最新实用统计分析软件包指南

卫生部卫生统计信息中心赵京华等编

一九九七年四月　　北京

前　　言

统计分析软件包是电子计算机技术与统计方法相结合的产物。近年来，随着计算机应用的普及，国内陆续引进了多种数据分析的软件包。由于这些软件种类、版本不同、参考资料不完整，也出现了许多应用上的问题。这既有分析目的与手段之间的矛盾，也有统计分析与计算机应用结合上的困难。实际上，这些软件包不是简单的重叠，而是相互补充。在同事和朋友们的热心鼓励下，我们收集整理成这本《最新实用统计分析软件包指南》，综合介绍这些它们的操作使用。我们感到任务艰巨，但又非常有意义。这有利于找出共性，取长补短，增强软件包的可操作性，同时更好地理解数据处理和统计分析问题，更好地实现统计工作者与计算机专业人员之间的构通，提高应用水平。

本书于九三年初拟稿，九四年七月完成初稿。这次的修订保持了原有的体例并有较大的删节和补充。我们的总目标仍然是力求简明扼要，旨在解决计算机和软件系统最常见问题，对其功能进行详尽的说明以供软件评价和选择参考，便于读者掌握其常用的术语，用最短的时间掌握这些软件包，进一步在数据分析方法与统计软件结合上建立一个系统的概念。本书写作过程中参考了国内外大量文献。虽然目前各统计软件包继续更新换代，但从过去二十几年统计软件包发展过程来看，我们笃信其精神不变。

本书分准备知识、通用软件包、专用软件包和数据、图形处理及辅助工具四篇。第一部分介绍统计和计算机基本概念。第二部分介绍SAS、SPSS、BMDP、SYSTAT、Stata、Splus、Minitab和Genstat。第三部分介绍TSP、GLIM、LISREL 和Epi Info。第四部分介绍软件包之间的数据交换和综合用例，统计绘图，及常用的编辑软件。本书的附录给出有关参考资料和软件包信息。

本书大部分章节由卫生部统计信息中心赵京华编写。第7章和第12 章主要由哈尔滨医科大学马进编写。上海师范大学龚秀芳、北京医科大学公共卫生学院黄湘宁参加了第2章、第4章及第5章中的多元统计部分的编写。本书所采用SAS 6.04，6. 07和汉化EPI INFO由卫生部卫生统计信息中心提供，软件的收集过程中给予帮助的有：中日友好医院段银康硕士、上海医科大学钱文娣讲师、卫生部国外贷款办赵宏雯硕士、预防医学科学院王公昊教授。卫生部统计中心陈育德主任、饶克勤副主任和黄东祖工程师、上海医科大学钱文娣讲师、军事医学科学院胡良平教授提出了建设性的修改意见。本书大部分内容使用中文WPS和中文emTeX录入，我们对软件的作者们表示衷心感谢。由于本书涉及的面很广，我们精力、时间所限，疏漏谬误之处必然很多，望读者不吝指正，以供订正。

编者

一九九七年四月　　北京

目　录

第一部分

准备知识

第一章　概述

§1.1　统计学和统计分析

统计学研究怎样以比较有效的方式收集、整理、分析带随机性的数据，并在此基础上，对所研究的问题作出统计性的推断，直至对可能作出的决策提供依据或建议[1]。简而言之，统计学研究数据的收集、整理、分析并做出推断[2]。

统计学的发展源远流长，社会各个领域和各门自然科学研究都离不开统计学方法。统计学在各个领域中的应用，不但促进了统计方法本身的发展，而且推动了各相关学科的发展。

统计分析是指用统计学的观点和方法对客观事物进行分析和研究，即数据的整理、描述和综合[2]。统计分析与信息提取存在密切关联，将收集的数据“去粗取精、去伪存真、由此及彼、由表及里”地加工处理，通过事物外在的数量表现，揭示事物可能存在的规律性，并结合各门专业知识加以解释。

统计分析常用的概念有：

1. 总体与样本。在实际应用中，把研究对象的全体叫做总体(population)，把每个研究单位叫做个体，把总体中个体的总数叫做总体容量。如果总体中包括有限个单位，就称为有限总体，否则为无限总体。从总体中随机抽出的一部分观察单位就称为概率样本(probability sample)，观察单位的数目为样本容量。统计分析的任务之一就是由概率样本推断总体。其它如自愿者组成的随意样本(haphazard sample)、指定的样本(representative samples)或分配样本(quota samples) 等均是非概率样本。

2. 误差。总体中的个体之间存在着差异，这种差异可以是由多方面原因引起。统计中的误差，是指测量值与真值之差或样本指标与总体指标之差。误差的出现，可以是实验仪器不准、研究对象不同质等导致的系统误差，也可以是由于疏忽造成的过失误差。由于观察单位间存在的个体差异，使得由样本指标与总体指标间存在差异，称为抽样误差。系统误差和过失误差是应该而且可以避免的。

3. 统计分布与统计量。对大量的带有随机性的统计资料的描述，我们得到资料波动变化的信息，统计学常用分布来刻划，如正态分布、二项分布及泊松分布等。设一个随机变量是一些独立同分布随机变量的函数，当样本给定时其值能够唯一确定，并且与总体的未知参数无关，则该随机变量称为统计量(statistics)。如果研究的总体只包含若干个未知参数则称其统计问题是参数性的，否则为非参的。不同的资料要用不同的统计方法来处理。总体与样本、参数与统计量的关系可用图 1.1来表示：

4. 变量分类。统计分析的原始数据通常是n行p列$n \times p$ 的矩形数据矩阵$X = \{x_{ij}\}, i = 1, ..., n, j = 1, ..., p$，它包含了关于n 个对象的p个变量的信息，变量的统计学的意义是指一种可以测量的特征，统计分析时首先要对测量的方法加以考察。变量根据其测量的尺度不同，有表 1.1的分类[8]：

   名义型(nominal)数据是一种纯粹的数据符号，没有量的概念，如婚姻状况中已婚、未婚、离婚、丧偶和分居，可任意记作1,2,3,4,5或A,B,C,D,E。有序型(ordinal)数据就是有先后次序，如由小到大的年龄分组。间隔型(interval)数据所包含的量不仅可以比较大

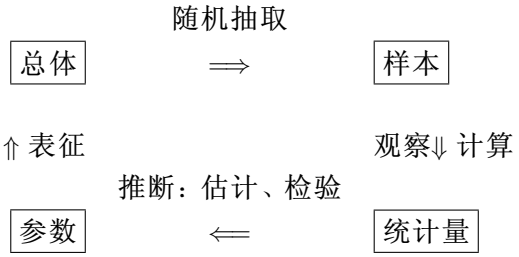


图 1.1 总体与样本、参数与统计量的关系[3]

表 1.1 各种变量的类型、意义与实例

名　　称	意　　义	举　例
名义型	类别间的地位相同	血型、是/否
有序型	类别可以排序	文化程度
间隔型	水平间的差值相同	温度、海拔高度
比率型	水平间比的等同	身高、体重

小，而且可以确定相差的量，温度是最典型的间隔型数据之一，即不仅可以比较温度的高低，而且可以说10℃比20℃低10℃。这一类数据可做复杂的四则运算，“间隔”的确定有一定的任意性，如摄氏和华氏两种温度的间隔就不相同。0点的确定是任意的，如0℃有确定的温度而不是没有温度。比率型(ratio)数据突出的特点是零点具有明确的含义，如重量尽管有不同的计量单位，但重量等于零时的概念很明确，而且任何一种计量单位可经一个比例常数简单地换算成另一种相应的单位。这些标度有几种情况[11]：a .尺子上的值可表为比；b .尺子上的距离有意义；c .尺子上的元可以自小到大排列。间隔尺度无固定的原点，如海拔高度和温度的取值，仅仅是与一个相对的值比较，仅有 b 、c 的性质。最后，仅有条件 c 成立的称做有序的，分类A,B,C,D不能说成立A-B=C-D。名义变量不具有前面的三个性质。

变量的其它一些分类方法也与上述尺度有关。当变量测量的尺子无限可分时，称做连续的(continuous) ，如长度由公里到米到厘米等等，否则是离散的(discrete) 。名义的也称计数的(qualitative) ，间隔和比率的也称计量的(quantitative)，有序的也称等级的。分类数据(categorical variable) 度量尺度是不确定的，可以是名义的，或者是有序的，或者是间隔的。

统计学研究的课题有：

1. 抽样技术与实验设计(sampling technique and design of experiment)。统计分析中大量带有随机性的数据通常来自观察和实验，如医学中的临床试验、围绕某个专题进行的抽样调查等。统计抽样要保证收集到的样本数据在性质上和数量上对总体有代表性；实验设计包括专业设计和统计学设计。资料的收集至关重要，做不好这一环，则统计分析必是“垃圾进去，垃圾出来(garbage in and garbage out，GIGO)”。

2. 描述性分析(descriptive analysis)。简而言之，描述性分析是对数据的综合与提练[6]。收集

的数据经过核实和归纳，计算出统计指标，或者结合统计图、统计表等手段表达出来，使对数据有一个概括的印象，就完成了数据的描述性分析。

复杂的分析常基于对数据基本的了解，统计原始数据的整理加工，传统上主要有三种方法：a .原始数据分组，从而获得各种统计图表；b .原始数据排序；c . 按统计推断的要求，把原始数据归纳为一个或几个数字特征。D.R.Cox and E. J. Snell(1981)指出，典型的数据质量检查如：a .直观或自动检查在逻辑上不一致或与先验知识相悖的情况；b . 检查主要变量的频数分布以找出少数有差异的数据；c .通过可能高度相关的变量对之间的散点图更敏感地检查；d .数据收集方法的检查找出测量上可能存在的偏差，如观察者间的差异，主要变量的编码出现的问题等；e . 寻找缺失的记录，包括省略掉的那些记录，缺失值常用一种简便的方法标识，如99,999等，任何分析都不应包括这些数据。

探索性分析可以检测数据的错误、寻找数据存在的模式或关系。常见的数据错误有：a . 数字位次搞错，1984 写成了9184；b . 某个变量的值用错，如用错变量，甚至数据集；c . 把变量和观察关系搞混，录入不是按照观察而是按照变量；d . 实验误差。常见的模式或关系有：a . 线性关系；b . 资料具有重复，宜采用方差分析；c .交叉或嵌套资料；d . 时间趋势，如按时间的自然排序；e 、边界点，在边界以外无观察；f . 改变系统的点，即数据中出现一个冲激，此时应分别配合模型；g .异常点(outlier)。h . 数据堆积(clump)，即数据可分成两个或多个点的集合，从而把拟合的方程拉向自己。其他的方面如被研究量的缺失，实验随机化的影响、趋势外推等。在探索性数据分析中，常出现的概念有：局部稳定性(resistance)是指对于数据中局部出现的不正常变化的不敏感性，中位数相对于均数就是如此。稳健性(robustness) 是指围绕特定概率模型假设下的偏离不敏感。

描述性分析对其它统计分析至关重要。统计指标在统计软件中往往以综合统计量(summary statistics) 的形式出现，它是进一步分析的基础。如描述集中趋势的均数(如算术均数、几何均数、调和均数)与中位数，众数表示最大频数所在的位置。描述离散情况的标准差、全距、百分位差、偏度、峰度等，它们从不同的角度综合了数据的特性。图形分析(graphical analysis)如线图、圆图、直方图和直条图，茎叶图、箱式图、星形图、统计地图等，用于直观描述资料特征。对于多维的情况，可以用样本的相关系数、协方差阵描述它们之间的关系。数据的整理还包括一些变换，如Box-Cox 转换等。探索性数据分析(exploratory data analysis, EDA) 丰富了描述性分析的内容。

3. 统计推断(statistical inference)。基于收集的数据，以及对数据整理分析的结果，对数据所来自于总体的情况作出一定的论断，即是统计推断，也就是由样本作出关于总体结论的过程。简言之，统计推断就是由数据下结论[2]。统计推断可以用图 1.2来示意：

推断要解决三个问题，即由样本算出估计量，把它作为总体参数的点估计；对估计量算得一个确切的区间作为总体参数的区间估计，所估计的区间称为可信区间；应用统计检验来判断样本信息支持总体参数值是否理想。与推断有关的是统计决策问题，它是基于收集的数据，使用统计推断理论中提供的种种方法，结合经济上可能的后果进行分析。常用的估计方法有最大似然估计、最小二乘估计、贝叶斯估计等。

可信区间基于样本资料和α水平，指出在$1-\alpha$ 的概率下该区间包含总体的参数，$1-\alpha$称为可信度，通常取为95%或99%。假设检验中把受到保护的假设定为原假设(null hy-

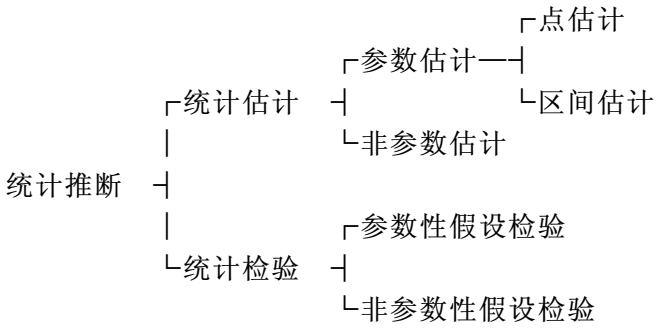


图 1.2 统计推断方法

pothesis，也称无效假设)，对立的假设定为备择假设(alternative hypothesis)。根据样本提供的信息计算检验统计量，在原假设下，检验统计量取该值或比该值更极端的概率可以算出，若概率值较小，就拒绝原假设。一般，取$\alpha = 0.05$ 做为小概率事件。

假设检验可能犯两类错误：当原假设为真但检验的结果否定其真实性，称为 Ⅰ 类错误，常用α来表示；当原假设为假而检验的结果不能够拒绝它，称 Ⅱ 类错误，记其概率为β。检验的能力或功效可以用$1 - \beta$表示。Ⅰ 类错误由研究者控制，等于设定的α水平。Ⅱ 类错误受一系列因素影响：备选假设离无效假设的距离较大时，功效要高；它也受α水平的影响，常常要在二者之间有一个折衷。通常的做法是控制 Ⅰ 类错误的条件下，使 Ⅱ 类错误尽可能地小。在多元分析中，研究者倾向于对多个参数进行推断，犯错误的概率就要大。若在给定的α水平下进行k个假设检验，则当所有无效假设为真时，以相同的水平完成k个检验，接受所有假设的概率$\geq 1 - k\alpha$；若各检验独立，则错误概率是$1 - (1 - \alpha)^k$，这样若在0.05水平上检验十个假设，且这些检验所获p值的分布独立同分布，则在原假设下得到一个特定检验为显著的概率为0.05而声明十个检验至少一个为显著的概率为0.401；若是20个检验，则后面的概率为0.642。因此检验多个假设或求多维可信区间，常常要控制任何一个 Ⅰ 类错误的概率为α，这被称做是控制实验的(experimentwise) 或分析的(analysiswise) α 水平。有许多方法处理这种情形，如Bonferroni 检验(若k个假设在α/k水平检验，则 Ⅰ 类错误的概率不大于α)、Scheffe 区间及其到Roy区间的推广、Tukey 检验及ominbus检验(对向量的检验)。在探索性研究或者当研究者想在资料中寻找有价值的东西时，由于有许多机会出现阳性结果，就应有控制 Ⅰ 类错误的保守方法。当对所研究的领域有更多的把握时，统计研究时的假设成为真正的假设，可以冒风险对每个检验用α水平，此时类似一般的假设检验。检验就是依照假设的(hypothesewise或comparisonwise 或parameterwise) 的α。

4. 多元分析(multivariate analysis)。我们对客观现象进行观察和实验，往往是多因素相互关联、相互影响的。对这些现象的分析，便是多元分析研究的对象。由于这些方法的背景和处理对象不同而有多种分类方法。传统的多元与逐步回归、多元方差协方差分析、典型相关分析、主成分与因子分析、聚类分析、时间序列分析，以及一些新的统计方法，如回归中的有偏估计、Logit 分析与Cox 回归和生存分析。多元方法处理问题的范围正逐渐扩大，如计量经济学分析和群体遗传学分析。

多元分析是数理统计最重要的分支之一，它的理论和方法在近半个世纪获得了飞

速的发展，尤其是最近十几年来由于计算机的普及，在许多领域和学科中，如生物、医学、地质、农业、工程技术、气象、社会经济等方面，得到日益广泛的应用，使得这个分支的发展更加活跃和深入。

统计研究主要的步骤是：

1. 初步的数据处理，即把数据归整为便于详细分析和数据检查的形式。2. 初步分析，目的在于搞清数据总的模式并提示下一步分析的方向，这可由简单的图表完成。3. 进一步分析，是论断的基础。4. 用准确、简明的的形式表达得到的结论[5]。即明确问题、收集资料、分析资料、展示结果[6]。

根据表 1.1对变量的分类，统计分析大致有表 1.2的分类[8]：

表 1.2 统计分析的分类

	自 变 量			
	名义的/有序的		间隔的/比率的	
因变量	一个变量	多于一个	一个变量	多于一个
无因变量	χ^2拟合度	关联度的测量	一元统计量	相关阵
		对数线性模型	描述统计	主成分分析
		χ^2独立检验	正态性检验	因子分析
				聚类分析
名义的/有序的				
一个变量	χ^2检验	对数线性模型	判别分析	判别分析
	Fisher 精确	LOGISTIC 回归	LOGISTIC回归	LOGISTIC回归
	检验		一元统计量	
			(两组t)	
多于一个	对数线性模型	对数线性模型	判别分析	判别分析
间隔的/比率的				
一个变量	t 检验	方差分析	线性回归	多元回归
	方差分析	多分类分析	相关	生存分析
	生存分析	生存分析	生存分析	
多于一个	方差分析	方差分析	典型相关	典型相关
	主成分ANOVA	主成分ANOVA		通径分析
	Hotelling T^2			结构分析
	轮廓分析			(LIRESL,EQS)

多元分析中因变量(dependent)与自变量(explanatory) 是由变量取值变化是否依其它变量而定，并没有尺度的区分。如临床上用新生儿月龄推算体重，月龄是自变量，体重是因变量。回归分析中有因变量，聚类分析无因变量，等等。A.J. Hartley(1983) 按照变量的地位分成两类。第一类中变量的地位相同，有主成分分析、因子分分析及聚类分析；第二类中各变量的地位不同，可以区分因变量和自变量，等等。

与探索性分析相应，确证性(confirmatory) 分析是预先指示某种假设或模型，利用数据

检查模型的拟合情况，因此有：剩余(residual)=数据值(data)-模型拟合值(model)，数据表示基于样本的观测，模型表示假设观测所包含的结构，剩余是两者的差。模型的好坏常用拟合优度检验(goodness of fit test)，它是检验原始数据的样本分布是否与已知分布相符合的方法。拟合优度检验常用χ^2检验和Kolmogorov-Smirnov 检验。

总之，从实验设计到资料的收集、整理和分析，是有机的统一体，软件包也遵循了这个框架。常用内容有：

描述统计(descriptive statistics)
概率(Probability)
概率分布(probability distribution)
估计(estimation)
显著性检验(significance tests)
回归(regression)
方差分析(ANOVA)
线性模型(General Linear Models)
广义线性模型(Generalized Linear Models)
抽样调查(sample survey)
实验设计(design of experiments)
临床实验(clinical trials)
多元分析(multivariate analysis)
时间序列分析(time series analysis)
质量控制及可靠性(quality control and reliability)

从实际的分析来看，大多数时间常常用于数据的整理上。现场收集的数据必须进行逻辑检查，不同项目间交叉有许多缺失值，在处理上有其特色。

§1.2 统计分析软件包的种类

§1.2.1 计算机软件和软件包

计算机软件是计算机赖以完成其功能的一系列指令、程序等的统称。软件包不是子程序的集合，而是建立在基本功能之上的一个体系结构。统计软件包的应用开始于六十年代，主要有SAS、SPSS及BMDP，以批处理方式运行于大型机上，它们只有简单的数据类型。伴随着计算机软硬件等一系列新技术的发展，八十年代初出现了S 语言，增加了图形、交互式处理和菜单驱动的接口。目前，统计软件包是群星夺目，它们综合了计算机和统计学发展的最新成果。G.E.P. Box (1969) 生动描述了这种情形：现在的事情就这么简单，尽管可以对自己在做什么一无所知，但只肖借助于计算机，你仍能用令人难以置信的速度得到不论是正确的还是错误的所有的东西。

作为统计分析与计算机的有机结合体，统计软件包反映了计算机在统计中的作用：数据存贮及操作、综合数据特征、探索性数据分析、推断应用中的计算、模拟研究[6]。大量的数据处理和统计分析离开统计软件包是不可想象的，统计分析软件包的应用，使深奥的分析技术走出了统计的“象牙塔”，使人们从大量繁琐的手工计算、重复编程中解脱出来。它既能帮助统计工作者和各专业研究者把统计理论应用于实践，引导理论工作者走向应用领

域，同时又在很大程度上帮助实际工作者从现实的直观背景中认识统计理论，最终提高应用水平和理论水平，可见软件包成了构通理论和实际工作者、统计工作者和计算机人员的桥梁，也遵循了“理论——实践——理论”循环往复的认识论规律。

§1.2.2 统计分析软件包的种类和特点

目前统计软件包的种类很多，我们大致把它们归为通用和专用两类，本书主要涉及以下几种：

SAS，是一个通用的软件包，它的数据处理和分析能力很强，同时也是一个数据管理与撰写报告的工具。它拥有强大的编程能力和丰富的统计方法，而且能够保存大部分中间结果，尤其适于比较复杂的数据处理和建模。SAS系统庞大，对非统计专业人员来说掌握比较困难。

SPSS，是一个典型的统计分析软件包，最初面向社会科学开发但其应用远不限于社会科学，其特点是操作简便，有良好的选择式菜单和运行提示。BMDP 是古老而著名的软件包，它采用模块化调用，各种统计分析功能由独立的模块实现，也采用了方便调用的菜单操作方式，但模块的独立调用也造成代码上的重复。SYSTAT 也是一个装卸方便、典型的模块化软件，其图形功能很强。SYSTAT 的LOGIT和生存分析要用附加的LOGIT、SURVIVAL 模块。

Stata，是一个数据处理、图形显示和统计分析高度集成化的软件包，它同时也可以使用附加模块，其源程序对用户透明，从而使它生机勃勃。它提供了最好的软件运行使用说明。

S-Plus，引进了目标编程，它是一个功能强大的函数语言，提供了丰富的数据分析和图形显示功能。

Minitab，是一个特别适于教学使用的软件包，包括了常规的统计方法，其宏定义允许用户进行功能拓展。

GLIM，是专为广义线性模型设计的软件包，广义线性模型具有广泛的理论与实际意义。Genstat 功能与用法上与GLIM 类似，其开发宗旨在于通用，故其功能扩大到处理一般统计问题。这两种软件对于进行新模型方法的研究很有用。

LISREL，是专用于结构方程模型分析的软件，该软件广泛应用于心理学、社会学、经济学等相关学科资料的分析。是确证型分析的典范。

MicroTSP，是一个时序和经济分析软件包。

Epi Info，特别适于现场资料的收集和流行学分析，它能与大多数统计软件包进行数据交换，英文版的Epi Info 软件已由有关单位汉化成功。

综观这些统计软件包，有如下功能特点：1.丰富的函数和强大的编程能力，通常多数操作可在软件包内完成；2.系统的统计方法，随着原有软件版本的更新，现代统计方法不断被吸收。相互借鉴，纳入一些专用的程序，如SAS/STAT的CALIS 和SPSS 的LISREL；3.数据处理功能强，如适应和生成各种复杂的数据格式，生成新的变量，缺失值处理等；4.丰富的图形功能，标准图形的格式交换。5.用户界面友好，可以交互式或批处理方式运行。如PC Windows版SPSS 和SAS 不仅考虑方便操作，同时也接受命令输入，用户可择其所好。另外，它们有教学程序、样本程序、多窗口和在线帮助、运行提示等；5. 程序可移植性，有相应微机、小型机、巨型机上应用的软件。

计算机系统上的改进促进了软件包的更新换代，如近年来PC Windows及软件包开发语言可移植性改善了软件包的可移植性。它们将随着统计分析技术和计算机技术而继续发展，但在数据的智能化处理方面仍待突破，如专家系统的开发。现有软件包强大的功能没有影

响新软件的出现，正说明一个软件包不能成为“万应灵药”，而是“兼收并蓄”，“吐故纳新”，不断发展。

统计软件包是方便研究与应用的手段，它也在理论的开拓与应用的深度和广度方面，给统计工作者提出了更高的要求。统计软件包的应用离不开现实的背景，不断的探索，同时也要求具有一定的的理论水平。正如达芬奇(Leonard da Vinci, 1452-1519) 所说的那样：没有理论的实践，就象一个水手没有舵手和指南针，只能随波逐流。

§1.3 统计分析软件包的应用

软件包使用有几个前提条件[7]：首先要有适用的程序或软件包以及人力；其次应有把统计问题化为计算问题的能力，以保证正确使用这些程序或软件包；再次，能够正确地解释结果；最后还应有分析的资源，包括计算机、充足的经费和时间等。不考虑分析的研究是屡见不鲜的，这往往造成资源的浪费。

软件包使用的便利和效率之间常常需要一个折衷。菜单式直观简明，但行命令式纠错方便、运行速度快。灵活性包括数据的操作、宏调用、与其它软件的接口等。用户支持包括软件包提供的用户指南、示范程序的提供情况、教学与培训情况等。专用程序的好处是能够在软件包尚未包括所用的统计方法时采用，其缺点是书写费时，程序功能需要增强时，程序员往往不在场。

统计软件包有效的使用，必须借助扎实的统计理论基础和实际应用经验。我们推荐“以小见大”，即开始阶段用小的问题进行试验，进而“举一反三”。对于比较复杂的问题，阅读说明书、专著和向有经验的用户学习是很简捷的。一些要点如：

- 交互式与批处理方式的进入与退出、文件的操作方法。如程序文件的存贮方法：这些软件包的程序文件有习惯上的扩展名如SAS,INC,INP,DO,CMD,GLM。提交执行则是SUBMIT、DO以及INC。填充执行如SAS 和Stata 的MENU。又如系统外壳命令的使用，SAS用命令行的x命令、SPSS/PC+和SYSTAT用DOS.、Stata用!命令。
- 教学程序与样本程序的使用。软件包推出时，往往随软件提供了运行实例，利用上述文件输入输出方法可方便使用。建议开始用软件包统计分析时，用一些熟悉的例子进行一些演算和对比，明确其有关术语和体例，更主要是能从中获得结果的解释方法。
- 软件包手册、专著和专家请教，往往更可以事半功倍。
- 许多材料可以取自Internet，本书附录给出了一些地址可供参考。

选用软件包之前，应明确实际的需要和软件包分析的特点，以回归分析为例，多数软件包能够进行模型拟合和回归诊断，两阶段最小二乘、伪变量回归、有偏估计如岭回归等。软件包之间可以相互取长补短，如微机上的SAS 6.04 没有COX 生存分析过程，而Stata 2.0就已经提供了该项功能。在回归分析中对记录进行加权时，SPSS相对于SAS和Stata有其独到之处。

问题的程序化能力对于软件包功能的有效发挥是至关重要的，如掌握数据库管理系统技术和高级语言乃至有关计算方法。软件包的功能可以受限，但使用数据库和高级语处理问题是没有限度的。著名的Fortran语言专用程序库有如IMSL、NAG。使用Fortran 时，要花很多时间搞清楚数据类型和格式，编程时的数据定义也要占据大量时间，调试也很费时。这些经验恰恰也是软件包程序化和调试的必要准备。SAS的编程采用模块调用，组合DATA

和PROC，思路非常清楚，调试也方便，出现错误时，系统以红色在记录窗口给出提示。系统能方便地与其它数据管理系统数据交换，如VAX/VMS 下与SPSS、BMDP 的交换。系统既有解释语言纠错的特点又有编译语言快速执行的优点。但如果用户已能熟练使用Fortran 语言，则对于软件包的使用大有好处，建议从事数据处理和分析的人能够了解一些Fortran 语言的有关知识。如Fortran对于记录的引用还是较SAS方便的，SYSTAT 还提供了Fortran 格式数据调用的子程序。

一开始接触软件包说明书，容易“只见树木，不见森林”,《指南》编写介绍其概貌，我们力求简明扼要，抓住共性。不同于传统的方式[8,9,10]，我们采用分别介绍的方式，各部分自成体系。这与新近出版的[12]体例相同但更充实。限于篇幅，我们仅对SAS 和SPSS详细介绍。

第二章　统计学基本知识

§2.1　调查设计与实验设计

资料收集是统计分析的第一步。如第1章所述，其基本方法是抽样和实验。调查设计中，观察者处于被动地位对感兴趣的事物进行研究，如观察吸烟与肺癌是否有关系；实验设计是在严格控制实验条件下，安排实验因素，排除非实验因素的干扰，如物理实验、动物实验和临床试验。实验数据是统计数据的重要来源之一，其处理方法自然是统计软件包处理的典型问题。这里介绍统计实验设计的原则及常见实验设计的概念，更详细的内容可以参阅有关书籍。

§2.1.1　调查设计

1. 普查(mass screening，census)。即全面调查，指对调查范围内全部对象进行调查的方法，其目的是掌握某一时点、一定范围内的研究对象的基本情况，如人口普查、专门人才普查等。人口普查是收集、编制、评价、分析及出版某范围的地区在一个特定时期人口及有关经济和社会资料的全过程，它是重要的国情调查，如1990年全国第四次人口普查。普查的优点是比较有把握地掌握研究对象的基本情况，避免抽样调查中的抽样误差，但需要大量的人、财、物投入，易出现遗漏、由于参与人员多而标准不易统一，大规模全面调查常常只适于做描述研究。

2. 抽样调查(sampling survey)。根据随机的原则，从研究对象的全体中抽取一定数量进行调查的一种调查方法。随机抽样方法有多种，如单纯随机抽样(simple random sampling, SRS)、系统(systmatic) 抽样、分层(stratified) 抽样、整群或集落(cluster) 抽样和多阶段(multistage) 抽样等，在实际应用中可根据具体情况操作和调整。其优点是省时、经济、易于操作，适用范围广，准确性也较好。为保证样本对于总体的代表性，所抽样本应足够大，抽样应随机，调查单位应同分布。

   实际工作中，还可以有其它类型的调查，如个案调查或典型调查，它是在全体研究对象中选取个别研究对象进行的调查。断面调查(cross-sectional study) 是在某一特定时间对调查对象及有关因素进行研究。流行病学研究中还常常用到病例—对照(case-control)研究和队列(cohort)研究(见本章“LOGISTIC 回归”)。

   调查应有计划、有组织、有步骤地进行，调查方案的内容应包括调查目的、调查对象、调查项目、调查表、调查方法、调查人员、调查的组织实施、质量控制与调查进度等。实施时，调查员要经过培训、以保持口径一致，最好进行预调查、复查和数据质量评价等。

   调查数据的分析方法近年来发展很快，其基本思想是考虑这些调查所具有的特征，如基于设计的多阶段抽样调查中的加权，如OSRISIS、SUDAAN、PC CARP ，Stata 5.0 提供许多这一类的方法。考虑数据地域分布的地理信息系统(GIS)方法(如SPLANCS)等。

§2.1.2　实验设计

㈠统计实验设计的原则

1. 对照(control)。为了排除非实验因素的干扰，在进行实验设计时应该设立对照组，并同实验组一样作相同的观察。如观察某种干预措施对儿童缺铁性贫血的影响，可选择一

组儿童给予这种干预，另一组不予干预，间隔一定时间后再对两组儿童的血红蛋白含量进行比较。

2. 随机化(randomization)。是指每一研究单位有均等的机会安排到某个观察组中去，可以消除实验实施时的系统偏差。若无区组，则随机化是随机交换实验的次序以及实际分配因素的水平。出现区组则应对每个区组内的实验进行随机化，然后对区组实验的次序进行随机化。

3. 重复(replication)。指实验组与对照组均要有足够的样本含量，这因为每个观察单位具有变异性，重复观察对于考察响应的平均水平与变动情况是有益的。一种实施办法是在基本的设计中每种情况下的组合做给定的实验数目，另一种是考虑到实验过程须保持一致的环境条件，如温度、湿度，这些条件称做噪声因素。

实验的可靠性是指在同一批对象在不同的时间或等价变量测量时，数据的一致性，也可以指同一总体抽出其它样本时的一致性[12]。

㈡常用实验设计[1]

