形来. 后面将要介绍的主成分分析等方法就是一些降维的方法.

习　题　一

1-1　为了研究人体的心肺功能，对 31 个成年男子测量了肺活量(OXY)，并且记录了他们的年龄(age)、体重(weight)，以及简单训练后的测试数据：跑 1.5 英里的时间(time)、休息时的脉搏(spulse)、跑步时的脉搏(rpulse)和跑步时记录的最大脉搏(mpulse)，共 7 项指标(数据见表 1.2).

(1) 分别绘制 OXY 与 time 和 age 的散布图，从图中可得出什么结论？

(2) 绘制 7 项指标的散布图矩阵，从这里能否直观看出一些结论；

(3) 绘制序号为 1,2,21,22 的 4 个人的轮廓图和雷达图；

(4) 绘制序号为 1,2,21,22 的 4 个人的调和曲线图(放在同一张图上).

表 1.2　肺活量与其他指标的数据

序号	age	weight	time	spulse	rpulse	mpulse	OXY
1	57	73.37	12.63	58	174	176	39.407
2	54	79.38	11.17	62	156	165	46.080
3	52	76.32	9.63	48	164	166	45.441
4	50	70.87	8.92	48	146	155	54.625
5	51	67.25	11.08	48	172	172	45.118
6	54	91.63	12.88	44	168	172	39.203
7	51	73.71	10.47	59	186	188	45.790
8	57	59.08	9.93	49	148	155	50.545
9	49	76.32	9.40	56	186	188	48.673
10	48	61.24	11.50	52	170	176	47.920
11	52	82.78	10.50	53	170	172	47.467
12	44	73.03	10.13	45	168	168	50.541
13	45	87.66	14.03	56	186	192	37.388
14	45	66.45	11.12	51	176	176	44.754
15	47	79.15	10.60	47	162	164	47.273
16	54	83.12	10.33	50	166	170	51.855

（续表）

序号	age	weight	time	spulse	rpulse	mpulse	OXY
17	49	81.42	8.95	44	180	185	49.156
18	51	69.63	10.95	57	168	172	40.836
19	51	77.91	10.00	48	162	168	46.672
20	48	91.63	10.25	48	162	164	46.774
21	49	73.37	10.08	76	168	168	50.388
22	44	89.47	11.37	62	178	182	44.609
23	40	75.07	10.07	62	185	185	45.313
24	44	85.84	8.65	45	156	168	54.297
25	42	68.15	8.17	40	166	172	59.571
26	38	89.02	9.22	55	178	180	49.874
27	47	77.45	11.63	58	176	176	44.811
28	40	75.98	11.95	70	176	180	45.681
29	43	81.19	10.85	64	162	170	49.091
30	44	81.42	13.08	63	174	176	39.442
31	38	81.87	8.63	48	170	186	60.055

第二章　多元正态分布及参数的估计

在多元统计分析中，多元正态分布占有相当重要的地位. 这是因为许多实际问题涉及到的随机向量服从正态分布或近似服从正态分布；当样本量很大时，许多统计量的极限分布往往和正态分布有关. 此外，对多元正态分布，理论与实践都比较成熟，已有一整套行之有效的统计推断方法. 基于这些理由，我们在介绍多元统计分析的种种具体方法之前，首先介绍多元正态分布的定义、性质及多元正态分布中参数的估计问题.

§2.1　随机向量

本课程所讨论的是多变量总体. 把 p 个随机变量放在一起得 $X=(X_1,X_2,\cdots,X_p)'$ 为一个 p 维随机向量，如果同时对 p 个变量作一次观测，得观测值：$(x_{11},x_{12},\cdots,x_{1p})\overset{\text{def}}{=\!=}X'_{(1)}$，它是一个样品. 观测 n 次得 n 个样品：$X'_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})$ $(i=1,2,\cdots,n)$，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，称为**样本数据阵**(或**样本资料阵**)，记为

$$X=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1p}\\ x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix}X'_{(1)}\\ X'_{(2)}\\ \vdots\\ X'_{(n)}\end{bmatrix}$$

$$\overset{\text{或 def}}{=\!=}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p).$$

矩阵 X 的第 i 行：$X'_{(i)}=(x_{i1},x_{i2},\cdots,x_{ip})$ $(i=1,2,\cdots,n)$表示对第 i

个样品的观测值，在具体观测之前，它是一个 p 维的随机向量. 矩阵 X 的第 j 列

$$\boldsymbol{X}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \quad (j = 1,2,\cdots,p)$$

表示对第 j 个变量的 n 次观测，在具体观测之前，它是一个 n 维随机向量；而样本数据阵 X 是一个随机阵.

在多元统计分析中涉及到的都是随机向量，或是多个随机向量放在一起组成的随机阵. 本节首先来回顾一下随机向量的有关内容.

一、随机向量的联合分布，边缘分布，条件分布

1. 联合分布

设 $X=(X_1,X_2,\cdots,X_p)'$ 是 p 维随机向量，称 p 元函数

$$F(x_1,\cdots,x_p) = P\{X_1 \leqslant x_1,\cdots,X_p \leqslant x_p\}$$

为 X 的**联合分布函数**.

若存在非负函数 $f(x_1,x_2,\cdots,x_p)$，使得随机向量 X 的联合分布函数对一切 $(x_1,x_2,\cdots,x_p)\in \mathbb{R}^p$ 均可表示为

$$F(x_1,\cdots,x_p) = \int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_p} f(x_1,\cdots,x_p)\mathrm{d}x_1\cdots\mathrm{d}x_p,$$

则称 X 为**连续型随机向量**，称 $f(x_1,x_2,\cdots,x_p)$ 为 X 的**联合概率密度函数**，简称为**多元密度函数**或**密度函数**.

多元密度函数 $f(x_1,x_2,\cdots,x_p)$ 满足以下两条性质：

(1) $f(x_1,\cdots,x_p)\geqslant 0$，对一切实数 $x_1,x_2,\cdots,x_p$；

(2) $\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1,\cdots,x_p)\mathrm{d}x_1\cdots\mathrm{d}x_p = 1.$

2. 边缘分布

称随机向量 X 的部分分量 $(X_{i_1},\cdots,X_{i_m})$ $(1\leqslant m<p)$ 的分布为**边缘分布**.

设 $X^{(1)}$ 为 r 维随机向量，$X^{(2)}$ 为 $p-r$ 维随机向量. 若 p 维随机向量 $X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}$，则 $X^{(1)}$ 的边缘分布为

$$f_1(x^{(1)})=f_1(x_1,\cdots,x_r)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\cdots,x_p)\mathrm{d}x_{r+1}\cdots\mathrm{d}x_p,$$

$X^{(2)}$ 的边缘分布为

$$f_2(x^{(2)})=f_2(x_{r+1},\cdots,x_p)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(x_1,\cdots,x_p)\mathrm{d}x_1\cdots\mathrm{d}x_r.$$

例 2.1.1　设二维随机向量 $X=(X_1,X_2)'$ 的联合密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\left[1+x_1x_2\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\right],$$

试求 X_1 和 X_2 关于随机向量 X 的边缘密度.

解　首先可验证 $f(x_1,x_2)$ 满足联合密度函数的两条性质. 再利用边缘密度的计算公式，有

$$\begin{aligned}f_1(x_1)&=\int_{-\infty}^{\infty}\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\left[1+x_1x_2\mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\right]\mathrm{d}x_2\\&=\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}x_1^2}\left[\int_{-\infty}^{\infty}\mathrm{e}^{-\frac{1}{2}x_2^2}\mathrm{d}x_2+x_1\mathrm{e}^{-\frac{1}{2}x_1^2}\int_{-\infty}^{\infty}x_2\mathrm{e}^{-x_2^2}\mathrm{d}x_2\right]\\&=\frac{1}{2\pi}\mathrm{e}^{-\frac{1}{2}x_1^2}[\sqrt{2\pi}+0]=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x_1^2},\end{aligned}$$

即 $X_1\sim N(0,1)$.

类似可得出 $X_2\sim N(0,1)$.

3. 条件分布

设 $X^{(1)}$ 为 r 维随机向量，$X^{(2)}$ 为 $p-r$ 维随机向量. 若 p 维随机向量 $X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}$，则当给定 $X^{(2)}$ 时，称 $X^{(1)}$ 的分布为**条件分布**. 当 X 的密度函数为 $f(x^{(1)},x^{(2)})$ 时，给定 $X^{(2)}$ 时 $X^{(1)}$ 的条件密度为

$$f_1(x^{(1)}|x^{(2)})=f(x^{(1)},x^{(2)})/f_2(x^{(2)}),$$

其中 $f_2(x^{(2)})$ 是 $X^{(2)}$ 的密度函数.

4. 独立性

设 $X_1,\cdots,X_p$ 是 p 个随机变量，X_i 的分布函数记为 $F_i(x_i)$ $(i=$

$1,\cdots,p$)；$F(x_1,\cdots,x_p)$是$(X_1,\cdots,X_p)'$的联合分布函数. 若对一切实数 $x_1,\cdots,x_p$,

$$F(x_1,\cdots,x_p) = F_1(x_1)\cdots F_p(x_p)$$

均成立，则称 $X_1,\cdots,X_p$ **相互独立**. 在连续型随机变量的情况下，$X_1,\cdots,X_p$ 相互独立，当且仅当 $X=(X_1,\cdots,X_p)'$的联合密度函数 $f(x_1,\cdots,x_p)$满足

$$f(x_1,\cdots,x_p) = f_1(x_1)\cdots f_p(x_p)$$

对一切实数 $x_1,\cdots,x_p$ 均成立，其中 $f_i(x_i)$是 X_i 的密度函数($i=1,\cdots,p$).

在例 2.1.1 中随机向量 X 的两个分量 X_1 和 X_2 互相不独立.

二、随机向量的数字特征

设 $X=(X_1,\cdots,X_p)'$, $Y=(Y_1,\cdots,Y_q)'$是两个随机向量.

1. 随机向量 X 的均值向量

若 $\mathrm{E}(X_i)=\mu_i$ 存在，则称

$$\mathrm{E}(X) = \begin{bmatrix} \mathrm{E}(X_1) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

为随机向量 X 的**均值向量**.

2. 随机向量 X 的协方差阵

若 X_i 和 X_j 的**协方差** $\mathrm{Cov}(X_i,X_j)$存在$(i,j=1,\cdots,p)$，则称

$$\begin{aligned} \mathrm{D}(X) &= \mathrm{E}[(X-\mathrm{E}(X))(X-\mathrm{E}(X))'] \\ &= \begin{bmatrix} \mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_p) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \cdots & \mathrm{Cov}(X_2,X_p) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_p,X_1) & \mathrm{Cov}(X_p,X_2) & \cdots & \mathrm{Cov}(X_p,X_p) \end{bmatrix} \\ &= (\sigma_{ij})_{p\times p} \overset{\text{def}}{=\!=} \Sigma \end{aligned}$$

为随机向量 X 的**协方差阵**.

3. 随机向量 X 和 Y 的协方差阵

若 X_i 和 Y_j 的协方差 $\mathrm{Cov}(X_i,Y_j)$ 存在($i=1,\cdots,p$; $j=1,\cdots,q$),则称

$$\mathrm{COV}(X,Y)=\mathrm{E}[(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))']$$

$$=\begin{bmatrix}\mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \cdots & \mathrm{Cov}(X_1,Y_q)\\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \cdots & \mathrm{Cov}(X_2,Y_q)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_p,Y_1) & \mathrm{Cov}(X_p,Y_2) & \cdots & \mathrm{Cov}(X_p,Y_q)\end{bmatrix}$$

为随机向量 X 和 Y 的协方差阵. 若

$$\mathrm{COV}(X,Y)=O \quad (\text{其中 } O \text{ 表示零矩阵}),$$

则称 X 与 Y **不相关**.

4. 随机向量 X 的相关阵

若 X_i 和 Y_i 的协方差 $\mathrm{Cov}(X_i,Y_i)$ 存在($i,j=1,2,\cdots,p$),称 $R=(r_{ij})_{p\times p}$ 为 X 的**相关阵**,其中

$$r_{ij}=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_i)}\sqrt{\mathrm{Var}(X_j)}}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad (i,j=1,2,\cdots,p).$$

这里

$$\mathrm{Var}(X_i)=\mathrm{Cov}(X_i,X_i)\xlongequal{\text{def}}\sigma_{ii},$$

为随机变量 X_i 的方差,而 $\sqrt{\sigma_{ii}}$ 为 X_i 的标准差($i=1,2,\cdots,p$).

若记 $V^{1/2}=\mathrm{diag}(\sqrt{\sigma_{11}},\cdots,\sqrt{\sigma_{pp}})$ 为**标准差矩阵**,则

$$\Sigma=V^{1/2}RV^{1/2} \quad \text{或} \quad R=(V^{1/2})^{-1}\Sigma(V^{1/2})^{-1}.$$

三、均值向量和协方差阵的性质

性质 1　设 X,Y 是随机向量,A,B 是常数矩阵,则

$$\mathrm{E}(AX)=A\mathrm{E}(X),$$
$$\mathrm{E}(AXB)=A\mathrm{E}(X)B,$$
$$\mathrm{D}(AX)=A\mathrm{D}(X)A',$$
$$\mathrm{COV}(AX,BY)=A\mathrm{COV}(X,Y)B'.$$

证明 我们只证明最后一公式：

$$\begin{aligned}\mathrm{COV}(AX,BY)&=\mathrm{E}[(AX-\mathrm{E}(AX))(BY-\mathrm{E}(BY))']\\&=\mathrm{E}[A(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))'B']\\&=A[\mathrm{E}(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))']B'\\&=A\mathrm{COV}(X,Y)B'.\end{aligned}$$

(证毕)

性质 2 若 X,Y 相互独立，则 $\mathrm{COV}(X,Y)=O_{p\times q}$；反之不一定成立.

性质 3 随机向量 $X=(X_1,X_2,\cdots,X_p)'$ 的协方差阵 $\mathrm{D}(X)=\Sigma$ 是对称非负定矩阵.

证明 因为 $\mathrm{Cov}(X_i,X_j)=\mathrm{Cov}(X_j,X_i)$，所以 $\Sigma=\Sigma'$. 对任给 $\alpha=(\alpha_1,\cdots,\alpha_p)'$，有

$$\begin{aligned}\alpha'\Sigma\alpha&=(\alpha_1,\cdots,\alpha_p)\mathrm{E}[(X-\mathrm{E}(X))(X-\mathrm{E}(X))']\begin{bmatrix}\alpha_1\\\vdots\\\alpha_p\end{bmatrix}\\&=\mathrm{E}[\alpha'(X-\mathrm{E}(X))\cdot(X-\mathrm{E}(X))'\alpha]\\&=\mathrm{E}[(\alpha'(X-\mathrm{E}(X)))^2]\geqslant 0,\end{aligned}$$

所以 $\Sigma\geqslant 0$，即 Σ 为非负定矩阵. **(证毕)**

性质 4 $\Sigma=L^2$，其中 L 为非负定矩阵.

证明 由于 $\Sigma\geqslant 0$ (非负定)，利用线性代数中实对称阵的对角化定理，存在正交矩阵 Γ，使得

$$\begin{aligned}\Sigma&=\Gamma\begin{bmatrix}\lambda_1&&0\\&\ddots&\\0&&\lambda_p\end{bmatrix}\Gamma'\quad(\text{其中 }\lambda_i\geqslant 0)\\&=\Gamma\begin{bmatrix}\sqrt{\lambda_1}&&0\\&\ddots&\\0&&\sqrt{\lambda_p}\end{bmatrix}\Gamma'\cdot\Gamma\begin{bmatrix}\sqrt{\lambda_1}&&0\\&\ddots&\\0&&\sqrt{\lambda_p}\end{bmatrix}\Gamma'\\&=L^2,\end{aligned}$$

其中 $L=\Gamma\mathrm{diag}(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_p})\Gamma'$，且 $L=L'$，所以 $L\geqslant 0$. **(证毕)**

当矩阵 $\Sigma>0$(正定)时，矩阵 L 也称为 Σ 的**平方根矩阵**，记为

$\Sigma^{1/2}$. 若令 $A=\Gamma\mathrm{diag}(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_p})$，则协方差阵 Σ 还有如下分解：$\Sigma=AA'$（A 为非退化方阵）.

§2.2　多元正态分布的定义与基本性质

在一元统计中，若 $U\sim N(0,1)$，则 U 的任意线性变换为 $X=\sigma U+\mu\sim N(\mu,\sigma^2)$. 利用这一性质，可以由标准正态分布来定义一般正态分布：若 $U\sim N(0,1)$，则称 $X=\sigma U+\mu$ 的分布为一般正态分布，记为 $X\sim N(\mu,\sigma^2)$. 此定义中，不必要求 $\sigma>0$，当 σ 退化为 0 时仍有意义. 把这种新的定义方式推广到多元情况，可得出多元正态分布的第一种定义.

定义 2.2.1　设 $U=(U_1,\cdots,U_q)'$ 为随机向量，$U_1,\cdots,U_q$ 相互独立且同 $N(0,1)$分布；设 μ 为 p 维常数向量，A 为 $p\times q$ 常数矩阵，则称 $X=AU+\mu$ 的分布为 p **元正态分布**，或称 X 为 p **维正态随机向量**，记为 $X\sim N_p(\mu,AA')$.

简单地说，由 q 个相互独立的标准正态随机变量的一些线性组合所构成的随机向量的分布，称其为**多元正态分布**.

在一元统计中，若 $X\sim N(\mu,\sigma^2)$，则 X 的特征函数为

$$\varphi(t)=\mathrm{E}(\mathrm{e}^{\mathrm{i}tX})=\exp\left[\mathrm{i}t\mu-\frac{1}{2}t^2\sigma^2\right].$$

将其推广到多维正态随机向量的情况有如下性质.

性质 1　设 $U=(U_1,\cdots,U_q)'$ 为随机向量，$U_1,\cdots,U_q$ 相互独立且同 $N(0,1)$分布；令 $X=AU+\mu$，则 X 的特征函数为

$$\Phi_X(t)=\exp\left[\mathrm{i}t'\mu-\frac{1}{2}t'AA't\right].$$

证明　根据随机向量特征函数的定义和性质，可知 X 的特征函数为

$$\begin{aligned}\Phi_X(t)&=\mathrm{E}(\mathrm{e}^{\mathrm{i}t'X})=\mathrm{E}(\mathrm{e}^{\mathrm{i}t'(\mu+AU)})\\&=\exp(\mathrm{i}t'\mu)\cdot\mathrm{E}(\mathrm{e}^{\mathrm{i}t'AU})\quad(\text{令 }s'=t'A=(s_1,\cdots,s_q))\\&=\exp(\mathrm{i}t'\mu)\cdot\mathrm{E}(\mathrm{e}^{\mathrm{i}(s_1U_1+\cdots+s_qU_q)})\end{aligned}$$

$$= \exp(\mathrm{i}t'\mu) \cdot \prod_{j=1}^{q} \mathrm{E}(\mathrm{e}^{\mathrm{i}s_jU_j}) \quad (\text{因 } U_1,\cdots,U_q \text{ 独立})$$

$$= \exp(\mathrm{i}t'\mu) \cdot \prod_{j=1}^{q} \exp\left(-\frac{1}{2}s_j^2\right) \quad (\text{因 } U_j \sim N(0,1))$$

$$= \exp\left(\mathrm{i}t'\mu - \frac{1}{2}s's\right) = \exp\left(\mathrm{i}t'\mu - \frac{1}{2}t'AA't\right). \qquad \textbf{(证毕)}$$

定义 2.2.2 若 p 维随机向量 X 的特征函数为

$$\Phi_X(t) = \exp\left[\mathrm{i}t'\mu - \frac{1}{2}t'\Sigma t\right] \quad (\Sigma \geqslant 0),$$

则称 X 服从 p 元正态分布，记为 $X \sim N_p(\mu,\Sigma)$.

性质 2 设 $X \sim N_p(\mu,\Sigma)$，B 为 $s\times p$ 常数矩阵，d 为 s 维常向量，令 $Z=BX+d$，则 $Z \sim N_s(B\mu+d, B\Sigma B')$.

证明 因 $\Sigma \geqslant 0$，Σ 可分解为：$\Sigma = AA'$，则由定义 2.2.1 知

$$X \overset{\mathrm{d}}{=\!=} AU + \mu \quad (A \text{ 为 } p\times q \text{ 实矩阵}),$$

其中 $U=(U_1,\cdots,U_q)'$，且 $U_1,\cdots,U_q$ 相互独立同 $N(0,1)$ 分布. 又

$$Z = BX + d \overset{\mathrm{d}}{=\!=} B(AU+\mu) + d = BAU + (B\mu + d).$$

由定义 2.2.1 可知，$Z \sim N_s(B\mu+d, (BA)(BA)')$，即

$$Z \sim N_s(B\mu + d, B\Sigma B'). \qquad \textbf{(证毕)}$$

性质 2 指出正态随机向量的任意线性组合仍服从正态分布.

推论 设 $X=\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-r}^{r} \sim N_p(\mu,\Sigma)$，将 μ,Σ 剖分为

$$\mu=\begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}_{p-r}^{r}, \quad \Sigma=\begin{bmatrix} \overset{r}{\Sigma_{11}} & \overset{p-r}{\Sigma_{12}} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{p-r}^{r},$$

则 $X^{(1)} \sim N_r(\mu^{(1)},\Sigma_{11})$，$X^{(2)} \sim N_{p-r}(\mu^{(2)},\Sigma_{22})$.

证明 取 $B_1=(I_r \vdots O)$（其中 I_r 为 r 阶单位矩阵，O 为 $r\times(p-r)$ 零矩阵），r 维向量 $d_1=0$，由性质 2 即得

$$X^{(1)} = B_1X + d_1 \sim N_r(\mu^{(1)},\Sigma_{11}).$$

---

① $\overset{\mathrm{d}}{=\!=}$ 表示两边的随机向量服从相同的分布.

类似地，取 $B_2=(O \vdots I_{p-r})$（其中 O 为 $(p-r)\times r$ 零矩阵），$p-r$ 维向量 $d_2=0$，则

$$X^{(2)} = B_2X + d_2 \sim N_{p-r}(\mu^{(2)},\Sigma_{22}). \qquad \text{（证毕）}$$

此推论指出，多元正态分布的边缘分布仍为正态分布. 但反之，若随机向量的任何边缘分布均为正态分布，也不一定能导出该随机向量服从多元正态分布（见例 2.1.1）.

性质 3　若 $X\sim N_p(\mu,\Sigma)$，则 $\mathrm{E}(X)=\mu$，$\mathrm{D}(X)=\Sigma$.

证明　因 $\Sigma\geqslant 0$，Σ 可分解为：$\Sigma=AA'$，则由定义 2.2.1 可知

$$X \overset{\mathrm{d}}{=\!=} AU + \mu \quad (A \text{ 为 } p\times q \text{ 实矩阵}),$$

其中 $U=(U_1,\cdots,U_q)'$，且 $U_1,\cdots,U_q$ 相互独立同 $N(0,1)$分布.

由一元正态分布的知识可知：$\mathrm{E}(U_i)=0$，$\mathrm{Var}(U_i)=1$ $(i=1,\cdots,q)$，$\mathrm{Cov}(U_i,U_j)=0$ $(i\neq j)$，故 $\mathrm{E}(U)=0_q$（0_q 表示 q 维零向量），$\mathrm{D}(U)=I_q$. 利用均值向量和协方差阵的有关性质可得：

$$\mathrm{E}(X) = \mathrm{E}(AU+\mu) = A\mathrm{E}(U) + \mu = \mu,$$

$$\mathrm{D}(X) = \mathrm{D}(AU+\mu) = \mathrm{D}(AU) = AI_qA' = \Sigma. \qquad \text{（证毕）}$$

此性质给出多元正态分布中参数 μ 和 Σ 的明确统计意义.

性质 4　设 $X=(X_1,\cdots,X_p)'$ 为 p 维随机向量，则

X 服从 p 元正态分布

$\Longleftrightarrow$ 对任一 p 维实向量 a，$\xi=a'X$ 是一维正态随机变量.

证明　$\Longrightarrow$（必要性）：若 $X\sim N_p(\mu,\Sigma)$，对任一实向量 $a=(a_1,\cdots,a_p)'$，取 $B=a'$，$d=0$，由性质 2 即得

$$\xi = a'X = \sum_{i=1}^{p} a_iX_i \sim N(a'\mu, a'\Sigma a).$$

$\Longleftarrow$（充分性）：因对任给实向量 $t\in R^p$，$\xi=t'X\sim$一元正态分布，可知 ξ 的各阶矩存在，故 $\mathrm{E}(X_i)$，$\mathrm{Cov}(X_i,X_j)$ $(i,j=1,\cdots,p)$存在. 记 $\mathrm{E}(X)=\mu$，$\mathrm{D}(X)=\Sigma$.

对任意给定的 $t\in R^p$，$\xi=t'X\sim N(t'\mu,t'\Sigma t)$，且 ξ 的特征函数为

$$\Phi_\xi(\theta) = \mathrm{E}(\mathrm{e}^{\mathrm{i}\theta\xi}) = \exp\left[\mathrm{i}\theta(t'\mu) - \frac{1}{2}\theta^2(t'\Sigma t)\right],$$

取 $\theta=1$，

$$\Phi_\xi(1) = E(e^{i\xi}) = E(e^{it'X}) = \Phi_X(t) = \exp\left[it'\mu - \frac{1}{2}t'\Sigma t\right],$$

由定义 2.2.2 可知，$X \sim N_p(\mu,\Sigma)$. （证毕）

定义 2.2.3 若 p 维随机向量 X 的任意线性组合均服从一元正态分布，则称 X 为 p **维正态随机向量**.

在概率论中大家知道，一元正态随机变量的密度函数是

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (\sigma > 0, -\infty < x < \infty).$$

这个式子又可改写为

$$f(x) = \frac{1}{(2\pi)^{1/2}|\sigma^2|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'(\sigma^2)^{-1}(x-\mu)\right].$$

作为一元正态随机变量的推广，以下来导出多维正态随机向量的联合密度函数.

性质 5 设 $X \sim N_p(\mu,\Sigma)$，且 $\Sigma > 0$（正定），则 X 的联合密度函数为

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right].$$

证明 因 $\Sigma > 0$，$\mathrm{rank}(\Sigma) = p$，由线性代数的知识知，存在 p 阶非奇异方阵 A，使得 $\Sigma = AA'$，且

$$X \overset{d}{=\!=} AU + \mu,$$

其中 $U = (U_1,\cdots,U_p)'$，且 $U_1,\cdots,U_p$ 相互独立同 $N(0,1)$ 分布.

U 的联合密度函数为

$$f_U(u) = \frac{1}{(2\pi)^{p/2}}\exp\left[-\frac{1}{2}u'u\right].$$

利用 U 的联合密度函数及随机向量的变换 $X = AU + \mu$ 的密度函数公式：

$$\begin{aligned} f_X(x) &= \frac{1}{(2\pi)^{p/2}}\exp\left[-\frac{1}{2}u'u\right]J(u \to x) \\ &= \frac{1}{(2\pi)^{p/2}}\exp\left\{-\frac{1}{2}[A^{-1}(x-\mu)]'[A^{-1}(x-\mu)]\right\}|\Sigma|^{-1/2} \\ &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]. \end{aligned}$$

这里积分变换的雅可比（Jacobian）行列式 $J(u \to x)$ 可利用线性变换

$x=Au+\mu$ 及 $J(x\to u)$来计算：因

$$J(x\to u)=\left|\frac{\partial x'}{\partial u}\right|_{+}^{①}=\begin{vmatrix}\frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_p}{\partial u_1}\\ \vdots & & \vdots\\ \frac{\partial x_1}{\partial u_p} & \cdots & \frac{\partial x_p}{\partial u_p}\end{vmatrix}_{+}$$

$$=|A'|_{+}=|AA'|^{1/2}=|\Sigma|^{1/2},$$

故　　$$J(u\to x)=\frac{1}{J(x\to u)}=|\Sigma|^{-1/2}.$$　　**(证毕)**

定义 2.2.4　若 p 维随机向量 $X=(X_1,X_2,\cdots,X_p)'$的联合密度函数为

$$f(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right],$$

其中 μ 是 p 维实向量，Σ 是 p 阶正定矩阵，则称 $X=(X_1,X_2,\cdots,X_p)'$服从(非退化的)p 元正态分布；也称 X 为 p **维正态随机向量**，简记 $X\sim N_p(\mu,\Sigma)$.

以上给出了多元正态分布的 4 种定义. 定义 2.2.4 是用密度函数给出的，它可看成为一元正态密度的直接推广；但在这个定义里要求 Σ 是正定矩阵，因而它给出的是非退化的正态分布的定义. 另三种定义中把 Σ 矩阵推广到非负定的情形，这三种定义是等价的.

例 2.2.1 (二元正态分布)　设 $X=\begin{bmatrix}X_1\\X_2\end{bmatrix}\sim N_2(\mu,\Sigma)$，记

$$\mu=\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\quad \Sigma=\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}>0$$

(即 $\sigma_1>0,\sigma_2>0,|\rho|<1$).

(1) 试写出 X 的联合密度函数和边缘密度函数；

(2) 试说明 ρ 的统计意义.

解　(1) 因 $|\Sigma|=\sigma_1^2\sigma_1^2(1-\rho^2)$，以及

$$\Sigma^{-1}=\frac{1}{\sigma_1^2\sigma_1^2(1-\rho^2)}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix},$$

---

① $|A|_{+}$表示矩阵 A 的行列式的绝对值.

因此二维正态随机向量 X 的联合密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$$

$$=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right.$$

$$\left.\left.-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

由性质 2 即得 $X_1\sim N(\mu_1,\sigma_1^2)$, $X_2\sim N(\mu_2,\sigma_2^2)$.

(2) 因 $\mathrm{Cov}(X_1,X_2)=\sigma_{12}=\rho\sigma_1\sigma_2$, 而 X_1 与 X_2 的相关系数

$$\rho(X_1,X_2)=\frac{\mathrm{Cov}(X_1,X_2)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(X_2)}}=\frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2}=\rho.$$

故二元正态分布的参数 ρ 就是两个分量的相关系数. 显然:

当 $\rho=0$ 时, $f(x_1,x_2)=f_1(x_1)\cdot f_2(x_2)$, 即 X_1 和 X_2 相互独立.

当 $|\rho|=1$ 时, $|\Sigma|=0$ (Σ 退化), 则存在非零向量 $t=(t_1,t_2)'$, 使得 $\Sigma t=0$, 从而 $t'\Sigma t=0$, 故而有

$$\mathrm{Var}[t'(X-\mu)]=t'\Sigma t=0.$$

这表示 $P\{t'(X-\mu)=0\}=1$, 即 $t_1(X_1-\mu_1)+t_2(X_2-\mu_2)=0$ 以概率 1 成立; 反之, 若 X_1 和 X_2 以概率 1 存在线性相关关系, 则 $|\rho|=1$.

当 $\rho>0$ 时我们称 X_1 和 X_2 存在正相关; 当 $\rho<0$ 时我们称 X_1 和 X_2 存在负相关.

为了对多元正态密度函数有更直观地了解, 下面的例子给出几组参数下二元正态密度函数的几何图形. 我们把具有等密度的点的轨迹称为等高线(面). 显然当 $p=2$ 时

$$f(x_1,x_2)=C$$

$$\Leftrightarrow\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\cdot\sigma_2}$$

$$+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2=a^2\quad(a\geqslant 0),$$

它是一族中心在 $(\mu_1,\mu_2)'$ 的椭圆. 一般的 p 元正态密度函数的等高面为

$$(x-\mu)'\Sigma^{-1}(x-\mu)=a^2 \quad (a\geqslant 0).$$

例 2.2.2　绘制二元正态密度函数的图形及其相应的等高线图形.

作图　我们采用 SAS 系统分别绘制 3 组不同参数时的二元正态密度函数及其相应的等高线图(取 $\mu_1=\mu_2=0$),如图 2.1 至图 2.3 所示.

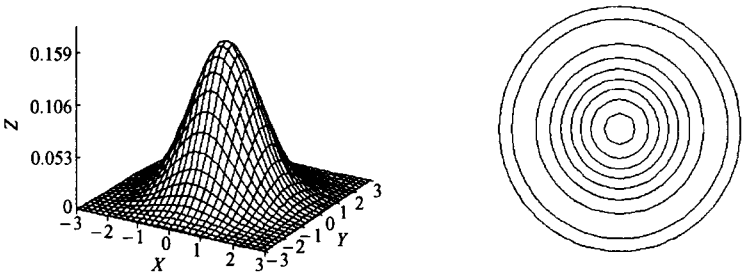

图 2.1　$\sigma_1^2=1,\sigma_2^2=1,\rho=0$ 时的二元正态密度函数及其等高线图

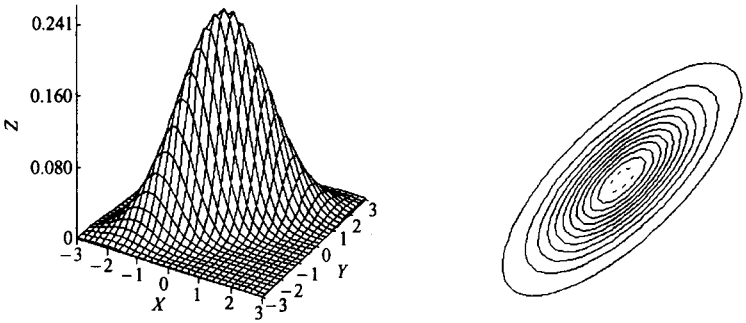

图 2.2　$\sigma_1^2=1,\sigma_2^2=1,\rho=0.75$ 时二元正态密度函数及其等高线图

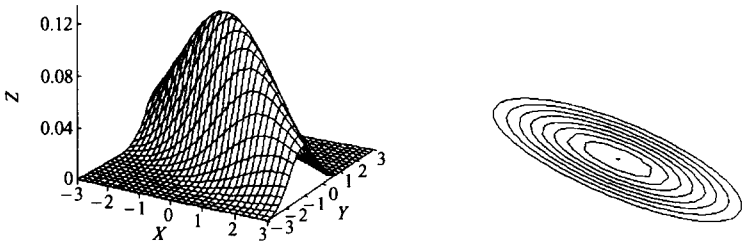

图 2.3　$\sigma_1^2=4,\sigma_2^2=1,\rho=-0.75$ 时二元正态密度函数及其等高线图

§2.3 条件分布和独立性

设 $X\sim N_p(\mu,\Sigma)$ $(p\geqslant 2)$，将 X,μ,Σ 剖分为

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\begin{matrix}r\\p-r\end{matrix},\quad \mu=\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix},$$

$$\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\begin{matrix}r\\p-r\end{matrix}>0,$$

一、独立性

定理 2.3.1 设 p 维随机向量 $X\sim N_p(\mu,\Sigma)$，

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\sim N_p\left(\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),$$

则

$$X^{(1)}\text{ 与 }X^{(2)}\text{ 相互独立}\iff \Sigma_{12}=O$$

（即 $X^{(1)}$ 与 $X^{(2)}$ 互不相关）.

证明 $\Longrightarrow$：已知 $X^{(1)}$ 与 $X^{(2)}$ 相互独立，则

$$\mathrm{COV}(X^{(1)},X^{(2)})=\Sigma_{12}=O.$$

$\Longleftarrow$：设 $\Sigma_{12}=O$，则 X 的联合密度函数为

$$\begin{aligned}f(x^{(1)},x^{(2)})&=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)'\begin{bmatrix}\Sigma_{11} & O\\O & \Sigma_{22}\end{bmatrix}^{-1}(x-\mu)\right\}\\&=\frac{1}{(2\pi)^{r/2}|\Sigma_{11}|^{1/2}}\exp\left\{-\frac{1}{2}(x^{(1)}-\mu^{(1)})'\Sigma_{11}^{-1}(x^{(1)}-\mu^{(1)})\right\}\\&\quad\cdot\frac{1}{(2\pi)^{(p-r)/2}|\Sigma_{22}|^{1/2}}\exp\left\{-\frac{1}{2}(x^{(2)}-\mu^{(2)})'\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})\right\}\\&=f_1(x^{(1)})\cdot f_2(x^{(2)}),\end{aligned}$$

所以 $X^{(1)}$ 与 $X^{(2)}$ 相互独立. **（证毕）**

推论 1 设 $r_i\geqslant 1$ $(i=1,\cdots,k)$，且 $r_1+r_2+\cdots+r_k=p$，有

$$X=\begin{bmatrix}X^{(1)}\\ \vdots\\ X^{(k)}\end{bmatrix}\begin{matrix}r_1\\ \vdots\\ r_k\end{matrix}\sim N_p\left(\begin{bmatrix}\mu^{(1)}\\ \vdots\\ \mu^{(k)}\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \cdots & \Sigma_{1k}\\ \vdots & & \vdots\\ \Sigma_{k1} & \cdots & \Sigma_{kk}\end{bmatrix}_{p\times p}\right)$$

则

$$X^{(1)},\cdots,X^{(k)} \text{ 相互独立} \Longleftrightarrow \Sigma_{ij}=O \quad (\text{一切 } i\neq j).$$

推论 2　设 $X=(X_1,\cdots,X_p)'\sim N_p(\mu,\Sigma)$，若 Σ 为对角矩阵，则 $X_1,\cdots,X_p$ 相互独立.

二、条件分布

首先来考虑二元正态的条件分布，即当 $p=2,r=1$ 时，由条件密度的定义知，当 X_2 给定时，X_1 的条件密度为

$$f_1(x_1|x_2)=\frac{f(x_1,x_2)}{f_2(x_2)},$$

而

$$\begin{aligned}f(x_1,x_2)&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right.\\&\quad\left.\left.-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}\\&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2\right.\right.\\&\quad+\rho^2\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\\&\quad\left.\left.+(1-\rho^2)\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}\\&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}\\&\quad\cdot\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)-\rho\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]^2\right\}\\&=\frac{1}{\sqrt{2\pi}\sigma_2}\exp\left\{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}\cdot\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\end{aligned}$$

$$\cdot \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\left[x_1-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right]^2\right\}$$

$$= f_2(x_2)\cdot f(x_1|x_2),$$

其中

$$f(x_1|x_2)=\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)\sigma_1^2}\cdot\left[x_1-\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)\right]^2\right\},$$

所以

$$(X_1|X_2)\sim N_1\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),\sigma_1^2(1-\rho^2)\right).$$

将其推广到 p 元情况，利用 Σ^{-1}的分块求逆公式(参见附录§4)：

$$\Sigma^{-1}=\left[\begin{array}{c:c}\Sigma_{11\cdot 2}^{-1} & -\Sigma_{11\cdot 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ \hdashline -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1} & \Sigma_{22}^{-1}+\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11\cdot 2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\end{array}\right],$$

其中 $\Sigma_{11\cdot 2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. 类似 $p=2$ 的方法，可证明

$$f(x^{(1)},x^{(2)})=f_2(x^{(2)})\cdot f_1(x^{(1)}|x^{(2)}),$$

且 $f_1(x^{(1)}|x^{(2)})$为 r 元正态密度函数.

定理 2.3.2 设 $X=\begin{bmatrix}X^{(1)}\\ X^{(2)}\end{bmatrix}\begin{matrix}r\\ p-r\end{matrix}\sim N_p(\mu,\Sigma)$ $(\Sigma>0)$，则当 $X^{(2)}$给定时，$X^{(1)}$的条件分布为

$$(X^{(1)}|X^{(2)})\sim N_r(\mu_{1\cdot 2},\Sigma_{11\cdot 2}),$$

其中

$$\mu_{1\cdot 2}=\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)}),$$

$$\Sigma_{11\cdot 2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

证明 作非奇异线性变换，令

$$Z=\begin{bmatrix}Z^{(1)}\\ Z^{(2)}\end{bmatrix}=\begin{bmatrix}X^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}X^{(2)}\\ X^{(2)}\end{bmatrix}=\left[\begin{array}{c:c}I_r & -\Sigma_{12}\Sigma_{22}^{-1}\\ \hdashline O & I_{p-r}\end{array}\right]\begin{bmatrix}X^{(1)}\\ X^{(2)}\end{bmatrix}$$

$$=BX.$$

由§2.2的性质2显然有

$$Z\sim N_p\left(\begin{bmatrix}\mu^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}\\ \mu^{(2)}\end{bmatrix},\begin{bmatrix}\Sigma_{11\cdot 2} & O\\ O & \Sigma_{22}\end{bmatrix}\right),$$

且因 $D(Z)=\begin{bmatrix}\Sigma_{11\cdot 2} & O\\ O & \Sigma_{22}\end{bmatrix}$，故 $Z^{(1)}$ 与 $Z^{(2)}$ 相互独立.

Z 的联合密度为

$$g(z^{(1)},z^{(2)})=g_1(z^{(1)})\cdot g_2(z^{(2)})=g_1(z^{(1)})\cdot f_2(z^{(2)}).$$

这里因 $Z^{(2)}=X^{(2)}$，故有 $g_2(z^{(2)})=f_2(z^{(2)})$（式内 $f_2(\cdot)$ 为 $X^{(2)}$ 的密度函数）.

因为 $Z=BX$，利用积分变换公式，可以用 $g(z)$ 来表示 X 的密度函数 $f(x)$，即

$$\begin{aligned}f(x^{(1)},x^{(2)})&=g(Bx)\cdot J(z\to x)\\&=g_1(x^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}x^{(2)})\cdot g_2(x^{(2)})\cdot\left|\frac{\partial z'}{\partial x}\right|_+\\&=g_1(x^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}x^{(2)})\cdot f_2(x^{(2)}),\end{aligned}$$

式中 $\left|\dfrac{\partial z'}{\partial x}\right|_+=|B'|=1$. 并注意到

$$Z^{(1)}\sim N_r(\mu^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)},\Sigma_{11\cdot 2})),$$

所以

$$\begin{aligned}f_1(x^{(1)}|x^{(2)})&=\frac{f(x^{(1)},x^{(2)})}{f_2(x^{(2)})}=g_1(x^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}x^{(2)})\\&=\frac{1}{(2\pi)^{r/2}|\Sigma_{11\cdot 2}|^{1/2}}\exp\Big[-\frac{1}{2}(x^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}x^{(2)}\\&\quad-(\mu^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}))'\Sigma_{11\cdot 2}^{-1}(x^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}x^{(2)}\\&\quad-(\mu^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}))\Big]\\&=\frac{1}{(2\pi)^{r/2}|\Sigma_{11\cdot 2}|^{1/2}}\exp\Big[-\frac{1}{2}(x^{(1)}-\mu_{1\cdot 2})'\Sigma_{11\cdot 2}^{-1}(x^{(1)}-\mu_{1\cdot 2})\Big],\end{aligned}$$

其中

$$\mu_{1\cdot 2}=\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)}),$$

$$\Sigma_{11\cdot 2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

（证毕）

推论　在定理 2.3.2 条件下可得：

(1) $X^{(2)}$ 与 $X^{(1)}-\Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ 相互独立；

(2) $X^{(1)}$ 与 $X^{(2)}-\Sigma_{21}\Sigma_{11}^{-1}X^{(1)}$ 相互独立；

(3) $(X^{(2)}|X^{(1)})\sim N_{p-r}(\mu_{2\cdot1},\Sigma_{22\cdot1})$，其中

$$\mu_{2\cdot1}=\mu^{(2)}+\Sigma_{21}\Sigma_{11}^{-1}(x^{(1)}-\mu^{(1)}),$$

$$\Sigma_{22\cdot1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

三、几个概念

1. 条件期望，回归系数，偏相关系数

设

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\begin{matrix}r\\p-r\end{matrix}\sim N_p\left(\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}\right);$$

又已知 $X^{(2)}$ 给定时 $X^{(1)}$ 的条件分布为

$$(X^{(1)}|X^{(2)})\sim N(\mu_{1\cdot2},\Sigma_{11\cdot2}).$$

则称

$$\mu_{1\cdot2}=\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})$$

为**条件期望**，记为 $\mathrm{E}(X^{(1)}|X^{(2)})$；并称 $\mu_{1\cdot2}$ 为 $X^{(1)}$ 对 $X^{(2)}$ 的**回归**，称

$$\Sigma_{12}\Sigma_{22}^{-1}\xlongequal{\text{def}}B$$

为**回归系数**. 记

$$\Sigma_{11\cdot2}=(\sigma_{ij\cdot r+1,\cdots,p})_{r\times r}\quad(i,j=1,\cdots,r).$$

称

$$r_{ij\cdot r+1,\cdots,p}=\frac{\sigma_{ij\cdot r+1,\cdots,p}}{\sqrt{\sigma_{ii\cdot r+1,\cdots,p}}\cdot\sqrt{\sigma_{jj\cdot r+1,\cdots,p}}}$$

为当 $X^{(2)}=(X_{r+1},\cdots,X_p)'$ 给定时，X_i 与 $X_j(i,j=1,2,\cdots,r)$ 的**偏相关系数**.

2. 全相关系数

设 $Z=\begin{bmatrix}X\\Y\end{bmatrix}\begin{matrix}p\\1\end{matrix}\sim N_{p+1}\left(\begin{bmatrix}\mu_X\\\mu_y\end{bmatrix},\begin{bmatrix}\Sigma_{XX}&\Sigma_{Xy}\\\Sigma_{yX}&\sigma_{yy}\end{bmatrix}\right)$，则称

$$R=\left(\frac{\Sigma_{yX}\Sigma_{XX}^{-1}\Sigma_{Xy}}{\sigma_{yy}}\right)^{1/2}$$

为 Y 与 $X=(X_1,X_2,\cdots,X_p)'$ 的**全相关系数**.

3. 最佳预测

在定理 2.3.2 条件下，我们考虑 $r=1$，记 $X^{(1)}=Y$，$g(x^{(2)})=\mathrm{E}(Y|X^{(2)})$，则对任意函数 $\varphi(\cdot)$，可以证明（见习题二的第 2-16 题）：

$$\mathrm{E}[(Y-g(x^{(2)}))^2]\leqslant \mathrm{E}[(Y-\varphi(x^{(2)}))^2].$$

即在均方差最小的准则下，条件期望 $g(x^{(2)})$ 是对 Y 的最佳预测函数.

§2.4　随机阵的正态分布

把来自 p 元总体的容量为 n 的随机样本排成一矩阵 X：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \overset{\text{def}}{=\!=} \begin{bmatrix} X'_{(1)} \\ X'_{(2)} \\ \vdots \\ X'_{(n)} \end{bmatrix}$$

$$\overset{\text{或 def}}{=\!=} (X_1, X_2, \cdots, X_p),$$

其中 $X_{(i)}(i=1,\cdots,n)$ 是来自 p 元总体的一个样品，则样本数据阵 X 就是一个随机阵. 讨论随机阵 X 的分布时，可考虑把 X 的行向量（即样品）一个接一个连接起来构成一个 np 维长向量，然后讨论这个长向量的分布.

一、拉直运算和克罗内克(Kronecker)积

1. 拉直运算

所谓拉直运算，就是将矩阵拉成一个长向量，通过它来建立矩阵和向量之间的联系. 设随机矩阵 X 是一个 $n\times p$ 矩阵，用 X 的列向量 $X_1, X_2, \cdots, X_p$ 组成一个 np 维向量，记为

$$\mathrm{Vec}(X)=\begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}=(x_{11}, x_{21}, \cdots, x_{n1}, \cdots, x_{1p}, x_{2p}, \cdots, x_{np})',$$

符号“Vec”称为**拉直运算**. 如果将矩阵 X 的行向量(样品)拉直为一个 np 维向量,用拉直运算的符号可记为

$$\mathrm{Vec}(X') = \begin{bmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{bmatrix} = (x_{11}, x_{12}, \cdots, x_{1p}, \cdots, x_{n1}, x_{n2}, \cdots, x_{np})'.$$

在多元统计分析中,经常需要考虑对称矩阵的拉直运算. 设 S 是 p 阶对称随机阵,在 S 矩阵中只包含 $p(p+1)/2$ 个不同的随机变量,故将其拉直为 p^2 维向量是不合适的,应拉成 $p(p+1)/2$ 维向量. 设 $S=(S_{ij})_{p\times p}$ 为 p 阶对称矩阵,令

$$\mathrm{Svec}(S) = (S_{11}, \cdots, S_{p1}, S_{22}, \cdots, S_{p2}, \cdots, S_{pp})'$$

为 $p(p+1)/2$ 维向量. 符号“Svec”称为**对称矩阵的拉直运算**.

2. 克罗内克积

设 $A=(a_{ij})$ 和 B 分别为 $n\times p$ 和 $m\times q$ 的矩阵,A 和 B 的克罗内克积 $A\otimes B$ 定义为

$$A \otimes B = (a_{ij}B) = \begin{bmatrix} a_{11}B & \cdots & a_{1p}B \\ \vdots & & \vdots \\ a_{n1}B & \cdots & a_{np}B \end{bmatrix},$$

它是 $mn\times pq$ 矩阵. 在多元统计分析中克罗内克积又称**矩阵的直积**,是一个有用的工具. 在下面的讨论中将用到矩阵的直积的一些性质(见参考文献[1]).

二、随机阵的正态分布

设 $X_{(i)}=(x_{i1}, \cdots, x_{ip})'$ $(i=1, \cdots, n)$ 为来自 p 元正态总体 $N_p(\mu, \Sigma)$ 的随机样本(独立同分布),记随机阵 $X=(x_{ij})_{n\times p}$,利用拉直运算及矩阵的直积的定义和性质,可知

$$\mathrm{Vec}(X') \sim N_{np}(\mathbf{1}_n \otimes \mu, I_n \otimes \Sigma),$$

事实上,np 维长向量 $\mathrm{Vec}(X')$ 的联合密度函数为

$$f(x_{(1)}, \cdots, x_{(n)})$$

① 本书中 $\mathbf{1}_p$ 表示向量元素均为 1 的 p 维常向量.

$$= \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x_{(i)}-\mu)'\Sigma^{-1}(x_{(i)}-\mu)\right]$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_{(i)}-\mu)'\Sigma^{-1}(x_{(i)}-\mu)\right]$$

$$= \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \cdot \exp\left[-\frac{1}{2}\begin{bmatrix} x_{(1)}-\mu \\ \vdots \\ x_{(n)}-\mu \end{bmatrix}' \begin{bmatrix} \Sigma & \cdots & O \\ \vdots & & \vdots \\ O & \cdots & \Sigma \end{bmatrix}^{-1} \begin{bmatrix} x_{(1)}-\mu \\ \vdots \\ x_{(n)}-\mu \end{bmatrix}\right].$$

由矩阵的直积的定义，np 维随机向量 $\mathrm{Vec}(X')$ 的均值向量和协方差阵分别为

$$\begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \mathbf{1}_n \otimes \mu, \quad \begin{bmatrix} \Sigma & \cdots & O \\ \vdots & & \vdots \\ O & \cdots & \Sigma \end{bmatrix} = I_n \otimes \Sigma.$$

当随机阵 X 按行拉直后，如果有

$$\mathrm{Vec}(X') \sim N_{np}(\mathbf{1}_n \otimes \mu, I_n \otimes \Sigma),$$

则称 X 服从**矩阵正态分布**，记作

$$X \sim N_{n\times p}(M, I_n \otimes \Sigma),$$

其中

$$\mathrm{Vec}(M') = \mathbf{1}_n \otimes \mu = (\mu_1, \cdots, \mu_p, \cdots, \mu_1, \cdots, \mu_p)'.$$

即

$$X \sim N_{n\times p}(M, I_n \otimes \Sigma) \Longleftrightarrow \mathrm{Vec}(X') \sim N_{np}(\mathrm{Vec}(M'), I_n \otimes \Sigma),$$

其中

$$M = \begin{bmatrix} \mu_1 & \cdots & \mu_p \\ \vdots & & \vdots \\ \mu_1 & \cdots & \mu_p \end{bmatrix} = \mathbf{1}_n \mu' = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} (\mu_1, \cdots, \mu_p).$$

随机阵正态分布有如下有用的性质：

设 $X \sim N_{n\times p}(M, I_n \otimes \Sigma)$，$A$ 为 $k\times n$ 常数矩阵，B 为 $q\times p$ 常数矩阵，D 为 $k\times q$ 常数矩阵，令 $Z = AXB' + D$，则

$$Z \sim N_{k\times q}(AMB' + D, (AA') \otimes (B\Sigma B')).$$

§2.5 多元正态分布的参数估计

考虑 p 元正态总体 $X\sim N_p(\mu,\Sigma)$，设 $X_{(i)}=(x_{i1},\cdots,x_{ip})'$ $(i=1,\cdots,n)$为 p 元正态总体 X 的简单随机样本，此时观测数据阵

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

是一个随机阵.

本节讨论参数 μ 和 Σ 的最大似然估计及其性质.

一、多元正态总体样本的数字特征

对于多元统计分析，我们引入以下多元正态总体样本的相关量.

(1) 样本均值向量 $\overline{X}$：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_{(i)}=(\overline{x}_1,\cdots,\overline{x}_p)'=\frac{1}{n}X'\mathbf{1}_n,$$

其中

$$\overline{x}_i=\frac{1}{n}\sum_{\alpha=1}^{n}x_{\alpha i}\quad (i=1,2,\cdots,p).$$

(2) 样本离差阵(又称交叉乘积阵)A：

$$A=\sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})'=X'X-n\overline{X}\,\overline{X}'$$

$$=X'\left[I_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right]X\xlongequal{\text{def}}(a_{ij})_{p\times p},$$

其中

$$a_{ij}=\sum_{\alpha=1}^{n}(x_{\alpha i}-\overline{x}_i)(x_{\alpha j}-\overline{x}_j)\quad (i,j=1,2,\cdots,p).$$

(3) 样本协方差阵 S：

$$S=\frac{1}{n-1}A=(s_{ij})_{p\times p}\quad \left(\text{或 } S^*=\frac{1}{n}A\right),$$

其中

$$s_{ii}=\sum_{a=1}^{n}(x_{ai}-\bar{x}_i)^2 \quad (i=1,2,\cdots,p)$$

称为变量 X_i 的**样本方差**；样本方差的平方根$\sqrt{s_{ii}}$称为变量 X_i 的**样本标准差**.

(4) 样本相关阵 R：

$$R=(r_{ij})_{p\times p},$$

其中

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}\overset{或}{=\!=}\frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}} \quad (i,j=1,2,\cdots,p).$$

二、μ,Σ 的最大似然估计

设 $X_{(i)}(i=1,\cdots,n)$为 p 元正态总体 $N(\mu,\Sigma)$的随机样本，以下用最大似然法来求参数 μ,Σ 的最大似然估计.

1. 似然函数 $L(\mu,\Sigma)$

把随机数据阵 X 按行拉直后形成的 np 维长向量 $\mathrm{Vec}(X')$的联合密度函数看成未知参数 μ,Σ 的函数，并称为**样本 $X_{(i)}(i=1,\cdots,n)$ 的似然函数**，记为 $L(\mu,\Sigma)$：

$$\begin{aligned}
L(\mu,\Sigma)&=\prod_{i=1}^{n}\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\Big[-\frac{1}{2}(x_{(i)}-\mu)'\Sigma^{-1}(x_{(i)}-\mu)\Big]\\
&=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\Big[-\frac{1}{2}\sum_{i=1}^{n}(x_{(i)}-\mu)'\Sigma^{-1}(x_{(i)}-\mu)\Big]\\
&=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\Big[-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}((x_{(i)}-\mu)'\Sigma^{-1}(x_{(i)}-\mu))\Big]\\
&=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\Big[-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}(\Sigma^{-1}(x_{(i)}-\mu)(x_{(i)}-\mu)')\Big]\\
&=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\exp\Big[\mathrm{tr}\Big(-\frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)'\Big)\Big]\\
&\overset{\mathrm{def}}{=\!=}\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\mathrm{etr}\Big(-\frac{1}{2}\Sigma^{-1}\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)'\Big),
\end{aligned}$$

其中

$$\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)'$$

$$=\sum_{i=1}^{n}(x_{(i)}-\overline{X}+\overline{X}-\mu)(x_{(i)}-\overline{X}+\overline{X}-\mu)'$$

$$=\sum_{i=1}^{n}(x_{(i)}-\overline{X})(x_{(i)}-\overline{X})'+n(\overline{X}-\mu)(\overline{X}-\mu)'$$

$$=A+n(\overline{X}-\mu)(\overline{X}-\mu)'.$$

由于 $\ln x$ 是 x 的单调函数，$L(\mu,\Sigma)$与 $\ln L(\mu,\Sigma)$有相同的最大值点. 以下只须讨论 $\ln L(\mu,\Sigma)$的最大值问题.

2. 迹的有关性质

在附录中介绍了迹的一些性质，下面的一条引理给出与迹有关的进一步的性质.

引理 2.5.1 设 B 为 p 阶正定矩阵，则

$$\mathrm{tr}B-\ln|B|\geqslant p,$$

且等号成立的充分必要条件是 $B=I_p$.

证明 因为 $B>O$，所以 B 的全部特征值 $\lambda_1,\cdots,\lambda_p>0$，且 $|B|=\lambda_1\cdots\lambda_p$. 利用不等式 $\ln(1+x)\leqslant x$(当 $x+1>0$)，可得

$$\ln|B|=\sum_{i=1}^{p}\ln\lambda_i=\sum_{i=1}^{p}\ln(1+\lambda_i-1)$$

$$\leqslant\sum_{i=1}^{p}(\lambda_i-1)=\mathrm{tr}(B)-p.$$

所以

$$\mathrm{tr}B-\ln|B|\geqslant p.$$

因不等式 $\ln(1+x)\leqslant x$ 中的等号仅当 $x=0$ 时成立，故引理给出的不等式仅当 $\lambda_i-1=0$ $(i=1,\cdots,p)$时成立，即 $B=I_p$.

反之，当 $B=I_p$ 时，$\ln|I_p|=0$，$\mathrm{tr}B=p$，故引理给出的不等式中的等号成立. **(证毕)**

3. 讨论 $\ln L(\mu,\Sigma)$的最大值点

首先利用迹的有关性质来讨论当给定 $\Sigma>0$ 时，$\ln L(\mu,\Sigma)$的最大值点. 经直接运算，有

$$\begin{aligned}\ln L(\mu,\Sigma) &= -\frac{np}{2}\ln 2\pi - \frac{n}{2}\ln|\Sigma| \\ &\quad - \frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1}\sum_{i=1}^{n}(x_{(i)}-\mu)(x_{(i)}-\mu)'\Big] \\ &= C - \frac{1}{2}\mathrm{tr}[\Sigma^{-1}A + n\Sigma^{-1}(\overline{X}-\mu)(\overline{X}-\mu)'] \\ &= C - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A) - \frac{n}{2}[(\overline{X}-\mu)'\Sigma^{-1}(\overline{X}-\mu)] \\ &\leqslant C - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A).\end{aligned}$$

以上不等式仅当 $\mu=\overline{X}$ 时等号成立，即对于固定的 $\Sigma>0$，有

$$\ln L(\overline{X},\Sigma) = \max_{\mu}\ln L(\mu,\Sigma).$$

进一步地可利用迹的有关性质及引理 2.5.1 来证明，当取 $\hat{\Sigma}=\frac{1}{n}A$ 时 $\ln L(\overline{X},\hat{\Sigma})=\max\limits_{\overline{X},\Sigma>0}\ln L(\overline{X},\Sigma)$：

$$\begin{aligned}\ln L(\overline{X},\Sigma) &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{tr}(\Sigma^{-1}A) \\ &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\Big[\ln|\Sigma| + \mathrm{tr}\Big(\Sigma^{-1}\frac{A}{n}\Big)\Big] \\ &= C_1 - \frac{n}{2}\Big[\mathrm{tr}\Big(\Sigma^{-1}\frac{A}{n}\Big) - \ln\Big|\Sigma^{-1}\frac{A}{n}\Big| + \ln\Big|\frac{A}{n}\Big|\Big] \\ &= C_1 - \frac{n}{2}\Big[\mathrm{tr}\Big(\Sigma^{-1/2}\frac{A}{n}\Sigma^{-1/2}\Big) - \ln\Big|\Sigma^{-1/2}\frac{A}{n}\Sigma^{-1/2}\Big| + \ln\Big|\frac{A}{n}\Big|\Big] \\ &\leqslant C_1 - \frac{np}{2} - \frac{n}{2}\ln\Big|\frac{A}{n}\Big|.\end{aligned}$$

若取 $B=\Sigma^{-1/2}\frac{A}{n}\Sigma^{-1/2}$ 是正定矩阵，由引理 2.5.1，以上不等式的等号仅当 $B=\Sigma^{-1/2}\frac{A}{n}\Sigma^{-1/2}=I_p$，即 $\Sigma=\frac{A}{n}$ 时成立. 所以

$$\begin{aligned}\ln L\Big(\overline{X},\frac{1}{n}A\Big) &= \max_{\overline{X},\Sigma>0}\ln L(\overline{X},\Sigma) \\ &= -\frac{np}{2}(1+\ln(2\pi)) - \frac{n}{2}\ln\Big|\frac{A}{n}\Big|,\end{aligned}$$

因而似然函数的最大值为

$$L\left(\overline{X},\frac{1}{n}A\right)=\left(\frac{n}{2\pi e}\right)^{np/2}|A|^{-n/2}.$$

定理 2.5.1 设 $X_{(i)}(i=1,\cdots,n)$是多元正态总体 $N_p(\mu,\Sigma)$的随机样本，$n>p$，则 μ,Σ 的最大似然估计为 $\hat{\mu}=\overline{X}$，$\hat{\Sigma}=\frac{1}{n}A$.

如果$|\hat{\Sigma}|=\left|\frac{1}{n}A\right|=0$ 怎么办？此时可以证明 $\mathrm{Sup}L(\hat{\mu},\hat{\Sigma})=\infty$，最大似然估计不存在. 但$|A|=0$ 的情况几乎不存在，因为可以证明，当 $n>p$ 时 $P\{A>0\}=1$(见定理 2.5.2).

三、最大似估计量的性质

前面已给出了参数 μ,Σ 的最大似然估计$\hat{\mu}=\overline{X},\hat{\Sigma}=\frac{1}{n}A$. 参数的最大似然估计有很多优良性标准，如无偏性、有效性、相合性等. μ 和 Σ 的最大似然估计是否具有这些好的性质呢？这是我们现在要讨论的问题.

设 $X_{(t)}=(x_{t1},\cdots,x_{tp})'$ $(t=1,\cdots,n)$独立同 $N_p(\mu,\Sigma)$分布，且 $\Sigma>0$，记

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{n}X_{(t)},\quad A=\sum_{t=1}^{n}(X_{(t)}-\overline{X})(X_{(t)}-\overline{X})'.$$

定理 2.5.2 设 $\overline{X}$ 和 A 分别为 p 元正态总体 $N_p(\mu,\Sigma)$的样本均值向量和样本离差阵，则

(1) $\overline{X}\sim N_p\left(\mu,\frac{1}{n}\Sigma\right)$；

(2) $A\overset{d}{=\!=}\sum_{t=1}^{n-1}Z_tZ_t'$，其中 $Z_1,\cdots,Z_{n-1}$独立同 $N_p(0,\Sigma)$分布；

(3) $\overline{X}$ 和 A 相互独立；

(4) $P\{A>0\}=1\iff n>p$.

证明 设 Γ 是 n 阶正交矩阵，具有以下形式

$$\Gamma=\begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & & \vdots \\ r_{(n-1)1} & \cdots & r_{(n-1)n} \\ 1/\sqrt{n} & \cdots & 1/\sqrt{n} \end{bmatrix}=(r_{ij})_{n\times n}.$$

令

$$Z=\begin{bmatrix}Z_1'\\ \vdots\\ Z_n'\end{bmatrix}=\Gamma\begin{bmatrix}X_{(1)}'\\ \vdots\\ X_{(n)}'\end{bmatrix}=\Gamma X,$$

即

$$Z_t=(X_{(1)},\cdots,X_{(n)})\begin{bmatrix}r_{t1}\\ \vdots\\ r_{tn}\end{bmatrix}\quad(t=1,\cdots,n)$$

为 p 维随机向量. 因 Z_t 是 p 维正态随机向量 $X_{(1)},\cdots,X_{(n)}$ 的线性组合，故 Z_t 也是 p 维正态随机向量，且

$$\mathrm{E}(Z_t)=\sum_{i=1}^{n}r_{ti}\mathrm{E}(X_{(i)})=\begin{cases}0, & \text{当 } t\neq n \text{ 时},\\ \sqrt{n}\,\mu, & \text{当 } t=n \text{ 时};\end{cases}$$

$$\begin{aligned}\mathrm{Cov}(Z_\alpha,Z_\beta)&=\mathrm{E}[(Z_\alpha-\mathrm{E}(Z_\alpha))(Z_\beta-\mathrm{E}(Z_\beta))']\\&=\sum_{i=1}^{n}r_{\alpha i}r_{\beta i}\Sigma=\begin{cases}O, & \text{当 } \alpha\neq\beta \text{ 时},\\ \Sigma, & \text{当 } \alpha=\beta \text{ 时}.\end{cases}\end{aligned}$$

(1) 因为 $Z_n=\dfrac{1}{\sqrt{n}}\sum_{\alpha=1}^{n}X_{(\alpha)}=\sqrt{n}\,\overline{X}\sim N_p(\sqrt{n}\,\mu,\Sigma)$，故有

$$\overline{X}=\frac{1}{\sqrt{n}}Z_n\sim N_p\left(\mu,\frac{1}{n}\Sigma\right).$$

(2) 因为

$$\sum_{\alpha=1}^{n}Z_\alpha Z_\alpha'=(Z_1,\cdots,Z_n)\begin{bmatrix}Z_1'\\ \vdots\\ Z_n'\end{bmatrix}=Z'Z$$

$$=X'\Gamma'\cdot\Gamma X=X'X=\sum_{\alpha=1}^{n}X_{(\alpha)}X_{(\alpha)}',$$

且

$$\begin{aligned}\sum_{\alpha=1}^{n-1}Z_\alpha Z_\alpha'&=\sum_{\alpha=1}^{n}X_{(\alpha)}X_{(\alpha)}'-Z_nZ_n'=\sum_{\alpha=1}^{n}X_{(\alpha)}X_{(\alpha)}'-n\overline{X}\,\overline{X}'\\&=\sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})'=A.\end{aligned}$$

(3) 因 $A=\sum_{\alpha=1}^{n-1}Z_\alpha Z_\alpha'$ 是 $Z_1,\cdots,Z_{n-1}$ 的函数，$\overline{X}$ 是 Z_n 的函数，而 $Z_1,\cdots,Z_{n-1}$ 与 Z_n 相互独立，故 A 与 $\overline{X}$ 也相互独立.

(4) 记 $B=(Z_1,\cdots,Z_{n-1})$，则 $A=BB'$，以下来证明：$P\{A>0\}=1$ 的充要条件是 $n>p$.

因为 $A=BB'$，B 是 $p\times(n-1)$ 矩阵. 显然 $\mathrm{rank}(A)=\mathrm{rank}(B)$. 当 A 为正定矩阵时 A 的秩是 p，故 B 的秩也是 p. 从而 $p<n$.

反之，设 $n>p$，我们来证明 $P\{A>0\}=1$，为此只须证 $P\{B$ 的前 p 列线性相关$\}=0$. 容易看出：

$$
\begin{aligned}
&P\{B\text{ 的前 }p\text{ 列线性相关}\}=P\{Z_1,\cdots,Z_p\text{ 线性相关}\}\\
&\leqslant\sum_{i=1}^{p}P\{Z_i\text{ 可表成 }Z_1,\cdots,Z_{i-1},Z_{i+1},\cdots,Z_p\text{的线性组合}\}\\
&=p\cdot P\{Z_1\text{ 可表成 }Z_2,\cdots,Z_p\text{ 的线性组合}\}\\
&=p\cdot \mathrm{E}[P\{Z_1\text{ 可表成 }z_2,\cdots,z_p\text{ 的线性组合}\\
&\qquad\qquad |Z_2=z_2,\cdots,Z_p=z_p\}]\\
&=p\cdot \mathrm{E}[P\{Z_1\text{ 落入由 }z_2,\cdots,z_p\text{ 张成的子空间}\\
&\qquad\qquad |Z_2=z_2,\cdots,Z_p=z_p\}]\\
&=p\cdot \mathrm{E}[P\{\text{存在 }p\text{ 维常向量 }\alpha\neq 0,\text{使 }\alpha' Z_1=0\\
&\qquad\qquad |Z_2=z_2,\cdots,Z_p=z_p\}]\\
&=p\cdot \mathrm{E}(0)=0.
\end{aligned}
$$

(证毕)

在证明过程中用到以下事实：由于 $Z_1\sim N_p(0,\Sigma)$，而 $\Sigma>0$(正定)，对常向量 $\alpha\neq 0$，$\alpha' Z_1\sim N(0,\alpha'\Sigma\alpha)$，且 $\alpha'\Sigma\alpha>0$，即 $P\{\alpha' Z_1=0\}=0$. 或者说 Z_1 取值落入任何维数小于 p 的子空间的概率是 0.

以下是 μ 和 Σ 的最大似然估计所具有的一些性质.

1. 无偏性

可以证明

$$
\mathrm{E}(\overline{X})=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_{i1})\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(x_{ip})\end{bmatrix}=\begin{bmatrix}\frac{1}{n}\sum_{i=1}^{n}\mu_1\\ \vdots\\ \frac{1}{n}\sum_{i=1}^{n}\mu_p\end{bmatrix}=\begin{bmatrix}\mu_1\\ \vdots\\ \mu_p\end{bmatrix}=\mu,
$$

故 $\overline{X}$ 是 μ 的无偏估计.

又

$$\mathrm{E}(A)=\mathrm{E}\Big(\sum_{\alpha=1}^{n-1}Z_\alpha Z_\alpha'\Big)=\sum_{\alpha=1}^{n-1}(\mathrm{E}(Z_\alpha Z_\alpha'))$$

$$=\Big(\sum_{\alpha=1}^{n-1}\mathrm{D}(Z_\alpha)\Big)=(n-1)\Sigma.$$

因而 Σ 的最大似然估计 $\hat{\Sigma}=\frac{1}{n}A$ 不是无偏估计. 为了得到无偏估计量,常作如下修正:

令 $S=\frac{1}{n-1}A$,则 S 是 Σ 的无偏估计. 常称 $\overline{X}=(\overline{x}_1,\cdots,\overline{x}_p)'$ 为**样本均值**; $S=\frac{1}{n-1}A$ 为样本协方差阵.

2. 有效性

可以证明 $\overline{X},S$ 是 μ,Σ 的"最小方差"无偏估计量,即 $\overline{X},S$ 是 μ,Σ 的有效估计量(见参考文献[2]).

3. 相合性(一致性)

可以证明当 $n\to\infty$ 时 $\overline{X},\hat{\Sigma}$ 是 μ,Σ 的强相合估计.

实际上,因 $\mathrm{E}(\overline{X})=\mu$,由强大数定律知

$$P\{\lim_n \overline{X}=\mu\}=1.$$

另一方面,因 $\hat{\Sigma}=\frac{1}{n}\sum_{\alpha=1}^{n-1}Z_\alpha Z_\alpha'$,而 $Z_1,\cdots,Z_{n-1}$ 相互独立同分布,共同分布是 $N_p(0,\Sigma)$,而 $\mathrm{E}(Z_\alpha Z_\alpha')=\Sigma(\alpha=1,\cdots,n-1)$. 再利用强大数定律知

$$P\{\lim_n \hat{\Sigma}=\Sigma\}=1.$$

4. 其他

还可以证明 $\overline{X},\hat{\Sigma}$ 是 μ,Σ 的充分统计量; $\overline{X}$ 是 μ 的极小极大估计量(最大风险达最小);且估计量具有渐近正态性.

四、参数函数的最大似然估计

为了从参数 μ,Σ 的最大似然估计来导出参数函数 $\varphi(\mu,\Sigma)$ 的最

大似然估计，我们首先介绍有关的概念与性质.

设参数向量 θ 的变化范围是 $\Theta\in\mathbb{R}^k$. $L(\theta)$是似然函数. 设 $w=g(\theta)$是 Θ 到 Θ^* 上的博雷尔(Borel)可测映射，这里 Θ^* 是 $\mathbb{R}^k$ 的子集. 对任何 $w\in\Theta^*$，令

$$M(w)=\sup_{\{\theta;\ g(\theta)=w\}}L(\theta).$$

定义 2.5.1 称 $M(w)$为函数 $g(\theta)$诱导出的**似然函数**.

定义 2.5.2 若 $\hat{w}$ 满足 $M(\hat{w})=\sup_{w}M(w)$，则称 $\hat{w}$ 是 $g(\theta)$的**最大似然估计**.

定理 2.5.3 若 $\hat{\theta}$ 是 θ 的最大似然估计，则 $\hat{w}=g(\hat{\theta})$是 $g(\theta)$ 的最大似然估计.

证明 任给 $w\in\Theta^*$，因 $g(\hat{\theta})=\hat{w}$，故有

$$M(w)=\sup_{\{\theta;\ g(\theta)=w\}}L(\theta)\leqslant\sup_{\theta\in\Theta}L(\theta)=L(\hat{\theta})\leqslant M(\hat{w}),$$

这就证明了 $\hat{w}=g(\hat{\theta})$是 $g(\theta)$的最大似然估计. **(证毕)**

既然多元正态分布 $N_p(\mu,\Sigma)$的参数 μ 和 Σ 有最大似然估计量 $\hat{\mu}=\overline{X}$，$\hat{\Sigma}=\frac{1}{n}A$，从定理 2.5.3 知，函数 $g(\mu,\Sigma)$的最大似然估计为 $g\left(\overline{X},\frac{1}{n}A\right)$.

例 2.5.1 设 p 维正态随机向量 $X=(X_1,\cdots,X_p)'$，X_i,X_j 的相关系数为

$$\rho_{ij}=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathrm{Var}(X_i)\cdot\mathrm{Var}(X_j)}}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\cdot\sigma_{jj}}},$$

其中 σ_{ij}是协方差阵 Σ 的第 i 行第 j 列的元素. 试求 ρ_{ij}的最大似然估计量 r_{ij}.

解 给定样本 $X_{(t)}(t=1,\cdots,n)$，则 Σ 的最大似然估计为

$$\frac{1}{n}\sum_{t=1}^{n}(X_{(t)}-\overline{X})(X_{(t)}-\overline{X})'=\frac{1}{n}A,$$

Σ 的元素 σ_{ij} 的最大似然估计

$$\hat{\sigma}_{ij}=\frac{1}{n}\sum_{t=1}^{n}(x_{ti}-\overline{x}_i)(x_{tj}-\overline{x}_j)=\frac{1}{n}a_{ij}.$$

由定理 2.5.3 知，相关系数 ρ_{ij} 的最大似然估计量 r_{ij} 为

$$r_{ij}=\frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\cdot\hat{\sigma}_{jj}}}=\frac{a_{ij}}{\sqrt{a_{ii}\cdot a_{jj}}}.$$

例 2.5.2　设

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\sim N_p\left(\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix}\right),$$

试求 $X^{(1)}$对 $X^{(2)}$的回归系数阵及 $X^{(2)}$给定时 $X^{(1)}$的条件协方差阵的最大似然估计量.

解　$X^{(2)}$给定时 $X^{(1)}$的条件分布为

$$(X^{(1)}|X^{(2)})\sim N_r(\mu_{1\cdot2},\Sigma_{11\cdot2}),$$

其中 $\mu_{1\cdot2}=\mu^{(1)}+\Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})$，且 $B=\Sigma_{12}\Sigma_{22}^{-1}$为 $X^{(1)}$对 $X^{(2)}$的回归系数阵；$\Sigma_{11\cdot2}=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$为条件协方差阵.

由样本 $X_{(t)}(t=1,\cdots,n)$，计算离差阵 A，且

$$A=\left[\begin{array}{c:c}A_{11}&A_{12}\\\hdashline A_{21}&A_{22}\end{array}\right],$$

其中 A_{11}为 r 阶方阵，A_{22}为 $p-r$ 阶方阵. 由定理 2.5.3 知，回归系数阵 B 的最大似然估计为

$$\hat{B}=\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}=\left(\frac{1}{n}A_{12}\right)\cdot\left(\frac{1}{n}A_{22}\right)^{-1}=A_{12}A_{22}^{-1}.$$

条件协方差阵 $\Sigma_{11\cdot2}$的最大似然估计为

$$\hat{\Sigma}_{11\cdot2}=\frac{1}{n}(A_{11}-A_{12}A_{22}^{-1}A_{21}).$$

习　题　二

2-1　设三维随机向量 $X\sim N_3(\mu,2I_3)$，已知

$$\mu=\begin{bmatrix}2\\0\\0\end{bmatrix},\quad A=\begin{bmatrix}0.5&-1&0.5\\-0.5&0&-0.5\end{bmatrix},\quad d=\begin{bmatrix}1\\2\end{bmatrix}.$$

试求 $Y=AX+d$ 的分布.

2-2　设 $X=(X_1,X_2)'\sim N_2(\mu,\Sigma)$，其中

$$\mu=\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix},\quad \Sigma=\sigma^2\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}.$$

(1) 试证明 X_1+X_2 和 X_1-X_2 相互独立；

(2) 试求 X_1+X_2 和 X_1-X_2 的分布.

2-3　设 $X^{(1)}$ 和 $X^{(2)}$ 均为 p 维随机向量，已知

$$X=\begin{bmatrix}X^{(1)}\\ X^{(2)}\end{bmatrix}\sim N_{2p}\left(\begin{bmatrix}\mu^{(1)}\\ \mu^{(2)}\end{bmatrix},\left[\begin{array}{c:c}\Sigma_1 & \Sigma_2\\ \hdashline \Sigma_2 & \Sigma_1\end{array}\right]\right),$$

其中 $\mu^{(i)}(i=1,2)$ 为 p 维向量，$\Sigma_i(i=1,2)$ 为 p 阶矩阵，

(1) 试证明 $X^{(1)}+X^{(2)}$ 和 $X^{(1)}-X^{(2)}$ 相互独立；

(2) 试求 $X^{(1)}+X^{(2)}$ 和 $X^{(1)}-X^{(2)}$ 的分布.

2-4　设 $X\sim N_3(\mu,\Sigma)$，其中

$$\mu=(\mu_1,\mu_2,\mu_3)',\quad \Sigma=\begin{bmatrix}1 & \rho & \rho\\ \rho & 1 & \rho\\ \rho & \rho & 1\end{bmatrix}\quad (0<\rho<1).$$

(1) 试求条件分布 $(X_1,X_2|X_3)$ 和 $(X_1|X_2,X_3)$；

(2) 给定 $X_3=x_3$ 时，试写出 X_1 和 X_2 的条件协方差.

2-5　设 $X\sim N_2(0,I_2)$，其中 $X=(X_1,X_2)'$. 试求当 X_1+X_2 给定时 X_1 的条件分布.

2-6　设 $X\sim N_3(\mu,\Sigma)$，其中

$$X=(X_1,X_2,X_3)',\quad \mu=(2,-3,1)',$$

$$\Sigma=\begin{bmatrix}1 & 1 & 1\\ 1 & 3 & 2\\ 1 & 2 & 2\end{bmatrix}.$$

(1) 试求 $3X_1-2X_2+X_3$ 的分布；

(2) 求二维向量 $a=(a_1,a_2)'$，使 X_3 与 $X_3-a'\begin{bmatrix}X_1\\ X_2\end{bmatrix}$ 相互独立.

2-7　设 $X\sim N_3(\mu,\Sigma)$，其中

$$X=(X_1,X_2,X_3)',\quad \mu=(-3,1,4)',$$

$$\Sigma=\begin{bmatrix}1 & -2 & 0\\ -2 & 5 & 0\\ 0 & 0 & 2\end{bmatrix}.$$

试问下列 5 对随机变量中哪几对是相互独立的，为什么？

(1) X_1 与 $2X_2$；　　(2) X_2 与 X_3；

(3) (X_1,X_2)与 X_3；　　(4) $\frac{1}{2}(X_1+X_2)$与 X_3；

(5) X_2 与 $X_2-\frac{5}{2}X_1-X_3$.

2-8　设 $X\sim N_p(\mu,\Sigma)$，A 为 $m\times p$ 常数矩阵，B 为 $k\times p$ 常数矩阵. 令 $Y=AX+d$，$Z=BX+c$. 证明

$$Y \text{ 与 } Z \text{ 独立} \iff A\Sigma B' = O_{m\times k}.$$

2-9　设

$$A=\begin{bmatrix} 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} & 1/\sqrt{4} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & 0 \\ 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & -3/\sqrt{12} \end{bmatrix}.$$

(1) 试证明 A 是一个正交矩阵(即 $AA'=I_4$)；

(2) 已知 $X\sim N_4(\mu\mathbf{1}_4,\sigma^2 I_4)$，设 $Y=(Y_1,Y_2,Y_3,Y_4)'=AX$，试证明：

① $Y_2^2+Y_3^2+Y_4^2=\sum_{i=1}^{4}(X_i-\overline{X})^2$，其中 $\overline{X}=\frac{1}{4}(X_1+X_2+X_3+X_4)$；

② Y_1,Y_2,Y_3,Y_4 相互独立；

③ $Y_1\sim N(2\mu,\sigma^2)$，$Y_i\sim N(0,\sigma^2)$ $(i=2,3,4)$.

2-10　设 $X\sim N_2(0,\Sigma)$，$\Sigma=\begin{bmatrix}4 & 2\\ 2 & 1\end{bmatrix}$，即 X 具有退化的正态分布. 试求一个矩阵 A，使 $X\overset{d}{=\!=}AU$，且 $U\sim N_2(0,I_2)$.

2-11　已知 $X=(X_1,X_2)'$的密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi}\exp\Big\{-\frac{1}{2}(2x_1^2+x_2^2+2x_1x_2 - 22x_1-14x_2+65)\Big\},$$

试求 X 的均值向量和协方差阵.

2-12　设 $X_1\sim N(0,1)$，令

$$X_2=\begin{cases}-X_1, & \text{当 } -1\leqslant X_1\leqslant 1,\\ X_1, & \text{其他.}\end{cases}$$

(1) 证明 $X_2\sim N(0,1)$;

(2) 证明(X_1,X_2)不是二元正态分布.

2-13 设 $X\sim N_p(\mu,\Sigma)$,A 为对称阵,试证明:

(1) $\mathrm{E}(XX')=\Sigma+\mu\mu'$;

(2) $\mathrm{E}(X'AX)=\mathrm{tr}(\Sigma A)+\mu' A\mu$;

(3) 当 $\mu=a\begin{bmatrix}1\\ \vdots\\ 1\end{bmatrix}\xlongequal{\text{def}}a\mathbf{1}_p$, $A=I_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p'$, $\Sigma=\sigma^2 I_p$ 时,试利用(1)和(2)的结果证明 $\mathrm{E}(X'AX)=\sigma^2(p-1)$.

若记 $X=(X_1,\cdots,X_p)'$,此时

$$X'AX=\sum_{i=1}^{p}(X_i-\overline{X})^2,$$

则

$$\mathrm{E}\Big[\sum_{i=1}^{p}(X_i-\overline{X})^2\Big]=\sigma^2(p-1).$$

2-14 试用对 μ,Σ 求微商的方法求总体 $N_p(\mu,\Sigma)$中参数 μ 和 Σ 的最大似然估计.

2-15 设 $X_{(1)},\cdots,X_{(n)}$为来自总体 $N_p(\mu,\Sigma)$的随机样本,若 $\mu=\mu_0$ 已知,试求总体 $N_p(\mu_0,\Sigma)$中参数 Σ 的最大似然估计.

2-16 设 $Z=(Y,X_1,\cdots,X_m)'$是 $m+1$ 维随机向量,$\mathrm{E}(Z)=0$,$\mathrm{D}(Z)=\Sigma$. 试证在一切的 m 元函数 $g(x_1,\cdots,x_m)$中,当 $g(x_1,\cdots,x_m)=\mathrm{E}(Y|X_1=x_1,\cdots,X_m=x_m)$时,$\mathrm{E}(Y-g(x_1,\cdots,x_m))^2$ 为极小.

2-17 设 $X\sim N_p(\mu,\Sigma)$,$\Sigma>0$,X 的密度函数记为 $f(x;\mu,\Sigma)$. 任给 $a>0$,试证明概率密度等高面

$$f(x;\mu,\Sigma)=a$$

是一个椭球面. 特别当 $p=2$ 且 $\Sigma=\sigma^2\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}$ $(\rho>0)$时,概率密度等高面就是平面上的一个椭圆. 试求该椭圆的方程、长轴和短轴.

2-18 设 $X_{(1)},\cdots,X_{(n)}$是来自 $N_p(\mu,\Sigma)$的随机样本,$c_i\geqslant 0$ $(i=$

$1,\cdots,n)$，$\sum_{i=1}^{n}c_i=1$，令 $Z=\sum_{i=1}^{n}c_iX_{(i)}$. 试证明：

(1) Z 是 μ 的无偏估计量；

(2) $Z\sim N_p(\mu,c'c\Sigma)$，其中 $c=(c_1,\cdots,c_n)'$；

(3) 当 $c=\frac{1}{n}\mathbf{1}_n$ 时，Z 的协方差阵在非负定的意义下达到极小.

2-19　为了了解某种橡胶的性能，今抽取 10 个样品，每个测量三项指标：硬度、变形和弹性，其数据如下表：

序号	硬度(X_1)	变形(X_2)	弹性(X_3)
1	65	45	27.6
2	70	45	30.7
3	70	48	31.8
4	69	46	32.6
5	66	50	31.0
6	67	46	31.3
7	68	47	37.0
8	72	43	33.6
9	66	47	33.1
10	68	48	34.2

试计算样本均值、样本离差阵、样本协方差阵和样本相关阵.

第三章　多元正态总体参数的假设检验

一元正态总体中，参数 μ,σ^2 的检验涉及到一个总体、两个总体，乃至多个总体的检验问题；推广到 p 元正态总体 $N_p(\mu,\Sigma)$，类似地，对参数向量 μ 和参数矩阵 Σ 涉及到的检验也有一个总体、两个总体，乃至多个总体的检验问题.

在一元统计中，用于检验 μ,σ^2 的抽样分布有 χ^2 分布、t 分布、F 分布等，它们都是由来自总体 $N(\mu,\sigma^2)$ 的随机样本导出的检验统计量. 推广到多元正态总体后，也有相应于以上三个常用分布的统计量：威沙特(Wishart)统计量，霍特林(Hotelling) T^2 统计量，威尔克斯(Wilks) Λ 统计量，讨论这些统计量的分布是多元统计分析所涉及的假设检验问题的基础.

§3.1　几个重要统计量的分布

一、正态变量二次型的分布

1. 分量独立的 n 维随机向量 X 的二次型

设 $X_i\sim N_1(\mu_i,\sigma^2)$ $(i=1,\cdots,n)$，且相互独立，记

$$X=\begin{bmatrix}X_1\\ \vdots\\ X_n\end{bmatrix},$$

则 $X\sim N_n(\mu,\sigma^2 I_n)$，其中 $\mu=(\mu_1,\cdots,\mu_n)'$.

X 的二次型具有以下一些结论：

结论 1　当 $\mu_i=0$ $(i=1,\cdots,n)$，$\sigma^2=1$ 时，则

$$\xi=X'X=\sum_{i=1}^{n}X_i^2\sim\chi^2(n);$$

当 $\mu_i=0$ $(i=1,2,\cdots,n)$，$\sigma^2\neq 1$ 时，则有

$$\frac{1}{\sigma^2}X'X\sim\chi^2(n)\quad(\text{或记为 } X'X\sim\sigma^2\chi^2(n)).$$

结论 2　当 $\mu_i\neq 0$ $(i=1,2,\cdots,n)$，$X'X$ 的分布常称为**非中心 χ^2 分布**.

定义 3.1.1　设 n 维随机向量 $X\sim N_n(\mu,I_n)$ $(\mu\neq 0)$，则称随机变量 $\xi=X'X$ 为服从 n 个自由度、非中心参数 $\delta=\mu'\mu=\sum_{i=1}^{n}\mu_i^2$ 的 χ^2 分布，记为 $X'X\sim\chi^2(n,\delta)$ 或 $X'X\sim\chi_n^2(\delta)$.

当 $X\sim N_n(\mu,\sigma^2I_n)$，$\mu\neq 0$，且 $\sigma^2\neq 1$ 时，令

$$Y_i=\frac{1}{\sigma}X_i.$$

显然

$$Y_i\sim N\left(\frac{\mu_i}{\sigma},1\right)\quad(i=1,\cdots,n),$$

则

$$Y'Y=\frac{1}{\sigma^2}X'X\sim\chi_n^2(\delta),$$

其中 $\delta=\frac{1}{\sigma^2}\mu'\mu$.

结论 3　设 $X\sim N_n(0_n,\sigma^2I_n)$，$A$ 为对称矩阵，且 $\operatorname{rank}(A)=r$，则二次型 $X'AX/\sigma^2\sim\chi^2(r)\Longleftrightarrow A^2=A$　(A 为**对称幂等矩阵**).

证明　$\Longrightarrow$：因 A 为对称矩阵，所以存在正交阵 Γ 使

$$\Gamma'A\Gamma=\operatorname{diag}(\lambda_1,\cdots\lambda_r,0\cdots 0).$$

令

$$Y=\Gamma'X\sim N_n(0_n,\sigma^2I_n),\quad X=\Gamma Y.$$

则

$$\xi=X'AX/\sigma^2=Y'\Gamma'A\Gamma Y/\sigma^2=\sum_{i=1}^{r}\lambda_iY_i^2/\sigma^2,$$

且 $Y_1,\cdots,Y_r$ 相互独立同 $N(0,\sigma^2)$ 分布. 故而 $Y_i^2/\sigma^2\sim\chi^2(1)$ $(i=1,\cdots,r)$，且相互独立. $\sum_{i=1}^{r}\lambda_iY_i^2/\sigma^2$ 的特征函数为

$$(1-2i\lambda_1t)^{-1/2}\cdot(1-2i\lambda_2t)^{-1/2}\cdot\cdots\cdot(1-2i\lambda_rt)^{-1/2}.$$

又知 $\xi=X'AX/\sigma^2\sim\chi^2(r)$，故 ξ 的特征函数为 $(1-2it)^{-r/2}$. 利用

$$(1-2it)^{r/2}=[(1-2i\lambda_1 t)(1-2i\lambda_2 t)\cdots(1-2i\lambda_r t)]^{1/2}$$

可得出 $\lambda_1=\lambda_2=\cdots=\lambda_r=1$，于是

$$\operatorname{diag}(1,\cdots,1,0,\cdots,0)=\Gamma'A\Gamma=\Gamma'A\Gamma\cdot\Gamma'A\Gamma=\Gamma'A^2\Gamma,$$

故 $A=A^2$，即 A 为对称幂等矩阵.

$\Leftarrow$：因 A 为对称幂等矩阵，而对称幂等矩阵的特征值非 0 即 1，且只有 r 个非 0 特征值，即存在正交矩阵 Γ，使

$$\Gamma'A\Gamma=\begin{bmatrix}I_r & O\\ O & O\end{bmatrix}.$$

令 $Y=\begin{bmatrix}Y_1\\ \vdots\\ Y_n\end{bmatrix}=\Gamma'X$（即 $X=\Gamma Y$），则

$$Y\sim N_n(0_n,\sigma^2\Gamma'I_n\Gamma)=N_n(0_n,\sigma^2I_n),$$

$$\frac{1}{\sigma^2}X'AX=\frac{1}{\sigma^2}Y'\Gamma'A\Gamma Y=\frac{1}{\sigma^2}Y'\begin{bmatrix}I_r & O\\ O & O\end{bmatrix}Y=\frac{1}{\sigma^2}\sum_{i=1}^{r}Y_i^2.$$

因为 $Y_i\sim N(0,\sigma^2)$ $(i=1,2,\cdots,r)$，且相互独立，所以

$$\xi=\frac{1}{\sigma^2}X'AX=\frac{1}{\sigma^2}\sum_{i=1}^{r}Y_i^2\sim\chi^2(r).\qquad\text{（证毕）}$$

结论 4 设 $X\sim N_n(\mu,\sigma^2I_n)$，$A=A'$，则

$$\frac{1}{\sigma^2}X'AX\sim\chi^2(r,\delta),$$

其中

$$\delta=\frac{1}{\sigma^2}\mu'A\mu\Longleftrightarrow A=A^2\quad\text{（对称幂等矩阵）},$$

且 $\operatorname{rank}(A)=r$ $(r\leqslant n)$.

结论 5 二次型与线性函数的独立性：设 $X\sim N_n(\mu,\sigma^2I_n)$，$A$ 为 n 阶对称矩阵，B 为 $m\times n$ 矩阵，令 $\xi=X'AX$，$Z=BX$（Z 为 m 维随机向量），若 $BA=O$，则 BX 和 $X'AX$ 相互独立.

证明 设 $\operatorname{rank}(A)=r>0$（当 $r=0$ 时 $A=O$，结论显然成立），存在正交矩阵 Γ 使

$$\Gamma'A\Gamma=\begin{bmatrix}D_r & O\\ O & O\end{bmatrix},\quad D_r=\begin{bmatrix}\lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_r\end{bmatrix},$$

其中 λ_i 是 A 的非零特征值($i=1,\cdots,r$).

因为

$$BA=B\Gamma\begin{bmatrix}D_r & O\\ O & O\end{bmatrix}\Gamma'=(C_1\vdots C_2)\begin{bmatrix}D_r & O\\ O & O\end{bmatrix}\Gamma'$$
$$=(C_1D_r\vdots O)\Gamma'=O,$$

其中 B 为 $m\times n$ 矩阵，O 为 $m\times n$ 零矩阵，C_1 为 $m\times r$ 矩阵，C_2 为 $m\times(n-r)$ 矩阵. 故有 $C_1D_r=O$，又 D_r 为对角矩阵，且 $\lambda_i\neq 0$，从而得 $C_1=O$.

令 $Y=\Gamma'X$，即 $X=\Gamma Y$. 则

$$Y=\begin{bmatrix}Y_1\\ \vdots\\ Y_n\end{bmatrix}\sim N_n(\Gamma'\mu,\sigma^2I_n),$$

即 $Y_1,\cdots,Y_n$ 独立. 因

$$X'AX=Y'\Gamma'A\Gamma Y=Y'\begin{bmatrix}D_r & O\\ O & O\end{bmatrix}Y=\sum_{i=1}^{r}\lambda_iY_i^2,$$

而

$$BX=B\Gamma Y=(C_1\vdots C_2)\begin{bmatrix}Y_1\\ \vdots\\ Y_n\end{bmatrix}=C_2\begin{bmatrix}Y_{r+1}\\ \vdots\\ Y_n\end{bmatrix},$$

由于 $Y_1,\cdots,Y_r$ 与 $Y_{r+1},\cdots,Y_n$ 相互独立，故 $X'AX$ 与 BX 相互独立.

(证毕)

结论 5 反之也成立，即：若 BX 和 $X'AX$ 相互独立，则 $BA=O$.

结论 6　两个二次型相互独立的条件：设 $X\sim N_n(\mu,\sigma^2I_n)$，$A$，$B$ 为 n 阶对称矩阵，则

$$AB=O\Longleftrightarrow X'AX\text{ 与 }X'BX\text{ 相互独立.}$$

2. 一般 p 维正态随机向量的二次型

p 维随机向量的二次型具有下述结论：

结论 1　设 $X\sim N_p(\mu,\Sigma)$，$\Sigma>0$，则 $X'\Sigma^{-1}X\sim\chi^2(p,\delta)$，其中 $\delta=\mu'\Sigma^{-1}\mu$.

证明　因 $\Sigma>0$，由正定矩阵的分解可得 $\Sigma=CC'$ (C 为非退化方

阵).

令 $Y=C^{-1}X$,即 $X=CY$. 则

$$Y \sim N_p(C^{-1}\mu, C^{-1}\Sigma(C^{-1})').$$

因 $\Sigma=CC'$,所以 $Y\sim N_p(C^{-1}\mu, I_p)$,且有

$$X'\Sigma^{-1}X = Y'C'\Sigma^{-1}CY = Y'Y \sim \chi^2(p,\delta),$$

其中 $$\delta = (C^{-1}\mu)'(C^{-1}\mu) = \mu'\Sigma^{-1}\mu.$$ (证毕)

结论 2 设 $X\sim N_p(\mu,\Sigma)$,$\Sigma>0$,A 为对称矩阵,$\mathrm{rank}(A)=r$. 则

$$(X-\mu)'A(X-\mu) \sim \chi^2(r) \Longleftrightarrow \Sigma A\Sigma A\Sigma = \Sigma A\Sigma.$$

证明 因 $\Sigma>0$,则有 $\mathrm{rank}(\Sigma)=p$,且存在正交矩阵 Γ 和 $\lambda_i(i=1,2,\cdots,p)$,使得

$$\Sigma = \Sigma^{1/2}\cdot\Sigma^{1/2},$$

其中 $\Sigma^{1/2}=\Gamma\mathrm{diag}(\sqrt{\lambda_1},\cdots,\sqrt{\lambda_p})\Gamma'$ 为 Σ 的平方根矩阵.

记

$$\Sigma^{-1/2} = \Gamma\mathrm{diag}\left(\frac{1}{\sqrt{\lambda_1}},\cdots,\frac{1}{\sqrt{\lambda_p}}\right)\Gamma',$$

显然有 $\Sigma^{1/2}\Sigma^{1/2}=I_p$.

令

$$Y = \Sigma^{-1/2}(X-\mu) \sim N_p(0_p, I_p),$$

这里

$$\begin{aligned}\mathrm{D}(Y) &= \mathrm{D}(\Sigma^{-1/2}(X-\mu)) = \Sigma^{-1/2}\cdot\Sigma\cdot(\Sigma^{-1/2})' \\ &= \Sigma^{-1/2}\cdot\Sigma^{1/2}\Sigma^{1/2}\cdot\Sigma^{-1/2} = I_p,\end{aligned}$$

$$(X-\mu)'A(X-\mu) = Y'\Sigma^{1/2}A\Sigma^{1/2}Y \xlongequal{\text{def}} Y'CY.$$

由本节的小节一、1 "分量独立的 n 维随机向量 X 的二次型"中结论 3 可知

$$Y'CY \sim \chi^2(p) \Longleftrightarrow C^2 = C,$$

即 $$\Sigma^{1/2}A\Sigma^{1/2}\cdot\Sigma^{1/2}A\Sigma^{1/2} = \Sigma^{1/2}A\Sigma^{1/2}.$$

将上式两边左右乘 $\Sigma^{1/2}$,即得 $\Sigma A\Sigma A\Sigma=\Sigma A\Sigma$. (证毕)

结论 3 设 $X\sim N_p(\mu,\Sigma)$,$\Sigma>0$,A 和 B 为 p 阶对称矩阵,则

$$(X-\mu)'A(X-\mu) \text{ 与 } (X-\mu)B(X-\mu) \text{ 独立}$$
$$\Longleftrightarrow \Sigma A\Sigma B\Sigma = O_{p\times p}.$$

3. 非中心 t 分布和非中心 F 分布

定义 3.1.2　设 $X\sim N(\delta,1)$ 与 $Y\sim\chi^2(n)$ 相互独立，令

$$T=\frac{X\sqrt{n}}{\sqrt{Y}},$$

则称 T 的分布为具有 n 个自由度、非中心参数为 δ 的非中心 t 分布，记为 $T\sim t(n,\delta)$.

定义 3.1.3　设 $X\sim\chi^2(m,\delta)$ 与 $Y\sim\chi^2(n)$ 独立，令

$$F=\frac{X/m}{Y/n},$$

则称 F 的分布为具有自由度为 m,n 和非中心参数为 δ 的 F 分布，记为 $F\sim F(m,n,\delta)$.

4. 非中心 χ^2 分布、非中心 t 分布和非中心 F 分布的应用

一元统计中，关于在一个正态总体 $N(\mu,\sigma^2)$ 的均值检验中，检验 H_0：$\mu=\mu_0$ 时，检验统计量为

$$T=\frac{\overline{X}-\mu_0}{\sqrt{s^2/n}}\overset{H_0下}{\sim}t(n-1),$$

否定域为 $\{|T|>\lambda\}$，其中 λ 满足：$P\{|T|>\lambda\}=\alpha$（显著性水平）.

当否定 H_0 时，可能犯**第一类错误**，且

第一类错误的概率 $=P\{\text{“以真当假”}\}=P\{|T|>\lambda|\mu=\mu_0\}$

$=$ 显著性水平 α；

当 H_0 相容时，可能犯**第二类错误**，且

第二类错误的概率 $=P\{\text{“以假当真”}\}=P\{|T|\leqslant\lambda|\mu\neq\mu_0\}$

$$\overset{\text{设}\mu=\mu_1\neq\mu_0}{=\!=\!=}P\left\{\left|\frac{\overline{X}-\mu_1+(\mu_1-\mu_0)}{\sqrt{s^2/n}}\right|\leqslant\lambda|\mu=\mu_1\right\}$$

$=\beta$.

此时检验统计量 $T\sim t(n-1,\delta)$（非中心参数 $\delta=\sqrt{n}(\mu_1-\mu_0)/\sigma$），利用非中心 t 分布可以计算第二类错误 β 的值，从而得到检验法的功效函数为 $1-\beta$.

类似地，非中心 χ^2 分布和非中心 F 分布在一元统计的相应检验中，将应用非中心分布来计算第二类错误.

二、威沙特(Wishart)分布

威沙特分布是一元统计中 χ^2 分布的推广. 多元正态总体 $N_p(\mu,\Sigma)$中,常用样本均值向量 $\overline{X}$ 作为 μ 的估计,样本协方差阵

$$S=\frac{1}{n-1}A$$

作为 Σ 的估计. 第二章的定理 2.5.2 已给出

$$\overline{X}\sim N_p\left(\mu,\frac{\Sigma}{n}\right).$$

一元统计中,用样本方差

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)}-\overline{X})^2$$

作为 σ^2 的估计,而且知道

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_{(i)}-\overline{X})^2\sim\chi^2(n-1).$$

推广到 p 元正态总体,样本协方差阵 $S=\frac{1}{n-1}A$ 及**随机阵** A(离差阵)的分布是什么?

设 $X_{(\alpha)}(\alpha=1,\cdots,n)$为来自总体 $N_p(0,\Sigma)$的随机样本,记 $X=(X_{(1)},\cdots,X_{(n)})'$为 $n\times p$ 样本数据阵. 考虑随机阵

$$W=\sum_{i=1}^{n}X_{(i)}X'_{(i)}=(X_{(1)},\cdots,X_{(n)})\begin{bmatrix}X'_{(1)}\\ \vdots\\ X'_{(n)}\end{bmatrix}=X'X$$

的分布. 当 $p=1$ 时(总体 $X\sim N_1(0,\sigma^2)$),

$$W=\sum_{i=1}^{n}X_{(i)}^2=(X_{(1)},\cdots,X_{(n)})\begin{bmatrix}X_{(1)}\\ \vdots\\ X_{(n)}\end{bmatrix}=X'X\sim\sigma^2\chi^2(n).$$

在一元正态总体情况下,

$$\xi=\frac{1}{\sigma^2}\sum_{i=1}^{n}X_{(i)}^2\sim\chi^2(n),$$

推广到 p 元正态总体时,随机阵 W 的分布是什么?

1. 威沙特分布的定义

定义 3.1.4　设 $X_{(\alpha)}\sim N_p(0,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立，记 $X=(X_{(1)},\cdots,X_{(n)})'$ 为 $n\times p$ 矩阵，则称随机阵

$$W=\sum_{\alpha=1}^{n}X_{(\alpha)}X_{(\alpha)}'=X'X$$

的分布为威沙特分布，记为 $W\sim W_p(n,\Sigma)$.

显然，$p=1$ 时，$X_{(\alpha)}\sim N(0,\sigma^2)$，此时

$$W=\sum_{\alpha=1}^{n}X_{(\alpha)}^2\sim\sigma^2\chi^2(n),$$

即 $W_1(n,\sigma^2)$就是 $\sigma^2\chi^2(n)$. 当 $p=1,\sigma^2=1$ 时，$W_1(n,1)$就是 $\chi^2(n)$.

一般地，设 $X_{(\alpha)}\sim N_p(\mu,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立，记

$$M=\begin{bmatrix}\mu_1 & \cdots & \mu_p\\ \vdots & & \vdots\\ \mu_1 & \cdots & \mu_p\end{bmatrix}=\mathbf{1}_n\mu',$$

则称 $W=X'X$ 服从非中心参数为 Δ 的非中心威沙特分布，记为 $W\sim W_p(n,\Sigma,\Delta)$，其中

$$\Delta=M'M=(\mathbf{1}_n\mu')'(\mathbf{1}_n\mu')=\mu\mathbf{1}_n'\mathbf{1}_n\mu'=n\mu\mu'.$$

当 $X_{(\alpha)}\sim N_p(\mu_\alpha,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立时，非中心参数

$$\Delta=\sum_{\alpha=1}^{n}\mu_\alpha\mu_\alpha'\quad\text{或}\quad\Delta=M'M.$$

这里

$$M=\begin{bmatrix}\mu_{11} & \cdots & \mu_{1p}\\ \vdots & & \vdots\\ \mu_{n1} & \cdots & \mu_{np}\end{bmatrix}=\begin{bmatrix}\mu_1'\\ \vdots\\ \mu_n'\end{bmatrix},$$

其中 p 为随机阵 W 的阶数，n 为自由度，一元统计中的 σ^2 对应 p 元统计中的协方差阵 Σ.

随机阵 W 的密度函数是威沙特于 1928 年推导出来的，故此分布称为威沙特分布.

2. 威沙特分布的性质

性质 1　设 $X_{(\alpha)}\sim N_p(\mu,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立，则样本离差

阵 A 服从威沙特分布，即

$$A=\sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})'\sim W_p(n-1,\Sigma).$$

证明 根据第二章的定理 2.5.2 知

$$A=\sum_{\alpha=1}^{n-1}Z_\alpha Z_\alpha',$$

而 $Z_\alpha\sim N_p(0,\Sigma)$ $(\alpha=1,\cdots,n-1)$相互独立，由定义 3.1.4 可知

$$A\sim W_p(n-1,\Sigma).\qquad\text{(证毕)}$$

由于威沙特分布是 χ^2 分布的推广，因此它还具有 χ^2 分布的一些其他性质.

性质 2 关于自由度 n 具有**可加性**：设 $W_i\sim W_p(n_i,\Sigma)$ $(i=1,\cdots,k)$相互独立，则

$$\sum_{i=1}^{k}W_i\sim W_p(n,\Sigma),\quad\text{其中 } n=n_1+\cdots+n_k.$$

性质 3 设 p 阶随机阵 $W\sim W_p(n,\Sigma)$，C 是 $m\times p$ 常数矩阵，则 m 阶随机阵 CWC' 也服从威沙特分布，即

$$CWC'\sim W_m(n,C\Sigma C').$$

证明 因 $W\overset{d}{=\!=}\sum_{\alpha=1}^{n}Z_\alpha Z_\alpha'\sim W_p(n,\Sigma)$，其中 $Z_\alpha\sim N_p(0,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立.

令 $Y_\alpha=CZ_\alpha$，则 $Y_\alpha\sim N_m(0,C\Sigma C')$. 故

$$\sum_{\alpha=1}^{n}Y_\alpha Y_\alpha'=\sum_{\alpha=1}^{n}CZ_\alpha\cdot Z_\alpha'C'$$

$$\overset{d}{=\!=}CWC'\sim W_m(n,C\Sigma C').\qquad\text{(证毕)}$$

特别地：

(1) $aW\sim W_p(n,a\Sigma)$ $(a>0$，为常数$)$.

在性质 3 中只须取 $C=\sqrt{a}\,I_p$，即得此结论.

(2) 设 $l'=(l_1,\cdots,l_p)$，则 $l'Wl=\xi\sim W_1(n,l'\Sigma l)$，即

$$\xi\sim\sigma^2\chi^2(n)\quad(\text{其中 }\sigma^2=l'\Sigma l).$$

在性质 3 中只须取 $C=l'$，即得此结论.

性质 4 分块威沙特矩阵的分布(习题三中第 3-4 题)：设

$X_{(\alpha)}\sim N_p(0,\Sigma)$ $(\alpha=1,\cdots,n)$相互独立，其中

$$\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\begin{matrix}r\\ p-r\end{matrix}.$$

又已知随机阵

$$W=\sum_{\alpha=1}^{n}X_{(\alpha)}X'_{(\alpha)}=\begin{bmatrix}W_{11} & W_{12}\\ W_{21} & W_{22}\end{bmatrix}\begin{matrix}r\\ p-r\end{matrix}\sim W_p(n,\Sigma),$$

则

(1) $W_{11}\sim W_r(n,\Sigma_{11})$，$W_{22}\sim W_{p-r}(n,\Sigma_{22})$；

(2) 当 $\Sigma_{12}=O$ 时，W_{11}与 W_{22}相互独立.

性质 5　设 $W\sim W_p(n,\Sigma)$，记 $W_{22\cdot1}=W_{22}-W_{21}W_{11}^{-1}W_{12}$，则

$$W_{22\cdot1}\sim W_{p-r}(n-r,\Sigma_{22\cdot1}),$$

其中 $\Sigma_{22\cdot1}=\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$，且 $W_{22\cdot1}$与 W_{11}相互独立.

性质 6　设随机阵 $W\sim W_p(n,\Sigma)$，则 $\mathrm{E}(W)=n\Sigma$.

性质 7　设 $X\sim N_{n\times p}(M,I_n\otimes\Sigma)$，$A$ 为 n 阶对称矩阵，则

$$X'AX\sim W_p(r,\Sigma,\Delta),$$

其中 $\Delta=M'AM\iff A^2=A$，且 $\mathrm{rank}(A)=r$.

这是一元统计中 n 维观测向量 X 的二次型分布在 p 维情况下的推广(证明见参考文献[2]).

性质 8　设 $X\sim N_{n\times p}(M,I_n\otimes\Sigma)$，$A$ 和 B 均为 n 阶对称幂等矩阵，则

$$X'AX\text{ 与 }X'BX\text{ 相互独立}\iff AB=O.$$

这是一元统计中$(p=1)n$ 维观测向量 X 的两个二次型相互独立的条件在 p 维情况下的推广(证明见参考文献[2]).

三、霍特林(Hotelling) T^2 分布

1. 霍特林 T^2 分布的定义

一元统计中，若 $X\sim N(0,1)$，$\xi\sim\chi^2(n)$，X 与 ξ 相互独立，则随机变量

$$t=\frac{X}{\sqrt{\xi/n}}\sim t(n).$$

下面把 $t^2=nX^2/\xi=nX'\xi^{-1}X$ 的分布推广到 p 元总体. 设总体 $X\sim N_p(0,\Sigma)$,随机阵 $W\sim W_p(n,\Sigma)$,我们来讨论 $T^2=nX'W^{-1}X$ 的分布.

定义 3.1.5 设 $X\sim N_p(0,\Sigma)$, 随机阵 $W\sim W_p(n,\Sigma)$ $(\Sigma>0, n\geqslant p)$,且 X 与 W 相互独立, 则称统计量 $T^2=nX'W^{-1}X$ 为**霍特林 T^2 统计量**,其分布称为**服从 n 个自由度的 T^2 分布**,记为

$$T^2\sim T^2(p,n).$$

更一般地,若 $X\sim N_p(\mu,\Sigma)$ $(\mu\neq 0)$,则称 T^2 的分布为**非中心霍特林 T^2 分布**,记为 $T^2\sim T^2(p,n,\mu)$.

2. 霍特林 T^2 分布的性质

性质 1 设 $X_{(\alpha)}(\alpha=1,\cdots,n)$是来自 p 元总体 $N_p(\mu,\Sigma)$的随机样本, $\overline{X}$ 和 A 分别是正态总体 $N_p(\mu,\Sigma)$的样本均值向量和样本离差阵,则统计量

$$\begin{aligned}T^2&=(n-1)[\sqrt{n}\,(\overline{X}-\mu)]'A^{-1}[\sqrt{n}\,(\overline{X}-\mu)]\\&=n(n-1)(\overline{X}-\mu)'A^{-1}(\overline{X}-\mu)\\&\sim T^2(p,n-1).\end{aligned}$$

证明 事实上,因 $\overline{X}\sim N_p\left(\mu,\frac{1}{n}\Sigma\right)$,则 $\sqrt{n}\,(\overline{X}-\mu)\sim N_p(0,\Sigma)$. 而 $A\sim W_p(n-1,\Sigma)$,且 A 与 $\overline{X}$ 相互独立. 由定义 3.1.5 知

$$T^2\sim T^2(p,n-1).\qquad\text{(证毕)}$$

性质 2 T^2 与 F 分布的关系: 设 $T^2\sim T^2(p,n)$,则

$$\frac{n-p+1}{np}T^2\sim F(p,n-p+1).$$

在一元统计中,若 $t=\dfrac{X}{\sqrt{\xi/n}}\sim t(n)$,则 $t^2=\dfrac{X^2/1}{\xi/n}\sim F(1,n)$.

当 $p=1$ 时,一元总体 $X\sim N(0,\sigma^2)$, $X_{(\alpha)}(\alpha=1,\cdots,n)$为来自总体 X 的随机样本,则

$$W\overset{\mathrm{d}}{=\!=}\sum_{\alpha=1}^{n}X_{(\alpha)}X'_{(\alpha)}=\sum_{\alpha=1}^{n}X_{(\alpha)}^2\sim W_1(n,\sigma^2)\quad(\text{即 }\sigma^2\chi^2(n)).$$

所以

$$\frac{n}{n}T^2=nX'W^{-1}X=\frac{nX^2}{W}=\frac{(X/\sigma)^2}{(W/\sigma^2 n)}\sim F(1,n).$$

一般地,

$$\frac{n-p+1}{p}\cdot\frac{T^2}{n}\overset{\mathrm{d}}{=\!=}\frac{n-p+1}{p}X'W^{-1}X$$
$$=\frac{n-p+1}{p}X'\Sigma^{-1}X\Big/\frac{X'\Sigma^{-1}X}{X'W^{-1}X}\overset{\mathrm{def}}{=\!=}\frac{n-p+1}{p}\cdot\frac{\xi}{\eta}$$
$$=\frac{\xi/p}{\eta/n-p+1}\sim F(p,n-p+1),$$

其中 $\xi=X'\Sigma^{-1}X\sim\chi^2(p,\delta)$ $(\delta=0)$. 还可证明

$$\eta=\frac{X'\Sigma^{-1}X}{X'W^{-1}X}\sim\chi^2(n-p+1),$$

且 ξ 与 η 独立(详细证明见参考文献[2]).

性质 3　设 $X_{(\alpha)}(\alpha=1,2,\cdots,n)$为来自 p 元总体 $N_p(\mu,\Sigma)$的随机样本. $\overline{X},A$ 分别为样本均值向量和样本离差阵. 记

$$T^2=n(n-1)\overline{X}A^{-1}\overline{X},$$

则

$$\frac{n-p}{p}\frac{T^2}{n-1}\sim F(p,n-p,\delta),$$

其中 $\delta=n\mu'\Sigma^{-1}\mu$.

一元统计中($p=1$ 时), t 统计量与参数 σ^2 无关. 类似地有以下性质.

性质 4　T^2 统计量的分布只与 p,n 有关,而与 Σ 无关.

设 $U\sim N_p(0,I_p),W_0\sim W_p(n,I_p),U$ 和 W_0 相互独立,则

$$nU'W_0^{-1}U\overset{\mathrm{d}}{=\!=}nX'W^{-1}X\sim T^2(p,n).$$

事实上,因 $X\sim N_p(0,\Sigma)$ $(\Sigma>0),W\sim W_p(n,\Sigma)$,则 $\Sigma^{-1/2}X\sim N_p(0,I_p)$,且 $\Sigma^{-1/2}W\Sigma^{-1/2}\sim W_p(n,I_p)$,因此

$$U\overset{\mathrm{d}}{=\!=}\Sigma^{-1/2}X,\quad W_0\overset{\mathrm{d}}{=\!=}\Sigma^{-1/2}W\Sigma^{-1/2}.$$

所以 $nU'W_0^{-1}U\overset{\mathrm{d}}{=\!=}nX'W^{-1}X\sim T^2(p,n)$.

性质 5　T^2 统计量对非退化变换保持不变.

设 $X_{(\alpha)}(\alpha=1,\cdots,n)$是来自 p 元总体 $N_p(\mu,\Sigma)$的随机样本, $\overline{X}_x$ 和 A_x 分别表示正态总体 X 的样本均值向量和样本离差阵,则由性质 1 有

$$T_x^2=n(n-1)(\overline{X}_x-\mu)'A_x^{-1}(\overline{X}_x-\mu)\sim T^2(p,n-1).$$

令 $Y_{(\alpha)}=CX_{(\alpha)}+d$ $(\alpha=1,\cdots,n)$,其中 C 为 $p\times p$ 非退化常数矩阵,d 为 p 维常向量,则可以证明(习题三中第 3-4 题)

$$T_y^2 = T_x^2.$$

四、威尔克斯(Wilks)Λ统计量及其分布

1. 威尔克斯 Λ 分布的定义

一元统计中,设 $\xi \sim \chi^2(m)$,$\eta \sim \chi^2(n)$,且相互独立,则

$$F = \frac{\xi/m}{\eta/n} \sim F(m,n).$$

在两个总体($N(\mu_1,\sigma_x^2)$和 $N(\mu_2,\sigma_y^2)$)方差齐性检验中(H_0: $\sigma_x^2=\sigma_y^2$),设 $X_{(i)}(i=1,\cdots,m)$为来自 $N(\mu_1,\sigma_x^2)$的随机样本,$Y_{(j)}(j=1,\cdots,n)$为来自 $N(\mu_2,\sigma_y^2)$的随机样本,取 σ_x^2 和 σ_y^2 的估计量(样本方差)分别为

$$s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_{(i)} - \overline{X})^2 \quad 和 \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{(i)} - \overline{Y})^2,$$

则检验统计量

$$F = \frac{s_x^2}{s_y^2} \overset{H_0下}{\sim} F(m-1,n-1).$$

在 p 元总体 $N_p(\mu,\Sigma)$中,协方差阵 Σ 的估计量为

$$\hat{\Sigma} = \frac{1}{n-1}A \quad \left(或\frac{1}{n}A\right).$$

在检验 H_0: $\Sigma_1=\Sigma_2$ 时,如何用一个数值来描述对矩阵的离散程度的估计呢?一般可用矩阵的行列式、迹或特征值等数量指标来描述总体的分散程度.

定义 3.1.6 设 $X \sim N_p(\mu,\Sigma)$,则称协方差阵的行列式$|\Sigma|$为 X **的广义方差**. 若 $X_{(\alpha)}(\alpha=1,\cdots,n)$为 p 元总体 X 的随机样本,A 为样本离差阵,则称$\left|\frac{1}{n}A\right|$或$\left|\frac{1}{n-1}A\right|$为**样本广义方差**.

有了广义方差的概念后,在多元统计的协方差阵齐性检验中,类似一元统计,可考虑两个广义方差之比构成的统计量——威尔克斯统计量的分布.

定义 3.1.7 设 $A_1 \sim W_p(n_1,\Sigma)$,$A_2 \sim W_p(n_2,\Sigma)$ $(\Sigma>0, n_1 \geqslant p)$,且 A_1 与 A_2 独立,则称广义方差之比

$$\Lambda = \frac{|A_1|}{|A_1 + A_2|}$$

为**威尔克斯统计量或 Λ 统计量**，其分布称为**威尔克斯分布**，记为

$$\Lambda \sim \Lambda(p, n_1, n_2).$$

当 $p=1$ 时，Λ 统计量的分布正是一元统计中的参数为 $n_1/2$，$n_2/2$ 的 β 分布(记为 $\beta(n_1/2, n_2/2)$).

2. Λ 统计量与 T^2 或 F 统计量的关系

在实际应用中，常把 Λ 统计量化为 T^2 统计量，进而化为 F 统计量. 然后我们利用熟悉的 F 统计量来解决多元统计分析中有关检验的问题.

结论 1　当 $n_2=1$ 时，设 $n_1=n>p$，则

$$\Lambda(p, n, 1) \overset{d}{=\!=} \frac{1}{1 + \frac{1}{n}T^2(p, n)},$$

或

$$T^2(p, n) = n \cdot \frac{1 - \Lambda(p, n, 1)}{\Lambda(p, n, 1)},$$

$$\frac{n - p + 1}{np}T^2 = \frac{n - p + 1}{p}\frac{1 - \Lambda}{\Lambda} \overset{d}{=\!=} F(p, n - p + 1).$$

证明　设 $X_{(\alpha)}(\alpha=1, \cdots, n, n+1)$ 相互独立同 $N_p(0, \Sigma)$ 分布，显然有

$$W_1 = \sum_{\alpha=1}^{n} X_{(\alpha)}X'_{(\alpha)} \sim W_p(n, \Sigma),$$

$$W = \sum_{\alpha=1}^{n+1} X_{(\alpha)}X'_{(\alpha)} \sim W_p(n+1, \Sigma).$$

由定义 3.1.7，知

$$\Lambda = \frac{|W_1|}{|W|} \sim \Lambda(p, n, 1),$$

又因 $W=W_1+X_{(n+1)} \cdot X'_{(n+1)}$，我们利用分块矩阵行列式的公式(见附录)，可得

$$|W| = |W_1 + X_{(n+1)}X'_{(n+1)}| = \left|\begin{array}{c:c} W_1 & -X_{(n+1)} \\ \hdashline X'_{(n+1)} & 1 \end{array}\right|_1^p$$

$$\overset{\text{分块求行列式公式}}{=\!=} |W_1|(1 + X'_{(n+1)}W_1^{-1}X_{(n+1)}).$$

所以

$$\Lambda=\frac{|W_1|}{|W|}=\frac{1}{1+X'_{(n+1)}W_1^{-1}X_{(n+1)}}$$

$$\overset{\mathrm{d}}{=\!=}\frac{1}{1+\frac{1}{n}T^2(p,n)}. \qquad \text{(证毕)}$$

结论 2 当 $n_2=2$ 时,设 $n_1=n>p$,则

$$\frac{n-p+1}{p}\frac{1-\sqrt{\Lambda(p,n,2)}}{\sqrt{\Lambda(p,n,2)}}\overset{\mathrm{d}}{=\!=}F(2p,2(n-p+1)).$$

结论 3 当 $p=1$ 时,则

$$\frac{n_1}{n_2}\frac{1-\Lambda(1,n_1,n_2)}{\Lambda(1,n_1,n_2)}\overset{\mathrm{d}}{=\!=}F(n_2,n_1).$$

利用 $\Lambda(1,n_1,n_2)$就是 $\beta(n_1/2,n_2/2)$,以及 β 分布与 F 分布的关系即得此结论.

结论 4 当 $p=2$ 时,则

$$\frac{n_1-1}{n_2}\cdot\frac{1-\sqrt{\Lambda(2,n_1,n_2)}}{\sqrt{\Lambda(2,n_1,n_2)}}\overset{\mathrm{d}}{=\!=}F(2n_2,2(n_1-1)).$$

结论 5 当 $n_2>2,p>2$ 时,可用 χ^2 统计量或 F 统计量近似.

博克斯(Box)(1949)给出以下结论:

设 $\Lambda\sim\Lambda(p,n_1,n_2)$,则当 $n\to\infty$时,

$$-r\ln\Lambda\sim\chi^2(pn_2),$$

其中 $r=n_1-\frac{1}{2}(p-n_2+1)$.

当 n 不太大时也有一些近似分布,我们将在相应的假设检验中介绍.

3. 两个重要结论

下面不加证明地给出两个很有用的结论.

结论 1 若 $\Lambda\sim\Lambda(p,n_1,n_2)$,则存在 $B_k\sim\beta\left(\frac{n_1-p+k}{2},\frac{n_2}{2}\right)$ $(k=1,\cdots,p)$相互独立,使得

$$\Lambda\overset{\mathrm{d}}{=\!=}B_1B_2\cdots B_p.$$

结论 2 若 $n_2<p$,则

$$\Lambda(p,n_1,n_2)\overset{d}{=\!=}\Lambda(n_2,p,n_1+n_2-p).$$

结论 2 是一元统计中 $F(n,m)\overset{d}{=\!=}\dfrac{1}{F(m,n)}$ 的推广.

§3.2　单总体均值向量的检验及置信域

本节讨论单个 p 元正态总体 $N_p(\mu,\Sigma)$ 的统计推断问题，包括均值向量的检验和均值的置信域问题. p 维正态随机向量的每一个分量都是一元正态变量，关于均值向量的推断问题能否化为 p 个一元正态的均值推断问题呢？显然这是不完全的. 因为 p 个分量之间往往有互相依赖的关系，分开进行统计推断，往往得不出正确的结论. 但我们可以构造出类似于一元统计中的统计量，用来对均值向量进行检验或求置信域.