1. 完全随机化设计(completely randomized design) 是一种最简单的实验设计，没有区组。先根据实验目的选择实验对象，然后用随机化方法将观察单位分别分到实验组和对照组，并给予不同的处理。

2. 随机区组设计(randomized block design) 是将类似的实验对象分到同一组中，称为区组(block)，各区组接受不同的处理，每区组包含的实验单元数为处理数。在每一区组中每个实验对象接受哪一种处理是随机的。配对实验设计是随机区组设计的特例，设计时某些特征相同的两个实验对象配成一对，一个作试验，另一个作对照。在随机区组设计中，若实验的处理数大于每一区组所能容纳的观察单位数时，用平衡不完全区组设计(BIB)，其特点是：所有区组的大小都一样；因子各水平出现次数都一样，每水平在每区组内最多只能出现一次；任一对水平同时出现的次数相同。

3. 交叉实验设计(crossover design) 是把实验对象分为两组，在实验的第一阶段，甲组接受处理，乙组作对照；在实验的第二阶段，乙组接受处理，甲组作为对照。实施时首先将所有对象按某种性状配对，然后随机决定每一对象的试验顺序，经过一个阶段后，处理组与对照组交换。

4. 拉丁方设计(Latin square design) 是使用k 个拉丁字母排成的k x k 方阵的三因素k 水平的设计，在拉丁方中，每个字母在每行每列中只出现一次，实施时先按两因素安排行和列，再按第三个因素的水平随机分到每个拉丁字母。

5. 析因实验设计(factorial experiment design) 是同时检验两种或两种以上因素效应的实验设计。实施方法是首先确定各因素的水平，然后按各因素的各水平的组合确定实验处理，每个受试对象随机地接受处理。其优点是可以了解因素的交互效应。由于随着k 的增大实验次数增加，常常采用部分(fractional) 析因设计。

6. 正交实验设计(orthogonal experimental design) 是在析因设计的基础上发展起来的，其基本思想是用尽可能少的实验次数达到实验的目的。实施时首先确定试验的因素和水平，然后依据正交表安排实验。如$L_4(2)^3$ 表示用4次试验安排一个3个因素，每因素两水平的试验。有时根据需要，可仅使用正交表的部分列来安排试验。

7. 裂区实验设计(split plot design) 是完全随机设计与随机区组设计相结合的设计，同样可做因素和交互效应的研究。在随机区组设计中，只能分析一个处理因素的效应和一个区组效应，若需要分析新的因素但区组对象数无法增加，则应采用裂区实验设计。

8. 嵌套或巢设计(nested design) 是将受试对象成组地分到某试验因素不同水平下的一种设计方法。当试验对象成群出现，不便于对每个对象进行随机化处理，只能成组地随机化分配。

9. 响应曲面设计[6,7] 响应曲面方法(response surface model RSM) 是通过一系列试验来获得响应变量的可靠测量，并且决定一个最合适的模型，最终决定实验因素的最佳配合。响应是测量到的量，其值假设可以通过改变因素的水平而变化。设响应真值可以由因素的某种函数式来表达，对其Taylor 展开式进行一些换算，可把这种关系以多项式的形式表示出来。在几何上，响应与因素的关系式可用超平面表示，也可以等值线图的形式表达。多数响应曲面的研究是一个序贯过程。首先，考虑可能影响响应的因素，然后进行实验，考查这些因素是否真正有影响，再涌现新的想法。

   以下以响应曲面为例进行较详细介绍。

   类似地，分析中的因素是指实验中设定取不同值的变量，一般来说，实验是用因素不同的水平取值来研究感兴趣的响应，直接关心的因素称设计因素，对反应有一定影响但没有直接兴趣的因素称为区组因素。实验的目的之一是避免设计因素与区组因素的混杂。使用正交混杂构造一个设计时，所有因素有相同水平q，q是素数或素数的幂次，一般取值为2，这并不意味着两水平因素的设计不能有两个以上的区组，相反可以用几个两水平的因子来标记两个以上水平的因素，以下的例子是用三个两水平的因素来标记一个8水平的因素。

P_1	P_2	P_3	F
0	0	0	0
0	0	1	1
0	1	0	2
0	1	1	3
1	0	0	4
1	0	1	5
1	1	0	6
1	1	1	7

   因素P_i 仅是用于直接导出因素F 的水平，因而称伪因素(pseudo factors), F 称做导出因素。一般来说，k个q水平的伪因素产生一个q^k 水平的导出因素。区组因素是导出因素，其相关联的P 称作区组伪因素(block pseudo-factor)。在正交混杂设计的构造中，q水平的因素，用q的m 次方个实验，可以区分其前面的m个因素与后面的组合，把前面的m个因素称实验标记因素(run-indexing factors)。设计的分辨(resolution) 能够决定可以独立于其他因素而估计的效应数目。如分辨为5 的设计所有的主效应与两因子交互可以估计。而分辨为4 的实验则某些两因素的交互含有混杂。一般说来，高的分辨需要更大规模的实验。Box 与Draper 列举了响应曲面设计的14 个特性，现列如下：

   1) 在研究区域R 内产生满意的分布。

2) 保证在某点X 处的拟合值$\hat{y}(X)$ 与此处的真值尽量接近。

3) 能方便地看出拟合不当(lack of fit)。

4) 能够进行数据转换。

5) 允许进行区组实验。

6) 允许递增次序的设计能够依次产生。

7) 提供误差的内部估计。

8) 在数据有较大的波动或偏离正态分布假设时不敏感。

9) 只需要最小量的实验单元数。

10) 提供一个简单的数据模式，因而能够地进行一些判断。

11) 计算简便。

12) 当自变量X 发生误差时，设计的行为令人满意。

13) 并不需要自变量不切实际的水平数。

14) 提供一个方差定常(constancy of variance) 假设的检验。

其中较为重要的为1)−3),5)−7),9),11)。

实验正交性(orthogonality) 的含义是拟合的模型的各项相互独立。可旋转性(rotability) 则保证响应的估计仅仅与因素离实验中心的距离有关。响应曲面最基本的问题是估计曲面峰点的坐标，最常用的是一阶设计和二阶设计，其一阶设计模型的形式是：

$$Y = \beta_0 + \beta X + \varepsilon$$

若ε的均值为零，则均值真值为：$\beta_0 + \beta X$

二阶设计包含因素的最高幂次为2，$Y = \beta_0 + X\beta + X'BX$，如：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon$$

这时，

$$B = \begin{pmatrix} \beta_{11} & \beta_{12}/2 \\ \beta_{12}/2 & \beta_{22} \end{pmatrix}$$

经过原点位置的移动和坐标的旋转，上式可表达为：

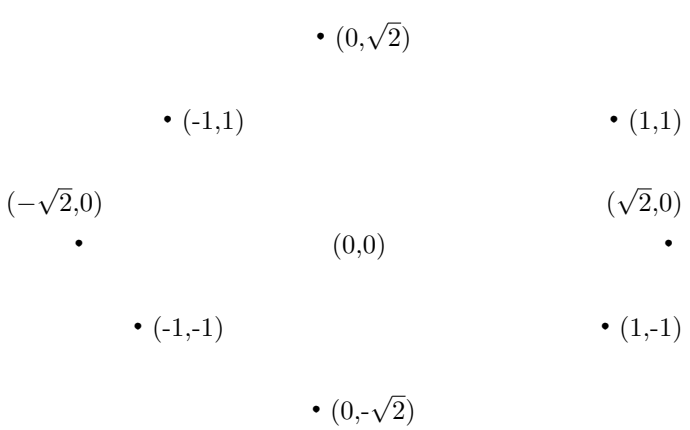
$$y - c_3 = \lambda_1 (x_1 - c_1)^2 + \lambda_2 (x_2 - c_2)^2, \lambda_1, \lambda_2 > 0$$

那么(c_1, c_2, c_3) 即是峰点的一个估计。因为估计二阶函数需要每个因素至少有3个水平，我们可以使用3^k的析因设计来实现，每个因素在-1, 0 和+1 三个水平，k=2 时有九个设计点。当k 增加时，实验的次数猛增，这时可用中心复合设计(central composite design, CCD)来解决，即从通常的析因设计开始，增加2k 个轴点(axial points) 并且在中心点多做几次实验。这种做法的实验次数一般较3^k 要少。

Box-Wilson 设计是一个中心复合设计，由三部分组成：一个完全的或部分的2^k析因设计，因素水平以-1和+1编码，称做析因部分。它有$n_0 \geq 1$ 个中心点，离设计中心α长度的两个轴点，称做设计的轴点。总的设计有$n = 2^k + 2k + n_0$个实验点。如一

个$n_0=1,\alpha=\sqrt{2},k=2$ 的设计为：

$$D=\begin{array}{c}\begin{array}{cc}x_1 & x_2\end{array}\\ \begin{pmatrix}-1&-1\\1&-1\\-1&1\\1&1\\\sqrt{2}&0\\-\sqrt{2}&0\\0&\sqrt{2}\\0&-\sqrt{2}\\0&0\end{pmatrix}\end{array}$$



取$\alpha=\sqrt{2}$设计是可旋转的，因为所有实验点在半径为$\sqrt{2}$ 的球上。在k=3时，α常取为$2^{3/4}=1.682$, $n_0=1$ 时，仅有15次实验，而$3^3=27$次。

Plackett-Burman 引入$n=k+1$ 的k 变量部分2水平析因设计，仅当n 为4 的倍数时可以实现。设计的目的是使设计在此时能以最可能的精度估计所有的主效应。若n 是2 的幂次，并且$n>k+1$，因子间的某些交互影响也能很好地估计。n 是2 的幂次时，Plackett-Burman 设计与标准的部分2水平析因设计等价。它们得到了$n=4,8,12,...,100$ 次实验下$k=3,7,11,...,99$ 个因素的安排方法。构造这种设计，可以这样做：择一由+1 和−1 构成的行，使其+1和−1 的数目分别为$(k+1)/2$ 和$(k-1)/2$，注意这儿由于$k+1$ 是4 的倍数而能够整除。以后列的构造可以从第一列移动一个位置，共移$k-1$ 次。最后，再追加一行皆为−1 的行而得到$n=k+1$ 的设计。现以$n=12,16,20,24$ 为例，设计矩阵的第一行可以是：

n=12 ＋＋－＋＋＋－－－＋－
n=16 ＋＋＋＋－＋－＋＋－－＋－－－
n=20 ＋＋－－＋＋＋＋－＋－＋－－－－－＋＋－
n=24 ＋＋＋＋＋－＋－＋＋－－＋＋－－＋－＋－－－－

要得到完整的设计，循环移动$(n-2)$次，再增加一行符号均为“－”号的行。所有这些设计均是具有复杂alias结构的分解Ⅲ型设计[2]。

Box-Behnken 设计是通过组合2水平析因设计和平衡不完全区组设计(BIB) 而获。如现有一个4 种处理因素6 个区组的BIB，每个处理在实验中出现3 次：

$$\begin{array}{c}\begin{array}{cccc}x_1&x_2&x_3&x_4\end{array}\\ \begin{pmatrix}\star&\star&&\\&&\star&\star\\\star&&&\star\\&\star&\star&\\&\star&&\star\\\star&&\star&\end{pmatrix}\end{array}\quad\begin{array}{c}\begin{array}{cc}x_i&x_j\end{array}\\ \begin{pmatrix}-1&-1\\1&-1\\-1&1\\1&1\end{pmatrix}\end{array}$$

现用右边的2^2 的析因设计去组合，在左边的星号用x_i 代替，右边的星号用x_j代替，没有星号的列填充零，最后再增加几个中点，现增加三个，即可得到一个27个点的设计。一般地，这种设计不是可旋转和区组正交的。

最优设计的准则。线性模型$Y = X\beta_{p\times 1}+\varepsilon$中$\beta$ 的p维可信区域$\beta : (\beta - b)'X'X(\beta - b) \le c$是一个椭圆。$p$个主轴长度的平方是$(X'X)^{-1}$的特征值。一些最有名的选择X的方法就可以从这些值获得。如A、D、E最优。如2^2 设计是正交的，将两因素x_1, x_2 数据进行变换，并取值为± 1，则设计是D最优的。为验证其正交性，记x_1, x_2 两个水平分别是(1,a)、(1,b)，则所有可能的组合是(1,1)，(1,a)，(1,b)，(a,b) ，现记相应的反应观察值是Y_1, Y_a, Y_b, Y_{ab}，关于这些点的响应曲面方程形式为[8]：

$$\begin{pmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

$$\hat{\beta} = (X`X)^{-1}X`Y = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_a \\ Y_b \\ Y_{ab} \end{pmatrix}$$

$\hat{\beta}_0, \hat{\beta}_1$ 可作为x_1, x_2 主效应的估计。

㈢实验研究中的误差控制与样本含量

误差有几种，即：抽样误差、系统误差、随机测量误差和过失误差。减小抽样误差的方法是使样本在性质上和数量上能够对总体具有代表性，样本含量的估计方法可见有关专著。系统误差往往偏向一个方向，从不同程度上干扰研究结果，应尽力加以避免。偶然造成的随机误差可通过重复测量来消除。过失误差不应出现。

本节最后一部分的内容在SAS/STAT ADX宏与SAS/QC涉及较多，Minitab 10也纳入上述实验设计方法。Epi Info 软件提供了流行病学研究估算样本大小的方法。

§2.2 基础统计分析方法

§2.2.1 基础统计方法

假设$X_1, X_2, ..., X_n$是来自分布$F(x;\theta)$的一个样本，其中分布函数$F(x;\theta)$的形式已知，参数θ未知。参数统计分析包括对参数θ的估计和检验。

记样本统计量$T_n(X_1, ..., X_n)$ 为参数θ的函数$g(\theta)$的一个估计量，它的优良性有一些评价标准，常用的有无偏性(unbiasness) 、优效性(efficiency)、相合性(consistency)与不变性(invariance)等。若$E_\theta[T_n] = g(\theta)$对所有$\theta$和$n$成立，则$T_n$是关于$g(\theta)$ 的无偏估计量，E_θ表示关于θ的期望。T_n是无偏估计量且达到Rao-Cram ér 不等式的下界，则为$g(\theta)$的优效估计量。T_n是无偏估计量且使得估计量方差最小，则称最小方差无偏估计量(MVUE)。统计量T_n是$g(\theta)$的(弱)相合估计量，通常是指当n 增大以概率收敛于$g(\theta)$。若一个估计量的任何函数关于θ的期望值为零，则该估计量是充分的。如当任何分布的期望值与方差是有限的，则样本均值$\sum X_{i=1}^n/n$为无偏估计量。样本方差公式

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

右端是$(X_1,...,X_n)$ 的二次型，方阵为$I-\frac{1}{n}(1,...,1)'(1,...,1)$，其秩为$n-1$, 故样本方差以$n-1$为分母，且是总体方差的无偏估计。

显然，“样本均值+某个常量”不相合而$[(n-1)/n]\sigma^2$ 则是相合估计量。

假设θ可以表为总体r阶矩$\alpha_1,...,\alpha_r$的函数$\theta=\theta(\alpha_1,...,\alpha_n),\alpha_i=E[X_1^i]$，记$a_1,...,a_r$为相应的样本矩，$a_i=\Sigma_{j=1}^n X_j^i/n$，则矩法是用$a_i$代替$\alpha_i,i=1,...,r$。因由大数定理，若总体的r 阶矩存在，则样本的r 阶矩是依概率收敛于总体的r 阶矩的，这启发我们用样本矩代替总体矩。设$F(x;\theta)$有概率密度函数$f(x;\theta)$，其似然函数为$\Pi_{i=1}^n f(X_i;\theta)$，为样本的联合分布，使其取值最大的估计称做极大似然估计。如正态分布均值与方差的极大似然估计也是其矩估计。矩估计可能不唯一，不同模型的极大似然函数表达式不同，求解时采用的数值方法也就不同，最常用的是牛顿－拉弗森(Newton-Raphson)方法。其他的参数估计方法有贝叶斯估计及最小二乘估计等。

检验统计量服从正态分布、t分布、χ^2分布、F分布时分别称作z检验(u 检验)、t检验、χ^2检验、F检验。

㈠单变量方法

常用描述分布位置的指标有：均值(算术均值、几何均值、调和均值等)、中位数(M)、众数。将样本变量值求和除以样本例数，即是算术均值(AM)。n 个样本变量值的乘积再开n次方就得到几何均值(GM)，变量的调和均值(HM) 是样本各变量值取倒数后求均值，三种均值间有关系：$HM\le GM\le AM$，几何均值与调和均值常用于描述偏态分布数据。一组变量中，50%分位点的数为中位数，众数是样本中出现频数最多的变量值，即频数分布图上对应峰值的变量值，对称分布如正态分布的算术均值、中位数与众数相同。

表示离散的指标有：方差(VAR)、标准差(STD)、绝对偏差中位数(MAD)、极差(R)、四分位差(IQR)。四分位差是样本75%分位点与25%分位点的差，半个四分位差与标准差大致相当。

设$X_1,...,X_n$ 是来自分布$F(x;\theta)$的样本，把它们按升序排列，并记$X_{(1)},...,X_{(n)}$ 就得到了顺序统计量(order statistic)。$X_{(1)}$与$X_{(n)}$称为极值，$X_{(n)}-X_{(1)}$称为极差或样本全距(sample range)。

上述两类指标之间存在着一定的关系，如正态资料：n 约小于12 ，$STD\approx R/\sqrt{n}$; $20<n<40, STD\approx R/4$; n 在100 左右时$STD\approx R/5; n>400$ 则$STD\approx R/6$; 变异系数是样本标准差和算术均值的比值，反映了样本关于均值变异的大小。

位置的一种稳健估计是把数据偏大和偏小的数据去掉的一定比例，这样就得到了截尾均值(α-trimmmed mean)，如20%的截尾均值是两端均去掉20% 后观测的均值。由于不使用仅仅一个数估计，因而一般要较中位数好，中位数是$0.5-\frac{1}{2n}$ 的截尾均值，近似为50% 的截尾均值。截尾均值可以看成一种加权平均，被去掉的数据权重是0，剩下参与估计的权重为1。同样我们可以对不同的观测构造不同的权，这就是M-估计的思想。

描述分类数据的指标常用率(rate)、构成比(percentage)及比(ratio)等。率是一种相对指标，用于某事件发生的频度，如出生率、发病率等。构成比用于描述具有某种特征的对象占所有对象的数目，如某小学的一个班级男生占56%，女生占44%。比例是两相关事物发生或出现次数的比值，如人口统计中的性比。

一元统计图主要有直方图(histogram)和条图(bar chart)、圆图(pie chart)、茎叶图(stem-and-leaf plot)、箱尾图(Box-and-whisker plot)。直方图和茎叶图、箱尾图常用于定量数据的描述，条图和圆图常用于描述定性数据。

茎叶图做法是依全距把变量分成不重叠的区间，其大小一般取作$k10^p, k = 0.2, 0.5, 1$，p的值可正可负。茎叶图的茎由这些区间构成，每区间的观察构成了叶，在整个数据范围内观察数的变化反映了分布的形状。区间有许多种分法。

箱尾图依数据的中位数画一条竖线，图中方框的位置相当于四分位差的位置，横线(Whisker)延至方框两边的半个四分差，更远的点即为异常值。

在SPSS/PC+中的箱尾图画一个方框代表四分极差，它用星号表示中位数。方框两端发出的须一直延伸到不是异常值的点。方框越大，观测越分散。若一组观测的最大最小值到盒子的两端距离小于一个四分极差，则从方框两端发出的尾延伸到最大和最小的观测值(用X表示)，若大于此距离但小于1.5个四分极差，则用记号O (outlier) 标记这个点，更远的点如超过3个四分极差用记号E(Extreme)表示。

正态分布是最常用的连续性分布，常用$N(\mu, \sigma^2)$表示，其中μ和σ^2分别为正态分布的均值和方差。据正态分布的样本$(X_1, ..., X_n)$ 可以得到均值和方差的极大似然估计$\hat{\mu} = \sum_i x_i/n$和$\hat{\sigma}^2 = (n-1)S^2/n$，$\sqrt{n}(\overline{X} - \mu)/\sigma$ 服从正态分布$N(0,1)$, 在σ未知时$\sqrt{n}(\overline{X} - \mu)/\mathrm{S}$ 服从自由度为$(n-1)$ 的t分布。同时，$(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$。总体方差95%可信区间为$[(n-1)S^2/\chi^2_{0.05;(n-1)}, (n-1)S^2/\chi^2_{0.95;(n-1)}]$ 内。

对于非正态分布数据，据中心极限定理，在样本数较大时，$\sqrt{n}(\overline{X} - \mu)/\sigma$ 服从标准正态分布，其中μ为总体均值，σ^2为总体方差，$\overline{X}$为样本均值。对总体均值μ可进行u 检验或t检验。方差具有正态分布$N(\sigma^2, 2\sigma^4/(n-1))$。因此可以使用检验$\sqrt{n}(S^2 - \sigma^2)/\sqrt{2S^4} \sim N(0,1)$。

离散性变量常用二项分布和泊松分布描述。考虑n次独立试验，每次试验中事件E以概率p发生，用随机变量X表示E发生的试验次数x，概率为：

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \equiv b(x; n, p), x = 0, 1, ..., n$$

$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}, n! = n(n-1)...1, 0! = 1$，期望为$np$，方差为$np(1-p)$。总体率为$\pi, n\pi =$常数$\lambda$时则用泊松分布近似二项分布。函数形式为：

$$P[X = x] = \frac{\lambda^x}{x!} e^{-\lambda}; \lambda > 0, x = 0, 1, 2, ...,$$

其期望与方差均为λ。

在大样本下可以使用近似$(\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1)$。在$p$值近于1 或0 时，$n$ 应至少为100。

【例2.1 】随机抽取某医院400 份病例，有60% 书写合格，则可信区间为：

$$\hat{p} \pm z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n} = 0.60 \pm 1.96\sqrt{(0.60)(0.40)/400} = (0.552, 0.648)$$

检验H: p=0.50 即合格与不合格各占50%

$$z = (\hat{p} - p)/\sqrt{p(1-p)/n} = (0.60 - 0.50)/\sqrt{(0.50)(0.50)/400} = 4.000$$

$p < 0.001$ 对原假设予以拒绝。

对于分组变量，随机变量具有g个类或格子，共观察 n 个对象，格子或分类的频率为$n_i, i = 1, ..., g, n = \Sigma_i n_i$，则描述$n_i$ 的分布为多项分布，n_i 的均值为np_i，方差为$np_i(1-p_i)$，协方差为：$-np_ip_j$。

检验统计量为Pearson χ^2 适合度检验：

$$\sum_{i=1}^{g} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{(g-1)}$$

似然比统计量为：$2\sum_{i=1}^{g} n_i \ln[n_i/np_i] \sim \chi^2_{(g-1)}$

【例2.2 】100 个献血者的血型的分布为：A：35，B：25，AB：5，O：35，问是否符合30:30:10:40的比例？

Pearson χ^2为：

$$(35-30)^2/30 + ... + (35-40)^2/30 = 5.208 < \chi^2_{0.10;3} = 6.251$$

似然比统计量为：2[35ln(35/30)+..+35ln(35/40)]=5.668

当n_i与N_i相比是小值时，分布为超几何分布。具有均值$n[N_i/N]$，方差$n[(N-n)/(N-1)][N_i/N][1-N_i/N]$，协方差$-n[(N-n)/(N-1)][N_i/N][N_j/N]$。

【例2.3 】一种遗传学指标的基因突变率为4/1,000,000，试看25,000 次中小于两个的概率？

这是一个二项分布资料，p很小，使用泊松分布近似

$$P(\leq 1) = P(0;0.1) + p(1;0.1) = e^{-0.1}(0.1)^0/0! + e^{-0.1}(0.1)^1/1! = 0.995321$$

探索性数据分析包括一些统计指标和图形表示，还包括寻找异常点、数据转换、研究模型拟合后的残差。

【例2.4 】表2.1为Hoaglin, D.C.(1983) 描述20 个数据的茎叶图和截尾均值：

表 2.1 20 个分析数据

分析数据		数据的茎叶图	
28	29	-4	4
26	22	-3	
33	24	-2	
24	21	-1	
34	25	-0	2
-44	30	0	
27	23	1	69
16	29	2	1234567899
40	31	3	0134
-2	19	4	0

这里茎的单位为10位数，叶为个位数。截尾均值如下：

$$\begin{aligned}
T_{(0.00)} &= (1/20)\Sigma_{i=1}^{20} x_{(i)} = 435/20 = 21.75 \\
T_{(0.05)} &= (1/18)\Sigma_{i=2}^{19} x_{(i)} = 407/18 = 24.39 \\
T_{(0.10)} &= (1/12)\Sigma_{i=3}^{16} x_{(i)} = 407/16 = 25.44 \\
T_{(0.20)} &= (1/12)\Sigma_{i=5}^{16} x_{(i)} = 308/12 = 25.67 \\
T_{(0.30)} &= (1/8)\Sigma_{i=7}^{14} x_{(i)} = 206/8 = 25.75 \\
T_{(0.40)} &= (1/4)\Sigma_{i=9}^{12} x_{(i)} = 102/4 = 25.50
\end{aligned}$$

频数分布(frequency distribution) 也表示了变量在取值范围内不同区间上绝对数目或相对频率或累积频率的变化。

偏度(skewness) 的测量，常用偏度系数，在箱尾图中Whisker 左右长度表示了偏度。$\gamma = E(X-\mu)^3/\sigma^3$。大的负值常表示左偏，否则为右偏。偏度系数用样本计算时公式为：

$$g = \frac{n\sum_i (x_i - \overline{x})^3}{(n-1)(n-2)S^3}$$

正态大样本时g的均值为0，方差为6/n。

为了去掉偏性，可采用Box-Cox 转换：

$$y = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0; \\ \ln(x) & \lambda = 0 \end{cases}, x > 0$$

Hoaglin, D.C.(1989) 等建议对称性转换的使用方法：a . 在尾部数据不重要时，使用对数转换；b . 尾部的对称较重要时，使用平方根转换；c . 对主要部分分布的偏性和极值的合理偏性间做平衡时，用4 次方根。

有时直接写出其极大似然形式，第 4 章第5节有一个在实验设计中的用例。

对称图(symmetry plot) 是用上下两端的第i 个观察绘图。对称的条件是$X_M - X_{(i)} = X_{(n-i+1)} - X_M$。图的斜率为1。中位数$X_M$ 的位置在数据数目为奇数时为$(n-1)/2$，为偶数时为$n/2$。

峰度(kurtosis) 指标。在对称分布中，指示分布中间部分的频率对分布的形状有意义。有

$$\delta = \frac{E(X-\mu)^4}{\sigma^4} - 3$$

此值为正时分布为突起的(leptokurtic)，否则为扁平的(platykurtic)。样本测量：

$$d = \frac{n(n+1)\sum (x_i - \overline{x})^4}{(n-1)(n-2)(n-3)S^4} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

具渐近分布$N(0, 24/n)$。

去掉峰度的方法通常使用修正的指数转换，其形式类似于Box-Cox 转换：

$$y = SIGN(x - X_M)\frac{(|x - X_M| + 1)^\lambda - 1}{\lambda}$$

$SIGN(.)$ 是符号函数，据函数参量的负值、零和正值分别取值-1, 0 和1。X_M 可以取做均值或中位数。

异常值的检测可以利用$|(x-\overline{x})/s|$并且使用2.70 做界值，其上界为$(n-1)/\sqrt{n}$，在$n < 9$ 时无异常值，修正界值为4 则$n < 18$ 无异常值，在小样本下，利用$X < (Q_1 - 1.5Q)$ 或$X > (Q_3 + 1.5Q)$ 进行比较，$Q = Q_3 - Q_1 \equiv IQR$。由Dixon, W. J.(1950) 的方法是基于顺序统计量，单一的异常值用公式$r_{10} = (X_{(n)} - X_{(n-1)})/R$，R是全距。在n=30时，$p = 0.01, 0.05, 0.10$分别对应$0.341, 0.260, 0.215$，其界值表见Dixon, W.J.(1965). Ratios involving extreme values, Ann. Math. Stat. 22，67-78。

正态性检验，有几种方法，常用的是正态图示、回归方法如Shapiro-Wilk 统计量、Filliben 统计量和D'Agostino 统计量、矩法检查如使用偏度峰度的检验。标准的拟合优度检验是卡方检验、Kolmogorov-Smirnov(K-S) 检验等。K-S 检验使用经验分布函数$F_n(x, \theta) = [X_i \leq x$的数目]/n，$\theta$ 是未知参数。检验统计量

$$D = |\frac{i}{n} - F_{(i)}|, i = 1, 2, ..., n, Z_{(i)} = \frac{(X_{(i)} - \overline{X})}{S}$$

是标化顺序统计量，$F_{(i)} = \Phi(Z_{(i)})$ 的最大值。

图的表示可用Q-Q 图，使用数据分布分位点$x_{(i)}$ 做横坐标，它也是经验分布函数的分位点，$\Phi^{-1}((i-3/8)/(n+0.25))$ 为纵坐标，正态分布时应是一条直线。

Wilks-Shapiro 检验或W-检验. 是基于顺序统计量的方差最优估计量与通常的方差估计的比值，设n个观察排成$x_{(i)} > x_{(i-1)}, i = 2, ..., n$，计算$b = \Sigma_i[x_{(n-i+1)} - x_{(i)}]a_{in}, i = 1, ..., [n/2]$，计算$W_n = b^2/[(n-1)S^2]$，其中S 是$x_i$ 的样本方差，W_n 的百分位点可查表，其值应近于1，否则当W_n 很小时应予拒绝，系数a_{in} 由查表而来，此检验计算较K-S 检验复杂。在样本含量小于50时，W 统计量是Shapiro-Francia W'统计量的良好逼近。

SAS 在样本数小于2000 时计算W-统计量，若样本数大于2000 ，打印Kolmogorov D-统计量，较大值的概率由$(\sqrt{n} - 0.01 + (0.85/\sqrt{n}))D$ 给出，其常用的界值为0.775(0.15), 0.819(0.10), 0.895(0.05), 0.955(0.025) 和1.035(0.01)。

一种简便的情况是在大样本时，考虑偏度与峰度两个指标的正态性，构造检验：$W = n[g/6 + d^2/24]$，它服从自由度为2 的χ^2分布。利用样本均值与方差的独立性，采用刀切法的方法，估计时去掉一个观察，算出均值和方差，每个观察依次轮换，形成了两个新的变量，据数理统计，两个量的相关应为0。因为方差非正态分布，求相关前先对方差开立方根，末了算得的相关采用Fisher 的正态转换。

这里以SPSS/PC+为例，对例2.4的数据计算有关的统计量。其程序为：

```
set length 300.
data list free /x.
begin data.
 28          29        -44          30
 26          22         27          23
 33          24         16          29
 24          21         40          31
 34          25         -2          19
end data.
SET SCREEN OFF.
EXAMINE X /PLOT ALL /MESTIMATOR ALL
          /STATISTICS DESCRIPTIVES EXTREME.
```

其结果如下，按列读取为：均数(21.75)、中位数(25.5)、5%截尾均值(24.39)、标准误(3.94)、方差(310.72)、标准差(17.63)、最小值(-44)、最大值(40)、极差(84)、四分极差(8.5)、偏度(-3.05)及其标准误(.51)、峰度(10.81)及其标准误(.99)。

Mean 21.75	Std Err 3.94	Min -44.00	Skewness -3.05
Median 25.50	Variance 310.72	Max 40.00	S E Skew .51
5% Trim 24.39	Std Dev 17.63	Range 84.00	Kurtosis 10.81
		IQR 8.50	S E Kurt .99

几种M-估计量如下：

Huber (1.34)	25.62	Tukey (4.69)	26.34
Hampel (1.70, 3.40, 8.50)	26.43	Andrew (1.34 * pi)	26.34

设$X_1, ..., X_n$是独立同分布随机变量，通过使$\sum_{i=1}^n \rho(X_i; \theta)$极小而得到参数$\theta$的估计称作M-估计。它包含一大类估计量，如选择$\rho(x_i, \theta) = -\ln f(x_i, \theta)$, $f(x_i, \theta)$为概率密度函数，我们就得

到了极大似然估计。对样本均值有$\rho(x;\theta)=(x-\theta)^2$，由$\min\sum(x-\theta)^2$ 推出$\theta=\sum_i x_i/n$。$\rho(x;\theta)=|x-\theta|$ 的解是样本中位数。R- 估计和L- 估计分别是秩次统计量(rank statistics)和顺序统计量的线性组合或函数，截尾均值($\alpha-$trimmed mean) 就是一种L-估计量。SAS PROC MEANS 中的L_1 是最小绝对偏差(least absolute deviation, LAD)，L_p估计与此相仿。与标准差相应，尺度的估计常用绝对偏差中位数(MAD)来表示，MAD=median$\{|x_i-M|\}$, M=median$\{x_i\}$。这里将以上几种M-估计量解释如下：