一、均值向量的检验

设总体 $X\sim N_p(\mu,\Sigma)$，随机样本 $X_{(\alpha)}(\alpha=1,\cdots,n)$. 检验

$$H_0:\ \mu=\mu_0(\mu_0\text{ 为已知向量}),\quad H_1:\ \mu\neq\mu_0.$$

1. 当 $\Sigma=\Sigma_0$ 已知时均值向量的检验

因

$$\overline{X}\sim N_p\left(\mu,\frac{1}{n}\Sigma_0\right),\quad \sqrt{n}\,(\overline{X}-\mu)\sim N_p(0,\Sigma_0),$$

利用二次型分布的结论，知

$$(\overline{X}-\mu)'\left(\frac{1}{n}\Sigma_0\right)^{-1}(\overline{X}-\mu)\sim\chi^2(p).$$

取检验统计量为

$$T_0^2=n(\overline{X}-\mu_0)'\Sigma_0^{-1}(\overline{X}-\mu_0)\overset{H_0下}{\sim}\chi^2(p).$$

按传统的检验方法，对给定的显著性水平 α，查 χ^2 分布临界值表得 λ_α，使 $P\{T_0^2>\lambda_\alpha\}=\alpha$，则否定域为 $\{T_0^2>\lambda_\alpha\}$.

由样本值 $x_{(\alpha)}(\alpha=1,\cdots,n)$，计算 $\overline{X}$ 及 T_0^2 值，若 $T_0^2>\lambda_\alpha$，则否定 H_0，否则 H_0 相容.

利用统计软件(如 SAS 系统),还可以通过计算显著性概率值(p 值)给出检验结果,且由此得出的结论更丰富.

假设在 H_0 成立情况下,随机变量 $T_0^2 \sim \chi^2(p)$,由样本值计算得到 T_0^2 的值为 d,同时可以计算以下概率值:

$$p = P\{T_0^2 \geqslant d\},$$

常称此概率值为**显著性概率值**,或简称为 p **值**.

对给定的显著性水平 α,当 $p<\alpha$ 时,则在显著性水平 α 下否定假设 H_0;在这种情况下,可能犯"以真当假"的第一类错误,且 α 就是犯第一类错误的概率.

当 $p \geqslant \alpha$ 时,则在显著性水平 α 下 H_0 相容;在这种情况下,可能犯"以假当真"的第二类错误,且犯第二类错误的概率 β 为

$$\beta = P\{T_0^2 \leqslant \lambda_\alpha \mid \text{当 } \mu = \mu_1 \neq \mu_0\},$$

其中检验统计量 $T_0^2 \sim \chi^2(p,\delta)$,非中心参数

$$\delta = n(\mu_1 - \mu_0)' \Sigma_0^{-1} (\mu_1 - \mu_0).$$

p 值的直观含义可这样看,检验统计量 T_0^2 的大小反映 $\overline{X}$ 与 μ_0 的偏差大小,当 H_0 成立时 T_0^2 值应较小.现由观测数据计算 T_0^2 值为 d;当 H_0 成立时统计量 $T_0^2 \sim \chi^2(p)$,由 χ^2 分布可计算该统计量 $\geqslant d$ 的概率值(即 p 值).比如 $p=0.02<\alpha=0.05$,这时出现一个比小概率标准($\alpha=0.05$)还要小的事件$\{T_0^2 \geqslant d\}$. 也就是说,在 $\mu=\mu_0$ 假设下,观测数据中极少情况会出现 T_0^2 的值大于等于 d 值,故在 0.05 显著性水平下有足够的证据否定原假设,即认为 μ 与 μ_0 有显著地差异.

又比如当 $p=0.22 \geqslant \alpha=0.05$ 时,表示在 $\mu=\mu_0$ 的假设下,观测数据中经常会出现 T_0^2 的值大于等于 d 值的情况,故在 0.05 显著性水平下没有足够的证据否定原假设,即认为 μ 与 μ_0 没有显著地差异.

2. 当 Σ 未知时均值向量的检验

当 $p=1$ 时(一元统计),取检验统计量为

$$t = \frac{(\overline{X} - \mu_0)\sqrt{n}}{\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_{(i)} - \overline{X})^2}} \sim t(n-1),$$

或等价地取检验统计量

$$t^2 = n(\overline{X} - \mu_0)'\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)} - \overline{X})^2\right)^{-1}(\overline{X} - \mu_0).$$

推广到多元，考虑统计量

$$T^2 = n(\overline{X} - \mu_0)'\left(\frac{1}{n-1}A\right)^{-1}(\overline{X} - \mu_0),$$

因而

$$\overline{X} \overset{H_0下}{\sim} N_p\left(\mu_0, \frac{1}{n}\Sigma\right), \quad \sqrt{n}(\overline{X} - \mu_0) \overset{H_0下}{\sim} N_p(0, \Sigma).$$

样本离差阵为

$$A = \sum_{\alpha=1}^{n}(X_{(\alpha)} - \overline{X})(X_{(\alpha)} - \overline{X})' \sim W_p(n-1, \Sigma).$$

由定义 3.1.5 可知

$$\begin{aligned} T^2 &= (n-1) \cdot [\sqrt{n}(\overline{X} - \mu_0)]'A^{-1}[\sqrt{n}(\overline{X} - \mu_0)] \\ &= (n-1)n(\overline{X} - \mu_0)'A^{-1}(\overline{X} - \mu_0) \sim T^2(p, n-1), \end{aligned}$$

再利用 T^2 与 F 分布的关系，检验统计量取为

$$F = \frac{(n-1) - p + 1}{(n-1)p}T^2 \overset{H_0下}{\sim} F(p, (n-1) - p + 1)$$

$$\overset{H_0下}{\sim} F(p, n-p).$$

例 3.2.1　人的出汗多少与人体内钠和钾的含量有一定的关系. 今测量了 20 名健康成年女性的出汗量(X_1)、钠的含量(X_2)和钾的含量(X_3)(数据见表 3.1). 试检验 H_0: $\mu = \mu_0 = (4, 50, 10)'$, H_1: $\mu \neq \mu_0$ ($\alpha = 0.05$).

表 3.1　成年女性的出汗量及其体内钠和钾含量的数据

序号	X_1	X_2	X_3	序号	X_1	X_2	X_3
1	3.7	48.5	9.3	2	4.7	65.1	8.0
3	3.8	47.2	10.9	4	3.2	53.2	12.0
5	3.1	55.5	9.7	6	4.6	36.1	7.9
7	2.4	24.8	14.0	8	7.2	33.1	7.6
9	6.7	47.4	8.5	10	5.4	54.1	11.3
11	3.9	36.9	12.7	12	4.5	58.8	12.3
13	3.5	27.8	9.8	14	4.5	40.2	8.4
15	1.5	13.5	10.1	16	8.5	56.4	7.1
17	4.5	71.6	8.2	18	6.5	52.8	10.9
19	4.1	44.1	11.2	20	5.5	40.9	9.4

解 记随机向量 $X=(X_1,X_2,X_3)'$，假定 $X\sim N_3(\mu,\Sigma)$. 检验 H_0: $\mu=\mu_0$, H_1: $\mu\neq\mu_0$. 取检验统计量为

$$F=\frac{n-p}{(n-1)p}T^2 \quad (p=3,n=20).$$

由样本值计算得：$\overline{X}=(4.64,45.4,9.965)'$，及

$$A=\begin{bmatrix} 54.708 & & \\ 190.190 & 3795.98 & \\ -34.372 & -107.16 & 68.926 \end{bmatrix}^{①},$$

$$A^{-1}=\begin{bmatrix} 0.0308503 & & \\ -0.0011620 & 0.0003193 & \\ 0.0135773 & -0.0000830 & 0.0211498 \end{bmatrix}.$$

进一步计算可得

$$D^2=(n-1)(\overline{X}-\mu_0)'A^{-1}(\overline{X}-\mu_0)=19\times(\overline{X}-\mu_0)'Y$$
$$=19\times 0.02563=0.48694,$$

其中 $Y=A^{-1}(\overline{X}-\mu_0)$. Y 也可通过解线性方程组 $AY=(\overline{X}-\mu_0)$ 得到.

$$T^2=n(n-1)(\overline{X}-\mu_0)'A^{-1}(\overline{X}-\mu_0)=9.7388,$$

$$F=\frac{n-p}{(n-1)p}T^2=2.9045.$$

对给定 $\alpha=0.05$，按传统的检验方法，可查 F 分布临界值表得 $\lambda_c=F_{3,17}(0.05)=3.2$. 比较由样本值计算得到的 F 值及临界值，因 $F=2.9045<3.2$，故 H_0 相容.

利用统计软件进行检验时，首先计算 p 值(此时检验统计量 $F\sim F(3,17)$)：

$$p=P\{F\geqslant 2.9045\}=0.06493.$$

因 $p=0.06493>0.05=\alpha$，故 H_0 相容. 在这种情况下，可能犯第二类错误，且犯第二类错误的概率为 $\beta=P\{F\leqslant 3.2\,|\,\mu=\overline{X}\}=0.3616$（假定总体均值 $\mu=\mu_1\neq\mu_0$，取 $\mu_1=\overline{X}$).

下面介绍构造检验法的似然比原理，并说明由一元统计推广得到的 T^2 统计量是检验 H_0 的似然比统计量.

① 因 A 为对称矩阵，故只列出下三角部分，以下同.

二、似然比统计量

在数理统计中关于总体参数的假设检验，通常是利用最大似然原理导出似然比统计量来进行检验. 在多元统计分析中几乎所有重要的检验都是利用最大似然比原理给出的. 下面我们回顾一下最大似然比原理.

设 p 元总体的密度函数为 $f(x,\theta)$，其中 θ 是未知参数，且 $\theta\in\Theta$（参数空间），又设 Θ_0 是 Θ 的子集，我们希望对下列假设：

$$H_0:\theta\in\Theta_0,\quad H_1:\theta\overline{\in}\Theta_0$$

作出判断，这就是假设检验问题. 称 H_0 为**原假设**（或**零假设**），H_1 为**对立假设**（或**备择假设**）.

从总体 X 抽取容量为 n 的样本 $X_{(t)}(t=1,\cdots,n)$. 把样本的联合密度函数

$$L(x_{(1)},\cdots,x_{(n)};\theta)=\prod_{t=1}^{n}f(x_{(t)};\theta)$$

记为 $L(X;\theta)$，并称它为**样本的似然函数**.

引入统计量

$$\lambda=\max_{\theta\in\Theta_0}L(X;\theta)\Big/\max_{\theta\in\Theta}L(X;\theta),$$

它是样本 $X_{(t)}(t=1,\cdots,n)$ 的函数，常称 λ 为**似然比统计量**. 由于 $\Theta_0\subset\Theta$，从而 $0\leqslant\lambda\leqslant1$.

由最大似然比原理知，如果 λ 取值太小，说明 H_0 为真时观测到此样本 $X_{(t)}(t=1,\cdots,n)$ 的概率比 H_0 为不真时观测到此样本 $X_{(t)}(t=1,\cdots,n)$ 的概率要小得多. 故有理由认为假设 H_0 不成立，所以从似然比出发，以上检验问题的否定域为

$$\{\lambda(X_{(1)},\cdots,X_{(n)})<\lambda_\alpha\}.$$

按传统的检验方法，λ_α 是由显著性性水平 α 确定的临界值，它满足当 H_0 成立时使得：

$$P\{\lambda(X_{(1)},\cdots,X_{(n)})<\lambda_\alpha\}=\alpha.$$

为了得到 λ_α，必须研究似然比统计量 λ 的抽样分布. 在一些特殊的情况下，可以得到 λ 的精确分布；但在很多情况下是得不到 λ 的精确分布的. 当样本量很大且满足一定正则条件时，$-2\ln\lambda$ 的抽样分布

与 χ^2 分布十分接近. 下面不加证明地给出一条很有用的结论.

定理 3.2.1 当样本容量 n 很大时,

$$-2\ln\lambda = -2\ln\left[\left(\max_{\theta\in\Theta_0} L(X;\theta)\right) \Big/ \max_{\theta\in\Theta} L(X;\theta)\right]$$

近似服从自由度为 f 的 χ^2 分布,其中 $f=\Theta$ 的维数$-\Theta_0$ 的维数.

本章将讨论的一些检验问题,就是利用似然比统计量的近似分布进行检验的方法. 下面我们来导出当 Σ 未知时检验均值向量 $\mu=\mu_0$ 的似然比统计量,并讨论它的分布.

设样本的似然函数为 $L(\mu,\Sigma)$. 检验均值向量 $\mu=\mu_0$ 的似然比统计量为

$$\lambda = \max_{\mu=\mu_0,\Sigma>0} L(\mu_0,\Sigma) \Big/ \max_{\mu,\Sigma>0} L(\mu,\Sigma).$$

在第二章 §2.5 中已经导出: 上面比式的分母当 $\mu=\overline{X}$, $\Sigma=\frac{1}{n}A$ 时达最大值,且最大值为

$$\max_{\mu,\Sigma>0} L(\mu,\Sigma) = (2\pi)^{-np/2}\left|\frac{1}{n}A\right|^{-n/2} e^{-np/2}.$$

由习题二第 2-15 题知,上面比式的分子当

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(X_{(i)}-\mu_0)(X_{(i)}-\mu_0)' = \frac{1}{n}A_0$$

时达最大值,且最大值为

$$\max_{\Sigma>0} L(\mu_0,\Sigma) = (2\pi)^{-np/2}\left|\frac{1}{n}A_0\right|^{-n/2} e^{-np/2}.$$

故

$$\lambda = \frac{|A_0|^{-n/2}}{|A|^{-n/2}} = \left(\frac{|A|}{|A_0|}\right)^{n/2}.$$

以下来推导似然比统计量 λ 与 T^2 的关系:

$$\begin{aligned} A_0 &= \sum_{i=1}^{n}(X_{(i)}-\mu_0)(X_{(i)}-\mu_0)' \\ &= \sum_{i=1}^{n}(X_{(i)}-\overline{X}+\overline{X}-\mu_0)(X_{(i)}-\overline{X}+\overline{X}-\mu_0)' \\ &= A + n(\overline{X}-\mu_0)(\overline{X}-\mu_0)', \end{aligned}$$

利用分块矩阵行列式的性质有:

$$|A_0| = |A + n(\overline{X}-\mu_0)(\overline{X}-\mu_0)'|$$

$$= \left| \begin{array}{c|c} A & -\sqrt{n}(\overline{X}-\mu_0) \\ \hline \sqrt{n}(\overline{X}-\mu_0)' & 1 \end{array} \right|$$

$$= |A| \cdot (1 + n(\overline{X}-\mu_0)'A^{-1}(\overline{X}-\mu_0)).$$

所以

$$\frac{|A|}{|A_0|} = \frac{1}{1+n(\overline{X}-\mu_0)'A^{-1}(\overline{X}-\mu_0)} = \frac{1}{1+\frac{1}{n-1}T^2},$$

其中

$$T^2 = (n-1)n(\overline{X}-\mu_0)'A^{-1}(\overline{X}-\mu_0) \overset{H_0下}{\sim} T^2(p,n-1).$$

否定域：

$$\{\lambda < \lambda_\alpha\} \Longleftrightarrow \{T^2 > T_\alpha^2\} \Longleftrightarrow \{F > F_\alpha\},$$

其中 $F = \frac{n-p}{p}\frac{T^2}{n-1} \overset{H_0下}{\sim} F(p,n-p)$.

三、置信域与联立置信区间

在一元统计中，讨论均值的假设检验问题本质上也等价于求均值的置信区间. 下面就单个多维正态总体均值向量的置信域的概念作为一元统计中置信区间的推广给出简单介绍.

1. 置信域

假设 $X_{(t)}(t=1,2,\cdots,n)$ 来自 p 元正态总体 $N_p(\mu,\Sigma)$ (Σ 未知)，由前面的讨论可知

$$T^2 = n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \sim T^2(p,n-1),$$

或者

$$F = \frac{n-p}{(n-1)p}T^2 \sim F(p,n-p).$$

任给置信度 $1-\alpha$，查 F 分布临界值表得 F_α 满足

$$P\{F \leqslant F_\alpha\} = 1-\alpha, \tag{3.2.1}$$

则均值向量 μ 的置信度为 $1-\alpha$ 的置信域为

$$T^2 = n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu) \leqslant \frac{(n-1)p}{n-p}F_\alpha.$$

该置信域是一个中心在 $\overline{X}$ 的椭球.

当检验假设 H_0：$\mu=\mu_0$ 时，若 μ_0 落入上述置信域内，即

$$T^2 = n(\overline{X} - \mu_0)'S^{-1}(\overline{X} - \mu_0) \leqslant \frac{(n-1)p}{n-p}F_\alpha,$$

则在显著性水平 α 下，H_0 相容；若 μ_0 没有落入上述置信域内，则否定 H_0. 可见在多元统计中，讨论均值向量的假设检验问题本质上也等价于求均值向量的置信域.

例 3.2.2 沿用例 3.2.1 的数据，试求 μ 的置信度为 95%的置信椭球.

解 由观测数据计算样本均值向量 $\overline{X}$ 和样本离差阵 A 及样本协方差阵 S

$$S = \frac{1}{n-1}A = \begin{bmatrix} 2.8794 & & \\ 10.0100 & 199.7884 & \\ -1.8090 & -5.6400 & 3.6277 \end{bmatrix},$$

S 的特征值 λ 和单位正交特征向量 l 分别为

$$\lambda_1 = 200.4625,\quad \lambda_2 = 4.5316,\quad \lambda_3 = 1.3014;$$

$$l_1 = (0.05084, 0.9983, -0.02907)',$$

$$l_2 = (-0.5737, 0.05302, 0.8173)',$$

$$l_3 = (0.8175, -0.02488, 0.5754)'.$$

记 $c^2 = \frac{(n-1)p}{n(n-p)}F_{0.05} = \frac{19\times 3}{20\times 17}\times 3.2 = 0.5365$. 由 S^{-1} 的谱分解式

$$S^{-1} = \sum_{i=1}^{3}\frac{1}{\lambda_i}l_i l_i',$$

并令 $Y_i = (\overline{X}-\mu)'l_i (i=1,2,3)$，则 μ 的置信度为 95%的置信椭球为

$$\frac{Y_1^2}{\lambda_1 c^2} + \frac{Y_2^2}{\lambda_2 c^2} + \frac{Y_3^2}{\lambda_3 c^2} \leqslant 1.$$

置信椭球的第一长轴半径为 $d_1 = \sqrt{\lambda_1}c = 10.3703$，方向沿 l_1；第二长轴半径为 $d_2 = \sqrt{\lambda_2}c = 1.5592$，方向沿 l_2；短轴半径为 $d_3 = \sqrt{\lambda_3}c = 0.8356$，方向沿 l_3. 第一长轴与短轴的比为 $d_1/d_3 = 12.4106$，即第一长轴的长度是短轴的 12 倍还多.

2. 联立置信区间

在构造均值向量 μ 的置信域 $n(\overline{X}-\mu_0)'S^{-1}(\overline{X}-\mu_0) \leqslant c^2$ 的同时，我们往往更需要考查 μ 的线性组合的联立置信区间.

设 $X\sim N_p(\mu,\Sigma)$，考虑 X 的线性组合

$$Z=a_1X_1+a_2X_2+\cdots+a_pX_p=a'X,$$

由多元正态分布的性质 2 可知：$Z\sim N(a'\mu,a'\Sigma a)$. 假设 $X_{(t)}$ $(t=1,2,\cdots,n)$为 p 元正态总体 $N_p(\mu,\Sigma)$的简单随机样本，则总体 Z 的样本为

$$Z_{(t)}=a'X_{(t)}\quad(t=1,2,\cdots,n),$$

且样本均值和样本方差分别为 $\overline{Z}=a'\overline{X}$，$s_z^2=a'Sa$，这里 $\overline{X}$ 和 S 分别是样本 $X_{(t)}(t=1,2,\cdots,n)$的样本均值和样本协方差阵.

对任意的 a，考虑 $a'\mu$ 的置信区间便能够得到所要的联立置信区间. 事实上，当 a 固定而 $\sigma_z^2=a'\Sigma a$ 未知时，$\mu_z=a'\mu$ 的置信度为 $1-\alpha$ 的置信区间可根据 t 统计量

$$t=\frac{\overline{Z}-\mu_z}{s_z/\sqrt{n}}=\frac{\sqrt{n}\,(a'\overline{X}-a'\mu)}{\sqrt{a'Sa}}$$

得到. 于是置信区间为

$$a'\overline{X}-t_{\alpha/2}\frac{\sqrt{a'Sa}}{\sqrt{n}}\leqslant a'\mu\leqslant a'\overline{X}+t_{\alpha/2}\frac{\sqrt{a'Sa}}{\sqrt{n}},\qquad(3.2.2)$$

其中 $t_{\alpha/2}$ 满足：$P\{|t|\leqslant t_{\alpha/2}\}=1-\alpha$ (这里 $t\sim t(n-1)$).

由(3.2.2)式可以给出 μ 的分量 $\mu_i(i=1,2,\cdots,p)$的置信区间，如取 $a=e_i=(0,\cdots,1,\cdots,0)'$，即取 e_i 第 i 个分量为 1，其余均为 0 的向量，则(3.2.2)式给出一元正态均值 $a'\mu=\mu_i$ 的置信区间. 显然通过选择不同的系数向量 a，便可得到 μ 的若干个线性组合的置信度为 $1-\alpha$ 的置信区间；但请注意，这时总的置信度不再是 $1-\alpha$，而比 $1-\alpha$ 低. 下面给出构造所有 $a'\mu$ 的联立置信区间估计的 Scheffe 方法.

对给定的样本 $X_{(t)}(t=1,2,\cdots,n)$和系数向量 a，若全体 $a'\mu$ 值的置信区间是由(3.2.2)式给出的，则不等式

$$t^2=\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}\leqslant t_{\alpha/2}^2$$

成立. 若让 a 变化，求所有 $a'\mu$ 的联立置信区间，那么应将(3.2.2)式的右边换上更大的常数才较为合理. 为此来求最大值

$$\max_{a\neq0}t^2=\max_{a\neq0}\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}.$$

根据附录中定理 7.1 有

$$\max_{a\neq 0}\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}=n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)=T^2, \tag{3.2.3}$$

且最大值在 a 与 $S^{-1}(\overline{X}-\mu)$ 成比例时达到.

定理 3.2.2 假设 $X_{(t)}(t=1,2,\cdots,n)$ 为来自 p 元正态总体 $N_p(\mu,\Sigma)$ ($\Sigma>0$ 未知)的随机样本,则对所有的 a,区间

$$[a'\overline{X}-d,a'\overline{X}+d]\quad\left(\text{其中 } d=\sqrt{\frac{(n-1)p}{n(n-p)}F_\alpha a'Sa}\right)$$

包含 $a'\mu$ 的概率为 $1-\alpha$ (其中 F_α 满足(3.2.1)式).

证明 由(3.2.3)式知, $T^2=n(\overline{X}-\mu)'S^{-1}(\overline{X}-\mu)\leqslant c^2$ 意味着对一切 a,有

$$\frac{n[a'(\overline{X}-\mu)]^2}{a'Sa}\leqslant c^2,$$

即对一切 a,有

$$a'\overline{X}-c\sqrt{\frac{a'Sa}{n}}\leqslant a'\mu\leqslant a'\overline{X}+c\sqrt{\frac{a'Sa}{n}}.$$

取 $c^2=\frac{(n-1)p}{n-p}F_\alpha$($F_\alpha$ 满足(3.2.1)式),故对所有 a,则有

$$P\{T^2\leqslant c^2\}=1-\alpha. \qquad \text{(证毕)}$$

由于置信概率由 T^2 分布确定,因此为方便起见,以后称定理 3.2.2 给出的联立置信区间为 T^2 **区间**. 在 T^2 区间中,若取 $a=e_i=(0,\cdots,1,\cdots,0)'$,我们便同时得到 $\mu_i(i=1,\cdots,p)$ 的**置信度**均为 $1-\alpha$ 的 T^2 区间

$$\overline{x}_i-c\sqrt{\frac{s_{ii}}{n}}\leqslant\mu_i\leqslant\overline{x}_i+c\sqrt{\frac{s_{ii}}{n}}\quad\left(\text{其中 } c=\sqrt{\frac{(n-1)p}{(n-p)}F_\alpha}\right), \tag{3.2.4}$$

其中 s_{ii} 为样本协方差阵 S 的第 i 个对角元素.

请注意: 如果在(3.2.2)式中取 $a=e_i(i=1,\cdots,p)$,即每次考虑一个分量的置信区间,则得到单个 $\mu_i(i=1,2,\cdots,p)$ 的置信度为 $1-\alpha$ 的置信区间

$$\bar{x}_i - t_{\alpha/2}\sqrt{\frac{s_{ii}}{n}} \leqslant \mu_i \leqslant \bar{x}_i + t_{\alpha/2}\sqrt{\frac{s_{ii}}{n}}. \tag{3.2.5}$$

比如在例 3.2.2 中，均值向量 μ 的第一个分量 μ_1，由(3.2.4)式可得置信度为 95%的置信限为[3.3972,5.8828]；由(3.2.5)式可得单个 μ_1 的置信度为 95%的置信限为[3.8459,5.4341].

若把这 p 个形如(3.2.5)式的区间合在一起构成 $\mu_i(i=1,2,\cdots,p)$的联立置信区间，其置信度比 $1-\alpha$ 低. 请读者仔细比较(3.2.4)和(3.2.5)式在统计意义上的差别.

§3.3　多总体均值向量的检验

一、两正态总体均值向量的检验

1. 两总体协方差阵相等(但未知)时均值向量的检验

设 $X_{(\alpha)}(\alpha=1,\cdots,n)$为来自总体 $X\sim N_p(\mu^{(1)},\Sigma)$的随机样本；$Y_{(\alpha)}(\alpha=1,\cdots,m)$为来自总体 $Y\sim N_p(\mu^{(2)},\Sigma)$的随机样本，且相互独立，$\Sigma$ 未知. 检验

$$H_0:\ \mu^{(1)}=\mu^{(2)},\quad H_1:\ \mu^{(1)}\neq\mu^{(2)}.$$

当 $p=1$ 时，因 $\bar{X}\sim N_1\left(\mu^{(1)},\frac{\sigma^2}{n}\right)$，$\bar{Y}\sim N_1\left(\mu^{(2)},\frac{\sigma^2}{m}\right)$，且相互独立，故有

$$\bar{X}-\bar{Y}\sim N_1\left(\mu^{(1)}-\mu^{(2)},\left(\frac{1}{n}+\frac{1}{m}\right)\sigma^2\right),$$

$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(X_{(i)}-\bar{X})^2+\sum_{j=1}^{m}(Y_{(j)}-\bar{Y})^2\right]\sim\chi^2(n-1+m-1).$$

取检验统计量为

$$t=\frac{(\bar{X}-\bar{Y})\Big/\sqrt{\frac{1}{n}+\frac{1}{m}}}{\sqrt{\left[\sum(X_{(i)}-\bar{X})^2+\sum(Y_{(j)}-\bar{Y})^2\right]\Big/(n+m-2)}}$$

$$\overset{H_0\text{下}}{\sim} t(n+m-2),$$

即

$$t^2=\frac{nm}{m+n}(\overline{X}-\overline{Y})'\left[\frac{\sum(X_{(i)}-\overline{X})^2+\sum(Y_{(j)}-\overline{Y})^2}{n+m-2}\right]^{-1}(\overline{X}-\overline{Y})$$

$$\overset{H_0下}{\sim} F(1,n+m-2).$$

推广到 p 元总体，检验统计量的形式类似，可考虑以下检验统计量 T^2：

$$T^2=\frac{nm}{n+m}(\overline{X}-\overline{Y})'\left(\frac{A_1+A_2}{n+m-2}\right)^{-1}(\overline{X}-\overline{Y}),$$

其中 A_1 和 A_2 是两总体的样本离差阵. 上式是一元统计中的偏差平方和 $\sum(X_{(i)}-\overline{X})^2$ 在 p 元情况下的推广.

以下来证明统计量 $T^2\sim T^2(p,n+m-2)$. 因

$$(\overline{X}-\overline{Y})\overset{H_0下}{\sim} N_p\left(0,\left(\frac{1}{n}+\frac{1}{m}\right)\Sigma\right),$$

$$\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})\sim N_p(0,\Sigma),$$

$$A_1=\sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})'\sim W_p(n-1,\Sigma),$$

$$A_2=\sum_{\alpha=1}^{m}(Y_{(\alpha)}-\overline{Y})(Y_{(\alpha)}-\overline{Y})'\sim W_p(m-1,\Sigma).$$

由威沙特分布的可加性知

$$A_1+A_2\sim W_p(n+m-2,\Sigma).$$

由 T^2 统计量的定义 3.1.5 可知

$$T^2=(n+m-2)\frac{nm}{n+m}(\overline{X}-\overline{Y})'(A_1+A_2)^{-1}(\overline{X}-\overline{Y})$$

$$\overset{H_0下}{\sim} T^2(p,n+m-2).$$

利用 T^2 与 F 的关系，检验统计量取为

$$F=\frac{(n+m-2)-p+1}{(n+m-2)\cdot p}T^2\sim F(p,n+m-p-1).$$

可以证明 T^2(或 F)统计量是检验以上假设 H_0 的似然比统计量.

例 3.3.1 为了研究日、美两国在华投资企业对中国经营环境的评价是否存在差异，今从两国在华投资企业中各抽出 10 家，让其对中国的政治、经济、法律、文化等环境进行打分，评分结果如表 3.2 所示(表中序号 1 至 10 为美国在华投资企业的代号，11 至 20 为日

本在华投资企业的代号. 数据来源于：国务院发展研究中心 APEC 在华投资企业情况调查).

表 3.2　日、美两国在华投资企业对中国经营环境的评价数据

序号	政治环境	经济环境	法律环境	文化环境
1	65	35	25	60
2	75	50	20	55
3	60	45	35	65
4	75	40	40	70
5	70	30	30	50
6	55	40	35	65
7	60	45	30	60
8	65	40	25	60
9	60	50	30	70
10	55	55	35	75
11	55	55	40	65
12	50	60	45	70
13	45	45	35	75
14	50	50	50	70
15	55	50	30	75
16	60	40	45	60
17	65	55	45	75
18	50	60	35	80
19	40	45	30	65
20	45	50	45	70

解　比较日、美两国在华投资企业对中国多方面的经营环境的评价是否有差异问题，就是两总体均值向量是否相等的检验问题. 记美国在华投资企业对中国 4 个方面的经营环境的评价为 4 元总体 X，并设 $X\sim N_4(\mu^{(1)},\Sigma)$. 日本在华投资企业对中国经营环境的评价为 4 元总体 Y，并设 $Y\sim N_4(\mu^{(2)},\Sigma)$. 来自两总体的样本容量 $n=m=10$. 检验

$$H_0:\ \mu^{(1)}=\mu^{(2)},\quad H_1:\ \mu^{(1)}\neq\mu^{(2)}.$$

取检验统计量为

$$F=\frac{n+m-p-1}{(n+m-2)p}T^2\quad (p=4,n=m=10),$$

由样本值计算得：

$$\overline{X} = (64,43,30.5,63)',$$

$$\overline{Y} = (51.5,51,40,70.5)',$$

$$A_1 = \sum_{\alpha=1}^{n}(X_{(\alpha)} - \overline{X})(X_{(\alpha)} - \overline{X})'$$

$$= \begin{bmatrix} 490 & & & \\ -170 & 510 & & \\ -120 & 10 & 322.5 & \\ -245 & 310 & 260.0 & 510 \end{bmatrix},$$

$$A_2 = \sum_{\alpha=1}^{m}(Y_{(\alpha)} - \overline{Y})(Y_{(\alpha)} - \overline{Y})'$$

$$= \begin{bmatrix} 502.5 & & & \\ 60.0 & 390 & & \\ 175.0 & 50 & 450 & \\ -7.5 & 195 & -100 & 322.5 \end{bmatrix}.$$

进一步计算可得：

$$D^2 = (n+m-2)(\overline{X} - \overline{Y})'(A_1 + A_2)^{-1}(\overline{X} - \overline{Y})$$

$$= 18 \times 0.3318055 = 5.9725,$$

$$T^2 = \frac{nm}{n+m}D^2 = 29.8625,$$

$$F = \frac{n+m-p-1}{(n+m-2)p}T^2 = 6.2214.$$

对给定显著性水平 $\alpha=0.01$，利用统计软件进行检验时，首先计算 p 值(此时检验统计量 $F \sim F(4,15)$)：

$$p = P\{F \geqslant 6.2214\} = 0.0037.$$

因 $p=0.0037<0.01=\alpha$，故否定 H_0，即日、美两国在华投资企业对中国经营环境的评价存在显著性差异. 在这种情况下，可能犯第一类错误，且犯第一类错误的概率为 0.01.

2. 两总体协方差阵不等时均值向量的检验

在一元统计中($p=1$ 时)，当 $\sigma_1^2 \neq \sigma_2^2$ 时，检验 H_0：$\mu^{(1)} = \mu^{(2)}$ 也没有很好的方法，以下介绍实用中的几种方法.

(1) 当 $n=m$ 时，作为成对数据进行处理：令

$$Z_{(i)} = X_{(i)} - Y_{(i)} \quad (i = 1,\cdots,n),$$

将两个总体化为单个 p 元总体 Z 的均值检验问题

$$H_0: \mu^{(1)} = \mu^{(2)} \iff H_0: \mu_Z = 0_p.$$

利用 § 3.2 中介绍的方法进行检验. 注意：在这里 $X_{(i)}, Y_{(i)}(i=1,\cdots,n)$ 相互独立的信息没有利用.

(2) 当 $n \neq m$ 时(不妨设 $n<m$)：想法也是将其化为单个 p 元新总体的均值检验问题. 若只取 n 对数据，按(1)的方法处理又将损失一些信息. 改进的办法是利用 $X_{(i)}(i=1,\cdots,n)$ 和 $Y_{(j)}(j=1,\cdots,m)$ 构造新总体 Z 的样本 $Z_{(i)}$，令

$$Z_{(i)} = X_{(i)} - \sqrt{\frac{n}{m}}Y_{(i)} + \frac{1}{\sqrt{nm}}\sum_{j=1}^{n}Y_{(j)} - \frac{1}{m}\sum_{j=1}^{m}Y_{(j)}$$
$$(i = 1,2,\cdots,n),$$

可以证明：

$$\mathrm{E}(Z_{(i)}) = \mu^{(1)} - \sqrt{\frac{n}{m}}\mu^{(2)} + \frac{1}{\sqrt{nm}}n\mu^{(2)} - \frac{1}{m}\cdot m\mu^{(2)}$$
$$= \mu^{(1)} - \mu^{(2)}\left(\sqrt{\frac{n}{m}} + 1 - \sqrt{\frac{n}{m}}\right) = \mu^{(1)} - \mu^{(2)},$$

$$\mathrm{COV}(Z_{(i)}, Z_{(j)}) = \begin{cases} \Sigma_1 + \frac{n}{m}\Sigma_2, & \text{当 } i = j \text{ 时} \\ O, & \text{当 } i \neq j \text{ 时} \end{cases} \xlongequal{\text{def}} \Sigma_Z\delta_{ij}.$$

所以 $Z_{(i)} \sim N_p(\mu^{(1)} - \mu^{(2)}, \Sigma_Z)$ $(i=1,\cdots,n)$，且相互独立. 利用前面介绍的单个正态总体均值向量的检验方法进行检验.

(3) 当 Σ_1, Σ_2 相差甚大时，可构造近似检验统计量进行检验(见参考文献[1]).

二、多个正态总体均值向量的检验——多元方差分析

设有 k 个 p 元正态总体 $N_p(\mu^{(t)}, \Sigma)$ $(t=1,\cdots,k)$，样品 $X_{(\alpha)}^{(t)}(t=1,\cdots,k, \alpha=1,\cdots,n_t)$ 是来自 $N^{(p)}(\mu^{(t)}, \Sigma)$ 的随机样本，检验 $H_0: \mu^{(1)} = \cdots = \mu^{(k)}$，$H_1$：至少存在 $i \neq j$ 使得 $\mu^{(i)} \neq \mu^{(j)}$(即 $\mu^{(1)},\cdots,\mu^{(k)}$ 中至少有一对不等).

当 $p=1$ 时，此检验问题就是一元方差分析问题，比如比较 k 个不同品牌的同类产品中某一个质量指标 X（如耐磨度）有无显著差异的问题. 我们把不同品牌对应不同总体（假定为正态总体），这种多组比较问题就是检验

H_0: $\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)}$; H_1: 至少存在 $i\neq j$ 使 $\mu^{(i)}\neq\mu^{(j)}$.

从第 i 个总体抽取容量为 n_i 的随机样本如下（$i=1,\cdots,k$；记 $n=n_1+n_2+\cdots+n_k$）：

$$\begin{array}{llll} X_{(1)}^{(1)}, & X_{(2)}^{(1)}, & \cdots, & X_{(n_1)}^{(1)}, \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ X_{(1)}^{(k)}, & X_{(2)}^{(k)}, & \cdots, & X_{(n_k)}^{(k)}. \end{array}$$

记

$$\overline{X}=\frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{n_t}X_{(j)}^{(t)},\quad \overline{X}^{(t)}=\frac{1}{n_t}\sum_{j=1}^{n_t}X_{(j)}^{(t)}\quad (t=1,\cdots,k).$$

当 $p=1$ 时，利用一元方差分析的思想来构造检验统计量. 记：

总偏差平方和 $\mathrm{SST}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{(j)}^{(i)}-\overline{X})^2$;

组内偏差平方和 $\mathrm{SSE}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{(j)}^{(i)}-\overline{X}^{(i)})^2$;

组间偏差平方和 $\mathrm{SSA}=\sum_{i=1}^{k}n_i(\overline{X}^{(i)}-\overline{X})^2$.

则有平方和分解公式：

$$\mathrm{SST}=\mathrm{SSA}+\mathrm{SSE}.$$

直观考察，若 H_0 成立，当总偏差平方和 SST 固定不变时，应有 SSA 小而 SSE 大，因而比值 SSA/SSE 应很小. 检验统计量取为

$$F=\frac{\mathrm{SSA}/(k-1)}{\mathrm{SSE}/(n-k)}\overset{H_0下}{\sim}F(k-1,n-k).$$

给定显著性水平 α，按传统检验方法，查 F 分布临界值表得 F_α 满足：$P\{F>F_\alpha\}=\alpha$，否定域 $W=\{F>F_\alpha\}$.

推广到 k 个 p 元总体 $N_p(\mu^{(i)},\Sigma)$（假定 k 个总体的协方差阵相等，且记为 Σ），记第 i 个 p 元总体的数据阵为

$$X^{(i)}=\begin{bmatrix} x_{11}^{(i)} & \cdots & x_{1p}^{(i)} \\ \vdots & & \vdots \\ x_{n_i1}^{(i)} & \cdots & x_{n_ip}^{(i)} \end{bmatrix}=\begin{bmatrix} X_{(1)}^{(i)\prime} \\ \vdots \\ X_{(n_i)}^{(i)\prime} \end{bmatrix} \quad (i=1,\cdots,k).$$

对总离差阵 T 进行分解：

$$\begin{aligned} T&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{(j)}^{(i)}-\overline{X})(X_{(j)}^{(i)}-\overline{X})' \\ &=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{(j)}^{(i)}-\overline{X}^{(i)}+\overline{X}^{(i)}-\overline{X})(X_{(j)}^{(i)}-\overline{X}^{(i)}+\overline{X}^{(i)}-\overline{X})' \\ &=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{(j)}^{(i)}-\overline{X}^{(i)})(X_{(j)}^{(i)}-\overline{X}^{(i)})' \\ &\quad+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{X}^{(i)}-\overline{X})(\overline{X}^{(i)}-\overline{X})' \\ &=\sum_{i=1}^{k}A_i+\sum_{i=1}^{k}n_i(\overline{X}^{(i)}-\overline{X})(\overline{X}^{(i)}-\overline{X})' \\ &=A+B, \end{aligned}$$

其中 $A=\sum_{i=1}^{k}A_i$ 称为**组内离差阵**，

$$B=\sum_{i=1}^{k}n_i(\overline{X}^{(i)}-\overline{X})(\overline{X}^{(i)}-\overline{X})'$$

称为**组间离差阵**.

根据直观想法及用似然比原理得到检验 H_0 的统计量为

$$\Lambda=\frac{|A|}{|A+B|}=\frac{|A|}{|T|}.$$

易见：

(1) 因 $A_i\sim W_p(n_i-1,\Sigma)$且相互独立$(i=1,\cdots,k)$，由可加性可得

$$A=\sum_{i=1}^{k}A_i\sim W_p(n-k,\Sigma) \quad (n=n_1+\cdots+n_k).$$

(2) 在 H_0 下，$T\sim W_p(n-1,\Sigma)$.

(3) 还可以证明在 H_0 下，$B\sim W_p(k-1,\Sigma)$，且 B 与 A 相互独立.

根据 Λ 分布的定义，可知

$$\Lambda = \frac{|A|}{|A+B|} \overset{H_0下}{\sim} \Lambda(p, n-k, k-1).$$

给定显著性水平 α，查威尔克斯分布临界值表，可得 λ_α，使

$$P\{\Lambda < \lambda_\alpha\} = \alpha,$$

故否定域 $W=\{\Lambda<\lambda_\alpha\}$. 当手头没有威尔克斯临界值表时，可用 χ^2 分布或 F 分布来近似，即由 Λ 的函数的近似分布进行检验（见参考文献[1]或[2]）.

例 3.3.2 为了研究某种疾病，对一批人同时测量了 4 个指标：β 脂蛋白(X_1)，甘油三酯(X_2)，α 脂蛋白(X_3)，前 β 脂蛋白(X_4). 按不同年龄、不同性别分为三组（20 至 35 岁的女性、20 至 25 岁的男性和 35 至 50 岁的男性），数据见表 3.3. 试问这三个组的 4 项指标间有无显著性差异($\alpha=0.01$)？

表 3.3 身体指标化验数据

X_1	X_2	X_3	X_4	组	X_1	X_2	X_3	X_4	组	X_1	X_2	X_3	X_4	组
260	75	40	18	1	310	122	30	21	2	320	64	39	17	3
200	72	34	17	1	310	60	35	18	2	260	59	37	11	3
240	87	45	18	1	190	40	27	15	2	360	88	28	26	3
170	65	39	17	1	225	65	34	16	2	295	100	36	12	3
270	110	39	24	1	170	65	37	16	2	270	65	32	21	3
205	130	34	23	1	210	82	31	17	2	380	114	36	21	3
190	69	27	15	1	280	67	37	18	2	240	55	42	10	3
200	46	45	15	1	210	38	36	17	2	260	55	34	20	3
250	117	21	20	1	280	65	30	23	2	260	110	29	20	3
200	107	28	20	1	200	76	40	17	2	295	73	33	21	3
225	130	36	11	1	200	76	39	20	2	240	114	38	18	3
210	125	26	17	1	280	94	26	11	2	310	103	32	18	3
170	64	31	14	1	190	60	33	17	2	330	112	21	11	3
270	76	33	13	1	295	55	30	16	2	345	127	24	20	3
190	60	34	16	1	270	125	24	21	2	250	62	22	16	3
280	81	20	18	1	280	120	32	18	2	260	59	21	19	3
310	119	25	15	1	240	62	32	20	2	225	100	34	30	3
270	57	31	8	1	280	69	29	20	2	345	120	36	18	3
250	67	31	14	1	370	70	30	20	2	360	107	25	23	3
260	135	39	29	1	280	40	37	17	2	250	117	36	16	3

解　比较三个组($k=3$)的 4 项指标($p=4$)间是否有差异问题，就是多总体均值向量是否相等的检验问题. 设第 i 组为 4 元总体 $N_4(\mu^{(i)},\Sigma)$ $(i=1,2,3)$，来自 3 个总体的样本容量 $n_1=n_2=n_3=20$. 检验：

H_0: $\mu^{(1)}=\mu^{(2)}=\mu^{(3)}$,　H_1: $\mu^{(1)},\mu^{(2)},\mu^{(3)}$ 至少有一对不相等.

因似然比统计量 $\Lambda\sim\Lambda(p,n-k,k-1)$，在此例中 $k-1=2$，可以利用 Λ 统计量与 F 统计量的关系，取检验统计量为 F 统计量：

$$F=\frac{(n-k)-p+1}{p}\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}\quad(k=3,p=4,n=60),$$

由样本值计算得：$\overline{X}=(259.08,84.12,32.37,17.8)'$，以及

$$\overline{X}^{(1)}=\begin{bmatrix}231.0\\89.6\\32.9\\17.1\end{bmatrix},\quad \overline{X}^{(2)}=\begin{bmatrix}253.50\\72.55\\32.45\\17.90\end{bmatrix},\quad \overline{X}^{(3)}=\begin{bmatrix}292.75\\90.20\\31.75\\18.40\end{bmatrix},$$

$$A=A_1+A_2+A_3=\sum_{t=1}^{3}\sum_{a=1}^{n_t}(X_{(a)}^{(t)}-\overline{X}^{(t)})(X_{(a)}^{(t)}-\overline{X}^{(t)})'$$

$$=\begin{bmatrix}125408.75\\23278.50 & 40466.95\\-3950.75 & -1937.75 & 2082.50\\1748.00 & 2166.30 & -26.90 & 1024.40\end{bmatrix},$$

$$T=\sum_{t=1}^{3}\sum_{a=1}^{n_t}(X_{(a)}^{(t)}-\overline{X})(X_{(a)}^{(t)}-\overline{X})'$$

$$=\begin{bmatrix}164474.580\\25586.417 & 44484.183\\-4674.833 & -1973.567 & 2095.933\\2534.000 & 2139.400 & -41.600 & 1041.600\end{bmatrix}.$$

进一步计算可得

$$\Lambda=\frac{|A|}{|T|}=\frac{7.8419\times10^{15}}{1.1844\times10^{16}}=0.6621,$$

$$f=\frac{n-k-p+1}{p}\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}}=\frac{54}{4}\frac{1-\sqrt{0.6621}}{\sqrt{0.6621}}=3.0907.$$

对给定 $\alpha=0.01$，利用统计软件(如 SAS 系统)，首先计算 p 值(此时检验统计量 $F\sim F(8,108)$)：

$$p = P\{F \geqslant 3.09007\} = 0.003538.$$

因 $p=0.003538<0.01=\alpha$，故否定 H_0，这表明三个组的指标之间有显著的差异. 在这种情况下，可能犯第一类错误，且犯第一类错误的概率为 0.01.

进一步地若还想了解三个组指标间的差异究竟是由哪几项指标引起的，可以对 4 项指标逐项用一元方差分析方法进行检验，我们将发现三个组指标间只有第一项指标 X_1 有显著差异.

事实上，用一元方差分析检验第一项指标 X_1 在三个组中是否有显著差异时，因

$$f_1=\frac{(t_{11}-a_{11})/(k-1)}{a_{11}/(n-k)}=\frac{(164474.58-125408.75)/2}{125408.75/57}$$

$$=8.8780,$$

其中 t_{11} 和 a_{11} 分别是 T 和 A 中的第一个对角元素，有

$p_1=P\{F_1\geqslant 8.8780\}=0.0004401$ (检验统计量 $F_1\sim F(2,57)$)，

因 $p_1=0.0004401$ 显著地小于 0.01，故第一项指标 X_1 在三个组中有显著差异.

§3.4 协方差阵的检验

一、单个 p 元正态总体协方差阵的检验

设 $X_{(\alpha)}(\alpha=1,\cdots,n)$ 为来自 p 元正态总体 $N_p(\mu,\Sigma)$ ($\Sigma>0$ 未知)的随机样本，检验

$$H_0:\ \Sigma=\Sigma_0(\Sigma_0>0\text{ 为已知矩阵}),\quad H_1:\ \Sigma\neq\Sigma_0.$$

1. 当 $\Sigma_0=I_p$ 时检验 $H_0:\ \Sigma=I_p$，$H_1:\ \Sigma\neq I_p$

利用似然比原理来导出似然比统计量 λ_1：

$$\lambda_1=\max_{\mu}L(\mu,I_p)\Big/\max_{\mu,\Sigma>0}L(\mu,\Sigma).$$

当 $\Sigma=I_p$ 成立时，似然函数在 $\mu=\bar{X}$ 达最大值，因此

λ_1 表示式的分子$= L(\overline{X}, I_p)$

$$= (2\pi)^{-np/2}|I_p|^{-n/2}\exp\left[-\frac{1}{2}\mathrm{tr}(I_p^{-1}A)\right],$$

λ_1 表示式的分母$= L\left(\overline{X}, \frac{1}{n}A\right) = (2\pi)^{-np/2}\left|\frac{1}{n}A\right|^{-n/2}\mathrm{e}^{-np/2}$

$$= (2\pi)^{-np/2}\left(\frac{\mathrm{e}}{n}\right)^{-np/2}|A|^{-n/2}.$$

所以似然比统计量

$$\lambda_1 = \exp\left\{-\frac{1}{2}\mathrm{tr}(A)\right\}|A|^{n/2}\left(\frac{\mathrm{e}}{n}\right)^{np/2},$$

其中
$$A = \sum_{\alpha=1}^{n}(X_{(\alpha)} - \overline{X})(X_{(\alpha)} - \overline{X})'.$$

利用定理 3.2.1 可知，当 n 很大且 H_0 成立时，$\xi = -2\ln\lambda_1$ 的近似分布为 $\chi^2\left(\frac{p(p+1)}{2}\right)$，利用检验统计量 ξ 来构造检验方法.

2. 当 $\Sigma_0 \neq I_p$ 时检验 $H_0: \Sigma = \Sigma_0$，$H_1: \Sigma \neq \Sigma_0$

因 $\Sigma_0 > 0$，存在非退化矩阵 $D_{p\times p}$，使 $D\Sigma_0 D' = I_p$. 令

$$Y_{(\alpha)} = DX_{(\alpha)} \quad (\alpha = 1, \cdots, n),$$

则

$$Y_{(\alpha)} \sim N_p(D\mu, D\Sigma D') \xlongequal{\text{def}} N_p(\mu^*, \Sigma^*).$$

检验

$$H_0: \Sigma = \Sigma_0 \iff H_0: \Sigma^* = I_p.$$

从新样本 $Y_{(\alpha)}(\alpha = 1, \cdots, n)$ 出发，检验 $H_0: \Sigma^* = I_p$ 的似然比统计量取为(以下记 $\exp(\mathrm{tr}A) \xlongequal{\text{def}} \mathrm{etr}(A)$)：

$$\lambda_2 = \exp\left\{-\frac{1}{2}\mathrm{tr}(A^*)\right\}|A^*|^{n/2}\left(\frac{\mathrm{e}}{n}\right)^{np/2}$$

$$= \mathrm{etr}\left(-\frac{1}{2}A^*\right)|A^*|^{n/2}\left(\frac{\mathrm{e}}{n}\right)^{np/2},$$

其中
$$A^* = \sum_{\alpha=1}^{n}(Y_{(\alpha)} - \overline{Y})(Y_{(\alpha)} - \overline{Y})' = DAD'.$$

若注意到 $D\Sigma_0 D' = I_p$，则似然比统计量 λ_2 还可以表示为

$$\lambda_2 = \mathrm{etr}\left(-\frac{1}{2}A\Sigma_0^{-1}\right)|A\Sigma_0^{-1}|^{n/2}\left(\frac{\mathrm{e}}{n}\right)^{np/2}.$$

研究似然比统计量 λ_2 的抽样分布是很困难的，通常根据定理3.2.1由 λ_2 的近似分布来构造检验法.

当样本容量 n 很大，在 H_0 成立时，$-2\ln\lambda_2$ 的极限分布为

$$\chi^2\left(\frac{p(p+1)}{2}\right).$$

除此以外，在不同适用范围下还有其他近似分布可用来构造检验法(见参考文献[1]或[2]).

3. 检验 H_0: $\Sigma=\sigma^2\Sigma_0$(σ^2 未知)

当 $\Sigma_0=I_p$ 时此检验常称为**球性检验**. 以下利用似然比原理来导出似然比统计量 λ_3:

$$\lambda_3 = \max_{\mu,\sigma^2>0}L(\mu,\sigma^2\Sigma_0)\Big/\max_{\mu,\Sigma>0}L(\mu,\Sigma).$$

当 σ^2 给定时，似然函数 $L(\mu,\sigma^2\Sigma_0)$ 在 $\mu=\bar{X}$ 达最大值，且

$$L(\bar{X},\sigma^2\Sigma_0) = (2\pi)^{-np/2}|\sigma^2\Sigma_0|^{-n/2}\exp\left[-\frac{1}{2}\mathrm{tr}((\sigma^2\Sigma_0)^{-1}A)\right]$$

$$= (2\pi)^{-np/2}(\sigma^2)^{-np/2}|\Sigma_0|^{-n/2}\mathrm{etr}\left(-\frac{1}{2\sigma^2}(\Sigma_0^{-1}A)\right).$$

令

$$\frac{\partial L(\bar{X},\sigma^2\Sigma_0)}{\partial\sigma^2} = (2\pi)^{-np/2}(\sigma^2)^{-np/2-2}|\Sigma_0|^{-n/2}\mathrm{etr}\left(-\frac{1}{2\sigma^2}(\Sigma_0^{-1}A)\right)$$

$$\cdot\left[-\frac{np}{2}\sigma^2+\frac{1}{2}\mathrm{tr}(\Sigma_0^{-1}A)\right]=0,$$

可得出 $\hat{\sigma}^2=\frac{1}{np}\mathrm{tr}(\Sigma_0^{-1}A)$. 从而有

$$\lambda_3\text{ 表示式的分子} = (2\pi)^{-np/2}\left(\frac{1}{np}\mathrm{tr}(\Sigma_0^{-1}A)\right)^{-np/2}|\Sigma_0|^{-n/2}\mathrm{e}^{-np/2},$$

$$\lambda_3\text{ 表示式的分母} = L\left(\bar{X},\frac{1}{n}A\right) = (2\pi)^{-np/2}\left(\frac{\mathrm{e}}{n}\right)^{-np/2}|A|^{-n/2}.$$

所以似然比统计量

$$\lambda_3 = \frac{|\Sigma_0^{-1}A|^{n/2}}{[\mathrm{tr}(\Sigma_0^{-1}A)/p]^{np/2}},$$

或等价于

$$W=(\lambda_3)^{2/n}=\frac{p^p|\Sigma_0^{-1}A|}{[\mathrm{tr}(\Sigma_0^{-1}A)]^p}.$$

当样本容量 n 很大，在 H_0 为真时有以下近似分布：

$$-\left((n-1)-\frac{2p^2+p+2}{6p}\right)\ln W \text{ 近似为 } \chi^2\left(\frac{p(p+1)}{2}-1\right).$$

二、多总体协方差阵的检验

设有 k 个总体 $N_p(\mu^{(t)},\Sigma_t)$ $(t=1,\cdots,k)$，$X_{(\alpha)}^{(t)}(t=1,\cdots,k;\ \alpha=1,\cdots,n_t)$为来自第 t 个总体 $N_p(\mu^{(t)},\Sigma_t)$的随机样本，记 $n=\sum_{i=1}^{k}n_i$. 检验 H_0：$\Sigma_1=\Sigma_2=\cdots=\Sigma_k\overset{\text{def}}{=\!=}\Sigma$，$H_1$：$\Sigma_1,\Sigma_2,\cdots,\Sigma_k$ 不全相等.

样本$\{X_{(\alpha)}^{(t)}\}$的似然函数为

$$L(\mu^{(1)},\Sigma_1,\cdots,\mu^{(k)},\Sigma_k)=\prod_{t=1}^{k}L_t(\mu^{(t)},\Sigma_t),$$

似然比统计量 λ_4 为

$$\lambda_4=\max_{\mu^{(i)},\Sigma>0}L(\mu^{(1)},\cdots,\mu^{(k)},\Sigma)\Big/\max_{\mu^{(i)},\Sigma_i>0}L(\mu^{(1)},\Sigma_1,\cdots,\mu^{(k)},\Sigma_k).$$

从而有

$$\begin{aligned}
\text{上式的分母}&=\max_{\mu^{(i)},\Sigma_i>0}L(\mu^{(1)},\Sigma_1,\cdots,\mu^{(k)},\Sigma_k)\\
&=\prod_{t=1}^{k}L_t\left(\overline{X}^{(t)},\frac{1}{n_t}A_t\right)\\
&=\prod_{t=1}^{k}(2\pi)^{-n_tp/2}\left|\frac{A_t}{n_t}\right|^{-n_t/2}\cdot\mathrm{e}^{-n_tp/2}\\
&=(2\pi)^{-np/2}\mathrm{e}^{-np/2}\prod_{t=1}^{k}\left|\frac{A_t}{n_t}\right|^{-n_t/2},\\
\text{上式的分子}&=\max_{\mu^{(i)},\Sigma>0}L(\mu^{(1)},\cdots,\mu^{(k)},\Sigma)\\
&=(2\pi)^{-np/2}\mathrm{e}^{-np/2}\left|\frac{A}{n}\right|^{-n/2}\quad(\text{其中 }A=A_1+\cdots+A_k),
\end{aligned}$$

则似然比统计量 λ_4 为

$$\lambda_4 = \left|\frac{A}{n}\right|^{-n/2} \Big/ \prod_{i=1}^{k}\left|\frac{A_i}{n_i}\right|^{-n_i/2}.$$

根据无偏性的要求进行修正，将 λ_4 中的 n_i 用 n_i-1 替代，n 用 $n-k$ 替代. 然后对 λ_4 取对数，可得到统计量：

$$M = -2\ln\lambda_4^* = (n-k)\ln\left|\frac{A}{n-k}\right| - \sum_{i=1}^{k}(n_i-1)\ln\left|\frac{A_i}{n_i-1}\right|.$$

当样本容量 n 很大时，在 H_0 为真时 M 有以下近似分布：

$$(1-d)M = -2(1-d)\ln\lambda_4^* \sim \chi^2(f),$$

其中

$$f = \frac{1}{2}p(p+1)(k-1),$$

$$d = \begin{cases} \dfrac{2p^2+3p-1}{6(p+1)(k-1)}\left[\displaystyle\sum_{i=1}^{k}\frac{1}{n_i-1} - \frac{1}{n-k}\right], & \text{当 } n_i \text{ 不全等,} \\ \dfrac{(2p^2+3p-1)(k+1)}{6(p+1)(n-k)}, & \text{当 } n_i \text{ 全相等.} \end{cases}$$

例 3.4.1 对例 3.3.2 表 3.3 中给出的身体指标化验数据，试判断三个组(即三个总体)的协方差阵是否相等($\alpha=0.10$)?

解 这是三个 4 元正态总体的协方差阵是否相等的检验问题. 设第 i 组为 4 维总体 $N_4(\mu^{(i)},\Sigma_i)$ $(i=1,2,3)$. 来自三个总体的样本容量 $n_1=n_2=n_3=20$. 检验

$$H_0: \Sigma_1=\Sigma_2=\Sigma_3, \quad H_1: \Sigma_1,\Sigma_2,\Sigma_3 \text{ 至少有一对不相等.}$$

在 H_0 成立时，取近似检验统计量为 $\chi^2(f)$ 统计量：

$$\xi = (1-d)M = -2(1-d)\ln\lambda_4^*.$$

由样本值计算三个总体的样本协方差阵：

$$S_1 = \frac{1}{n_1-1}A_1 = \frac{1}{n_1-1}\sum_{\alpha=1}^{n_1}(X_{(\alpha)}^{(1)}-\overline{X}^{(1)})(X_{(\alpha)}^{(1)}-\overline{X}^{(1)})'$$

$$= \frac{1}{19}\begin{bmatrix} 30530 & & & \\ 6298 & 15736.8 & & \\ -1078 & -796.8 & 955.8 & \\ 198 & 1387.8 & 90.2 & 413.8 \end{bmatrix},$$

$$S_2=\frac{1}{n_2-1}A_2=\frac{1}{n_2-1}\sum_{\alpha=1}^{n_2}(X_{(\alpha)}^{(2)}-\overline{X}^{(2)})(X_{(\alpha)}^{(2)}-\overline{X}^{(2)})'$$

$$=\frac{1}{19}\begin{bmatrix}51705.0 & & & \\ 7021.5 & 12288.95 & & \\ -1571.5 & -807.95 & 364.95 & \\ 827.0 & 321.10 & -5.10 & 133.8\end{bmatrix},$$

$$S_3=\frac{1}{n_3-1}A_3=\frac{1}{n_3-1}\sum_{\alpha=1}^{n_3}(X_{(\alpha)}^{(3)}-\overline{X}^{(3)})(X_{(\alpha)}^{(2)}-\overline{X}^{(3)})'$$

$$=\frac{1}{19}\begin{bmatrix}43173.75 & & & \\ 9959.00 & 12441.2 & & \\ -1301.25 & -333.0 & 761.75 & \\ 723.00 & 457.4 & -112.00 & 476.8\end{bmatrix}.$$

进一步计算可得

$$|S|=\left|\frac{1}{57}A\right|=742890016,\quad |S_1|=791325317,$$

$$|S_2|=145821806,\quad |S_3|=1.08116\times 10^9,$$

$$M=22.6054,\quad d=0.1006,\quad f=20,$$

$$\xi=(1-d)M=20.3316.$$

对给定 $\alpha=0.10$，利用统计软件(如 SAS 系统)，首先计算 p 值(此时检验统计量 $\xi\sim\chi^2(20)$)：

$$p=P\{\xi\geqslant 20.3316\}=0.4374.$$

因 $p=0.4374>0.10=\alpha$，故 H_0 相容，这表明三个组的协方差阵之间没有显著的差异.

三、多个正态总体的均值向量和协方差阵同时检验

设有 k 个总体 $N_p(\mu^{(t)},\Sigma_t)$ $(t=1,\cdots,k)$，$X_{(\alpha)}^{(t)}(t=1,\cdots,k;\alpha=1,\cdots,n_t)$ 为来自第 t 个总体 $N_p(\mu^{(t)},\Sigma_t)$ 的随机样本. 检验：

$$H_0:\ \mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)},\quad \text{且 } \Sigma_1=\Sigma_2=\cdots=\Sigma_k,$$

H_1：$\mu^{(i)}(i=1,\cdots,k)$ 或 $\Sigma_i(i=1,\cdots,k)$ 至少有一对不相等.