Hampel 估计是一种"redescending" M-估计，用三个常数(a,b,c) 来表征。标化观测值绝对值大于c时赋权重为零，0—a 之间的值赋权为1，a—b 和b—c 之间的权随离零的距离而定, 大于c的观测权为0，此处a=1.7，b=3.4，c=8.5。Andrew 估计量也是一种redescending M-估计量。它对于各记录赋的权重没有急剧的变化，而是用一个平滑的正弦曲线来决定各记录的权，标化值绝对值大于c=1.34 π 的记录赋权重为0。Tukey 的biweight 估计量对于标化值大于c=4.685的观测为零，其它权重与离开中心点的距离成反比例。Huber 估计量对标化值小于c=1.339的记录赋权为1，具有较大绝对值记录随离开零的距离增大权重减少。Tukey 的hinges 是每一半数据中点上的值，用于计算盒式图中的四分极差。

本例正态性图示如图 2.1：

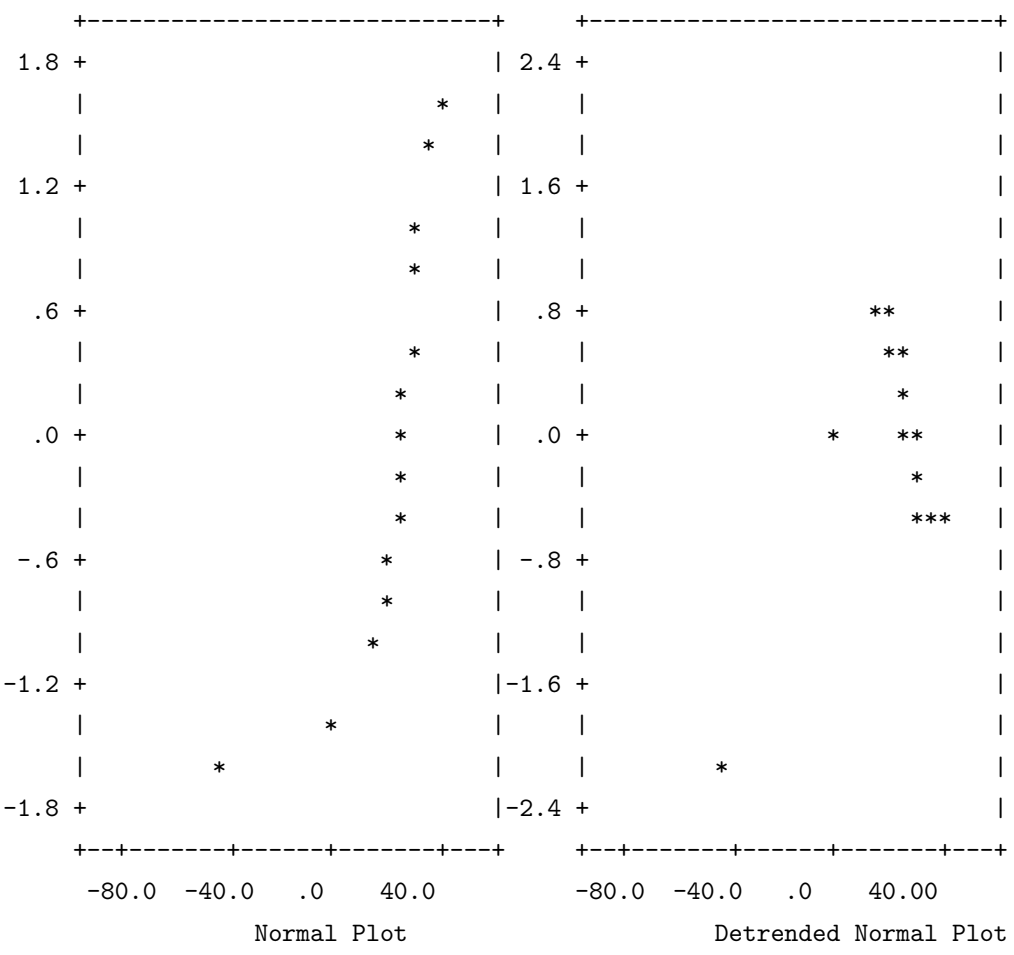


图 2.1 例2.4正态图示

正态性检验统计量：

	Statistic	df	Significance
Shapiro-Wilks	.6460	20	$< .0100$
K-S (Lilliefors)	.2380	20	.0042

箱尾图为图 2.2：

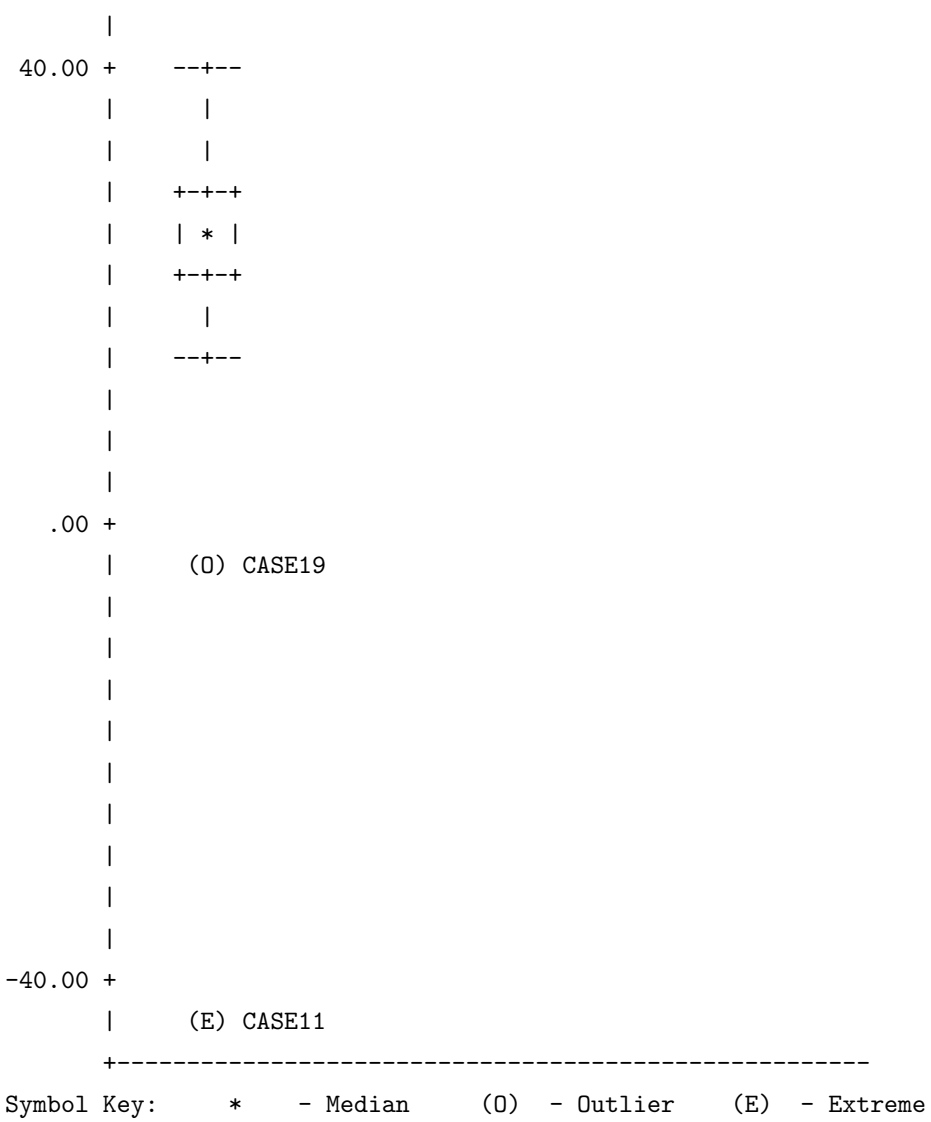


图 2.2 例2.4的箱尾图

正态概率图(P-P 图) 用于检查资料与正态的偏离情况，各点的累积比例与标准正态分布的累积比例绘图。若资料来自于正态分布，则点应近于直线。

去趋势正态图(detrended normal plot) 是观察值为期望值之差而做的图。若样本来自于正态分布，则所有点应聚集在零周围的水平带中，不应该有模式存在。

本例正态图示和检验统计量均提示该数据不符合正态性分布。

进行检验时，所采用的方法与均值与方差是已知还是未知以及观察的数目有关，可用

下面的思路[15]：首先，进行的检验是关于率、均值还是方差？

关于率的检验：是一个率还是两个率？

一个率的检验，$z=(p-\pi)/\sqrt{\pi(1-\pi)/n}$

两个率的检验，使用$\chi^2=\Sigma(o-e)^2/e$

关于方差的检验：是一个还是两个？

一个方差的检验，使用$\chi^2_{(n-1)}=(n-1)S^2/\sigma^2$

两个方差的检验，使用$F_{v_1,v_2}=S_1^2/S_2^2$

关于均值的检验：是一个或是两个？

一个均值检验：样本数≥30？

是，方差已知时使用$z=(\overline{X}-\mu)/[\sigma/\sqrt{n}]$

方差未知时使用样本标准差代替总体标准差用上式进行 z 检验。

否，变量是正态时使用$t_{n-1}=(\overline{X}-\mu)/[S/\sqrt{n}]$

变量非正态时使用非参检验

两个均值检验：样本数≥30？

是，方差已知时使用$z=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$

方差未知时用样本方差代替总体方差继续使用上式。

否，两个变量是正态的吗？

是，方差相等时使用，$t_{n_1+n_2-2}=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S^2/n_1+S^2/n_2}}$

其中$S=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

方差不齐时使用t'检验。

$$t_f=\frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{S^2/n_1+S^2/n_2}$$

其自由度为(Satterthwaite's approximation)：

$$f=\frac{|S_1^2/n_1+S_2^2/n_2|^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$$

否，使用Mann-Whitney-wilcoxon 检验法。

以上方法，在各统计软件包中实现方式不一，如SAS PROC TTEST 提供两种t检验，方差不齐时采用t'检验。

多个样本均值的检验使用方差分析，方差分析中两两比较时有Fisher(LSD)、Duncan、Student-Newman-Keuls(SNK)、Tukey(Honestly Significant Different , HSD) 和Scheffë 法，这些检验的效率越来越低，而一类错误的机会也逐渐减小。Fisher 法逐个控制比较(comparisonwise)的一类误差；Tukey 的HSD 法控制整个比较过程(experimentwise)的 I 类误差，所以进行所有对子的比较，而出错机会是5%；Duncan 法处于两者之间，其界值取决于均值在排列中距离的远近，相邻均值处理同LSD，否则界值增加，但总小于Tukey法；Scheffë 检验最保守，但却可用于其它方式的比较。如：$H:\mu_1=(\mu_2+\mu_3)/2$ 可以等价地写成，$\mu_1-(\mu_2+\mu_3)/2=0$，检验误差为$1+(-1/2)^2+(-1/2)^2=3/2$，在SAS 中使用CONTRAST 语句进行此类检验，SAS 共有十七种比较的方法而在SPSS 中有七种。比较方法的选用取决于具体问题，如在条件较严格的物理实验中常用Fisher 法而在一过性事件和社会科学中Tukey 方法反而常用。

在SPSS/PC+中的Tukey-B 方法也是一种两两比较方法，把均值由小到大排列，然后对排列中每种比较求得一个距离，该法使用Tukey HSD 和SNK 的均值计算每步差的取值范围。

对分类数据分析，若仅对一个分类变量，则数据通常用频数表表示，它列出变量的取值和发生频数；若有两个或以上分类变量，一个对象的profile 定义为它在各分类上的取值，这样的数据可以把对象的profile 连同其频数一起列成频数表；若恰好有两个分类变量，则常用两维列联表(contingency table)的形式，其行列由每个分类变量不同取值构成。行列的交叉成为格点(cell) ，格点记录了相应profile的频数；对于多个分类变量，用多维列联表表示，在SAS FREQ和CATMOD 用不同的方法表示。

㈡两变量方法

1. 两变量的图示常用散点图、分组的箱尾图、茎叶图(back-to-back stem- and-leaf plot)等。

为了检验两个变量是否正态分布，可采用以下统计量：

$$\begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}' \begin{pmatrix} S_2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}^{-1} \begin{pmatrix} X-\overline{X} \\ Y-\overline{Y} \end{pmatrix}$$

应服从$\chi^2(2)$分布。

分类数据中，双向分类数据的χ^2检验假设：①有N次相同的实验；②每次试验有k种可能的结果；③K 个结果的概率保持不变；④实验是独立的；⑤K个格子的预计反应数应至少为5。有关列联表稳健性的讨论可见Hoaglin, D.C .(1983)，多维列联表使用对数线性模型来处理，见第13章。

一个行数为R 列数为C 的$R\times C$ 列联表独立性使用Pearson χ^2来检验，表示$R\times C$列联表行列因素相关的程度有许多统计量，其中之一是ϕ，其公式是：$\phi=\sqrt{\chi^2/N}$，其下界为0，当观测值与期望值相同时为0，而其上界是$\sqrt{\min(r-1,c-1)}$，两行或两列时，上界为1，故最常用。对于观察格子为A，B，C，D时，四格表$\phi=\frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$。

Cramer's $V=\frac{\phi}{\sqrt{\min(r-1,c-1)}}$，其范围是$0\sim 1$。

列联表系数$CC=\sqrt{\frac{\chi^2}{\chi^2+N}}$，$\chi^2$值最小时该量最大，上界随表的增大而趋于1。

现在看一个吸烟与与肺癌关系的的例子。x:1=不吸烟，0=吸烟；Y:1=死于其它疾病，0=死于肺癌。原始数据如下：

X：0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1

Y：1 1 0 0 0 0 0 0 0 1 1 0 1 1 1 1 0 1 1 1 0

可以算得其积矩(Pearson)相关系数r 和列联表系数 ϕ 为0.302。按列联表形式算得$\chi^2=1.818$。显然，χ^2是x与y两变量关系的显著性而列联表系数是其大小。$\phi=\sqrt{\chi^2/N}$，但只用于2x2表，对于更大的表使用V 统计量。SAS 程序如下：

```
data phi;
input x y @@;
cards;
0 1   0 0   1 0   1 0
0 1   0 0   1 1   1 1
0 0   0 0   1 1   1 1
0 0   0 1   1 1   1 1
0 0   0 1   1 1   1 0
proc print;
proc corr;
```

```
run;
proc freq;
   table x*y/chisq expected nopercent nocol norow;
run;
```

```
 X        Y
 Frequency|
 Expected |       0|       1|  Total
 ---------+--------+--------+
0 |      6 |      4 |     10
  |    4.5 |    5.5 |
 ---------+--------+--------+
1 |      3 |      7 |     10
  |    4.5 |    5.5 |
 ---------+--------+--------+
 Total      9      11       20

STATISTICS FOR TABLE OF X BY Y

Statistic   DF  Value       Prob
------------------------------------------------------
Chi-Square                          1     1.818       0.178
Likelihood Ratio Chi-Square         1     1.848       0.174
Continuity Adj. Chi-Square          1     0.808       0.369
Mantel-Haenszel Chi-Square          1     1.727       0.189
Fisher's Exact Test (Left)                              0.965
(Right)                0.185
(2-Tail)                              0.370
Phi Coefficient                    0.302
Contingency Coefficient            0.289
Cramer's V                         0.302

Sample Size = 20
WARNING:  50% of the cells have expected counts less
          than 5. Chi-Square may not be a valid test.
```

两率比较的功效：

设$p_A = p_B \approx 60\%$，期望$p_A - p_B = 6\%$，试问样本量不同取值下的检验功效？

$p_A - p_B$的方差为$\frac{\pi_A(1-\pi_A)}{(n/2)} + \frac{\pi_B(1-\pi_B)}{(n/2)} \approx 0.6 \times 0.4 \times (4/n) = 0.96/n$

$$Z = \frac{(p_A - p_B) - (\pi_A - \pi_B)}{\sqrt{0.96/n}} \sim N(0,1)$$

当$p_A - p_B = 0.06, H_0 : \pi_A - \pi_B = 0$　功效为：

$$p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} \geq z_{\alpha/2}\right]$$

$$\alpha = 0.05, Z_{0.025} = 1.96$$

有

$$p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} > 1.96 - 0.06\sqrt{n/0.96}\right] + p\left[\frac{|p_A - p_B| - 0.06}{\sqrt{0.96/n}} < -1.96 - 0.06\sqrt{n/0.96}\right]$$

$$\begin{aligned} &= P[Z > 1.96 - 0.06\sqrt{n/0.96}] + p[z < -1.96 - 0.06\sqrt{n/0.96}] \\ &= \Phi[1.96 - 0.06\sqrt{n/0.96}] + \Phi[-1.96 - 0.06\sqrt{n/0.96}] \end{aligned}$$

当$n = 50, \approx 0.07$; $n = 200, \approx 0.14$。

作为列联表探索性数据分析的用例，这里给出一个中位数平滑的例子(V. A. Sposito, On Median Polish and L1 Estimators, Comp. Stat. and Data Anal., V5. N3.，1989)。

列联表为$\begin{pmatrix} 1 & 8 & 3 \\ 5 & 9 & 2 \\ 6 & 4 & 7 \end{pmatrix}$　每行中位数$3, 5, 6$

以行开始进行第一个半步，每行减去行中位数，结果为：

```
 -2   5   0  | 3          -2    1     0  | -2
  0   4  -3  | 5           0    0    -3  |  0
  0  -2   1  | 6           0   -6     1  |  1
 ------------+---        ----------------+---
  0   4   0  |            0    4     0  |  5
```

5 是前一行效应估计的中位数，由于第二半步时每行、列的中位数为0，过程停止，得：行效应$\overline{\alpha} = (-2, 0, 1)$，列效应$\overline{\beta} = (0, 4, 0)$。对应于模型：

$$Y_{ij} = \mu_i + \beta_j + \varepsilon_{ij}\ \varepsilon_{ij} = \begin{pmatrix} -2 & 1 & 0 \\ 0 & 0 & -3 \\ 0 & -6 & 1 \end{pmatrix}$$

软件包Statgraphics 能够进行上述计算。

2. 两变量分析最常用的手段是变量的线性相关。设随机向量(ξ, η) 服从二维正态分布，参数为$(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$，其中$\rho$ 是ξ 和η 的相关系数，现在$(X_1, X_1), ..., (X_n, Y_n)$是来自$(\xi, \eta)$ 的简单随机样本，相关系数的公式是：

$$r = \frac{\sum_{i=1}^n (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum_{i=1}^n (X - \overline{X})^2 \sum_{i=1}^n (Y - \overline{Y})^2}}$$

据Schwartz 不等式，变量的相关系数应在$[-1, 1]$ 内。相关系数是否有显著性意义，可利用r-检验，查其界值表判断其显著性；或者使用t-检验，其公式是$t = r\sqrt{(n-2)/(1-r^2)}$；最后，还可以使用Fisher 的z-转换进行检验。根据相关为组内相关或组间相关的不同，对子数的校正方法也不相同。

在线性回归分析中，应该注意的是回归诊断技术。其主要内容为：

误差项是否满足独立性、等方差性、正态性。

选择线性模型是否合适，是否存在曲线等关系？

是否有异常样品存在，即异常点？

回归模型是否过多依赖于某些样品，即模型稳健性如何？

自变量之间是否高度相关，即多重共线性？

【例2.5】表 2.2是Anscomb构造的数据(Anscomb, F.J., 1973, Graphs in statistical analysis. Am. Statist.,27,17-21)，前四列可分为三组，第一列是共用的自变量x，最后两列是第四组的x,y。

表 2.2 Anscomb(1973)设计的四个数据例子

x1	y1	y2	y3	x4	y4
10.0	8.04	9.14	7.46	8.0	6.58
8.0	6.95	8.14	6.77	8.0	5.76
13.0	7.58	8.74	12.74	8.0	7.71
9.0	8.81	8.77	7.11	8.0	8.84
11.0	8.33	9.26	7.81	8.0	8.47
14.0	9.86	8.10	8.84	8.0	7.04
6.0	7.24	6.13	6.08	8.0	5.25
4.0	4.26	3.10	5.39	19.0	12.50
12.0	10.84	9.13	8.15	8.0	5.56
7.0	4.82	7.26	6.42	8.0	7.91
5.0	5.68	4.74	5.73	8.0	6.89

相应的SAS回归程序为：

```
data anscomb;
input x1 y1 y2 y3 x4 y4;
cards;
......
proc reg data=anscomb;
L1:model y1-y3=x1;
L4:model y4=x4;
run;
```

结果算得四组回归结果是相同的, y=3+0.5x, $R^2 = 0.667$，可见通常的检验方法对四组数据的区分是无能为力的，而由残差图可清楚地看出残差所隐含的模式。

从SAS、SPSS/PC+、Stata 等可以获得回归诊断统计量。如曲线关系利用标化残差与x的点图来发现；方差不一致利用标化残差对x 或y 预测值点图的方法；残差间的相关，使用Durbin-Watson 统计量。这些诊断统计量在线性回归有最简单的形式，现将模型$y = \beta_0 + \beta_1 x + \varepsilon$有关的诊断量列如下，

1. 杠杆(leverage)：

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{\sum (x_j - \overline{x})^2}$$

可见x离均值较远时h_{ij}就大，提示一个距离的解释。

$$\hat{y}_i = b_0 + b_1 x_i = \sum h_{ij} y_j$$

它反映了对y估计值的影响，$\frac{(n-2)(h_{ii}-1/n)}{1-h_{ii}}$ 服从$t(n-2)$ 分布，因而实算值大于界值时可认为是高杠杆点。

2. 残差(residual)：普通残差为$e_i = y_i - \hat{y}_i$，删除残差为

$$e_{(-i)i} = y_i - \hat{y}_{(-i)i} = \frac{e_i}{1 - h_{ii}}$$

反映了预测值与实际值的差，故也称预测残差，下标$-i$ 表示去掉第i 个观察后的结果。

预测残差平方和$PRESS = \sum e^2_{(-i)i}$ 与$\sum e^2$ 的比反映了缺失观察的影响，可用于变量的筛选。

标准化残差和学生化残差(studentized residual)：由于$E(e_i) = 0, V(e_i) = \sigma^2(1 - h_{ii})$，直接比较残差是不合适的，用标准化残差$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$。记其最大的一个为$R_n = MAX|r_i|$，其上界值为：

$$\frac{(n-2)F_{1,(n-2)}(\alpha)}{n - 3 + F_{1,(n-2)}(\alpha)}$$

另一种学生化残差是$t_i = \frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}$

$$s^2_{(-i)} = \frac{(n-2)s^2 - \frac{e_i^2}{1-h_{ii}}}{n-3}$$

因s^2并不与e_i独立，(x_i, y_i) 的统计量就可以写成下式：

$$t = \frac{e_i\sqrt{(n-3)}}{\sqrt{(n-2)s^2(1-h_{ii}) - e_i^2}}$$

应注意的是，小的值仍可以是高影响点，用相同的界值进行所有的比较时就有可能使Ⅰ类误差增大，考虑用Bonferroni 修正。要保证所有单侧检验水平是α，则界值是α，对于双侧检验则是$\alpha/2n$。

3. 库克距离(Cooks' D)：

$$D_i = \sum \frac{(\hat{y}_{(-i)j} - \hat{y}_j)^2}{2s^2}, i, j = 1, ..., n$$

或：$D_i = \frac{r_i^2 h_{ii}}{2(1-h_{ii})}, i = 1, ..., n$ 可见其受r_i与h_{ii} 的影响。

Belsey, Kuh 和Welsch (1980) 提出了DF 簇方法：

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(-i)i}}{s_{(-i)}/\sqrt{h_{ii}}} = \sqrt{\frac{h_{ii}}{(1-h_{ii}}}\frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}$$

可见它与D_i是可比的，仅仅以$s_{(-i)}$代替了$\sqrt{2}s$。与此相仿，有$DFBETA_{1i} = \frac{b_1 - b_{(-i)1}}{C_1 s_{(-i)}}, i = 1, ..., n, C_1^2 = \frac{n}{n\sum x^2 - (\sum x)^2}$，是$b_1$ 的方差除以σ^2 。$DFBETA_{0i}$ 亦与此相仿。Welsch 建议用加权最小二乘求取回归系数，使用的权w_i当$DFFITS_i \leq 0.34$ 时为1，否则为$0.34/|DFFITS_i|$。

考虑观察值对于方差协方差阵的影响，原来估计值与估计值的比值COVRATIO 为：

$$\frac{n-1}{(n-2)+t_i}(1-h_{ii})$$

在$1/n \le h_{ii} \le 2/n$ 及$|t_i| \le 2$ 时在$[1-3/n, 1+3/n]$ 内。否则当$|COVRATIO-1| > 3/n$ 时，(x_i, y_i) 点应予以研究。另外，$h_{ii} \to 0$ 时，COVRATIO 比值近于1，较大较负的t_i 可导致更大的COVRATIO。

4. 自相关的检验(Durbin-Watson test)

检验统计量

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

在n比较大时近似有：

$$\sum_{i=1}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2 \approx \sum_{i=2}^{n} e_i^2$$

故$d \approx 2-2r = 2(1-r)$, r 是一阶自回归模型AR(1) 的参数，AR(1) 的模型是：

$$e_i = \rho e_{i-1} + u_i, u_i \sim N(0, \sigma^2), V(e) = \frac{\sigma^2}{1-\rho^2}$$

$\rho_s = \rho^s$ 是自相关系数，d可能的取值为(0,4)，上述结论便于记忆相关的符号。

$H: \rho = 0 \leftrightarrow A: \rho > 0$, 在$d < d_L^\star$ 时拒绝，在$d > d_U^\star$ 时接受；

$H: \rho = 0 \leftrightarrow A: \rho < 0$, 在$d > 4 - d_L^\star$ 时拒绝，在$d < 4 - d_U^\star$时接受。

一般地，由于Box 和Jenkins 自相关和移动平均模型的形式是：

$$ARMA(p,q): e_t = \phi_1 e_{t-1} + ... + \phi_p e_{t-p} + v_t + \theta_1 v_{t-1} + ... + \theta_q v_{t-q}$$

它由两部分组成：

自相关AR(p)：$e_t = \phi_1 e_{t-1} + ... + \phi_p e_{t-p}$

移动平均MA(q)：$e_t = v_t + \theta_1 v_{t-1} + ... + \theta_q v_{t-q}$

它们属于时间序列分析的内容，SAS/ETS、SPSS/PC+ Trend、SYSTAT Series 及TSP 软件均可进行时间序列分析。

序列间的相关还可用游程检验(Wald-Wolfowitz) 来进行。检验是针对按一定顺序排列的二分类变量，连续出现一种结果(符号)为一个游程，全部游程数记为r。记出现正号的数目为$n_1, n_2, n = n_1 + n_2$，据界值表，r 过大或过少均表示符号的变化是不随机的。

$$\mu = \frac{2n_1 n_2}{n} + 1, \quad \sigma^2 = \frac{2(n_1 n_2)(2n_1 n_2 - n)}{n^2(n-1)}$$

$$z = \frac{|r-\mu| - 0.5}{\sigma}$$

游程检验可由SAS/QC 的SHEWHART 过程完成；在SYSTAT 中，命令RUNS 计算游程检验。

多数回归诊断技术提供的最好方法还是删除观察量，较好的方法是采用稳健回归的方法。这些方法的研究近年来很活跃。除此以外，非线性回归与相关、校准(calibration) 方法亦经常采用。

§2.2.2 非参统计方法

属于单样本方法有符号检验(sign test)、Wilcoxon 符号秩次检验、Kendall 检验、Spearman 检验、K-S 检验，两样本方法有Wilcoxon-Mann-Whitney 检验、Kolmogorov-Smirnov 检验，多样本检验用Kruskal-Wallis等方法[36]。

符号检验是最简单的一种，它只考虑符号，而Wilcoxon 符号秩次检验考虑了秩的绝对值大小Spearman检验与Kendall检验是关于相关系数的。K-S检验可用于分布的拟合，与χ^2检验的区别是它仅适于连续型分布，其精确的界值是已知的，对于任意样本大小均更有效。

Wald-Wolfowitz 游程检验也是一种非参检验，其思想是：两个样本来自于同一个总体，则将两样本的观测值混合按小到大排列，用一个游程表示同一组数据的数列，如果游程数很少，表明两样本来自不同的总体。

这些检验结果的判断一般据查表和正态分布、χ^2分布近似方法，软件包多基于近似方法，此处作主要介绍。

㈠ Mann-Whitney 秩和检验

有时称Mann-Whitney U 或Wilcoxon Rank Sum W Test，是一种非参检验，检验两个容量分别为m和n的独立样本是否来自同一总体，它使用了观察的秩次，故较中位数检验(median test)更为有效。

H:总体1与总体2的相对频数分布是相同的。

A:总体1的频数分布相对于总体2向左右移动。

检验统计量为U，在$U \leq U_c$时拒绝原假设。

做法是先把$m+n$ 个观察自小到大排序，每个样本的观察值序号之和称为秩和，记为T_1与T_2。编秩时相同的秩次取其平均值。小样本时两者较小一个$\geq T_U$ 或$\leq T_L$ 时拒绝，T_U 和T_L可由专门的表查出；大样本时，m 与n 的值均应至少为10，可用Z-检验。

【例2.6】测得铅作业与非铅作业两组工人的血铅值($\gamma/100g$) 如表2.3，判断两组工人血铅值之间是否存在差别？

秩和(期望值)为：T_2=59.5 (63.0), T_1=93.5 (90.0)。

H: 铅作业工人与非铅作业工人血铅的分布是相同的

A: 两组工人血铅分布是不同的。

计算统计量为：$U = nm + \frac{m(m+1)}{2} - T_1 = (10)(7) + \frac{7(8)}{2} - 93.5 = 4.5$

U 的期望值是mn/2，方差是nm(n+m+1)/12，所以采用正态近似，

$$z = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}} = \frac{4.5 - (10)(7)/2}{(10)(7)(10+7+1)/12} = -2.97$$

计算结果亦列于上表。Wilcoxon 秩和检验与Mann-Whitney 检验是等价的，有时又称Mann-Whitney-Wilcoxon 检验。结果可由SAS PROC NPAR1WAY 指定Wilcoxon 选项而得，检验又是Kruskal-Wallis 两组时的情况，χ^2=8.8813, 自由度1, P=0.0029。使用SPSS/PC+，结果类似，校正相同秩次后，Z值为-2.9801, P=0.0029，均表明两组工人血铅值存在显著差异。

㈡Wilcoxon 符号秩次检验: 配对检验。

即Wilcoxon matched-pairs signed-ranks test，是一种两样本非参方法，检验两个变量的分布是否相同。它不对分布的形状作任何假设，算出变量差值的绝对值，并由小到大排列，检验根据正差和负差之和进行。

H: 总体1与总体2的相对频数分布相同。

A: 总体1的频数分布相对于总体2向左右移动。

表 2.3 两组工人血铅值的秩和检验

工人号	组号	血铅	秩次
1	1	5	1.5
2	1	5	1.5
3	1	6	3
4	1	7	4
5	1	9	5
6	1	12	6
7	1	13	7
8	1	15	8
9	1	18	10.5
10	1	21	13
11	2	17	9
12	2	18	10.5
13	2	20	12
14	2	25	14
15	2	34	15
16	2	43	16
17	2	44	17

差值为零者忽略不计，使用差为负的秩和或差为正的秩和，在秩和少于界值时拒绝原假设。在组数足够大如$N \geq 25$时采用正态逼近的办法。

【例2.7 】用两种饲料喂8只大白鼠后，测定其肝中维生素A的含量(国际单位/mg)，正常组(normal)与维生素缺乏组(deff)结果见表2.4，问不同饲料的效果有无差别？

大样本的算式如下：

$$Z = \frac{T_{-} - [n(n+1)/4]}{\sqrt{[n(n+1)(2n+1)/24]}} = \frac{1-(8)(9)/4}{\sqrt{(8)(9)(17)/24}} = -4.20$$

在H 下，$Z > Z.05 = 1.96$，故应拒绝原假设。

应用SPSS/PC+中语句NPAR TESTS /WILCOXON normal deff，有Z = -2.3805，P =.0173。使用语句T-TEST /PAIRS normal deff. 进行配对t 检验，均值分别为3.3188和2.5063，差的均值为0.8123，标准误为0.193，t=4.21，P=0.004 。在SAS中可以使用PROC MEANS进行这个检验，语句为proc means t prt var std stderr; var d;run;。