记

$$\bar{X}^{(t)}=\frac{1}{n_t}\sum_{j=1}^{n_t}X_{(j)}^{(t)},\quad \bar{X}=\frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{n_t}X_{(j)}^{(t)},\quad n=\sum_{t=1}^{k}n_t,$$

$$A_t=\sum_{j=1}^{n_t}(X_{(j)}^{(t)}-\bar{X}^{(t)})(X_{(j)}^{(t)}-\bar{X}^{(t)})',\quad A=\sum_{t=1}^{k}A_t,$$

$$T=\sum_{t=1}^{k}\sum_{j=1}^{n_t}(X_{(j)}^{(t)}-\bar{X})(X_{(j)}^{(t)}-\bar{X})'$$

$$=A+\sum_{t=1}^{k}n_t(\bar{X}^{(t)}-\bar{X})(\bar{X}^{(t)}-\bar{X})'.$$

则检验以上假设 H_0 **的似然比统计量为**

$$\lambda_5=\frac{\prod_{t=1}^{k}|A_t|^{n_t/2}}{|T|^{n/2}}\cdot\frac{n^{np/2}}{\prod_{t=1}^{k}n_t^{n_tp/2}}.$$

若用 Λ 表示当协方差阵均相同时检验 k 个总体均值向量是否相等的似然比统计量，将发现这里的似然比统计量 $\lambda_5=\Lambda\cdot\lambda_4$. 在实际应用中我们采用类似的修正方法，在 λ_5 中用 n_t-1 替代 n_t，用 $n-k$ 替代 n. 修正后的统计量记为 λ_5^*：

$$\lambda_5^*=\frac{\prod_{t=1}^{k}|A_t|^{\frac{n_t-1}{2}}}{|T|^{\frac{n-k}{2}}}\cdot\frac{(n-k)^{\frac{(n-k)p}{2}}}{\prod_{t=1}^{k}(n_t-1)^{\frac{(n_t-1)p}{2}}}.$$

当样本容量 n 很大，在 H_0 为真时 λ_5^* 有以下近似分布：

$$-2(1-b)\ln\lambda_5^*\sim\chi^2(f),$$

其中

$$f=\frac{1}{2}p(p+3)(k-1),$$

$$b=\left(\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right)\left(\frac{2p^2+3p-1}{6(p+3)(k-1)}\right)$$

$$-\frac{p-k+2}{(n-k)(p+3)}.$$

例 3.4.2 对例 3.3.2 表 3.3 中给出的身体指标化验数据，试判断三个组(即三个总体)的均值向量和协方差阵是否全都相等($\alpha=$

0.05)?

解　这是三个 4 元正态总体的均值向量和协方差阵是否同时相等的检验问题. 取近似检验统计量为近似 χ^2 统计量:

$$\xi = -2(1-b)\ln\lambda_5^* \sim \chi^2(f).$$

由样本值计算这三个总体的样本协方差阵(见例 3.4.1),以及所有样本的总离差阵 T(见例 3.3.2). 进一步计算可得

$$\left|\frac{1}{n-k}T\right| = \left|\frac{1}{57}T\right| = 1.12198 \times 10^9, \quad |S_1| = 791325317,$$

$$|S_2| = 145821806, \quad |S_3| = 1.08116 \times 10^9,$$

$$M_5 = -2\ln\lambda_5^* = 46.1067, \quad b = 0.06433, \quad f = 28,$$

$$\xi = (1-b)M_5 = 43.1408.$$

对给定 $\alpha=0.05$,利用统计软件(如 SAS 系统),首先计算 p 值(此时检验统计量 $\xi \sim \chi^2(28)$):

$$p = P\{\xi \geqslant 43.1408\} = 0.03373.$$

因 $p=0.03373<0.05=\alpha$,故否定 H_0,这表明三个组的均值向量和协方差阵之间有显著的差异. 在这种情况下,可能犯第一类错误,且犯第一类错误的概率为 0.05.

§3.5　独立性检验

设总体 $X \sim N_p(\mu,\Sigma)$,将 X 剖分为 k 个子向量,而 μ 和 Σ 也相应剖分为

$$X = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(k)} \end{bmatrix} \begin{matrix} p_1 \\ \vdots \\ p_k \end{matrix}, \quad \mu = \begin{bmatrix} \mu^{(1)} \\ \vdots \\ \mu^{(k)} \end{bmatrix} \begin{matrix} p_1 \\ \vdots \\ p_k \end{matrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} \\ \vdots & & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} \end{bmatrix} \begin{matrix} p_1 \\ \vdots \\ p_k \end{matrix},$$

其中 $p=p_1+\cdots+p_k$,且知 p_t 维子向量 $X^{(t)} \sim N_{p_t}(\mu^{(t)},\Sigma_{tt})$ $(t=1,\cdots,k)$. 若 k 个随机子向量相互独立,则可把 p 维(高维)随机向量的问题化为 k 个低维随机向量的问题来处理,这在处理多元统计分析的许多问题中将带来极大的方便.

在第二章中,我们已介绍过若 $X^{(1)},\cdots,X^{(k)}$ 相互独立 $\Longleftrightarrow \Sigma_{ij}=O$

(对一切 $i\neq j$). 因此检验 $X^{(1)},\cdots,X^{(k)}$ 是否相互独立的问题等价于检验对任意两个子向量,协方差阵 Σ_{ij} 是否等于 O(对一切 $i\neq j$).

在正态总体下,独立性检验可化为检验:

H_0: $\Sigma_{ij}=O$(一切 $i\neq j$), H_1: $\Sigma_{ij}\neq O$,至少有一对 $i\neq j$.

设 $X_{(\alpha)}$($\alpha=1,\cdots,n,n>p$)为来自总体 X 的随机样本. 将 $X_{(\alpha)}$, 样本均值向量 $\overline{X}$ 和样本离差阵

$$A=\sum_{j=1}^{n}(X_{(j)}-\overline{X})(X_{(j)}-\overline{X})'$$

作相应剖分为

$$X_{(\alpha)}=\begin{bmatrix}X_{(\alpha)}^{(1)}\\ \vdots\\ X_{(\alpha)}^{(k)}\end{bmatrix}\begin{matrix}p_1\\ \vdots\\ p_k\end{matrix},\quad \overline{X}=\begin{bmatrix}\overline{X}^{(1)}\\ \vdots\\ \overline{X}^{(k)}\end{bmatrix}\begin{matrix}p_1\\ \vdots\\ p_k\end{matrix},$$

$$A=\begin{bmatrix}A_{11} & \cdots & A_{1k}\\ \vdots & & \vdots\\ A_{k1} & \cdots & A_{kk}\end{bmatrix}\begin{matrix}p_1\\ \vdots\\ p_k\end{matrix}.$$

应用似然比原理,在 H_0 成立时, $X_{(\alpha)}^{(i)}\sim N_{p_i}(\mu^{(i)},\Sigma_{ii})$ ($i=1,\cdots,k$; $\alpha=1,\cdots,n$),且相互独立,故样本的似然函数为

$$L(\mu,\Sigma)=\prod_{i=1}^{k}L_i(\mu^{(i)},\Sigma_{ii}).$$

当 $\hat{\mu}^{(i)}=\overline{X}^{(i)}$, $\hat{\Sigma}_{ii}=\frac{1}{n}A_{ii}$时, $L_i(\mu^{(i)},\Sigma_{ii})$达最大. 所以似然比统计量表示式的分子为

$$\max_{\mu,\Sigma_{ij}=O}L(\mu,\Sigma)=\prod_{i=1}^{k}(2\pi)^{-np_i/2}\left|\frac{A_{ii}}{n}\right|^{-n/2}\mathrm{e}^{-np_i/2}$$

$$=(2\pi)^{-np/2}\cdot\mathrm{e}^{-np/2}\prod_{i=1}^{k}\left|\frac{A_{ii}}{n}\right|^{-n/2}.$$

似然比统计量为

$$\lambda=\prod_{i=1}^{k}\left|\frac{A_{ii}}{n}\right|^{-n/2}\Big/\left|\frac{1}{n}A\right|^{-n/2}=\left(\frac{|A|}{\prod_{i=1}^{k}|A_{ii}|}\right)^{n/2}\overset{\text{def}}{=\!=}V^{n/2},$$

其

$$\ln\lambda=\frac{n}{2}\ln V.$$

博克斯(Box)证明了，在 H_0 成立下当 $n\to\infty$ 时，

$$-b\ln V \sim \chi^2(f),$$

其中

$$b = n - \frac{3}{2} - \frac{p^3 - \sum_1^k p_a^3}{3\left(p^2 - \sum_1^k p_a^2\right)},$$

$$f = \frac{1}{2}\left[p(p+1) - \sum_{a=1}^{k} p_a(p_a+1)\right].$$

例 3.5.1　试检验例 3.2.1 女性汗液数据中随机向量 X 的三个分量是否相互独立($\alpha=0.05$).

解　记随机向量 $X=(X_1,X_2,X_3)'$，假定 $X\sim N_3(\mu,\Sigma)$，且记 $\Sigma=(\sigma_{ij})_{3\times 3}$. 检验

$$H_0:\ \sigma_{12}=0,\ \sigma_{13}=0,\ \sigma_{23}=0,\quad H_1:\ \sigma_{12},\sigma_{13},\sigma_{23}\ \text{不全为 0}.$$

取检验统计量为

$$\xi = -b\ln\left(\frac{|A|}{\prod_{i=1}^{k}|A_{ii}|}\right).$$

当 X 的三个分量相互独立，且样本容量 n 很大时，ξ 近似于 $\chi^2(f)$. 由表 3.1 的样本值计算样本离差阵 A，可得：

$$A = \begin{bmatrix} 54.708 & & \\ 190.190 & 3795.98 & \\ -34.372 & -107.16 & 68.926 \end{bmatrix}.$$

此例中 $n=20, p=3, p_1=p_2=p_3=1, k=3$. 进一步计算可得：

$$b = 17.166667,\quad f = 3,$$

$$V = \frac{|A|}{\prod_{i=1}^{k}|A_{ii}|} = \frac{8108729}{54.708\times 3795.98\times 68.926}$$

$$= \frac{8108729}{14313791} = 0.5665,$$

$$\xi = -b\ln V = -17.1667\times\ln(0.5665) = 9.7555.$$

对给定显著性水平 $\alpha=0.05$，用统计软件 SAS 系统计算时，通过计算

p 值进行检验：

$$p = P\{\xi \geqslant 9.7555\} = 0.02076.$$

因 $p=0.02076<0.05=\alpha$，故否定 H_0，即随机向量的三个分量不相互独立. 在这种情况下，可能犯第一类错误，且犯第一类错误的概率为 0.05.

§3.6 正态性检验

在均值向量和协方差阵的检验中，以及以后将介绍的一些统计方法中都是假定样本来自 p 元正态总体. 所作统计推断的结论是否正确，在某种意义上取决于实际总体与正态总体接近的程度如何？因此建立一些方法来检验多元观测数据与多元正态数据的差异是否显著是十分必要的.

设 $X_{(\alpha)}=(X_{\alpha 1},\cdots,X_{\alpha p})'$ $(\alpha=1,\cdots,n)$ 是来自 p 元总体 X 的随机样本，试问总体 X 是否服从 $N_p(\mu,\Sigma)$ 分布？

若总体 $X=(X_1,\cdots,X_p)'\sim N_p(\mu,\Sigma)$，利用多元正态分布的一些性质可知以下结论(记 $\mu=(\mu_1,\cdots,\mu_p)'$，$\Sigma=(\sigma_{ij})_{p\times p}$)：

结论 1 每个分量 $X_i\sim N(\mu_i,\sigma_{ii})$ $(i=1,\cdots,p)$；

结论 2 任意两个分量 $(X_i,X_j)\sim$ 二元正态分布；

结论 3 设 $l=(l_1,\cdots,l_p)'$ 为任给的 p 维常向量，令 $\xi=l'X$，则

$$\xi \sim N_1(l'\mu,l'\Sigma l);$$

结论 4 令 $\eta=(X-\mu)'\Sigma^{-1}(X-\mu)$，则 $\eta\sim\chi^2(p)$；

结论 5 正态随机向量 X 的概率密度等高线为椭球.

若总体 X 为多元正态总体，必具有以上所列的几条性质. 如果 X 具有以上这些性质，也不一定能得出 X 为 p 元正态分布. 但如果经过检验，比如发现某个分量 X_i 与正态分布有显著差异，即可得出 p 元总体 X 与 p 元正态分布也有显著差异. 利用以上性质，要来构造出好的满意的多元正态的整体性检验是十分困难的. 在实际应用中如果经过从多方面得到的检验结果与正态分布均无显著性差异，也就认为该总体 X 与 p 元正态无显著差异.

关于多元数据的正态性检验问题，常转化为多个一元或二元数据的正态性检验，而一元数据的正态性检验已有一些方法；或者先求 X 分量的线性组合，化为一元数据的正态性检验等. 这些方法虽不是严格的，但一元或二元数据是正态的，而多元非正态的病态数据在实际应用中并不常见.

一、一维边缘分布的正态性检验

设 p 维随机向量 $X=(X_1,\cdots,X_p)'$，检验分量 $X_i\sim N(\mu_i,\sigma_i^2)$ $(i=1,\cdots,p)$. 若要把 p 元正态性检验化为 p 个一元数据的正态性检验，常用的检验方法有以下几种(见参考文献[4]，[7]).

(1) χ^2 检验法：这是适用于连续型或离散型随机变量分布的拟合优度检验方法，也称为**皮尔逊**(Pearson) χ^2 **检验法**.

(2) 科尔莫戈罗夫(Kolmogorov)检验法：这是适用于连续型分布的拟合优度检验方法，当然也适用于正态性检验.

以下几种方法是仅适用于正态分布的检验法.

(3) 偏峰检验法.

(4) W(Wilks)检验和 D 检验.

(5) Q-Q(Quantile-Quantile)图检验法.

(6) P-P(Probability-Probability)图检验法：这是与 Q-Q 图检验法类似的图示检验法. Q-Q 图检验法绘制散点 $(q_i,x_{(i)}^*)$ 的散布图，其中 $q_i=\Phi^{-1}(p_i)$ 为正态总体的 p_i 分位数，

$$p_i=\frac{i-0.5}{n}\quad(i=1,\cdots,n);$$

$x_{(i)}^*$ 为样本的 p_i 分位数. 如果总体 X 为一元正态总体，这些散点应散布在一条直线上.

另外，我们还可以绘制另一对数据点：$(p_i,F(x_{(i)}^*))$ $(i=1,\cdots,n)$ 的散布图，记 $F_n(\cdot)$ 为经验分布函数. 因为

$$p_i=F_n(x_{(i)}^*)\approx F(x_{(i)}^*)=\Phi\left(\frac{x_{(i)}^*-\mu}{\sigma}\right),$$

故这些散点也应散布在一直线上. 这里 $F(x_{(i)}^*)$ 是正态总体 X($X\sim$

$N(\mu,\sigma^2)$)的分布函数在点 $x^*_{(i)}$ 上的值，即随机事件$\{X\leqslant x^*_{(i)}\}$的概率(Probability)值，而 p_i 是由经验分布函数 $F_n(x)$ 得到的样本分布函数在点 $x^*_{(i)}$ 上的值，故称此散布图为 P-P 图，利用此图得到的检验法称为 P-P 图检验法.

(7) "3σ"原则检验法：如果总体 $X\sim N(\mu,\sigma^2)$，根据"3σ"原则，可知

$$p_k=P\{\mu-k\sigma<X<\mu+k\sigma\}=\begin{cases}0.683, & \text{当 } k=1 \text{ 时},\\ 0.954, & \text{当 } k=2 \text{ 时},\\ 0.997, & \text{当 } k=3 \text{ 时}.\end{cases}$$

若 $X_{(1)},X_{(2)},\cdots,X_{(n)}$ 是来自总体 X 的样本(假设样本容量 n 很大)，又 X 为正态分布，则样品点落入区域$(\mu-k\sigma,\mu+k\sigma)$的比例 $\hat{p}_k$ 与以上列出的概率 p_k 应是相差不多的. 利用大数定律可知，$\hat{p}_k$ 近似为正态分布，由此得出的检验法如下：

由样本值 $x_{(i)}(i=1,\cdots,n)$ 首先计算样本均值 $\bar{x}$ 和样本标准差 s，然后统计落入区域$(\bar{x}-s,\bar{x}+s)$或$(\bar{x}-2s,\bar{x}+2s)$的样品点个数，并计算占样品总数 n 的比例 $\hat{p}_1$ 或 $\hat{p}_2$，如果

$$|\hat{p}_1-0.683|>3\sqrt{\frac{0.683\times 0.317}{n}}=\frac{1.396}{\sqrt{n}},$$

或

$$|\hat{p}_2-0.954|>3\sqrt{\frac{0.954\times 0.046}{n}}=\frac{0.628}{\sqrt{n}},$$

则认为总体 X 与正态总体有偏离，当 $\hat{p}_1$ 或 $\hat{p}_2$ 都较小时，表示总体 X 的分布比正态分布有较重的尾部.

(8) A^2 和 W^2 统计量检验法：这是由 SAS 软件系统提供的检验法，使用的方法主要是经验分布拟合优度检验法，由样本计算得到的经验分布函数 $F_n(x)$ 可作为总体分布函数 $F(x)$ 的估计，所以考虑用 $F_n(x)$ 与原假设指定的分布函数 $F_0(x)$ 间的差异来检验原假设. 以下三种统计量都是用以度量指定分布函数 $F_0(x)$ 与经验分布函数 $F_n(x)$ 这两个函数之间的差异：

柯尔莫戈罗夫-斯米尔诺夫(Smirnov)统计量：

$$D = \sup_x |F_n(x) - F_0(x)|.$$

Anderson-Darling 统计量：

$$A^2 = n\int_{-\infty}^{+\infty}(F_n(x) - F_0(x))^2[F_0(x)(1 - F_0(x))]^{-1}\mathrm{d}F_0(x).$$

Cramer-von Mises 统计量：

$$W^2 = n\int_{-\infty}^{+\infty}(F_n(x) - F_0(x))^2\mathrm{d}F_0(x).$$

当原假设成立时，上面三个统计量应取较小的值. 这三个统计量取很大数值时是极端情况，故度量这三个统计量取极端情况的相应的 p 值若小于给定的显著性水平 α，则有足够证据否定正态性假定（见参考文献[3]，35～39).

二、二元数据的正态性检验

设 $X=(X_1,X_2,\cdots,X_p)'$ 为 p 维随机向量，X 的任意两个分量的 n 次观测数据记为 $X_{(i)}=(X_{i1},X_{i2})'$ $(i=1,\cdots,n)$. 下面介绍检验二元观测数据是否来自二元正态分布的方法.

1. 等概椭圆检验法

若二维随机向量 $X=(X_1,X_2)'\sim N_2(\mu,\Sigma)$，则 X 的概率密度函数等高线

$$f(x_1,x_2) = a \iff (X-\mu)'\Sigma^{-1}(X-\mu) = b^2,$$

上式右边是中心在(μ_1,μ_2)由$(X-\mu)'\Sigma^{-1}(X-\mu)=b^2$ 决定的椭圆. 由本章 § 3.1 中所介绍的知识可知

$$D^2 = (X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(2).$$

对给定 $p_0\in(0,1)$，则存在 d_0，使

$$P\{D^2 \leqslant d_0\} = p_0.$$

比如取 $p_0=1/2$，由统计软件计算或者查有关临界值表，可得 $d_0=1.386$；当 $p_0=0.25$ 时 $d_0=0.575$；$p_0=0.75$ 时 $d_0=2.773$，….而

$$P\{(X-\mu)'\Sigma^{-1}(X-\mu) \leqslant 1.386\} = 0.5$$

表示样品点 $X_{(i)}$落入由 1.386 指定的椭圆内的概率为 1/2. 利用这一

结论，可对二元观测数据的正态性进行检验，具体步骤请看下面例子.

例 3.6.1 考查由第一章表 1.1 中给出的 12 名学生的数学成绩(X_4)和物理成绩(X_5)所组成的二维向量$(X_4,X_5)'$的观测数据，试问这批二元数据可否认为是来自二元正态总体.

解 首先由二元数据计算样本均值和样本协方差阵：

$$\overline{X}=(80.7500,81.9167)',$$

$$S=\frac{1}{n-1}A=\frac{1}{11}\begin{bmatrix}4710.2500 & \\ 3995.7500 & 5288.9167\end{bmatrix}$$

$$=\begin{bmatrix}428.2046 & \\ 363.2500 & 480.8106\end{bmatrix},$$

$$S^{-1}=\begin{bmatrix}0.006503 & \\ -0.004913 & 0.005792\end{bmatrix}.$$

接着计算第 i 个观测点(学生)的两门课程成绩 $X_{(i)}$到中心点 $\overline{X}=(\overline{X}_4,\overline{X}_5)'=(80.7500,81.9167)'$的 D_i^2(第五章称这样定义的 D_i^2 为马氏距离)：

$$D_i^2=(X_{(i)}-\overline{X})'S^{-1}(X_{(i)}-\overline{X})\quad(i=1,2,\cdots,12).$$

其结果分别为

0.8832, 0.7787, 0.6965, 0.7891, 2.1882, 2.3849,

0.8768, 2.0337, 0.2691, 5.0465, 0.7892, 5.2641.

由上可知，马氏距离 $D_i^2\leqslant 1.386$ 的个数为 7 个，占样品总数的比例为 58.33%. 这与 0.5 相差不多，且因样本容量 $n=12$ 较小，因此可以认为所得结果不能拒绝数据是来自二元正态性的假设.

这个检验法显然比较粗糙，下面我们将介绍的一种判断数据联合正态性原则的比较正规的方法，就是对 n 个观测点的 $D_i^2(i=1,\cdots,n)$作 χ^2 图检验.

2. 二元数据的 χ^2 图检验法

因二元数据的 χ^2 图检验法与 p 元数据的 χ^2 图检验法其原理完全相同，故关于此检验方法的介绍请参阅下面 p 元数据的 χ^2 图检验方法.

三、p 元数据的正态性检验

设 $X_{(\alpha)}=(X_{\alpha 1},\cdots,X_{\alpha p})'$ $(\alpha=1,\cdots,n)$ 为来自 p 元总体 X 的随机样本. 检验

$$H_0: X\sim N_p(\mu,\Sigma),\quad H_1: X \text{ 不服从 } N_p(\mu,\Sigma).$$

1. χ^2 统计量的 Q-Q 图检验法(或 P-P 图检验法)

这是由正态分布的性质的结论 4 构造的检验法. 在 H_0 下,将样品 X 到总体中心 μ 的马氏距离 $D^2(X,\mu)$ 记为 D^2,则有

$$D^2=(X-\mu)'\Sigma^{-1}(X-\mu)\sim\chi^2(p).$$

以下构造的检验方法就是检验统计量 D^2 是否有 $D^2\sim\chi^2(p)$ 成立. 直观的想法是: 由样品 $X_{(\alpha)}$ 计算 $D_\alpha^2(\alpha=1,\cdots,n)$,对 D_α^2 排序:

$$D_{(1)}^2\leqslant D_{(2)}^2\leqslant\cdots\leqslant D_{(n)}^2.$$

统计量 D^2 的经验分布函数取为

$$F_n(D_{(t)}^2)=\frac{t-0.5}{n}\xlongequal{\text{def}}p_t\approx H(D_{(t)}^2\mid p),$$

其中 $H(D_{(t)}^2\mid p)$ 表示 $\chi^2(p)$ 的分布函数在 $D_{(t)}^2$ 的值.

设 χ^2 分布的 p_t 分位数为 χ_t^2,显然 χ_t^2 满足: $H(\chi_t^2\mid p)=p_t$,即 χ^2 分布的 p_t 分位数 $\chi_t^2=H^{-1}(p_t\mid p)$. 又由经验分布得到样本的 p_t 分位数 $D_{(t)}^2=F_n^{-1}(p_t)$. 若 $H(x\mid p)\approx F_n(x)$,应有

$$D_{(t)}^2\approx\chi_t^2.$$

绘制点 $(D_{(t)}^2,\chi_t^2)$ 的散布图,当 X 为正态总体时,这些点应散布在一条直线上. 这种检验法其实就是 χ^2 分布的 Q-Q 图检验法.

类似地,也可以绘制点 $(p_t,H(D_{(t)}^2\mid p))$ 的散布图,当 X 为正态总体时,这些点也应散布在一条直线上. 这种检验法其实就是 χ^2 分布(有时表示为"卡方分布")的 P-P 图检验法.

具体检验步骤如下:

(1) 由 n 个 p 维样品点 $X_{(\alpha)}(\alpha=1,\cdots,n)$ 计算样本均值 $\overline{X}$ 和样本协方差阵 S:

$$S=\frac{1}{n-1}\sum_{\alpha=1}^{n}(X_{(\alpha)}-\bar{X})(X_{(\alpha)}-\bar{X})'.$$

(2) 计算样品点 $X_{(t)}$ 到 $\bar{X}$ 的马氏距离：

$$D_t^2=(X_{(t)}-\bar{X})'S^{-1}(X_{(t)}-\bar{X})\quad(t=1,\cdots,n).$$

(3) 对马氏距离 D_t^2 按从小到大的次序排序：

$$D_{(1)}^2\leqslant D_{(2)}^2\leqslant\cdots\leqslant D_{(n)}^2.$$

(4) 计算 $p_t=\dfrac{t-0.5}{n}$ $(t=1,2,\cdots,n)$及 χ_t^2，其中 χ_t^2 满足：

$$H(\chi_t^2|p)=p_t\quad(\text{或计算 }H(D_{(t)}^2|p)\text{ 的值}).$$

(5) 以马氏距离为横坐标，χ^2 分位数为纵坐标作平面坐标系，用 n 个点$(D_{(t)}^2,\chi_t^2)$绘制散布图，即得到 χ^2 分布的 Q-Q 图；或者用另 n 个点$(p_t,H(D_{(t)}^2|p))$绘制散布图，即得 χ^2 分布的 P-P 图.

(6) 考察这 n 个点是否散布在一条通过原点，斜率为 1 的直线上，若是，接受数据来自 p 元正态总体的假设；否则拒绝正态性假设.

2. 主成分检验法

设 $X_{(i)}=(X_{i1},X_{i2},\cdots,X_{ip})'$ $(i=1,\cdots,n)$为来自 p 元总体 $X=(X_1,\cdots,X_p)'$的观测数据(样本)，检验

$$H_0:X\sim N_p(\mu,\Sigma),\quad H_1:X\text{ 不服从 }N_p(\mu,\Sigma).$$

设样本协方差阵 S 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$，相应的特征向量为 $l_1,l_2,\cdots,l_p$，记 $l_t=(l_{1t},l_{2t},\cdots,l_{pt})'$. 令

$$Z_t=l_{1t}X_1+l_{2t}X_2+\cdots+l_{pt}X_p\quad(t=1,2,\cdots,p),$$

即新变量 $Z_1,\cdots,Z_p$ 是 $X_1,\cdots,X_p$ 的线性组合. 则可以证明：$Z_1,\cdots,Z_p$ 是相互独立的；p 元观测数据提供的信息大部分可由前几个新变量所提供. 这时 p 元数据的正态性检验可化为几个相互独立的新变量的一元数据的正态性检验. 这些新变量在第七章中被称为**主成分**. 故此检验法称为**主成分检验法**.

如果正态性假设不能成立，一般应考虑对数据进行变换，使非正态数据更接近正态，然后对变换后的数据进行统计分析. 有关变换的方法请见参考文献[5]、[6]或[7].

习　题　三

3-1　设 $X\sim N_n(\mu,\sigma^2 I_n)$，$A$ 为 n 阶对称幂等矩阵，且 $\mathrm{rank}(A)=r$ $(r\leqslant n)$. 证明 $\frac{1}{\sigma^2}X'AX\sim\chi^2(r,\delta)$，其中 $\delta=\frac{1}{\sigma^2}\mu'A\mu$.

3-2　设 $X\sim N_n(\mu,\sigma^2 I_n)$，$A,B$ 为 n 阶对称矩阵. 若 $AB=O$，证明 $X'AX$ 与 $X'BX$ 相互独立.

3-3　设 $X\sim N_p(\mu,\Sigma)$，$\Sigma>0$，A 和 B 为 p 阶对称矩阵，试证明

$$(X-\mu)'A(X-\mu)\text{ 与 }(X-\mu)B(X-\mu)\text{ 相互独立}$$

$$\Longleftrightarrow \Sigma A\Sigma B\Sigma=O_{p\times p}.$$

3-4　试证明威沙特分布的性质 4 和霍特林 T^2 分布的性质 5.

3-5　对单个 p 元正态总体 $N_p(\mu,\Sigma)$均值向量的检验问题，试用似然比原理导出检验 H_0：$\mu=\mu_0$（$\Sigma=\Sigma_0$ 已知）的似然比统计量及分布.

3-6　（均值向量的各分量间结构关系的检验）设总体

$$X\sim N_p(\mu,\Sigma)\quad(\Sigma>0),$$

$X_{(\alpha)}(\alpha=1,\cdots,n)$ $(n>p)$为来自 p 元正态总体 X 的样本，记 $\mu=(\mu_1,\cdots,\mu_p)'$. C 为 $k\times p$ 常数矩阵$(k<p)$，$\mathrm{rank}(C)=k$，r 为已知 k 维向量. 试给出检验 H_0：$C\mu=r$ 的检验统计量及分布.

3-7　设总体 $X\sim N_p(\mu,\Sigma)$ $(\Sigma>0)$，$X_{(\alpha)}(\alpha=1,\cdots,n)$ $(n>p)$为来自 p 元正态总体 X 的样本，样本均值为 $\overline{X}$，样本离差阵为 A. 记 $\mu=(\mu_1,\cdots,\mu_p)'$. 为检验 H_0：$\mu_1=\mu_2=\cdots=\mu_p$，H_1：$\mu_1,\mu_2,\cdots,\mu_p$ 至少有一对不相等，令

$$C=\begin{bmatrix}1&-1&0&\cdots&0\\1&0&-1&\cdots&0\\\vdots&\vdots&\vdots&&\vdots\\1&0&0&\cdots&-1\end{bmatrix}_{(p-1)\times p},$$

则上面的假设等价于

$$H_0:C\mu=0_{p-1},\quad H_1:C\mu\neq 0_{p-1},$$

其中 0_{p-1} 为 $p-1$ 维零向量. 试求检验 H_0 的似然比统计量和分布.

3-8 假定人体尺寸有这样的一般规律：身高(X_1)，胸围(X_2)和上半臂围(X_3)的平均尺寸比例是 6∶4∶1. 假设 $X_{(\alpha)}(\alpha=1,\cdots,n)$ 为来自总体 $X=(X_1,X_2,X_3)'$ 的随机样本，并设 $X\sim N_3(\mu,\Sigma)$. 试利用表 3.4 中男婴这一组数据来检验其身高、胸围和上半臂围这三个尺寸(变量)是否符合这一规律(写出假设 H_0，并导出检验统计量).

表 3.4 某地区农村两周岁婴儿的体格测量数据

性别	身高(X_1)	胸围(X_2)	上半臂围(X_3)
男	78	60.6	16.5
男	76	58.1	12.5
男	92	63.2	14.5
男	81	59.0	14.0
男	81	60.8	15.5
男	84	59.5	14.0
女	80	58.4	14.0
女	75	59.2	15.0
女	78	60.3	15.0
女	75	57.4	13.0
女	79	59.5	14.0
女	78	58.1	14.5
女	75	58.0	12.5
女	64	55.5	11.0
女	80	59.2	12.5

3-9 对单个 p 元正态总体 $N_p(\mu,\Sigma)$ 协方差阵的检验问题，试用似然比原理导出检验 H_0：$\Sigma=\Sigma_0$ 的似然比统计量及分布.

3-10 对两个 p 元正态总体 $N_p(\mu^{(1)},\Sigma)$ 和 $N_p(\mu^{(2)},\Sigma)$ 均值向量的检验问题，试用似然比原理导出检验 H_0：$\mu^{(1)}=\mu^{(2)}$ 的似然比统计量及分布.

3-11 表 3.4 给出 15 名两周岁婴儿的身高(X_1)，胸围(X_2)和上半臂围(X_3)的测量数据. 假设男婴的测量数据 $X_{(\alpha)}(\alpha=1,\cdots,6)$ 为来自总体 $N_3(\mu^{(1)},\Sigma)$ 的随机样本；女婴的测量数据 $Y_{(\alpha)}(\alpha=1,\cdots,9)$ 为来自总体 $N_3(\mu^{(2)},\Sigma)$ 的随机样本. 试利用表 3.4 中的数据检验 H_0：$\mu^{(1)}=\mu^{(2)}(\alpha=0.05)$.

3-12　地质勘探中，在 A，B，C 三个地区采集了一些岩石，测量其部分化学成分，其数据见表 3.5. 假定这三个地区岩石的成分遵从 $N_3(\mu^{(i)},\Sigma_i)$ $(i=1,2,3)$ $(\alpha=0.05)$.

(1) 检验 H_0：$\Sigma_1=\Sigma_2=\Sigma_3$；$H_1$：$\Sigma_1,\Sigma_2,\Sigma_3$ 不全等；

(2) 检验 H_0：$\mu^{(1)}=\mu^{(2)}$，H_1：$\mu^{(1)}\neq\mu^{(2)}$；

(3) 检验 H_0：$\mu^{(1)}=\mu^{(2)}=\mu^{(3)}$，$H_1$：存在 $i\neq j$，使 $\mu^{(i)}\neq\mu^{(j)}$；

(4) 检验三种化学成分相互独立.

表 3.5　岩石部分化学成分数据

	SiO_2	FeO	K_2O
A 地区	47.22	5.06	0.10
	47.45	4.35	0.15
	47.52	6.85	0.12
	47.86	4.19	0.17
	47.31	7.57	0.18
B 地区	54.33	6.22	0.12
	56.17	3.31	0.15
	54.40	2.43	0.22
	52.62	5.92	0.12
C 地区	43.12	10.33	0.05
	42.05	9.67	0.08
	42.50	9.62	0.02
	40.77	9.68	0.04

3-13　对表 3.3 给出的三组观测数据分别检验是否来自 4 元正态分布.

(1) 对每个分量检验是否是一元正态？

(2) 利用 χ^2 图检验法对三组观测数据分别检验是否来自 4 元正态分布.

第四章　回归分析

回归分析方法是多元统计分析的各种方法中应用最广泛的一种.它是处理多个变量间相互依赖关系的一种数理统计方法.变量间的相互依赖关系在实际问题中是大量存在的,回归分析是研究这种相互依赖关系的有效数学方法.

回归分析方法是在众多相关的变量中,根据实际问题的要求,考查其中一个或几个变量与其余变量的依赖关系.如果只要考查某一个变量(常称为响应变量、因变量或指标)与其余多个变量(称为自变量或因素)的相互依赖关系.我们称为**多元回归问题**.如果要同时考查 p 个因变量与 m 个自变量的相互依赖关系,我们称为**多因变量的多元回归问题**(或简称为**多对多回归**).

在一元统计分析中讨论的多元回归是只考虑一个因变量的回归问题.多元统计分析中讨论的回归问题是指有多个因变量的回归问题,它自然把一元统计中的回归作为特例.因多元回归问题在实际应用中更为广泛,它涉及的统计推断结论能够推广到多因变量的多元回归的问题中.本章首先不加证明地介绍经典多元线性回归、逐步回归的一些结论,然后讨论多因变量的多元线性回归和双重筛选逐步回归.

§4.1　经典多元线性回归

多元回归分析是研究因变量 Y 与 m 个自变量 $x_1, x_2, \cdots, x_m$ 的相关关系,而且总是假设因变量 Y 为随机变量,而 $x_1, x_2, \cdots, x_m$ 为一般变量.

$$= \frac{1}{n}(Y - Cb)'(Y - Cb) = \frac{1}{n}Q(b).$$

但因 $\hat{\sigma}^2$ 不是 σ^2 的无偏估计量，通常取

$$s^2 = \frac{1}{n - m - 1}Q(b)$$

作为 σ^2 的估计量，它是 σ^2 的无偏估计量.

定理 4.1.2　设 $\mathrm{rank}(C) = m + 1 \leqslant n$，则 $\mathrm{E}(s^2) = \sigma^2$.

5. 参数函数 $\alpha'\beta$ 的估计

在回归分析中，求出参数 β 的最小二乘估计 b 并不是我们的目的. 我们的目的是要估计 Y，而 Y 是 β 的线性函数：

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m = (1, x_1, \cdots, x_m)\beta = \alpha'\beta,$$

即估计参数 β 的线性函数 $\alpha'\beta$.

很自然地我们用 $\alpha'b$ 作为 $\alpha'\beta$ 的估计. 因 b 是 β 的最大似然估计量，故 $\alpha'b$ 也是 $\alpha'\beta$ 的最大似然估计量，它具有最大似然估计量的一切优良性. 特别要强调的，$\alpha'b$ 是 $\alpha'\beta$ 的最小方差线性无偏估计.

二、回归方程和回归系数的显著性检验

在实际问题中，我们事先并不能判定因变量 Y 与自变量 $x_1, x_2, \cdots, x_m$ 之间确有线性关系. 在求出回归系数 β 的估计之前，回归模型(4.1.2)只是一种假定，尽管这种假定常常不是没有根据的，但在求出线性回归方程后，对 Y 与 $x_1, x_2, \cdots, x_m$ 之间是否有线性关系还需进行统计检验，以给出肯定或者否定的结论.

我们假定 $\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$，如果 Y 与 $x_1, x_2, \cdots, x_m$ 之间均无线性相关关系，则(4.1.2)模型中 $x_i (i = 1, 2, \cdots, m)$ 的系数 β_i 应均为 0. 故检验 Y 与 $x_1, x_2, \cdots, x_m$ 是否线性相关的问题就等价于检验假设

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0,$$

为了选择合适的检验统计量，我们首先介绍平方和分解公式.

1. 平方和分解公式

引理 4.1.1　对任给定的观测数据阵

$$\left[\begin{array}{c|cccc} y_1 & x_{11} & x_{12} & \cdots & x_{1m} \\ y_2 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ y_n & x_{n1} & x_{n2} & \cdots & x_{nm} \end{array}\right],$$

恒有公式：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2, \quad (4.1.4)$$

其中

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,$$

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_mx_{im} \quad (i=1,2,\cdots,n).$$

而

$$\hat{\beta}=\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_m\end{bmatrix}=(C'C)^{-1}C'Y$$

是 β 的最小二乘估计. 公式(4.1.4)称为**平方和分解公式**.

平方和分解公式(4.1.4)等号的左边 $\sum_{i=1}^{n}(y_i-\bar{y})^2$ 体现了 Y 的观测值 $y_1,y_2,\cdots,y_n$ 总波动大小，称为**总偏差平方和**，记作 l_{yy}(或 TSS). (4.1.4)式等号右边的第二项 $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 体现了 n 个估计值 $\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$ 的波动大小，它是由于 Y 与自变量 $x_1,x_2,\cdots,x_m$ 之间确有线性关系并通过 $x_1,x_2,\cdots,x_m$ 的变化而引起，我们称它为**回归平方和**，记为 U (或 MSS)；(4.1.4)式等号右边第一项 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ 称为**残差平方和**，记为 Q(或 ESS). 在模型(4.1.2)假定下，即

$$E(Y)=\beta_0+\beta_1x_1+\cdots+\beta_mx_m,$$

Q 是由于随机误差引起的. 实际上，模型(4.1.2)只是一种假定，自变量 $x_1,x_2,\cdots,x_m$ 和 Y 的关系除了线性关系外，可能还有非线性的关

系. Q 是除了 $x_1,x_2,\cdots,x_m$ 对 Y 的线性关系之外的一切其他因素(包括 $x_1,x_2,\cdots,x_m$ 对 Y 的非线性关系及随机误差)引起的,故 Q 也称为**剩余平方和**. 利用以上记号,(4.1.4)式可简写为:

$$l_{yy}=Q+U \quad 或 \quad \mathrm{TSS}=\mathrm{ESS}+\mathrm{MSS}. \tag{4.1.5}$$

2. 回归方程的显著性检验(或称相关性检验)

由最小二乘准则求回归系数的计算过程中,并不一定知道 Y 与自变量 $x_1,x_2,\cdots,x_m$ 是否有线性关系. 如果不存在线性关系,那么得到的回归方程是毫无意义的. 在一元回归中,若 $\beta_1=0$,则一般地说,Y 并没有随 x_1 的变化而线性的变化. 因此对回归方程的显著性检验就是检验以下假设是否成立:

$$H_0:\ \beta_1=\beta_2=\cdots=\beta_m=0. \tag{4.1.6}$$

由平方和分解公式及 MSS 和 ESS 的意义,若 MSS(回归平方和或模型平方和)比 ESS(残差平方和或误差平方和)大得多,则 Y 的总偏差 TSS 主要由 $x_i(i=1,\cdots,m)$ 的变化引起的,即所考察的这些自变量对 Y 的影响是显著的,也就是假设(4.1.6)不成立. 利用比值 MSS/ESS 就可以构造检验假设(4.1.6)的检验统计量.

定理 4.1.3　在模型(4.1.3)下有

(1) $\hat{\beta}\sim N_{m+1}(\beta,\sigma^2(C'C)^{-1})$;

(2) $\frac{1}{\sigma^2}Q\sim\chi^2_{n-m-1}$;

(3) $\hat{\beta}$ 与 Q 相互独立;

(4) $H_0:\ \beta_1=\beta_2=\cdots=\beta_m=0$ 成立时,$\frac{U}{\sigma^2}=\frac{\mathrm{MSS}}{\sigma^2}\sim\chi^2_m$.

利用定理 4.1.3,为检验 H_0,构造的检验统计量为

$$F=\frac{U/m}{Q/(n-m-1)}=\frac{\mathrm{MSS}/m}{\mathrm{ESS}/(n-m-1)}=\frac{\mathrm{MMS}(\text{模型均方})}{\mathrm{MSE}(\text{均方误差})}.$$

在 H_0 成立时,检验统计量 $F\sim F(m,n-m-1)$,其中 m 和 $n-m-1$ 分别称为**模型的自由度**和**误差的自由度**.

利用 n 组观测数据,计算检验统计量 F 的值(记为 f_0)及显著性概率(p 值),p 值是指在 H_0 下,利用 F 的分布规律,计算出检验统

计量 F 大于等于 f_0 的概率. 若得出的 p 值很小(小于显著性水平 α),依统计思想,小概率事件在一次实践中一般不会发生. 如果发生小概率事件,将否定前提假定 H_0;否则 H_0 相容.

3. 正规方程的等价形式及 U 的计算公式

回归模型(4.1.1)可以改写为

$$\begin{cases} y_i - \bar{y} = \beta_0^* + \beta_1(x_{i1} - \bar{x}_1) + \cdots + \beta_m(x_{im} - \bar{x}_m) + \varepsilon_i \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (i = 1,2,\cdots,n), \\ \varepsilon = (\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)' \sim N_n(0,\sigma^2 I_n). \end{cases} \tag{4.1.7}$$

它与原模型(4.1.1)没有本质差别,只不过 $\beta_0 = \beta_0^* - \sum_{i=1}^{m} \beta_i \bar{x}_i + \bar{y}$.

模型(4.1.7)的特点是对观测数据 $(x_{i1},x_{i2},\cdots,x_{im},y_i)$ $(i=1,2,\cdots,n)$ 做了中心化处理. 下面将说明在模型(4.1.7)下得到的正规方程的形式. 记

$$\widetilde{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1m} - \bar{x}_m \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2m} - \bar{x}_m \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nm} - \bar{x}_m \end{bmatrix},$$

$$B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \quad \widetilde{Y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \quad \widetilde{C} = (\mathbf{1}_n \vdots \widetilde{X}), \quad \widetilde{\beta} = \begin{bmatrix} \beta_0^* \\ B \end{bmatrix}.$$

则模型(4.1.7)的矩阵形式为

$$\begin{cases} \widetilde{Y} = \widetilde{C}\widetilde{\beta} + \varepsilon, \\ \varepsilon \sim N_n(0,\sigma^2 I_n), \end{cases}$$

正规方程为: $\widetilde{C}'\widetilde{C}\widetilde{\beta} = \widetilde{C}'\widetilde{Y}$. 又

$$\widetilde{C}'\widetilde{C} = (\mathbf{1}_n \vdots \widetilde{X})'(\mathbf{1}_n \vdots \widetilde{X}) = \begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\widetilde{X} \\ \widetilde{X}'\mathbf{1}_n & \widetilde{X}'\widetilde{X} \end{bmatrix} \overset{\text{def}}{=\!=} \begin{bmatrix} n & O_{1\times m} \\ O_{m\times 1} & L \end{bmatrix},$$

其中

$$L=\widetilde{X}'\widetilde{X}=(l_{ij})_{m\times m},$$

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

而

$$\widetilde{C}'\widetilde{Y}=(\mathbf{1}_n\mid\widetilde{X})'\widetilde{Y}=\begin{bmatrix}\mathbf{1}_n' & \widetilde{Y}\\ \widetilde{X}' & \widetilde{Y}\end{bmatrix}\xlongequal{\text{def}}\begin{bmatrix}0\\ l\end{bmatrix},$$

其中 $l=(l_{1y},l_{2y},\cdots,l_{my})'$，$l_{iy}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(y_k-\bar{y})$. 于是正规方程可写为：

$$\begin{bmatrix}n & O_{1\times m}\\ O_{m\times 1} & L\end{bmatrix}\begin{bmatrix}\beta_0^*\\ B\end{bmatrix}=\begin{bmatrix}0\\ l\end{bmatrix},$$

由此可得出 $\hat{\beta}_0^*=0$，故正规方程的另一等价形式为：

$$LB=l,\tag{4.1.8}$$

其中 $L=\widetilde{X}'\widetilde{X}$，$l=\widetilde{X}'\widetilde{Y}$. 方程(4.1.8)是 m 元线性方程组，解之得 B 的最小二乘估计为：$\hat{B}=L^{-1}l$，且 $\hat{B}\sim N_m(B,\sigma^2L^{-1})$.

数据中心化后的线性回归模型(4.1.7)可表为

$$\begin{cases}\widetilde{Y}=\widetilde{X}B+\varepsilon,\\ \varepsilon\sim N_n(0,\sigma^2I_n).\end{cases}\tag{4.1.9}$$

因为 $U=(\hat{Y}-\bar{y}\mathbf{1}_n)'(\hat{Y}-\bar{y}\mathbf{1}_n)$，在模型(4.1.9)下 $(\hat{Y}-\bar{y}\mathbf{1}_n)=\widetilde{X}\hat{B}$，所以回归平方和 U 有以下计算公式：

$$U=(\widetilde{X}\hat{B})'\widetilde{X}\hat{B}=\hat{B}'L\hat{B}=\hat{B}'l$$

$$=\hat{\beta}_1l_{1y}+\hat{\beta}_2l_{2y}+\cdots+\hat{\beta}l_{my}.$$

利用平方和分解公式，还可得残差平方和的计算公式为：

$$Q=l_{yy}-U.$$

4. 回归系数的显著性检验

对回归方程进行显著性检验，若否定 H_0，仅表示 $\beta_1,\beta_2,\cdots,\beta_m$ 不全为 0，但并不排除有某个 β_i 为 0. 若 $\beta_i=0$，说明自变量 x_i 对因变量 Y 的影响不明显，应从回归模型中删除. 因此对回归系数 $\beta_i(i=1,2,\cdots,m)$ 是否为 0，进行逐个检验是很必要的，即检验以下的假设：

$$H_0^{(i)}: \beta_i = 0 \quad (i = 1,2,\cdots,m). \tag{4.1.10}$$

为构造检验以上假设的检验统计量，我们引进偏回归平方和的概念. 它是刻画某个自变量对 Y 作用大小的统计量.

定义 4.1.2 设 U 为 $x_1,\cdots,x_m$ 对 Y 的回归平方和；$U(i)$为去掉 x_i 后余下的 $m-1$ 个自变量对 Y 的回归平方和. 则称 $P_i=U-U(i)$(或 $P_i=Q(i)-Q$)为变量 x_i 的**偏回归平方和**.

P_i 表示去掉自变量 x_i 后回归平方和减少(或残差平方和增加)的数值，由定义可知，P_i 这个数值大，说明 x_i 重要，这个数值小，说明 x_i 不重要.

可以证明 P_i 的计算公式为

$$P_i = \hat{\beta}_i^2 / l^{ii} \quad (i = 1,2,\cdots,m), \tag{4.1.11}$$

其中 l^{ii} 为 L^{-1}的第 i 个对角元素，而 $L=\widetilde{X}'\widetilde{X}$，$\widetilde{X}$ 是中心化的数据阵. 检验 H_0: $\beta_i=0$ $(i=1,2,\cdots,m)$的检验统计量选为

$$F_i = \frac{P_i}{Q/(n-m-1)}, \quad \text{或} \quad t_i = \frac{\hat{\beta}_i / \sqrt{l^{ii}}}{\sqrt{Q/(n-m-1)}}.$$

已知 $Q/\sigma^2 \sim \chi^2(n-m-1)$，而 $P_i=\hat{\beta}_i^2/l^{ii}$. 又已知

$$\hat{B} = (\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_m)' \sim N_m(B, L^{-1}\sigma^2),$$

所以

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 l^{ii}).$$

在 $H_0^{(i)}$ 成立时 $\dfrac{\hat{\beta}_i}{\sigma\sqrt{l^{ii}}} \sim N(0,1)$，即 $\dfrac{\hat{\beta}_i^2}{\sigma^2 l^{ii}} \sim \chi^2(1)$，且与 Q 相互独立，所以

$$F_i = \frac{P_i}{Q/(n-m-1)} \sim F(1, n-m-1),$$

或

$$t_i = \frac{\hat{\beta}_i / \sqrt{l^{ii}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1).$$

给定显著性水平 α，由样本观测数据计算 Q、P_i 及检验统计量的值(记为 f_i)，并计算显著性概率值(p 值)：$p=P\{F_i \geqslant f_i\}$. 若 $p<\alpha$，否定 H_0，即认为 x_i 对 Y 的作用是显著的(x_i 在回归方程中是显著的)；否则 H_0 相容.

5. 建立“最优”回归方程

所谓“最优”回归方程是指包含所有在显著性水平 α 下对 Y 作用显著的变量，而不包含在显著性水平 α 下对 Y 作用不显著的变量的回归方程.

经对 m 个变量逐个做检验后，若 m 个变量在给定的显著性水平 α 下对 Y 作用都是显著的，即认为所得方程是“最优”回归方程. 若有不显著变量，则每次只能剔除一个，然后由余下的变量和 Y 再做回归，然后再逐个检验，每次只许剔除一个最不重要的变量. 重复以上步骤，直至方程中的变量都是重要的为止. 这时得到的方程即为“最优”回归方程. 利用此方程可对生产过程作预报或进行控制.

例 4.1.1(水泥数据) 设某种水泥在凝固时所释放的热量 Y(卡/克)与水泥中下列四种化学成分有关：

x_1——$3CaO \cdot Al_2O_3$ 的成分(%)；

x_2——$3CaO \cdot SiO_2$ 的成分(%)；

x_3——$4CaO \cdot Al_2O_3 \cdot Fe_2O_3$ 的成分(%)；

x_4——$2CaO \cdot SiO_2$ 的成分(%).

共观测了 13 组数据(见表 4.1). 试求出 Y 与 x_1, x_2, x_3, x_4 的回归方程，并对该回归方程和各个回归系数进行检验.

表 4.1　水泥数据

序号	x_1	x_2	x_3	x_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

解 使用 SAS/STAT 软件中最常用的 REG 过程来完成经典多元线性回归分析中的估计和检验问题. REG 过程产生的主要结果见输出 4.1.1.

输出 4.1.1 REG 过程产生的主要输出结果

建立水泥数据的多元线性回归模型

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	2667.89944	666.97486	111.479	0.0001
Error	8	47.86364	5.98295		
C Total	12	2715.76308			

Root MSE	2.44601	R-square	0.9824
Dep Mean	95.42308	Adj R-sq	0.9736
C.V.	2.56333		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	62.405369	70.07095921	0.891	0.3991
X1	1	1.551103	0.74476987	2.083	0.0708
X2	1	0.510168	0.72378800	0.705	0.5009
X3	1	0.101909	0.75470905	0.135	0.8959
X4	1	-0.144061	0.70905206	-0.203	0.8441

输出 4.1.1 给出以下几方面结果：

(1) 回归方程：

$$\hat{Y}=62.4054+1.5511x_1+0.5102x_2+0.1019x_3-0.1441x_4.$$

(2) 回归方程显著性检验的结果：由该输出中方差分析(Analysis of Variance)表可得出，平方和分解式为：

$$2715.76308 = 2667.89944 + 47.86364;$$

均方误差为 MSE=47.86364/8=5.98295，它是模型中误差方差 σ^2 的估计；该表还给出检验统计量 F 值为 111.479，p 值为 0.0001，这表示拟合的模型是高度显著的，该模型解释了这组数据总变差中的主要部分.

(3) 回归系数显著性检验的结果：该输出中参数估计(Parameter Estimates)表不仅给出回归方程的系数，并给出检验 $H_0^{(i)}$：$\beta_i=0$ $(i=0,1,\cdots,m)$的结果. 见该表的最右边一列“Prob>|T|”(即显著性概率 p 值)，若给定 $\alpha=0.05$，常数项(或称截距项)和 4 个自变量的 p 值均$\geqslant\alpha$，这与回归方程高度显著产生矛盾. 从后面的

讨论将看到此现象是因为 4 个自变量间存在较强的相关性. 为了得到“最优”回归方程,应从方程中删除最不重要的自变量(如 x_3,因 x_3 的 $p=0.8959$ 为最大),重新建立 Y 与其余自变量的回归方程后再检验. 我们将在 § 4.2 中介绍变量选择问题.

(4) 有关的回归统计量: 决定系数 $R^2=0.9824$(或复相关系数 $R=\sqrt{0.9824}$),标准差 σ 的估计量(Root MSE)为 2.4460,回归平方和 $U=2667.8994$,残差平方和 $Q=47.8636$.

三、预报与控制

在模型(4.1.3)的假定下,由观测数据求得参数 β 的估计值,从而得到回归方程:

$$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m, \tag{4.1.12}$$

并经过检验,设以上方程就是“最优”回归方程.

1. 预测 y_0 点的预报区间(区间估计)

设给定点$(x_{01},x_{02},\cdots,x_{0m})$处 Y 的观测值 y_0 是随机变量,它满足

$$y_0=\beta_0+\beta_1x_{01}+\cdots+\beta_mx_{0m}+\varepsilon_0.$$

但 y_0 未知. 很自然地我们把$(x_{01},x_{02},\cdots,x_{0m})$代入回归方程(4.1.12)得到 Y 的回归值(或称预报值)

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_mx_{0m}.$$

作为 y_0 的估计,它是 y_0 的一个最小方差线性无偏估计量.

但由回归方程仅得出 y_0 的点估计 $\hat{y}_0$ 并没有给出估计的精度. 我们进一步来讨论 y_0 的区间估计问题. 首先给出一条有关的定理.

定理 4.1.4　设给定点$(x_{01},x_{02},\cdots,x_{0m})$处因变量 Y 的观测值为 y_0 及样本$(x_{j1},x_{j2},\cdots,x_{jm},y_j)$ $(j=1,2,\cdots,n)$满足模型:

$$\begin{cases} y_j=\beta_0+\beta_1x_{j1}+\cdots+\beta_mx_{jm}+\varepsilon_j \quad (j=1,2,\cdots,n), \\ y_0=\beta_0+\beta_1x_{01}+\cdots+\beta_mx_{0m}+\varepsilon_0, \\ \varepsilon_1,\cdots,\varepsilon_n,\varepsilon_0\sim N(0,\sigma^2),\text{且相互独立}. \end{cases}$$

则

(1) $\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_mx_{0m}=(1,x_{01},x_{02},\cdots,x_{0m})\hat{\beta}\overset{\text{def}}{=\!=}$

$x_0'\hat{\beta}$ 是 y_0 的最小方差线性无偏估计量，且

$$\hat{y}_0 \sim N(x_0'\beta, \sigma^2 x_0'(C'C)^{-1}x_0);$$

(2) $y_0-\hat{y}_0 \sim N(0, \sigma^2(1+x_0'(C'C)^{-1}x_0)$；

(3) 统计量 t 为

$$t=\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+x_0'(C'C)^{-1}x_0}} \sim t(n-m-1) \quad \left(\text{其中 } \hat{\sigma}=\sqrt{\frac{Q}{n-m-1}}\right).$$

利用以上定理，可得出 y_0 的预报区间.

给定置信度 $1-\alpha$，选用定理 4.1.4 给出的统计量 t，因 $t \sim t(n-m-1)$，查 t 分布表得临界值 t_α，使

$$P\{|t| < t_\alpha\} = 1-\alpha,$$

即
$$P\{|y_0-\hat{y}_0| < t_\alpha\hat{\sigma}\sqrt{1+x_0'(C'C)^{-1}x_0}\} = 1-\alpha.$$

称 $d=t_\alpha\hat{\sigma}\sqrt{1+x_0'(C'C)^{-1}x_0}$ 为**预报半径**，则 y_0 的置信度为 $1-\alpha$ 的置信区间为 $[\hat{y}_0-d, \hat{y}_0+d]$. 该区间以 $\hat{y}_0$ 为中心，d 为半径. 若预报半径 d 小，则预报就精确. 由 d 的定义可以看出：

(1) 若 t_α 小(即 $1-\alpha$ 小)，则 d 也小；

(2) 因 $\hat{\sigma}=\sqrt{\frac{Q}{n-m-1}}$，当 Q 小时，d 也小；

(3) 利用分块求逆公式有

$$x_0'(C'C)^{-1}x_0 = (1, x_{01}, \cdots, x_{0m}) \begin{bmatrix} \frac{1}{n}+\overline{X}'L_{XX}^{-1}\overline{X} & -\overline{X}'L_{XX}^{-1} \\ -L_{XX}^{-1}\overline{X} & L_{XX}^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0m} \end{bmatrix}$$

$$= \frac{1}{n} + (\overline{x}_1 - x_{01}, \cdots, \overline{x}_m - x_{0m}) L_{XX}^{-1} \begin{bmatrix} \overline{x}_1 - x_{01} \\ \vdots \\ \overline{x}_m - x_{0m} \end{bmatrix}.$$

当样本容量 n 充分大，$\overline{x}_i - x_{0i} \approx 0$ $(i=1,2,\cdots,m)$ 时 d 小. 在实际问题中，常近似地认为 $y_0-\hat{y}_0 \sim N(0, \hat{\sigma}^2)$，当 $\alpha=0.05$ 时，预报区间为 $[\hat{y}_0-2\hat{\sigma}, \hat{y}_0+2\hat{\sigma}]$；当 $\alpha=0.01$ 时，预报区间为 $[\hat{y}_0-3\hat{\sigma}, \hat{y}_0+3\hat{\sigma}]$.

2. $E(y_0)$ 的预报区间

以上给出单个 y_0 的预报区间，类似可以讨论 $E(y_0)$ 的预报区

间. 由 $E(y_0)$的点估计 $\hat{y}_0$ 的分布容易得出 $E(y_0)$的置信度为 $1-\alpha$ 的置信区间为

$$[\hat{y}_0 - d_1, \hat{y}_0 + d_1] \quad (\text{其中 } d_1 = t_\alpha \hat{\sigma} \sqrt{x_0'(C'C)^{-1}x_0}).$$

该区间以 $\hat{y}_0$ 为中心，d_1 为半径(显然 $d_1 < d$).

3. 控制问题

控制问题实际上是预报的反问题. 如实际问题要求 y_0 落在一定的范围内：$A < y_0 < B$，问如何控制自变量 $x_1, x_2, \cdots, x_m$ 的取值，这就是控制问题.

给置信度 $1-\alpha$ (α 称为显著性水平)，则近似地有(当 $\alpha = 0.05$)

$$P\{\hat{y}_0 - 2\hat{\sigma} < y_0 < \hat{y}_0 + 2\hat{\sigma}\} = 0.95.$$

解不等式：

$$\begin{cases} \hat{y}_0 + 2\hat{\sigma} < B, \\ \hat{y}_0 - 2\hat{\sigma} > A. \end{cases}$$

如果不等式有解，即得自变量 $x_{01}, x_{02}, \cdots, x_{0m}$ 的控制范围.

在实际问题中，常常希望通过控制 m 个变量中的某一个(或少数几个)来满足对 y_0 的要求.

§4.2　回归变量的选择与逐步回归

在实际问题中，影响因变量 Y 的因素(自变量)可能很多，人们希望从中挑选出影响显著的自变量来建立回归关系式，这就涉及到自变量选择问题.

在回归方程中若漏掉对 Y 影响显著的自变量，那么建立的回归式用于预测时将会产生较大的偏差. 但回归式中若包含的变量太多，且其中有些对 Y 影响不大，显然这样的回归式不仅使用不方便，而且反而会影响预测的精度. 因而选择合适的变量用于建立一个“最优”的回归方程是十分重要的问题.

一、变量选择问题

什么是“最优”回归方程？直观考虑应该是方程中包含的所有自

变量对因变量 Y 的影响都是显著的；而不包含在方程中的变量对 Y 的影响是不显著的(可忽略). 也就是从自变量集$\{x_1, x_2, \cdots, x_m\}$中选出适当的子集$\{x_{i_1}, x_{i_2}, \cdots, x_{i_l}\}$($l \leqslant m$)，使得建立 Y 与 $x_{i_1}, x_{i_2}, \cdots, x_{i_l}$ 的回归方程就是这样的“最优”回归方程(或“最优”回归子集). 这就是回归变量的选择问题.

1. 选择“最优”回归子集的方法

在 SAS/STAT 软件的 REG 过程中，选择变量子集的方法有八种，可分为三类：

1) “最优”子集的变量筛选法

该方法包括逐步回归法(逐步筛选法)(STEPWISE)、向前引入法(FORWARD)和向后剔除法(BACKWARD).

向前引入法是从回归方程仅包含常数项开始，把自变量逐个引入回归方程. 具体地说，先在 m 个自变量中选择一个与因变量线性关系最密切的变量，记为 x_{i_1}，然后在剩余的 $m-1$ 个自变量中，再选一个 x_{i_2}，使得$\{x_{i_1}, x_{i_2}\}$联合起来二元回归效果最好，第三步在剩下的 $m-2$ 个自变量中选择一个变量 x_{i_3}，使得$\{x_{i_1}, x_{i_2}, x_{i_3}\}$联合起来回归效果最好，……如此下去，直至得到“最优”回归方程为止.

在向前引入法中的终止条件，一般有下列三种筛选方法：

(1) 给定显著性水平 α，当某一步对将被引入变量的回归系数作显著性检查时，若 $p \geqslant \alpha$，则引入变量的过程结束，所得方程即为“最优”回归方程；

(2) 指定方程中自变量个数 m_0，当方程中引入的变量个数$=m_0$ 时，逐步引入过程结束；

(3) 给定选择自变量的某个准则(后面将介绍)，逐个引入变量，于是我们得到的分别含有一个，两个，……及全部自变量的 m 个回归方程，按给定准则从中选出一个最优的回归方程. 这个方法有一个明显的缺点，就是由于各自变量之间可能存在着相关关系，因此后续变量的选入可能会使前面已选入的自变量变得不重要. 这样最后得到的“最优”回归方程可能包含一些对 Y 影响不大的自变量.

向后剔除法与向前引入法正相反，首先将全部 m 个自变量引入

回归方程，然后逐个剔除对因变量 Y 作用不显著的自变量. 具体地说，从回归式 m 个自变量中选择一个对 Y 贡献最小的自变量，比如是 x_{j_1}，将它从回归方程中剔除；然后重新计算 Y 与剩下的 $m-1$ 个自变量的回归方程，再剔除一个贡献最小的自变量，比如 x_{j_2}，依次下去，直到得到“最优”回归方程为止. 在向后剔除法中终止条件与向前引入法类似.

向后剔除法的缺点有二：一是计算量大，特别当自变量个数 m 很大，其中不显著变量又很多时，其计算量比向前引入法大得多；二是前面剔除的变量有可能因以后变量的剔除变为相对重要的变量，这样最后得到的“最优”回归方程中有可能漏掉相对重要的变量.

逐步回归法是上述两个方法的综合. 向前引入法中被选入的变量，将一直保留在方程中. 向后剔除法中被剔除的变量，将永远排除在方程之外. 这两种方法在某些情况下会得到不合理的结果，于是产生了一个自然的想法，被选入的变量当它的作用在新变量引入后变得微不足道时，可以将它剔除；被剔除的变量，当它的作用在新变量引入情况下变得重要时，也可将它重新选入回归方程. 这样一种以向前引入法为主，变量可进可出的筛选变量方法，称为逐步回归法.

在应用上，逐步回归法面临的一个较大的困难是引入或删除时的显著性水平 α_{in} 或 α_{out} 的选择，若 α_{in} 和 α_{out} 都选得大，最后所得方程含较多的自变量；相反，方程所含的自变量则偏少. 理论上为保证筛选过程有限步停止，要求 $\alpha_{\text{in}} \leqslant \alpha_{\text{out}}$，但在很多实际应用中，人们一般令 $\alpha_{\text{in}} = \alpha_{\text{out}}$. 显然逐步回归法最终所得“最优”回归方程与显著水平 α 的选择有关，并不能保证所挑选出的回归方程在某种准则下是最优. 但从长期实践看，一般地逐步回归法所选出的回归方程是较好的，加之计算量少，因而到目前为止它仍是被广泛使用的变量选择方法.

2）计算量很大的全子集法

通过计算所有**可能回归子集**后按变量选择的标准选择最优回归方程. 选择方法包括 R^2 选择法(RSQUARE)、C_p 选择法(CP)和修正 R^2 选择法(ADJRSQ).

设有 m 个自变量，这 m 个自变量的任一个子集都可以和 Y 建立回归方程，为了寻找最符合要求的回归方程，一个自然的想法是将

m 个自变量所有可能的组合，一一与因变量 Y 建立回归方程，然后根据实际的需要，按某一选择变量准则，逐一比较所有可能回归子集，找出最优回归方程. 这就是**全子集法**. 全子集法的最大优点是能够得到在某准则下的最优回归方程，这是上面 1)中所提到的三种方法所不及的. 但是此法计算量很大，当自变量个数为 m 时，包含一个自变量的回归子集有 m 个，包含两个自变量的回归子集有 $C_m^2(=m\times(m+1)/2)$个，一般包含 k 个自变量的回归子集有 C_m^k 个，那么所有可能回归子集共有 2^m-1 个. 如 $m=10$ 时，共有 1023 个，当 m 较大时，数字 2^m-1 大得惊人，若没有一个巧妙的算法，即使是用计算机也难以承受. 目前不少学者提出了一些保留全子集法的优点，又可以大大减少计算量的算法，从而使全子集法成为一般工作中也能够使用的普通方法.

3) 计算量适中的选择法

不需要计算所有可能回归子集，但比较的子集个数多于 1)中所提到三种筛选方法的一些选择法，如最小 R^2 增量法(MINR)和最大 R^2 增量法(MAXR). 这两个选择法的细节请参见参考文献[18]和[20].

2. 变量选择的几个准则

回归变量的选择问题在实用中和理论上都是十分重要的. 这个问题最大的困难就是如何比较不同选择(即不同子集)的优劣，即最优选择的标准. 从不同的角度出发，可以有不同的比较准则，在不同的准则下，“最优”回归方程也可能不同.

评价一个回归方程：$\hat{Y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m$ 的好坏标准通常从以下几个方面来考虑：

(1) 残差平方和 Q 愈小愈好，或者说，复相关系数 $R=\sqrt{U/l_{yy}}$ 越靠近 1 越好；

(2) 剩余标准差 $s=\sqrt{Q/(n-m-1)}$越小越好；

(3) 回归方程中包含的自变量的个数 m 越少越好.

如果按(1)，Q 愈小愈好(或 R 愈大愈好)的原则来选择自变量子集，则毫无疑问应该选全部自变量，这与(3)矛盾. 显然(1)和(3)不

能独立地作为选择变量的准则，希望给出类似于(2)的同时兼顾 Q，m 都小的准则.

所有自变量个数为 k 的回归子集有 C_m^k 个. 通常记为 $A(k)$. 显然子集 $A(k)$中变量个数为 k.

设回归子集 $A(k)=\{i_1,i_2,\cdots,i_k\}\subset\{1,2,\cdots,m\}$. 相应的回归模型为

$$\begin{cases} y_t=\beta_0+\beta_{i_1}x_{ti_1}+\cdots+\beta_{i_k}x_{ti_k}+\varepsilon_t \quad (t=1,\cdots,n), \\ \varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\sim N(0,\sigma^2) \text{ 且相互独立.} \end{cases} \tag{4.2.1}$$

记

$$\beta(A(k))=(\beta_0,\beta_{i_1},\cdots,\beta_{i_k})',$$

$$C(k)=\left[\begin{array}{c:ccc} 1 & x_{1i_1} & \cdots & x_{1i_k} \\ 1 & x_{2i_1} & \cdots & x_{2i_k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{ni_1} & \cdots & x_{ni_k} \end{array}\right], \quad Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

则回归模型(4.2.1)可表为

$$\begin{cases} Y=C(k)\beta(A(k))+\varepsilon, \\ \varepsilon\sim N_n(0,\sigma^2I_n). \end{cases} \tag{4.2.2}$$

模型(4.2.2)中参数的最小二乘估计为

$$\hat{\beta}(A(k))=[C(k)'C(k)]^{-1}C(k)'Y;$$

残差平方和为

$$Q(A(k))=(Y-\hat{Y})'(Y-\hat{Y}),$$

其中 $\hat{Y}=C(k)\hat{\beta}(A(k))$；决定系数(即复相关系数 R 的平方)为

$$R^2(A(k))=1-\frac{Q(A(k))}{l_{yy}}.$$

对于变量个数为 k 的 C_m^k 个回归子集 $A(k)$中，设子集 $L(k)$满足

$$Q(L(k))=\min_{A(k)}Q(A(k))\xlongequal{\text{def}}Q_k. \tag{4.2.3}$$

即子集 $L(k)$是自变量个数为 k 的所有回归子集中残差平方和最小的一个子集.

对于不同的变量选择准则，选择最优子集的方法都可分两步进行，首先对固定变量个数 k ($k=1,2,\cdots,m$)，求满足(4.2.3)式的子集 $L(k)$；然后按不同的准则确定变量个数 k.

比较不同子集优劣的准则常见的有以下几种(记 n 为观测个数，k 为子集模型中自变量个数).

准则 1 均方误差 s^2 最小.

选择子集 A，使均方误差：

$$s^2(A)=\frac{Q_k}{n-k-1} \text{ 达最小.}$$

显然均方误差 $s^2(A)$是由子集 A 确定的回归模型中 σ^2 的无偏估计量.

准则 2 C_p 统计量准则.

一般称 Y 与 m 个自变量 $x_1,x_2,\cdots,x_m$ 的 n 次观测数据满足的回归模型(4.1.1)为全回归模型. 如果 m 个自变量中有部分对 Y 的影响不显著，这类自变量是无用变量，不妨设 $x_{i_{k+1}},\cdots,x_{i_m}$ 是无用变量. 这时称 Y 与 $x_{i_1},x_{i_2},\cdots,x_{i_k}$(有用变量)的回归模型(4.2.1)为**选回归模型**(简称选模型).

对样本点

$$c_{(t)}=(1,x_{ti_1},\cdots,x_{ti_k}\mid x_{ti_{k+1}},\cdots,x_{ti_m})'$$

$$\xlongequal{\text{def}}(x_t(k)'\mid x_t(m-k)') \quad (t=1,2,\cdots,n).$$

当选用模型(4.2.1)时由回归方程可得样本点 $c_{(t)}$的估计值为 $\hat{y}_t(k)$，它与理论值 $c'_{(t)}\beta$ 的偏差平方和记为 J_k，

$$J_k=\frac{\sum_{t=1}^{n}[x_t(k)'\hat{\beta}(A(k))-c'_{(t)}\beta]^2}{\sigma^2}.$$

可以证明，

$$\mathrm{E}(J_k)=\mathrm{E}(Q(A(k))/\sigma^2+2(k+1)-n,$$

其中 σ^2 为未知参数，一般取 $s^2=\frac{Q(A(m))}{n-m-1}$作为 σ^2 的估计.

定义 4.2.1 记 $p=k+1$，称统计量

$$C_p=\frac{Q(A(p-1))}{s^2}+2p-n$$

为 C_p **统计量**，其中 $s^2=\dfrac{Q(A(m))}{n-m-1}$.

C_p 统计量准则 1：根据 $\mathrm{E}(C_p)=\mathrm{E}(J_{p-1})$，选回归子集 A，使得子集 A 的 C_p 值 $\dfrac{Q_k}{s^2}+2(k+1)-n$ 达到最小.

C_p 统计量准则 2：当选模型(4.2.1)正确时，有 $\mathrm{E}(C_p)\approx p$（当 n 比 m 大得多时），选回归子集 A，使得子集 A 的 C_p 与 p 的差值

$$|C_p-p|=\left|\frac{Q_k}{s^2}+(k+1)-n\right|$$

达到最小.

C_p 统计量准则 3：绘制 C_p 值随 p $(1\leqslant p\leqslant m+1)$变化的图形——$C_p$ 图，综合以上两个 C_p 统计量准则，选 r 使点 (p,C_p) 接近 $C_p=p$ 的直线，且 C_p 值最小的子集 $L(r)$.

准则 3　修正 R^2(记为 $\widetilde{R}^2$)准则.

令

$$\widetilde{R}^2=1-\frac{n-i}{n-k-i}(1-R^2),$$

其中当模型含截距项 β_0 时 $i=1$，否则 $i=0$. 选回归子集 A，使得 $\widetilde{R}^2$ 达到最大.

理论上我们认为满足以下两条原则的方程是最优的，即：(1) 当增加变量时，不能使 R^2 显著提高；(2) 变量个数尽可能少. 修正 R^2 准则就是在此理论基础上提出的具体准则.

注意，当 $R^2<k/(n-1)$时，$\widetilde{R}^2<0$. 这说明相应的子集变量和 Y 的关系不密切.

准则 4　预测均方误差及平方和最小的准则(记 $p=k+1$).

(1) J_p 统计量：这是基于 n 个观测点预测偏差及方差和最小的准则. 即选择子集 A，使得

$$J_p(A)=\frac{n+k+1}{n-k-1}Q_k=(n+p)s^2(A)\text{ 达最小.}$$

(2) S_p 统计量：这是基于 n 个观测点的平均预测均方最小的准则. 即选择子集 A，使得

$$S_p(A)=\frac{Q_k/(n-p)}{n-p-1}=\frac{s^2(A)}{n-p-1}\text{ 达最小.}$$

(3) PRESS 统计量：记 $\hat{y}_i(i)$为删去第 i 个点后用其余 $n-1$ 组观测数据来建立回归方程，并用于预测第 i 个观测点的预测值. 即选择子集 A，使得 PRESS 统计量

$$\mathrm{PRESS}(A)=\sum_{i=1}^{n}(y_i-\hat{y}_i(i))^2 \text{ 达到最小.}$$

准则 5 AIC，SBC 或 BIC 准则.

该法则是赤池弦次等提出的，为同时兼顾 Q_k，k 都小的一类信息量准则.

定义 4.2.2 分别定义 **AIC，SBC 或 BIC 统计量**为($p=k+1$)：

$$\mathrm{AIC}(A(k))=n\ln\frac{Q(A(k))}{n}+2p,$$

$$\mathrm{SBC}(A(k))=n\ln\frac{Q(A(k))}{n}+p\ln n,$$

$$\mathrm{BIC}(A(k))=n\ln\frac{Q(A(k))}{n}+2(p+2)q-2q^2,$$

其中 $q=\dfrac{\hat{\sigma}^2}{Q(A(k))/n}$，$\hat{\sigma}^2=s^2$ 是全回归模型中 σ^2 的估计量.

以上我们采用在 REG 过程中所给出的定义. 对于定义 4.2.2 中的 $Q(A(k))/n$ 我们也可以更换成对应于选模型中 σ^2 的估计，即用 $Q(A(k))/(n-k-1)$ 代替 $Q(A(k))/n$. 选回归子集 $L(r)=\{x_{i_1},\cdots,x_{i_r}\}$，使 AIC(或 SBC 或 BIC)统计量达到最小的回归子集 $L(r)$ 为 AIC(或 SBC 或 BIC)准则下的最优回归子集.

二、逐步回归分析

逐步回归分析是目前被广泛使用的回归分析方法. 全子集法虽最终可得到在某准则下的最优回归方程；但此法计算量大，即使采用一些巧妙算法，当自变量个数 m 特别大(如 $m>30$)时，计算量仍很大. 在这种情况下，人们普遍地应用逐步回归法来解决实际问题. 逐步回归法吸收了向前引入法和向后剔除法的优点，克服了它们的缺点，计算量小，且最终保证得到相对于某显著性水平 α 下的“最优”回归方程.

1. 逐步回归的基本思想和基本步骤

以上介绍的选择回归子集的几种方法中，最常用的方法是逐步筛选变量的逐步回归法. 逐步回归的基本思想和基本步骤如下：

基本思想：逐个引入自变量. 每次引入对 Y 影响最显著的自变量，并对方程中的老变量逐个进行检验，把变为不显著的变量逐个从方程中剔除掉，最终得到的方程中既不漏掉对 Y 影响显著的变量，又不包含对 Y 影响不显著的变量.

基本步骤：首先给出引入变量的显著性水平 α_{in} 和剔除变量的显著性水平 α_{out}；然后按图 4.1 的框图筛选变量.

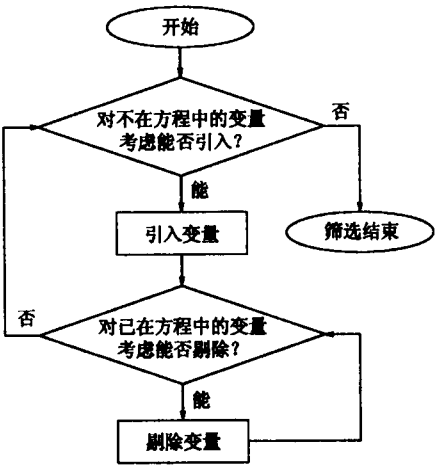

图 4.1　逐步回归的基本步骤

2. 逐步筛选法的基本步骤

设因变量 Y 与 m 个自变量 $x_1, x_2, \cdots, x_m$ 满足多元线性回归模型. 从逐步回归的基本思想和图 4.1 给出的变量筛选的过程可知，逐步筛选变量的过程主要包括两个基本步骤：一是从回归方程中考虑剔除不显著变量的步骤；二是从不在方程中的变量考虑引入新变量的步骤. 下面分别讨论这两方面的基本步骤.

(1) 考虑可否剔除变量的基本步骤. 假设已引入回归方程的变量为 $x_{i_1}, x_{i_2}, \cdots, x_{i_r} (r \leqslant m)$.

① 计算已在方程中的变量 x_{i_k} 的偏回归平方和 P_{i_k} 及偏 $R^2_{i_k}$：

$$\begin{aligned} P_{i_k} &= Q(i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_r) - Q(i_1,\cdots,i_r) \\ &= U(i_1,\cdots,i_r) - U(i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_r), \\ \text{偏 } R^2_{i_k} &= R^2(i_1,\cdots,i_r) - R^2(i_1,\cdots,i_{k-1},i_{k+1},\cdots,i_r) \\ &= P_{i_k}/l_{yy} \quad (k=1,\cdots,r), \end{aligned} \tag{4.2.4}$$

其中 $Q(\cdots)$(或 $U(\cdots)$或 $R^2(\cdots)$)表示包含括号中这些变量的回归模型的残差平方和(或回归平方和或决定系数). 度量回归方程中变量重要程度的统计量可采用偏回归平方和的大小,也可以采用偏 R^2 的大小. 在 REG 过程中,筛选变量时使用的统计量为偏 R^2. 以下介绍时,我们使用偏回归平方和 P_{i_k} 作为变量 x_{i_k} 重要性的度量. 设

$$P_{i_0} = \min(P_{i_1},\cdots,P_{i_r}),$$

即相应的变量 x_{i_0} 是方程中对 Y 影响最小的变量.

② 检验 x_{i_0} 对 Y 的影响是否显著. 对变量 x_{i_0} 进行回归系数的显著性检验,即检验 H_0: $\beta_{i_0}=0$,检验统计量为

$$F_{i_0} = \frac{P_{i_0}}{Q(i_1,\cdots,i_r)/(n-r-1)}, \tag{4.2.5}$$

及

$$p = P\{F \geqslant F_{i_0}\} \quad (\text{其中 } F \sim F(1,n-r-1)).$$

若 $p \geqslant \alpha_{\text{out}}$,则剔除 x_{i_0},重新建立 Y 与其余 $r-1$ 个变量的回归方程,然后再检验方程中最不重要的变量可否剔除,直到方程中没有变量可剔除后,转入考虑能否引入新变量的步骤. 若 $p < \alpha_{\text{out}}$,不能剔除 x_{i_0},转入考虑能否引入新变量的步骤.

(2) 考虑可否引入新变量的基本步骤. 假设已入选 r 个变量,不在方程中的变量记为 $x_{j_1},\cdots,x_{j_{m-r}}$.

① 计算不在方程中的变量 x_{j_k} 的偏回归平方和 P_{j_k} 及偏 $R^2_{j_k}$:

$$\begin{aligned} P_{j_k} &= Q(i_1,\cdots,i_r) - Q(i_1,\cdots,i_r,j_k), \\ \text{偏 } R^2_{j_k} &= P_{j_k}/l_{yy} \quad (k=1,2,\cdots,m-r), \end{aligned} \tag{4.2.6}$$

并设

$$P_{j_0} = \max(P_{j_1},\cdots,P_{j_{m-r}}),$$

即不在方程中的变量 x_{j_0} 是对 Y 影响最大的变量.

② 检验变量 x_{j_0} 对 Y 的影响是否显著. 对变量 x_{j_0} 作回归系数的显著性检验，即检验 H_0：$\beta_{j_0}=0$，检验统计量为

$$F_{j_0}=\frac{P_{j_0}}{Q(i_1,\cdots,i_r,j_0)/(n-r-2)},\tag{4.2.7}$$

及

$$p=P\{F\geqslant F_{j_0}\}\quad(\text{其中 } F\sim F(1,n-r-2)).$$

若 $p<\alpha_{\text{in}}$，则引入 x_{j_0}，并转入考虑可否剔除变量的步骤. 若 $p\geqslant\alpha_{\text{in}}$，则逐步筛选变量的过程结束.

假设用逐步回归法得到 r 个变量 $x_{i_1},x_{i_2},\cdots,x_{i_r}$，再建立 Y 与这 r 个变量的回归方程，这就是用逐步回归法得到的“最优”回归方程.

例 4.2.1(水泥数据)　设某种水泥在凝固时放出的热量 Y(卡/克)与水泥中四种化学成分 $x_1,\cdots,x_4$ 有关. 共观测了 13 组数据(见表 4.1)，试用逐步回归法求“最优”回归方程.

解　使用 SAS/STAT 软件的 REG 过程来完成逐步回归计算. 假设引入变量的显著性水平 $\alpha_{\text{in}}=0.10$，剔除变量的显著性水平 $\alpha_{\text{out}}=0.10$(一般取 $\alpha_{\text{in}}=\alpha_{\text{out}}$).

输出的计算结果首先给出筛选变量的过程：第一步引入 x_4，一元回归模型的 $R^2=0.6745$；第二步引入 x_1，Y 与 x_4,x_1 的二元回归模型的 $R^2=0.9725$；第三步引入 x_2，Y 与 x_4,x_1 和 x_2 的三元回归模型的 $R^2=0.9823$；因引入新变量后原变量 x_4 变得不重要了，故第四步剔除 x_4，Y 与 x_1,x_2 的二元回归模型的 $R^2=0.9787$. 经过这四步后，筛选变量的过程结束，所得到的“最优”回归方程中包含两个变量，即为：

$$Y=52.5774+1.4683x_1+0.6623x_2.$$

例 4.2.2(水泥数据)　试用全子集法求水泥在凝固时放出的热量 Y(卡/克)与四种化学成分 $x_1,\cdots,x_4$ 的最优回归方程.

解　使用 REG 过程中可完成所有可能回归子集的计算. 所有可能回归子集共有 $2^4-1=15$ 个，各回归子集中回归系数的估计结果见输出 4.2.1. 在输出 4.2.1 中，首先按回归方程所包含变量的

个数为1、2、3和4个的顺序给出15个所有可能回归子集的参数估计；再对每种变量个数，按 R^2 值从大到小的次序列出回归子集的 R^2(第2列)及回归系数的最小二乘估计. 如第1列(包含变量的个数)为2的第2行，给出包含变量 x_1 和 x_4 的回归子集，该回归子集的

$$R^2 = 0.9725,$$

由参数估计给出的回归方程为：

$$\hat{Y} = 103.09738 + 1.43996x_1 - 0.61395x_4.$$

几种最优准则的统计量见输出4.2.2. 从计算结果可以看出：按 C_p 及SBC统计量最小的准则得到最优回归子集为 $\{x_1, x_2\}$，该子

输出 4.2.1 所有可能回归子集中回归系数的估计

Number in Model	R-Square	Parameter Estimates: Intercept	x1	x2	x3	x4
1	0.6745	117.56793	.	.	.	-0.73816
1	0.6663	57.42368	.	0.78912	.	.
1	0.5339	81.47934	1.86875	.	.	.
1	0.2859	110.20266	.	.	-1.25578	.
2	0.9787	52.57735	1.46831	0.66225	.	.
2	0.9725	103.09738	1.43996	.	.	-0.61395
2	0.9353	131.28241	.	.	-1.19985	-0.72460
2	0.8470	72.07467	.	0.73133	-1.00839	.
2	0.6801	94.16007	.	0.31090	.	-0.45694
2	0.5482	72.34899	2.31247	.	0.49447	.
3	0.9823	71.64831	1.45194	0.41611	.	-0.23654
3	0.9823	48.19363	1.69589	0.65691	0.25002	.
3	0.9813	111.68441	1.05185	.	-0.41004	-0.64280
3	0.9728	203.64196	.	-0.92342	-1.44797	-1.55704
4	0.9824	62.40537	1.55110	0.51017	0.10191	-0.14406

输出 4.2.2 所有可能回归子集中几种最优准则的统计量

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	MSE	SBC
1	0.6745	0.6450	138.7308	58.8516	80.35154	59.98154
1	0.6663	0.6359	142.4864	59.1780	82.39421	60.30789
1	0.5339	0.4916	202.5488	63.5195	115.06243	64.64937
1	0.2859	0.2210	315.1543	69.0674	176.30913	70.19730
2	0.9787	0.9744	2.6782	25.4200	5.79045	27.11484
2	0.9725	0.9670	5.4959	28.7417	7.47621	30.43655
2	0.9353	0.9223	22.3731	39.8526	17.57380	41.54743
2	0.8470	0.8164	62.4377	51.0371	41.54427	52.73199
2	0.6801	0.6161	138.2259	60.6293	86.88801	62.32417
2	0.5482	0.4578	198.0947	65.1167	122.70721	66.81153
3	0.9823	0.9764	3.0182	24.9739	5.33030	27.23368
3	0.9823	0.9764	3.0413	25.0112	5.34562	27.27099
3	0.9813	0.9750	3.4968	25.7276	5.64846	27.98735
3	0.9728	0.9638	7.3375	30.5759	8.20162	32.83568
4	0.9824	0.9736	5.0000	26.9443	5.98295	29.76903

集正是用逐步回归法得到的；而按其余统计量的有关准则得到最优回归子集为$\{x_1,x_2,x_4\}$.

§4.3　多因变量的多元线性回归

前面介绍的回归模型，因变量仅有一个，自变量可以是多个，简称为**多元线性回归模型**. 在实际问题中，经常要同时考察多个自变量对多个因变量的相关关系. 如环境科学中，在同一时间地点，抽取了大气样品，测得多种污染气体，如 CO，SO_2 等的浓度. 大气样品中多种污染气体组成一个多维的随机向量，作为因变量；而大气中各污染气体的含量又与污染源的排放量以及气象因子(风向，风速，湿度等)有关，这就是一个多个因变量、多个自变量的回归问题. 再如工厂中要同时考察某产品的产量和质量指标，质量指标还可分为若干项，这样产量、质量等指标就是一个多维随机向量，作为因变量；而影响产品产量、质量的因素也有多个，这又是一个多对多的回归问题. 实际问题中，这种考察多个因变量与多个自变量的依赖关系的问题是大量存在的.

多对多的回归问题，当然也可以化为多个多元线性回归问题来解决. 但多个因变量之间一般存在某种相关关系. 如多种污染气体是来自同一大气样品，它们之间可能有某种相关关系，若分别对各种污染气体求其与污染源、气象因子的回归关系式，将会丢失一部分它们之间相互联系的信息. 在介绍了多元线性回归分析和逐步回归分析后，我们还要进一步来讨论多对多的回归模型.

一、模型和最小二乘估计

1. 多因变量的多元线性回归模型

设有 m 个自变量：$x_1,x_2,\cdots,x_m$，p 个因变量：$Y_1,Y_2,\cdots,Y_p$，假设它们之间有线性关系. 今有 n 组自变量与因变量的实测数据$(x_{t1},x_{t2},\cdots,x_{tm};y_{t1},y_{t2},\cdots,y_{tp})$ $(t=1,2,\cdots,n)$，数据阵分别用 X,Y 表示：

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix},\quad Y=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}.$$

设 n 组数据满足如下关系式

$$y_{tj}=\beta_{0j}+\beta_{1j}x_{t1}+\cdots+\beta_{mj}x_{tm}+\varepsilon_{tj}$$
$$(t=1,2,\cdots,n;j=1,2,\cdots,p).$$

记

$$\beta=\begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0p} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \vdots & \vdots & & \vdots \\ \beta_{m1} & \beta_{m2} & \cdots & \beta_{mp} \end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix} \beta'_{(0)} \\ \beta'_{(1)} \\ \vdots \\ \beta'_{(m)} \end{bmatrix}\overset{\text{def}}{=\!=}(\beta_1,\beta_2,\cdots,\beta_p),$$

$$E=\begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix}\overset{\text{def}}{=\!=}\begin{bmatrix} \varepsilon'_{(1)} \\ \varepsilon'_{(2)} \\ \vdots \\ \varepsilon'_{(n)} \end{bmatrix}\overset{\text{def}}{=\!=}(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_p),$$

则有

$$Y=(\mathbf{1}_n\vdots X)\beta+E=C\beta+E,$$

其中 C 为 $n\times(m+1)$ 矩阵；且假定 $\varepsilon_{(i)}=(\varepsilon_{i1},\varepsilon_{i2},\cdots,\varepsilon_{ip})'$ $(i=1,2,\cdots,n)$ 是相互独立的，其均值向量为 0，协方差阵相等，均为 Σ. 进一步可假定 $\varepsilon_{(i)}\sim N_p(0,\Sigma)$ $(i=1,2,\cdots,n)$.

定义 4.3.1 称模型

$$\begin{cases} Y=(\mathbf{1}_n\vdots X)\beta+E=C\beta+E, \\ \varepsilon_{(i)}\sim N_p(0,\Sigma)\ (i=1,2,\cdots,n)\ \text{相互独立}; \end{cases}\tag{4.3.1}$$

或

$$\begin{cases} Y=(\mathbf{1}_n\vdots X)\beta+E=C\beta+E, \\ \mathrm{E}(\varepsilon_{(i)})=0,\quad \mathrm{D}(\varepsilon_{(i)})=\Sigma\ (i=1,2,\cdots,n)\ \text{相互独立} \end{cases}\tag{4.3.2}$$

为多个因变量与多个自变量的**线性回归模型**，其中 Y 和 E 是随机

阵，$\beta=(\beta_{ij})$，$\Sigma=(\sigma_{ij})$是未知参数矩阵，X 是已知矩阵，$C=(\mathbf{1}_n \vdots X)$，且 $\mathrm{rank}(C)=m+1$.

2. 参数矩阵 β 的最小二乘估计

与一个因变量的多元线性回归分析一样，采用最小二乘法来求 β 的估计. 为此，我们来考察残差平方和 Q. 由(4.3.2)式知

$$E=Y-(\mathbf{1}_n \vdots X)\beta=(\varepsilon_{ij})_{n\times p}$$

$$(i=1,2,\cdots,n;\ j=1,2,\cdots,p),$$

残差平方和 $Q=\sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{ij}^2$.

因模型(4.3.2)等价于“拉直”后的模型：

$$\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_p\end{bmatrix}=\underset{np\times(m+1)p}{\begin{bmatrix}C&O&\cdots&O\\O&C&\cdots&O\\\vdots&\vdots&&\vdots\\O&O&\cdots&C\end{bmatrix}}\underset{(m+1)p\times 1}{\begin{bmatrix}\beta_1\\\beta_2\\\vdots\\\beta_p\end{bmatrix}}+\underset{np\times 1}{\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_p\end{bmatrix}},\tag{4.3.3}$$

其中

$$Y_j=\underset{n\times 1}{\begin{bmatrix}y_{1j}\\y_{2j}\\\vdots\\y_{nj}\end{bmatrix}},\quad \beta_j=\underset{(m+1)\times 1}{\begin{bmatrix}\beta_{0j}\\\beta_{1j}\\\vdots\\\beta_{mj}\end{bmatrix}},\quad \varepsilon_j=\underset{n\times 1}{\begin{bmatrix}\varepsilon_{1j}\\\varepsilon_{2j}\\\vdots\\\varepsilon_{nj}\end{bmatrix}}$$

$$(j=1,2,\cdots,p).$$

记

$$D=\begin{bmatrix}C&\cdots&O\\\vdots&&\vdots\\O&\cdots&C\end{bmatrix}=\begin{bmatrix}\mathbf{1}_nX&\cdots&O\\\vdots&&\vdots\\O&\cdots&\mathbf{1}_nX\end{bmatrix},$$

则(4.3.3)式可简记为

$$\mathrm{Vec}(Y)=D\mathrm{Vec}(\beta)+\mathrm{Vec}(E),$$

其中 D 为 $np\times(m+1)p$ 矩阵，β 为 $(m+1)\times p$ 矩阵，E 为 $n\times p$ 矩阵. 在模型(4.3.3)下

$$Q[\mathrm{Vec}(\beta)]=\sum_{i=1}^{n}\sum_{j=1}^{p}\varepsilon_{ij}^2=[\mathrm{Vec}(E)]'[\mathrm{Vec}(E)]$$

$$= [\mathrm{Vec}(Y) - D\mathrm{Vec}(\beta)]'[\mathrm{Vec}(Y) - D\mathrm{Vec}(\beta)]$$
$$= [\mathrm{Vec}(Y)]'[\mathrm{Vec}(Y)] - 2[\mathrm{Vec}(\beta)]'D'[\mathrm{Vec}(Y)]$$
$$+ [\mathrm{Vec}(\beta)]'D'D[\mathrm{Vec}(\beta)].$$

令

$$\frac{\partial Q[\mathrm{Vec}(\beta)]}{\partial[\mathrm{Vec}(\beta)]} = 0$$

(见附录中§8的(8.2)和(8.3)式),则正规方程组

$$D'D[\mathrm{Vec}(\beta)] = D'[\mathrm{Vec}(Y)]$$

的解为

$$\mathrm{Vec}(\hat{\beta}) = (D'D)^{-1}D'\mathrm{Vec}(Y).$$

因

$$Q[\mathrm{Vec}(\beta)] = [\mathrm{Vec}(Y) - D\mathrm{Vec}(\beta)]'[\mathrm{Vec}(Y) - D\mathrm{Vec}(\beta)]$$
$$= [\mathrm{Vec}(Y) - D\mathrm{Vec}(\hat{\beta})]'[\mathrm{Vec}(Y) - D\mathrm{Vec}(\hat{\beta})]$$
$$+ [\mathrm{Vec}(\hat{\beta}) - \mathrm{Vec}(\beta)]'D'D[\mathrm{Vec}(\hat{\beta}) - \mathrm{Vec}(\beta)],$$

所以

$$Q[\mathrm{Vec}(\hat{\beta})] = \min_{\text{一切}\mathrm{Vec}(\beta)} Q[\mathrm{Vec}(\beta)].$$

这说明正规方程的解 $\mathrm{Vec}(\hat{\beta})$是参数向量 $\mathrm{Vec}(\beta)$的最小二乘估计.

又

$$\mathrm{Vec}(\hat{\beta}) = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (D'D)^{-1}D'\mathrm{Vec}(Y)$$

$$= \begin{bmatrix} C'C & O & \cdots & O \\ O & C'C & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & C'C \end{bmatrix}^{-1} \begin{bmatrix} C' & O & \cdots & O \\ O & C' & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & C' \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix}$$

$$= \begin{bmatrix} (C'C)^{-1}C'Y_1 \\ (C'C)^{-1}C'Y_2 \\ \vdots \\ (C'C)^{-1}C'Y_p \end{bmatrix},$$

即 $\hat{\beta}_j=(C'C)^{-1}C'Y_j(j=1,2,\cdots,p)$，其中 Y_j 是第 j 个因变量的 n 次观测值. 可见，在模型(4.3.3)下参数的最小二乘估计与 § 4.1 中一个因变量的回归模型(4.1.2)的结果完全相同. 也就是说，在多对多的回归模型下，回归系数矩阵的最小二乘估计等于对各因变量分别建立回归模型时所得的估计量. 这两者的一致性在某种意义下降低了多对多回归模型的地位，因此，必须设法提取其他信息，才能显示多对多回归模型的优越性，这将在 § 4.4 介绍.

为了方便，下面把在“拉直”后的模型(4.3.3)下的正规方程及回归系数的估计“压缩”为矩阵形式.

正规方程

$$\begin{bmatrix} C'C & O & \cdots & O \\ O & C'C & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & C'C \end{bmatrix} \mathrm{Vec}(\beta) = \begin{bmatrix} C' & O & \cdots & O \\ O & C' & \cdots & O \\ \vdots & \vdots & & \vdots \\ O & O & \cdots & C' \end{bmatrix} \mathrm{Vec}(Y)$$

等价于 $C'C(\beta_1,\beta_2,\cdots,\beta_p)=C'(Y_1,Y_2,\cdots,Y_p)$，即

$$C'C\beta = C'Y \quad (\text{其中 } C=(\mathbf{1}_n \vdots X) \text{ 为 } n\times(m+1) \text{ 矩阵}).$$

把 $(m+1)\times p$ 的参数矩阵 β 分为两块：$b_{(0)}$ 为 $1\times p$ 矩阵，B 为 $m\times p$ 矩阵，则参数矩阵 β 的估计可表为

$$\hat{\beta} = \begin{bmatrix} \hat{b}_{(0)} \\ \hat{B} \end{bmatrix} = (C'C)^{-1}C'Y.$$

由分块求逆公式有(假定 $\mathrm{rank}(C)=m+1$)

$$(C'C)^{-1} = \begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'X \\ X'\mathbf{1}_n & X'X \end{bmatrix}^{-1} = \begin{bmatrix} \dfrac{1}{n} + \overline{X}'L_{XX}^{-1}\overline{X} & -\overline{X}'L_{XX}^{-1} \\ -L_{XX}^{-1}\overline{X} & L_{XX}^{-1} \end{bmatrix},$$

其中

$$L_{XX} = X'\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)X, \quad \overline{X} = \frac{1}{n}X'\mathbf{1}_n = (\bar{x}_1,\cdots,\bar{x}_m)'.$$

记

$$\overline{Y} = \frac{1}{n}Y'\mathbf{1}_n = (\overline{y}_1, \cdots, \overline{y}_p)',$$

$$L_{YY} = Y'\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)Y = Y'\left(I_n - \frac{1}{n}J\right)Y,$$

$$L_{XY} = X'\left(I_n - \frac{1}{n}J\right)Y = X'Y - n\overline{X}\,\overline{Y}' = L_{YX}',$$

其中

$$J = \mathbf{1}_n\mathbf{1}_n' = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}.$$

于是分块估计的表达式为

$$\begin{bmatrix} \hat{b}_{(0)} \\ \hat{B} \end{bmatrix} = \begin{bmatrix} \left(\frac{1}{n} + \overline{X}'L_{XX}^{-1}\overline{X}\right)n\overline{Y}' - \overline{X}'L_{XX}^{-1}X'Y \\ -L_{XX}^{-1}\overline{X}(n\overline{Y})' + L_{XX}^{-1}X'Y \end{bmatrix}$$

$$= \begin{bmatrix} \overline{Y}' - \overline{X}'L_{XX}^{-1}L_{XY} \\ L_{XX}^{-1}L_{XY} \end{bmatrix},$$

而且称 $\hat{b}_{(0)}, \hat{B}$ 满足的方程

$$\begin{cases} L_{XX}\hat{B} = L_{XY}, \\ \hat{b}_{(0)} = \overline{Y}' - \overline{X}'\hat{B} \end{cases} \tag{4.3.4}$$

为正规方程.

3. 参数矩阵 Σ 的估计

以上求得 β 的最小二乘估计量 $\hat{\beta} \xlongequal{\text{def}} (b_{ij})_{(m+1)\times p}$，即得 p 个因变量的回归方程：

$$\hat{Y}_j = b_{0j} + b_{1j}x_1 + \cdots + b_{mj}x_m \quad (j = 1, 2, \cdots, p).$$

于是得 n 组资料的预报值为

$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1p} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2p} \\ \vdots & \vdots & & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \cdots & \hat{y}_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} b_{01} & b_{02} & \cdots & b_{0p} \\ b_{11} & b_{12} & \cdots & b_{1p} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix}$$

$$= C\hat{\beta},$$

即

$$\hat{Y}=(\mathbf{1}_n X)\begin{bmatrix}\hat{b}_{(0)}\\ \hat{B}\end{bmatrix}=\mathbf{1}_n\hat{b}_{(0)}+X\hat{B}$$

$$=\mathbf{1}_n\overline{Y}'+\left(I_n-\frac{1}{n}J\right)X\hat{B}.$$

实测值 Y 与预报值 $\hat{Y}$ 之差 $Y-\hat{Y}$ 就称为**残差**. 可以用它构造 Σ(误差向量 $\varepsilon_{(i)}$的协方差阵)的估计量. 残差

$$Y-\hat{Y}=Y-\mathbf{1}_n\overline{Y}'-\left(I_n-\frac{1}{n}J\right)X\hat{B}$$

$$=\left(I_n-\frac{1}{n}J\right)Y-\left(I_n-\frac{1}{n}J\right)XL_{XX}^{-1}L_{XY}$$

$$=\left(I_n-\frac{1}{n}J\right)(Y-XL_{XX}^{-1}L_{XY}),$$

或

$$Y-\hat{Y}=Y-C\hat{\beta}=(I_n-C(C'C)^{-1}C')Y$$

$$=(I_n-H)Y.$$

令 $Q=(Y-\hat{Y})'(Y-\hat{Y})$为 $p\times p$ 矩阵. 当 $p=1$ 时(即多元线性回归模型),数值 $Q=\sum_{j=1}^{n}(y_j-\hat{y}_j)^2$ 称为**残差平方和**(或剩余平方和). 对一般的 p,Q 是 $p\times p$ 矩阵,它是残差平方和的推广,称为**残差阵**. Q 有以下计算公式:

$$Q=(Y-\hat{Y})'(Y-\hat{Y})=L_{YY}-L_{YX}L_{XX}^{-1}L_{XY}$$

$$=Y'(I_n-H)Y. \tag{4.3.5}$$

很自然地,我们用残差阵 Q 作为随机误差向量 $\varepsilon_{(i)}$的协方差阵 Σ 的估计,考虑到无偏性,常取 Σ 的估计为

$$\hat{\Sigma}=\frac{1}{n-m-1}Q.$$

4. $\hat{\beta}$,$\hat{\Sigma}$ 的统计性质

定理 4.3.1　在多对多回归模型(4.3.2)下

(1) $\hat{\beta}=\begin{bmatrix}\hat{b}_{(0)}\\ \hat{B}\end{bmatrix}$是 β 的无偏估计量;

(2) $\hat{\Sigma}=\frac{1}{n-m-1}Q$ 是 Σ 的无偏估计量.

证明 以下只证明(2),因

$$Q=Y'(I_n-H)Y\xlongequal{\text{def}}Y'PY,$$

其中 $P=I_n-H$ 为 $n\times n$ 对称幂等矩阵. 因而

$$\begin{aligned}\mathrm{E}(Q)&=\mathrm{E}(Y'PY)\\&=\mathrm{E}[(Y-\mathrm{E}(Y)+\mathrm{E}(Y))'P(Y-\mathrm{E}(Y)+\mathrm{E}(Y))]\\&=\mathrm{E}(E'PE)+\mathrm{E}(Y)'P\mathrm{E}(Y)\\&=\mathrm{E}(E'PE)\quad(\text{因 }\mathrm{E}(Y)=C\beta,\text{而 }PC=O)\\&=\mathrm{E}\left[\begin{bmatrix}\varepsilon_1'\\\vdots\\\varepsilon_p'\end{bmatrix}P[\varepsilon_1,\cdots,\varepsilon_p]\right]\\&=(\mathrm{E}[\varepsilon_i'P\varepsilon_j])_{(p\times p)}\quad(i,j=1,2,\cdots,p).\end{aligned}$$

在模型(4.3.1)下, $\varepsilon_{(i)}=(\varepsilon_{i1},\cdots,\varepsilon_{ip})'\sim N_p(0,\Sigma)$ $(i=1,2,\cdots,n)$相互独立,且 $\Sigma=(\sigma_{kl})_{p\times p}$,即

$$\mathrm{Cov}(\varepsilon_{ik},\varepsilon_{il})=\mathrm{E}(\varepsilon_{ik}\varepsilon_{il})=\sigma_{kl}\quad(k,l=1,2,\cdots,p).$$

于是

$$\begin{aligned}\mathrm{E}(\varepsilon_i'P\varepsilon_j)&=\mathrm{E}\left[(\varepsilon_{1i},\cdots,\varepsilon_{ni})P\begin{bmatrix}\varepsilon_{1j}\\\vdots\\\varepsilon_{nj}\end{bmatrix}\right]\\&=\mathrm{E}\left[\mathrm{tr}\left[P\begin{bmatrix}\varepsilon_{1j}\\\vdots\\\varepsilon_{nj}\end{bmatrix}(\varepsilon_{1i},\cdots,\varepsilon_{ni})\right]\right]\\&=\mathrm{tr}\left[P\mathrm{E}\begin{bmatrix}\varepsilon_{1j}\varepsilon_{1i}&\cdots&\varepsilon_{1j}\varepsilon_{ni}\\\vdots&&\vdots\\\varepsilon_{nj}\varepsilon_{1i}&\cdots&\varepsilon_{nj}\varepsilon_{ni}\end{bmatrix}\right]\quad(\text{因 }\varepsilon_{(i)}\text{ 相互独立})\\&=\mathrm{tr}(P\mathrm{diag}(\sigma_{ij},\cdots,\sigma_{ij}))\\&=\sigma_{ij}\ \mathrm{tr}(P)=\sigma_{ij}(n-m-1).\end{aligned}$$

故有

$$\mathrm{E}\left(\frac{1}{n-m-1}Q\right)=\Sigma. \qquad \text{(证毕)}$$

引理 4.3.1　在模型(4.3.2)下,记 $\hat{\beta}=(b_{ij})_{(m+1)\times p}$,则

$$\mathrm{Cov}(b_{ik},b_{jl})=\sigma_{kl}e_{i+1}'(C'C)^{-1}e_{j+1}$$
$$(i,j=0,1,2,\cdots,m;\ k,l=1,2,\cdots,p),$$

其中 e_i 是第 i 个分量为 1,其余分量全为 0 的单位向量.

证明　因 $b_{ik}=e_{i+1}'\hat{\beta}e_k$, $b_{jl}=e_{j+1}'\hat{\beta}e_l$,故有

$$\begin{aligned}\mathrm{Cov}(b_{ik},b_{jl})&=\mathrm{E}[(e_{i+1}'\hat{\beta}e_k-\mathrm{E}(e_{i+1}'\hat{\beta}e_k))\\&\quad\cdot(e_{j+1}'\hat{\beta}e_l-\mathrm{E}(e_{j+1}'\hat{\beta}e_l))']\\&=\mathrm{E}[e_{i+1}'(\hat{\beta}-\mathrm{E}(\hat{\beta}))e_k\cdot e_l'(\hat{\beta}-\mathrm{E}(\hat{\beta}))'e_{j+1}]\\&=e_{i+1}'(C'C)^{-1}C'\cdot\mathrm{E}[(Y-\mathrm{E}(Y))e_ke_l'(Y-\mathrm{E}(Y))']\\&\quad\cdot C(C'C)^{-1}e_{j+1}\\&=e_{i+1}'(C'C)^{-1}C'(\mathrm{E}(\varepsilon_k\varepsilon_l'))C(C'C)^{-1}e_{j+1}\\&=e_{i+1}'(C'C)^{-1}C'\mathrm{diag}(\sigma_{kl},\cdots,\sigma_{kl})C(C'C)^{-1}e_{j+1}\\&=\sigma_{kl}e_{i+1}'(C'C)^{-1}e_{j+1}.\end{aligned} \qquad \text{(证毕)}$$

引理 4.3.2　在模型(4.3.2)下,记 $\hat{\beta}=\begin{bmatrix}\hat{b}_{(0)}\\ \hat{B}\end{bmatrix}\begin{matrix}1\\ m\end{matrix}$, $L_{XX}^{-1}=(l^{ij})$,而

$$\hat{B}=(b_{ij})_{m\times p}\xlongequal{\text{def}}(\hat{b}_1,\cdots,\hat{b}_p)\xlongequal{\text{或 def}}\begin{bmatrix}\hat{b}_{(1)}\\ \vdots\\ \hat{b}_{(m)}\end{bmatrix},$$

则

(1) $D(\hat{b}_{(0)}')=\left(\frac{1}{n}+\overline{X}'L_{XX}^{-1}\overline{X}\right)\Sigma$;

(2) $\mathrm{COV}(\hat{b}_{(i)}',\hat{b}_{(j)}')=l^{ij}\Sigma$ $(i,j=1,2,\cdots,m)$;

(3) $\mathrm{COV}(\hat{b}_i,\hat{b}_j)=\sigma_{ij}L_{XX}^{-1}(i,j=1,2,\cdots,p)$.

证明　由协方差阵的定义及引理 4.3.1 即得以上结论. (证毕)

定理 4.3.2　在模型(4.3.1)下,设 $n>m+1$,

$$\mathrm{rank}(C)=\mathrm{rank}(\mathbf{1}_n\vdots X)=m+1,$$

则

(1) $\hat{\beta}$ 遵从矩阵正态分布;

(2) $Q\sim W_p(n-m-1,\Sigma)$;

(3) $\hat{\beta}$ 与 Q 相互独立.

证明 (1) 我们有

$$\hat{\beta}=(C'C)^{-1}C'Y=(C'C)^{-1}C'(C\beta+E)=\beta+(C'C)^{-1}C'E.$$

由模型(4.3.1)的假定可知:

$$E\sim N_{n\times p}(O,I_n\otimes\Sigma)\quad(\text{其中}O\text{为}n\times p\text{零矩阵})$$

$$\Longleftrightarrow \mathrm{Vec}(E')\sim N_{np}(0,I_n\otimes\Sigma)\quad(\text{其中}0\text{为}np\text{维零向量}).$$

利用随机阵正态分布的性质即得:

$$\hat{\beta}\sim N_{(m+1)\times p}(\beta,(C'C)^{-1}C'\cdot[(C'C)^{-1}C']'\otimes I_p\Sigma I_p'),$$

即

$$\hat{\beta}\sim N_{(m+1)\times p}(\beta,(C'C)^{-1}\otimes\Sigma)$$

$$\Longleftrightarrow \mathrm{Vec}(\hat{\beta}')\sim N_{(m+1)p}(\mathrm{Vec}(\beta'),(C'C)^{-1}\otimes\Sigma).$$

(2) 由 Q 的计算公式(4.3.5)有:

$$Q=Y'(I_n-H)Y\overset{\text{def}}{=\!=}Y'PY,$$

又知 $Y\sim N_{n\times p}(C\beta,I_n\otimes\Sigma)$, $P=I_n-H=I_n-C(C'C)^{-1}C'$ 是对称幂等矩阵,且 $\mathrm{rank}(P)=n-m-1$,由威沙特分布的性质 7 得

$$Q=Y'PY\sim W_p(n-m-1,\Sigma,\Delta),$$

其中 $\Delta=[C\beta]'PC\beta=O$ (因 $PC=O$). 所以

$$Q\sim W_p(n-m-1,\Sigma).$$

(3) 因 $\hat{\beta}=(C'C)^{-1}C'Y$,考虑

$$\hat{\beta}'(C'C)\hat{\beta}=Y'C(C'C)^{-1}(C'C)(C'C)^{-1}C'Y=Y'HY,$$

其中 $H=C(C'C)^{-1}C'$ 为对称幂等矩阵,且有 $PH=PC(C'C)^{-1}C'=O$,由威沙特分布的性质 8 可知

$$Q=Y'PY\quad\text{与}\quad\hat{\beta}'(C'C)\hat{\beta}=Y'HY\text{相互独立},$$

故而有 Q 与 $\hat{\beta}$ 相互独立. (证毕)

定理 4.3.3 在模型(4.3.1)下,记号同引理 4.3.2,则

(1) $\hat{b}_{(0)}'\sim N_p\left(b_{(0)}',\left(\frac{1}{n}+\overline{X}'L_{XX}^{-1}\overline{X}\right)\Sigma\right)$;

(2) $\hat{b}_{(i)}'\sim N_p(b_{(i)}',l^{ii}\Sigma)$ $(i=1,2,\cdots,m)$;

(3) $\hat{b}_j\sim N_m(b_j,\sigma_{jj}L_{XX}^{-1})$ $(j=1,2,\cdots,p)$.

证明　由引理 4.3.2 及定理 4.3.2 即得以上结论.　**(证毕)**

二、回归系数的显著性检验

在多因变量的多元线性回归中,同样要考查某个自变量 x_i 对 p 个因变量的影响是否显著的问题. 若 x_i 对 p 个因变量的作用不显著,那么在模型(4.3.1)中 x_i 的回归系数 $\beta_{(i)}=0_p$. 判断变量 x_i 对 p 个因变量作用是否显著的问题,即要检验假设 $H_0^{(i)}$: $\beta_{(i)}=0_p(i=1,2,\cdots,m)$.

更一般地,可同时考查几个自变量对 p 个因变量是否有影响的问题,即考虑模型:

$$\begin{cases} Y=(\mathbf{1}_n \mid X_1)\begin{bmatrix} b_{(0)} \\ B_1 \end{bmatrix}+X_2B_2+E, \\ \varepsilon_{(i)} \sim N_p(0,\Sigma)\ (i=1,2,\cdots,n)\ \text{相互独立}, \end{cases} \tag{4.3.6}$$

其中 $C=(\mathbf{1}_n \mid X)=(\mathbf{1}_n \mid X_1 \mid X_2)$, X_1 为 $n\times m_1$ 给定矩阵,X_2 为 $n\times m_2$ 给定矩阵,且 $m_1+m_2=m$.

记 $B=\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$,其中 B_1 为 $m_1\times p$ 参数矩阵,B_2 为 $m_2\times p$ 参数矩阵 $(m_1+m_2=m)$,且 $\operatorname{rank}(C)=\operatorname{rank}(\mathbf{1}_n \mid X_1 \mid X_2)=m+1$. 检验假设 H_0: $B_2=O$. 这就是要检验一部分变量(即 $x_{m_1+1},\cdots,x_m$)是否对 p 个因变量没有显著影响.

1. 检验 $H_0^{(i)}$: $\beta_{(i)}=0_p$

首先来讨论某个自变量 x_i 对 $Y_1,\cdots,Y_p$ 的作用是否显著问题. 利用定理 4.3.2 和定理 4.3.3 即可得出检验 $H_0^{(i)}$ 的统计量. 由定理 4.3.3 知

$$\hat{\beta}_{(i)}=\hat{b}'_{(i)} \sim N_p(\beta_{(i)},l^{ii}\Sigma),$$

记 $E_i=\dfrac{1}{\sqrt{l^{ii}}}I_p$,则在 $H_0^{(i)}$ 下

$$E_i\,\hat{\beta}_{(i)} \sim N_p(0,\Sigma).$$

由定理 4.3.2 及 Q 的计算公式(4.3.5)知

$$Q = L_{YY} - L_{YX}L_{XX}^{-1}L_{XY} \sim W_p(n-m-1,\Sigma),$$

且 Q 与 $\hat{\beta}_{(i)}$ 相互独立. 由第三章 T^2 的定义 3.1.5 知,统计量

$$\begin{aligned} T^2 &= (n-m-1)(E_i\hat{\beta}_{(i)})'Q^{-1}(E_i\hat{\beta}_{(i)}) \\ &= (n-m-1)\hat{\beta}'_{(i)}Q^{-1}\hat{\beta}_{(i)}/l^{ii} \\ &\sim T^2(p,n-m-1) \quad (\text{在 } H_0^{(i)} \text{ 成立时}). \end{aligned}$$

于是检验统计量为

$$F = \frac{(n-m-1)-p+1}{(n-m-1)p}T^2 = \frac{n-m-p}{p}\frac{\hat{\beta}'_{(i)}Q^{-1}\hat{\beta}_{(i)}}{l^{ii}}$$
$$\sim F(p,n-m-p) \quad (\text{在 } H_0^{(i)} \text{ 成立时}).$$

显然,当 $H_0^{(i)}$ 成立时有 $\hat{\beta}_{(i)} \approx 0_p$,于是数值 $V_i = \dfrac{\hat{\beta}'_{(i)}Q^{-1}\hat{\beta}_{(i)}}{l^{ii}}$ 应较小,常称 V_i 为变量 x_i 对 p 个因变量 $Y_1,\cdots,Y_p$ 的"贡献". 当 $p=1$ 时,$V_i = \hat{\beta}_{(i)}^2/l^{ii}Q = P_i/Q$ (P_i 为 x_i 的偏回归平方和).

给定显著性水平 α,由样本观测数据计算 V_i 及 $f_i = \dfrac{n-m-p}{p}V_i$,并计算显著性概率值($p$ 值)$=P\{F \geqslant f_i\}$,若 $p<\alpha$,则否定 $H_0^{(i)}$,表示 x_i 对 p 个因变量的作用显著;否则,x_i 对 p 个因变量的作用不显著.

2. 检验 H_0: $B_2=O$

在模型(4.3.6)下,记

$$C_1 = (\mathbf{1}_n \vdots X_1), \quad C = (\mathbf{1}_n \vdots X_1 \vdots X_2) = (C_1 \vdots X_2),$$

残差阵

$$Q = Y'(I_n - C(C'C)^{-1}C')Y = Y'(I_n - H)Y.$$

当 H_0 成立(即 $B_2=O$)时,模型变为(设 $\operatorname{rank}(\mathbf{1}_n \vdots X_1)=m_1+1$)

$$\begin{cases} Y = (\mathbf{1}_n \vdots X_1)\begin{bmatrix} b_{(0)} \\ B_1 \end{bmatrix} + E = C_1\beta(1) + E, \\ \varepsilon_{(i)} \sim N_p(0,\Sigma)\ (i=1,2,\cdots,n) \text{ 相互独立}, \end{cases} \tag{4.3.7}$$

其相应的残差阵为

$$Q_1 = Y'(I_n - C_1(C_1'C_1)^{-1}C_1')Y = Y'(I_n - H_1)Y,$$

其中 $H_1 = C_1(C_1'C_1)^{-1}C_1'$.

首先计算 $Q_1 - Q$ 的表达式. 因 $C=(C_1 \vdots X_2)$,记

$$D = X_2'(I_n - H_1)X_2,$$

故有

$$(C'C)^{-1} = \begin{bmatrix} C_1'C_1 & C_1'X_2 \\ X_2'C_1 & X_2'X_2 \end{bmatrix}^{-1} = \begin{bmatrix} (C_1'C_1)^{-1} & O \\ O & O \end{bmatrix} + \begin{bmatrix} (C_1'C_1)^{-1}C_1'X_2 \\ -I_{m_2} \end{bmatrix} D^{-1}(X_2'C_1(C_1'C_1)^{-1} \vdots -I_{m_2}).$$

因此

$$Q = Y'\left(I_n - (C_1 \vdots X_2)(C'C)^{-1}\begin{bmatrix} C_1' \\ X_2' \end{bmatrix}\right)Y$$

$$= Y'(I_n - C_1(C_1'C_1)^{-1}C_1')Y - Y'(I_n - C_1(C_1'C_1)^{-1}C_1')X_2D^{-1}X_2'(I_n - C_1(C_1'C_1)^{-1}C_1')Y,$$

即得　$Q_1 - Q = Y'(I_n - H_1)X_2D^{-1}X_2'(I_n - H_1)Y.$

另一方面，

$$\hat{\beta} = (C'C)^{-1}C'Y = \begin{bmatrix} (C_1'C_1)^{-1}C_1'Y - (C_1'C_1)^{-1}C_1'X_2D^{-1}X_2'(I_n - H_1)Y \\ D^{-1}X_2'(I_n - H_1)Y \end{bmatrix} \begin{matrix} m_1 + 1 \\ m_2 \end{matrix}.$$

于是 $\hat{B}_2 = D^{-1}X_2'(I_n - H_1)Y.$

所以

$$Q_1 - Q = Y'(I_n - H_1)X_2D^{-1}X_2'(I_n - H_1)Y = \hat{B}_2'D\hat{B}_2 = \hat{B}_2'X_2'(I_n - H_1)X_2\hat{B}_2. \quad (4.3.8)$$

定理 4.3.4　在模型(4.3.1)下，有

(1) $Q \sim W_p(n-m-1, \Sigma)$；

(2) 在模型(4.3.7)下(即 H_0 成立时)，$Q_1 - Q \sim W_p(m_2, \Sigma)$；

(3) Q 与 $Q_1 - Q$ 相互独立.

证明　(1) 定理 4.3.2 已证明.

(2) 因 $Q_1 - Q = Y'(I_n - H_1)X_2D^{-1}X_2'(I_n - H_1)Y$，记 $R = (I_n - H_1)X_2$ 为 $n \times m_2$ 矩阵，$D = X_2'(I_n - H_1)X_2$ 为 m_2 阶矩阵，其中 $H_1 =$

$C_1(C_1'C_1)^{-1}C_1'$. 则

$$Q_1 - Q = Y'RD^{-1}R'Y \xlongequal{\text{def}} Y'BY.$$

在模型(4.3.7)下(即 H_0 成立时),

$$Y \sim N_{n\times p}(C_1\beta(1), I_n \otimes \Sigma),$$

且容易验证 $B=RD^{-1}R'$ 是对称幂等矩阵. 而

$$R'R = X_2'(I_n - H_1)'(I_n - H_1)X_2 = X_2'(I_n - H_1)X_2 = D,$$

故

$$\begin{aligned}\operatorname{rank}(B) &= \operatorname{tr}(B) = \operatorname{tr}(RD^{-1}R') = \operatorname{tr}(D^{-1}R'R)\\ &= \operatorname{tr}(D^{-1}D) = \operatorname{tr}(I_{m_2}) = m_2.\end{aligned}$$

由威沙特分布的性质 7 可得

$$Q_1 - Q = Y'BY \sim W_p(m_2, \Sigma, \Delta),$$

其中 $\Delta=[C_1\beta(1)]'BC_1\beta(1)=[C_1\beta(1)]'RD^{-1}R'C_1\beta(1)=O$,这是因为 $R'C_1=X_2'(I_n-H_1)C_1=O$,所以 $Q_1 - Q \sim W_p(m_2, \Sigma)$.

(3) 下面来证明 Q 与 Q_1-Q 相互独立.

已知 $Q_1-Q=Y'BY$, $Q_1=Y'(I_n-H_1)Y$,从而

$$\begin{aligned}Q &= Y'PY = Q_1 - (Q_1 - Q) = Y'(I_n - H_1)Y - Y'BY\\ &= Y'(I_n - H_1 - B)Y,\end{aligned}$$

故有

$$PB = (I_n - H_1 - B)B = -H_1B = -C_1(C_1'C_1)^{-1}C_1'B = O$$

(因 $C_1'R=O$). 由威沙特分布的性质 8 可知

$Q = Y'PY$ 与 $Q_1 - Q = Y'BY$ 相互独立. **(证毕)**

下面来导出检验 H_0 的似然比统计量.