㈢Kruskal-Wallis H 检验: 比较K 个总体相对频数分布。

H:K个总体的相对频数分布是相同的

A:至少有两个总体的相对频数分布是不同的

检验统计量：

$$H = \frac{12}{n(n+1)} \Sigma_i \frac{T_i^2}{n_i} - 3(n+1)$$

其中n_i= 第i个样本的观察数, T_i=第i个样本的秩和, n=Σn_i为总的样本数目

表 2.4 配对资料符号秩和检验两种鼠肝中维生素A含量

OBS	正常饲料	维生素E缺乏组	差值(d)	秩次
1	3.55	2.45	1.10	6
2	2.00	2.40	-0.40	-1
3	3.00	1.80	1.20	7
4	3.95	3.20	0.75	3
5	3.80	3.25	0.55	2
6	3.75	2.70	1.05	5
7	3.45	2.50	0.95	4
8	3.05	1.75	1.30	8

$T_+ = 35, T_- = 1$。由界值表，$\alpha < 0.05$，应拒绝原假设。

假设: 1. K 个样本独立并随机地由各自的总体抽出。2. 为使卡方逼近较为稳妥，每个样本应至少有 5 个或更多的观察。3. 秩次重复时，它们的秩次排列好象它们没有重复时的秩次求和取平均而得。

【例2.8 】研究社会经济状况与在校成绩的关系，将社会经济状况分为三等。看入校新生绩点成绩间的差别。原始数据如表2.5：

表 2.5 社会经济状况与在校成绩的关系

下　等	中　等	上　等
2.87 (10)	3.23 (16)	2.25 (5)
2.16 (3.5)	3.45 (18)	3.13 (14)
3.14 (15)	2.76 (8)	2.44 (6)
2.51 (7)	3.77 (20)	3.27 (17)
1.80 (2)	2.97 (11)	2.81 (9)
3.01 (12.5)	3.53 (19)	1.36 (1)
2.16 (3.5)	3.01 (12.5)	
$T_1 = 53.5$	$T_2 = 104.5$	$T_3 = 52$

$H = 6.13 > \chi^2_{.05;(3-1)} = \chi^2_{2;0.05} = 5.99, P < 0.05$ 拒绝原假设。

㈣ Friedman 检验: 随机区组设计

H: K个总体的相对频数分布是相同的

A: 至少有两个总体的相对频数分布是不同的

检验统计量：

$$F = \frac{12}{bk(k+1)} \Sigma_i \frac{T^2}{n_i} - 3b(k+1)$$

其中b=实验中使用的区组数, k=处理数, T_i=第i种处理的秩和。在$F > \chi^2_{\alpha;(k-1)}$界值时拒绝原假设。

假设：1. K个处理随机分配给每个区组内的K个实验单元。2. 为保证χ^2分布的适度，区组数或处理数均应超过 5 。3. 秩次重复时，它们的秩次排列好象它们没有重复时的秩次求

和取平均而得。

【例2.9 】一个食品公司进行了一个单盲试验，让 6 个受试者随机地品尝三种不同的咖啡 A 、B 、C ，并把三种咖啡的优劣排序，得表2.6结果：

表 2.6 6 个受试者评判三种咖啡的结果

受试者	A	B	C
1	1	3	2
2	2	3	1
3	1	2	3
4	1	3	2
5	2	3	1
6	1	3	2
	$T_a = 8$	$T_b = 17$	$T_c = 11$

$$F = \frac{12}{(6)(3)(3+1)}[(8)^2 + (17)^2 + (11)^2] - 3(6)(3+1) = 7.0 > 5.99, P < 0.05$$

利用SPSS 和Minitab 可以进行这个检验，SAS 的示范程序中含有这个检验的例子。

㈤ Spearman 和Kindall 秩和相关检验

Spearman 相关是利用变量对间秩次的相关来说明两个变量的关系，由于编秩时信息的损失，它仅是一个单调性检验，并不是一个真正的线性关联。计算时分别对两个变量进行排秩，把编成的秩次仿Pearson 积矩相关进行计算，就得到Spearman 相关。在相同秩次较多时，建议用秩次代入积矩相关的公式中进行计算。利用两各观察秩次的差(d) 时，可用下面的公式，在相同秩次较多时应予校正。

H: 样本X与样Y是独立的

A: 较大的X趋于同较大的Y配对

记X与Y秩次为R_i, T_i，其均值皆为$\frac{n+1}{2}$, 检验统计量为：

$$r_s = 1 - \frac{2(2+1)}{n-1} + \frac{2}{n(n^2-1)}\Sigma_i R_i T_i$$

也用下面的形式：检验统计量: $r_s = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)}$，大样本时$r_s\sqrt{n-1}$ 服从标准正态分布

肯德尔τ系数$\tau = 2\frac{\sum_{i<j} i_{ij}}{n(n-1)}$

$i_{ij} = SIGN[(x_i - x_j)(y_i - y_j)], SIGN(.)$ 是符号函数。对于(x_i, y_i)和(x_j, y_j)，若当$x_i > x_j$时$y_i > y_j$，则是一致的(concordance)，否则是不一致的(discordance)。肯德尔相关系数反映了数据这种一致与不一致的情况，该统计量在观测重复数较大时除以下式来校正：

$$(\frac{n(n-1)}{2} - n_x)^{0.5}(\frac{n(n-1)}{2} - n_y)^{0.5}$$

其中n_x与n_y是x与y的重复数。

大样本时用$3\tau\sqrt{\frac{n(n-1)}{2(2n+5)}} \sim N(0,1)$

【例2.10 】肝癌病因研究中，某地调查十个乡的肝癌死亡率(1/20万)与某食物中黄曲霉毒素相对含量的关系，数据列于表2.7，试分析两者是否存在相关？

表 2.7 黄曲霉素相对含量与肝癌死亡率

乡编号	黄曲霉素相对含量		肝癌死亡率		
	X	d	Y	d	秩次差
1	3.7	4	46.6	7	-3
2	1.0	2	18.9	2	0
3	1.7	3	14.4	1	2
4	0.7	1	21.5	3	-2
5	4.0	5	27.3	4	1
6	5.1	6	64.6	9	-3
7	5.5	7	46.3	6	1
8	5.7	8	34.2	5	3
9	5.9	9	77.6	10	-1
10	10.0	10	55.1	8	2

Spearman 相关分析结果

检验统计量:

$$r_s = 1 - \frac{6\Sigma_i d_i^2}{n(n^2-1)} = 1 - \frac{6(42)}{10(10^2-1)} = 0.7545$$

P=0.0133, 相关有显著意义。

以上几个非参检验在软件包中实现很方便，例2.10的SAS程序为：

```
data list;
   input x y ;
cards;
 3.7   46.6
 1.0   18.9
 1.7   14.4
 0.7   21.5
 4.0   27.3
 5.1   64.6
 5.5   46.3
 5.7   34.2
 5.9   77.6
 10.0  55.1
proc corr pearson spearman kendall nosimple;
   var x y;
run;
```

结果如下：Pearson 相关系数0.69754, P=0.0249，Spearman 相关系数0.74545, P=0.0133, Kendall Tau b 0.51111, P=0.0397。

也可以利用PROC RANK进行变量x、y的排序，生成的排序变量直接用于Pearson 相关公式得出Spearman相关系数。

§2.3 多元分析

§2.3.1 均值的检验

1. 单样本检验

设$X_1, ..., X_n$ 是独立同分布、来自p维正态分布$N_p(\mu, \Sigma)$的样本，其中均值向量μ和协方差矩阵Σ未知，现要检验假设H: $\mu = \mu_0$，A:$\mu \neq \mu_0$，

检验统计量为$T^2 = n(\overline{X} - \mu_0)' S^{-1}(\overline{X} - \mu_0)$ ，其中样本均值向量$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, 样本协方差矩阵$S = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})'}{n-1}$。在假设H成立的条件下，$\frac{n-p}{p(n-1)} T^2 \sim F_{p,n-p}$。因为在显著水平为$\alpha$时，$\frac{n-p}{p(n-1)} T^2 \geq F_{p,n-p;\alpha}$则拒绝假设H，其中$F_{p,n-p;\alpha}$表示自由度为$p, n-p$的F分布的右侧分位点。

【例2.11】8 个人服用某种药物后，将其血糖和血压记录于表 2.8，研究是否由于药物作用造成的特性改变为某一定量零。

表 2.8 八个人对某种药物的反应结果[21]

编号	1	2	3	4	5	6	7	8
血糖	30	90	-10	10	30	60	0	40
收缩压	-8	7	-2	0	-2	0	-2	1
舒张压	-1	6	4	2	5	3	4	2

现在，样本均值向量$\overline{X} = \begin{pmatrix} 31.25 \\ -0.75 \\ 3.125 \end{pmatrix}$ 方差矩阵$S = \begin{pmatrix} 1069.64 & 82.5 & 16.964 \\ 82.5 & 17.357 & 6.393 \\ 16.964 & 6.393 & 4.694 \end{pmatrix}$

检验$H: \mu = 0$ 对$A: \mu \neq 0$

$$T^2 = n\overline{X}' S^{-1} \overline{X} = 79.064$$

$\frac{n-p}{p(n-1)} T^2 = [\frac{5}{3(7)} 79.064 = 18.825 > F_{3,5;0.05} = 5.41$

则拒绝H,即药物作用造成的改变量不为零。

总体均值μ可用$\overline{X}$来估计。相应于单变量的区间估计，均值向量μ的$100(1-\alpha)\%$可信区域为：

$$\{\mu : n(\overline{X} - \mu)' S^{-1}(\overline{X} - \mu) \leq \frac{(n-1)p}{n-p} F_{p,n-p;\alpha}\}$$

现仅就本例中的收缩压和舒张压改变来考虑，95%可信区域为：

$$0.116(\mu_1 + 0.75)^2 - 0.314(\mu_1 + 0.75)(\mu_2 - 3.125) + 0.427(\mu_2 - 3.125)^2 \leq 1.50$$

在多变量时，也可对μ 的所有线性函数的同时可信区间感兴趣，使用Bonferroni 法。本例为$(-5.18 \leq \mu_1 \leq 3.68)$ 和$(0.825 \leq \mu_2 \leq 5.425)$

上述过程在BMDP 3D 程序为：

```
/PROBLEM   TITLE IS 'BLOOD DATA'.
           VARIABLES ARE 3.
/VARIABLES NAMES ARE SUGAR,SYST,DIAS.
/TEST      VARIABLES ARE SUGAR,SYST,DIAS.
           HOTELLING.
```

```
  /PRINT     DATA.
             COVARIANCE.
             CORRELATION.
  /END
```

上述问题在SAS GLM 中处理，程序为：

```
* Hotelling T square;
data;
input sugar syst dias @@;
a=1;
cards;
30  -8  -1   30  -2   5
90   7   6   60   0   3
-10 -2   4   0   -2   4
10   0   2   40   1   2
proc corr cov;
   var sugar syst dias;
run;
proc glm;
   class a;
   model sugar syst dias=a/noint;
   manova h=a;
quit;
```

PROC CORR 印出变量间的协方差阵(选项COV)，MANOVA 语句据因素A对变量行检验。

其均值与协方差矩阵分别为：

$$\overline{X} = \begin{pmatrix} 31.250 \\ -0.750 \\ 3.125 \end{pmatrix}$$

及

$$S = \begin{pmatrix} 1069.642857 & 82.500000 & 16.964286 \\ 82.500000 & 17.357143 & 6.392857 \\ 16.964286 & 6.392857 & 4.696429 \end{pmatrix}$$

过程默认打印单变量的检验，变量SUGAR：F 值为7.30，P 值0.0305；变量SYST：F=0.26, P=0.6263；变量DIAS：F=16.63，P=0.0047；自由度均为1,7。

多元方差分析(MANOVA)的结果：

```
    Characteristic Roots and Vectors of: E Inverse * H, where
     H = Type III SS&CP Matrix for A   E = Error SS&CP Matrix
Characteristic Percent      Characteristic Vector  V'EV=1
   Root
                              SUGAR           SYST          DIAS
```

```
11.294801079    100.00    0.01087150    -0.14412966    0.23692200
0.0000000000      0.00   -0.01076657     0.02062595    0.11261594
0.0000000000      0.00    0.00193692     0.08070481    0.00000000
```

```
            Manova Test Criteria and Exact F Statistics for
                 the Hypothesis of no Overall A Effect
       H = Type III SS&CP Matrix for A   E = Error SS&CP Matrix
                          S=1    M=0.5    N=1.5
Statistic                   Value        F   Num DF Den DF  Pr > F
Wilks' Lambda           0.08133519  18.8247    3        5  0.0037
Pillai's Trace          0.91866481  18.8247    3        5  0.0037
Hotelling-Lawley Trace 11.29480108  18.8247    3        5  0.0037
Roy's Greatest Root    11.29480108  18.8247    3        5  0.0037
```

后面两部分结果指明是矩阵$E^{-1}H$ 的特征值和特征向量，Wilks 的似然比为0.081335，对应的F 值为18.8247，对比F(3,5) 界值，概率为0.0037 ，表明药物作用的存在。S, M, N 的意义在SAS 说明书中有说细的说明，如：SAS/STAT User's Guide, Release 6.03 Edtion, pp 16-17。Wilks' Lambda 是一种多元显著性检验。取值范围为0—1，其较大的值提示均值不存在差异，当所有均值相同时取值为1。有时也称U统计量。Hotelling 迹是根据特征值之和进行的多元显著性检验。

相应的SPSS/PC+程序如下：

```
SET MORE OFF.
data list free /sugar syst dias.
begin data.
30  -8  -1   30  -2   5
90   7   6   60   0   3
-10 -2   4   0   -2   4
10   0   2   40   1   2
end data.
manova sugar syst dias.
```

系统自动按析因分析处理，结果与SAS相同。单变量F检验自由度为1,7，F值和P值与SAS相同。

2. 两样本配对T^2检验

记$d_i, i = 1, .., .n$是两配对样本的差，设$d_i \sim N_p(\delta, \Sigma)$

检验$H : \delta = 0$, 对$A : \delta \neq 0$

若$\frac{n(n-p)}{p(n-1)}\overline{d}' S_d \overline{d} \geq F_{p,n-p,\alpha}$ 则拒绝H。

其中$\overline{d} = \frac{\sum d_i}{n}$, $(n-1)S_d = \sum_{i=1}^{n}(d_i - \overline{d})(d_i - \overline{d})'$

以下数据是Maindonald,JH[16]的一个例子。

```
 -0.2  1.6  1.3
  8.2 11.1  1.1
```

```
-1.9 -2.2  0.9
 4.4  6.2  2.5
 1.5  4.6  2.0
 2.1  2.7  0.3
 1.7  1.6  1.8
-1.5 -0.2  3.0
 2.3  6.9  3.4
```

使用SAS分析语句与前面例子相同。

```
data;
input d1-d3;cards;
.... 数据行....
proc glm;
class a;
model d1 d2 d3=a/noint;
manova h=a;
run;
```

结果Hotelling-Lawley Trace=4.27933866，F=8.5587，自由度为3，6，P= 0.0138，可以认为各差值在0.05水平上有差异。类似地，SPSS/PC+语句是：

```
data list free /d1 d2 d3.
begin data.
.... 数据行....
end data.
manova d1 to d3/print cellinfo(means).
```

结果如下：

```
EFFECT .. CONSTANT
Multivariate Tests of Significance (S = 1, M = 1/2, N = 2 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .81058    8.55868       3.00       6.00       .014
Hotellings      4.27934    8.55868       3.00       6.00       .014
Wilks            .18942    8.55868       3.00       6.00       .014
Roys             .81058
```

3. 两样本均值检验

设$X \sim N_p(\mu_1, \Sigma)$, $Y \sim N_p(\mu_2, \Sigma)$，$X$ 与Y 的独立样本分别为$X_1, ..., X_{n_1}$及$Y_1, ..., Y_{n_2}$。

检验$H : \mu_1 = \mu_2$ 对$A : \mu_1 \neq \mu_2$

现有μ_1, μ_2 及Σ 的估计量

$$\overline{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}, \quad S_1 = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})(X_i - \overline{X})'}{n_1 - 1}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}, \quad S_2 = \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y})(Y_i - \overline{Y})'}{n_2 - 1}$$

$$S_p = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$$

$$T^2 = \frac{(\overline{X}-\overline{Y})' S_p^{-1} (\overline{X}-\overline{Y})}{(1/n_1 + 1/n_2)}$$

若由样本观察值计算得到$\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2 \geq F_{p,n_1+n_2-p-1;\alpha}$ 则以显著水平α拒绝H。

对任一$a \neq 0, a'(\mu_1-\mu_2)$ 的$(1-\alpha)100\%$ 可信区间为：

$$a'(\overline{X}-\overline{Y}) - [T_\alpha^2(\frac{1}{n_1}+\frac{1}{n_2})a'S_p a]^{0.5} \leq a'(\mu_1-\mu_2) \leq a'(\overline{X}-\overline{Y}) + [T_\alpha^2(\frac{1}{n_1}+\frac{1}{n_2})a'S_p a]^{0.5}$$

其中

$$T_\alpha^2 = \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1;\alpha}$$

【例2.12】8 只狗分至两组接受不同的实验处理，第一组是控制组，第二组每只狗腿上置一金属盘，然后测量狗腿的张力与压力，其数据见表 2.9。

表 2.9 两组狗对某种处理的实验结果[21]

	控制组(X)				实验组(Y)			
编号	1	2	3	4	1	2	3	4
张力	131.5	145	191	150	40.5	80	50	90
压力	9	12	30	36	54	74.5	64.5	60.5

控制组与实验组的均值分别是$\overline{X} = \begin{pmatrix} 141.875 \\ 21.75 \end{pmatrix}$和$\overline{Y} = \begin{pmatrix} 65.125 \\ 63.375 \end{pmatrix}$

合差协方差阵的估计值为$S = \begin{pmatrix} 309.90 & 86.36 \\ 86.36 & 124.99 \end{pmatrix}$

现在，

$$T^2 = \frac{(\overline{X}-\overline{Y})' S^{-1} (\overline{X}-\overline{Y})}{\frac{1}{n_1}+\frac{1}{n_2}} = 116.7$$

$$T_{0.05}^2 = \frac{p(n_1+n_2-2)}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1;0.05} = 13.9$$

拒绝两实验结果相同的假设。

可信区间为$(\overline{x}_1-\overline{y}_1) \pm (13.9 \times 0.5 \times 309.9)^{0.5} = 76.75 \pm 46.41$ 和$(\overline{x}_2-\overline{y}_2) \pm (13.9 \times 0.5 \times 124.99)^{0.5} = -41.625 \pm 29.47$

使用Bonferroni 可信区间为

$$(\overline{x}_1-\overline{y}_1) \pm 3.03\sqrt{(0.5 \times 309.9)} = 76.75 \pm 37.7 \quad (\overline{x}_2-\overline{y}_2) \pm 3.03\sqrt{(0.5 \times 124.99)} = -41.625 \pm 24.0$$

其中$t_{0.0125,7} = 3.03$

BMDP 程序为：

```
/PROBLEM   TITLE IS 'DOG DATA'.
           VARIABLES ARE 3.
/VARIABLES NAMES ARE STRIDE, STRAIN, TREAT.
           GROUPS IS TREAT.
```

```
  /GROUP     CODE(3) ARE 1,2.
             NAMES(3) ARE CONTROL, TREAT.
  /TEST      VARIABLES ARE STRIDE,STRAIN.
             GROUPS ARE 1,2.
             HOTELLING.
  /PRINT     DATA.
             COVARIANCE.
             CORRELATION.
  /END
```

SAS GLM 相应的程序如下：

```
* tests difference between two populations;
data dogs;
input stride strain treat @@;
cards;
131.5  9.01 1  40.5  54.02 2
145.0 12.01 1  80.0  74.52 2
141.0 30.01 1  50.0  64.52 2
150.0 36.01 1  90.0  60.52 2
proc glm;
   class treat;
   model stride strain=treat;
   manova h=treat;
quit;
```

一元方差分析结果：STRIDE：均值103.5，F= 38.2, P=0.0008；STRAIN：均值42.57，F=27.74，P=0.0019，F的自由度均是1,6。

多元方差分析结果：

```
  Characteristic   Percent       Characteristic Vector  V'EV=1
       Root
                                          STRIDE         STRAIN
    19.455518487    100.00           0.02253459    -0.03337111
    0.0000000000      0.00           0.01258064     0.02319117
         Manova Test Criteria and Exact F Statistics for
            the Hypothesis of no Overall TREAT Effect
 H = Type III SS&CP Matrix for TREAT   E = Error SS&CP Matrix
                        S=1    M=0    N=1.5
Statistic                  Value        F   Num DF Den DF  Pr > F
Wilks' Lambda           0.04888656  48.6388      2      5  0.0005
Pillai's Trace          0.95111344  48.6388      2      5  0.0005
Hotelling-Lawley Trace 19.45551849  48.6388      2      5  0.0005
Roy's Greatest Root    19.45551849  48.6388      2      5  0.0005
```

拒绝处理结果相同的假设。

相应的SPSS/PC+程序如下：

```
data list free/ stride strain treat.
begin data.
131.5  9.01 1 40.5  54.02 2
145.0 12.01 1 80.0  74.52 2
141.0 30.01 1 50.0  64.52 2
150.0 36.01 1 90.0  60.52 2
end data.
manova stride strain by treat(1,2)
   /DESIGN treat /print homogeneity(all) ERROR(COVARIANCES SSCP).
```

Cochran和Bartlett-Box检验均显示齐性。多元Box M 检验、F检验和χ^2 检验的结果如下：

Boxs M =	5.06310	
F WITH (3,6479) DF =	1.07931,	P = .357 (近似)
Chi-Square with 3 DF =	3.23476,	P = .357 (近似)

其余结果同上。

在SAS中，PROC DISCRIM 可用于方差阵齐性检验。对于协方差阵不等的两正态总体均值检验，是多元的Behrens-Fisher 问题，可采用Scheffë 或Yao 方法处理，见文献[21]。

4. 推广的t检验

单变量t检验的一个推广是p种处理与其中单一响应变量行比较，每个对象或实验单元在连续的时间接受这p个处理，第j个观察是$X_j = (x_{1j}, ..., x_{pj})', j = 1, ..., n, x_{ij}$是第$i$种处理对第$j$个对象的反应。为了比较，考虑：

$$\begin{pmatrix} \mu_1 - \mu_2 \\ ... \\ \mu_1 - \mu_p \end{pmatrix} = \begin{pmatrix} 1 & -1 & ... & 0 \\ ... & ... & ... & ... \\ 1 & ... & ... & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ ... \\ \mu_p \end{pmatrix} = C_1\mu$$

或

$$\begin{pmatrix} \mu_2 - \mu_1 \\ ... \\ \mu_p - \mu_{p-1} \end{pmatrix} = \begin{pmatrix} -1 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ ... \\ \mu_p \end{pmatrix} = C_2\mu$$

C_1, C_2 称为对比矩阵，有p-1个行线性无关，每个都是一个对比向量，对比向量各元素之和为0。处理的均值相同时$C_1\mu = C_2\mu = 0$。事实上在处理间无差别时的假设就成了C μ =0，C是对比矩阵的任何一种择取。这时有均值$C\overline{X}$，方差CSC'，T^2统计量就是：

$$T^2 = n(C\overline{X})'(CSC')^{-1}(C\overline{X}) \sim \frac{(n-1)(p-1)}{n-p+1} F_{p-1,n-p+1;\alpha}$$

【例2.13】下面是J. Atlee 关于19条狗的试验数据。19 只狗开始给予苯巴比妥，然后每条狗在两种不同压力的CO_2上加上卤烷(Halothane)，最终测到狗的心跳间隔(毫秒)。用C表示CO_2，C+和C-表示其高低两个水平；用H表示卤烷，用H+和H-表示其高低两水平，原始数据列于表 2.10：

表 2.10 J. Atlee 19 条狗的实验数据

狗号	C+H-	C-H-	C+H+	C-H+	狗号	C+H-	C-H-	C+H+	C-H+
1	426	609	556	600	11	349	382	473	497
2	253	236	392	395	12	429	410	488	547
3	359	433	349	357	13	348	377	447	514
4	432	431	522	600	14	412	473	472	446
5	405	426	513	513	15	347	326	455	468
6	324	438	507	539	16	434	458	637	524
7	310	312	410	456	17	364	367	432	469
8	326	326	350	504	18	420	395	508	531
9	375	447	547	548	19	397	556	645	625
10	286	286	403	422					

其均值与协方差矩阵分别为：

$$\overline{X}=\begin{pmatrix}368.21053\\404.63158\\479.26316\\502.89474\end{pmatrix}\quad S=\begin{pmatrix}2819.29 & 3568.42 & 2943.50 & 2295.36\\3568.42 & 7963.13 & 5303.99 & 4065.46\\2943.50 & 5303.99 & 6851.32 & 4499.64\\2295.36 & 4065.46 & 4499.64 & 4878.99\end{pmatrix}$$

记其理论均值为μ_1,μ_2,μ_3,μ_4。考虑以下几种效应：

卤烷H $(\mu_3+\mu_4)-(\mu_1+\mu_2)$

CO_2 $(\mu_1+\mu_3)-(\mu_2+\mu_4)$

H-C 的交互 $(\mu_1+\mu_4)-(\mu_2+\mu_3)$

即

$$C=\begin{pmatrix}-1 & -1 & 1 & 1\\1 & -1 & 1 & -1\\1 & -1 & -1 & 1\end{pmatrix}=\begin{pmatrix}C_1'\\C_2'\\C_3'\end{pmatrix}$$

$T^2=n(C\overline{X})'(CSC')^{-1}(C\overline{X})=116$

在$\alpha=0.05$时，$\frac{(n-1)(p-1)}{n-p+1}F_{p-1,n-p+1;\alpha}=[18(3)/16]F_{3,16;0.05}$ =10.94。因此拒绝处理相同的假设，卤烷效应的95%可信区为：

$$(\overline{X}_3+\overline{X}_4)-(\overline{X}_1+\overline{X}_2)\pm\sqrt{10.94}\sqrt{C_1'SC_1/19}=209.31\pm73.70$$

其它两个可信区间分别是-60.05 ± 54.70 和-12.79 ± 65.97。

5.重复数据的检验

当对同一实验单元进行多次测量，测量的结果彼此相关。若这些测量性质不同，如重量、长度、宽度，则用多元方差分析等方法；若测量是在实验因素如时间、药物不同剂量等不同水平上进行，则由重复测量(repeated measures)方差分析处理。

随机顺序下可假设其协方差阵是$\Sigma=\sigma^2[I(1-\rho)+\rho ee']$, $e=(1,...,1)'$。即组内相关(intraclass correlation)模型。

重复测量分析有许多特点：因为误差中除去了个体间差异的影响，所以效率要高，精度提高，实验所需对象数目减少。分析技术可以是一元或者多元，文献建议观察数少于处理数＋10时使用一元分析。它假设数据的分布符合多元正态、各观察独立并且是球形(sphericity，这在多元分析方法中不需要)。球形要求所有重复测量对的方差相同，尽管在SPSS/PC+中有这样的检验，却不推荐。若不满足球形，则Ⅰ类误差增大。

Greenhouse 和Geisser(1959) 利用 ε 对重复测量结果进行调整。当球形假设成立时，$\varepsilon = 1$，最差时为1/(k-1)，k为处理数。SAS 输出它的Huynh 和Feldt(1976) 修正，当重复测量设计中偏离球形假设时，分子和分母自由度均乘以该值，使用调整后的自由度计算观察显著性水平。Huynh 和Feldt 证明Greenhouse-Geisser ε比较保守，特别对于小样本更是如此。常用的轮廓分析或形象分析(profile analysis)，特色是构造合适的对比矩阵，SYSTAT 的MGLH 模块和SAS、SPSS/PC+都能进行，这要求数据被很好地标化，以两组做为例：

$$\mu_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1p})', \mu_2 = [\mu_{21}, \mu_{22}, ..., \mu_{2p}]'$$

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2}$$

要检验$H : \mu_1 = \mu_2$即两总体具有相同的均值，有三个检验：H_1. 轮廓是相似的吗?H_2.若轮廓相似，是重合的吗? H_3.若轮廓重合，轮廓的各水平是相同的吗?

$H_1 : \mu_{1i} - \mu_{1,i-1} = \mu_{2i} - \mu_{2,i-1}, i = 2, ..., p$ 或$C(\mu_1 - \mu_2) = 0$ 相对于$A_1 : C(\mu_1 - \mu_2) \neq 0$

$$C_{(p-1)\times p} = \begin{pmatrix} -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & -1 & 1 \end{pmatrix}$$

从而$Ce = 0, e = (1, ..., 1)'$，以上检验即$H_1 : C(\mu_1 - \mu_2) = \gamma e$相对于$A_1 : C(\mu_1 - C\mu_2) \neq \gamma e$，$\gamma$为轮廓间的平均差异。检验统计量为：

$$\frac{n_1 + n_2 - p}{(n_1 + n_2 - 2)(p-1)}(\overline{X}_1 - \overline{X}_2)'C'[(\frac{1}{n_1} + \frac{1}{n_2})CSC']^{-1}C(\overline{X}_1 - \overline{X}_2)$$

与$F_{p-1,n_1+n_2-p;\alpha}$ 相比较

H_2与A_2即$H_2 : \gamma = 0$相对于$A_2 : \gamma \neq 0$

$$T^2 = (\overline{X}_1 - \overline{X}_2)'C'[(\frac{1}{n_1} + \frac{1}{n_2})CSC']^{-1}C(\overline{X}_1 - \overline{X}_2)$$

检验统计量$(\frac{1}{n_1}+\frac{1}{n_2})^{-1}(e'S^{-1}(\overline{X}-\overline{X}_2))^2(e'S^{-1}e)^{-1}(1+\frac{T^2}{n_1+n_2-2})^{-1}$ 与$F_{1,n_1+n_2-p-1;\alpha}$ 相比较，γ的极大似然估计为: $\hat{\gamma} = \frac{e'S^{-1}(\overline{X}_1-\overline{X}_2)}{e'S^{-1}e}$

H_3与A_3即$H_3 : \mu_1 = \delta e, \mu_2 = \xi e$，$\delta, \xi$未知，或$H_3 : C(\mu_1+\mu_2) = 0$ 相对于$A_3 : \mu_1 \neq \delta e, \mu_2 \neq \xi e$，有

$$\frac{(n_1 + n_2 - p)(n_1 + n_2)}{(n_1 + n_2 - 2)(p-1)}\overline{X}'C'[CSC']^{-1}C\overline{X}$$

与$F_{p-1,n_1+n_2-p;\alpha}$比较，$\overline{X} = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$。

第 4 章给出了一个利用PROC GLM 进行三组轮廓分析的例子，第 5 章也给出了相应的SPSS/PC+程序。有关协方差阵的其它检验可见如[21]。

§2.3.2 回归分析

1. 回归分析是研究应用最为广泛的多元分析技术。

因变量Y 和p 个自变量$X_1, ..., X_p$ 线性回归模型是

$$y = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)$$

n 次观测$(y_i, x_{i1}, ..., x_{ip}), i = 1, ..., n$，具有如下线性关系：

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, i = 1, .., n$$

这里假定$\varepsilon_i \sim N(0, \sigma^2)$, 其回归系数的最小二乘估计为：$\hat{\beta} = (X'X)^{-1}X'Y \equiv HY$，$X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix}$，$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}$。

2. 有关回归方程的统计量

⑴R^2和校正后的R^2

R^2是度量一个线性模型拟合优度的常用统计量，它不仅是自变量X和因变量Y复相关系数的平方，还是因变量Y与其预测值相关系数的平方。由于样本R^2 对模型拟合好坏趋于做出一个乐观估计，使用校正R^2 以更好地刻划模型拟合情况。校正后的R^2为：

$R_a^2 = R^2 - \frac{(1-R^2)p}{n-p-1}$，其中$p$ 为自变量个数，n 为观测个数。

⑵方差分析表

其F检验用于检验线性回归方程的零假设：$H : \beta_1 = ... = \beta_p = 0$,即因变量与所有自变量无线性关系。F统计量可以表为：F=回归均方/残差均方$\sim F_{p,n-p-1}$

因此，可对R^2 给予另一种解释：R^2是能被模型所解释的因变量的那一部分比例，即：R^2=1-残差平方和/总平方和，R_a^2=1-[残差平方和/(n-p-1)]/[总平方和/(n-1)]

⑶R^2 改变量

R^2 改变量是评价自变量相对重要性的一个常用指标，即考察当一个变量进入方程时R^2 的增量$R^2 - R(i)^2$，其中$R(i)^2$ 是当除第i个自变量外其它自变量均包括在方程中时的复相关系数的平方。显然R^2 的改变量大则意味着第i个变量提供其它已在方程中的变量所不能提供的信息量最大。

⑷条件数(condition number)

方阵$X'X$ 的条件数定义为$k = \lambda_1/\lambda_p$，其中λ_1 和λ_P 分别是$X'X$ 的最大和最小特征根，常用来刻划多重共线性是否存在以及严重程度。经验上。若k 取值在范围$100 \sim 1000$，则认为存在中等或较强的多重共线性；若$K > 1000$，则认为存在严重的多重共线性。

3. 关于自变量的统计量

⑴t统计量

对零假设$H : \beta_i = 0$，可用具有自由度为$1, (n - p - 1)$ 的F 统计量进行检验，但由于具有k个自由度的t值平方等于具有$1, k$ 自由度的F 值，故可用t统计量检验上述假设。

⑵标准化回归系数

一般来说，回归方程中各自变量测量单位不同, 因此不能将其系数大小视为变量重要性标志。标准化回归系数使系数在一定程度上具有可比性, 它是当所有变量用其标准化形式时自变量的系数，可以直接从回归系数计算。

$\tilde{\beta}_i = \beta_i S_i / S_y, S_i$ 是第i 个自变量的标准差，S_y 是因变量Y 的标准差。

⑶方差估计

仅仅按照每个自变量系数的t值的显著性判断对预测的重要性是危险的，因为那些具有较大方差的系数估计是不可靠的。所以，我们常常关心回归系数方差的估计量。

⑷变量容许性

记R_i^2 为当第i 个自变量被视为因变量且它与其它自变量计算产生回归方程时的复相关系数，R_i^2 较大同时表明第i 个自变量几乎是其它自变量的线性组合。$1 - R_i^2$ 反映了其它自变量未解释的变异比例，称为第i 个自变量的容许性。容许性是描述自变量之间相互依赖的常用指标。一个容许性小的变量进入方程，不仅导致估计的方差增大，还会引起一些计算上的问题。此外，即使某一备选变量具有可接受的容许性从而可以进入方程，但由此可能导致原来已在方程中的那些变量的容许性变得不可接受的低。因此，在逐步回归中的每一步，都应当重新计算方程中全部变量的容许性。在SPSS REGRESSION中，通过TOLERANCE 设置变量容许性，其默认值为0.01。

⑸部分相关系数与偏相关系数

对于R^2 的改变量$R^2 - R(i)^2$ 带有正负号的平方根称为部分相关系数，部分相关系数是当其它自变量的线性效应从X_i 中消除之后Y 和X_i 的相关系数。另一个重要的系数是$PR_i^2 = [R^2 - R(i)^2]/[1 - R(i)^2]$ 其分子是部分相关系数的平方，分母是当除第i 个自变量以外所有其它自变量包含在方程中时解释变异的比例。带有正负号的PR_i^2 平方根称为偏相关系数，可解释为当其它自变量的线性效应从X_i 和Y 两者中消除后第i个自变量与因变量的相关系数。注意在绝对意义上部分相关系数不大于偏相关系数。

⑹关于未入选变量的统计量

对于尚未进入方程的自变量，可通过如下统计量考察其性质：如果该变量进入方程，其回归系数；对该系数为零的假设的t检验及概率水平；与因变量的偏相关系数与容许性等。由这些统计量可帮助判断下一步应该进入的变量。

通常，对于一个具体问题，我们事先并不知道上述模型是否合适。因此，有必要利用观测数据对模型假设的合理性给予考察，这种考察主要是通过残差分析进行的。若对数据而言模型是合适的，则残差E_i 作为e_i 的估计具有与e_i 类似的特征。

残差分析的内容很丰富，现只就SPSS REGRESSION 中有关残差分析的三个主要方面作一简要介绍，其一是利用残差分析考察上述模型假设的合理性，其二是探查对回归分析产生较大作用的异常点的强影响点。最后是多元共线性问题。设$e = (I - H)Y$, 它服从正态分布$N(0, (I-H)\sigma^2)$，方差估计用$(I-H)s^2$代替，就有标准化残差，$e_i/s\sqrt{1-h_{ii}}, h_{ii}$是H 的第$i$ 个对角元，在SAS 中称为STUDENT。据Belsley, Kuh and Welsch (1980) 的建议，使用$s_{(-i)}$代替s_i，在SAS 中称为RSTUDENT。

4. 模型假设合理性考察

⑴线性假设。方法一：作“残差图——预测值”图(即以预测值为横轴，残差为纵轴，将各观测个体相应的点绘于图上)。如果直线性和方差齐性得以满足，则预测值与残差值之间应不存在任何关系，即在假设满足时，残差应随机分布在通过0 的水平直线所展开的带状区域内(通常是|残差| ≤ 2)，如果从图上可看出任何变化模式，就应怀疑上述假设是否满足。方法二：作“残差——某自变量图”，同样若假设得以满足，残差应随机分布于水平带内。特别地，可考虑以未入选方程的那些自变量作残差图，若发现残差不是随机分布，可考虑将该变量包括进方程内。

⑵等方差性假设。如上所述，可利用作图的方法，若方差随自变量或预测值的增长而增长(或减小)，则应当怀疑Y 对所有X 值均为等方差这一假设。在SPSS REGRESSION 中，利用SCATTERPLOT 子命令可以对指定的一对变量作其散点图，以考察其直线性和等方差性假设。

⑶误差独立性假设。只要数据是按顺序(如时间) 收集，就应当作“残差—顺序变量”图，这是因为即使时间并未作为模型中的一个变量，它也可能影响残差。

如果残差与收集顺序无关，在上述图形上不应该发现任何变化模式，当误差项是正相关时，残差在一段上为正，另一段则为负；当误差呈负相关时，残差符号变化很频繁。利用Durbin-Watson 统计量可以检验相邻残差项是否为序列相关。残差自相关可用e_i与$e_{(i-1),i}$ 为下标的图来表现，用Durbin-Watson 检验。

⑷正态性假设。方法一，作残差直方图。方法二，作观测残差累积分布—期望残差累积分布“图(P-P 图)”。显然，当两者一致时，应产生一条直线，其中期望残差作横轴，观测残差为纵轴。正态性检验的Q-Q 图正常时为一条过原点的线，其斜率依赖于残差的标准差。另外可以使用W-统计量及Shapiro-Francia 统计量，它是观察次序残差与正态次序统计量的相关系数的平方。

⑸偏回归图。偏回归图是考察合理性的另一重要工具。对第j个自变量的偏回归图由两个残差构成，第一个是删除了第j个自变量后，其余自变量对因变量所做回归的残差(常称为偏回归残差)，第二个是自变量j与其余自变量回归所做回归的残差。在SPSS REGRESSION 中，利用PARTIALPLOT 子命令可对指定自变量作偏回归图。

⑹数据修正。当发现数据不符合模型假设时，可考虑对数据作适当变换。当直线性假设不符时，可根据描点和实际知识背景知识对数据作某些变换。当等方差假设不符时，可考虑对因变量作变换，使变换后的数据的误差方差相等。当误差独立性不成立时，可采用“两步估计法”。

5. 检查异常点和强影响点

⑴对某些观测个体，若其残差明显比其它观测个体的残差大很多，则称为异常点。由于异常点具有绝对值较大的残差，故可以直接使用残差图探查异常点。其次，可以通过直方图检查异常点。SPSS REGRESSION 对学生化残差的绝对值大于3.16 的观测个体直方图中都用”Out” 标记出来，最后还可以用逐点残差异常图来获得详细信息。

⑵马氏距离和中心化杠杆。马氏距离反映了某个观测点到观测中心的距离。第i 个观测个体的中心化杠杆值定义为：$L_i = D_i/(n-1)$，D_i 是第i 个观测点到中心的马氏距离。L_i 的取值范围从$-1/n$ 到$(n-1)/n$，均值为n/p。L_i 的取值越大则对回归的影响越大。

⑶强影响点。在一组数据中，对参数的估计具有特别大影响的观测个体称为强影

响点。这样的点被删除后回归直线与删除前相比有很大不同。识别强影响点的方法之一是当怀疑某个点是强影响点时删除该点，重新计算残差，考察其变化情况。一个点被删除后所计算的残差称为删除残差，删除残差除以其标准误则产生学生化删除残差。对回归效果有潜在影响的点是在空间上离$\overline{X}$ 较远的点，此距离可用h_{ii}来表示，它是投影阵H 的第i 个元，因为$\Sigma h_{ii}=p$，故其均值为p/n，Belsley, Kuh & Welsch (1980) 建议用$2p/n$做为它的界值(CUTOFF)，其上界是$3p/n$。有四种反映点的影响的统计量，它们是通过去掉该点来反映的。Cook's D 反映对回归系数估计值的影响，DFFITS 反映对预测值的影响，DFBETAS 反映对特定回归系数的影响，COVRATIO 反映对参数估计量方差—协方差阵的影响。前面三个可以想象成对去掉观察i 后对k 个线性无关的回归系数的影响，即$k(\beta-\beta_{-i})$ 可以写成与一般线性模型类似的二次型。三种度量对应不同的k值。建议用$2\sqrt{p/n}$做DFFITS 界值，其上界为$\sqrt{p}$，对Cook 距离近似用$4/n$ 做界值，其上界为1 ，对DEBETAS 建议用$2/\sqrt{n}$做界值，其上界为1，对于COVRATIO，建议值为$1\pm 3p/n$做界值。Cook 距离反映了当第i 个点被删除后在所有残差中的变化，常用于识别强影响点，其定义为：$C_i=\Sigma_{k=1}^{N}(\hat{y}_{k(i)}-\hat{y}_k^2)/[ps^2]$ 其中p 为自变量个数，s^2 是残差方差的估计量。很明显，强影响点时库克距离较大。

⑷异常点和强影响点的处理。发现异常点，应当根据专业知识的数据收集情况对其进行慎重分析处理。若发现是由于数据失误导致异常或强影响点，应当删除观测数据。若发现数据确系系统身产生，则应予以保留。考虑用一些稳健方法进行参数估计。

6. 多元共线性问题

若$|X'X|$不满秩，即多元共线性(multicollinearity) 问题，正规方程将有非正常解，为此出现了岭回归及主成分回归等相关方法。共线性诊断中的条件指标(condition index) 是矩阵条件数(condition number) 的推广，第i 个条件指标是最大特征值与第i 个特征值之比，Belsley 建议在10 附近为有相关的影响，大小100 为严重共线性，看一下有几个条件指标指示共线性的存在，从每个共线性中去掉一变量，若去掉变量后拟合效果太差，则应在拟合和共线性之间进行折衷。共线性诊断也可采用有方差扩张因子，若其值大于10时有共线性存在。

共线性处理常用方法：1. 适当的变量转换，增加一些观察，但这往往并不现实；2.从模型中去掉一些变量，两个自变量高度相关，模型中通常包括一个就够了；3.进行主成分回归。许多回归诊断结果提示的最好方法是删除记录，另外可采用稳键回归(robust regression) 等方法。岭回归(ridge regression)是一种有偏估计。对于回归方程$y=\beta_0+X\beta+\varepsilon,\beta_0$ 的估计是y的均值，而β的岭估计是：

$$\beta(k)=(X'X+kI)^{-1}X'y$$

$$Var[\beta(k)]=(X'X+kI)^{-1}(X'X)(X'X+kI)^{-1}\sigma^2$$

据Hoerl 与Kennard，存在$k>0$使$E[\hat{\beta}(k)-\beta]^2<E[\beta-\beta]^2$ k=0 时即通常的最小二乘估计，在k 增大时，方差扩张因子减少而$\beta_{(k)}$ 的偏倚增大，k 通常可取如$0.005\sim 0.2$，每个回归系数对k 做图给出岭迹(ridge trace)并由此确定k的取值。不同的作者有不同的做法。也可使用方差扩张因子VIF进行k 的选择，此时VIF 应于范围$1\sim 10$。

回归诊断在SAS(REG)、SPSS(REGRESSION)、Stata(regress) 等软件包均可以得到，SAS有专用过程RIDGE 进行岭回归分析，RIDGE 也可做为PROC REG 的选项进行岭回归分析。在SPSS 中使用RIDGEREG 宏定义进行岭回归分析。启用方法：

RIGDGEREG DEP=因变量/enter自变量/start= /stop= /inc= /k=后面选项设定k 值。

描述、分析和研究因素间的相互关联和影响，可以通过一些统计指标如相关系数和典型相关，还可以通过考查其相关的结构。据Karlin, S. (1983)，常用的方法有：通径分析、结构方程模型以及方差分量分析，它们都基于线性的假设。结构方程模型在第14章介绍，第4章有CALIS的例子。结构方程模型与通径分析方法有密切的联系。

多元线性模型包含多元回归模型、多元方差分析模型用协方差分析模型等。在SAS 中有专门的过程MIXED 处理混合模型数据。方差分量分析是基于一般线性模型，估计各变异来源的方差组分大小，可用于遗传学分析等。由平衡资料估计方差分量即为方差分析法(ANOVA)，即把方差分析表中的均方作为其期望值的估计。对于非平衡资料估计方差分量采用Henderson 法Ⅰ、Ⅱ、Ⅲ及极大似然法(ML)、约束极大似然法(REML)、MINQUE、MIVQUE和I-MINQUE 等，在SAS 中方差分量分析使用过程VARCOMP 实现。

§2.3.3 方差分析

许多研究，要归结到几个样本均数的比较。设有k个样本均数，要比较它们的差别，若取检验水准$\alpha = 0.05$，对这些均数用t—检验两两比较，共有$k(k-1)/2$个比较，总结论的检验水准就成为$1-(1-\alpha)^k$，可见这样做既不经济，效率也低。方差分析正是处理这一类问题的统计学方法，方差分析有时也称变异数分析。

方差分析的应用条件是：各个样本来自正态总体；各个样本是相互独立的随机样本；各总体方差相等。

方差分析的基本思想是把所有数据的总变异(离均差来方和)分解成几个部分，然后对各部分的变异进行比较。完全随机设计或单因素设计，是把受试对象完全随机地分配到各个处理组中去。处理组可以为两组或多组，各组样本含量可以相等，也可以不等。完全随机设计方差分析把总变异区分为处理组间变异和组内变异。配伍组设计，也称随机区组设计，是扩展的配对设计。配伍组设计的方差分析可以把总变异分为处理组间变异、配伍间变异和误差三个部分，较完全随机设计提高了效率。若研究的因素很多，可以使用析因设计或正交设计。

使用MANOVA可控制整个实验水平上的误差，同时考虑了因变量间的关联。其基本假设是独立性、方差协方差阵相等、正态性。MANOVA需要更多的样本量，受异常值的影响也更大，其假设因变量间的线性组合。显著性检验准则常用的有四种，即Roy最大特征根、Wilks λ、Hotelling迹、Pillai准则。它们之间最基本的不同是对“不同维”上因变量的差异的评定方法。Roy准则利用第一个特征根来评价，这样其功效和特异度比较好，最适于因变量在某一准上存在强相关的情形，同时也是违背准则时受影响最大的。

对于方差分析假设的检验常用的如Bartlett检验、进行必要的数据转换等。

所谓平衡是指在分类变量的交叉下的记录个数相同；表2.11是一个不平衡设计的例子。

表 2.11 2×2 设计中的格子均值

数目	第一列	第二列	数值	第一列	第二列
第一行	2	200	第一行	10	20
第二行	200	2	第二行	30	40

第一行总均值=19.90。在行均值之间约差10，行效应存在；列均值之间亦约差10，列效应也是有效的，可以化出交互项，效应也存在，原因在于模型受了大样本的影响。表2.12也是一个不平衡设计的例子，负值表示缺失，括号内是每个格子中观察值的数目。

表 2.12 不平衡设计的例子

		因	子	B
数据(n_{ij})		1	2	3
因子 A	1	2,4,6	4,6	5
		(3)	(2)	(1)
	2	12,8	11,7	-1
		(2)	(2)	(0)

方差分析模型是$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

表 2.12 的SAS运算结果列如表2.13：

表 2.13 表 2.12的运算结果

来源	自由度	I	II	III	IV
A	1	60.00	57.02	54.55	54.55
B	2	0.32	0.32	0.21	1.50
A×B	1	2.18	2.18	2.18	2.18

对线性模型$Y = X\beta + \varepsilon, E(Y) = X\beta$，分析的基本目的是在可能的情况下估计或检验$\beta$或其线性组合，这可以由观测Y的线性组合来做到。又设β的线性组合是$L\beta$，则应有Y的线性组合，使其期望为$L\beta$，这就要求这样的组合存在，也就是说能找到X的相应的线性组合，因而X的行也就成为L的发生集，而且因为$X = X(X'X)^{-}(X'X), X'X$的行也成为L的发生集。根据设计是否平衡等因素，L取为不同的形式。如 I 型平方和是用修正扫描算子计算$X'X$的g2逆求解正规方程组的副产品，可用许多方法得到，一种方法是对$X'X$ 向前的Doolittle分解，跳过任何对角元为零的情形。

SAS 能输出四种平方和，I 型为顺序平方和，使用Searle的记号，$SS(A) = R(\alpha|\mu), SS(B) = R(\beta|\alpha,\mu), SS(AB) = R(\gamma|\alpha,\beta,\mu)$ 等等，其各SS的大小依赖于效应进入模型的顺序，在效应的排列很好时是有用的，如用于多项式回归可以看出是否需要继续引入高次项。在多项式回归中，相应于正交多项式检验，项的贡献可以很清楚。其它的特点如：所有效应的SS之和与模型SS 相同，若剩余是独立正态分布则各SS是独立的。对于不平衡资料，其假设是格点的函数，结果一般与平衡资料不同。II 、III、IV 型称为偏平方和，“偏”的含义是指进行了其它效应的调整，即每一种均调整了其它的分类效应，调整准则不同，检验与效应进入模型的顺序无关。如第 II 种$SS(A) = R(\alpha|\beta,\mu), SS(B) = R(\beta|\alpha,\mu), SS(AB) = R(\gamma|\alpha,\beta,\mu)$，一般不具有平衡设计中的那种相等分布(equitable distribution) 或正交(orthogonality)特性。III、IV 与 II 的区别是高阶交互或嵌套效应的系数也进行调整以满足正交性条件(III) 或等分布(IV)，这些效应的系数不再依赖于格子数n_{ij}，只要$n_{ij} \neq 0$时III、IV 是相同的。出现$n_{ij} = 0$时，III具有正交

特性，检验好象是针对“效应和为零”，而Ⅳ具有平衡特性，用非零格子的子集作为平衡数据集，结果不唯一。对于平衡资料，四种平方和相同；在模型没有交互时，Ⅱ=Ⅲ；在所有格子非空(all-cell-filled data)时Ⅲ=Ⅳ。在PROC GLM中，可估计函数由选择项E1-E4给出，几种平方和由S1-S4给出。

SPSS用MANOVA和ANOVA进行方差分析，并区分独立(UNIQUE)、顺序(SEQUENTIAL)等离均差平方和。

§2.3.4 主成分分析

在实际应用中，为了全面分析问题，提出的指标(或变量)往往很多，每个指标都在不同程度上反映了所要研究的课题的某些信息，由于指标之间常常具有一定的相关性。因此希望找到较少的几个彼此不相关的指标，来代替原来的指标并且尽可能地反映原来指标的信息，这就是主成分分析的思想方法。

1. 主成分的定义及求法

设$X=(X_1,X_2,...,X_p)$为p维随机变量，有二阶矩存在，记$\mu=E(X),\Sigma=Var(X)$，考虑它的线性变换$Z_i=l_i'X,i=1,...,p$, 其中$l_1'=(l_{11},l_{12},...,l_{1p})$，易见

$$Var(Z_i)=l_i'\Sigma l_i,\quad Cov(Z_i,Z_j)=l_i'\Sigma l_j,i\neq j$$

因此$Var(Z_1)$越大，表明Z_1包含的信息就越多，若$Z_1=l_1'X$满足

$$l_1'l_1=1,Var(Z_1)=\max Var(l'X)$$

则称Z_1是X的第一主成分，Z_1是X的所有线性变换$l'X$ 中最能综合原p个变量信息的一个线性变换，其中，使得方差达到最大的向量即为主成分系数。如果第一个主成分不足以代表原p个变量的信息，考虑第二主成分Z_2，为了有效地代表原变量的信息，第一主成分Z_1 已有的信息就不需要出现在第二主成分Z_2中，即$Cov(Z_1,Z_2)=0$。由此可得$l_1'l_2=0$或$l_2'l_1=0$。因此，若满足

$$l_2'l_2=1,l_2'l_1=0,Var(Z_2)=\max_l Var(l'X)$$

则称Z_2是X的第二主成分，一般地，如果$Z_i=l_i'X$ 满足

$$l_i'l_i=1,l_i'l_j=0,j=1,...,i-1,Var(Z_i)=\max_l Var(l'X)$$

则称Z_i是X的第i个主成分。

假定$\lambda_i,i=1,2,...,p$为方差矩阵$Var(X)=\Sigma$ 的p个特征根并且它们由大到小的排列为$\lambda_1\geq\lambda_2\geq...\geq\lambda_p\geq0$，则$X$的第$i$个主成分的系数向量$l_i$就是第$i$个特征根$\lambda_i$ 所对应的正则化特征向量，其中$l_i'l_i=1,l_i'l_j=0,i\neq j,i,j=1,...,p$

若记所有主成分构成的向量为Z, 相应的主成分系数矩阵为L,则上述线性变换即为$Z=L'X$, 而且可以得到以下结论：

(1). $L'L=I_p$, 即L 是正交阵，因此，主成分代表原变量空间中的垂直向量，或者说，主成分是对原变量进行了一次正交变换。

(2). Z的分量间互不相关，即相关系数矩阵$Cov(Z_i,Z_j)=0$。

(3). Z的p个分量是按方差大小，由大到小排列的。

(4). $Var(Z) = L'\Sigma L = diag(\lambda_1, \lambda_2, ..., \lambda_p)$。因此，由特征根$\lambda_i, i = 1, 2, ..., p$的大小可知它所包含信息的多少。一般，称$\lambda_k/\Sigma_{i=1}^{p}\lambda_i$为第k个主成分的方差贡献率，$\Sigma_{i=1}^{k}\lambda_k/\Sigma_{i=1}^{p}\lambda_i$称为前$k(k \leq p)$个主成分的累积方差贡献率。通常在实际应用中，当前个主成分的累积方差贡献率超过85%时，则用这前k 个主成分就可以描述原来p个变量的信息了。

(5). 相关系数矩阵$\rho(Z_k, X_i) = \sqrt{\lambda_k}l_{ki}/\sqrt{\sigma_{ii}}$其中$\sigma_{ii}$为$\Sigma$的第$i$个主对角元素，表示原变量在主成分中的负荷量，在实际应用中，为了消除量纲的影响，往往把原变量标准化，标准化后的协差阵即为原变量的相关阵。此时，要观察原变量的负荷量，只需观察l_{ki}的系数，也即Z_k的系数。

2. 样本主成分

在实际问题中，$X' = (X_1, ..., X_n)$的协差阵或相关阵常常是未知的，于是抽取随机样本，得到观察数据阵X，从X可得样本协差阵或样本相关阵，并求得它们的特征根以及对应的正则化特征向量，得到的主成分，即为样本主成分。若记$Z_i = l_i'X$为的第i个样本主成分，$Z_{ki} = l_i'X_k, i = 1, ..., p, k = 1, ..., n$, 称$Z_k = (Z_{k1}, Z_{k2}, ..., Z_{kp})', k = 1, ..., n$ 为主成分得分。

3. 主成分回归

在回归分析中，当自变量$X_1, X_2, ..., X_p$之间存在多重共线性关系时，$|X'X|$接近于0，此时用通常的最小二乘估计求得的回归方程就可能出现一些不符合实际的情况，主成分回归是一种可以选择的方法。设自变量$X_1, ..., X_p$ 和因变量Y有对应关系，首先对自变量进行中心化，仍记作$X = \{x_{ij}\}$，则有线性回归方程：

$$\hat{y} = \hat{c}_1x_1 + ... + \hat{c}_px_p$$

其中回归系数$\hat{c}_j, j = 1, ..., p$是通常的最小二乘解。

记矩阵X的奇异值分解$X'X = U\Lambda U'$ 中的正交阵为U, X 的主成分为$w_1, ..., w_p$, 现求Y对这些主成分变量的回归，得：

$$\hat{y} = \hat{b}_1w_1 + ... + \hat{b}_pw_p$$

则$\hat{b}$与$\hat{c}$有关系$\hat{b} = U'\hat{c}$，且$\Sigma\hat{y}^2 = \Sigma\hat{\lambda}b^2$

主成分回归的回归系数只与其相应的主成分有关，与其它主成分无关；主成分回归平方和与原来的回归平方和相等，且等于各主成分对 y 的回归平方贡献之和，此性质可用于变量的筛选。若主成分有明确的实际意义，则把主成分看成单个自变量，要减少参加回归的主成分，仅去掉若干个主成分即可；若主成分没有明确的意义，或仍希望用原始变量来求回归方程，则首先找出贡献最小者，再据其组合系数U值的大小进行取舍，因为该主成分对回归贡献小，则其对应的起主要作用的原始变量贡献也应较小，应予以舍弃。SAS 用PRINCOMP过程进行主成分分析。

§2.3.5 因子分析

1. 因子分析是识别代表大量相关变量相互关系的一组少量(通常是不可测的) 因子的统计技术。因子分析试图用最小个数的不可测的所谓公因子的线性函数与特殊因子来对原

来观测的变量进行描述，这样做的目的，是尽可能合理地解释存在于原变量间的相关性，并且简化变量的维数与结构。

例如: 教师想从各门课程考试的成绩中了解学生的“理解能力”,“计算能力”,“记忆能力”等等。成绩是可以测定的，而这些“能力”是不可直接测定的，习惯上称此为公共因子(common factor)，显然，成绩的好坏受这些公共因子的影响，而每门课都有其特殊性，因此,它还受到一个不可测的特殊因子(unique factor)的影响，因子分析就是根据一些变量$X_1...X_p$(相当于各门课程的成绩) 来得到公共因子$f_1...f_q(q<p)$(相当于“理解能力”,“记忆能力”等等)的统计方法。

2. 因子模型

(1). 初始因子模型

设$X=(X_1,X_2,...,X_p)'$为p维可观测随机变量，而且X已中心化并仍记作X,由q个公共因子$f=(f_1,...,f_q)'$所支配，其相应的因子模型为：

$$X_{p\times 1}=A_{p\times q}f_{q\times 1}+\varepsilon_{p\times 1}$$

这里假定：

$$E(X)=0,Var(X)=\Sigma>0,E(f)=0,Var(f)=I_q,q<p$$

$$E(\varepsilon)=0,Cov(f,\varepsilon)=0$$

$$Var(\varepsilon)=D=diag(e_1^2,e_2^2,...,e_p^2)$$

特殊因子$\varepsilon_i,i=1,...,p$彼此不相关且具有单位方差，每个公共因子至少对两个变量有贡献，否则它将成为特殊因子。

据模型有: $\Sigma=AA'+D\equiv H+D$

记$\Sigma=(\sigma_{ij})_{p\times p},A=(a_{ij})_{p\times q},H=(h_{ij})_{p\times p}$ 且$h_{ij}=\Sigma a_{ik}a_{kj},h_{ii}=h_i^2$ 则

$$\sigma_{ii}=h_i^2+e_i^2,\sigma_{ij}=h_{ij},i\neq j,i,j=1,...,p$$

由此可知：

1) a_{ij} 反映了X_i 与f_j 之间的相关。

$a_{ij}=Cov(X_i,f_i)$，且当$Var(X_i)=1$ 时$a_{ij}=\rho(X_i,f_j)$，因此通常称a_{ij} 为第i个变量X_i在第j个公共因子f_j上的载荷量，A为公因子$f_1,...,f_q$的载荷矩阵。

2) $h_i^2,i=1,...,p$ 反映了公因子对X_i的影响作用，称此为公因子方差或称共性估计值。

3) $g_k^2=\Sigma_{i=1}^p a_{ik}^2,k=1,...,q$ 反映了第k个公共因子f_k对X的各分量X_i的方差贡献之和，是衡量每个公因子相对重要的一个尺度。

值得注意的是，以上模型只是对均值为0的随机变量而言，若X为标准化随机变量，则模型中的Σ应是相关矩阵R。

(2). 旋转后的因子模型

初始因子模型建立后，每个变量在公共因子上的载荷量往往没有很明显的差别，因此，不易对公共因子作出解释，这时，需要对载荷矩阵进行进一步简化，使得各列元素

向0和1两极分化，但保持各变量的公因子方差$h_i^2, i = 1, ..., p$ 不变，这种变换方法称为因子的旋转。

旋转的主要思想在于获得一个简单的结构，希望每个因子都对部分变量有非零载荷，以便对因子作出解释。希望每个变量也仅对部分因子有非零载荷，这样使因子之间相互不同(因为若几个因子在某个变量上均有较高载荷，则很难解释清楚这些因子各有何区别、特点)。SPSS提供了几种转换方法，最常用的是方差极大旋转法(VARIMAX)，它试图尽可能减少在一个因子上具有较高载荷的变量个数，使因子便于被解释。另外两种方法是四次极大化(QUARTIMAX) 和等方差极大化(EQUAMAX) 方法。此外，为了便于简化因子载荷矩阵，也可考虑采用斜交旋转。

a. 正交旋转(othogonal rotation)

在初始因子模型中，若在A阵后面乘上$q \times q$的正交阵Γ，f 的前面乘上Γ'，此时，相当于对公共因子进进行正交旋转，正交旋转后的因子模型为：

$$X = Af + \varepsilon = A\Gamma\Gamma' f \equiv A^\star f^\star + \varepsilon$$

$A^\star$ 中的元素向0和1两极分化，$f^\star$为旋转后的公因子便于解释，因为

$$f^\star = \Gamma' f, E(f^\star) = \Gamma' E(f) = 0$$

$$Var(f^\star) = \Gamma' Var(f)\Gamma = I_q$$

$$Cov(f^\star, \varepsilon) = Cov(\Gamma' f, \varepsilon) = \Gamma' Cov(f, \varepsilon) = 0$$

所以，旋转后的公共因子也是不相关的。

b. 斜交旋转(oblique rotation)

若在公共因子之前乘上的非奇异矩阵，此时，相当于对公共因子进行斜交旋转，斜交旋转后的因子模型为$X = Af + \varepsilon = AB^{-1}Bf + \varepsilon = AB^{-1}f^{\star\star} + \varepsilon$

由于

$$E(f^{\star\star}) = E(Bf) = BE(f) = 0$$

$$Var(f^{\star\star}) = Var(Bf) = BB'$$

不一定为单位阵。

因此，公共因子通过斜交旋转后，变为相关的因子，尽管如此，斜交旋转常产生比正交旋转更有用的模型。

(3). 因子得分模型

无论是初始因子模型，还是旋转后的因子模型，它们都是用可观测变量的线性组合来表示公共因子，并计算这些公共因子的估计值，这种估计值叫做因子得分。对第k个观测个体，第j个因子的得分估计为：$f_{jk} = \Sigma_{i=1}^{p} w_{ji} X_{ik}$为第$k$ 个观测个体第i 个变量的标准化值，w_{ji} 是第j 个因子第i 个变量的因子得分系数，$W = \{w_{ij}\}$称为得分矩阵。

SPSS FACTOR 提供的因子得分系数估计的方法有：Anderson-Rubin 方法、回归方法和Bartlett 方法。

3. 参数估计

估计载荷矩阵和特殊因子的协差阵常用的方法常有以下几种：

(1). 主成分分析法。从p维可观测变量$X_{p\times 1}$的观测数据阵，可得到的样本协差阵，假设由大到小排列的特征根所对应的正则化特征向量为$l_1,...,l_p$，则当最后$p-q$个特征根较少时，S可近似地分解为

$$S=\lambda_1 l_1 l_1'+...+\lambda_q l_q l_q'+D=AA'+D$$

上述的A和D即为因子模型中载荷矩阵和特殊因子协方差阵的估计。

由于A中第j列元素与主成分的函数只相差一个常数，$\sqrt{\lambda_j}, j=1,...,q$，故这个估计方法通常称为主成分分析法。当量纲不同时，可把原变量标准化，类似地从相关阵R出发求得A和D的估计，将其进行谱分解而获得特征向量$V_1,V_2,...,V_p$及其特征根$\lambda_1\geq\lambda_2\geq ...\lambda_P\geq 0$，因此对事先设定的临界值$t$，若$q$ 满足：$\sum_{i=1}^{q}\lambda_i/\sum_{i=1}^{p}\lambda_i\geq t \quad \sum_{i=1}^{q-1}\lambda_i/\sum_{i=1}^{p}\lambda_i < t$