在模型(4.3.6)下,似然函数(即 $\operatorname{Vec}(Y')$的联合密度函数)为

$$\begin{aligned}L(\beta,\Sigma) &= \left(\frac{1}{2\pi}\right)^{np/2}\frac{1}{|\Sigma|^{n/2}}\\ &\quad\cdot\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(Y_{(i)} - \beta'c_{(i)})'\Sigma^{-1}(Y_{(i)} - \beta'c_{(i)})\right]\\ &= \left(\frac{1}{2\pi}\right)^{np/2}\frac{1}{|\Sigma|^{n/2}}\exp\left[-\frac{1}{2}\operatorname{tr}\Sigma^{-1}\right.\end{aligned}$$

$$\cdot \sum_{i=1}^{n}(Y_{(i)}-\beta' c_{(i)})(Y_{(i)}-\beta' c_{(i)})'\Big]$$

$$=(2\pi)^{-np/2}(|\Sigma|)^{-n/2}\exp\Big[-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(Y-C\beta)'(Y-C\beta)\Big].$$

当 $\beta=\hat{\beta}$，$\Sigma=\frac{1}{n}Q$ 时，

$$L\Big(\hat{\beta},\frac{Q}{n}\Big)=\max_{\beta,\Sigma}L(\beta,\Sigma)$$

$$=(2\pi)^{-np/2}\Big(\Big|\frac{Q}{n}\Big|\Big)^{-n/2}\exp\Big[-\frac{1}{2}\mathrm{tr}\Big(\frac{1}{n}Q\Big)^{-1}Q\Big]$$

$$=(2\pi)^{-np/2}\Big(\Big|\frac{Q}{n}\Big|\Big)^{-n/2}\exp\Big[-\frac{1}{2}np\Big].$$

在模型(4.3.7)下，似然函数

$$L(\beta(1),\Sigma)=(2\pi)^{-np/2}(|\Sigma|)^{-n/2}$$

$$\cdot\exp\Big[-\frac{1}{2}\mathrm{tr}\Sigma^{-1}(Y-C_1\beta(1))'(Y-C_1\beta(1))\Big],$$

其中 $\beta(1)=\begin{bmatrix}b_{(0)}\\B_1\end{bmatrix}$，$C_1=(\mathbf{1}_n\vdots X_1)$；且当 $\beta(1)=\hat{\beta}(1)$，$\Sigma=\frac{1}{n}Q_1$ 时，

$$L\Big(\hat{\beta}(1),\frac{Q_1}{n}\Big)=\max_{\beta(1),\Sigma}L(\beta(1),\Sigma)$$

$$=(2\pi)^{-np/2}\Big(\Big|\frac{Q_1}{n}\Big|\Big)^{-n/2}\exp\Big[-\frac{1}{2}np\Big].$$

因此似然比统计量为

$$\lambda=\frac{\max L(\beta(1),\Sigma)}{\max L(\beta,\Sigma)}=\frac{|Q_1/n|^{-n/2}}{|Q/n|^{-n/2}}$$

$$=\frac{|Q_1|^{-n/2}}{|Q|^{-n/2}}=\Big(\frac{|Q|}{|Q+(Q_1-Q)|}\Big)^{n/2},$$

等价于

$$U=\frac{|Q|}{|Q+(Q_1-Q)|}=\frac{|Q|}{|Q+\hat{B}_2'D\hat{B}_2|}.$$

在 H_0 成立时，$Q_1-Q=\hat{B}_2'D\hat{B}_2\sim W_p(m_2,\Sigma)$，又因 $Q\sim W_p(n-m-1,\Sigma)$，且 Q 与 Q_1-Q 相互独立，由第三章定义 3.1.7 知

$$U\sim\Lambda(p,n-m-1,m_2).$$

直观地看，若 H_0 成立，则 U 值应近似等于 1；若 U 值太小，则应否定假设 H_0. 对给定的显著性水平 α，由样本资料计算 U 值为 u，利用检验统计量 U 的分布，计算显著性概率值(p 值)$=P\{U\leqslant u\}$. 当 $p<\alpha$ 时，否定 H_0，即认为 m_2 个自变量 $x_{m_1+1},\cdots,x_m$ 对 p 个因变量的作用显著；当 $p\geqslant\alpha$ 时，H_0 相容，即认为 m_2 个自变量 $x_{m_1+1},\cdots,x_m$ 对 p 个因变量的作用不显著.

当 $m_2=1$ 时，X_2 是 n 维向量，$D=X_2'(I_n-H_1)X_2$ 是一数值，记为 d，而 $\hat{B}_2=\hat{\beta}'_{(m)}$ 为 $1\times p$ 矩阵，所以

$$U=\frac{|Q|}{|Q+\hat{\beta}_{(m)}d\,\hat{\beta}'_{(m)}|}.$$

利用分块求行列式的公式有：

$$\begin{vmatrix}1 & -d\,\hat{\beta}_{(m)}\\ \hat{\beta}_{(m)} & Q\end{vmatrix}=|Q+d\,\hat{\beta}_{(m)}\,\hat{\beta}'_{(m)}|$$

$$=|Q||1+d\,\hat{\beta}'_{(m)}Q^{-1}\,\hat{\beta}_{(m)}|,$$

因此

$$U=\frac{1}{1+d\,\hat{\beta}'_{(m)}Q^{-1}\,\hat{\beta}_{(m)}}. \tag{4.3.9}$$

另一方面，当 $m_2=1$ 时，由第三章 §3.1 中的结论知

$$\Lambda(p,n-m-1,1)=\frac{1}{1+\dfrac{1}{n-m-1}T^2(p,n-m-1)}. \tag{4.3.10}$$

比较(4.3.9)和(4.3.10)式得

$$T^2(p,n-m-1)=(n-m-1)d\,\hat{\beta}'_{(m)}Q^{-1}\,\hat{\beta}_{(m)}$$

$$=(n-m-1)\frac{1-U}{U},$$

即 U 统计量可化为 T^2 统计量；再根据第三章的有关定理知

$$F=\frac{(n-m-1)-p+1}{(n-m-1)p}T^2(p,n-m-1)$$

$$\sim F(p,n-m-p),$$

即

$$F=\frac{n-m-p}{p}\frac{1-U}{U}\sim F(p,n-m-p).$$

这表明 U 统计量在筛选变量过程中是很重要的统计量.

例 4.3.1　设发电量 Y_1,工业总产值 Y_2 与钢材产量 x_1,水泥产量 x_2,机械工业总产值 x_3,棉纱产量 x_4,机制纸产量 x_5 之间有线性相关关系. 现收集了 1949 到 1978 年共 30 年的数据(见表 4.2). 试用 REG 过程求出 Y_1,Y_2 与 x_1,x_2,x_3,x_4,x_5 的关系式.

表 4.2　发电量与经济发展数据

年	x_1	x_2	x_3	x_4	x_5	Y_1	Y_2
1949	0.9	0.8	0.14	6.63	0.24	1.47	7.31
1950	1.0	2.1	0.15	7.07	0.46	1.25	7.42
1951	2.9	6.3	0.33	7.60	1.02	2.05	11.13
1952	5.0	4.4	0.78	12.88	1.61	2.49	16.08
1953	8.2	13.3	1.18	15.86	1.63	3.16	22.86
1954	13.1	16.8	1.56	18.79	1.93	3.87	29.52
1955	23.8	17.8	2.11	14.63	2.31	4.50	34.54
1956	34.8	27.8	3.09	19.79	3.32	6.09	41.22
1957	35.4	22.1	3.58	16.50	4.44	6.78	47.54
1958	47.0	32.2	7.31	26.22	7.18	10.73	60.00
1959	62.6	33.2	9.61	28.00	8.77	17.65	78.00
1960	68.0	55.6	12.85	27.56	9.89	26.84	96.20
1961	35.3	24.4	6.76	10.95	5.58	24.20	52.37
1962	31.3	17.9	5.08	10.15	6.03	20.08	37.77
1963	35.2	24.8	5.54	14.23	7.18	19.28	40.07
1964	45.3	37.8	7.14	20.38	8.80	22.89	50.36
1965	49.5	78.8	11.20	26.56	10.45	28.94	65.33
1966	59.7	101.6	15.89	33.18	12.51	39.05	83.64
1967	47.8	74.9	10.86	23.90	11.42	39.09	68.16
1968	17.7	40.2	5.10	17.56	9.03	26.81	41.64
1969	36.0	73.3	13.14	27.20	8.05	37.19	67.30
1970	62.0	138.6	25.54	36.28	10.30	54.09	103.57
1971	97.0	247.0	31.31	41.53	14.18	77.39	135.80
1972	95.2	270.0	28.79	40.24	15.19	84.02	118.10
1973	118.4	233.5	28.03	38.20	15.77	88.39	119.62
1974	99.9	205.0	26.50	31.54	12.29	86.32	112.39
1975	151.0	288.0	38.61	46.87	17.36	107.94	144.41
1976	108.0	262.2	31.46	38.62	15.10	102.76	130.66
1977	162.5	358.6	46.21	52.48	20.48	118.84	175.10
1978	238.2	454.8	55.86	55.96	26.40	139.30	214.44

解 此例因变量个数 $p=2$，自变量个数 $m=5$，观测数据 $n=30$. 仍使用 REG 过程来完成多因变量的多元线性回归计算.

输出的结果中给出两个因变量的回归方程如下：

$$\begin{cases}\hat{Y}_1 = 8.9911 - 0.1675x_1 + 0.1724x_2 + 1.7036x_3 - 0.7622x_4 + 1.9756x_5,\\ \hat{Y}_2 = 4.3224 + 0.2757x_1 - 0.1341x_2 + 2.2313x_3 + 0.9880x_4 + 1.0502x_5.\end{cases}$$

两回归方程经检验都是高度显著的($p<0.0001$)；Y_1 与 $x_1,\cdots,x_5$ 的回归系数在显著性水平 $\alpha=0.10$ 下也都是显著的，复相关系数 $R_1=0.9901$（决定系数 $R_1^2=0.9804$）；误差标准差（Root MSE）$s_1=6.25355$.

Y_2 与 $x_1,\cdots,x_5$ 的回归系数除 x_5 外在显著性水平 $\alpha=0.05$ 下也都是显著的，复相关系数 $R_2=0.9933$（决定系数 $R_2^2=0.9867$）；误差标准差（Root MSE）$s_2=6.56271$.

使用 REG 过程还可以完成几个自变量对于因变量的作用是否显著的检验. 如输出 4.3.1 给出三个自变量 x_3,x_4,x_5 对 Y_1,Y_2 的影响是否显著的检验统计量. 除 Λ 统计量外还给出其他几个统计量，结论都是否定 $B_2=O$ 的假定，即自变量 x_3,x_4,x_5 对 Y_1,Y_2 的影响是显著的.

输出 4.3.1 多变量检验统计量

Multivariate Statistics and F Approximations

S=2 M=0 N=10.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.17390860	10.72	6	46	<.0001
Pillai's Trace	1.08953122	9.57	6	48	<.0001
Hotelling-Lawley Trace	3.23532937	12.16	6	28.955	<.0001
Roy's Greatest Root	2.66743672	21.34	3	24	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

§4.4 多因变量的逐步回归

本节讨论多个因变量时关于自变量的逐步筛选方法，它的基本思想及基本步骤和一个因变量情况下逐步回归的基本思想和基本步

$$
\begin{aligned}
d &= u'u - u'C(C'C)^{-1}C'u = u'(I_n - H)u \\
&= u'u - u'(\mathbf{1}_n \vdots X)\left[\begin{array}{c|c} \dfrac{1}{n} + \overline{X}' L_{XX}^{-1}\overline{X} & -\overline{X}' L_{XX}^{-1} \\ \hline -L_{XX}^{-1}\overline{X} & L_{XX}^{-1} \end{array}\right]\begin{bmatrix} \mathbf{1}_n' \\ X' \end{bmatrix} u \\
&= u'\left(I_n - \frac{1}{n}J\right)u + u'\frac{1}{n}JXL_{XX}^{-1}X'\left(I_n - \frac{1}{n}J\right)u \\
&\quad - u'XL_{XX}^{-1}X'\left(I_n - \frac{1}{n}J\right)u \\
&= L_{uu} - u'\left(I_n - \frac{1}{n}J\right)XL_{XX}^{-1}X'\left(I_n - \frac{1}{n}J\right)u \\
&= L_{uu} - L_{uX}L_{XX}^{-1}L_{Xu}.
\end{aligned}
$$

类似可得 $u'(I_n - H)Y = L_{uY} - L_{uX}L_{XX}^{-1}L_{XY}$. 所以

$$\hat{b}_{(u)} = d^{-1}(L_{uY} - L_{uX}L_{XX}^{-1}L_{XY}).$$

而

$$
\begin{aligned}
\begin{bmatrix} \hat{b}_{(0)}(u) \\ \hat{B}(u) \end{bmatrix} &= \begin{bmatrix} \hat{b}_{(0)} \\ \hat{B} \end{bmatrix} - (C'C)^{-1}C'u\hat{b}_{(u)} \\
&= \begin{bmatrix} \hat{b}_{(0)} \\ \hat{B} \end{bmatrix} - \left[\begin{array}{c|c} \dfrac{1}{n} + \overline{X}' L_{XX}^{-1}\overline{X} & -\overline{X}' L_{XX}^{-1} \\ \hline -L_{XX}^{-1}\overline{X} & L_{XX}^{-1} \end{array}\right]\begin{bmatrix} \mathbf{1}_n' u \\ X'u \end{bmatrix}\hat{b}_{(u)} \\
&= \begin{bmatrix} \hat{b}_{(0)} - \overline{u}\hat{b}_{(u)} + \overline{X}' L_{XX}^{-1}L_{Xu}\hat{b}_{(u)} \\ \hat{B} - L_{XX}^{-1}L_{Xu}\hat{b}_{(u)} \end{bmatrix},
\end{aligned}
$$

所以

$$
\begin{aligned}
\hat{B}(u) &= \hat{B} - L_{XX}^{-1}L_{Xu}\hat{b}_{(u)}, \\
\hat{b}_{(0)}(u) &= \hat{b}_{(0)} - \overline{u}\hat{b}_{(u)} + \overline{X}' L_{XX}^{-1}L_{Xu}\hat{b}_{(u)} \\
&= \hat{b}_{(0)} - \overline{u}\hat{b}_{(u)} + \overline{X}'(\hat{B} - \hat{B}(u)) \\
&= \overline{Y}' - \overline{X}'\hat{B}(u) - \overline{u}\hat{b}_{(u)}.
\end{aligned}
$$

又在模型(4.3.6)及(4.3.7)下，已证明

$$Q_1 - Q = \hat{B}_2' X_2'(I_n - H_1)X_2\hat{B}_2.$$

对应于模型(4.4.2)及(4.4.1)有：$\hat{B}_2 = \hat{b}_{(u)}$，$X_2 = u$，$C_1 = C$. 故有

$$Q - Q(u) = \hat{b}'_{(u)}u'(I_n - H)u\hat{b}_{(u)}$$
$$= \hat{b}'_{(u)}d\hat{b}_{(u)} = d\hat{b}'_{(u)}\hat{b}_{(u)},$$

所以

$$Q(u) = Q - d\hat{b}'_{(u)}\hat{b}_{(u)}. \qquad \text{(证毕)}$$

2. 检验 $H_0: b(u)=O_{1\times p}$

根据§4.3的公式(4.3.10)选检验统计量

$$T^2(p,n-r-2) = (n-r-2)d\hat{b}_{(u)}Q^{-1}_{(u)}\hat{b}'_{(u)},$$

其中 $d=u'u-u'C(C'C)^{-1}C'u=u'(I_n-H)u$；又

$$(C'_uC_u)^{-1} = \begin{bmatrix} C'C & C'u \\ u'C & u'u \end{bmatrix}^{-1} = \begin{bmatrix} * & * \\ * & d^{-1} \end{bmatrix}_{(r+2)\times(r+2)},$$

所以

$$d^{-1} = e'_{r+2}(C'_uC_u)^{-1}e_{r+2}$$
$$= (C'_uC_u)^{-1} \text{ 的第 } r+2 \text{ 个对角元素.}$$

利用定理4.4.1及附录中§4的定理4.2可得

$$\hat{b}_{(u)}Q^{-1}_{(u)}\hat{b}'_{(u)} = \hat{b}_{(u)}(Q - d\hat{b}'_{(u)}\hat{b}_{(u)})^{-1}\hat{b}'_{(u)}$$
$$= \hat{b}_{(u)}\left(Q^{-1} + \frac{dQ^{-1}\hat{b}'_{(u)}\hat{b}_{(u)}Q^{-1}}{1 - d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}\right)\hat{b}'_{(u)}$$
$$= \frac{\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}{1 - d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}},$$

因而

$$T^2 = (n-r-2)\frac{d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}{1 - d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}.$$

在 H_0 成立时 $T^2 \sim T^2(p,n-r-2)$，从而

$$F = \frac{(n-r-2)-p+1}{p}\frac{T^2}{n-r-2}$$
$$= \frac{n-r-p-1}{p}\frac{d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}{1 - d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}$$
$$\sim F(p,n-r-p-1).$$

给定显著性水平 α，由样本观测值计算 T^2、F 值(记为 f)及显著性概率值(p 值)：

$p=P\{F\geqslant f\}$　(检验统计量 $F\sim F(p,n-r-p-1)$)，

若 $p<\alpha$，否定 H_0，即变量 x_u 对 p 个因变量的作用显著；若 $p\geqslant\alpha$，则 H_0 相容，即变量 x_u 对 p 个因变量的作用不显著.

利用似然比原理，可引入统计量 U，且由 U 和 T^2 的关系有：

$$\begin{aligned}U&=\frac{1}{1+d\hat{b}_{(u)}Q^{-1}(u)\hat{b}'_{(u)}}\\&=\frac{1}{1+\dfrac{T^2}{n-r-2}}=\frac{1}{1+\dfrac{d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}{1-d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}}}\\&=1-d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}\overset{\text{def}}{=\!=}1-V_u,\end{aligned}$$

其中

$$V_u=d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}. \tag{4.4.4}$$

显然 $V_u=d\hat{b}_{(u)}Q^{-1}\hat{b}'_{(u)}$ 是变量 u 对 p 个因变量的“贡献”.

3. 在模型(4.4.1)下检验 $H_0^{(i)}$：$b_{(i)}=O_{1\times p}$

在模型(4.4.1)下检验 $H_0^{(i)}$：$b_{(i)}=O_{1\times p}$ $(i=1,2,\cdots,r)$，其中 $b_{(i)}$ 是参数矩阵 B 中第 i 个行向量：

$$B=\begin{bmatrix}b_{(1)}\\\vdots\\b_{(r)}\end{bmatrix}.$$

由 §4.3 中小节**二**“回归系数的显著性检验”的讨论可知，在 $H_0^{(i)}$ 成立时统计量

$$\begin{aligned}T^2&=(n-r-1)b_{(i)}Q^{-1}b'_{(i)}/l^{ii}\\&\sim T^2(p,n-r-1),\end{aligned}$$

其中 l^{ii} 为 L_{XX}^{-1} 的第 i 个对角元素. 于是在 $H_0^{(i)}$ 成立时，

$$F=\frac{n-r-p}{p}\frac{T^2}{n-r-1}\sim F(p,n-r-p).$$

给定显著性水平 α，由样本值计算检验统计量 F 的值(记为 f)及显著性概率值(p 值)：

$$p = P\{F \geqslant f\} \quad (\text{检验统计量 } F \sim F(p, n-r-p)),$$

若 $p<\alpha$，否定 $H_0^{(i)}$，即变量 x_i 对 p 个因变量的作用显著；若 $p\geqslant\alpha$，则 $H_0^{(i)}$ 相容，即变量 x_i 对 p 个因变量的作用不显著.

二、多因变量逐步回归的步骤及算法

设有 p 个因变量与 m 个自变量，观测数据阵为

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad Y=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}.$$

1. 准备工作

(1) 考虑是否对原始数据进行标准化. 由于变量 $x_i(i=1,2,\cdots,m)$ 和 $Y_j(j=1,2,\cdots,p)$ 所取单位不同，取值范围不同，为了减少量纲的影响及减少计算误差，经常对数据进行标准化. 标准化的方法有多种，在这里我们采用标准差标准化，即令

$$x_{ti}^{*}=\frac{x_{ti}-\bar{x}_i}{s_i(x)} \quad (i=1,2,\cdots,m;\ t=1,2,\cdots,n),$$

其中 $\bar{x}_i=\dfrac{1}{n}\sum\limits_{t=1}^{n}x_{ti}$，$s_i(x)=\sqrt{\dfrac{1}{n-1}\sum\limits_{t=1}^{n}(x_{ti}-\bar{x}_i)^2}$. 令

$$y_{tj}^{*}=\frac{y_{tj}-\bar{y}_j}{s_j(y)} \quad (j=1,2,\cdots,p;\ t=1,2,\cdots,n),$$

其中 $\bar{y}_j=\dfrac{1}{n}\sum\limits_{t=1}^{n}y_{tj}$，$s_j(y)=\sqrt{\dfrac{1}{n-1}\sum\limits_{t=1}^{n}(y_{tj}-\bar{y}_j)^2}$.

(2) 计算 $m+p$ 阶矩阵 L. 设中心化后的数据阵记为 $\widetilde{X}$ 和 $\widetilde{Y}$，记 $L=L^{(0)}$，则

$$L^{(0)}=\begin{bmatrix} \widetilde{X}'\widetilde{X} & \widetilde{X}'\widetilde{Y} \\ \widetilde{Y}'\widetilde{X} & \widetilde{Y}'\widetilde{Y} \end{bmatrix} \overset{\text{def}}{=\!=} (l_{ij}^{(0)})=\begin{bmatrix} L_{XX}^{(0)} & L_{XY}^{(0)} \\ L_{YX}^{(0)} & L_{YY}^{(0)} \end{bmatrix}\begin{matrix} m \\ p \end{matrix}$$

为 $(m+p)\times(m+p)$ 矩阵. 如果数据已做标准化变换，则矩阵 $L^{(0)}$ 就是 $m+p$ 个变量的相关阵.

(3) 给定引入变量时的显著性水平 α_{in} 和剔除变量时的显著性水平 α_{out}(要求 $\alpha_{in}\leqslant\alpha_{out}$).

2. 逐步筛选自变量

下面我们从 $L^{(0)}$ 出发利用消去变换进行多因变量逐步回归计算.

第 1 步：考虑从 m 个自变量 $x_1,\cdots,x_m$ 中能否引入变量的步骤和公式.具体如下：

(1) 计算自变量 $x_i(i=1,2,\cdots,m)$ 对因变量的贡献，由公式(4.4.4)可得

$$V_i = d_i\hat{b}_{(i)}Q^{-1}\hat{b}'_{(i)} \quad (i=1,2,\cdots,m),$$

此时方程中变量个数 $r=0$，故 $d_i=l_{ii}^{(0)}$，$\hat{b}_{(i)}=l_{iY}^{(0)}/l_{ii}^{(0)}$，其中

$$l_{iY}^{(0)} = (l_{i(m+1)}^{(0)},l_{i(m+2)}^{(0)},\cdots,l_{i(m+p)}^{(0)}), \quad Q = L_{YY}.$$

于是

$$V_i = l_{ii}^{(0)}\frac{l_{iY}^{(0)}}{l_{ii}^{(0)}}L_{YY}^{-1}\left(\frac{l_{iY}^{(0)}}{l_{ii}^{(0)}}\right)' = \frac{l_{iY}^{(0)}L_{YY}^{-1}l_{Yi}^{(0)}}{l_{ii}^{(0)}}.$$

(2) 选 V_i 最大者，记为 V_{i_1}，即 $V_{i_1}=\max\limits_{i=1,\cdots,m}V_i$.

(3) 检验 x_{i_1} 对因变量的作用是否显著(即检验 H_0：$\beta_{(i_1)}=0_p$ 或 $b_{(i_1)}=O_{1\times p}$).因检验统计量

$$F_1=\frac{n-p-1}{p}\frac{d_{i_1}\hat{b}_{(i_1)}Q^{-1}\hat{b}'_{(i_1)}}{1-d_{i_1}\hat{b}_{(i_1)}Q^{-1}\hat{b}'_{(i_1)}}=\frac{n-p-1}{p}\frac{V_{i_1}}{1-V_{i_1}}$$

$$\sim F(p,n-p-1) \quad (\text{当 } b_{(i_1)}=O_{1\times p}\text{ 时}).$$

由 $L^{(0)}$ 矩阵出发计算检验统计量 F_1 的值(记为 f_1)及显著性概率值(p 值)：

$$p = P\{F_1\geqslant f_1\} \quad (\text{其中 } F_1\sim F(p,n-p-1)).$$

若 $p<\alpha_{in}$，则引入变量 x_{i_1}，并对 $L^{(0)}$ 作消去变换得

$$L^{(1)} = T_{i_1}[L^{(0)}] = \begin{bmatrix} L_{XX}^{(1)} & L_{XY}^{(1)} \\ L_{YX}^{(1)} & L_{YY}^{(1)} \end{bmatrix},$$

且从 $L^{(1)}$ 中，可得

$$\hat{b}_{(i_1)} = l_{i_1Y}^{(1)}(L_{XY}^{(1)}\text{ 的第 }i_1\text{ 行}), \quad Q(i_1) = L_{YY}^{(1)}.$$

若 $p \geqslant \alpha_{in}$，则自变量的筛选停止.

第 k 步：考虑能否剔除变量的步骤和公式. 不妨设已引入回归方程的变量记为 $x_1,\cdots,x_r(r \leqslant m)$. 每引入或剔除一个自变量作一次消去变换，$L^{(0)}$ 经若干次消去变换后化为 $L^{(r)}=T_r\cdots T_1[L^{(0)}]$. 记

$$L^{(r)}=\begin{bmatrix} L_{XX}^{(r)} & L_{XY}^{(r)} \\ L_{YX}^{(r)} & L_{YY}^{(r)} \end{bmatrix}=\begin{bmatrix} L_{XX}^{(r)}(1,1) & L_{XX}^{(r)}(1,2) & L_{XY}^{(r)}(1) \\ L_{XX}^{(r)}(2,1) & L_{XX}^{(r)}(2,2) & L_{XY}^{(r)}(2) \\ L_{YX}^{(r)}(1) & L_{YX}^{(r)}(2) & L_{YY}^{(r)} \end{bmatrix}\begin{matrix} r \\ m-r \\ p \end{matrix}.$$

利用消去变换的性质可知：

$$L^{(r)}=\begin{bmatrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{bmatrix},$$

其中

$$\begin{aligned} L_{11} &= L_{XX}^{-1}(1,1), \quad L_{12}=L_{XX}^{-1}(1,1)L_{XX}(1,2), \\ L_{13} &= L_{XX}^{-1}(1,1)L_{XY}(1), \\ L_{21} &= -L_{XX}(2,1)L_{XX}^{-1}(1,1), \\ L_{22} &= L_{XX}(2,2)-L_{XX}(2,1)L_{XX}^{-1}(1,1)L_{XX}(1,2), \\ L_{23} &= L_{XY}(2)-L_{XX}(2,1)L_{XX}^{-1}(1,1)L_{XY}(1), \\ L_{31} &= -L_{YX}(1)L_{XX}^{-1}(1,1), \\ L_{32} &= L_{YX}(2)-L_{XX}(2,1)L_{XX}^{-1}(1,1)L_{XX}(1,2), \\ L_{33} &= L_{YY}-L_{YX}(1)L_{XX}^{-1}(1,1)L_{XY}(1). \end{aligned}$$

显然，模型 $Y=(\mathbf{1}_n \mid X(r))\begin{bmatrix} b_{(0)} \\ B \end{bmatrix}+E$ 中参数矩阵 B 的最小二乘估计为

$$\hat{B}=L_{XX}^{-1}(1,1)L_{XY}(1)=L_{XY}^{(r)}(1).$$

残差阵

$$Q=L_{YY}-L_{YX}(1)L_{XX}^{-1}(1,1)L_{XY}(1)=L_{YY}^{(r)}.$$

以下是能否剔除变量的具体步骤：

(1) 计算自变量 $x_i(i=1,2,\cdots,r)$ 对 p 个因变量的贡献：

$$V_i=d_i\hat{b}_{(i)}Q^{-1}\hat{b}_{(i)}' \quad (i=1,2,\cdots,r),$$

其中 $d_i=1/l_{ii}^{(r)}$，$\hat{b}_{(i)}=l_{iY}^{(r)}$，$Q=L_{YY}^{(r)}$. 又因为 x_i 为已入选的变量，由消去变换的性质知

$$(l_{iY}^{(r)})' = -\,l_{Yi}^{(r)} \quad (i=1,\cdots,r),$$

所以

$$V_i = l_{iY}^{(r)}(L_{YX}^{(r)})^{-1}(-\,l_{Yi}^{(r)})/l_{ii}^{(r)}$$
$$= -\,\frac{l_{iY}^{(r)}(L_{YY}^{(r)})^{-1}l_{Yi}^{(r)}}{l_{ii}^{(r)}} \quad (i=1,2,\cdots,r).$$

(2) 选 V_i 最小者，记为 V_{i_0}，即 $V_{i_0}=\min\limits_{i=1,\cdots,r} V_i$.

(3) 检验变量 x_{i_0} 是否可以剔除(即检验 H_0：$\beta_{(i_0)}=0_p$ 或 $b_{(i_0)}=O_{1\times p}$). 计算检验统计量

$$f_2=\frac{(n-r-1)-p+1}{p}\,\frac{T_{i_0}^2}{n-r-1}=\frac{n-p-r}{p}V_{i_0},$$

及

$$p=P\{F_2\geqslant f_2\} \quad (\text{其中 } F_2\sim F(p,n-p-r)).$$

若 $p\geqslant\alpha_{\text{out}}$，则剔除变量 x_{i_0}，并对 $L^{(r)}$ 作消去变换得 $L^{(r+1)}=T_{i_0}[L^{(r)}]$，且以 $L^{(r+1)}$ 为当前矩阵，重复第 k 步的几个步骤，直到没有变量可剔除为止；若 $p<\alpha_{\text{out}}$，则转入考虑能否引入新变量的步骤.

第 $k+1$ 步：考虑能否引入新变量的步骤和公式. 不妨设从未入选回归方程的变量为 $x_{r+1},\cdots,x_m$；当前矩阵为 $L^{(r)}$. 考虑可否引入新变量的步骤如下：

(1) 计算自变量 $x_j(j=r+1,\cdots,m)$ 对 p 个因变量的贡献：

$$V_j=d_j\hat{b}_{(j)}Q^{-1}\hat{b}_{(j)}' \quad (j=r+1,\cdots,m).$$

利用模型(4.4.1)和(4.4.2)下参数最小二乘估计的关系(4.4.3)可知，$\hat{b}_{(j)}=d_j^{-1}(l_{jY}-l_{jX}L_{XX}^{-1}L_{XY})$. 所以

$$V_j=d_jd_j^{-1}(l_{jY}-l_{jX}L_{XX}^{-1}L_{XY})Q^{-1}(l_{Yj}-L_{YX}L_{XX}^{-1}l_{Xj})d_j^{-1}$$
$$=\frac{l_{jY}^{(r)}(L_{YY}^{(r)})^{-1}(l_{jY}^{(r)})'}{l_{jj}^{(r)}} \quad (j=r+1,\cdots,m).$$

上式中

$$d_j=X_j'[I_n-C(C'C)^{-1}C']X_j=l_{jj}-l_{jX}L_{XX}^{-1}l_{Xj}=l_{jj}^{(r)},$$

其中 $C=(\mathbf{1}_n\,\vdots\,X(r))$，$X_j$ 为 $X(r)$ 矩阵的第 j 列. 又因为 $x_j(j=r+1,\cdots,m)$ 为未入选变量，由消去变换的性质可知

$$l_{jY}^{(r)} = (l_{Yj}^{(r)})'.$$

故 $V_j>0$ $(j=r+1,\cdots,m)$.

(2) 选 V_j 的最大者，记为 V_{j_0}，即 $V_{j_0}=\max\limits_{j=r+1,\cdots,m} V_j$.

(3) 检验变量 x_{j_0} 是否可以引入回归方程(即检验 H_0: $\beta_{(j_0)}=0_p$ 或 $b_{(j_0)}=O_{1\times p}$). 计算检验统计量

$$f_1=\frac{(n-r-2)-p+1}{p}\frac{T_{j_0}^2}{n-(r+1)-1}$$

$$=\frac{n-p-r-1}{p}\frac{V_{j_0}}{1-V_{j_0}},$$

及

$$p=P\{F_1\geqslant f_1\}\quad (F_1\sim F(p,n-p-r-1)).$$

若 $p<\alpha_{\text{in}}$，则引入变量 x_{j_0}，并对 $L^{(r)}$ 作消去变换得 $L^{(r+1)}=T_{j_0}[L^{(r)}]$，且以 $L^{(r+1)}$ 为当前矩阵，转入考虑能否剔除老变量的步骤；若 $p\geqslant\alpha_{\text{in}}$，则逐步筛选自变量的过程结束.

3. 给出计算结果

设筛选自变量的过程结束时，入选的自变量为 $x_{i_1},x_{i_2},\cdots,x_{i_r}$ $(r\leqslant m)$，矩阵 $L^{(0)}$ 经多次消去变换后化为 $L^{(r)}=T(i_1,i_2,\cdots,i_r)(L^{(0)})$. 记

$$L^{(r)}=\begin{bmatrix} L_{XX}^{(r)} & L_{XY}^{(r)} \\ L_{YX}^{(r)} & L_{YY}^{(r)} \end{bmatrix}\begin{matrix} m \\ p \end{matrix}.$$

(1) Y_j 与 $x_{i_1},x_{i_2},\cdots,x_{i_r}$ 的回归方程为

$$\hat{Y}_j=\hat{\beta}_{0j}+l_{i_1(m+j)}^{(r)}x_{i_1}+\cdots+l_{i_r(m+j)}^{(r)}x_{i_r},$$

其中 $$\hat{\beta}_{0j}=\bar{y}_j-\sum_{t=1}^{r}l_{i_t(m+j)}^{(r)}\bar{x}_{i_t}(j=1,2,\cdots,p).$$

(2) 协方差阵 Σ 的无偏估计为：$\hat{\Sigma}=\frac{1}{n-r-1}L_{YY}^{(r)}$.

(3) 考虑第 j 个因变量 Y_j 对 $x_{i_1},x_{i_2},\cdots,x_{i_r}$ 的多对一回归模型，回归方程见(1)中所示方程. 残差平方和：

$$Q_j=l_{(m+j)(m+j)}^{(r)}\quad (j=1,2,\cdots,p).$$

复相关系数：

$$R_j=\sqrt{1-\frac{Q_j}{l^{(0)}_{(m+j)(m+j)}}}\quad(j=1,2,\cdots,p).$$

例 4.4.1（例 4.3.1 的继续）　试用逐步筛选的方法求 Y_1,Y_2 与 $x_1,x_2,\cdots,x_5$ 的关系式.

解　取显著性水平 $\alpha_{in}=\alpha_{out}=0.05$，利用消去变换对自变量作筛选，最终入选的变量为 x_3,x_4,x_5，回归关系式为

$$\hat{Y}_1=8.499+2.841x_3-0.849x_4+1.348x_5,$$

$$\hat{Y}_2=5.293+1.725x_3+1.005x_4+1.973x_5.$$

复相关系数 $R_1=0.9855,R_2=0.9900$.

§4.5　双重筛选逐步回归

在多因变量的逐步回归方法中，引入（或剔除）自变量的准则是考察此变量对 p 个因变量的“贡献”大小，如果某一变量 x_i 只对因变量 Y_{i_0} 影响显著，而对其余变量作用不显著时，那么对 x_i 作显著性检验，很可能 x_i 不能引入方程. 在最终得到的回归方程组中，有的回归方程可能不是“最优”的，如在 Y_{i_0} 的回归方程中，重要变量 x_i 就没有被引入.

在多因变量与多个自变量的回归问题中，实际情况可以理解为这样，因变量的一部分与自变量的一部分有密切关系. 例如（不妨设为）$Y_1,Y_2,\cdots,Y_{p_1}$ 与 $x_{i_1},x_{i_2},\cdots,x_{i_k}(k\leqslant m)$ 有密切的关系，另一部分 $Y_{p_1+1},\cdots,Y_{p_2}$ 与 $x_{j_1},x_{j_2},\cdots,x_{j_l}(l\leqslant m)$ 有密切的关系，……最后一部分 $Y_{p_r+1},\cdots,Y_p$ 与 $x_{h_1},x_{h_2},\cdots,x_{h_t}(t\leqslant m)$ 有密切的关系. 显然各部分的因变量中不能有共同的变量；而各部分的自变量中可以有共同的变量，因为同一个自变量 x_i 可能对许多不同的 Y_i 甚至全部的 Y_i 都有密切关系. 因此就提出了一个问题，是否有一种逐步的算法，既能依因变量和自变量的关系来将因变量进行分组，又能使每个自变量对各组因变量的影响都能反映出来. 这就是本节将要介绍的双重筛选逐步回归问题.

一、基本理论和公式

1. 筛选自变量的基本公式

设考查 p_1 个因变量组成的因变量组与自变量的相关关系;且在某一步骤引入方程的自变量为 $x_{i_1},x_{i_2},\cdots,x_{i_r}$. 由 §4.4 可得出筛选自变量的几个基本公式(设当前矩阵为 $L^{(r)}=(l_{ij}^{(r)})$).

(1) 变量 x_i 对 p_1 个因变量组的贡献

$$V_i=\frac{l_{iY}^{(r)}(L_{YY}^{(r)})^{-1}l_{Yi}^{(r)}}{l_{ii}^{(r)}}\quad(i=1,2,\cdots,m).$$

注意:当 x_i 是已入选的变量时,$(l_{iY}^{(r)})'=-l_{Yi}^{(r)}$,故 $V_i<0$;当 x_i 未入选时,$V_i>0$.

(2) 考虑是否引入变量 x_{j_0} 时,计算统计量

$$F_1=\frac{(n-(r+1)-1)-p_1+1}{p_1}\frac{T_{j_0}^2}{n-(r+1)-1}$$

$$=\frac{n-r-p_1-1}{p_1}\frac{V_{j_0}}{1-V_{j_0}}.$$

(3) 考虑是否剔除变量 x_{i_0} 时,计算统计量

$$F_2=\frac{(n-r-1)-p_1+1}{p_1}\frac{T_{i_0}^2}{n-r-1}=\frac{n-r-p_1}{p_1}(-V_{i_0}).$$

2. 筛选因变量的理论和公式

设 m_1 个自变量(不妨设为 $x_1,\cdots,x_{m_1}$)与 p_1 个因变量(不妨设为 $Y_1,\cdots,Y_{p_1}$)的 n 次观测资料满足以下模型:

$$\begin{cases}Y=(\mathbf{1}_n\vdots X)\begin{bmatrix}b_{(0)}\\B\end{bmatrix}+E,\ \operatorname{rank}(\mathbf{1}_n\vdots X)=m_1+1,\\ \varepsilon_{(i)}\sim N_{p_1}(0,\Sigma)\ (i=1,2,\cdots,n)\text{ 相互独立.}\end{cases}\tag{4.5.1}$$

如果添加一个因变量 $Y_j(j>p_1)$,其相应的观测值为 $Y_j=(y_{1j},y_{2j},\cdots,y_{nj})'$,设 m_1 个自变量与 p_1+1 个因变量的 n 次观测资料满足模型

$$\begin{cases}(Y \mid Y_j) = (\mathbf{1}_n \mid X)\begin{bmatrix} b^*_{(0)} \\ B^* \end{bmatrix} + E^*, \ \mathrm{rank}(\mathbf{1}_n \mid X) = m_1 + 1, \\ \varepsilon^*_{(i)} \sim N_{p_1+1}(0, \Sigma^*) \ (i = 1,2,\cdots,n) \text{ 相互独立}, \end{cases} \tag{4.5.2}$$

其中

$$\Sigma^* = \begin{bmatrix} \Sigma & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{p_1}_{1},$$

$$\begin{bmatrix} b^*_{(0)} \\ B^* \end{bmatrix} = \left[\begin{array}{ccc:c} \beta_{01} & \cdots & \beta_{0p_1} & \beta_{0(p_1+1)} \\ \beta_{11} & \cdots & \beta_{1p_1} & \beta_{1(p_1+1)} \\ \vdots & & \vdots & \vdots \\ \beta_{m_1 1} & \cdots & \beta_{m_1 p_1} & \beta_{m_1(p_1+1)} \end{array}\right] = \left[\begin{array}{c:c} b_{(0)} & r_0 \\ B & r \end{array}\right]^{1}_{m_1}.$$

定理 4.5.1　在模型(4.5.2)下有

$$\begin{cases} \hat{b}^*_{(0)} = (\overline{Y}' \mid \overline{y}_j) - \overline{X}' \hat{B}^*, \\ \hat{B}^* = L_{XX}^{-1}(L_{XY} \mid L_{XY_j}), \\ Q^* = \left[\begin{array}{c:c} L_{YY} - L_{YX}L_{XX}^{-1}L_{XY} & L_{YY_j} - L_{YX}L_{XX}^{-1}L_{XY_j} \\ L_{Y_jY} - L_{Y_jX}L_{XX}^{-1}L_{XY} & L_{Y_jY_j} - L_{Y_jX}L_{XX}^{-1}L_{XY_j} \end{array}\right]^{p_1}_{1}. \end{cases}$$

证明　在模型(4.5.2)下有

$$\begin{bmatrix} \hat{b}^*_{(0)} \\ \hat{B}^* \end{bmatrix} = (C'C)^{-1}C'(Y \mid Y_j)$$

$$= \begin{bmatrix} \frac{1}{n}\mathbf{1}_n' - \frac{1}{n}\mathbf{1}_n' X L_{XX}^{-1} X' \left(I_n - \frac{1}{n}J\right) \\ L_{XX}^{-1}X'\left(I_n - \frac{1}{n}J\right) \end{bmatrix} (Y \mid Y_j).$$

所以

$$\hat{b}^*_{(0)} = (\overline{Y}' \mid \overline{y}_j) - \overline{X}' L_{XX}^{-1}(L_{XY} \mid L_{XY_j}) = (\overline{Y}' \mid \overline{y}_j) - \overline{X}' \hat{B}^*,$$

$$\hat{B}^* = L_{XX}^{-1}(L_{XY} \mid L_{XY_j}),$$

$$Q^* = (Y \mid Y_j)'(I_n - C(C'C)^{-1}C')(Y \mid Y_j)$$

$$= \left[\begin{array}{c:c} Y'(I_n - H)Y & Y'(I_n - H)Y_j \\ Y_j'(I_n - H)Y & Y_j'(I_n - H)Y_j \end{array}\right] \quad (H = C(C'C)^{-1}C')$$

$$=\left[\begin{array}{c|c} L_{YY}-L_{YX}L_{XX}^{-1}L_{XY} & L_{YY_j}-L_{YX}L_{XX}^{-1}L_{XY_j} \\ \hline L_{Y_jY}-L_{Y_jX}L_{XX}^{-1}L_{XY} & L_{Y_jY_j}-L_{Y_jX}L_{XX}^{-1}L_{XY_j} \end{array}\right]. \quad \text{(证毕)}$$

(1) 考察因变量 Y_j 可否引入的统计量. 在模型(4.5.1)下,引入威尔克斯统计量 $\Lambda_1=\dfrac{|Q|}{|L_{YY}|}$,它反映了变量 $x_1,x_2,\cdots,x_{m_1}$ 与 $Y_1,Y_2,\cdots,Y_{p_1}$ 之间的线性关系是否密切,Λ_1 值越小,表明 $x_1,x_2,\cdots,x_{m_1}$ 与 $Y_1,Y_2,\cdots,Y_{p_1}$ 的线性相关程度越高.

在模型(4.5.2)下,$\Lambda_2=\dfrac{|Q^*|}{|L_{(Y\vdots Y_j)(Y\vdots Y_j)}|}$,反映 $x_1,x_2,\cdots,x_{m_1}$ 与 $Y_1,Y_2,\cdots,Y_{p_1},Y_j$ 之间相关关系的密切程度,且 $\Lambda_2\leqslant\Lambda_1$.

考察比值 $\dfrac{\Lambda_2}{\Lambda_1}$,显然 $\dfrac{\Lambda_2}{\Lambda_1}\leqslant 1$. 若 $\dfrac{\Lambda_2}{\Lambda_1}\approx 1$,这说明引入因变量 Y_j 对回归方程没有显著影响. 若 $\dfrac{\Lambda_2}{\Lambda_1}\ll 1$,即 $\Lambda_2\ll\Lambda_1$,表明引入因变量 Y_j 对回归方程有显著影响.

令统计量 $U_j=1-\dfrac{\Lambda_2}{\Lambda_1}$,则 U_j 的大小就是因变量 Y_j 对回归方程的"贡献". 下面来推导 U_j 的表达式:

$$\begin{aligned} L_{(Y\vdots Y_j)(Y\vdots Y_j)} &= (Y\vdots Y_j)'\left(I_n-\frac{1}{n}J\right)(Y\vdots Y_j) \\ &= \left[\begin{array}{c|c} Y'\left(I_n-\frac{1}{n}J\right)Y & Y'\left(I_n-\frac{1}{n}J\right)Y_j \\ \hline Y_j'\left(I_n-\frac{1}{n}J\right)Y & Y_j'\left(I_n-\frac{1}{n}J\right)Y_j \end{array}\right] \\ &= \left[\begin{array}{c|c} L_{YY} & L_{YY_j} \\ \hline L_{Y_jY} & L_{Y_jY_j} \end{array}\right]\begin{matrix} p_1 \\ 1 \end{matrix}. \end{aligned}$$

记

$$\begin{aligned} L_{YY}(X) &= L_{YY}-L_{YX}L_{XX}^{-1}L_{XY}, \\ L_{YY_j}(X) &= L_{YY_j}-L_{YX}L_{XX}^{-1}L_{XY_j}, \\ L_{Y_jY}(X) &= L_{Y_jY}-L_{Y_jX}L_{XX}^{-1}L_{XY}, \\ L_{Y_jY_j}(X) &= L_{Y_jY_j}-L_{Y_jX}L_{XX}^{-1}L_{XY_j}, \end{aligned}$$

则

$$Q^* = \left[\begin{array}{c:c} L_{YY}(X) & L_{YY_j}(X) \\ \hdashline L_{Y_jY}(X) & L_{Y_jY_j}(X) \end{array}\right]_1^{p_1},$$

且

$$\begin{aligned} |Q^*| &= |L_{YY}(X)| \cdot |L_{Y_jY_j}(X) - L_{Y_jY}(X)L_{YY}^{-1}(X)L_{YY_j}(X)| \\ &= |L_{YY}(X)| \cdot |L_{Y_jY_j}(X \vdots Y)|, \end{aligned}$$

其中

$$\begin{aligned} L_{Y_jY_j}(X \vdots Y) &= L_{Y_jY_j} - L_{Y_j(X \vdots Y)}L^{-1}_{(X \vdots Y)(X \vdots Y)}L_{(X \vdots Y)Y_j} \\ &= L_{Y_jY_j} - (L_{Y_jX} \vdots L_{Y_jY})\left[\begin{array}{c:c} L_{XX} & L_{XY} \\ \hdashline L_{YX} & L_{YY} \end{array}\right]^{-1}\left[\begin{array}{c} L_{XY_j} \\ \hdashline L_{YY_j} \end{array}\right] \\ &= \cdots\cdots(\text{展开整理}) \\ &= L_{Y_jY_j}(X) - L_{Y_jY}(X)L_{YY}^{-1}(X)L_{YY_j}(X). \end{aligned}$$

所以

$$\begin{aligned} U_j &= 1 - \frac{\Lambda_2}{\Lambda_1} = 1 - \frac{|Q^*|}{|L_{(Y \vdots Y_j)(Y \vdots Y_j)}|} \cdot \frac{|L_{YY}|}{|Q|} \\ &= 1 - \frac{|L_{YY}(X)| \cdot |L_{Y_jY_j}(X \vdots Y)| \cdot |L_{YY}|}{|L_{YY}| \cdot |L_{Y_jY_j} - L_{Y_jY}L_{YY}^{-1}L_{YY_j}| \cdot |L_{YY}(X)|} \\ &= 1 - \frac{L_{Y_jY_j}(X \vdots Y)}{L_{Y_jY_j}(Y)} \quad (j = p_1 + 1, \cdots, p). \end{aligned}$$

利用 Λ 统计量与 F 统计量的关系，可以证明

$$F_1 = \frac{n - p_1 - m_1 - 1}{m_1} \frac{U_j}{1 - U_j} \sim F(m_1, n - p_1 - m_1 - 1).$$

利用统计量 F_1 可检验 Y_j 可否引入.

(2) 在模型(4.5.2)下检验 Y_j 可否剔除的统计量. 在模型(4.5.2)下，威尔克斯统计量

$$\Lambda_2 = \frac{|Q^*|}{|L_{(Y \vdots Y_j)(Y \vdots Y_j)}|}.$$

当剔除 Y_j 后得模型(4.5.1)，且 $\Lambda_1 = \dfrac{|Q|}{|L_{YY}|}$. 记 $U_j = \dfrac{\Lambda_1 - \Lambda_2}{\Lambda_2} = \dfrac{\Lambda_1}{\Lambda_2} - 1$，并称 U_j 为 Y_j 的“贡献”. 则

$$F_2 = \frac{n - p_1 - m_1}{m_1} U_j \sim F(m_1, n - p_1 - m_1),$$

利用 F_2 可检验 Y_j 可否剔除.

3. 因变量筛选的另一模型

在考虑因变量的筛选时,我们也可以把 $x_1, x_2, \cdots, x_m$ 和 $Y_1, Y_2, \cdots, Y_p$ 所处的地位交换一下,即把 m 个变量 $x_1, x_2, \cdots, x_m$ 作为 m 维随机向量,来考察它与 $Y_1, Y_2, \cdots, Y_p$ 的依赖关系.

假设变量 $x_1, x_2, \cdots, x_{m_1}$ 与变量 $Y_1, Y_2, \cdots, Y_{p_1}$ 的 n 次观测数据满足模型:

$$\begin{cases} X_{n\times m_1} = (\mathbf{1}_n \vdots Y) \begin{bmatrix} b_{(0)} \\ B \end{bmatrix}_{(1+p_1)\times m_1} + E, \\ \varepsilon_{(i)} \sim N_{m_1}(0, \Sigma_X) \ (i = 1,2,\cdots,n) \text{ 相互独立.} \end{cases} \tag{4.5.3}$$

记

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p_1} \\ y_{21} & y_{22} & \cdots & y_{2p_1} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np_1} \end{bmatrix}, \quad Y_j = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{nj} \end{bmatrix} \quad (j = p_1 + 1, \cdots, p).$$

(1) 考虑逐个剔除 $Y_i (i=1,\cdots,p_1)$ 后模型(4.5.3)变为:

$$\begin{cases} X_{n\times m_1} = (\mathbf{1}_n \vdots \widetilde{Y}(i)) \begin{bmatrix} \tilde{b}_{(0)} \\ \widetilde{B}(i) \end{bmatrix}_{p_1\times m_1} + E, \\ \varepsilon_{(i)} \sim N_{m_1}(0, \Sigma_X) \ (i = 1,2,\cdots,n) \text{ 相互独立,} \end{cases} \tag{4.5.4}$$

其中 $\widetilde{Y}(i)$ 为从 Y 中删去第 i 列数据后的数据阵. $\widetilde{B}(i)$ 为从 B 中删去第 i 个行参数 $b_{(i)}$ 后的参数矩阵. 检验 Y_i 能否从方程中剔除即检验 $H_0^{(i)}: b_{(i)} = O_{1\times m_1}$. 根据似然比原理选统计量:

$$U = \frac{|Q(p_1)|}{|Q(p_1 - 1)|} = \frac{|Q(p_1)|}{|Q(p_1) + [Q(p_1 - 1) - Q(p_1)]|}$$
$$\sim \Lambda(m_1, n - p_1 - 1, 1),$$

其中 $Q(p_1)$ 表示包含 p_1 个变量时模型(4.5.3)的残差阵. 由第三章§3.1 的有关结论可得

$$T^2(m_1,n-p_1-1)=(n-p_1-1)\frac{1-U}{U}$$

$$=(n-p_1-1)\,d\hat{b}_{(i)}Q^{-1}(p_1)\hat{b}'_{(i)}.$$

记 $u_i=d\hat{b}_{(i)}Q^{-1}(p_1)\hat{b}'_{(i)}(i=1,2,\cdots,p_1)$，则称 u_i 为变量 Y_i 对变量 x_1，$x_2,\cdots,x_{m_1}$ 的贡献. 所以

$$F=\frac{n-m_1-p_1}{m_1}\frac{1-U}{U}=\frac{n-m_1-p_1}{m_1}u_i$$

$$\sim F(m_1,n-m_1-p_1).$$

利用 F 统计量可检验假设 $H_0^{(i)}$.

(2) 考虑引入变量 $Y_j(j=p_1+1,\cdots,p)$ 后模型(4.5.3)变为：

$$\begin{cases} X_{n\times m_1}=(\mathbf{1}_n\vdots Y\vdots Y_j)\begin{bmatrix} b_{(0)}\\ B\\ b_{(j)}\end{bmatrix}_{(p_1+2)\times m_1}+E,\\ \varepsilon_{(i)}\sim N_{m_1}(0,\Sigma_X)\ (i=1,2,\cdots,n)\ \text{相互独立},\end{cases}\tag{4.5.5}$$

检验 Y_j 可否引入方程中即检验 $H_0^{(j)}$：$b_{(j)}=O_{1\times m_1}$.

计算 $Y_j(j=p_1+1,\cdots,p)$ 对变量 $x_1,x_2,\cdots,x_{m_1}$ 的贡献 u_j，

$$u_j=d_j\hat{b}_{(j)}Q^{-1}(p_1)\hat{b}'_{(j)}\quad(j=p_1+1,\cdots,p).$$

而

$$\frac{T^2}{n-p_1-2}=\frac{d_j\hat{b}_{(j)}Q^{-1}(p_1)\hat{b}'_{(j)}}{1-d_j\hat{b}_{(j)}Q^{-1}(p_1)\hat{b}'_{(j)}}=\frac{u_j}{1-u_j},$$

则统计量

$$F=\frac{(n-p_1-2)-m_1+1}{m_1}\frac{T^2}{n-p_1-2}$$

$$=\frac{n-p_1-m_1-1}{m_1}\frac{u_j}{1-u_j}$$

$$\sim F(m_1,n-p_1-m_1-1).$$

利用 F 统计量可检验假设 $H_0^{(j)}$.

由上可见，关于因变量的两类不同模型，用来筛选因变量的统计量是一样的.

二、双重筛选逐步回归的基本步骤

设自变量(因子)为 $x_1,x_2,\cdots,x_m$,因变量(预报量)为 $Y_1,Y_2,\cdots,Y_p$,记为 $x_{m+1},x_{m+2},\cdots,x_{m+p}$. 它们共有 n 次观测数据,其数据阵 $X=(x_{ij})_{n\times(m+p)}$,其中 x_{ij} 表示第 j 个变量的第 i 次观测值.

1. 准备工作

计算 m 个自变量,p 个因变量 n 次观测数据的平均值;计算 $m+p$ 个变量的相关阵(即标准化数据的样本协方差阵). 规定筛选自变量和因变量的显著性水平 α_X 和 α_Y($0<\alpha_X,\alpha_Y<1$). 一般规定筛选自变量时引进和剔除变量的显著性水平相等,且记为 α_X;规定筛选因变量时引进和剔除变量的显著性水平均为 α_Y.

2. 双重逐步筛选过程

第一步:选入一个因变量 Y_{j_1}. 考虑 $Y_j(j=1,2,\cdots,p)$ 与 $x_i(i=1,2,\cdots,m)$ 的一元回归. 从 $p\times m$ 个回归平方和中选最大者,相应的因变量作为 Y_{j_1};也可以就取 Y_1 或任一个 $Y_j(j=1,\cdots,p)$ 作为 Y_{j_1}.

假设已计算了 k 步,入选的自变量有 m_1 个(不妨设为 $x_1,\cdots,x_{m_1}$),因变量有 p_1 个(不妨设为 $Y_1,\cdots,Y_{p_1}$). 每引入(或剔除)一个因变量 Y_{j_l} 时,对前面得到的 R 矩阵的相应块作消去变换.

第 $k+1$ 步:筛选自变量(筛选因子).

(1) 计算各个自变量对 p_1 个因变量的"贡献";

(2) 考虑可否剔除自变量. 对已入选的变量 x_i,选出对 $Y_1,Y_2,\cdots,Y_{p_1}$ 贡献最小的变量,记为 x_{i_0},并检验 x_{i_0} 可否剔除. 若不能剔除变量,转入下面的(3)考虑能否引入新变量;若可以剔除变量 x_{i_0},则对当前 R 矩阵的相应块作消去变换,并计算 m_1-1 个自变量与 p_1 个因变量的回归模型下的威尔克斯统计量 $\Lambda(m_1-1,p_1)$. 然后重复第 $k+1$ 步,继续考虑自变量的筛选.

(3) 考虑可否引入新变量. 对未入选的变量 x_j,选出对 $Y_1,Y_2,\cdots,Y_{p_1}$ 贡献最大的变量,记为 x_{j_0},并检验 x_{j_0} 可否引入. 若 x_{j_0} 不能被引入,则自变量的筛选过程结束;若可以引入变量 x_{j_0},则对当前 R 矩阵的相应块作消去变换,并计算 m_1+1 个自变量与 p_1 个因变量的回

归模型下的威尔克斯统计量 $\Lambda(m_1+1,p_1)$. 然后重复第 $k+1$ 步，继续考虑自变量的筛选.

第 $k+2$ 步：筛选因变量(筛选预报量). 仍假设此时入选的自变量为 m_1 个(不妨设为 $x_1,\cdots,x_{m_1}$)，因变量为 p_1 个(不妨设为 $Y_1,\cdots,Y_{p_1}$). 每引入(或剔除)一个自变量时，对前面得到的 R 矩阵的相应块作消去变换.

(1) 计算各因变量 Y_j 对 m_1 个自变量贡献.

(2) 考虑可否剔除因变量. 对已入选的因变量 Y_j，选出对 $x_1,x_2,\cdots,x_{m_1}$ 贡献最小的变量，记为 Y_{i_0}，并检验 Y_{i_0} 可否剔除. 若 Y_{i_0} 不能剔除，转入考虑引入新因变量；若可以剔除 Y_{i_0}，则对当前 R 矩阵的相应块作消去变换，并计算 m_1 个自变量与 p_1-1 个因变量的回归模型下的威尔克斯统计量 $\Lambda(p_1-1,m_1)$. 然后重复第 $k+2$ 步，继续考虑因变量的筛选.

注意：当 $p_1=1$ 时，考虑可否剔除的步骤应该跳过，直接考虑可否引入新因变量.

(3) 考虑可否引入新因变量. 对未入选的因变量 Y_j，选出对 $x_1,x_2,\cdots,x_{m_1}$ 贡献最大的变量，记为 Y_{j_0}，并检验 Y_{j_0} 可否引入. 若 Y_{j_0} 不能被引入，因变量的筛选过程结束；重复第 $k+1$ 步，考虑自变量的筛选. 如果自变量既没有可剔除的，又没有可引入的，则双重筛选过程结束，转入计算本组回归模型的有关结果. 若因变量 Y_{j_0} 可以引入，则对当前 R 矩阵的相应块作消去变换，并计算 m_1 个自变量与 p_1+1 个因变量的回归模型下的威尔克斯统计量 $\Lambda(p_1+1,m_1)$. 然后重复第 $k+2$ 步，继续考虑因变量的筛选.

在以上给出的双重逐步筛选过程中，自变量和因变量的地位是同等的. 在引入一个因变量后，对自变量进行筛选，找出对这一因变量影响显著的自变量组 $\{x_{i_1},x_{i_2},\cdots,x_{i_r}\}$；然后考虑因变量的筛选，这相当于把 $x_1,x_2,\cdots,x_m$ 和 $Y_1,Y_2,\cdots,Y_p$ 的地位作一交换. 类似地，用逐步筛选法筛选因变量，设 $\{Y_{j_1},Y_{j_2},\cdots,Y_{j_l}\}$ 为对 r 个变量 $\{x_{i_1},x_{i_2},\cdots,x_{i_r}\}$ 影响显著的变量组，接着再筛选自变量，找出对 l 个因变量影响显著的自变量组，这一过程直至某步当自变量筛选后，没有因变量

可剔除，同时也没有因变量可引入时，双重逐步筛选过程结束.

在张尧庭和方开泰编著的《多元统计分析引论》一书中，关于双重筛选逐步回归方法的筛选过程是：每当引入一个因变量，立即转入自变量的筛选；每当剔除一个因变量，也立即转入自变量的筛选. 显然这一过程突出了因变量，即逐次考察一个因变量的选入(或剔除)对于影响显著的自变量组的变化.

以上两种筛选过程有区别，计算结果有时也不完全一致，特别当 p 较大时，对因变量的分组结果可能不相同. 如果希望对因变量的分组细一些，一般采用《多元统计分析引论》中介绍的筛选过程.

3. 计算该多因变量组回归模型的有关结果

假设最终入选的因变量为 $Y_1, Y_2, \cdots, Y_{p_1}$，自变量为 $x_1, x_2, \cdots, x_{m_1}$. 观测数据阵 $X(1)$ 和 $X(3)=Y(1)$ ($X(1)$ 为 $n\times m_1$ 矩阵，$X(3)$ 为 $n\times p_1$ 矩阵)满足以下模型：

$$\begin{cases} Y(1) = (\mathbf{1}_n \vdots X(1))\begin{bmatrix} b_{(0)} \\ B(1) \end{bmatrix} + E, \ \mathrm{rank}(\mathbf{1}_n \vdots X(1)) = m_1 + 1, \\ \varepsilon_{(i)} \sim N_{p_1}(0, \Sigma) \ (i = 1, 2, \cdots, n) \text{ 相互独立.} \end{cases} \tag{4.5.6}$$

(1) 计算回归方程组. 但注意到原始观测数据阵 X 已经标准化处理了，故得到的回归方程应还原为原变量的回归方程.

(2) 计算模型(4.5.6)的残差阵 Q.

(3) 计算模型(4.5.6)的威尔克斯统计量 $\Lambda(m_1, p_1)$. 当 $p_1=1$ 时，

$$\Lambda(m_1, 1) = \frac{|Q(m_1, 1)|}{|l_{Y_1Y_1}|} = 1 - \frac{U(m_1)}{l_{Y_1Y_1}} = 1 - R^2,$$

其中 R 是 Y_1 与 $x_1, \cdots, x_{m_1}$ 的复相关系数，R^2 称为该模型的决定系数. $\Lambda(m_1, p_1)$ 值越小，表明 m_1 个自变量和 p_1 个因变量的关系越密切.

4. 计算下一组回归方程组

从原始数据阵中删去已入选的因变量的数据(注意，自变量的数据均不删)，重复以上 **2** 和 **3** 两小节中的步骤，考虑 $p-p_1$ 个因变量与 m 个自变量的双重筛选逐步回归，即可求得第二组，第三组，……

第 L 组的全部因变量的回归方程. 至此 p 个因变量，m 个自变量的双重筛选逐步回归计算全部结束.

例 4.5.1(马尾松毛虫的虫情预测)　马尾松毛虫危害极其严重，有时使整片森林树木死光. 利用上几代虫情、各种气象因子、防治否、坡向地形，以及虫龄等因素对下几代虫情进行预测预报是松毛虫综合防治中一项有意义的工作. 某县为对松毛虫的发生情况进行分析，在高山、丘陵与平地共均匀分布的设 20 个点，每月上、下旬分别调查统计有虫株率与虫口密度；并记录气象因子的资料. 试用双重筛选方法建立虫情预报公式(见参考文献[16]).

解　预测的指标(因变量)有：

Y_1——本月上旬有虫株率(%)，

Y_2——本月上旬虫口密度(虫数/株)，

Y_3——本月下旬有虫株率(%)，

Y_4——本月下旬虫口密度(虫数/株).