则因子载荷矩阵估计为：

$$A=(\sqrt{\lambda_1}V_1,\sqrt{\lambda_2}V_2,...,\sqrt{\lambda_S}V_q)$$

V_i是λ_i 相应的特征向量。特殊因子方差估计为：$\hat{\delta}_i^2=1-\Sigma_{j=1}^{q}a_{ij}^2$，公共因子方差为$\hat{h}^2=\Sigma_{j=1}^{q}a_{ij}^2$

(2). 主因子分析法。也称主轴析因法，这是对主成分法所作的一种修正，从相关阵出发求得A与D的估计。由于标化变量的协差阵即为原变量的相关阵R，因而有$R=AA'+D$，而且若有特殊因子方差的一个初始估计$\hat{\delta}^2$，则由$R-D=AA'$可知先验公因子方差的估计为，$\hat{h}_i^2=1-\hat{\delta}^2$。

记$R^\star=R-D$(称之为约相关阵)，则$R^\star$的对角元是$h_i^2, i=1,...,p.$ 由$R^\star$可得其特征根$\lambda_1^\star\geq\lambda_2^\star\geq ...\geq\lambda_p^\star>0$，取前$q$个较大的特征根，并计算其对应的特征向量，则$R^\star$可近似地分解成$R^\star=AA'$，其中$A=(\sqrt{\lambda_1}l_1,...,\sqrt{\lambda_q}l_q)$，由此可求得

$$\delta_i^2=1-\sum_{j=1}^{q}a_{ij}^2, i=1,...,p, D=diag(\delta_1^2,...,\delta_p^2)$$

在实际应用中，由于D是未知，用D的初始估计来求得A和D的估计也是一个近似解，因此，常用迭代主因子法来获得一个更好的解，即用上面得到的D 作为特殊因子协差阵的估计，重复上述步骤，直到得到稳定解，当特殊因子方差的初始值为0时，主因子分析法即为主成分分析法。

若已知先验公因子方差h_i^2的初始估计值，同样也可推得A和D 的主因子解，常用的初始估计有以下几种方法：

取h_i^2为第i个变量与其它所有变量的多重相关系数的平方

取h_i^2为第i个变量与其他变量相关系数绝对值的最大值

取$h_i^2=1$(此时，主因子法等于主成分法)

(3). 极大似然法。假定$f\sim N_q(0,I_q),\varepsilon\sim N(0,D)$，随机变量$X_1,...,X_n$ 为来自正态总体$N_p(\mu,\Sigma)$的简单随机样本，则样本似然函数为$L(\mu,\Sigma)$，取$\mu=\overline{X}$(样本均值)，$\Sigma=AA'+D$，

则$L(\mu,\Sigma)$为A、D的函数，使似然函数达到最大，A、D 从式$diag(S) = diag(AA' + D)$和$A = S(AA' + D)^{-1}A$求得，其中S为样本协差阵。

(4) 广义最小二乘法。载荷矩阵A和特殊方差矩阵D由极小化下列目标函数产生：$tr(S-\Sigma)'H(S-\Sigma)$，其中$\Sigma = D + AA', H = \Sigma^{-1}$或为其相合估计，S 为样本协方差阵。

(5) 未加权最小二乘法。与最小二乘法类似，但用于极小化的目标函数不作加权。

除了上述因子提取方法外，还有α方法和象因子法。

4. 模型的检验

因子模型建立以后，可用似然比检验的方法检验其是否合适：

$H : \Sigma = AA' + D$ 相对于$A : \Sigma \neq AA' + D$

若$\lambda = |S|^{n/2}/|\Sigma|^{n/2}$ 的值很小，则似然比检验拒绝H，其中的S为样本协差阵，由渐近理论可知-2ln λ 的分布是自由度为f的χ^2分布，其中$f = 0.5[(p-q)^2-(p+q)]$。Bartlett(1951)建议用如下的近似值

$$-[n-1-(1/6)(2p+5)-(2/3)q]\ln(|S|/|\Sigma|) = \chi_f^2$$

若$\chi_f^2 \geq \chi_{f;\alpha}^2$，则拒绝$H$，此处$\chi_{f;\alpha}^2$为显著性水平为$\alpha$，自由度为$f$的$\chi^2$临界值，若拒绝H，则认为所选的模型不合适，必须增加一个公共因子以重新求得载荷矩阵A的估计，然后再次检验模型，重复这个步骤，直到模型合适为止。

由于因子分析的目的之一就是试图获得能够解释变量关联的因子，因此在一个合适的因子模型中，变量必须相关。SPSS FACTOR 提供了下列考察变量间相关的途径。

(1) Bartlett 检验：用以检验关于相关阵为单位阵的假设，若不能拒绝该假设，则不宜使用因子模型。

(2) 偏相关系数：当其余变量的线性影响消除后，两个变量之间的偏相关系数应当较小，这可用反象相关(anti-image correlation) 来刻画。

(3) Kaiser-Meyer-Oklin (KMO) 指数：

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\left[\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} b_{ij}^2\right]}$$

其中r_{ij} 是变量i 和j 的样本相关系数，b_{ij} 是度量i 和j 的偏相关系数。若KMO 值较小，说明不宜做因子分析，根据Kaiser 的划分，$KMO > 0.9$ 为优秀，$0.8 < KMO < 0.9$ 为良好，$0.7 < KMO < 0.8$ 为中等，$KMO < 0.5$ 不能接受。

因子分析通常的步骤是：⑴计算并考察所有变量所构成的相关矩阵；⑵提取能代表数据的因子；⑶通过旋转变换使因子更具可解释性；⑷对每个观测个体计算因子得分。

确证型因子分析大致步骤为：设计一个理论模型，是构造因果关系的路径图并将其转化为一系列结构方程模型和度量模型，选择矩阵类型和模型估计，最后是对所识别的模型进行评价，看其适合度如何，进行模型解释和模型修正。

R.A. Johnson [18]建议因子分析用以下步骤， 1.首先进行主成分因子分析，绘出因子得分图并找出可疑观察，计算标准得分；进行最大方差旋转； 2.进行极大似然法因子分析和方

差极大化旋转；3.比较因子分析的结果：因子载荷聚合的模式是相同的吗？绘图比较主成分法和极大似然法的得分； 4.对于其他数目的共因子重复上述步骤，看是否有其它因子对于数据解释有用； 5.把大的数据集分成两部分，分别进行分析。

SAS 的因子分析过程为FACTOR，使用CALIS进行确证型分析。BMDP用4M进行因子分析。

§2.3.6 典型相关分析

1. 是研究两组随机变量间的相关关系的一种方法，其中，每组变量可能包含有多个变量。设$X=(X_1,...,X_p),Y=(Y_1,...,Y_q)$分别是$p$维和$q$ 维随机向量(假定$p\le q$)，典型相关分析就是要研究X与Y之间的相关性，当$p=q=1$时，X与Y之间的关系的大小用相关系数去衡量。当$p=1$，$q=n$ 时，X 与Y之间的关系大小用复相关系数去衡量，当$p\neq 1,q\neq 1$时，X与Y之间的相关大小用典型相关系数去衡量。

2. 总体典型相关

   令
   $$E\begin{pmatrix}X\\Y\end{pmatrix}=0,Cov\begin{pmatrix}X\\Y\end{pmatrix}=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$$
   其中Σ_{11},Σ_{22}分别为$p\times p,q\times q$ 阶正定矩阵，Σ_{12}为$p\times q$阶矩阵，且$\Sigma_{12}=\Sigma_{21}'$我们用X与Y的线性组合$\alpha'X$与$\beta'Y$之间的相关性来描述X与Y 之间的相关性。

   X与Y之间的第一典型相关(系数)为X的线性组合$\alpha'X$与Y的线性组合$\beta'Y$二者之间的极大相关，也就是
   $$\rho_1=\frac{\alpha_1'\Sigma_{12}\beta_1}{\sqrt{(\alpha_1'\Sigma_{11}\alpha_1)(\beta_1'\Sigma_{22}\beta_1)}}=\max_{\alpha,\beta}\frac{\alpha'\Sigma_{12}\beta}{\sqrt{(\alpha'\Sigma_{11}\alpha)(\beta'\Sigma_{22}\beta)}}$$
   可以验证ρ_1^2即为$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$的最大特征根，$\alpha_1$为相应特征向量，而$\beta_1$为$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$的最大特征根所对应的特征向量，称$V_1=\alpha'X,W_1=\beta'Y$ 为第一对典型变量，α_1,β_1为典型系数或典型权数，称方差为1的典型变量的系数为正则化典型系数。

   类似地，因$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$与$\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 非负定，且$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$与$\Sigma_{11}^{-0.5}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-0.5}$有相同特征根，若令$A=\Sigma_{11}^{-0.5}\Sigma_{12}\Sigma_{22}^{-0.5}$, 记$\rho_1^2\ge\rho_2^2\ge,...,\ge\rho_p^2$ (假定$p\le q$)为AA'和$A'A$的有序特征根，$\alpha_1,\alpha_2,...,\alpha_p$为对应于$AA'$的有序特征根的$p$维特征向量，$\beta_1,\beta_2,...,\beta_p$为对应于$A'A$ 的有序特征根的q维特征向量，则称$V_i=\alpha_i'X,W_i=\beta_i'Y,i=1,...,p$ 为第i对典型变量，它们之间的相关为第i个典型相关。注意V_i 与W_i 各自的变量之间不相关并且
   $$Var(\alpha_i'X)=Var(\beta_i'Y)=1,Cov(\alpha_iX,\beta_iY)=\rho_i,i=1,2,...,p$$

3. 样本典型相关

   在实际问题中，Σ往往是未知的，故要通过样本来估计。假定从总体中抽取一个大小为n的样本($n>p+q$)，每个样本有X,Y两组指标，若S为其样本协差阵，则将S分割如$S=\begin{pmatrix}S_{11}&S_{12}\\S_{21}&S_{22}\end{pmatrix}$，其中$S_{11},S_{12}$ 分别是$p\times p$ 和$q\times q$ 阶矩阵，$S_{12}=S_{21}'$。

   令$\lambda_1^2\ge\lambda_2^2\ge...\ge\lambda_p^2$为$S_{11}^{-1}S_{12}S_{22}^{-1}S_{21}$的有序特征根(假定$p\le q$)则称$1\ge\lambda_1\ge\lambda_2...\ge\lambda_p\ge 0$为样本典型相关(系数)。

4. 典型相关的显著性检验

求出典型变量对和典型相关系数后，把具有显著性意义的典型相关系数所对应的典型变量对保留下来，并给予合理的解释，若第i个典型相关系数近似为0，那么,这一对典型变量对于解释原来两组变量间的相关性就没有意义，因此，必须对ρ_i进行显著性检验，若$(X \quad Y)' \sim N_{p+q}(0,\Sigma)$，则可以对典型相关系数作$\chi^2$检验，检验假设为：

$H_{0i}:\rho_i=0, i=1,...,p$

检验以上假设可用Bartlett关于大样本χ^2的统计量，取统计量

$$Q_i=-[n-i-0.5(p+q+1)]\Sigma_{k=i}^{p}\ln(1-\lambda_k^2)$$

其中λ_k是由样本观测数据得到的第k个典型相关系数

对较大的样本量n，在H_{0i}为真时$Q_i \sim \chi^2(f_i)$, 其中$f_i=(p-i+1)(q-i+1)$，对给定的显著性水平α，当$Q_i > \chi_\alpha^2(f_i)$时，拒绝H_{0i}，即认为$\rho_i \neq 0$，于是再对ρ_{i+1}作检验，依次下去，若某一个$Q_j < \chi_\alpha^2(f_j)$, 则接受原假设，认为$\rho_j=0$，则有$\rho_{j+1}=\rho_{j+2}=...=\rho_p=0$。注意应首先对$\Sigma_{12}=0$ 进行检验。

5. 分析

在得到典型相关变量对后，可以用各变量的载荷反映其在相应的典型相关变量对中的作用。

载荷是变量组$X=(X_1,X_2,...,X_p)'$的任一指标$X_i, i=1,...,p$ 与其线性组合$V_j=\alpha_j'X$的相关系数，$\gamma_{X_iV_j}=\frac{Cov(X_i,V_j)}{\sqrt{Var(X_i)}\sqrt{Var(V_j)}}, i,j=1,...,p$ 称为X_i在V_j中的载荷。类似地，Y_i与W_j的相关系数：$\gamma_{Y_iV_j}=\frac{Cov(Y_i,W_j)}{\sqrt{Var(Y_i)}\sqrt{Var(W_j)}}, i,j=1,...,q$ 称为Y_i在W_j中的载荷。

现虑一个$n\times p$阶矩阵X，分块为g个$n_j\times p$阶阵$X^j, j=1,...,g$，X^j 的n_j个行来自群体$\Pi_j, j=1,...,g$。用$n\times(g-1)$ 阶矩阵Y表示X的分群标志：

$$y_{ij}=\begin{cases}1 & x_i\in\Pi_j;\\ 0 & x_i\notin\Pi_j\end{cases} \qquad i=1,...,n, j=1,...,g-1$$

可把判别分析问题化为典型分析问题，首先引进指示变量矩阵Y，然后求X与Y的最大典型相关系数λ_1和相应的典型向量a_1，即可得到判别函数$Y=a_1X$。

SAS 典型相关分析过程为CANCORR。SPSS 有专用的宏定义CANCORR，其使用方法为：CANCORR SET1=变量表1 /SET2=变量表2。BMDP相应的程序是6M。

§2.3.7 判别分析

1. 是根据待判个体的某些特定指标的观测值判断其类别归属的统计分析技术。从数学上看，就是对具有分布函数$F_i(x)$ 的k 个母体$G_i, i=1,...,k$ 判定给定的待判个体x 来自那个母体。常见的判别分析方法有Fisher 线性判别、Bayes 判别、距离判别、核密度法等。在SPSS/PC+ DSCRIMINANT 中的基本假定是参与判别分析的母体具有正态等方差分布，以线性判别函数和Bayes 分类规则进行判别。

以下简要介绍Fisher 线性判别函数、Bayes 分类规则以及在DSCRIMINANT 中使用的一些概念。

2. 普通判别(非逐步判别)

(a) Fisher 线性判别函数

对于p 维观测$X=(X_1,X_2,...,X_p)'$，线性判别函数可表为：

$$D(X)=\beta_0+\beta_1X_1+...+\beta_pX_p$$

Fisher 线性判别函数的系数b_i 的选取原则是使上述函数$D(X)$ 的取值尽可能在不同类别中不同，换言之，是选取使比率λ=组间平方和/ 组内平方和达到最大的系数b_i。

对观测个体X_0，代入上述Fisher 线性判别函数后，可得其判别得分$D(X_0)$。Fisher 准则只提供了确定判别系数的准则，并未涉及对个体的分类规则(allocation rule)。

(b) Bayes 分类规则

利用判别得分，DSCRIMINANT 基于Bayes 分类规则将待判个体判给某一类别。由Bayes 逆概率公式，在已知某个观测个体X_0 具有得分$D(x_0)$ 的条件下其属于母体i 的概率

$$P(G_i|D(X_0))=\frac{P(D(X_0)|G_i)P(G_i)}{\Sigma_{j=1}^{k}P(D(X_0)|G_j)P(G_j)}$$

其中$P(G_i)$ 为母体G_i 的先验概率(prior probability)。

先验概率是关于各母体的一种先验知识，其估计可由许多方法得到。对混合抽样问题，可用所抽到的每个母体中个体的比例作为对先验概率的估计。对于固定抽样问题，则只能考虑采用其他途径估计。最后，若各母体等可能出现或对其先验信息完全未知，则可对所有母体采用等先验概率。在DSCRIMINANT 中，为上述三种形式的先验概率在PRIORS 子命令中设计了相应的输入方式。

根据Bayes 分类规则，基于其判别得分$D(X_0)$，待判个体X_0将按最大后验概率$P(G_i|D(X_0))$确定其属于哪一个母体，即分类为规则：

若

$$P(G_i|D(X_0))=\max_j P(G_j|D(X_0))$$

则将观测个体X_0 (其判别得分为$D(X_0)$) 判为属于母体G_i。

(c) 训练样本、待判样本

一个典型的判别问题通常分为两步，首先根据训练样本(其所包含的观测个体的类别归属已知) 构造判别函数，其次对待判样本(不知其观测个体属于哪个母体) 或对验证样本(通过对其判别结果对错判率进行估计)进行判别。在DSCRIMINANT 中，利用SELECT 子命令可将训练样本和待判样本(或验证样本) 的个体标识出来。

(d) 标准化和典型判别函数

标准化和非标准化判别系数：非标准化系数是在变量采用原有度量单位时求得的判系数，而标准化系数(如同在回归分析中那样) 是当变量按零均值单位方差标准化后的系数。标准化的意义在于标准化矫正了由于不同测量单位所引起的偏差，从而使判别系数的大小反映出其各变量的不同重要程度。

典型判别系数：当线性判别函数按照Fisher 准则(使组间平方和与组内平方和之比最大) 求得时，它恰好等于典型相关分析中的典型相关系数，因此，它又称做典型判别系数。

(e) 多类判别函数个数

对多类判别，可建立多个判别函数，对P 个预测变量，K 类判别问题，通常有m 个函数，m 满足$m \leq \min(k-1, p)$。一般说来，在Fisher 线性判别函数中，以最大待征根对应的判别能力为最大，因此常只用它进行判别，但有时在多类判别中，问题较为复杂，也可建立多个判别函数。在DSCRIMINANT 中，采用FUNCTION 子命令决定判别函数的个数。

3. 逐步判别

逐步判别中有两个关键要素，变量选择准则和变量选择方法。

DSCRIMINANT 中提供了五种变量选择准则，可利用子命令METHOD 选择使用，下面将五个准则加以介绍。

设有k 个母体$G_1, G_2, ..., G_k$，p 个预测变量，记n_i 为第i 个母体的样本含量，$\overline{X}_{ij}$ 为第j 个母体第i 个变量的均值，$\overline{X}_i$ 为所有母体合并后第i 个变量的均值，w_{ij} 为组内协方差阵的逆阵第(i,j) 个元素，常用的变量选择选准则有以下几种：

①Rao 的V 统计量

$$V = (n-k)\sum_{i=1}^{p}\sum_{j=1}^{p} w_{ij}\sum_{l=1}^{k} n_l(\overline{X}_{il} - \overline{X}_i)(\overline{X}_{jl} - \overline{X}_j)$$

也称Lawley-Hotelling 迹，V 的渐近分布为$\chi^2_{p(k-1)}$。

显然，组内均值的差值越大，V 就越大，因此，考察一个变量贡献大小可看将其加入到模型后V 的增量有多大，其增量显著性检验可基于χ^2分布进行。

②Mahalanobis 距离(马氏距离)

母体a 和母体b 之间的样本距离定义为：

$$D^2_{ab} = (n-k)\Sigma_{i=1}^{p}\Sigma_{j=1}^{p} w_{ij}(\overline{X}_{ia} - \overline{X}_{ib})(\overline{X}_{ja} - \overline{X}_{jb})$$

变量选择时首先计算所有母体之间的两两距离，其次，对具有最小距离的两个母体，选择具有最大D^2 的变量入选模型。

③组间F 统计量

基于马氏距离，可构造两母体均值相等的零假设检验，其对应的统计量为：

$$F = \frac{n-1-p}{p(n-2)}\frac{n_a n_b}{n_a + n_b}D^2_{ab}$$

F 值可用于变量选择,在每一步,选择具有最大F 值的变量入选。

④未解释变异和

由于马氏距离与回归分析中的R^2成比例，即：$R^2 = cD^2$，对于每一对母体a 和b，从回归分析角度来看未解释变异为$1 - R^2_{ab}$，其中R^2_{ab} 为复相关系数的平方。在变量的选择时，选取使“未解释变异和”最小的变量入选。

⑤Wilks 的λ统计量

选择使Wilks λ最小的变量入选模型是最为常用和读者最为熟悉的一种选择准则。

在DSCRIMINANT 中提供三种常用的变量选择的算法：前进法、后退法和逐步选择法。其原理与逐步回归分析一样，只是变量入选的准则不同。三种变量选择法可利用ANALYSIS 子命令任选其一。

下面以逐步选择法为例说明其实施步骤：(i) 在所有变量$x_1, ..., x_p$ 中挑选对选择准则而言最大可能接受值的变量入选；(ii) 当第一个变量入选后，对尚未入选变量，根据准则重新计算评判，选出下一个具有最大可能接受值的变量入选；(iii) 对已入选变量，应重新评价以考察其重要程度是否因其他入选变量的进入而发生变化，及时删除满足删除准则的变量；(iv) 如此继续，直至既不能入选新变量又不能删除已入选变量为止。此时，利用最终选入的变量构成判别函数。

在上述算法过程中，DSCRIMIANT 将在每一步显示一个当前已入选变量表，在最后一步，给出一个综览表，以说明各次入选和删除变量的过程以及统计量的显著水平。

SAS 的判别分析过程为DISCRIM，典型判别分析的过程是CANDISC。

§2.3.8 聚类分析

1. 是在一组观测个体类别归属未知的条件下，根据其观测指标在数值上的特征进行归类的统计分析技术。常见的聚类方法有：系统聚类法，动态聚类法、分解法、有序聚类法等。在进行聚类分析时，应明确：使用哪些变量？变量的距离如何确定？使用哪些准则进行类的归并？SPSS/PC+ CLUSTER 命令提供了系统聚类法，以下略加介绍。

2. 相似系数和距离

观测个体的归类是基于对个体之间关系的某种度量进行的。常用的度量有相似系数和距离，即根据观测个体之间的相似或距离远近进行归类，SPSS/PC+ CLUSTER 中采用的相似系数和距离介绍如下：

设$X' = (x_1, x_2, ..., x_m)$，和$Y' = (y_1, y_2, ..., y_m)$ 为两个具有m 个变量指标的观测个体。以下的求和与取最大的下标范围均是1.,,,m。

①SEUCLID (平方欧氏距离) $d(X,Y) = \Sigma(x_i - y_i)^2$

②EUCLID (欧氏距离) $d(X,Y) = \sqrt{\Sigma(x_i - y_i)^2}$

③COSINE (夹角余弦——相似系数) $C(X,Y) = \Sigma x_i y_i / \sqrt{\Sigma x_i^2 \Sigma y_i^2}$

④BLOCK (绝对值距离) $d(X,Y) = \Sigma|x_i - y_i|$

⑤CHEBYSHEV (切比雪夫距离) $d(X,Y) = \max|x_i - y_i|$

⑥POWER(P,R) (绝对值幂距离) $d(X,Y) = \sqrt[r]{\Sigma|x_i - y_i|^p}$

其中，当r=p 时为明考夫斯基(Minkowski) 距离，r=p=1 时为绝对值距离，r=p=2 时为欧氏距离。

3. 系统聚类法及类间距离

系统聚类法的基本思想是，首先将n个待聚观测个体看作n小类，然后规定个体之间的距离或相似系数(SPSS/PC+ CLUSTER 中选择上述六种之一) 以及规定各类之间的距离，选择距离最小的两类作为新的一类，然后重新计算所有各类之间的距离，再次选择距离最小的两类并为一类，如此继续，直至并为一类为止。这一归类过程可用一张聚类图或谱系图形象地表示出来。

因此，系统聚类法实施的另一个重要前提是定义度量各类之间距离的方法，用$G_1, G_2, ...,$表示它们分别有观测数$n_1, n_2, ...,$的类，d_{ij} 表示观测个体i与j 的距离，D_{pq} 表示G_p 与G_q 的距离，SPSS/PC+ CLUSTER 中所提供的度量类间距离方法有：

①类间平均法(BAVERAGE) $D_{pq}^2 = (1/n_p n_q)\Sigma d_{ij}^2, i \in G_p, j \in G_q$

②类内平均法(WAVERAGE) $D_{pq}^2 = [\Sigma d_{ij}^2 + \Sigma d_{kl}^2]/[C_{np}^2 + C_{nq}^2], i, j \in G_p, k, l \in G_q$

③最短距离法(SINGLE) $D_{pq} = \min d_{ij}, i \in G_p, j \in G_q$

④最长距离法(COMPLETE) $D_{pq} = \max d_{ij}, i \in G_p, j \in G_q$

⑤重心法(CENTROID) $D_{pq} = d_{\overline{x}_p \overline{x}_q}$ 其中$\overline{x}_p$ 和$\overline{x}_q$ 分别为类G_p 和G_q 的均值重心且必须用平方欧氏距离，若将G_p 和G_q 合并为新的一类G_r 时，其新的(均值)重心定义为：$x_r = (1/n_r)(n_p \overline{x}_p + n_q \overline{x}_q), n_r = n_p + n_q$

⑥中间类法(MEDIAN) $\tilde{D}_{pq} = \tilde{d}_{XX}$

应当注意的是，与重心法不同之处在于在中间类法中，当G_p 和G_q 并为一类时，其新的(中间类法) 重心定义为：$\tilde{X} = 0.5(\overline{X}_p + \overline{X}_q)$

⑦WARD 法设将 n 个观测个体分成 k 类，$G_1, G_2, ..., G_k$，用x_{it} 表示G_t 中的第i 个样品(注意x_{it} 为m 维向量)，首先计算类G_t中观测个体的离差平方和：$S_t = \Sigma_{i=1}^{n_t}(\overline{x}_{it} - x_t)'(\overline{x}_{it} - x_t)$. 当$k$固定时，要选择使$S = \Sigma_{t=1}^{k} S_t$极小的分类。

多个样本点聚类时，如果数目很大，宜采用K-means聚类。

SAS 使用CLUSTER和FASTCLUS、ACECLUS进行系统聚类和K-means聚类，在BMDP中相关的模块为1M、2M、3M和KM。

§2.3.9 分类数据分析

与连续数据分析有许多类似之处，这里主要介绍对数线性模型和logistic 回归分析，对数线性模型在第13章也有一些讨论。

1. 对数线性模型

对数线性模型与回归模型有相似之处，在对数线性模型中，所有被用于分类的变量均作为自变量，而因变量为交叉表中各点上的理论频数。

例如，对于变量X 和Y 的$r \times s$ 列联表第i行第j 列格的模型为：

$$\ln(m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

其中λ_i^X 和λ_j^Y 称为主效应，λ_{ij}^{XY} 称为交互效应，m_{ij} 为格点(i, j) 的理论频数。

对数线性模型可分为分层(hierarchical) 和不分层两大类，其中分层模型更为常用，它是指如果模型中出现了某一组变量的某种交互项，则必存在这些变量所有可能组合的低阶项。

例如，在$r \times s$列联表中，若分层模型中出现λ_{ij}^{XY}，则必出现λ_i^X 和λ_j^Y，若模型不包括λ_j^Y，则必不包括λ_{ij}^{XY}。

SPSS 的HILOGLINEAR 用于处理分层对数线性模型，以下对它作一介绍。

分层对数线性模型有两个重要特例：饱和模型指包含了所有变量的主效应和这些变量所有可能的交互效应项的对数线性模型；独立变量模型指不包含变量的交互效应

项的对数线性模型称为独立变量模型。例如上例中不包含λ_{ij}^{XY}时则为独立变量模型。显然，饱和模型和独立变量模型是一般分层对数线性模型的两个极端情形。

(a) 模型参数及估计

对数线性模型参数有一个重要性质：在其余变量不变的条件下，任意一变量的所有效应之和为零。如上述$r \times s$列联表中：

$$\Sigma_{i=1}^{r}\lambda_i^X = 0, \Sigma_{i=1}^{s}\lambda_i^Y = 0, \Sigma_{i=1}^{r}\lambda_{ij}^{XY} = 0, \Sigma_{i=1}^{s}\lambda_{ij}^{XY} = 0$$

$i = 1, ..., r, j = 1, ..., s$

上述性质在对数线性模型的参数估计时非常有用。在HILOGLINEAR 中，只提供与参数自由度相等的参数估计，其余参数依上述性质自行推导。

如上所述，对于分层对数线性模型，指定了最高阶交互效应项，实际上就指定了整个模型的结构。在HILOGLINEAR 中，利用DESIGN了命令指定饱和模型生成类中的某模型的最高阶交互项，以指定该模型的结构。

HILOGLINEAR 在给出参数的估计值时，是按照上述最高阶交互项中最左边的变量顺序每取一值，最右边变量依次取其与自由度相应的所有可能值的方式给出参数估计。

例如，在变量为X和Y的3×3表中，若指定最高阶交互项为XY，则顺序给出下列参数的估计。

$\lambda_{11}^{XY}, \lambda_{12}^{XY}, \lambda_{21}^{XY}, \lambda_{22}^{XY}, \lambda_1^X, \lambda_2^X, \lambda_1^Y, \lambda_2^Y$

其余参数均可由约束方程导出。

HILOGLINEAR 采用迭代比例拟合算法为多维列联表拟合分层对数线性模型。对迭代的控制可通过HILOGLINEAR中的CRITERIA 子命令设置估计精度和最大迭步数实现。

(b) 检验

常见的关于拟合的假设检验有以下两种：χ^2 拟合优度检验

$$\chi^2 = \sum\sum \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \sim \chi^2_{(r-1)(s-1)}$$

似然比χ^2 检验

$$2\sum\sum n_{ij} \ln(\frac{n_{ij}}{m_{ij}})$$

式中n_{ij} 是实际观测频数，$\hat{m}_{ij}$ 为其相应的拟合值，下标i, j的求和是对表中所有格点而言，在大样本情况下这两种统计量等价。

(c) 残差与诊断

模型拟合好坏还可通过考察基于模型产生的期望格点与观测格点的差值—即残差进行，易见，当模型拟合较好时，残差应较小。与回归分析类似，考察标准化残差—它由残差除以其标准差构成—要比直接考察残差更为合理。

标准化残差=[观测格点数- 期望格点数]/$\sqrt{\text{期望格点数}}$

若模型是合适的，则标准化残差将有渐近标准正态分布。因此，当标准化残差绝对值大于1.96，表明拟合很不好。残差的诊断类似于回归分析中的情形。(1) 观察“标准化残差—格点观察频数”图和“标准化残差——格点期望频数”图；在HILOGLINEAR 中，通过PLOT 子命令中的RESID 作上述图形显示。(2) 观察正态概率图，该图是以正态概率为纵柚，以标准化残差为横轴构成的图。显然，当模型拟合合适时，图形应为一条直线。在HILOGLINEAR 中，通过PLOT 子命令中的NORMPROB 产生上述图形显示。关于残差诊断的进一步讨论，可参见“回归残差分析”一节。

(d) 模型选择

⑴模型分块选择法

类似于回归分析中利用复相关系数的改变量来刻划新增变量的贡献，在分层对数线性模型中通常利用似然比χ^2统计量检验一个模型在增加某新的效应项时其对应的贡献大小，从而决定模型的选择。

在HILOGLINEAR 中将自变动给出两类假设检验：关于所有k阶及其更高阶效应为零的假设和关于k阶效应为零的假设。

⑵后退删除法

由于在对数线性模型中后退法比前进法对模型选择更合适，所以HILOGLINEAR只提供了后退法。其初始模型可以是任何分层对数线性模型，程序将计算所有最高阶交互项卡方值的观测显著性水平，对于其观测值水平大于保留准则值的效应项，若其删除导致似然比卡方最小显著性改变，则可给予删除。注意在这一步中，为确保是分层模型，只考察对应于生成类的效应项。类似地，在删除某个效应项基础上，再次比较观测显著性水平和似然比卡方以决定其它的效应项是否删除。后退删除法通过METHOD 子命令中的BACKWARD 选择项指定。

2. LOGISTIC 回归

是一种用于结果为二分类数据的多元分析方法。流行病学分析致病因素与结果的关系，有两种方法经常使用，一种称定群或队列研究，在调查开始时，将人群分成两组，分别暴露于某种因素，另一组用做对照，此后随访观察他们在一个时期内的结局，称为前瞻性定群研究；否则若用现有对象，追溯所研究暴露因素的影响，就是历史性定群研究。第二种是调查病人和非病人暴露于于某危险因素的方法，称做病例对照研究，若病例或样本均是人群中的随机样本，则病例与对照使用的样本数目不一定相等，称为成组病例对照研究；否则对每一病例按特定条件如年龄、性别、住址等找出对照，就是配对病例—对照研究，对照可以是一个或多，称做1: 1 配对和1:M 配对。LOGISTIC 回归是分析这一类数据的有力方法，SPSS/PC+ 用LOGISTIC REGRESSION 命令，SAS 和SYSTAT 也都可以做成组或等级分组的LOGISTIC 回归分析，SAS 还可进行条件LOGISTIC 回归。现从一般多元回归分析出发，模型对二分类数据的因变量Y_i 预计取值应是一个$0 \sim 1$ 的概率P_i，通常的回归效果必然不佳，现做LOGIT 变换$\ln(\frac{P}{1-P})$，有：

$$\ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

或

$$P_i = \frac{\exp(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{k} \beta_i x_{ij})}, \quad i = 1, ..., n$$

即成组的LOGISTIC 模型。其中$(k+1)$ 维向量$(\beta_0,\beta_1,...,\beta_k)$ 是待估计参数，模型对于$\beta'X$ 是单调的。现设某个二分类的反应变量Y 在事件发生时取值为$'A'$，未发生时取值为$'B'$，又设有一个做为回归变量的危险因素X在出现时取值为1，不出现时取值为0，根据LOGISTIC 模型

$$P(Y ='\ A'|X=1)=\frac{\exp(\alpha+\beta)}{1+\exp(\alpha+\beta)}$$

$$P(Y ='\ B'|X=0)=\frac{\exp(\alpha)}{1+\exp(\alpha)}$$

其中α是截距，β是回归系数，对于具有危险因子的个体来说，事件的比数(odds) 定义为$\frac{P(Y='A'|X=1)}{P(Y='B'|X=1)}=\frac{P(Y='A'|X=1)}{1-P(Y='A'|X=1)}=\exp(\alpha+\beta)$；类似地，对于没有危险因素的个体比数为$\exp(\alpha)$，比数比(odds ratio)就是这两个比数的比，即$I=\frac{\exp(\alpha+\beta)}{\exp(\alpha)}=\exp(\beta)$。由此可见回归系数$\beta$表示了因素由0 变到1 的对数比数的变化。当因素的编码是a 与b 时$I=\exp((b-a)\beta),\beta$则是因素变化一个单位时对数比数的变化。在SAS PROC LOGISTIC 中，因素的比数比由$\hat{I}=\exp(\hat{\beta})$ 给出，可信限由$\exp(\hat{\beta}\pm z_{\alpha/2}s(\hat{\beta}))$ 算出。

LOGISTIC 似然函数正是Bernoulli 变量的似然函数，由下式表达：

$$\prod_i P_i^{Y_i}(1-P_i)^{1-Y_i}=\exp(\sum_i Y_i(\beta_0+\sum_{j=1}^k\beta_i x_{ij}))\prod_i(1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij}))^{-1}$$

上式表明，LOGISTIC 模型是$(k+1)$ 个参数的指数簇分布，具有充分统计量$s_0=\sum y_i,s_1=\sum y_i x_{i1},...,s_k=\sum y_i x_{ik}$。对数似然函数关于$\beta_j$的一阶导数为$V_j(\beta)=s_j-\sum x_{ij}\frac{\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}{1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}$, 即$s_j$ 的观察值减去期望值；β 的协方差阵由信息矩阵的逆给出，信息矩阵是：

$$M_{jl}(\beta)=\sum x_{ij}\frac{\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})}{(1+\exp(\beta_0+\sum_{j=1}^k\beta_i x_{ij})^2},j=0,1,...,k,l=0,1,...,k$$

从$\beta_0=\ln(\frac{s_0}{n-s_0}),\beta_1...,\beta_k=0$ 开始，用$\beta+[M(\beta)]^{-1}V(\beta)$ 不断修正至收敛。

针对回归系数的检验，根据方程估计中的正态近似，常用正态 z 检验(Wald's test)，计分检验(score test) 以及似然比检验(likelihood ratio test)，有关说明可详参Rao, C.R. (1973)。

计分检验统计量由对数似然函的一阶与二阶偏导计算，正态性检验是利用参数估计值与其近似标准误的比值，Wald's 检验则是该量的平方。

设有模型$logit\quad p(Y=1|X)=\beta_0$ 及$logit\quad p(Y=1|X)=\beta_0+\beta_1X_1$，它们的对数似然函数分别为$L(\hat{\beta}_0)$ 和$L(\hat{\beta})$，针对假设$\hat{\beta}_1=0$ 计算$-2(L(\hat{\beta}_0)-L(\beta))$，该统计量具有近似自由度为1 的$\chi^2$分布。

在SAS PROC LOGISTIC 中引入了Hosmer-Lemeshow 统计量，其原理如下：在求得回归系数后，代入原公式得到每个反应的概率，把这些概率从小到大进行排列，然后按如下规则分成大约十组：记N 为观察个体总数，M 是每个亚组的个体数$M=[0.1N+.5],[x]$是x 的整数部分，若原始数据是未分组的，则各组都有不同M 取值的个体；若资料是分组的，则观察个体按观察的组界进行分组，对应第一个观察的个体分到第一组，设第一个观察有n_1 个体，第二个观察有n_2个体，当$n_1<M$ 及$n_1+[0.5n_2]\le M$ 时第二个观察也

放到第一组，一般来说若$(j-1)$ 个观察已放到第k 组，而第k 组有c 个个体，第j 个观察的个体在$c \le M, c+[0.5n_j] \le M$ 时被放到第k 组，否则放到下一组，另外若最后一组的个体数目不超过$[0.5 \times N]$，则把最后两组合并。如此分得g 组，通过观察与计算频数的$2 \times g$表，计算Hosmer-Lemeshow 拟合优度检验，统计量为：

$$\chi^2_{hw} = \sum_{i=1}^{g} \frac{(O_i - N_i\overline{\pi}_i)^2}{N_i\overline{\pi}_i(1-\overline{\pi}_i)}$$

其中N_i 是第i 组中的个体数，O_i 是第i 组中的事件数，$\overline{\pi}_i$ 是第i 组的事件结果的平均估计概率，统计量与$\chi^2(g-2)$分布进行比较。

SAS 对于模型中包含分类变量的情况，可输出多个系数，克服了把分类变量作为连续型的不合理性。如种族变量race 有“1.黑人、2.白人、3.西班牙人、4.其他”样的分类，做为连续量处理则不合理，而用哑变量(dummy variable) 对变量进行编码，不同的软件编码方式不同，如：

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}$$

第一种使用参照组编码(reference cell coding)，第二种是与总均值的编码(deviation from means coding)，第三种为正交编码。在伪变量回归中，常数项表示了对照的均值或者是总观察的均值，或是组均值未加权的值，SAS 把最后一个分类作参照组，而GLIM则取第一组作参照，Hosmer, D.W.和S. Lemeshow (1989) 对此进行了说明。

SAS 分类数据分析的主要过程是FREQ和CATMOD。

3. 其它方法

对应分析(corresponding analysis)是用图形方式表达两维列联表关系的方法。行和列用图上的点来表示，点的位置表示了关系，点的坐标也是典型相关模型分数的一种方式。优点：首先，多个分类变量可以经过列联表数据简明地表达，这一手段使得研究者利用现有数据或收集一般名义的或不容易测量的数据进行分析。因为因子分析只能针对所有分析变量均为区间类型时的情形。其次，因为对应分析不仅勾画了行与列的关系，也表达了它们自身分类间的关系，比如：列的许多特征比较接近，则它们的轮廓比较相象，这样就构成了一组特征，因而与主成分分析中的因素相当。最后，也是最重要的，对应分析给行与列分类提供维数相当的表达。缺点或局限性：首先，这一方法是描述性的因而不适于假设检验。若要定量描述分类间的关系，应当使用对数线性模型等方法。对应分析最适于探索性分析。其次，同其它降维方法一样，并不存在确定一个合适维数的方法，研究者应在可以解释与数据表达的简明之间进行权衡。最后，对应分析对于异常值仍比较敏感。conjoint 分析特别用于分析应答者如何对某种想法、产品或服务的趋向性。其基本假设是消费者是在综合了这些特征之后进行评价，在这一意义上它与析因设计相当。在了解了应答者的趋向性之后，我们可以对这一倾向性进行剖析，估价单个特征的贡献多大。因此，它有别于判别分析和回归分析，因为后者只是根据多个特征与总的倾向性进行关联分析。这一区别可以称作“组合——分解(decomposional /compositional)”的不同。

§2.3.10 生存分析

1. 在医学和可靠性分析中，某个对象的观察常常与时间有关，如观察某种肿瘤由诊断到死亡的病程，某种产品从使用到失效的时间。这些数据常用生存分析来处理。不同于以往前的统计分析，生存分析研究取值大于零，并且有截尾的随机变量。

   现用T 表示失效时间，记其分布函数为$F(t) = P(T \leq t)$，表示了到时刻t 失效的概率；生存函数为$S(t) = P(T > t) = 1 - F(t) = exp(-H(t))$，表示t 时刻仍然存活的概率，其中H(t)=$\int_0^t h(u)du$称为累积风险函数。如果变量$T$的密度函数$f(t)$存在，那么风险函数$h(t) = f(t)/S(t) = f(t)/[1 - F(t)]$,它表示在时刻$t$活着的条件下，时刻$t$后失效的概率。

2. Kaplan-Meier 模型是一种非参方法，设有k个观察时间$t_1 < t_2 < ... < t_k$, 每个观察时间t_i有n_i个观察对象，其中有d_i个失效，$i = 1, 2, ..., k$, 生存函数的估计为

   $$\hat{S(t)} = \Pi_{t_j<t}\frac{n_j - d_j}{n_j}, \quad S.E.[\hat{S(t)}] = \hat{S(t)}\sqrt{\Sigma_{t_j<t}\frac{d_j}{n_j(n_j - d_j)}}$$

   其误差估计公式又称做Greenwood 公式。

3. log-rank检验

   现要对两组处理进行比较，两组分别有M 和N 个观察，看其生存情况是否相同，假设在$M + N$ 个对象的失效时间是不重复时，可依据每个时间区间把两组的情况列于表2.14：

表 2.14 生存率比较计算表

处理组	t_j 时的失效	t_{j-0}时处于风险的数目
第一组	m_j	M_j
第二组	n_j	N_j

在边缘值给定下n_j 的条件分布是超几何分布，即在有限总体$M_j + N_j$ 中有$m_j + n_j$个有某种特征，然后从N_j 的样本中观察到n_j 个具有特征。因此n_j的条件期望与方差是：

$$E_j = \frac{N_j(m_j + n_j)}{M_j + N_j}$$

$$V_j = \frac{M_J N_j(m_j + n_j)(M_j + N_j - m_j - n_j)}{(M_j + N_j - 1)(M_j + N_j)/2}$$

记$w_1, w_2, ..., w_J$ 是一些已知的常数，假设$n_1, n_2, ..., n_J$ 相互独立且具有正态分布，其矩由上式给出，则$\Sigma w_j(n_j - E_j)$ 也具有正态分布。进一步有$Q(w) = (\Sigma_j w_j(n_j - E_j))^2/\Sigma_j V_j w_j^2$服从自由度为1 的卡方分布。

当$w_j = 1$ 时，就是比例风验检验：$Q_{PH} = (O - E)^2/V, O = \Sigma n_j$ 是第二处理组有观察失效数目，E 是相应的期望值，这一检验拥有众多的名字：log- rank、Mantel-Heanszel、Generalized Savage 以及exponential order scores test。

当$w_j = M_j + N_j$时，就成为generalized Wilcoxon 检验。这一方法可以推广到多组。另外，对于风险函数的检验也是很重要的。第6章介绍了对Gehan关于白血病数据分析的结果。

表 2.15 比较多组生存情况的计算表

处理组	t_j 失效数	t_{j-0} 的失效数
0	n_{0j}	N_{0j}
1	n_{1j}	N_{1j}
.	. .	
k	n_{kj}	N_{kj}
总计	$n_{.j}$	$N_{.j}$

当组数多于两个，有类似的公式。现把各处理组的失效情况列如表 2.15 ：

$$n_j = (n_{1j}, n_{2j}, ..., n_{kj})', E_j = (E_{1j}, E_{2j}, ..., E_{kj})'$$

$$E_{ij} = N_{ij} n_{.j} / N_{.j}$$

$$V_{il} = [n_{.j} N_{ij} (N_{.j} - n_{.j})(N_{.j}\delta_{il} - N_{lj})]/[(N_{.j} - 1)N_{.j}^2]$$

其中δ_{il} 是Kronecker 记号。第k组的广义比例风险统计量

$$Q_{PH} = (O - E)' V^{-1} (O - E)$$

在H：$h_{0(t)} = h_{1(t)} = ...h_{k(t)}$ 的假设下，Q_{PH}近似服从自由度为k 的χ^2分布。

4. 参数模型最简单的是指数分布，$f(t) = \lambda \exp(-\lambda_t), h(t) = \lambda, S(t) = \exp(-\lambda_t), t > 0$。常用的参数模型列于表2.16：

表 2.16 几种生存分布的风险函数与生存函数

分　　布	h(t)	S(t)
指数(exponential)	λ	$\exp(-\lambda t)$
威布尔(Weibull)	λt^{γ}	$\exp(-\lambda/(\gamma+1)t^{\gamma+1})$
Gompertz	$\lambda \exp(\gamma t)$	$\exp[-(\lambda/\gamma)(\exp(\gamma t) - 1)]$
伽马(Gamma)	单调或定常	
对数正态(log normal)	增至最大，然后下降	

各模型的参数的估计通常采用极大似然方法。设有n个样本在某时刻观察到d个失效，记$t_i, i = 1, ..., n$为观察到的相应的失效或截尾时间，假定$t_1, ..., t_d$为失效时间，则似然函数为：

$$L = \Pi_{i=1}^{d} f(t_i) \Pi_{i=d+1}^{n} S(t_i)$$

标准误由参数的Fisher 信息矩阵而得。

对指数分布，似然函数为

$$L(\lambda) = \lambda^d \exp(\lambda \sum t_i)$$

极大似然估计为$\lambda = d/\sum t_i$，利用正态近似，有相应的检验统计量：

$$(\ln\hat{\lambda} - \ln\lambda)/\sqrt{(1/d)}$$

两个指数分布相等的检验统计量是：

$$(\ln\hat{\lambda}_1 - \ln\hat{\lambda}_2)/\sqrt{1/d_1 + 1/d_2}$$

均为正态性z检验。

生存分析的寿命表法在SAS PHREG、BMDP1L 和SPSS/PC+ 实现。

5. Cox 回归模型

Cox 模型是一种半参数或称比例风险模型(proportional hazard model, PH model)，其风险函数为$h(t;x) = \lambda(t)\exp(\beta' x), \lambda(t)$ 是一个任意的函数，称为基线风险函数。X称为协变量，β是未知参数；又因为$h(t;x_1)/h(t;x_2) = \exp[\beta'(x_1 - x_2)]$ 对所有时刻t 成立，所以是比例风险。设观察到k个失效时间$t_1 < t_2 ... < t_k$, 令x_i为其相应的协变量值，Cox偏似然函数为

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta x_i)}{\sum_{j\in R(t_i)}\exp(\beta x_j)}$$

其中$R(t_i)$是在先于时刻t_i还没有失效的样本集。参数β的统计推断基于Cox偏似然函数。

当协变量是时间的函数，则模型称时间协变量的Cox 模型。当失效出现重复时，问题要复杂，常用的处理方法如Breslow 方法、离散logistic 方法、Efron 方法等，这些方法可由SAS PHREG 和SPSS SURVIVAL 实现。

SAS生存分析的过程为LIFEREG、LIFETEST和PHREG。SPSS 过程为SURVIVAL、KM和COXREG，分别进行寿命表、Kaplan-Meier和Cox生存分析。BMDP2L可以进行固定协变量和时变协变量的Cox回归分析。

第三章　计算机应用概要

§3.1　简单历史与硬软件知识

人们探索自动计算的历史源远流长。早在公元前2000 年人们就使用了算盘，公元1642年，Blaise Pascal 创造了用于税务计算的加法器。1670年，Gottfried Von Leibniz 创造了四则和开方运算的机器。1842年，Charles Baobage 设计了微积分的机器计算，Ada Augusta 编写了它的程序。1890年，Herrnan Hollerith 设计了记录人口普查数据的系统。1937-38年，John V. Atanasoff 和他的助手Clifford Berry设计与建造了第一台电子数字计算机(Atanasoff-Berry-Computer , ABC)。1940年John Von Neumann 建议把程序与数据存于计算机，他也阐述了计算机硬件新的概念。

1946 年，J. Presper Eckert 和John W. Mauchly 设计建造了ENIAC(Elec -tronic Numerical Integrator And Computer) 计算机，它重达30 吨，共用了18,000 个真空管，占地30x55 平方英尺。UNIAC 于1951 年商业化，并成功分析了美国总统竞选的抽样数据，UNIAC揭开了现代计算机发展的序幕。1957年，John Backus 及其IBM 的同事完成了第一个Fortran 编译器。1958 年，IBM 7090 首先以晶体管为开关设备，Seymour Cray 当年建造了第一个完全的晶体管计算机CDC1604。1959年，COBOL 着手开发，数学家Grace Hopper 起了关键作用。1964 年，第一台使用集成电路的计算机IBM 360 问世。1965年DEC 引入第一台小型计算机(minicomputer)。1965年，Dartmouth 大学John Kemeny领导了BASIC 的开发。1969 年，Intel 公司的Ted Hoff 开发了Intel4004 芯片。1975年，第一台微机Altair问世。1976-77年，Steve Wozniak 与Steve Jobs 制成第一台Apple计算机；DEC VAX-11/780 问世。1978-79 年，Dan Bricklin 和Bob Frankston编制了第一个电子报表软件VicCalc 用于Apple Ⅱ 。1980 年，Bill Gates 着手研制MS-DOS。1983 年，Mitch Kapper 开发了Lotus 1-2-3，1984年Apple公司引入Macintosh，它于1988 年的产品NeXT 可以记录与处理声音。

人们把以电子管为主要逻辑元件的计算机称作第一代(1946-1957)；以晶体管为主要逻辑元件(1958-1964)为第二代，以中小规模集成电路为逻辑元件(1965 -1970)为第三代；以大规模集成电路组成第四代。工业发达国家亦组织人力和财力积极研制第五代电子计算机，计算机正向智能化发展。目前计算机主要用于数据处理和信息管理、科学计算、计算机辅助设计、过程控制和人工智能等领域。

计算机系统由硬件(hardware) 和软件(software) 组成。前者是指组成计算机各个部分的设备，后者是指使计算机运行并完成交给它完成的各种任务的一些指令、程序等。软件又分为系统软件、应用软件，后者基于系统软件开发。计算机的中心部位是中央处理机(CPU)，包括算术及逻辑单元(ALU)、输入输出控制逻辑、寄存器(registers)、控制单元(指令寄存器IR、处理器状态字PSW、栈指针SP)。在总线结构下，CPU 经总线(bus) 与存贮器、输入输出接口连接。总线有地址总线、控制总线和数据总线。主机加上外设，就构成了一个微型计算机。利用大规模集成电路技术，把中央处理器集成在一个芯片上，称为微处理器。七十年代以来，半导体存贮器逐渐取代了磁芯存贮器。其可以分为随机存贮器RAM 和只读存贮器ROM。前者用于存放各种现场的输入输出数据及中间结果，以及与外存交换信息和用作堆栈，后者只能读出，故一般用于存放固定的程序，如微型计算机的管理、监控程序、汇编程序等。把CPU、一定量的存贮器，以及输入输出接口电路，集成在一个芯片上，构成单片计算机。或把CPU、RAM 和ROM、输入输出接口装在一块印刷电路板上，即构成单板计算机。

目前，以IBM (国际商用机器公司) 的系列微型计算机及其兼容机是应用的主流。所谓兼容(compatible)是指一种计算机上的软件，可以不加改动地在另一种计算机上使用。IBM PC机于1981 年推出，之后IBM又相继推出PC/II 、PC/XT (Extended Technology，扩展型)、PC/AT(增强型)、XT/370、3270-PC 等，形成了IBM PC 机系列，后两种主要与IBM 的大型机联网用。XT 机于1983 年3 月推出，PC/AT 于1984 年引入，PS/2 于1987 年4 月推出。

IBM PC 选用了Intel 8088 作为CPU，8088 采用20 位寻址，最多可访问1 MB 字节空间，其中640 KB 供用户使用，称作常规存贮。IBM XT 使用硬盘和EGA 图形板。在PS/2 系列中30、40、50型是80286 机，80 型是80386 机。给BIOS 保留了额外的64KB 高位存贮，用于开机自检(POST) 和OS/2 功能(称A-BIOS)。PC/AT 是为Intel 的80286设计的，可访问16MB的RAM。此时CPU 应于保护(protected) 状态下运行，这个名称源于系统一次运行多个程序，而每个程序相互被保护而互不干扰。因为Intel 设计286 先于DOS 的流行，故此芯片对DOS并不特别友好，为了保证向后兼容，芯片留一与8088 兼容的状态，称为实方式(real mode)，使DOS 于286 下运行，但仅限于传统的1MB。启动286 时先进保护状态，之后转入实方式。

下图给出了IBM PC 机微处理器的发展过程[1]。

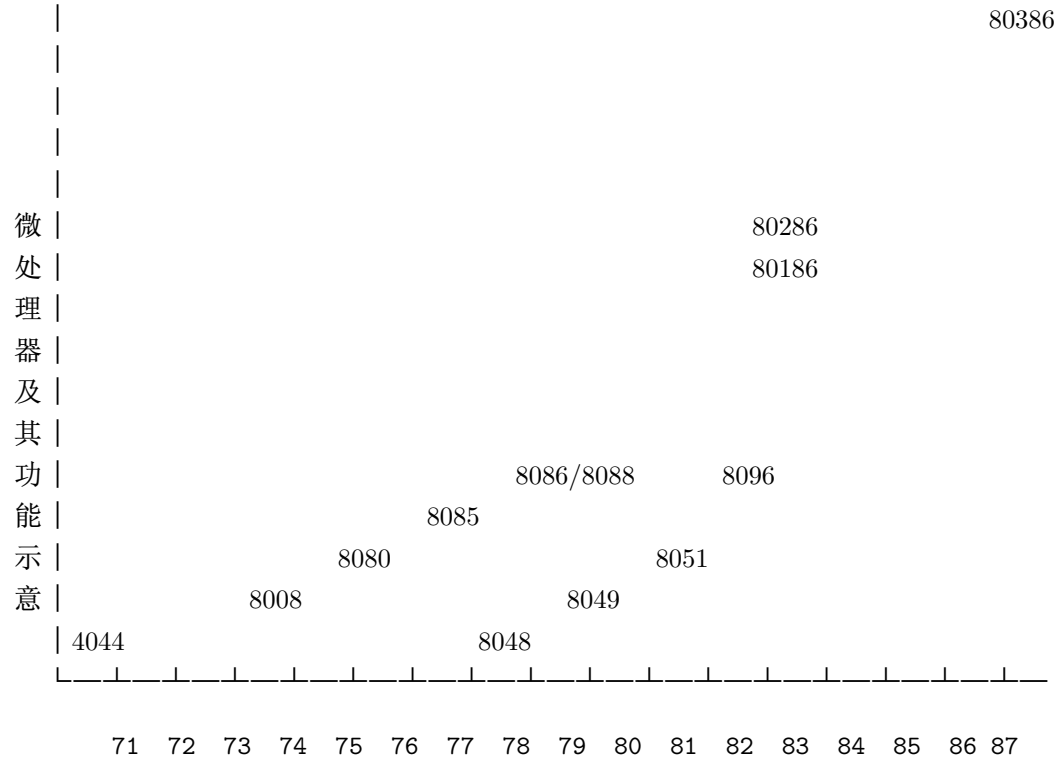


图 3.1 Intel 的微处理器

80386 处理器于1987年引入，80486 处理器于1989 年4 月推出，达15MIPS，是80386的四倍，实际上相当于一个80386 加上数学协处理器。1993 年又推出pentium 芯片，相当于80586。

在计算机中，数的表示有二进制(Binary)、浮点数(Floating point)、二进制编码的十进制(Binary coded Decimal) 十进制(Decimal)、八进制(Octal)、十六进制(Hex decimal) 用字母数字型(alphanumeric，如EBCDIC 及ASCII)等。

表 3.1 磁盘操作系统MS-DOS 的发展过程

版本号	时间	应　　用　　特　　点
DOS1.0	81.10	PC 机第一个操作系统，仅支持单面软盘
DOS1.1	82.10	广泛用于PC 兼容机，支持双面软盘
DOS2.0	83.3	PC/XT所用操作系统，支持硬盘
DOS3.0	84.8	PC/AT(286)所用操作系统，支持1.2MB软盘与大容量硬盘
DOS3.1	84.11	支持Microsoft 网络服务系统
DOS3.2	86.3	支持3.5 英寸软盘，驱动器中固化了盘格式化程序
DOS3.3	87.4	支持虚拟盘，支持硬盘分区且可支持PS/2 系统
DOS4.0	88.6	支持大于32M 的单一分区及许多强大功能
DOS5.0	91.6	改进了内存管理、方便、安全、高效、适于MS Windows
DOS6.0	93.3	提供硬盘压缩、碎片消除和备份、防病毒、内存优化工具

如多数软件包运行后产生分页符，它是第12个ASCII码，因此编辑软件包生成的结果时，利用编辑软件如PEII中的alt结合小键盘数字加以替换，避免不必要的走纸。本章稍后一部分介绍了针对SAS上的应用，PE II 软件在第 1 6 章介绍。

IBM PC微机字长是16或32位，形成一个字节。常用的进位关系是1K=1024字节1兆(MB)=1024K字节，1千兆(GB)=1024MB字节，本章最后部分给出了详细的计算机数制表。

习惯上，把前缀加在数量单位前面，如常规内存为640KB，386 机内存寻址为64TB，386机一个段地址为4GB，最初IBM PC 机的速度为4. 77MHz ，PC SAS/BASE 6.03 约占5MB等。

§3.2 MS-DOS 运行原理和常用命令

微机常用的操作系统有MS-DOS、OS/2 和XENIX 等，目前，多数计算机采用Microsoft 公司开发的DOS 磁盘操作系统(以下简称DOS)。同微处理器的发展一样，DOS 经历了一段时间的研究和实践，见表3.1[2]。目前又推出了DOS 5.0 和6.0，Microsoft WINDOWS 3.0 于1990年推出，其成功成为MS-DOS未来的策略。1995年，Windows第4版即Windows95问世，强化了设备和程序管理、网络支持等。

WINDOWS应用软件帮助信息的引用：HELP、EDIT、COPY 至CLIPBOARD ，进入WINDOWS工具WRITE可以存成有格式文件或者标准文本文件，在SAS下可以直接提交。

在DOS 中常用的设备是控制台(CON)、并行口(LPT1、LPT2)、串行口(COM1、COM2)、打印机(PRN)、空文件(NUL)，它们可以在DOS 支持下的高级语言中使用。

DOS 由几个主要部分组成，每一部分对应用程序提供特定支持，这些部分分别为基本输入输出(DOS BIOS)、核(DOS Kernel) 和外壳(DOS Shell)。在MS DOS 中，这三个模块分别对应三个文件，其文件名为IO.SYS、MSDOS. SYS 和COMMAND.COM，在PC DOS 中则为IBMBIO.COM、IBMDOS.COM 和COMMAND.COM。IO. SYS、MSDOS.SYS、IBMBIOS.COM 和IBMDOS.COM 是隐含文件，在执行DIR命令时看不到，但用PCTOOLS 和NORTON 等软件可以看到。IBMBIOS.COM 由两部分组成，即系统初始化程序SYSINT 和标准设备驱动程序。IBMDOS.COM 完成DOS 的文件管理和系统调用功能。COMMAND.COM 是DOS命令的解释程序，是操作系统和用户的接口，其功能是：命令解释；开始应用的调用. 走

访DOS 的路径及错误号；批文件处理. 有CALL, ECHO, GOTO, IF, PAUSE, REM, SHIFT 命令。COMMAND.COM 有三个部分，即：

A.初始化部分. 确定内存，把暂存放于最高地址，在指示/P 时，尚建立一环境区及执行AUTOEXEC.BAT。环境大小由/E 所指示（在DOS 4.0 有一系列错误讯息）

B.驻留服务部分. 具有三种基本功能: a. 执行程序并于结束时返回控制；b. 提供隐含的错误处理，最常见是"Abort, Retry, Ignore or Fail ?"；c. 释放命令暂存。

C.暂存部分(transient). 是它最大的部分，提供内部命令及批命令、命令行提示等。TRANSIENT 的含义是DOS 可以在必要的时候覆盖此区，它提示A¿、C¿，并有重新定位功能把.COM 和.EXE 文件装入内存，为程序建立PSP。

D，MS DOS 启动的顺序是先开外设如显示器，再开主机(关机顺序与之相反)。以DOS4.0为例，系统启动后，首先进行初始化(SYSINT)，有以下步骤：

a.自举(BOOT-UP). 即开电后的自检或者称POST，用存于ROM 的程序进行，检验处理器、内存、显示器、键盘、驱动器及按装的适配卡。无错误时再检查驱动器，有盘则把第一扇区内容读入内存，否则即转入第一个硬盘的活动的分区，错误时，可以出现磁带BASIC。

b.初始化(INITIALIZATION). IBMDOS 初始化建立默认的数据表，再打开标准设备CON, AUX, PRN 及驱动器，然后把控制转入CONFIG.SYS。

c.CONFIG.SYS 的处理. 在无CONFIG.SYS 时则停止，否则CONFIG.SYS 被存于内存，标记拷入内存顶端。此过程删除所有REM 语句。

继续有i. XMEAEM.SYS 检查(可使386 机于虚拟状态工作)；ii. 以后设备驱动程序的检查，每个驱动程序仅有一个头部，是EXE 型或是COM 型；iii. 配置参数的检查，即BREAK,BUFFERS, COUNTRY, FCBS, FILES, LASTDRIVE, SHELL,STACK，SWITCHES 参数；iv.系统扩充检查，即AUTOEXEC.BAT (DOS 4.0 则是INSTALL)。

d. 开始COMMAND.COM 的解释此时DIR, COPY, REN, ERASE 可以使用。

e. 系统停止(SYSTEM HALT). 若用SHELL* 开始的结束后，则提示 "TOP LEVEL COMMAND INTERPRETER TERMINATED, SYSTEM HALTED"，你应再启动。

系统的扩展可用程序APPEND, ASSIGN, SHELLB, FASTOPEN, GRAPHICS, IFSFUNC, KEYB, NLSFUNC, PRINT, MODE, SHARE。

DOS 对用户提供两个可修改的系统文件，即自动批处理文件AUTOEXEC.BAT 和系统配置文件CONFIG. SYS，批处理文件是计算机内部与外部命令的组合，AUTOEXEC.BAT 常用以下的设置：

```
a. PATH C;\;C:\DOS;C:\USER
```

即把路径设于根下、DOS命令的目录以及用户自己的目录。一般来讲，安装统计软件包以前，系统已做了一些路径设置，制作软件包的运行批文件时，可以包括下面的指令：

PATH %PATH%[C:\软件包目录]

如PATH %PATH%;C:\BMDP 是告诉系统在原先设定搜索路径的基础上填加BMDP 的路径。

b. PROPMT PG 即把路径放于通常的大于号内提示：如C:\DOS>_

其它常用的命令如ECHO、REM、APPEND、SUBST 等设置。

CONFIG.SYS 的常设置有：

FILES=打开的文件把柄数(handles)，如在SAS 中推荐用60。

BUFFERS=打开的缓冲区数目,如常用15。可在DOS3.3中用FASTOPEN 命令来设。

COUNTRY=国家信息。

DEVICE=设备驱动程序名，如DEVICE=ANSI.SYS，在设定国家信息、扩展内存等要用到，实际上可以是用MASM, LINK 及EXE2BIN 文件做成的特殊的程序。在WINDOWS的CONFIG.SYS文件中含有语句DEVICE=HIMEM.SYS，是用HIMEM.SYS 使用内存64KB 的HMA。

CONFIG.SYS	其它有关的设置内容有：
DEVICEHIGH	设备驱动程序名
DOS	扩展/扩充内存管理
DRIVPARM	设置一个物理驱动器号
FCBS	置打开的文件控制块数目
INSTALL	为内存驻留程序提供一个更好的装载手段
LASTDRIVE	置最后驱动器
SET	可以用于设置环境变量以便与AUTOEXEC.BAT等应用程序接口
SHELL	装载并启动命令处理(外壳)程序
STACKS	设置动态使用的数据栈
SWITCHES	阻止系统使用扩展键盘