对以上几个指标可能有影响的因素(自变量)有：

$$\left.\begin{array}{l} x_1 \text{——} \cos\left(\dfrac{2\pi}{12}i\right) \\ x_2 \text{——} \sin\left(\dfrac{2\pi}{12}i\right) \end{array}\right\} (i=1,2,\cdots,12)，(x_1,x_2)\text{表示月份，}$$

x_3——上月的气温(℃)，

x_4——上月相对湿度(%)，

x_5——上月雨量(mm)，

x_6——上月气压(mbar)，

x_7——上月蒸发量(mm)，

x_8——上月日照时数，

x_9——上月上旬有虫株率(%)，

x_{10}——上月上旬虫口密度(虫数/株)，

x_{11}——上月下旬有虫株率(%)，

x_{12}——上月下旬虫口密度(虫数/株)，

x_{13}——上月防治否，这是定性变量，若 $x_{13}=1$ 表示上月进行防治；若 $x_{13}=0$ 表示上月没有防治.

以下是考虑交互作用后引入的自变量：

x_{14}——上月防治与否(x_{13})和上月上旬有虫株率(x_9)的交互作用：$x_{14}=x_{13}\times x_9$，

x_{15}——x_{13}和上月上旬虫口密度(x_{10})的交互作用：$x_{15}=x_{13}\times x_{10}$，

x_{16}——x_{13}和上月下旬有虫株率(x_{11})的交互作用：$x_{16}=x_{13}\times x_{11}$，

x_{17}——x_{13}和上月下旬虫口密度(x_{12})的交互作用：$x_{17}=x_{13}\times x_{12}$，

x_{18}——x_{13}和上月相对湿度(x_4)的交互作用：$x_{18}=x_{13}\times x_4$.

此例因变量共有 $p=4$ 个，自变量个数 $m=18$. 部分原始数据见表 4.3 (1976 年 1 月至 1980 年 11 月 20 个点的平均资料).

表 4.3 部分原始数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
1	0.866	0.5	12.6	67	49.3	1021.1	128.5	161.4	3.3	0.05	3.3
2	0.5	0.866	13.6	69	2.5	1020.2	130.1	159.0	3.3	0.037	7.0
3	0.0	1.0	15.7	81	21.3	1015.0	93.1	94.1	7.0	0.06	1.7
4	−0.5	0.866	16.5	85	44.8	1014.3	87.0	49.7	8.0	0.11	8.3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

序号	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	Y_1	Y_2	Y_3	Y_4
1	0.05	0	0	0	0	0	0	3.3	0.037	7.0	0.06
2	0.06	0	0	0	0	0	0	7.0	0.06	1.7	0.03
3	0.03	0	0	0	0	0	0	8.0	0.11	8.3	0.26
4	0.26	1	8.0	0.11	8.3	0.26	85	9.0	0.69	4.6	0.07
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

由调查资料发现，有虫株率和虫口密度这两个指标很重要，如虫口密度较大，而有虫株率较低时，是幼龄的群集期；当这两个指标同时下降到较低点时，是世代交替期；当两指标同时上升时，是成灾速发期等. 要预测的指标是 4 个，它们之间可能有联系，自变量(因素)共 18 个，有的因素对指标影响大，有的可能影响很小，用双重筛选方法将指标分组以构造预报公式. 计算结果如下：当取 $F_x=2$、$F_y=1$ 时(计算该例的软件通过规定筛选因变量和自变量的临界值 F_y 和 F_x 来筛选变量)，将 4 个指标分为两组.

第一组包括 Y_1 和 Y_3，其回归方程分别为：

$$\hat{Y}_1=11.117-0.036x_8+0.874x_9-0.168x_{18},$$

复相关系数 $R_1=0.88$；

$$\hat{Y}_3=22.548-0.095x_8+0.832x_9-0.300x_{18},$$

复相关系数 $R_3=0.86$.

第二组包括 Y_2 和 Y_4，其回归方程分别为：

$$\hat{Y}_2=-131.838+17.806x_1+1.579x_4+0.553\ x_9+0.323x_{12},$$

复相关系数 $R_2=0.79$；

$$\hat{Y}_4=-159.4+23.087x_1+1.952x_4+0.505\ x_9+0.347x_{12},$$

复相关系数 $R_4=0.74$.

筛选的结果将 Y_1 和 Y_3 归为一组，即本月的有虫株率主要与上月日照时数(x_8)、上月上旬有虫株率(x_9)、上月防治与否和相对湿度交互作用(x_{18})相关. Y_2 与 Y_4 归为一组，本月的虫口密度与上月上旬有虫株率(x_9)及下月下旬虫口密度(x_{12})均显著相关；且与月份及相对湿度也显著相关.

当取 $F_x=F_y=8$ 时，4 个指标分为四组. 因 F_y 很大，相当于对 4 个指标分别作逐步回归. 所得回归方程分别为：

$\hat{Y}_1=4.614+0.869x_9$，　　复相关系数 $R_1=0.87$；

$\hat{Y}_2=6.164+0.820x_9$，　　复相关系数 $R_2=0.81$；

$\hat{Y}_3=-3715+0.466\ x_9+0.402\ x_{12}$，　复相关系数 $R_3=0.73$；

$\hat{Y}_4=7.270+0.636\ x_{12}$，　　复相关系数 $R_4=0.63$.

由以上计算结果可见，影响本月虫情的最本质因素是上月的虫情. 当进行虫情预报时，仅侧重气象因子的作用是不够全面的，因为虫口原来的状况的重要性超过了气象因子，而且人为措施在害虫数量变动中也起了重要作用，这些都必须考虑在内.

为了检验预测公式的精确度，我们用未参加计算的 1980 年 12 月的资料代入以上两种筛选临界值的回归方程，计算结果和实测值见表 4.4. 由表可见，Y_1，Y_2，Y_3 的预测结果比 Y_4 好.

表 4.4　预测结果

因变量	实测值	预测值	
		$F_x=2,F_y=1$	$F_x=F_y=8$
Y_1	42.4	42.86	39.03
Y_2	41.0	47.82	38.64
Y_3	7.3	9.36	6.95
Y_4	2.3	10.05	4.05

用双重筛选逐步回归还可以对松毛虫虫情进行较长期的预报.例如不仅对本月的松毛虫发生进行分析与预测,同时对今后两个月或更长时间的虫情也可进行分析与预测,当然这要求有长期的虫情发生的历史资料.

习 题 四

4-1 设

$$\begin{cases} y_1 = a + \varepsilon_1, \\ y_2 = 2a - b + \varepsilon_2, \\ y_3 = a + 2b + \varepsilon_3, \end{cases} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} \sim N_3(0, \sigma^2 I_3).$$

(1) 试求参数 a,b 的最小二乘估计;

(2) 试导出检验 $H_0: a=b$ 的似然比统计量,并指出当假设成立时,这个统计量的分布是什么?

4-2 在多元线性回归模型(4.1.3)中($p=1$),试求出参数向量 β 和 σ^2 的最大似然估计.

4-3 设 Y 与 x_1,x_2,x_3 有相关关系,其 8 组观测数据见表 4.5.

表 4.5 观测数据

序号	x_1	x_2	x_3	Y
1	38	47.5	23	66.0
2	41	21.3	17	43.0
3	34	36.5	21	36.0
4	35	18.0	14	23.0
5	31	29.5	11	27.0
6	34	14.2	9	14.0
7	29	21.0	4	12.0
8	32	10.0	8	7.6

(1) 设 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon$,试求回归方程及决定系数 R^2 和均方误差 s^2;

(2) 考虑二次回归模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2$$

$$+\beta_6 x_3^2+\beta_7 x_1x_2+\beta_8 x_1x_3+\beta_9 x_2x_3+\varepsilon,$$

用逐步回归法筛选变量($\alpha_{in}=\alpha_{out}=0.05$),并写出决定系数 R^2 和均方误差 s.

4-4　试对第一章表 1.2 给出的肺活量数据建立肺活量(OXY)与其他 6 个变量的回归方程:

(1) 求肺活量(OXY)与其他 6 个变量的回归方程,并写出决定系数 R^2 和均方根(Root MSE)s;

(2) 用逐步回归方法建立"最优"回归方程($\alpha=0.15$ 和 $\alpha=0.05$);

(3) 用全子集法在修正 R^2 准则下求最优回归方程.

4-5　考虑 Y 与 $x_1,x_2,\cdots,x_m$ 的逐步回归,由

$$A^{(0)}=\left[\begin{array}{c|c} X'X & X'Y \\ \hline Y'X & Y'Y \end{array}\right]$$

出发,第一步引入 x_{i_1},记 $A^{(1)}=T_{i_1}(A^{(0)})$;第二步引入 x_{i_2},记 $A^{(2)}=T_{i_2}(A^{(1)})$. 证明第三步不可能剔除变量(用反证法).

4-6　称观测向量 Y 和估计向量 $\hat{Y}$ 的相关系数 R 为全相关系数,即

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}\quad\left(\text{其中 }\bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i\right).$$

试证明:(1) $\bar{\hat{y}}=\bar{y}$;

(2) $R^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2\Big/\sum_{i=1}^{n}(y_i-\bar{y})^2$;

(3) 残差平方和 $Q(\hat{\beta})=(1-R^2)\sum_{i=1}^{n}(y_i-\bar{y})^2$.

4-7　在多因变量的多元线性回归模型中,给定 $Y_{n\times p}$,$X_{n\times m}$,且 $\mathrm{rank}(X)=m$. 记 $C=(\mathbf{1}_n\mid X)$. 则

$$\begin{aligned}Q(\beta)&=(Y-C\beta)'(Y-C\beta)\\&=(Y-C\hat{\beta})'(Y-C\hat{\beta})+(\hat{\beta}-\beta)'C'C(\hat{\beta}-\beta),\end{aligned}$$

其中 $\hat{\beta}=(C'C)^{-1}C'Y$.

4-8 在多对多的回归模型中，令 $Q(\beta)=(Y-C\beta)'(Y-C\beta)$. 试证明 $\hat{\beta}=(C'C)^{-1}C'Y$ 在下列四种意义下达最小：

(1) $\mathrm{tr}Q(\hat{\beta})\leqslant \mathrm{tr}Q(\beta)$；

(2) $Q(\hat{\beta})\leqslant Q(\beta)$；

(3) $|Q(\hat{\beta})|\leqslant |Q(\beta)|$；

(4) $\mathrm{ch}_1(Q(\hat{\beta}))\leqslant \mathrm{ch}_1(Q(\beta))$，其中 $\mathrm{ch}_1(A)$ 表示 A 的最大特征值.

以上 β 是 $(m+1)\times p$ 的任意矩阵.

4-9 设多对多回归模型为

$$\begin{cases} Y=(1_n \vdots X_1 \vdots X_2)\begin{bmatrix} b_{(0)} \\ B_1 \\ B_2 \end{bmatrix}+E=C\beta+E, \\ E\sim N_{n\times p}(O, I_n\otimes \Sigma), \end{cases}$$

其中 X_1 和 X_2 均为 $n\times q$ 的数据阵，E 和 O 均为 $n\times p$ 的数据阵，B_1 和 B_2 均为 $q\times p$ 参数矩阵.

(1) 试写出以上模型中当 $B_1=B_2\overset{\text{def}}{=\!=}B$ 时，B 的最小二乘估计和 Σ 的无偏估计量；

(2) 试导出检验 H_0：$B_1=B_2$ 的似然比统计量.

4-10 考虑洛河在某河段河水受污染情况. 考察的指标(因变量)有两个，Y_1 表示 BOD 浓度；Y_2 表示氧亏浓度. 而 Y_1，Y_2 又与以下几个因素(自变量)有关：

x_1——初始断面的 BOD 浓度 L_0，

x_2——初始断面的氧亏浓度 C_0，

x_3——水温 T，

x_4——河流流量 Q，

x_5——排污口流量 g，

x_6——污水 BOD 浓度 l，

x_7——流过该河段所需时间 t.

共观测了 15 组数据(见表 4.6)，试用逐步回归或双重筛选逐步回归求出 BOD 浓度 Y_1、氧亏浓度 Y_2 与 $x_1,x_2,\cdots,x_7$ 的回归方程.

表 4.6　水污数据

序号	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y_1	Y_2
1	6.88	−0.25	27.0	67.4784	1.1232	477.0	0.083	9.35	−2.66
2	6.08	−2.21	27.5	47.7792	1.1232	193.0	0.083	12.30	−4.02
3	2.14	−3.04	26.0	47.7792	1.1232	404.0	0.083	15.60	−4.59
4	5.02	−0.73	26.0	85.6224	1.1232	363.0	0.073	5.88	−3.96
5	7.89	−2.26	26.0	85.6224	1.1232	363.0	0.069	6.34	−3.02
6	2.38	−1.65	15.0	149.0400	1.5552	428.0	0.104	4.00	−1.74
7	1.86	−1.35	15.8	149.0400	1.5552	428.0	0.104	3.76	−1.47
8	1.02	−2.12	17.1	149.4720	1.3824	428.0	0.104	3.98	−2.33
9	1.22	−1.92	17.5	149.4720	1.3824	428.0	0.104	3.98	−2.19
10	0.90	−0.27	17.0	362.8800	0.9936	202.0	0.104	2.78	0.33
11	2.58	−0.09	17.0	362.8800	0.9936	202.0	0.104	1.88	0.23
12	2.78	−1.17	13.5	326.5920	0.9936	114.0	0.104	2.56	−0.74
13	2.10	−1.30	13.5	326.5920	0.9936	114.0	0.104	2.72	−0.80
14	2.32	−0.60	14.5	364.6080	0.8640	57.3	0.104	1.64	−0.62
15	2.96	−0.60	14.5	364.6080	0.8640	57.3	0.104	2.36	−0.32

第五章　判别分析

判别分析是用于判断样品所属类型的一种统计分析方法. 在生产、科研和日常生活中经常遇到如何根据观测到的数据资料对所研究的对象进行判别归类的问题. 例如：在医学诊断中，一个病人肺部有阴影，医生要判断他患的是肺结核、肺部良性肿瘤还是肺癌？这里由肺结核病人、良性肿瘤病人、肺癌病人组成三个总体，病人来源于这三个总体之一. 判别分析的目的是通过测得病人的指标(阴影的大小、边缘是否光滑、体温多少……)来判断他应该属哪个总体(即判断他患的什么病).

在气象学中，根据已有气象资料(气温、气压、湿度等)来判断明天是阴天还是晴天，是有雨还是无雨. 在经济学中，根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属的类型. 股票持有者根据某种股票近期的变化情况判断此种股票价格下一周是上升还是下跌. 在市场预测中，根据以往调查所得的种种指标，判断下季度(或下个月)产品是畅销、平常或滞销？在考古学中，根据挖掘出来的人头盖骨的高、宽等特征来判别其民族或性别. 在环境科学中，根据某地区的气象条件，以及大气污染元素浓度等来判断该地区是属严重污染、一般污染还是无污染. 在地质勘探中，需要从岩石标本的多种特征来判断地层的地质年代，是有矿还是无矿，是富矿还是贫矿. 在农林虫害预报中，根据以往的虫情及多种气象因子判别一个月后的虫情是大发生、中发生或正常. 在体育运动中，根据运动员的多项运动指标来判定游泳运动员的“苗子”是适合练蛙泳、仰泳还是自由泳等等.

总之，判别分析是应用性很强的一种多元统计方法，已渗透到各个领域. 但不管是哪个领域，判别分析问题都可以这样描述：设有 k 个 m 维总体 $G_1, G_2, \cdots, G_k$，其分布特征已知(如已知分布函数分别为

$F_1(x),F_2(x),\cdots,F_k(x)$，或知道来自各个总体的训练样本）. 对给定的一个新样品 X，我们要判断它来自哪个总体.

在进行判别归类时，由假设的前提，判别的依据及处理的手法不同，可得出不同判别方法. 如距离判别，贝叶斯(Bayes)判别，费希尔(Fisher)判别，逐步判别，序贯判别等. 本章介绍几个常用的判别方法.

§5.1　距离判别

距离判别的基本思想是：样品和哪个总体距离最近，就判断它属哪个总体. 距离判别也称为**直观判别法**.

一、马氏距离

已知有两个类 G_1 和 G_2，比如 G_1 是设备 A 生产的产品，G_2 是设备 B 生产的同类产品. 设备 A 的产品质量高(如考察指标为耐磨度 X)，其平均耐磨度 $\mu^{(1)}=80$，反映设备精度的方差 $\sigma_1^2=0.25$；设备 B 的产品质量稍差，其平均耐磨度 $\mu^{(2)}=75$，反映设备精度的方差 $\sigma_2^2=4$. 今有一产品 X_0，测得耐磨度 $x_0=78$，试判断该产品是哪一台设备生产的？

直观地看，x_0 与 $\mu^{(1)}$(设备 A)的绝对距离近些，按距离最近的原则是否应把该产品 X_0 判断为设备 A 生产的？

下面考虑一种相对于分散性的距离. 记 X_0 与 G_1 或 G_2 的相对平方距离为 $d_1^2(x_0)$ 或 $d_2^2(x_0)$，则有：

$$d_1^2(x_0)=\frac{(x_0-\mu^{(1)})^2}{\sigma_1^2}=\frac{(78-80)^2}{0.25}=16,$$

$$d_2^2(x_0)=\frac{(x_0-\mu^{(2)})^2}{\sigma_2^2}=\frac{(78-75)^2}{4.00}=2.25.$$

因为 $d_2(x_0)=1.5<4=d_1(x_0)$，按这种距离准则应判 X_0 为设备 B 生产的. 从图 5.1 可以看出，设备 B 生产的产品质量较分散，出现 x_0 为 78 的可能性较大；而设备 A 生产的产品质量较集中，出现 x_0 为 78

的可能性较小. 判断 X_0 为设备 B 生产的产品更合理.

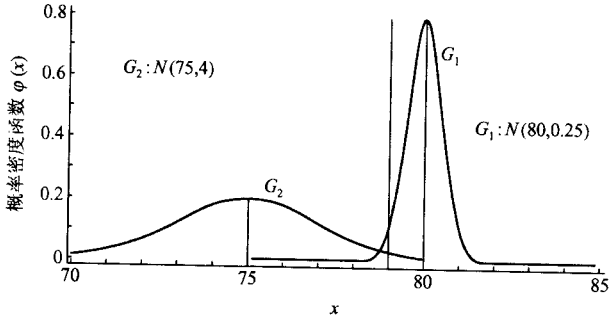

图 5.1 利用马氏距离对两个一元正态总体判别归类的示意图

一般地,我们假设总体 G_1 的分布为 $N(\mu^{(1)},\sigma_1^2)$,总体 G_2 的分布为 $N(\mu^{(2)},\sigma_2^2)$,则利用相对距离的定义,可以找出分界点 μ^*(不妨设 $\mu^{(2)}<\mu^{(1)}$),令

$$\frac{(x-\mu^{(1)})^2}{\sigma_1^2}=\frac{(x-\mu^{(2)})^2}{\sigma_2^2}\Rightarrow x=\frac{\mu^{(1)}\sigma_2+\mu^{(2)}\sigma_1}{\sigma_1+\sigma_2}\xlongequal{\text{def}}\mu^*,$$

而按这种距离最近的判别准则为:$\begin{cases}\text{判 } X\in G_1, & x>\mu^*,\\ \text{判 } X\in G_2, & x\leqslant\mu^*.\end{cases}$

因只有一个指标,这时判别函数为:$Y=Y(x)=x$. 此例中 $\mu^*=79$,因 $x_0=78<\mu^*$,故判 $X_0\in G_2$. 下面给出一般 m 元总体中这种相对距离——马氏(全称:马哈拉诺比斯(Mahalanobis))距离的定义.

定义 5.1.1(马氏距离) 设总体 G 为 m 元总体(考察 m 个指标),均值向量为 $\mu=(\mu_1,\mu_2,\cdots,\mu_m)'$,协方差阵为 $\Sigma=(\sigma_{ij})_{m\times m}$,则样品 $X=(x_1,x_2,\cdots,x_m)'$ 与总体 G 的**马氏距离**定义为

$$d^2(X,G)=(X-\mu)'\Sigma^{-1}(X-\mu).$$

当 $m=1$ 时,

$$d^2(x,G)=\frac{(x-\mu)'(x-\mu)}{\sigma^2}=\frac{(x-\mu)^2}{\sigma^2}.$$

二、两总体的距离判别

先考虑两个总体($k=2$)的情况. 设有两个总体 G_1 和 G_2,已知来

自 G_i $(i=1,2)$的训练样本为

$$X_{(t)}^{(i)}=(x_{t1}^{(i)},x_{t2}^{(i)},\cdots,x_{tm}^{(i)})' \quad (i=1,2;t=1,2,\cdots,n_i),$$

其中 n_i 是取自 G_i 的样品个数，则总体 G_i 的均值向量 $\mu^{(i)}$ 的估计量为

$$\overline{X}^{(i)}=\left(\frac{1}{n_i}\sum_{t=1}^{n_i}x_{t1}^{(i)},\cdots,\frac{1}{n_i}\sum_{t=1}^{n_i}x_{tm}^{(i)}\right)'=(\overline{x}_1^{(i)},\overline{x}_2^{(i)},\cdots,\overline{x}_m^{(i)})'.$$

总体 G_i 的协方差阵 Σ_i 的估计 S_i（称为**组内协方差阵**）为

$$S_i=\frac{1}{n_i-1}A_i=(s_{lj}^{(i)})_{m\times m},$$

其中 $A_i=\sum_{t=1}^{n_i}(X_{(t)}^{(i)}-\overline{X}^{(i)})(X_{(t)}^{(i)}-\overline{X}^{(i)})'$ 称为**组内离差阵**；

$$s_{lj}^{(i)}=\frac{1}{n_i-1}\sum_{t=1}^{n_i}(x_{tl}^{(i)}-\overline{x}_l^{(i)})(x_{tj}^{(i)}-\overline{x}_j^{(i)})$$
$$(l,j=1,\cdots,m).$$

当假定 $\Sigma_1=\Sigma_2\overset{\text{def}}{=\!=}\Sigma$ 时，反映分散性的协方差阵 Σ 的估计为

$$S=\frac{1}{n-k}\sum_{i=1}^{k}A_i=(s_{lj})_{m\times m},$$

并称 S 为**合并样本协方差阵**，其中

$$s_{lj}=\frac{1}{n-k}\sum_{i=1}^{2}\sum_{t=1}^{n_i}(x_{tl}^{(i)}-\overline{x}_l^{(i)})(x_{tj}^{(i)}-\overline{x}_j^{(i)})$$
$$(l,j=1,2,\cdots,m).$$

问题是对任给定的 m 维样品 $X=(x_1,x_2,\cdots,x_m)'$，要判断它来自哪个总体.

1. $\Sigma_1=\Sigma_2$ 时的判别方法

一个最直观的想法是，分别计算样品 X 到两个总体的距离 $d_1^2(X)$ 和 $d_2^2(X)$（或记为 $d^2(X,G_1)$ 和 $d^2(X,G_2)$），并按**距离最近准则**判别归类，判别准则为[①]：

$$\begin{cases}\text{判 } X\in G_1, & \text{当 } d^2(X,G_1)<d^2(X,G_2)\text{ 时},\\ \text{判 } X\in G_2, & \text{当 } d^2(X,G_1)\geqslant d^2(X,G_2)\text{ 时};\end{cases}$$

或

① 本章一般采用第一种形式的准则.

$$\begin{cases} \text{判 } X \in G_1, & \text{当 } d^2(X,G_1) < d^2(X,G_2) \text{ 时}, \\ \text{判 } X \in G_2, & \text{当 } d^2(X,G_1) > d^2(X,G_2) \text{ 时}, \\ \text{待判}, & \text{当 } d^2(X,G_1) = d^2(X,G_2) \text{ 时}, \end{cases}$$

这里的距离是指马氏距离. 利用马氏距离的定义及两总体协方差阵相等的假设,可以简化马氏距离的计算公式:

$$\begin{aligned} d^2(X,G_i) &= (X - \overline{X}^{(i)})' S^{-1} (X - \overline{X}^{(i)}) \\ &= X' S^{-1} X - 2\left[(S^{-1}\overline{X}^{(i)})' X - \frac{1}{2} (\overline{X}^{(i)})' S^{-1} \overline{X}^{(i)} \right] \\ &= X' S^{-1} X - 2Y_i(X) \quad (i = 1,2), \end{aligned}$$

其中 $Y_i(X)$ 是 X 的线性函数. 对给定样品 X,为计算 X 到各总体的马氏距离,只须计算 $Y_i(X)$:

$$Y_i(X) = (S^{-1}\overline{X}^{(i)})' X - \frac{1}{2} (\overline{X}^{(i)})' S^{-1} \overline{X}^{(i)} \quad (i = 1,2).$$

$Y_i(X)$ 称为**线性判别函数**, $a_i = S^{-1}\overline{X}^{(i)}$ 称为**判别系数向量**, $c_i = -\frac{1}{2} (\overline{X}^{(i)})' S^{-1} \overline{X}^{(i)}$ 称为**常数项**.

若考察这两个马氏距离之差,经计算可得:

$$\begin{aligned} d_2^2(X) - d_1^2(X) &= 2\left(X - \frac{1}{2} (\overline{X}^{(1)} + \overline{X}^{(2)}) \right)' S^{-1} (\overline{X}^{(1)} - \overline{X}^{(2)}) \\ &\xlongequal{\text{def}} 2W(X), \end{aligned}$$

其中

$$W(X) = (X - X^*)' S^{-1} (\overline{X}^{(1)} - \overline{X}^{(2)}),$$

$$X^* = \frac{1}{2} (\overline{X}^{(1)} + \overline{X}^{(2)}).$$

则判别准则还可以写为: $\begin{cases} \text{判 } X \in G_1, & \text{当 } W(X) > 0 \text{ 时}, \\ \text{判 } X \in G_2, & \text{当 } W(X) \leqslant 0 \text{ 时}. \end{cases}$

$W(X)$ 是 X 的线性函数,即 $W(X) = a'(X - X^*)$,其中 $a = S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})$. $W(X)$ 也称为**线性判别函数**,a 为**判别系数**.

$W(X)$ 把 m 维空间 $\mathbb{R}^m$ 划分为两个部分: $D_1 = \{X: W(X) > 0\}$ 和 $D_2 = \{X: W(X) \leqslant 0\}$,即 D_1, D_2 是 $\mathbb{R}^m$ 的一个划分. 显然,判别方法的最终结果是得到 $\mathbb{R}^m$ 中的一个划分. 由判别函数 $W(X)$ 得到划分 D_1, D_2,当样品 X 落入 D_1 时判 $X \in G_1$;当 X 落入 D_2 时判 $X \in G_2$.

下面考察 $m=1$ 的特殊情况，并设两总体为正态总体，已知其分布为 $N(\mu^{(1)},\sigma^2)$和 $N(\mu^{(2)},\sigma^2)$(两总体的方差相同，记为 σ^2)，这时判别函数为

$$W(x)=\left(x-\frac{\mu^{(1)}+\mu^{(2)}}{2}\right)\frac{1}{\sigma^2}(\mu^{(1)}-\mu^{(2)})=a(x-\bar{\mu}),$$

其中 $\bar{\mu}=\dfrac{\mu^{(1)}+\mu^{(2)}}{2}$，$a=\dfrac{\mu^{(1)}-\mu^{(2)}}{\sigma^2}$. 不妨设 $\mu^{(1)}>\mu^{(2)}$，则 a 为正数，$W(x)$的符号取决于 $x>\bar{\mu}$ 或 $x\leqslant\bar{\mu}$. 当 $x>\bar{\mu}$ 时判样品 $X\in G_1$；当 $x\leqslant\bar{\mu}$ 时判样品 $X\in G_2$. 从图 5.2 可以看出，用这种判别法会发生错判，如 X 来自 G_1，但却落入 D_2，被判为属于 G_2. 错判的概率为图 5.2 中阴影左半部分的面积，并记为 $P(2|1)$. 类似有 $P(1|2)$.

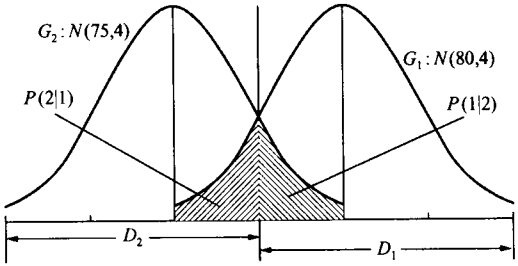

图 5.2　两个一元正态总体距离判别法($\sigma_1=\sigma_2$)的示意图

经计算可得 $P(2|1)=P(1|2)=1-\Phi\left(\dfrac{\mu^{(1)}-\mu^{(2)}}{2\sigma}\right)$. 比如当 $\mu^{(1)}=80,\mu^{(2)}=75,\sigma=2$ 时，$P(2|1)=0.1056$. 由错判概率的公式及图 5.2 可见，当两总体均值靠得很近(即 $|\mu^{(1)}-\mu^{(2)}|$很小)时，则错判概率很大，这时作判别分析是没有意义的. 因此只有当两总体的均值有显著性差异时，作判别分析才有意义.

2. $\Sigma_1\neq\Sigma_2$ 时的判别方法

当两总体协方差阵不等时，按距离判别准则先分别计算 X 到两个总体的距离 $d^2(X,G_1)$和 $d^2(X,G_2)$，然后按距离最近准则判别归类，或者类似地计算判别函数 $W(X)$，并用于判别归类. 令

$$W(X)=d^2(X,G_2)-d^2(X,G_1)=\cdots\overset{\text{def}}{=\!=}Z(X)-Z_0,$$

其中 $Z(X)$ 为 X 的二次函数(因 $\Sigma_1 \neq \Sigma_2$), Z_0 是一常数(具体表达式省略了). 判别准则仍可以写为:

$$\begin{cases} \text{判 } X \in G_1, & \text{当 } W(X) > 0 \text{ 时}, \\ \text{判 } X \in G_2, & \text{当 } W(X) \leqslant 0 \text{ 时}. \end{cases}$$

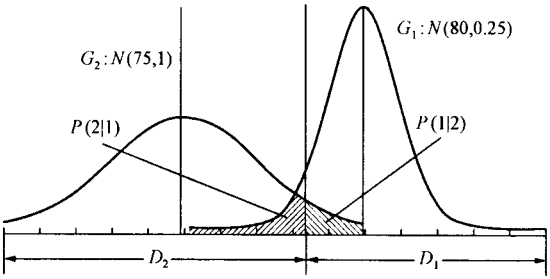

图 5.3 两个一元正态总体距离判别法($\sigma_1 \neq \sigma_2$)的示意图

当 $m=1$,两总体为正态总体时,记 G_i 的均值为 $\mu^{(i)}$,方差为 σ_i^2 $(i=1,2)$,这时马氏距离的平方根为

$$d_i(x) = \frac{|x-\mu^{(i)}|}{\sigma_i} \quad (i=1,2).$$

不妨设 $\mu^{(2)} < \mu^{(1)}$,当观测值 x 满足: $\mu^{(2)} < x < \mu^{(1)}$ 时,

$$d_2(x) - d_1(x) = \frac{x-\mu^{(2)}}{\sigma_2} - \frac{\mu^{(1)}-x}{\sigma_1} = \frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}(x-\mu^*),$$

其中

$$\mu^* = \frac{\mu^{(1)}\sigma_2 + \mu^{(2)}\sigma_1}{\sigma_1+\sigma_2},$$

它是 $\mu^{(1)}, \mu^{(2)}$ 的加权平均值(见图 5.3). 它把直线分为两部分: $D_1 = \{x > \mu^*\}$ 和 $D_2 = \{x \leqslant \mu^*\}$. 这时判别准则为

$$\begin{cases} \text{判 } X \in G_1, & \text{当 } x > \mu^* \text{ 时}, \\ \text{判 } X \in G_2, & \text{当 } x \leqslant \mu^* \text{ 时}. \end{cases}$$

当 $\sigma_1 = \sigma_2$ 时, $\mu^* = \bar{\mu}$. μ^* 或 $\bar{\mu}$ 常称为**阈值点**(分界点),阈值如何选取很重要,取得不当,错判概率将明显增加.

例 5.1.1 (盐泉含钾性判别) 某地区经勘探证明, A 盆地是一个钾盐矿区, B 盆地是一个钠盐(不含钾)矿区,其他盐盆地是否含

钾盐有待作出判断. 今从 A 和 B 两盆地各抽取 5 个盐泉样品；从其他盆地抽得 8 个盐泉样品，18 个盐泉的特征数值见表 5.1. 试对后 8 个待判盐泉进行含钾性判别.

表 5.1　盐泉的特征数值

盐泉类别	序号	$K\cdot10^3/Cl$ (X_1)	$Br\cdot10^3/Cl$ (X_2)	$K\cdot10^3/\Sigma$ 盐 (X_3)	K/Br (X_4)	类别号
第一类：含钾盐泉（A 盆地）	1	13.85	2.79	7.80	49.60	A
	2	22.31	4.67	12.31	47.80	A
	3	28.82	4.63	16.18	62.15	A
	4	15.29	3.54	7.50	43.20	A
	5	28.79	4.90	16.12	58.10	A
第二类：含钠盐泉（B 盆地）	6	2.18	1.06	1.22	20.60	B
	7	3.85	0.80	4.06	47.10	B
	8	11.40	0.00	3.50	0.00	B
	9	3.66	2.42	2.14	15.10	B
	10	12.10	0.00	5.68	0.00	B
待判盐泉	1	8.85	3.38	5.17	26.10	
	2	28.60	2.40	1.20	127.00	
	3	20.70	6.70	7.60	30.20	
	4	7.90	2.40	4.30	33.20	
	5	3.19	3.20	1.43	9.90	
	6	12.40	5.10	4.43	24.60	
	7	16.80	3.40	2.31	31.30	
	8	15.00	2.70	5.02	64.00	

解　把 A 盆地和 B 盆地看作两个不同的总体，并假定两总体协方差阵相等. 本例中变量个数 $m=4$，两类总体各有 5 个训练样品（$n_1=n_2=5$），另有 8 个待判样品. 使用 SAS/STAT 软件中的 DISCRIM 过程进行判别归类.

计算结果，首先给出两组间的平方距离（即马氏距离）为 37.02876，检验 H_0：$\mu^{(1)}=\mu^{(2)}$ 的 F 统计量为 14.46436，相应的 $p=0.0059<0.01$，这说明 A 和 B 两盆地的盐泉特征有显著差异，因此讨论判别归类问题是有意义的.

然后得出线性判别函数为：

$$Y_1(X)=-42.2473+7.6741X_1+5.5488X_2$$
$$-13.9631X_3+1.1813X_4,$$
$$Y_2(X)=-5.1627+2.9311X_1+1.3570X_2$$

$$-5.3738X_3+0.4558X_4.$$

回判结果给出对来自A或B盆地的10个盐泉样品都判对了；对8个待判样品判别的结果为：第2,3,6,7,8五个盐泉为含钾盐泉，其余三个不含钾，即为含钠盐泉.

三、多总体的距离判别

设有 k 个 m 元总体：$G_1,G_2,\cdots,G_k$ $(k>2)$. 它们的均值向量和协方差阵分别为 $\mu^{(i)}$, Σ_i $(i=1,2,\cdots,k)$. 对任给定的 m 元样品 $X=(x_1,x_2,\cdots,x_m)'$，要判断它来自哪个总体.

多个总体的情况，按距离最近的准则对 X 进行判别归类时，首先计算样品 X 到 k 个总体的马氏距离 $d_i^2(X)$ $(i=1,2,\cdots,k)$，然后进行比较，把 X 判归距离最小的那个总体. 设 $i=l$ 时，若

$$d_l^2(X)=\min_{i=1,\cdots,k}\{d_i^2(X)\},$$

则 $X\in G_l$.

计算马氏距离 $d_i^2(X)$ $(i=1,2,\cdots,k)$时，类似地可考虑 $\Sigma_1=\Sigma_2=\cdots=\Sigma_k$ 或 Σ_i 不全相等的两种情况，并用样本统计量作为 $\mu^{(i)}$和 Σ_i 的估计进行计算.

§5.2 贝叶斯(Bayes)判别法及广义平方距离判别法

距离判别只要求知道总体的特征量(即参数)——均值和协方差阵，不涉及总体的分布类型. 当参数未知时，就用样本均值和样本协方差阵来估计. 距离判别方法简单，结论明确，是很实用的方法. 但该方法也有缺点：一是该判别法与各总体出现的机会大小(先验概率)完全无关；二是判别方法没有考虑错判造成的损失，这是不合理的. 贝叶斯判别法正是为解决这两方面问题而提出的判别方法.

贝叶斯的统计思想总是假定对研究的对象已有一定的认识，常用先验概率分布来描述这种认识；然后抽取一个样本，用样本来修正已有的认识(先验概率分布)，得到后验概率分布. 各种统计推断都通

过后验概率分布来进行. 将贝叶斯思想用于判别分析就得到贝叶斯判别法.

在正态总体的假设下,按贝叶斯判别的思想,在错判造成的损失相等时得到的判别函数,其实就是马氏距离判别在考虑先验概率及协方差阵是否相等情况下的推广,故在 SAS/STAT 软件的 DISCRIM 过程中称为**广义平方距离判别法**.

所谓判别方法,就是给出空间 $\mathbb{R}^m$ 的一种划分: $D=\{D_1,D_2,\cdots,D_k\}$. 一种划分对应一种判别方法,不同的划分就是不同的判别方法. 贝叶斯判别法也是给出空间 $\mathbb{R}^m$ 的一种划分.

一、先验概率(先知知识)

设有 k 个总体 $G_1,G_2,\cdots,G_k$. 假设事先对所研究的问题有一定的认识,这种认识常用先验概率来描述,即已知这 k 个总体各自出现的概率(验前概率)为 $q_1,q_2,\cdots,q_k$(显然 $q_i>0,q_1+q_2+\cdots+q_k=1$). 比如研究人群中得癌($G_1$)和没有得癌($G_2$)两类群体的问题,由长期经验知: $q_1=0.001$, $q_2=0.999$. 这组验前概率 $q_1,\cdots,q_k$ 称为**先验概率**.

先验概率是一种权重(比例). 所谓"先验"是指先于我们抽取样品作判别分析之前. 贝叶斯判别法要求给出 q_i $(i=1,2,\cdots,k)$的值. q_i 的赋值方法有以下几种:

(1) 利用历史资料及经验进行估计. 例如某地区成年人中得癌症的概率为 $P(\text{癌})=0.001\overset{\text{def}}{=\!=}q_1$,而 $P(\text{无癌})=0.999\overset{\text{def}}{=\!=}q_2$.

(2) 利用训练样本中各类样品占的比例 n_i/n 做为 q_i 的值,即 $q_i=n_i/n$ $(i=1,\cdots,k)$,其中 n_i 是第 i 类总体的样品数,而 $n=n_1+\cdots+n_k$. 这时要求训练样本是通过随机抽样得到的,各类样品被抽到的机会大小就是验前概率.

(3) 假定 $q_1=q_2=\cdots=q_k=1/k$.

二、广义平方距离

在马氏距离判别的基础上,进一步考虑先验概率及各组内协方

差阵的不同，定义样品 X 到总体 G_t $(t=1,\cdots,k)$的广义平方距离 $D_t^2(X)$或 $D^2(X,G_t)$为：

$$D_t^2(X)=D^2(X,G_t)=d_t^2(X)+g_1(t)+g_2(t),$$

其中

$$g_1(t)=\begin{cases}\ln|S_t|, & \text{若各组的协方差阵 } \Sigma_i \text{ 不全相等},\\ 0, & \text{若各组的协方差阵 } \Sigma_i \text{ 全相等};\end{cases}$$

$$g_2(t)=\begin{cases}-2\ln|q_t|, & \text{若先验概率不全相等},\\ 0, & \text{若先验概率全相等},\end{cases}$$

其中 S_t 为第 t 类的组内样本协方差阵. 由以上公式可见，当 $d_t^2(X)$不变，而某个 q_t 大(即总体 G_t 出现的机会大)时，则 $g_2(t)$变小，故广义平方距离 $D_t^2(X)$也变小，进而判 X 为 G_t 的可能性大.

利用广义平方距离的判别法为：

判 $X\in G_t$，　当 $D_t^2(X)<D_i^2(X)$ 时 $(i\neq t,i=1,\cdots,k)$.

三、后验概率(条件概率)

标准的贝叶斯判别法应该计算后验概率分布. 即计算当样品 X 已知时，它属于 G_t 的概率，记为 $P(G_t|X)$(或 $P(t|X)$)，这个概率作为判别归类的准则，其概率意义更为直观. 假定总体 G_t 的概率密度函数 $f_t(x)$ $(t=1,\cdots,k)$给定，由条件概率的定义可以导出：

$$P(t|X)=P\{X\in G_t|X\text{ 已知}\}=\frac{q_tf_t(x)}{\sum_{i=1}^{k}q_if_i(x)}.$$

若假设 G_i $(i=1,\cdots,k)$为正态总体，其密度函数 $f_i(x)$为

$$f_i(x)=(2\pi)^{-m/2}|\Sigma_i|^{-1/2}\exp(-0.5d_i^2(x)),$$

则 X 属于第 t 组的后验概率为：

$$P(t|X)=\frac{\exp(-0.5D_t^2(x))}{\sum_{i=1}^{k}\exp(-0.5D_i^2(x))},$$

其中 $D_i^2(x)$是 X 到第 i 组的广义平方距离. 采用后验概率的判别准则为

判 $X\in G_t$，　当 $P(t|X)>P(i|X)$ 时 $(i\neq t, i=1,\cdots,k)$.

在正态假设下按后验概率最大进行归类的准则，等价于按广义平方距离最小准则进行归类. 由下面的介绍将知道，按后验概率最大准则归类的判别法就是贝叶斯判别法的一种情况. 一般地，贝叶斯判别法既考虑先验概率的不同，还考虑了错判损失的大小，在这里我们假定错判损失相等.

四、贝叶斯判别准则

所谓**贝叶斯判别准则**，就是给出空间 $\mathbb{R}^m$ 的一个划分：$D=\{D_1, D_2,\cdots,D_k\}$，使得当通过这个划分 D 来判别归类时，所带来的平均损失达到最小.

1. 错判概率和错判损失

当样品 $X\in G_i$，但用判别法 D 判别归类时，却把 X 判归 G_j(即 X 落入区域 D_j，$j\neq i$)，即判错了，我们用 $P(j|i;D)$(或简记为 $P(j|i)$)表示用判别法 D 把实属 G_i 的样品错判为 G_j 的概率. 显然

$$P\{j|i;D\}=\int_{D_j}\cdots\int f_i(x_1,\cdots,x_m)\mathrm{d}x_1\cdots\mathrm{d}x_m$$

$$=\int_{D_j} f_i(X)\mathrm{d}X \quad (j\neq i). \tag{5.2.1}$$

错判概率的估计方法有以下几种：

(1) 利用训练样本作为检验集，即用判别方法对已知类别的样品进行回判，统计判错的个数及占样品总数的比率，作为错判率的估计. 此法得出的估计一般偏低.

(2) 当训练样本足够大时，可留出一些已知类别的样品不参加建立判别准则，而是作为**检验集**，并把错判的比率作为错判率的估计. 此法当检验集较小时，估计的方差大.

(3) 舍一法(或称交叉确认法)，每次留出一个已知类别的样品，而用其余 $n-1$ 个样品建立判别准则，然后对留出的这一个已知类别的样品进行判别归类. 对训练样本中 n 个样品按此法逐个归类后，最后把错判的比率作为错判率的估计.

以上三种估计方法的估计结果在 SAS/STAT 软件的 DISCRIM 过程中都可以得到.

用 $L(j|i;D)$表示样品实属第 i 个总体 G_i,今用判别法 D 判别时将其错判为属 $G_j(j\neq i)$时所造成的损失;在不会引起混淆时,简记为 $L(j|i)$.

在实际问题中,错判的损失可以给出定性的分析,但很难用数值来表示.但应用贝叶斯判别准则时,要求定量地给出 $L(j|i)$. $L(j|i)$的赋值法常用以下两种:

(1) 由经验人为赋值.例如 L(判癌|得肺结核)=10, L(判肺结核|得癌症)=1000.

(2) 假定各种错判损失都相等,即令

$$L(j|i;D)=\begin{cases}1, & i\neq j\\ 0, & i=j\end{cases}\xlongequal{\text{def}}1-\delta_{ij}.$$

2. 关于先验概率的平均损失

有了先验概率的概念后,判别法 D 关于先验概率的错判平均损失 $g(D)$定义为

$$g(D)=\sum_{t=1}^{k}q_t\sum_{j=1}^{k}P(j|t)L(j|t)\xlongequal{\text{def}}\sum_{t=1}^{k}q_t r_t(D),\quad(5.2.2)$$

其中 $r_t(D)$表示实属 G_t 的样品被错判为其他总体的损失.

3. 什么是贝叶斯判别准则

定义 5.2.1 设有 k 个总体: $G_1,G_2,\cdots,G_k$,相应的先验概率为 $q_1,q_2,\cdots,q_k$ $(q_i>0,q_1+\cdots+q_k=1)$. 如果有判别法 D^*,使得 D^*带来的平均损失 $g(D^*)$达最小,即

$$g(D^*)=\min_{\text{一切}D}g(D),$$

则称判别法 D^***符合贝叶斯判别准则**,或称 D^*为**贝叶斯判别的解**.

4. 符合贝叶斯准则的判别法(贝叶斯判别的解)

定理 5.2.1 设有 k 个总体: $G_1,G_2,\cdots,G_k$,已知 G_i 的联合密度函数为 $f_i(X)$,先验概率为 $q_i(i=1,\cdots,k)$,错判损失为 $L(j|i)$,则贝叶斯判别的解 $D^*=\{D_1^*,\cdots,D_k^*\}$为

$$D_t^*=\{X|h_t(X)<h_j(X),j\neq t,j=1,\cdots,k\}\quad(t=1,\cdots,k),$$

其中

$$h_j(X)=\sum_{i=1}^{k}q_iL(j|i)f_i(X),\tag{5.2.3}$$

它表示把样品 X 判归 G_j 的平均损失.

证明　由(5.2.1)、(5.2.2)和(5.2.3)式得:

$$\begin{aligned}g(D^*)&=\sum_{i=1}^{k}q_i\sum_{t=1}^{k}L(t|i)\int_{D_t^*}f_i(X)\mathrm{d}X\\&=\sum_{t=1}^{k}\int_{D_t^*}\sum_{i=1}^{k}q_iL(t|i)f_i(X)\mathrm{d}X\\&=\sum_{t=1}^{k}\int_{D_t^*}h_t(X)\mathrm{d}X.\end{aligned}$$

若 $D=\{D_1,\cdots,D_k\}$ 是 $\mathbb{R}^m$ 上的任一种划分,则它带来的平均损失为

$$g(D)=\sum_{j=1}^{k}\int_{D_j}h_j(X)\mathrm{d}X,$$

于是

$$\begin{aligned}g(D^*)-g(D)&=\sum_{t=1}^{k}\int_{D_t^*}h_t(X)\mathrm{d}X-\sum_{j=1}^{k}\int_{D_j}h_j(X)\mathrm{d}X\\&=\sum_{j=1}^{k}\sum_{t=1}^{k}\int_{D_t^*\cap D_j}[h_t(X)-h_j(X)]\mathrm{d}X.\end{aligned}$$

由 D^* 的定义知,在 D_t^* 上恒有 $h_t(X)<h_j(X)$ $(j=1,\cdots,k)$,所以

$$g(D^*)-g(D)\leqslant 0,$$

即

$$g(D^*)=\min_{\text{一切}D}g(D).$$

由定义 5.2.1 知, D^* 是贝叶斯判别的解. **(证毕)**

以上定理是贝叶斯判别法的基本定理. 它给出了具体的判别方法: 对样品 X,分别计算 k 个 $h_j(X)$ $(j=1,\cdots,k)$,选其最小者,即可判定样品来自相应的总体. 当错判损失都相等时,判别方法还可以由以下的推论给出.

推论　当 $L(j|i)=1-\delta_{ij}$ 时(即错判损失都相等), 则贝叶斯判别的解 $D^*=\{D_1^*,\cdots,D_k^*\}$ 为

$$D_t^*=\{X|q_tf_t(X)>q_jf_j(X),j\neq t,j=1,\cdots,k\}\quad(t=1,\cdots,k),\tag{5.2.4}$$

其中 $f_j(X)$是 G_j 的联合概率密度函数.

证明 由假设可知

$$h_t(X)=\sum_{i=1}^{k}q_iL(t|i)f_i(X)=\sum_{i\neq t}q_if_i(X)=C(X)-q_tf_t(X),$$

其中 $C(X)=\sum_{i=1}^{k}q_if_i(X)$ 为依赖 X 的数值.

由定理 5.2.1,可得(5.2.4)式. **(证毕)**

例 5.2.1 试导出 $k=2$ 时的贝叶斯判别的解.

解 由(5.2.3)式得

$$h_1(X)=q_2f_2(X)L(1|2),\quad h_2(X)=q_1f_1(X)L(2|1),$$

从而

$$D_1=\{X|q_2f_2(X)L(1|2)<q_1f_1(X)L(2|1)\},$$
$$D_2=\{X|q_1f_1(X)L(2|1)\leqslant q_2f_2(X)L(1|2)\}.$$

若令判别函数为

$$W(X)=\frac{f_1(X)}{f_2(X)},\quad d=\frac{q_2L(1|2)}{q_1L(2|1)},\tag{5.2.5}$$

则贝叶斯判别准则为

$$\begin{cases}X\in G_1, & \text{若 } W(X)>d,\\ X\in G_2, & \text{若 } W(X)\leqslant d,\end{cases}$$

这与距离判别准则有相似的形式.

5. 正态总体的贝叶斯判别法

设 G_i 为正态总体 $N_m(\mu^{(i)},\Sigma_i)$ $(i=1,\cdots,k)$,并假定错判损失相等,先验概率为 $q_1,q_2,\cdots,q_k$.

(1) 当 $\Sigma_1=\Sigma_2=\cdots=\Sigma_m\overset{\text{def}}{=\!=}\Sigma$ 时,设总体 G_i 的概率密度函数为 $f_i(X)$,则

$$q_if_i(X)=\frac{q_i}{(2\pi)^{m/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(X-\mu^{(i)})'\Sigma^{-1}(X-\mu^{(i)})\right\},$$

$$\ln q_if_i(X)=-\frac{1}{2}[\ln|\Sigma|+m\ln(2\pi)+X'\Sigma^{-1}X]$$
$$+\ln q_i-\frac{1}{2}(\mu^{(i)})'\Sigma^{-1}\mu^{(i)}+X'\Sigma^{-1}\mu^{(i)}$$

表 5.2　胃癌检验的生化指标值

类别		序号	血清铜蛋白 X_1	蓝色反应 X_2	尿吲哚乙酸 X_3	中性硫化物 X_4
胃癌患者	胃癌患者	1	228	134	20	11
		2	245	134	10	40
		3	200	167	12	27
		4	170	150	7	8
		5	100	167	20	14
非胃癌患者	萎缩性胃炎患者	6	225	125	7	14
		7	130	100	6	12
		8	150	117	7	6
		9	120	133	10	26
		10	160	100	5	10
	非胃炎患者	11	185	115	5	19
		12	170	125	6	4
		13	165	142	5	3
		14	135	108	2	12
		15	100	117	7	2

注：X_3, X_4 是原始数据的 100 倍.

$d^2(2|1)$表示 $\bar{X}^{(2)}$到 G_1 的平方距离. 若 $\Sigma_1=\Sigma_2$，则

$$d^2(2|1) = d^2(1|2);$$

但此例中协方差阵 Σ_1 与 Σ_2 不等，因此

$$d^2(2|1)=22.1219,\quad d^2(1|2)=486.03104.$$

输出结果又给出三个总体间两两配对的组间广义平方距离，用记号 $D^2(2|1)$表示 $\bar{X}^{(2)}$到 G_1 的广义平方距离. 此例中协方差阵 Σ_1 与 Σ_2 不等，因此 $D^2(2|1)=43.06467$，$D^2(1|2)=498.2681$，且知

$$D^2(1|1)=20.9428=\ln|S_1|.$$

最后，输出结果给出回判的结果：三个类中 15 个样品都判对了；判别矩阵汇总了判别归类的结果，并指出错判的比率为 0.

§5.3　费希尔(Fisher)判别

一、费希尔判别的基本思想

费希尔判别的基本思想是**投影**. 将 k 组 m 元数据投影到某一个

方向，使得投影后组与组之间尽可能地分开. 而衡量组与组之间是否分开的方法借助于一元方差分析的思想. 利用方差分析的思想来导出判别函数，这个函数可以是线性的，也可以是很一般的函数. 因线性判别函数在实际应用中最方便，本节仅讨论线性判别函数的导出.

设从总体 G_t $(t=1,\cdots,k)$分别抽取 m 元样本如下：

$$X_{(i)}^{(t)} = (x_{i1}^{(t)},\cdots,x_{im}^{(t)})' \quad (t=1,\cdots,k;i=1,\cdots,n_t).$$

令 $a=(a_1,\cdots,a_m)'$ 为 m 维空间的任一向量，$u(x)=a'X$ 为 X 向以 a 为法线方向上的投影. 上述 k 个组中的 m 元数据投影后为

$$G_1: a'X_{(1)}^{(1)},\cdots,a'X_{(n_1)}^{(1)}, \quad \text{记 } \overline{X}^{(1)} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{(j)}^{(1)},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots \qquad \cdots\cdots\cdots\cdots\cdots\cdots$$

$$G_k: a'X_{(1)}^{(k)},\cdots,a'X_{(n_k)}^{(k)}, \quad \text{记 } \overline{X}^{(k)} = \frac{1}{n_k}\sum_{j=1}^{n_k} X_{(j)}^{(k)}.$$

每个总体的数据投影后均为一元数据. 对这 k 组一元数据进行一元方差分析，其组间平方和为

$$\begin{aligned} B_0 &= \sum_{t=1}^{k} n_t(a'\overline{X}^{(t)} - a'\overline{X})^2 \\ &= a'\left[\sum_{t=1}^{k} n_t(\overline{X}^{(t)} - \overline{X})(\overline{X}^{(t)} - \overline{X})'\right]a \\ &= a'Ba, \end{aligned}$$

其中 $\overline{X}^{(t)}$ 和 $\overline{X}$ 分别为 G_t 的样本均值和总样本均值，并记

$$\overline{X} = \frac{1}{n}\sum_{t=1}^{k}\sum_{j=1}^{n_t} X_{(j)}^{(t)},$$

而 B 为组间离差阵：

$$B = \sum_{t=1}^{k} n_t(\overline{X}^{(t)} - \overline{X})(\overline{X}^{(t)} - \overline{X})'.$$

合并的组内平方和为

$$A_0 = \sum_{t=1}^{k}\sum_{j=1}^{n_t}(a'X_{(j)}^{(t)} - a'\overline{X}^{(t)})^2$$

$$= a'\Big[\sum_{t=1}^{k}\sum_{j=1}^{n_t}(X_{(j)}^{(t)} - \overline{X}^{(t)})(X_{(j)}^{(t)} - \overline{X}^{(t)})'\Big]a$$

$$= a'Aa,$$

其中合并的组内离差阵(或称交叉乘积阵)A 为

$$A = \sum_{t=1}^{k}\sum_{j=1}^{n_t}(X_{(j)}^{(t)} - \overline{X}^{(t)})(X_{(j)}^{(t)} - \overline{X}^{(t)})'.$$

因此,若 k 个总体(类)的均值有显著差异,则比值

$$\frac{a'Ba}{a'Aa} \xlongequal{\text{def}} \Delta(a)$$

应充分大. 利用方差分析的思想,此问题化为求投影方向 a,使 $\Delta(a)$ 达极大值. 显然使 $\Delta(a)$达极大的解 a 不唯一. 若 a 使 $\Delta(a)$达极大,则 Ca (C 是任意不为零常数)也使 $\Delta(\cdot)$达极大,故对 a 附加一约束条件,即选取 a,使 $a'Aa=1$. 因此,问题又化为求 a,使 $\Delta(a)=a'Ba$ 在 $a'Aa=1$ 条件下达极大.

二、线性判别函数的求法

已知 a 是在 $a'Aa=1$ 条件下使 $\Delta(a)=a'Ba$ 达极大的方向,称 $u(X)=a'X$ 为**线性判别函数**. 以下利用拉格朗日乘子法来求条件极值问题的解. 令

$$\varphi(a) = a'Ba - \lambda(a'Aa - 1),$$

解方程组

$$\begin{cases}\dfrac{\partial\varphi}{\partial a} = 2(B - \lambda A)a = 0,\\[2ex] \dfrac{\partial\varphi}{\partial\lambda} = 1 - a'Aa = 0.\end{cases} \tag{5.3.1}$$

由方程组(5.3.1)的第一式可知,λ 是 $A^{-1}B$ 的特征根,a 是相应的特征向量,且可以证明 $\lambda=\Delta(a)$. 事实上,由 $Ba=\lambda Aa$ 两边左乘 a',得 $\Delta(a)=a'Ba=\lambda a'Aa=\lambda$.

因此,以上条件极值问题化为求 $A^{-1}B$ 的最大特征值和相应特征向量问题.

设 $A^{-1}B$ 的非零特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>0$,相应的满足约束

条件的特征向量为 $l_1,l_2,\cdots,l_r$,取 $a=l_1$ 时可使 $\Delta(a)$达最大,且最大值为 λ_1. $\Delta(a)$的大小可衡量判别函数 $u(X)=a'X$ 的判别效果,故称 $\Delta(a)$为**判别效率**. 综上所述得如下结论.

结论 1 在费希尔准则下,线性判别函数 $u(X)=a'X$ 的解 a,即为特征方程 $|A^{-1}B-\lambda I|=0$ 的最大特征根 λ_1 所对应的满足 $l_1'Al_1=1$ 的特征向量 l_1;且相应的判别效率 $\Delta(l_1)=\lambda_1$.

在有些问题中,仅用一个线性判别函数不能很好地区分 k 个总体,这时可用第二大特征值 λ_2,它所对应的满足 $l_2'Al_2=1$ 的特征向量 l_2,建立第二个线性判函数 $l_2'X$;如还不够,还可建立第三个线性判别函数 $l_3'X$;依次类推.

定义 5.3.1 设 $A^{-1}B$ 的非零特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>0$,其相应的满足约束条件的特征向量为 $l_1,l_2,\cdots,l_r$,称

$$P_1=\lambda_1\Big/\sum_{i=1}^{r}\lambda_i$$

为线性判别函数 $u_1(X)=l_1'X$ 的**判别能力**;称

$$P_{(l)}=(\lambda_1+\cdots+\lambda_l)\Big/\sum_{i=1}^{r}\lambda_i$$

为前 l 个($l\leqslant r$)线性判别函数 $u_1(X)=l_1'X,\cdots,u_l(X)=l_l'X$ 的**累计判别能力**.

三、费希尔判别准则

设 $A^{-1}B$ 的非零特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r>0$,其相应特征向量为 $l_1,l_2,,\cdots,l_r$ $(r\leqslant\min(m,k-1))$.

1. 判别准则 I ($r=1$ 情况)

如果只有一个判别函数 $u(X)=l'X$,它将 m 元数据投影到一维直线上. 例如 $k=2$ 情况,线性判别函数只有一个.

例 5.3.1 若 $k=2$,试求费希尔线性判别函数及其相应的判别效率.

解 当 $k=2$ 时,两总体的组间离差阵 B 为

$$B=n_1(\overline{X}^{(1)}-\overline{X})(\overline{X}^{(1)}-\overline{X})'+n_2(\overline{X}^{(2)}-\overline{X})(\overline{X}^{(2)}-\overline{X})',$$

利用 $\overline{X}=\frac{1}{n_1+n_2}(n_1\overline{X}^{(1)}+n_2\overline{X}^{(2)})$ 可得

$$B=\frac{n_1n_2}{n_1+n_2}(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})'. \tag{5.3.2}$$

合并的组内离差阵 A 为 $A=A_1+A_2$，其中

$$A_t=\sum_{i=1}^{n_t}(X_{(i)}^{(t)}-\overline{X}^{(t)})(X_{(i)}^{(t)}-\overline{X}^{(t)})' \quad (t=1,2).$$

由于 B 的秩为 1，故特征方程 $|A^{-1}B-\lambda I|=0$ 的非零特征根只有一个. 事实上，因为

$$A^{-1}B=\frac{n_1n_2}{n_1+n_2}A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})',$$

利用线性代数有关结论：AB 和 BA 的非零特征根相同知，$A^{-1}B$ 的非零特征根等同于

$$\frac{n_1n_2}{n_1+n_2}(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})=\frac{n_1n_2}{n_1+n_2}d^2, \tag{5.3.3}$$

其中

$$d^2=(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)}).$$

(5.3.3)式为一个数值，它就是欲求的特征根 λ. 记 l 为对应于 λ 的在条件 $l'Al=1$ 下的特征向量，它满足 $Bl=\lambda Al$，即

$$\frac{n_1n_2}{n_1+n_2}(\overline{X}^{(1)}-\overline{X}^{(2)})(\overline{X}^{(1)}-\overline{X}^{(2)})'l$$

$$=\frac{n_1n_2}{n_1+n_2}(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})\cdot Al.$$

若取 $l=\frac{1}{d}A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$，不难看出它满足以上方程，且 $l'Al=1$. 于是得费希尔线性判别函数为

$$u(X)=\frac{1}{d}X'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)}),$$

其相应的判别效率为

$$\Delta(l)=\frac{n_1n_2}{n_1+n_2}(\overline{X}^{(1)}-\overline{X}^{(2)})'A^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)}),$$

这里线性判别系数 l 与两总体间的马氏距离判别法的线性判别系数

$a=S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$相差一个倍数. 注意这里

$$S=\frac{1}{n_1+n_2-2}A.$$

下面以 $k=2$ 为例来导出按距离准则判断样品归类的判别法. 设两总体的样本均值为 $\overline{X}^{(1)},\overline{X}^{(2)}$,则线性判别函数值为

$$\overline{u}^{(1)}=l'\overline{X}^{(1)},\quad \overline{u}^{(2)}=l'\overline{X}^{(2)}.$$

在§5.1曾介绍过两种阈值点 $\bar{\mu}$ 和 μ^*,这两种都可以用来决定阈值点 $\bar{u}$ 和 u^*. 在这里

$$\bar{u}=\frac{1}{2}(l'\overline{X}^{(1)}+l'\overline{X}^{(2)})=\frac{1}{2}l'(\overline{X}^{(1)}+\overline{X}^{(2)}),\quad (5.3.4)$$

(5.3.4)式适用于投影后两总体的方差相等的情况. 若方差不等,记

$$u_{(i)}^{(t)}=l'X_{(i)}^{(t)}\quad (t=1,2;i=1,\cdots,n_t),$$

投影后总体 G_t $(t=1,2)$的样本方差为

$$\begin{aligned}\hat{\sigma}_t^2&=\frac{1}{n_t-1}\sum_{i=1}^{n_t}(u_{(i)}^{(t)}-\bar{u}^{(t)})^2\\&=\frac{1}{n_t-1}l'\Big[\sum_{i=1}^{n_t}(X_{(i)}^{(t)}-\overline{X}^{(t)})(X_{(i)}^{(t)}-\overline{X}^{(t)})'\Big]l\\&=\frac{1}{n_t-1}l'A_tl=l'S_tl.\end{aligned}$$

这时阈值点 u^* 为

$$u^*=\frac{\hat{\sigma}_2\bar{u}^{(1)}+\hat{\sigma}_1\bar{u}^{(2)}}{\hat{\sigma}_1+\hat{\sigma}_2},\quad (5.3.5)$$

判别准则为(不妨设 $l'\overline{X}^{(1)}>l'\overline{X}^{(2)}$)

$$\begin{cases}\text{判 } X\in G_1, & \text{当 } u(X)>\bar{u}(\text{或 } u^*);\\ \text{判 } X\in G_2, & \text{当 } u(X)<\bar{u}(\text{或 } u^*);\\ \text{待判}, & \text{当 } u(X)=\bar{u}(\text{或 } u^*).\end{cases}$$

2. 判别准则Ⅱ ($r>1$ 情况)

先取判别效率最大(记为 λ_1)的线性判别函数 $u_1(X)=l_1'X$, k 个总体的均值向量在 l_1 上的投影为 $\bar{u}_1^{(t)}=l_1'\overline{X}^{(t)}$ $(t=1,\cdots,k)$. 对样品 $X=(x_1,\cdots,x_m)'$,计算它在 l_1 上投影: $u_1(X)=l_1'X$. 若存在唯一的 i_1,使

$$\frac{|u_1(X)-\overline{u}_1^{(i_1)}|}{\hat{\sigma}_{i_1}}=\min_{t=1,\cdots,k}\frac{|u_1(X)-\overline{u}_1^{(t)}|}{\hat{\sigma}_t}$$

时，判 $X\in G_{i_1}$. 如果存在 j 个总体 $G_{k_1},\cdots,G_{k_j}$ $(j>1)$，使其与 $u_1(X)$ 距离相等且为最小，记序号集 $L=\{k_1,\cdots,k_j\}$，则再取判别效率为 λ_2 (次大)的判别函数 $u_2(X)=l_2'X$，当存在唯一的 i_2，使

$$\frac{|u_2(X)-\overline{u}_2^{(i_2)}|}{\hat{\sigma}_{i_2}}=\min_{t\in L}\frac{|u_2(X)-\overline{u}_2^{(t)}|}{\hat{\sigma}_t}$$

时，则判 $X\in G_{i_2}$，其中 $\overline{u}_2^{(t)}=l_2'\overline{X}^{(t)}$. 如果第二个判别函数仍不能判别样品 X 所属总体，则还可以取第三个线性判别函数，依此类推. 这个准则借用了序贯判别的思想.

3. 判别准则Ⅲ

如果有 r 个非零特征根 $(1\leqslant r\leqslant m)$，相应的有 r 个线性判别函数 $u_1(X),\cdots,u_r(X)$. 这时相当于把原来 m 个变量综合成 r 个新变量. 在实用中常取 $l\leqslant r$，且满足 $(\lambda_1+\cdots+\lambda_l)/(\lambda_1+\cdots+\lambda_l+\cdots+\lambda_r)\geqslant P_0$ (一般取 $P_0\geqslant 0.7$). 这样 m 元总体的判别问题即化为 l 元总体的判别问题，一般地新变量个数比原变量个数减少了. 由于特征向量线性无关，故 l 个新变量互不相关. 然后对 l 元数据按 §5.1 的距离判别准则来进行判别归类.

例 5.3.2　试对表 5.2 中胃癌检验的生化指标值用费希尔判别的方法进行判别归类.

解　首先使用典型判别(CANDISC)过程，由第十章将介绍的典型相关分析方法求出两个典型变量(即 $u_1(X)$ 和 $u_2(X)$)：

$$u_1(X)=0.0100X_1+0.04018X_2+0.1764X_3+0.03055X_4,$$

$$u_2(X)=-0.003880X_1-0.05462X_2+0.1600X_3+0.06206X_4.$$

然后计算典型变量的得分，也就是用中心化后的观测数据代入以上 $u_1(X)$ 和 $u_2(X)$ 的关系式中所得到的值. 如果绘制第一和第二典型变量得分的散布图，还可以直观地看出，这三个类基本上是能够分开的，特别是第 1 类与其他两类.

接着调用判别归类(DISCRIM)过程，由典型判别方法产生的两个典型变量的得分(这时把 4 元总体简化为 2 元总体)进行判别归

类. 首先给出用两个典型变量得分进行判别归类时计算的两两配对的组间距离及均值差异的显著性检验的结果. 记 $\nu^{(i)}$ $(i=1,2,3)$ 为变换后第 i 个 2 元总体的均值向量，如检验 $H_0^{(12)}$：$\nu^{(1)}=\nu^{(2)}$ 的 $p=0.0019<\alpha=0.05$；检验 $H_0^{(13)}$：$\nu^{(1)}=\nu^{(3)}$ 的 $p=0.0010<\alpha=0.05$；检验 $H_0^{(23)}$：$\nu^{(2)}=\nu^{(3)}$ 的 $p=0.3231>\alpha=0.05$. 这表明变换后第 1 类与第 2 和第 3 类之间有显著性差异，而第 2 类与第 3 类之间的差异就不显著. 接着给出用费希尔判别方法对 15 个观测进行判别的结果，从输出中可以看到判错的个数为 3 个(把原属于第 1 类的第 4 号观测判归为第 3 类；把原属于第 2 类的第 8 号观测判归为第 3 类；把原属于第 3 类的第 11 号观测判归为第 2 类).

如果假定三个类的协方差阵不等，由 DISCRIM 过程对 15 个观测进行判别的结果为错判了两个(即第 8 号和第 11 号观测).

§5.4 判别效果的检验及各变量判别能力的检验

以上几节介绍的判别准则，都是根据已知观测值(即训练样本)，建立判别函数，并由判别函数给出空间 $\mathbb{R}^m$ 的一个划分 D(即判别法). 建立在样本基础上的判别法则，其判别能力显然与样本是否来自不同的总体有关；也与所考察的 m 个判别指标变量是否能区分 k 个不同的总体(组)有关.

假设总体 G_t 的分布为 $N_m(\mu^{(t)},\Sigma_t)$ $(t=1,2,\cdots,k)$，$X_{(i)}^{(t)}(t=1,\cdots,k;i=1,2,\cdots,n_t)$ 为来自 G_t 的 m 元样本.

一、两总体判别效果的检验

先考虑 $k=2$ 的简单情况. 所谓判别效果的检验，就是检验两总体的均值是否有显著性差异. 一般我们提出的原假设 H_0 为两总体的均值是相等的. 如果 H_0 被否定，则说明两个总体 G_1 和 G_2 确实可以区分，建立的判别准则是有意义的. 如果 H_0 不能被拒绝，说明两个总体均值的差异不显著，此时来讨论判别分析是自欺欺人，毫无意

义，除非考虑其他新的判别变量.

假设 G_i 为 $N(\mu^{(i)},\Sigma)$ $(i=1,2)$. 检验两总体的均值是否有显著性差异(即检验 H_0: $\mu^{(1)}=\mu^{(2)}$)时，根据第三章的结论，首先计算两总体样本均值 $\overline{X}^{(1)}$ 与 $\overline{X}^{(2)}$ 之间的马氏距离 $d^2(1,2)$:

$$d^2(1,2)=(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)}),$$

其中 S 是合并样本协方差阵. 然后，由马氏距离 d^2 构造检验统计量——F 统计量:

$$F=\frac{(n_1+n_2-m-1)n_1n_2}{m(n_1+n_2)(n_1+n_2-2)}d^2(1,2),$$

其中 n_i 是第 i 个总体的样品个数$(i=1,2)$. 在两总体均值相等的假设成立下，F 统计量服从分子自由度为 m，而分母自由度为 n_1+n_2-m-1 的 F 分布. 利用样本可计算 F 统计量的值，由该值还可求出显著性概率值(p 值). 若 p 值小于给定的显著性水平 α (常取 $\alpha=0.05$)，则否定两总体的均值向量是相等的假设，即对这两总体讨论判别问题是有意义的. 若 p 值大于等于给定的显著性水平 α，则两总体的均值没有显著性的差异. 这时讨论两总体的判别问题是没有意义的. 如果盲目地应用以上介绍的方法进行判别归类，则错判的机会将很大.

二、k 个总体判别效果的检验($k>2$)

当 $k>2$ 时，判别效果的检验问题包括以下两方面：首先检验 k 个类的均值向量是否全都相等(即检验 H_0: $\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)}$)；若不全相等，则进一步对 k 个总体两两配对，然后逐对检验这两个总体的均值是否有显著差异(检验 $H_0^{(ij)}$: $\mu^{(i)}=\mu^{(j)}, i\neq j$)，也就是检验这两总体的判别效果是否显著. 具体方法仍是通过计算各总体间的马氏距离及 F 统计量，并利用 p 值的大小来判断其判别效果.

1. 检验 H_0: $\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)}$

假设 k 个总体的协方差阵相同：$\Sigma_1=\Sigma_2=\cdots=\Sigma_k\overset{\text{def}}{=\!=}\Sigma$. 根据第三章多元方差分析的方法. 我们把样本的总离差阵 T 分解为：

$$T=\sum_{t=1}^{k}\sum_{j=1}^{n_t}(X_{(j)}^{(t)}-\overline{X})(X_{(j)}^{(t)}-\overline{X})'=A+B,\quad(5.4.1)$$

其中

$$A=\sum_{t=1}^{k}A_t=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(X_{(i)}^{(t)}-\overline{X}^{(t)})(X_{(i)}^{(t)}-\overline{X}^{(t)})'\quad(5.4.2)$$

称为合并组内离差阵，

$$B=\sum_{i=1}^{k}n_i(\overline{X}^{(i)}-\overline{X})(\overline{X}^{(i)}-\overline{X})'$$

称为组间离差阵.

利用似然比原则可导出检验 H_0 的似然比统计量

$$\Lambda=\frac{|A|}{|A+B|}=\frac{|A|}{|T|}.$$

根据 Λ 分布的定义，可知(记 $n=n_1+n_2+\cdots+n_k$)

$$\Lambda=\frac{|A|}{|A+B|}\overset{H_0下}{\sim}\Lambda(m,n-k,k-1).$$

给定显著性水平 α，查威尔克斯分布临界值表，可得 λ_α，使

$$P\{\Lambda\leqslant\lambda_\alpha\}=\alpha,$$

故否定域 $W=\{\Lambda\leqslant\lambda_\alpha\}$. 当 $k=2$ 或 3 时，可把 Λ 分布转化为 F 分布，更一般地情况可用 χ^2 分布或 F 分布来近似，即由 Λ 函数的近似分布进行检验(见参考文献[1]或[2]).

2. 分别检验 $H_0^{(ij)}$: $\mu^{(i)}=\mu^{(j)}$

把 k 个总体两两配对，逐对检验，辨明各对的判别效果. 具体方法同小节—“两总体判别效果的检验”. 计算中作了如下处理(假定 $\Sigma_1=\cdots=\Sigma_k\overset{\text{def}}{=\!=}\Sigma$)，检验 $H_0^{(ij)}$: $\mu^{(i)}=\mu^{(j)}$ $(i,j=1,\cdots,k,i\neq j)$时，取

$$F_{ij}=\frac{n-k-m+1}{m(n-k)}\frac{n_in_j}{n_i+n_j}d_{ij}^2,$$

其中

$$d_{ij}^2=(\overline{X}^{(i)}-\overline{X}^{(j)})'S^{-1}(\overline{X}^{(i)}-\overline{X}^{(j)}),$$

$$\overline{X}^{(i)}-\overline{X}^{(j)}\overset{H_0^{(ij)}下}{\sim}N_m\left(0,\left(\frac{1}{n_i}+\frac{1}{n_j}\right)\Sigma\right),$$

$$(n-k)S=A_1+A_2+\cdots+A_k\sim W_m(n-k,\Sigma),$$

且 $\overline{X}^{(i)}$，A_i 分别是第 i 个总体 G_i 的样本均值和离差阵. 故

$$T_{ij}^2=\frac{n_in_j}{n_i+n_j}d_{ij}^2 \overset{H_0^{(ij)}下}{\sim} T^2(m,n-k),$$

$$F_{ij}=\frac{n-k-m+1}{m(n-k)}T_{ij}^2 \overset{H_0^{(ij)}下}{\sim} F(m,n-k-m+1).$$

从而可以利用 F 统计量对假设 $H_0^{(ij)}$ 进行检验.

三、各变量判别能力的检验

当检验 k 个类的均值向量是否全都相等(即检验 H_0：$\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)}$)时，若否定假设 H_0(即表明 k 个总体的均值向量之间有显著性差异)，也并不能保证其各分量的均值向量有显著差异. 若第 i 个分量间没有显著差异时，说明相应的变量 X_i 对判别分类不起作用，应该剔除. 关于各变量判别能力的检验问题是筛选判别变量的理论基础，也是下面介绍逐步判别的理论依据.

1. 变量判别能力的度量

以上检验 H_0：$\mu^{(1)}=\mu^{(2)}=\cdots=\mu^{(k)}$ 时，引入检验统计量

$$\Lambda_{(m)}=\frac{|A|}{|T|},$$

$\Lambda_{(m)}$ 的值越小，表明 m 个指标(变量)对 k 个总体的判别效果越好. 用附录中 §9 介绍的消去变换法可以求行列式的值：

$$|A|=\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{vmatrix}=a_{11}\left|\begin{array}{c|ccc} 1 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1m}}{a_{11}} \\ \hline 0 & a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{array}\right|$$

$$=a_{11}\begin{vmatrix} a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & & \vdots \\ a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{vmatrix}=\cdots=a_{11}a_{22}^{(1)}\cdots a_{mm}^{(m-1)},$$

其中(记 $a_{ij}=a_{ij}^{(0)}$)

$$a_{ii}^{(i-1)}=a_{ii}^{(i-2)}-\frac{a_{i(i-1)}^{(i-2)}\cdot a_{(i-1)i}^{(i-2)}}{a_{(i-1)(i-1)}^{(i-2)}}\quad(i=2,3,\cdots,m).$$

类似地有$|T|=t_{11}t_{22}^{(1)}\cdots t_{mm}^{(m-1)}$. 所以

$$\Lambda_{(m)}=\frac{|A|}{|T|}=\frac{a_{11}}{t_{11}}\frac{a_{22}^{(1)}}{t_{22}^{(1)}}\cdots\frac{a_{mm}^{(m-1)}}{t_{mm}^{(m-1)}}\xlongequal{\text{def}}U_{(1,2,\cdots,m)}.$$

以上行列式的计算是按自然顺序做消去变换. 由消去变换的性质可知,亦可不按自然顺序进行. 设$(i_1,i_2,\cdots,i_m)$是$(1,2,\cdots,m)$的任一排列,第k次以(i_k,i_k)为主元做消去变换$(k=1,2,\cdots,m)$,于是有:

$$\Lambda_{(m)}=\frac{a_{i_1i_1}}{t_{i_1i_1}}\frac{a_{i_2i_2}^{(1)}}{t_{i_2i_2}^{(1)}}\cdots\frac{a_{i_mi_m}^{(m-1)}}{t_{i_mi_m}^{(m-1)}}\xlongequal{\text{def}}U_{(i_1,i_2,\cdots,i_m)}.$$

$\Lambda_{(m)}$的大小可以用来度量m个指标$X_1,X_2,\cdots,X_m$对k个总体的判别效果,$\Lambda_{(m)}$越小,判别效果越好.

如果只考虑$m-1$个变量,不妨设为$X_1,X_2,\cdots,X_{m-1}$,则

$$\Lambda_{(m-1)}=\frac{|A_{m-1}|}{|T_{m-1}|}=\frac{a_{11}}{t_{11}}\frac{a_{22}^{(1)}}{t_{22}^{(1)}}\cdots\frac{a_{(m-1)(m-1)}^{(m-2)}}{t_{(m-1)(m-1)}^{(m-2)}}$$

$$\xlongequal{\text{def}}U_{(1,2,\cdots,m-1)}.$$

显然

$$U_{(1,2,\cdots,m)}=U_{(1,2,\cdots,m-1)}\cdot\frac{a_{mm}^{(m-1)}}{t_{mm}^{(m-1)}}.$$

记

$$U_{m|(1,\cdots,m-1)}=\frac{a_{mm}^{(m-1)}}{t_{mm}^{(m-1)}},$$

并称它为给定$X_1,X_2,\cdots,X_{m-1}$时,变量X_m的判别能力,它是变量X_m判别能力的一个度量. 它的值愈小,变量X_m的判别能力越强. 在以上记号下,有递推公式:

$$U_{(1,2,\cdots,m)}=U_{(1,2,\cdots,m-1)}\cdot U_{m|(1,\cdots,m-1)}.$$

类似地,可定义变量X_i判别能力的度量$U_{i|(1,\cdots,i-1,i+1,\cdots,m)}(i=1,\cdots,m)$,且

$$U_{(1,2,\cdots,m)}=U_{(1,\cdots,i-1,i+1,\cdots,m)}\cdot U_{i|(1,\cdots,i-1,i+1,\cdots,m)}.$$

变量判别能力的度量采用删去该变量后考察判别能力的变化,若变化小表示该变量对区分k个总体不起作用,否则该变量对区分k个总体是重要的.

2. 变量判别能力的检验(附加信息检验)

若已知 r 个变量 $X_{i_1},\cdots,X_{i_r}$ $(r<m)$对 k 个总体的判别效果显著. 相应的度量这 r 个变量判别能力的统计量为

$$U_{(i_1,\cdots,i_r)}=\frac{a_{i_1i_1}}{t_{i_1i_1}}\frac{a^{(1)}_{i_2i_2}}{t^{(1)}_{i_2i_2}}\cdots\frac{a^{(r-1)}_{i_ri_r}}{t^{(r-1)}_{i_ri_r}},$$

在此基础上考虑添加另一变量 $X_{i_{r+1}}$ 后,对应的 U 统计量为

$$U_{(i_1,\cdots,i_r,i_{r+1})}=U_{(i_1,\cdots,i_r)}\cdot U_{i_{r+1}|(i_1,\cdots,i_r)},$$

其中

$$U_{i_{r+1}|(i_1,\cdots,i_r)}=\frac{a^{(r)}_{i_{r+1}i_{r+1}}}{t^{(r)}_{i_{r+1}i_{r+1}}}$$

表示在已引入 r 个变量 $X_{i_1},\cdots,X_{i_r}$ 之后,再添加 $X_{i_{r+1}}$ 对 U 统计量的影响. 为检验 $X_{i_{r+1}}$ 对 k 个总体的判别效果能否提供附加信息(即它新提供的信息是否被包含在 $X_{i_1},\cdots,X_{i_r}$ 提供的信息之中)需作统计检验.

请注意,这时不是简单地检验 H_0: $\mu^{(1)}_{i_{r+1}}=\mu^{(2)}_{i_{r+1}}=\cdots=\mu^{(k)}_{i_{r+1}}$(即 k 个均值向量的第 i_{r+1} 个分量是否相等). 而是首先要从 $X_{i_{r+1}}$ 中将 $X_{i_1},\cdots,X_{i_r}$ 提供的信息扣除,再检验扣除后的均值向量是否相等. 这里需要利用条件均值的概念. 令

$$\mu^{(t)}_{i_{r+1}|(i_1,\cdots,i_r)}=\mathrm{E}[X^{(t)}_{ji_{r+1}}|X^{(t)}_{ji_1},\cdots,X^{(t)}_{ji_r}]$$
$$(t=1,2,\cdots,k;j=1,2,\cdots,n_t),$$

其中 $X^{(t)}_{ji_u}(u=1,2,\cdots,r+1)$表示第 t 个总体 G_t 中第 j 个样品 $X^{(t)}_{(j)}=(X^{(t)}_{j1},\cdots,X^{(t)}_{ji_r},\cdots,X^{(t)}_{jm})$的第 i_u 个分量. 附加信息的检验是

$$H_0:\ \mu^{(1)}_{i_{r+1}|(i_1,\cdots,i_r)}=\mu^{(2)}_{i_{r+1}|(i_1,\cdots,i_r)}=\cdots=\mu^{(k)}_{i_{r+1}|(i_1,\cdots,i_r)}.\tag{5.4.3}$$

在总体为正态分布假设下,因正态分布的条件分布仍为正态分布,因此检验(5.4.3)式仍可以用威尔克斯统计量. 由第二章 § 2.3 定理 2.3.2,不难证明,(5.4.3)式的似然比统计量为

$$U_{i_{r+1}|(i_1,\cdots,i_r)}=\frac{a^{(r)}_{i_{r+1}i_{r+1}}}{t^{(r)}_{i_{r+1}i_{r+1}}}.\tag{5.4.4}$$

可以证明:

$$U_{(i_1,\cdots,i_r)}=\Lambda_{(r)}\sim\Lambda(r,n-k,k-1),$$

$$U_{(i_1,\cdots,i_r,i_{r+1})}=\Lambda_{(r+1)}\sim\Lambda(r+1,n-k,k-1),$$

$$U_{i_{r+1}|(i_1,\cdots,i_r)}\xlongequal{\text{def}}\Lambda_{i_{r+1}|(r)}\sim\Lambda(1,n-k-r,k-1).$$

利用 Λ 统计量与 F 统计量的关系，有

$$F=\frac{n-k-r}{k-1}\frac{1-U_{i_{r+1}|(i_1,\cdots,i_r)}}{U_{i_{r+1}|(i_1,\cdots,i_r)}}\overset{H_0\text{下}}{\sim}F(k-1,n-k-r).$$

利用 F 统计量对假设 H_0 作统计检验：若否定 H_0，表示变量 $X_{i_{r+1}}$ 对 k 个总体的判别能力是显著的(在显著性水平 α 下)；否则，变量 $X_{i_{r+1}}$ 对 k 个总体的区分不能提供附加信息，这个变量应剔除.

§5.5 逐步判别

一、逐步判别法的基本思想

前面我们讨论了用全部 m 个变量 $X_1,X_2,\cdots,X_m$ 来建立判别函数，用以对样品进行判别归类的几种方法. 在这 m 个变量中，有的变量对区分 k 个总体的判别能力可能很强，有的可能很微弱. 如果不加区别地把 m 个变量全部用来建立判别函数，必然增加大量的计算，还可能因为变量间的相关性引起计算上的困难(病态或退化等)及计算精度的降低. 另一方面由于一些对区分 k 个总体的判别能力很小的变量的引入，产生干扰，致使建立的判别函数不稳定，反而影响判别效果，因此自然提出一个变量的选择问题. 即如何从 m 个变量中挑选出对区分 k 个总体有显著判别能力的变量，来建立判别函数，用以判别归类.

类似于回归分析，判别分析的变量选择方法也有向前法、后退法和逐步筛选法. 这里仅介绍逐步筛选法.

逐步判别的基本思想和逐步回归是类似的. 逐个引入变量，每次把一个判别能力最强的变量引入判别式，每引入一个新变量，对判别式中的老变量逐个进行检验，如其判别能力因新变量的引入而变得不显著，应把它从判别式中剔除. 这种通过逐步筛选变量使得建立的

判别函数中仅保留判别能力显著的变量的方法，就是**逐步判别法**.

二、逐步筛选变量的基本步骤

记合并组内离差阵 $A=(a_{ij})$，总离差阵 $T=(t_{ij})$，A,T 的定义见(5.4.1)和(5.4.2)式.

1. 可否引入变量进入判别式

(1) 考察变量 X_i $(i=1,\cdots,m)$对 k 个总体的判别能力(此时判别式中变量个数 $r=0$). 变量 X_i 的判断能力 $U_{(i)}$ 为

$$U_{(i)}=\frac{a_{ii}}{t_{ii}} \quad (i=1,\cdots,m),$$

设 $U_{(i_1)}=\min\limits_{i=1,\cdots,m} U_{(i)}$.

(2) 检验 X_{i_1} 对 k 个总体的判别效果是否显著，即检验：

$$H_0:\ \mu_{i_1}^{(1)}=\mu_{i_1}^{(2)}=\cdots=\mu_{i_1}^{(k)},$$

其中 $\mu_{i_1}^{(t)}$ 为总体 G_t 的均值向量 $\mu^{(t)}$ 的第 i_1 个分量. 在 H_0 成立时

$$U_{(i)}\sim \Lambda(1,n-k,k-1),$$

由 $U_{(i)}$ 可构造检验统计量

$$F=\frac{1-U_{(i_1)}}{U_{(i_1)}}\frac{n-k}{k-1}=\frac{t_{i_1i_1}-a_{i_1i_1}}{a_{i_1i_1}}\frac{n-k}{k-1}$$

$$\overset{H_0下}{\sim} F(k-1,n-k).$$

对给定的显著性水平 $\alpha=0.05$，按传统的检验方法，可查 F 分布临界值表得 F_α，使 $P\{F>F_\alpha\}=\alpha$. 比较由样本值计算得到的 F 值及临界值 F_α. 若 $F\leqslant F_\alpha$ 时表明判别能力“最强”的变量 X_{i_1} 对 k 个总体判别效果并不显著，逐步筛选变量的过程停止，这时所考察的 m 个变量不能区分 k 个总体，应考虑引入新变量.

若 $F>F_\alpha$，把变量 X_{i_1} 引入判别式，并对矩阵 A,T 做消去变换：

$$A^{(1)}=T_{i_1}(A),\quad T^{(1)}=T_{i_1}(T).$$

利用统计软件进行检验时(以后我们均采用此种方法描述统计检验的步骤)，首先由样本值计算得到的 F 统计量的值 f，并计算 p 值：

$$p=P\{F\geqslant f\}\quad (\text{其中 } F\sim F(k-1,n-k)).$$

若 $p\geqslant\alpha$，则 H_0 相容，没有变量引入；若 $p<\alpha$，则把变量 X_{i_1} 引入判别式，并对矩阵 A,T 做消去变换.

2. 考虑能否剔除变量的步骤

设判别式中已有变量 $X_{i_1},X_{i_2},\cdots,X_{i_r}$ $(r>1)$. 矩阵 A,T 经若干次消去变换后化为 $A^{(r)},T^{(r)}$.

(1) 计算判别式中变量 X_{i_j} 在其余 $r-1$ 个变量给定时的判别能力(即威尔克斯统计量). 记

$$U_{i_j|(i_1,\cdots,i_{j-1},i_{j+1},\cdots,i_r)}\xlongequal{\text{def}}U_{i_j|(r-1)},$$

$$U_{i_j|(r-1)}=\frac{a_{i_ji_j}^{(r-1)}}{t_{i_ji_j}^{(r-1)}}=\frac{t_{i_ji_j}^{(r)}}{a_{i_ji_j}^{(r)}}\quad(j=1,\cdots,r),$$

设 $U_{i_0|(r-1)}=\max\limits_{j=1,\cdots,r}U_{i_j|(r-1)}$.

(2) 检验 X_{i_0} 在其余 $r-1$ 个变量给定时对 k 个总体的判别效果是否显著，即检验

$$H_0:\mu_{i_0|(r-1)}^{(1)}=\mu_{i_0|(r-1)}^{(2)}=\cdots=\mu_{i_0|(r-1)}^{(k)}.$$

在 H_0 成立时威尔克斯统计量

$$U_{i_0|(r-1)}\sim\Lambda(1,n-k-r+1,k-1),$$

由 $U_{i_0|(r-1)}$ 可构造统计量

$$F=\frac{1-U_{i_0|(r-1)}}{U_{i_0|(r-1)}}\frac{n-(r-1)-k}{k-1}$$

$$\overset{H_0\text{F}}{\sim}F(k-1,n-k-r+1).$$

对给定的显著性水平 $\alpha=0.05$，首先由观测样本计算 F 统计量的值 f，并计算 p 值：

$$p=P\{F\geqslant f\}\quad(\text{其中 }F\sim F(k-1,n-k-r+1)).$$

若 $p<\alpha$，则否定 H_0，表明判别能力“最弱”的变量 X_{i_0} 对 k 个总体判别效果都是显著的，不能剔除，转入考虑可否引入新变量的步骤；若 $p\geqslant\alpha$，则 H_0 相容，表明因新变量的引入使判别式中原有的变量 X_{i_0} 变为不能提供附加信息(即判别效果不显著)，应剔除 X_{i_0}，并对 $A^{(r)},T^{(r)}$ 做消去变换：

$$A^{(r+1)} = T_{i_0}(A^{(r)}), T^{(r+1)} = T_{i_0}(T^{(r)}),$$

然后再继续考虑能否再剔除变量.

3. 考虑能否引入新变量的步骤

设判别式中已有 r 个变量 $X_{i_1}, X_{i_2}, \cdots, X_{i_r}$，考虑能否从其余 $m-r$ 个变量 $X_{j_1}, X_{j_2}, \cdots, X_{j_{m-r}}$ 中选出在给定 $X_{i_1}, X_{i_2}, \cdots, X_{i_r}$ 的条件下，其判别效果显著的变量.

(1) 对不在判别式中的变量 $X_{j_1}, X_{j_2}, \cdots, X_{j_{m-r}}$ 计算在 $X_{i_1}, X_{i_2}, \cdots, X_{i_r}$ 给定时的判别能力(威尔克斯统计量). X_{j_α} 的威尔克斯统计量为

$$U_{j_\alpha|(r)} = \frac{a^{(r)}_{j_\alpha j_\alpha}}{t^{(r)}_{j_\alpha j_\alpha}} \quad (\alpha = 1, \cdots, m-r),$$

设 $U_{j_0|(r)} = \min\limits_{\alpha=1,\cdots,m-r} U_{j_\alpha|(r)}$.

(2) 检验 H_0: $\mu^{(1)}_{j_0|(r)} = \mu^{(2)}_{j_0|(r)} = \cdots = \mu^{(k)}_{j_0|(r)}$. 在 H_0 成立时威尔克斯统计量

$$U_{j_0|(r)} \sim \Lambda(1, n-k-r, k-1),$$

由 $U_{j_0|(r)}$ 可构造统计量

$$F = \frac{1-U_{j_0|(r)}}{U_{j_0|(r)}} \frac{n-k-r}{k-1} \overset{H_0\text{下}}{\sim} F(k-1, n-k-r).$$

对给定的显著性水平 α，首先由观测样本计算 F 统计量的值 f，并计算 p 值：

$$p = P\{F \geqslant f\} \quad (\text{其中 } F \sim F(k-1, n-k-r)).$$

若 $p \geqslant \alpha$，则 H_0 相容，变量 X_{j_0} 不能引入判别式，筛选变量的过程结束；若 $p < \alpha$，则把变量 X_{j_0} 引入判别式，并对矩阵 $A^{(r)}, T^{(r)}$ 做消去变换：

$$A^{(r+1)} = T_{j_0}(A^{(r)}), \quad T^{(r+1)} = T_{j_0}(T^{(r)}),$$

然后转入考虑能否剔除老变量的步骤.

三、逐步判别的计算方法

设样本 $X^{(t)}_{(j)} = (X^{(t)}_{j1}, X^{(t)}_{j2}, \cdots, X^{(t)}_{jm})'$ $(t=1, \cdots, k; j=1, 2, \cdots, n_t)$,

记 $n=n_1+n_2+\cdots+n_k$.

1. 准备工作

(1) 计算各总体(类)的样本均值 $\overline{X}^{(t)}$ $(t=1,\cdots,k)$和总样本均值 $\overline{X}$.

(2) 计算样本的合并组内离差阵 A 和总离差阵 T.

(3) 规定显著性水平 α (如 $\alpha=0.05$).

2. 逐步筛选变量

假设已计算了 L 步($L\geqslant 0$),在判别式中选入了 L 个变量(用 L 表示入选变量的个数,且表示入选变量的集合,如 $L=\{i_1,\cdots,r_L\}$);合并组内离差阵 A 和总离差阵 T 经若干次消去变换化为 $A^{(L)},T^{(L)}$.

(1) 计算所有变量的判别能力 $U_{(i)}$ $(i=1,2,\cdots,m)$:

$$U_{(i)}=\begin{cases}\dfrac{a_{ii}^{(L)}}{t_{ii}^{(L)}}\overset{\text{def}}{=\!=}U_{i|(L)}, & \text{当 } i\overline{\in} L,\\ \dfrac{t_{ii}^{(L)}}{a_{ii}^{(L)}}\overset{\text{def}}{=\!=}U_{i|(L-1)}, & \text{当 } i\in L,\end{cases}$$

设 $U_{i_0|(L-1)}=\max\limits_{i\in L}U_{i|(L-1)}$, $U_{j_0|(L)}=\min\limits_{i\overline{\in}L}U_{i|(L)}$.

(2) 为检验 X_{i_0} 可否从判别式中剔除,计算检验统计量

$$F_1=\frac{1-U_{i_0|(L-1)}}{U_{i_0|(L-1)}}\frac{n-(L-1)-k}{k-1}$$

$$=\frac{a_{i_0i_0}^{(L)}-t_{i_0i_0}^{(L)}}{t_{i_0i_0}^{(L)}}\frac{n-L-k+1}{k-1}$$

的值;再由得到的 F_1 统计量值 f_1 计算 p 值:

$$p=P\{F_1\geqslant f_1\}\quad(\text{其中 } F_1\sim F(k-1,n-k-L+1)).$$

若 $p<\alpha$,不能剔除,转入考虑可否引入新变量的步骤(3);若 $p\geqslant\alpha$,应剔除 X_{i_0},记 $r=i_0$,并转到步骤(4).

(3) 为检验 X_{j_0} 可否引入判别式,计算检验统计量

$$F_2=\frac{1-U_{j_0|(L)}}{U_{j_0|(L)}}\frac{n-L-k}{k-1}$$

$$=\frac{t_{j_0j_0}^{(L)}-a_{j_0j_0}^{(L)}}{a_{j_0j_0}^{(L)}}\frac{n-k-L}{k-1}$$

的值;再由得到的 F_2 统计量值 f_2 计算 p 值:

$$p = P\{F_2 \geqslant f_2\} \quad (\text{其中 } F_2 \sim F(k-1, n-k-L)).$$

若 $p<\alpha$,则把 X_{j_0} 引入判别式,记 $r=j_0$,转入步骤(4);若 $p\geqslant\alpha$,没有变量可引入,逐步筛选变量的过程结束.转入进行判别归类.

(4) 计算当前变量 X_r 的威尔克斯统计量,并对 $A^{(L)}, T^{(L)}$ 做消去变换:

当 X_r 为入选变量时(即 $X_r=X_{j_0}$),$U_{(L+1)}=U_{(L)}\dfrac{a_{rr}^{(L)}}{t_{rr}^{(L)}}$,

当 X_r 为剔除变量时(即 $X_r=X_{i_0}$),$U_{(L-1)}=U_{(L)}\dfrac{t_{rr}^{(L)}}{a_{rr}^{(L)}}$;

对 $A^{(L)}$ 和 $T^{(L)}$ 同时做以 (r,r) 为主元的消去变换:

$$A^{(L+1)} = T_r(A^{(L)}), \quad T^{(r+1)} = T_r(T^{(L)}).$$

具体变换公式请见附录中 §9 的有关部分.

(5) 重复步骤(1)~(4),直到判别式中没有变量可剔除,且不在判别式中的变量也没有可引入时,逐步筛选变量的计算过程结束.

3. 判别归类

设逐步筛选变量的过程结束后,A, T 变为 $A^{(L)}$ 和 $T^{(L)}$;选入判别式的变量有 L 个,即 $X_{i_1}, X_{i_2}, \cdots, X_{i_L}$.接着对选出的判别能力强的 L 个变量,使用前几节介绍的各种方法(如距离判别准则,贝叶斯判别准则等)来建立判别函数并给出判别准则.如果按正态总体下的贝叶斯判别准则,由当前的矩阵 $A^{(L)}$ 和 $T^{(L)}$ 可以很方便的计算出判别函数,给出判别准则,并检验这 L 个变量对 k 个总体的判别效果.

例 5.5.1 (胃癌的鉴别)　对表 5.2 的病例资料,试用逐步判别方法建立判别准则,并对 15 个样品进行判别归类.

解　利用 SAS/STAT 软件中的逐步判别(STEPDISC)过程逐步筛选变量,然后利用 DISCRIM 过程进行判别归类.

STEPDISC 过程用于筛选对区分 k 个类能力强的变量集.使用中常要求:(1) 指定筛选变量的方法,如逐步筛选法;(2) 规定引入变量到判别式和剔除变量的显著性水平 α,默认值均为 0.15;(3) 规定最终判别式中变量个数;(4) 规定筛选过程的最大步数.

逐步筛选变量的第一步,首先给出各个变量对于区分 3 个类的

偏 R^2,F 统计量及 p 值，p 值最小者(0.0060)即最能区分 3 个类的变量,其中 X_2 第一个被引入判别式;然后由多元统计量给出此时判别式中这些变量的判别效果.逐步筛选变量的第二、第三步输出的结果同第一步类似.

用逐步筛选法选出的两个变量 X_2 和 X_3 来建立判别准则,并给出回判结果.在回判结果中我们看到,只有两个病例判错了,即把来自第 2 类的第 6 号观测判为第 3 类;把来自第 3 类的第 15 号观测判为第 2 类.

习 题 五

5-1 已知总体 $G_i(m=1)$的分布为 $N(\mu^{(i)},\sigma_i^2)$ $(i=1,2)$,按距离判别准则为(不妨设 $\mu^{(1)}>\mu^{(2)}$)

$$\begin{cases} x\in G_1, & 若\ x>\mu^*, \\ x\in G_2, & 若\ x\leqslant\mu^*, \end{cases}$$

其中 $\mu^*=\dfrac{\sigma_1\mu^{(2)}+\sigma_2\mu^{(1)}}{\sigma_1+\sigma_2}$.试求错判概率 $P(2|1)$和 $P(1|2)$.

5-2 设三个总体 G_1,G_2 和 G_3 的分布分别为：$N(2,0.5^2)$,$N(0,2^2)$和 $N(3,1^2)$.试问样品 $x=2.5$ 应判归哪一类?

(1) 按距离判别准则;

(2) 按贝叶斯判别准则$\left(\text{取 } q_1=q_2=q_3=\dfrac{1}{3},L(j|i)=\begin{cases}1,i\neq j\\0,i=j\end{cases}\right)$.

5-3 设总体 G_i 的均值为 $\mu^{(i)}$ $(i=1,2)$,同协方差阵为 Σ.记

$$\bar{\mu}=\frac{1}{2}(a'\mu^{(1)}+a'\mu^{(2)}) \quad (\text{其中 } a=\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})),$$

试证明：

(1) $E(a'X|G_1)>\bar{\mu}$;

(2) $E(a'X|G_2)<\bar{\mu}$.

5-4 设有两个正态总体 G_1 和 G_2,已知$(m=2)$

$$\mu^{(1)}=\begin{bmatrix}10\\15\end{bmatrix},\quad \mu^{(2)}=\begin{bmatrix}20\\25\end{bmatrix},$$

$$\Sigma_1=\begin{bmatrix}18 & 12\\12 & 32\end{bmatrix},\quad \Sigma_2=\begin{bmatrix}20 & -7\\-7 & 5\end{bmatrix},$$

先验概率 $q_1=q_2$；而 $L(2|1)=10$，$L(1|2)=75$. 试问样品

$$X_{(1)}=\begin{bmatrix}20\\20\end{bmatrix}\quad 及\quad X_{(2)}=\begin{bmatrix}15\\20\end{bmatrix}$$

各应判归哪一类？

(1) 按费希尔判别准则；

(2) 按贝叶斯判别准则$\left(\text{假定 }\Sigma_2=\Sigma_1=\begin{bmatrix}18 & 12\\12 & 32\end{bmatrix}\right)$；

(3) 已知样品 $x=(20,20)'$，试计算后验概率 $P(G_i|x)$ $(i=1,2)$.

5-5　已知 $X_{(i)}^{(t)}(t=1,2;\ i=1,\cdots,n_i)$为来自 G_t 的样本. 记

$$d=\overline{X}^{(1)}-\overline{X}^{(2)},$$

其中 $\overline{X}^{(i)}=\dfrac{1}{n_i}\sum\limits_t X_{(t)}^{(i)}$ $(i=1,2)$；

$$S=\frac{1}{n_1+n_2-2}(A_1+A_2).$$

试证明：$a=S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$使比值$(a'd)^2/a'Sa$ 达最大值，且最大值为马氏距离 D^2(其中 $D^2=(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$).

5-6　在两个 p 元正态总体 $N_p(\mu^{(i)},\Sigma)$ $(i=1,2)$下，设 $\mu^{(1)}$，$\mu^{(2)}$，Σ 均为已知. 又设线性判别函数为

$$W(X)=(X-\bar{\mu})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),\quad \bar{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)}),$$

判别准则为：

$$\begin{cases}判\ X\in G_1, & 当\ W(X)>0,\\ 判\ X\in G_2, & 当\ W(X)\leqslant 0.\end{cases}$$

试求错判概率 $P(2|1)$和 $P(1|2)$.

5-7　已知两个总体的分布为 $N_p(\mu^{(i)},\Sigma)$ $(i=1,2)$. 又设 $\mu^{(1)}$，$\mu^{(2)}$，Σ 均已知，先验概率为 q_1 和 q_2 $(q_1+q_2=1)$，错判损失为 $L(1|2)$和 $L(2|1)$. 试写出贝叶斯判别准则和距离判别准则，并说明

它们之间的关系.

5-8　用逐步判别法选择判别变量的过程中(已知训练样本总容量 $n=30, k=3$,考察的变量个数 $m=4$). 已知在第一步引入变量 X_3 后合并组内离差阵 A 和总离差阵 T 分别化为

$$A^{(1)}=T_3(A)=\begin{bmatrix} 28571.5 & 683.4 & -1.123 & 9464.3 \\ 683.4 & 114.9 & -0.519 & 1230.0 \\ 1.123 & 0.519 & 0.0027 & 3.845 \\ 9464.3 & 1230.0 & -3.845 & 15375.8 \end{bmatrix},$$

$$T^{(1)}=T_3(T)=\begin{bmatrix} 28884.9 & 671.2 & -1.172 & 9233.8 \\ 671.2 & 148.3 & -0.347 & 1877.6 \\ 1.172 & 0.347 & 0.0018 & 0.508 \\ 9233.8 & 1877.6 & -0.508 & 27925.9 \end{bmatrix}.$$

试问下一步可否引入变量? 引哪一个?

5-9　设在某地区抽取了 14 块岩石标本,其中 7 块含矿,7 块不含矿. 对每块岩石测定了 Cu,Ag,Bi 三种化学成分的含量,得到的数据如表 5.3.

表 5.3　岩石化学成分的含量数据

类型	序号	Cu	Ag	Bi	类型	序号	Cu	Ag	Bi
	1	2.58	0.90	0.95		8	2.25	1.98	1.06
	2	2.90	1.23	1.00		9	2.16	1.80	1.06
含	3	3.55	1.15	1.00	不	10	2.33	1.74	1.10
	4	2.35	1.15	0.79	含	11	1.96	1.48	1.04
矿	5	3.54	1.85	0.79	矿	12	1.94	1.40	1.00
	6	2.70	2.23	1.30		13	3.00	1.30	1.00
	7	2.70	1.70	0.48		14	2.78	1.70	1.48

(1) 假定两类样本服从正态分布,试用广义平方距离判别法进行判别归类(先验概率取为相等,并假定两类样本的协方差阵相等);

(2) 今得一块标本,并测得其 Cu,Ag,Bi 的含量分别为 2.95, 2.15 和 1.54,试判断该标本是含矿还是不含矿?

5-10　已知某研究对象分为三类,每个样品考察 4 项指标,各类的观测样品数分别为 7,4,6;另外还有 3 个待判样品(所有观测数据见表 5.4). 假定样本均来自正态总体.

(1) 试用马氏距离判别法进行判别分析，并对 3 个待判样品进行判别归类.

(2) 使用其他的判别法进行判别分析，并对 3 个待判样品进行判别归类，然后比较之.

表 5.4　判别分类的数据

样品号	X_1	X_2	X_3	X_4	类别号
1	6.0	−11.5	19.0	90.0	1
2	−11.0	−18.5	25.0	−36.0	3
3	90.2	−17.0	17.0	3.0	2
4	−4.0	−15.0	13.0	54.0	1
5	0.0	−14.0	20.0	35.0	2
6	0.5	−11.5	19.0	37.0	3
7	−10.0	−19.0	21.0	−42.0	3
8	0.0	−23.0	5.0	−35.0	1
9	20.0	−22.0	8.0	−20.0	3
10	−100.0	−21.4	7.0	−15.0	1
11	−100.0	−21.5	15.0	−40.0	2
12	13.0	−17.2	18.0	2.0	2
13	−5.0	−18.5	15.0	18.0	1
14	10.0	−18.0	14.0	50.0	1
15	−8.0	−14.0	16.0	56.0	1
16	0.6	−13.0	26.0	21.0	3
17	−40.0	−20.0	22.0	−50.0	3
1	−8.0	−14.0	16.0	56.0	
2	92.2	−17.0	18.0	3.0	
3	−14.0	−18.5	25.0	−36.0	

5-11　某城市的环保监测站于 1982 年在全市均匀地布置了 14 个监测点，每日三次定时抽取大气样品，测量大气中二氧化硫、氮氧化物和飘尘的含量. 前后 5 天，每个取样点(监测点)每种污染元素实测 15 次，取 15 次实测值的平均作为该取样点大气污染元素的含量(数据见表 5.5). 表中最后一列给出的类号是使用第六章将介绍的聚类分析方法分析得到的结果(第 1 类为严重污染地区，第 2 类为一般污染地区，第 3 类为基本没有污染地区).

(1) 试用广义平方距离判别法建立判别准则(假设三个总体为多元正态总体，其协方差阵相等，先验概率取为各类样本的比例)，并

列出回判结果.

(2) 该城市另有两个单位在同一期间测定了所在单位大气中这三种污染元素的含量(见表 5.5 中最后两行),试用马氏距离判别方法判断这两个单位的污染情况属哪一类.

表 5.5　大气污染数据

样品 \ 污染元素	二氧化硫 (X_1)	氮氧化物 (X_2)	飘　尘 (X_3)	类别
1	0.045	0.043	0.265	2
2	0.066	0.039	0.264	2
3	0.094	0.061	0.194	2
4	0.003	0.003	0.102	3
5	0.048	0.015	0.106	3
6	0.210	0.066	0.263	1
7	0.086	0.072	0.274	2
8	0.196	0.072	0.211	1
9	0.187	0.082	0.301	1
10	0.053	0.060	0.209	2
11	0.020	0.008	0.112	3
12	0.035	0.015	0.170	3
13	0.205	0.068	0.284	1
14	0.088	0.058	0.215	2
15	0.101	0.052	0.181	
16	0.045	0.005	0.122	

第六章　聚类分析

聚类分析又称群分析，它是研究对样品或指标进行分类的一种多元统计方法. 所谓的“类”，通俗地说就是相似元素的集合.

在实际问题中，经常遇到分类问题，例如对某城市按大气污染的轻重分成几类区域；对某年级学生按各科的学习情况分为几种类型；对学生在中学期间学习的科目按培养运算能力、培养推理能力、培养记忆能力等分成几组；对人体测量的几十个部位的尺寸按反映人体高矮，反映人体胖瘦及人体畸形的部位分为几类；在经济学中根据人均国民收入、人均工农业产值、人均消费水平等多种指标对世界上所有国家的经济发展状况进行分类等等. 随着生产技术和科学的发展，在许多领域中都将遇到分类问题.

什么是分类？它只不过是将一个观测对象指定到某一类(组). 分类的问题可以分成两种：一种是对当前所研究的问题已知它的类别数目及各类的特征(例如分布规律，或来自各类的训练样本)，我们的目的是：要将另一些未知类别的个体正确地归属于其中某一类，这是第五章判别分析所要解决的问题. 另一种是事先不知道研究的问题应分为几类，更不知道观测到的个体的具体分类情况，我们的目的是：需要通过对观测数据所进行的分析处理，选定一种度量个体接近程度的统计量，确定分类数目，建立一种分类方法，并按接近程度对观测对象给出合理的分类. 后一种问题在实际中大量存在，它正是聚类分析所要解决的问题.

§6.1　聚类分析的方法

聚类分析是实用多元统计分析的一个新的分支，正处于发展阶段，理论上虽不很完善，但由于它能够解决许多实际问题，因此这个

方法很受人们的重视，特别是和其他方法联合起来使用往往效果更好. 例如对一批观测对象先用聚类分析进行分类，然后用判别分析的方法建立判别准则，用以对新的观测对象判别归类.

聚类分析的功能是建立一种分类方法，它将一批样品或变量，按照它们在性质上的亲疏、相似程度进行分类.

聚类分析的内容十分丰富，按其聚类的方法可分为以下几种：

(1) 系统聚类法：开始每个对象自成一类，然后每次将最相似的两类合并，合并后重新计算新类与其他类的距离或相近性测度. 这一过程一直继续直到所有对象归为一类为止. 并类的过程可用一张谱系聚类图描述.

(2) 调优法(动态聚类法)：首先对 n 个对象初步分类，然后根据分类的损失函数尽可能小的原则对其进行调整，直到分类合理为止.

(3) 最优分割法(有序样品聚类法)：开始将所有样品看成一类，然后根据某种最优准则将它们分割为二类、三类，一直分割到所需的 K 类为止. 这种方法适用于有序样品的分类问题，也称为有序样品的聚类法.

(4) 模糊聚类法：利用模糊集理论来处理分类问题，它对经济领域中具有模糊特征的两态数据或多态数据具有明显的分类效果.

(5) 图论聚类法：利用图论中最小支撑树的概念来处理分类问题，创造了独具风格的方法.

(6) 聚类预报法：利用聚类方法处理预报问题，在多元统计分析中，可用来作预报的方法很多，如回归分析和判别分析. 但对一些异常数据，如气象中的灾害性天气的预报，使用回归分析或判别分析处理的效果都不好，而聚类预报弥补了这一不足，这是一个值得重视的方法.

聚类分析根据分类对象的不同又分为 R 型和 Q 型两大类，R 型是对变量(指标)进行分类，Q 型是对样品进行分类.

R 型聚类分析的目的有以下几方面：

(1) 可以了解变量间及变量组合间的亲疏关系；

(2) 对变量进行分类；

(3) 根据分类结果及它们之间的关系，在每一类中选择有代表

性的变量作为重要变量，利用少数几个重要变量进一步作分析计算，如进行回归分析或Q型聚类分析等.

Q型聚类分析的目的主要是对样品进行分类.分类的结果是直观的，且比传统分类方法更细致、全面、合理.当然使用不同的分类方法通常会得到不同的分类结果.对任何观测数据都没有惟一"正确的"的分类方法.实际应用中，常采用不同的分类方法，对数据进行分析计算，以便对分类提供具体意见，并由实际工作者决定所需要的分类数及分类情况.

本章重点介绍在实际问题中应用最广泛的系统聚类法，且主要讨论Q型聚类分析问题.

§6.2　距离与相似系数

为了对样品(或变量)进行分类，就必须研究它们之间的关系.描述样品间亲疏相似程度的统计量很多，目前用得最多的是距离和相似系数，这两个统计量与变量的类型密切相关，我们首先回顾一下变量的类型.

根据变量取值的不同，变量可分为两大类：定量变量和定性(属性)变量.

定量变量就是我们通常所说的连续变量，例如长度、重量、产量、人口、温度等，它们是由测量或计数、统计所得到的量，这类变量具有数值特征，称为定量变量.

定性变量并非真有数量上的变化，而只有性质上的差异，例如天气(阴、晴)，性别(男、女)，职业(工人、教师、干部、农民等)，质量(一等、二等、三等)，矿石的质量(富、中、贫)等.这些变量都是定性变量，在这类变量中还可以再分为两种：有序变量(没有明确的数量关系，只有次序关系，如质量的等级)和名义变量(变量值是几个没有次序关系的不同状态，如性别、职业等).

不同类型的变量在定义距离或相似性测度时有很大差异.在实际应用中更多遇到的是定量数据的聚类分析问题.下面先介绍定量

变量数据在聚类分析之前进行数据变换的一些方法.

一、数据的变换方法

设有 n 个样品,每个样品测得 m 项指标(变量),得观测数据 x_{ij} $(i=1,\cdots,n;j=1,\cdots,m)$. 通常将数据列成表 6.1 的形式. 表中:

表 6.1 观测数据及特征量

样品 \ 变量	X_1	$\cdots$	X_j	$\cdots$	X_m
$X_{(1)}$	x_{11}	$\cdots$	x_{1j}	$\cdots$	x_{1m}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_{(i)}$	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{im}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X_{(n)}$	x_{n1}	$\cdots$	x_{nj}	$\cdots$	x_{nm}
均值	$\bar{x}_1$	$\cdots$	$\bar{x}_j$	$\cdots$	$\bar{x}_m$
标准差	s_1	$\cdots$	s_j	$\cdots$	s_m
极差	R_1	$\cdots$	R_j	$\cdots$	R_m

均值 $\bar{x}_j=\dfrac{1}{n}\sum_{t=1}^{n}x_{tj}\ (j=1,2,\cdots,m)$,

标准差 $s_j=\sqrt{\dfrac{1}{n-1}\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}\ (j=1,2,\cdots,m)$,

极差 $R_j=\max\limits_{t=1,\cdots,n}x_{tj}-\min\limits_{t=1,\cdots,n}x_{tj}\ (j=1,2,\cdots,m)$.

我们所考察的 m 个不同变量,一般都有不同的量纲,不同的数量级单位,不同的取值范围. 为了使不同量纲,不同取值范围的数据能够放在一起进行比较,通常需要对数据进行变换处理. 常用的变换方法有以下几种.

1. 中心化变换

称变换

$$x_{ij}^{*}=x_{ij}-\bar{x}_j\quad(i=1,2,\cdots,n;j=1,\cdots,m)$$

为**中心化变换**. 变换后数据的均值为 0,而协方差阵不变,即协方差阵为

$$S^* = S = (s_{ij})_{m\times m},$$

其中 $s_{ij} = \dfrac{1}{n-1}\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j) = \dfrac{1}{n-1}\sum_{t=1}^{n}x_{ti}^*x_{tj}^*$.

中心化变换是一种方便地计算样本协方差阵的变换.

2. 标准化变换

称变换

$$x_{ij}^* = \frac{x_{ij}-\bar{x}_j}{s_j} \quad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots,m \end{pmatrix}$$

为**标准化变换**. 变换后的数据,每个变量的样本均值为 0,标准差为 1,而且标准化变换后的数据$\{x_{ij}^*\}$与变量的量纲无关.

3. 极差标准化变换

称变换

$$x_{ij}^* = \frac{x_{ij}-\bar{x}_j}{R_j} \quad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots,m \end{pmatrix}$$

为**极差标准化变换**. 变换后的数据,每个变量的样本均值为 0,极差为 1,且$|x_{ij}^*|<1$,在以后的分析计算中可以减少误差的产生;同时变换后的数据也是无量纲的量.

4. 极差正规化变换(规格化变换)

称变换

$$x_{ij}^* = \frac{x_{ij}-\min\limits_{1\leqslant t\leqslant n}x_{tj}}{R_j} \quad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots,m \end{pmatrix}$$

为**极差正规化变换**. 变换后的数据 $0\leqslant x_{ij}^*\leqslant 1$;极差为 1,也是无量纲的量.

5. 对数变换

称变换

$$x_{ij}^* = \ln(x_{ij}) \quad (\text{要求 } x_{ij}>0,\ i=1,2,\cdots,n;j=1,2,\cdots,m)$$

为**对数变换**. 它可将具有指数特征的数据结构变换为线性数据结构.

此外,还有平方根变换,立方根变换等. 它们的主要作用是把非线性数据结构变为线性数据结构,以适应某些统计方法的需要.

二、样品间的距离和相似系数

描述样品间的亲疏程度最常用的是距离. 设观测数据 x_{ij} ($i=1,2,\cdots,n;j=1,\cdots,m$)列成表 6.1 的形式. n 个样品看成 m 维空间中的 n 个点,用 d_{ij} 表示样品 $X_{(i)}$ 和 $X_{(j)}$ 之间的距离,一般要求:

(1) $d_{ij}\geqslant 0$,对一切 i,j;当 $d_{ij}=0 \Longleftrightarrow X_{(i)}=X_{(j)}$;

(2) $d_{ij}=d_{ji}$,对一切 i,j;

(3) $d_{ij}\leqslant d_{ik}+d_{kj}$,对一切 i,j,k (三角不等式).

对于定量变量,常用的距离有以下几种.

1. 闵科夫斯基(Minkowski)距离

称

$$d_{ij}(q)=\left[\sum_{t=1}^{m}|x_{it}-x_{jt}|^{q}\right]^{1/q}\quad (i,j=1,2,\cdots,n) \tag{6.2.1}$$

为**闵科夫斯基距离**.

(1) 绝对值距离:在(6.2.1)式中,当 $q=1$ 时的一阶闵科夫斯基距离为

$$d_{ij}(1)=\sum_{t=1}^{m}|x_{it}-x_{jt}|\quad (i,j=1,2,\cdots,n),$$

称它为**绝对值距离**.

(2) 欧氏距离:在(6.2.1)式中,当 $q=2$ 时的二阶闵科夫斯基距离为

$$d_{ij}(2)=\sqrt{\sum_{t=1}^{m}|x_{it}-x_{jt}|^{2}}\quad (i,j=1,2,\cdots,n),$$

称它为**欧氏距离**.

欧氏距离是聚类分析中使用最广泛的距离. 但该距离与各变量的量纲有关;没有考虑指标间的相关性;也没有考虑各变量方差的不同. 如从欧氏距离的定义中易见,变差大的变量在距离中的作用(贡献)就会大,这是不合适的. 简单的处理方法就是对各变量加权,比如用 $1/s^2$ 作为权重可得出"统计距离"(或方差加权距离):

$$d_{ij}^{*}(2)=\sqrt{\sum_{t=1}^{m}\left(\frac{x_{it}-x_{jt}}{s_t}\right)^{2}}\quad (i,j=1,2,\cdots,n).$$

(3) 切比雪夫距离：当 q 趋于 ∞ 时，称

$$d_{ij}(\infty)=\max_{1\leqslant t\leqslant m}|x_{it}-x_{jt}| \quad (i,j=1,2,\cdots,n)$$

为**切比雪夫距离**.

2. 兰氏距离(要求 $x_{ij}>0$)

兰氏距离是由 Lance 和 Williams 最早提出的，故称为**兰氏距离**，其定义为

$$d_{ij}(L)=\frac{1}{m}\sum_{t=1}^{m}\frac{|x_{it}-x_{jt}|}{(x_{it}+x_{jt})} \quad (i,j=1,2,\cdots,n).$$

这是一个无量纲的量，克服了闵氏距离与各指标的量纲有关的缺点，且兰氏距离对大的奇异值不敏感，这样使得它特别适合高度偏倚的数据. 但兰氏距离也没有考虑变量间的相关性.

闵氏距离和兰氏距离都是假定变量之间相互独立，即在正交空间中讨论距离. 但在实际问题中，变量之间往往存在着一定的相关性，为克服变量之间相关性的影响，可以采用马氏距离.

3. 马氏距离

样品 $X_{(i)}$ 和 $X_{(j)}$ 的马氏距离为

$$d_{ij}(M)=(X_{(i)}-X_{(j)})'S^{-1}(X_{(i)}-X_{(j)}),$$

其中 S^{-1} 为样本协方差阵的逆矩阵.

马氏距离虽然可以排除变量之间相关性的干扰，并且不受量纲的影响，但是在聚类分析处理之前，如果用全部数据计算均值和协方差阵来求马氏距离，效果不是很好. 比较合理的办法是用各个类的样本来计算各自的协方差阵，同一类样品间的马氏距离应当用这一类的协方差阵来计算，但类的形成需要依赖于样品间的距离，而样品间合理的马氏距离又依赖于类，这就形成了一个恶性循环. 因此在实际聚类分析中，马氏距离也不是理想的距离.

为了克服变量间相关性的影响，我们引入斜交空间距离.

4. 斜交空间距离

由于变量之间存在着不同程度的相关关系，在这种情况下，用正交空间距离来计算样品间的距离，易产生形变，从而使得用聚类分析进行分类时的谱系结构发生变形.

在 m 维空间中，为使具有相关性变量的谱系结构不发生变形，采用由下式定义的**斜交空间距离**，即

$$d_{ij}=\left[\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}(x_{ik}-x_{jk})(x_{il}-x_{jl})r_{kl}\right]^{1/2} \quad (i,j=1,2,\cdots,n),$$

在数据标准化处理下，式中的 r_{kl} 为变量 X_k 和 X_l 之间的相关系数.

5. 相似系数

样品间的亲疏程度除了用距离描述外，也可用相似系数来表示. 参见下面小节三“变量间的相似系数和距离”中的定义.

6. 定性变量样品间的距离或相似系数

以上介绍的样品间的距离或相似系数都是对定量指标定义的. 现介绍定性变量(名义或有序变量)的距离或相似系数的定义方法.

在数量化理论中，常把定性变量叫做**项目**，而把定性变量的各种不同取“值”叫做**类目**. 例如性别是项目，而男或女是这个项目的类目；体形也是一个项目，而适中、胖、瘦、壮等是这个项目的类目. 性别只能取男或女中的一个类目，不能兼取；而体形可以是适中且壮，即可兼取两个类目.

设样品 $X_{(i)}$ 的取值为

$$(\delta_i(k,1),\delta_i(k,2),\cdots,\delta_i(k,r_k)) \quad (i=1,\cdots,n;k=1,\cdots,m),$$

其中 n 为样品的个数，m 为项目的个数，r_k 是第 k 个项目的类目数. 比如在表 6.2 中，当 $k=1$ 时 $X_{(i)}$ 的取值为(1,0,0,0)，这里 $r_1=4$，$\delta_i(1,1)=1$，$\delta_i(1,l)=0$ $(l\neq 1)$. 若

$$\delta_i(k,l)=\begin{cases}1, & \text{当第 } i \text{ 个样品中第 } k \text{ 个项目的定性数据}\\ & \text{为第 } l \text{ 个类目时},\\ 0, & \text{否则},\end{cases}$$

则称 $\delta_i(k,l)$ 为第 k 项目之 l 类目在第 i 个样品中的反应.

设两个样品分别为 $X_{(i)}$ 和 $X_{(j)}$，若 $\delta_i(k,l)=\delta_j(k,l)=1$，则称这两个样品在第 k 个项目的第 l 类目上 1—1 配对；若 $\delta_i(k,l)=\delta_j(k,l)=0$，则称这两个样品在第 k 个项目之 l 类目上 0—0 配对；若 $\delta_i(k,l)\neq\delta_j(k,l)$，则称为**不配对**.

记 m_1 为 $X_{(i)}$ 和 $X_{(j)}$ 在 m 个项目的所有类目中 1—1 配对的总数；m_0 为 0—0 配对的总数；m_2 为不配对总数. 显然，