用例：

```
country=001,437, c:\dos\country.sys
device=c:\dos\ansi.sys
device=c:\dos\display.sys con=(ega,437)
```

又例，

```
files=50
country=033
device=display.sys con=(EGA,863,2)
device=printer.sys lpt1=(4201,863,2)
drivparm=/d:1
device=ramdrive.sys /e
```

DOS 的命令有内部命令与外部命令两种，前者在DOS 启动后的任何时间均可行，后者则隐含于带有.COM 或.EXE 扩名的文件中。DOS 命令处理时优先处理内部命令，内部命令含在COMMAND.COM 的暂驻部分，位于内存高端。若非内部命令，则在当前目录以及PATH指定的目录中寻找，顺序是.COM，.EXE，.BAT。执行时，DOS 在常驻部分之上的可用内存空间最低端建立一个程序段前缀(PSP)，所括命令行参数、文件控制块、环境块地址以及系统内部使用的附加信息。若是.EXE 文件，则系统还需要再定位(relocation)，最后，被加载程序位于PSP 上方，CS:IP 指向程序的第一条指令，程序正常运行。DOS内部命令见下表：

一个园点(.) 表示当前目录，两个连续的园点表示上级目录，反斜杠(\) 表示根目录。进入下级目录用命令CD ¡目录¿，返回则可用CD..至上级目录，以CD\ 返回根目录，CD\ < 目录> 进入根下的另一目录。

APPEND　开始是一个外部命令，执行后成为内部的。它可以把目录文件作为环境的一部分。如WordStar 运行时要同时存在WS.COM、WSOVLYL.OVR 及WSMSGS .OVR 三个文件，仅仅设好PATH 还不够，可以用以下命令放于AUTOEXEC.BAT 文件：

```
APPEND /E
```

表 3.2 DOS 的内部命令

命令名	功能说明
BREAK	设ON 或OFF，置DOS 中断开关
CALL	调用批文件，允许批文件间的调用
CHCP	改变国家代码
CHDIR 或CD	改变目录
CLS	清屏并使光标位于屏幕左上角
COPY	文件拷贝
CTTY	改变主控制台
d:	改变驱动器
DATE	显示和改变系统日期
DEL 或ERASE	文件删除
DIR	列目录
EXIT	退出DOS 外壳返回应用程序
FOR	用于批文件中指令的循环操作
GOTO	控制批命令中的分支
LOADHIGH 或LH	将程序装入高地址区
MKDIR 或MD	创建目录
PATH	建立访问路径
PAUSE	批命令执行暂停
PROMPT	设系统提示符
REM	显示注释信息
RENAME 或REN	文件改名
RMDIR 或RD	空目录删除
SET	设DOS 环境
SHIFT	批处理命令文件中的移位替换
TIME	显示和设置系统时间
TRUENAME	显示完整的文件指示信息
TYPE	显示文件内容
VER	显示DOS 的版本号
VERIFY	验证写盘数据
VOL	显示磁盘卷名

表 3.3 DOS 的外部命令

命令名	功能说明
APPEND	允许打开指定目录上的或当前目录上的数据文件
ASSIGN	分派驱动器请求
ATTRIB	置文件属性
BACKUP	硬盘文件备份
CHKDSK	磁盘检查与修复
COMMAND	加载命令处理命令
COMP	磁盘文件比较
DEBUG	DOS 命令文件调试程序
DISKCOMP	磁盘比较
DISKCOPY	整盘拷贝
DOSKEY	命令行编辑程序，键盘宏定义
DOSSHELL	执行DOS 外壳命令
EDIT	全屏幕编辑器
EMM386	扩展/扩充内存管理
EDLIN	行编辑
EXE2BIN	生成二进制文件(.COM、.SYS)
EXPAND	释放DOS系统压缩文件
FASTOPEN	缩短打开文件及驱动器时间
FC	文件内容比较
FDISK	硬盘分区管理
FIND	字串查找
FORMAT	磁盘格式化
GRAFTABL	装入附加图符表
GRAPHICS	拷贝屏幕图形
HELP	DOS 系统命令求助
JOIN	驱动器与目录的链接
KEYB	装入键盘替换程序
LABEL	设置磁盘标号
LINK	DOS 目标文件连接程序
LOADFIX	把用户程序装入常规内存第一个64KB并执行之
MEM	显示内存使用情况
MIRROR	磁盘信息记录程序
MODE	设置设备操作方式
MORE	屏幕显示过滤
PRINT	假脱机文件打印
QBASIC	DOS 5.0/6.0 BASIC
RECOVER	磁盘文件恢复
REPLACE	文件拷贝和替换
RESTORE	备份文件复原
SELECT	选择国别代码并生成系统
SETVER	改变MS-DOS向程序报告的版本号
SHARE	安装网络文件共享
SORT	文件排序

```
APPEND C:\WS
PATH C:\WS
```

则于DOS 的任何目录下都可以使用WS 命令进行编辑了。使用FoxBase 以及第18章介绍的WPS 亦可以用此办法，可以在上述APPEND 命令行参数中加上它们各自的目录，相互之间用分号(;) 间隔开。

各命令在使用时应当注意其命令行参数，命令行参数是在执行DOS 命令时在命令的同行上打出的内容。各命令行参数用法可详参DOS 手册。

常用DOS 文件扩展名的含义如：

.BAT BATch	批处理文件
.COM COMmand	系统命令文件
.CPI Code Page Information	代码页信息
.EXE EXEcutable	系统可执行文件
.PIF Program Information File	程序信息文件，用于WINDOWS
.SYS SYStem	系统文件，部分如KEYBOARD.SYS 例外

多个DOS 命令可以组合到一个命令行上执行，这要借助于DOS 的管道命令和再定向功能，DOS 的管道(pipe) 和再定向(redirection)，前者的— 表示并列执行，如：< 表示文件输入、> 表示文件输出、>> 表示追加到文件尾部，

C:\ >TYPE C:\DBASE\READ.ME | MORE	分页显示READ.ME 的内容
C:\ >DIR C:\DOS >PRN:	把DOS目录下的文件目录在打印机上列出
C:\ >TEST <DATA	把DATA 作为TEST文件的输入
C:\ >MORE <C:\DBASE\READ.ME	把READ.ME的内容分页在屏幕上显示

DOS 命令按命令功能的性质进行分类，有常用命令、过滤命令、配置命令、设备命令、批命令五类。常用命令指未被括入其它类型的DOS命令，过滤命令指从管道、重定向文件或设备的输入，配置命令主要指CONFIG.SYS中的命令，设备驱动程序是一种配置命令，支持的设备如：显示器、键盘、打印机、磁盘驱动器和辅助设备，批命令是在批文件中执行的内部命令。这种分法使用不多。

§3.3 MS-DOS 的内存管理与软件配置

§3.3.1 内存映象

使用统计软件包要与内存打交道，DOS的CHKDSK命令及在DOS 4.X 的MEM 命令可显示内存使用情况。使用DECnet PCSA系统中的MEMMAN 更可看出系统驻留程序及其执行的命令行参数。

1. MS-DOS的内存映象如下：

2. 内存的第一部分为常规内存(conventional)，大小为640 KB，映象如下：

剩余空间随系统和文件、缓冲区设置的大小而定，典型的是450-580 KB。

3. 上位存贮区(Upper memory Block, UMB) 起于640KB，至于1MB，共384KB，DOS 用于存贮ROM BIOS 即基本的输入输出管理和盒式BASIC，其映象如下：

高位内存区(HMA) 起于FFFF:10H，大小为64KB-16B。当CPU处于实模式并且第21条地址线(A20线)被激活的状态时，CPU就可以访问一块65520字节的内存，这块内存就叫做HMA。HMA的存在与CPU的寻址方式有关。HMA使用时必须安装XMS 的设备驱动程序。对于它的控制，Microsoft使用eXtended Memory Speci- fication 或XMS。驱动程序HiMem.sys可使用64KB，用后其它程序

16MB或4GB			
	Extended memory		扩展内存
1088KB		110000h	
1088KB	High memory area	10FFFFh	高位内存区域
1024KB		FFFFFh	
	High memory		上位存贮区
640KB		A000h	共384KB
640KB		9FFFh	
	Low memory		常规内存
0KB		0001h	

图 3.2 80286 及80386 内存映象

不能再使用它。在DOS 5.0 中，可把操作系统大部分命令移到HMA从而把较多的空间留给应用程序。

扩展内存(extended memory) 是1MB-16MB 或4GB 的存贮，位于HMA之上，多用于存放数据，故其最低地址是1024KB+64KB=1088KB。AutoCAD 386 和Lotus 1 -2-3 3.0 可在386 机访问4GB，VDISK.SYS 可至15MB、Windows 3.0 可在386 机访问4GB，VDISK.SYS 可至15MB、Windows 3.0 可至16MB。

使用扩展内存的另一种管理方法是扩充内存或EMS(Expanded Memory Speci －fication) 是在高存贮区64KB 进行16KB 为单位的(页)式存取方法，使用时可以不关心其具体地址，因为所有操作是针对1MB内存中的页(物理页)帧进行的。这些页是连续的，一般至少有4个，组成一个页帧(page frame)，经过这些物理页与逻辑页交换，这也是传统的内存管理方法。

有两种方法可以得到EMS，即使用EMS内存卡或通过内存映射(在386或486上)，有些286在系统开机设置时可以将扩充内存定义为EMS。

EMS使用时必须安装一设备驱动程序，如:QEMM.SYS。

针对扩展内存的应用程序并非一致。Quanterdeck Office 与Phar Lap Software 联合开发开发一个协议称虚拟控制程序接口(VCPI) 可允许AutoCAD 386、Lotus 1-2-3 V3.0 在扩展内存下使用而不与EMS 干扰。但却不能在多应用中使用。为此，IBM 推出了DPMI。1990 年夏天，Microsoft 与其他七个软件厂家包括Quarterdeck 和Phar Lap 推出DPMI 0.9，可允许多任务系统象Windows 3. X 及OS/2 运行多个使用扩展内存区域的DOS 程序。

在EMS3.2 版的EMS 管理仅是在特殊的内存板上以16KB 为单位进行移进移出即页式访问(paging)，这样Lotus1-2-3 可访问8MB 的数据。鉴于这种限制，QuadRAM、AST 及Ashton-Tate 采用增强的EMS 或EEMS 。可使程序与数据访至32KB 存贮，交换亦增至1MB，用DESQview等软件可一次运行许多程序。80年代末，LIM 财团括入了EEMS 的改进并引入LIM EMS 4.0。允许把扩展内存调入EMS 而不必使用其他的硬件。

在AutoCAD 10 中可用4MB 做为I/O 交换空间，Lotus1-2-3可用32MB数据，PCTOOLS 用512KB 作覆盖，WordPerfect 5.X 可用384KB。

虚拟存贮管理(Virtual Memory Manager) 或VMM 是4KB 的存贮，在扩展内存与硬盘间进行页式访问。

常规内存是在所有计算机中的基本内存，大多数计算机有640K的常规内存，如果没有特殊指令需要使用其它内存的话，程序一般使用常规内存。

640KB	Drivers	9FFFh	驱动程序
	TSRs		驻留程序
	Programs and Data	0B85h	程序和数据
	Buffers, files		缓冲区、文件
	COMMAND.COM, etc.	0A98h	
		0A97h	
	DOS		操作系统
		0070h	
		006Fh	
	System area		系统区域
		0050h	
		004Fh	
	BIOS data area		
		0040h	
		003Fh	
	Interrupt area		中断向量表
		0000h	

图 3.3 DOS 常规内存映象图

§3.3.2 软件配置

只有进行了恰当的配置，才能充分发挥计算机资源和应用软件的效能。使用扩展内存多要借助于设备驱动程序，它在系统配置文件config.sys 中指定。也可通过执行DOS 命令文件，如：

CONFIG.SYS 内容为：
device=[d:][path] 驱动程序名[选项]
C:\ >TYPE CONFIG.SYS
device=c:\dos\limsim.sys 1024
DOS 5.0 在安装了HiMem.sys 后，即借助EMM386.EXE 使用上位存贮器。如：
C:\ >TYPE CONFIG.SYS
device=c:\dos\himem.sys
device=c:\dos\emm386.exe noems
DOS=HIGH，UMB
或：device=c:\dos\himem.sys
device=c:\dos\emm386.exe 512 FRAME=D000 RAM
等等。

DOS 设置运行环境的另一个重要文件是autoexec.bat，该文件在DOS 引导时盘的根目录下自动被执行。在DOS6.0下有专门的程序memmaker进行内存管理，掌握它的使用是很有必要的。

如上节所述，当DOS 运行应用程序时，首先在内存可用空间的最低端建立一个程序段前

1024KB		DFFFh	
	System ROM (64KB)		系统只读存贮
960KB		F000h	
960KB		EFFFh	
	EMS page frame (64KB)		EMS 页帧
896KB		E000h	
896KB		DFFFh	
	Free (72KB)		自由空间
824KB		CE00h	
824KB		CDFFh	
	Disk ROM (8KB)		只读磁盘管理
816KB		CC00h	
816KB		CBFFh	
	Free (24KB)		自由空间
792KB		C600h	
792KB		C5FFh	
	Video ROM (24KB)		只读视频显示
768KB		C000h	
768KB		BFFFh	
	VGA text (32KB)		VGA 文本区
736KB		B800h	
736KB		B7FFh	
	Free (32KB)		自由空间
704KB		B000h	
704KB		AFFFh	
	VGA Graphics (64KB)		VGA 图形区
640KB		A000h	

图 3.4 DOS 高位存贮区映象

缀PSP 包括程序定位和运行的有关信息。这些信息包括DOS 和程序本身使用的入口以及加载程序使用的参数。在PSP+2CH 字段包括环境块的段地址。DOS的环境是由一系列字符串组成的。如SET、PROMPT、PATH 和COMPSPEC=。可在DOS 提示下用SET 看其设置，BMDP在运行前需要指示环境变量DNEWS=BMDP 路径。仅仅使用PATH 时，可以显示当前的路径设置，可以结合前面介绍的%PATH% 量制做BMDP运行批文件。为了避免环境不足，DOS提供CONFIG.SYS中使用SHELL设定，如SHELL=C:\COMMAND.COM /P /E:256。

在运行DOS 可执行文件(扩展名为.EXE、.COM 和.BAT) 只要打入文件的名字就可以了，有时还要加上一些其它的指示，称为命令行参数(command line parameters)。参数也放在程序的PSP 中。这样运行诸如SAS [SAS 程序名] 或SPSS [SPSS/PC+ 程序] 就成为可行的了。

如SPSS/PC+ 只要指明了路径，就可以在任何目录下使用。因此，建议在DOS 的自动执行文件AUTOEXEC.BAT 或SPSS/PC+ 的运行批处理文件中加上PATH \SPSS;命令。指示许多路径时，不同的目录用分号分开。

如：C:\ >TYPE AUTOEXEC.BAT
ECHO OFF
PROMPT PG
PATH C:\;C:\DOS;D:\SPSS;D:\WP51

§3.4 语言编程

从某种意义上讲，语言编程能力和软件使用能力密切相关，建议在学习数据处理和统计分析软件的同时，掌握高级语言如Fortran 的有关概念。

概率统计计算，主要是利用计算数学的方法解决统计问题。统计软件的形成，是用算法语言把数值算法表达出来，如微机SAS 和dBASEIII 均是用C 语言写成的。如SYSTAT 中求特征值采用QL 法。BMDP、Systat、Splus、Genstat 等均提供了与Fortran或c语言的接口。不少情况下，使用软件包需要指定所采用的计算方法，若对之不熟悉，只好望而却步。IMSL、NAG是标准的Fortran或c语言算法库，中科院的SASD 也是使用Fortran 写成的。

§3.4.1 有关概念

相对于机器语言(machine language)，Fortran一类高级语言(high- level language，HLL)更接近于自然语言，它们可以进一步分成解释语言与编译语言。介于两者之间的是汇编语言(assembler language)。高级语言要通过软件把指令转成机器码。第四代语言有以下基本特征：⑴用户友好(user friendliness)；⑵数据可及性(data accessibility)；⑶处理的弹性(flexibility of processing)；⑷开发及编程效率的改善(development and programming productivity improvement)。第四代语言的例子是SAS 语言。四代语言的大致划分为：机器水平与汇编水平为第一代；面向过程非结构化的高级语言，即算法水平为第二代，如：BASIC、Fortran Ⅱ、COBOL、Algol-60、PL/I、APL，再分为通用的如PL/I 以及专用的如Logo；面向过程结构化高级语言即算法与问题水平是第三代语言，如Ada 与C；第四代语言(4GL) 是非过程式的，不写实现算法，如数据库查询语言SQL 只要求在什么地方(文件) 按什么条件(约束) 查出什么数据。事实上查找是按数据库的结构并根据约束比较，或者使用搜索算法实现。但这些工作系统代替用户完成了。面向目标水平的语言可以进行符号操作，如LISP、PROLOG。解释语言能够直接执行源代码，编译语言则先要转成目标代码，然后连接为可执行代码。

§3.4.2 高级程序语言

1. 高级语言种类很多，如Fortran、BASIC、C、COBOL 、LISP 等。BASIC(Beginner's All-Purpose Symbolic Instruction Codes) 使用最简便。Fortran (FORmula TRANslation) 最早用于科学计算，是典型的编译语言。COBOL(COmmon Business Oriented Language) 是面向商务的应用语言，使用略窄。C 语言现多用于软件制做，其程序代码简捷，可移植性好，能充分利用系统的资源。LISP 多用于专家系统的开发，与统计分析有关的是XLISP-STAT，在第17章涉及的AutoCAD 软件使用专用的AutoLISP。现将这几种语言概况简述如下。

   【BASIC 】1964 年由Dartmouth 大学J.G.Kemeny 与T.E.Kurty 两位教授创立，以后又不断修改补充。是解释语言的典型代表，有BASIC,BASICA与GWBASIC 以及最近的QuickBASIC、True BASIC 等。一般来说，它们运行占用的内存小，计算与调试方便，集编辑、数据管理、图形等功能于一身，亦有丰富的指令调用系统资源，也

可以用SHELL命令执行DOS 命令或者退至DOS 外壳下，还有专门的MKDIR、CHDIR及RMDIR 命令，其CLS命令亦与DOS 系统命令相同。调试时常删除或添加语句，一段时间后就可以用RENUM 命令整理一下。在True BASIC 及QuickBASIC等环境下可以不用行号，在QuickBasic 中也可以采用递归。DOS 5.0 和6.0的QBASIC 拥有QuickBASIC的特性。本章稍后部分和第 1 4 章各有一个使用BASIC的例子。

【PASCAL 】由瑞士苏黎世工程技术大学的Niklaus Wirth 教授于1968年设计的，用十七世纪著名的法国哲学家和数学家Blaise Pascal 名字命名。是结构化程序的代表，其基本结构是头部的说明、文件及类型说明和BEGIN 与END内的程序体。有自定义类型及集合变量类型等。有典型的顺序、选择及循环模块化结构。Turbo Pascal 严格遵守由K.Jensen 和N.Wirth 在《Pascal 用户手册和报告》中定义的标准，并有许多扩充。不少软件如EPI INFO 也是用TURBO PASCAL 写成的。

【FORTRAN 】是由IBM及一些计算机用户于1957年开发，是一个过程化的通用语言。它拥有丰富的函数，常用于科学计算，也是数学和统计工作者最常用的语言，各种版本的Fortran 程序可经编译的目标文件做成库文件供以后连接调用，如国内的SASD。特别指出的是，在一些杂志上经常有各种算法的Fortran 子程序，一些软件如VAX/VMS、SYSTAT、BMDP还给出调用时的格式。LISREL 本身就是用Fortran 写成的。

开始使用时，往往感到它的格式描述难于掌握，同时它的参数传递、文件操作也可以造成困难。但掌握了它的格式后、参数传递方式后，对于使用一些统计软件是有益的，如GLIM、SAS 等。由于运行时需要编译，就不象BASIC 语言那样直观，MS-Fortran加上$DEBUG 一类调试语句、设置断点、打印语句，则当程序运行时可以指出出错的行号、前后的内容很方便，另外结合MAKE、CODEVIEW 等工具能够大大改善调试的效果，总之微机上Fortran 的编译连接很费时间。

【C 】由D.M.Ritchie 于1972 年所提出，是一个广泛用于软件开发的语言，较好解决了实用、移植、高效之间的矛盾。适当地考虑了背景机的特点，适于编写系统程序；具有良好的模块化结构；具有丰富的运算符、实用的表达式和先进的数据结构与控制结构，表达能力强而且灵活；书写简炼，易学易写。但某些运算符的优先次序不符合日常习惯。常用的如Quick C、Turbo C 等等。

C++ 由美国Bell实验室的Bjarne Stroupstrup 于80年代开发，其目标是对C进行改进，对数据抽象以及目标编程更好的支持同时保持与C 的兼容。最著名的是Free Software Foundation 的GNU C++，其它的如Borland C++和Symantec C++。

微机上的dBASEⅢ、SAS 就是用C 语言写成的。第16 章有一个用于计算机——四通文件转换的C语言实用小程序。

【COBOL 】1959年由CODASYL(Conference On DAta SYstem Language) Committee 开发，于1968年为ANSI等组织所认可。COBOL 是一个面向商用的语言。由标识区、设备区、数据区和过程区大的结构组成，

在某种意义上，所有语言是等价的，而每种语言总使某种问题的处理变得容易。对于数据处理和分析工作者来说，兴趣可以不在语言特征上，如它们调用汇编语言的格式等，关心更多的是用有限的条件，解决自己的问题，以取长补短。这时需要与计算机专业人员相结合。

了解常用的统计算法和程序语言，可以在没有条件获得最新软件的情况下实现现有软件所没有的功能，软件包的使用并不能代替自己软件的学习。Press, W. (1992) 等的Numerical Recipe 包括统计计算，还提供Fortran 和C 两种版本。

2. 使用可首先从BASIC 一类解释语言开始，如DOS中的QBASIC。其优点是可以边写程序边调试。可以提高使用的兴趣，循序渐进。Fortran 一类编译语言程序则一般有以下的顺序：

源程序的编辑 —→ 编译 —→ 目标文件的连接 —→ 执行

程序的编译用专门的编译程序进行，设在微机上用编辑软件生成程序TEST .FOR，MS-Fortran 77 3.X 软件的操作如下：

```
C>PAS1 TEST.FOR,,;
C>PAS1
C>PAS2
C>LINK TEST,,;
```

首先生成.OBJ目标文件，最后生成TEST.EXE可执行文件，执行时只要在系统下打入文件名TEST即可。LINK 还可以用自动应答文件：设有文件RESP，内容是：

```
TEST
TEST
TEST
```

则可以用LINK @RESP 来完成上述操作。三行分别对应执行文件名、列表文件名和映象文件名。

在MS-Fortran 4.0 则是FL 命令。若要写子程序供BMDP 等软件使用，则可仅编译用户书写的子程序生成. OBJ 文件。除此之外还有许多维护工具如CODEVIEW、MAKE 等。如用MAKE，可用以下的描述文件：

```
TEST.obj:TEST.for
  FL /Fs /c TEST.for
TEST.EXE:TEST.obj
  FL /Fe TEST TEST.obj
```

前一行是文件的来源描述，后一行则是产生的方法。

使用系统提供的库管理程序可以生成自己的运行库文件，如：

C>LIB file1+file2

Fortran 5.0 提供了结构记录，do/enddo控制等，功能大大增强。在VAX 下，使用Fortran/link启用编译和链接，在Unix系统下使用f77或f90等。相对于PC 的make，它们的功能更强。

3. 语言要素一般有以下几种情况：

A .变量类型

B .赋值与算术表达式、关系及逻辑表达式

C .无条件分支与条件分支、

D .循环

E .数组及其它数据结构

F .子程序和函数

G .输入输出和图形

H .系统调用

每一种高级语言都有自己的变量及其类型。常用的如：整数、实数、浮点数、单精度、双精度及记录类型、集合类型等，它们决定了语言使用的界限。在8087 等浮点处理器中还要细分，如长、短整数和实数、符号数，逻辑变量，指针变量。

不同类型的变量间可进行算术与逻辑运算和字符操作。

函数与子程序用于大量的重复调用时，使用哑元传递非常方便。这样可以只要子程序及其相应的调用说明。有些语言可以递归调用。

输入、输出操作主要是指文件的操作。

多数数据处理和统计分析软件具有句法分析(Parsing)功能，Stata还提供了相应的工具。

4. 程序设计第一步要明确问题，对被解决的问题，有一个总的轮廓。例如要分哪些步骤，用什么参数，指标等等，进而构造一些模块。软件的实现大体上有设计、编程、调试与测试验证几个阶段。比较典型的方法如流程图法，模块程序设计和结构程序设计。七十年代初，Boehm 和Jacobi 提出并证明任何程序都可以由三种基本结构构成. 这三种结构是：顺序结构，条件结构和循环结构，每个结构只有一个入口和一个出口。三种结构可以任意组合和嵌套。PASCAL 语言具有顺序、条件、循环、递归及选择几种结构，是比较典型的。面向目标(OOP)的设计方法又提供了许多新的概念如类、传递、多态等。软件工程(software engineering) 用于一般描述程序开发、检测、修正以保证产生可靠和有效的程序的科学方法。程序设计与统计分析一样，既是一门科学又是一门艺术。人们长期的编程实践，已形成一套通用的处理方法，编程中应加以注意。如求和(计数)、乘法(求幂)及数(表) 的排序、问题的迭代求解等。简单的如平均数的计算(求和) ，阶乘的计算Fibonaaci 级数的计算(递归)。

一般说来，在使用软件包处理常见的问题时时，软件的处理过程好象一个黑箱，但应用简便。如Fortran 在开始使用时，往往难于掌握它的格式描述，同时它的参数传递、文件操作也可以造成困难。但掌握了它的格式后，对于使用一些统计软件是有益的，如GLIM、SAS 等。由于运行时需要编译，就不象BASIC 语言那样直观，在微机上编译连接也很费时间。

了解数值分析的原则和方法。如在一元二次方程$ax^2+bx+c=0$ 求解时，第一个根利用公式$x_1=(-a/2)(b+sign(\sqrt{(b^2-4ac)},b))$得出，第二个根利用韦达定理$x2=c/a/x_1$得到，从而避免了大数吃掉小数。在计算多项式$p(x)=a_1x^n+a_2x^{n-1}+...+a_{n+1}$ 的值时，直接计算需要N(N+1)/2 次乘法和N 次加法，采用秦九韶法$p(x)=x(a_n+x(...(a_2+a_1x)...)))+a_{n+1}$ 就仅仅用N次乘法和N 次浮点加法，从而大大减少了计算量。在计算复数据离散傅利叶变换时，直接计算需要N^2 次运算，而采用快速傅利叶变换FFT 就减为$N\times log2(N)$ 次，也是算法上成功的典范之一。概率统计计算有许多专门的文献，如：

Journal of the Association for Computing Machinery (JACM)
Communications of the ACM (CACM)
ACM Transactions on Mathematical Software (TOMS)
ACM Computing Reviews (Comp. Rev.)
ACM Computing Surveys (Comp. Sur.)
Numerische Mathematik (Numer. Math.)
Journal of the Society for Industrial and Applied Mathematics (SIAM)
SIAM Journal on Applied Mathematics (SIAM J. Appl. Math.)
SIAM Reviews (SIAM Rev.)
SIAM Journal on Numerical Analysis (SIAM J. Num. Anal.)
Mathematical Tables and Other Aids to Computation (MTAC)
Mathematics of Computation (Math. Comp.)
Journal of Statistical Computation and Simiulation (JSCS)
Journal of the American Statistical Association (JASA)
Annals of Statistics (Ann. Stat.)
The American Statistician (Amer. Stat.)
Technometrics (Techno.)
Communications in Statistics (Com. Stat.)
Journal of Royal Statistical Society (JRSS)
Computer Journal (Comp. J.)
Journal of Optimization Theory and Application (JOTA)
Journal of the Institute of Mathematics and its Applications
Computational Statistics and Data Analysis
Computers in Biomedicine

这些文献在不同的年份会有不同的名称，如Applied Statistics，Series C (JRSSC) 有关的算法和实现方法。

了解一些数据结构和存贮的知识，有一个数据在机算机上存储的映象，是很必要的。如统计上常用的是对称矩阵，在样本量很大时，把大的N x N 矩阵用一个(N+1)*N/2 维的一维数组存储，就可以大大节省内存，因而使所处理的问题的规模增加。中科院推出的SASD 软件，多采用这种方法。Press 等人指出，大多数报导指示的标准线性方程和矩阵运算包的错误，缘于用户在传递矩阵逻辑或物理维数时不恰当。

软件需要达到目的并且可行，如计算量、计算机容量需求等。实时处理时需要高效率，就须结合汇编程序以节约代码。Fortran 最大的特点是子程序相对独立，便于调用；BASIC则调试方便。除此以外，还要考虑通用性、可移植性与可读性。最简单的例子是在排序程序中加上IF(N.EQ.1) RETURN 语句。在矩阵求逆中有IF((N.EQ.1).and.(A(1).NE.0)) THEN A(1)=1/A(1)也有同样效果。另外是尽量用通用的语句，少用扩展的语言成分。程序中加注注释、采用自然语言的变量和过程名、编写详细的使用说明和提供运行范例等。利用模块化结构，便于进行功能替代。掌握一些计算机系统方面的知识，改善界面从而便于推广。

不小算法若精度设定不当，常常会出现死机的情况。最优化拟牛顿法中的Stewart 法有限差分需要考虑计算机的精度，这时可以用Forsythe，G.E.(1977) 等人介绍的方法

自动得出。某些计算机BASIC 采用角度或是弧度不能区分时，上海医科大学等的POMS程序集针对Apple II 和TRS-80 等计算机的BASIC，用SIN(4) 的符号预先加以判别，问题就很好解决了。有许多时候理论上可行的，实际上未必可行，如区间法求根就可能出现这种情形。所以Forsythe，G.E. 介绍的Brent法一元函数求根程序被众多的作者引用。有时候，合理使用计算机的内存与外存也是很重要的。

程序调试往往借助于断点设置了解注意运行过程中的信息，从小问题到大问题的“以小见大”方法往往事半功倍。

优良的程序还源于不断的积累，不少程式已经以特有的形式固定下来。许多专著给出了同一过程不同算法语言下的形式，如中科院计算所的SASD。Internet 有许多算法的源程序。有了这种基础之后，就可以把别人的和自己的新思路在短的时间内复制或程序化。也有根据软件包写成的统计专著。

“工欲善其事，必利其器”。计算机系统是统计软件运行的平台，为了用好软件包，应当熟悉和更新计算机系统基本知识，不断实践。

§3.5 Unix 系统简介

Unix的第一个版本是1969年PDP 7上，起于Bell实验室(即后来的AT&T) Ken Thompson的思想，PDP 是DEC 公司的产品，而第一个真正的系统则是1971 年由Thompson, Dennis Ritchie,及Rudd Cadaday开发与修正而出现在DEC PDP 11上。其设计思想是要高度模块化、高效和富于灵活性。Unix用C语言实现，从而给用户提供了一套完整的命令集，有助于开发“shells”并生成是命令。从用户的角度来看，Unix是一个同心圆，最内圆是系统内核(kernel)而外部的圆是一些外壳，用户可以用exit命令退出这些外壳并且也可以用简单的命令(如csh与tcsh) 建立新的外壳，在所建立的外壳内工作。MS-MicroSoft为PC开发的相应系统称为Xenix，加州大学伯克莱分校开发的称为BSD。AT&T则继续开发出Unix 2.0、System Ⅲ并一直到System V，版本2、3、4分别称为SVR2、SVR3、SVR4。Linux是最近开发出的系统，是免费的。另外HP公司、DEC公司及Sun公司相应的HP-UX、Ultrix及Solaris等。

Unix系统起决定作用的是其文件系统，事实上Unix中任何东西都是文件，从用户的角度看主要有三种：普通磁盘文件、目录和特殊文件。普通文件分文本、二进制文件。目录可以看成对文件名特征(物理位置、日期等)映射到的一张表，这些概念与MS-DOS有些类似，但以/提示目录层次结构，如：/usr/local/bin。

man man 给出主要的命令集、ls、cp、rm、mv、mkdir、cd、rmdir、chmod、egrep。

常用DOS命令与Unix系统命令对照如下：

其它的Unix命令如：

alias/unalias 用于别名定义，从而方便命令的录入。grep 命令用于搜索字串。ln 用于建立逻辑文件与物理文件的连接。^Z 用于程序暂停，打%继续。

rlogin 用于实现远程登录。使用xterm实现Unix X-终端到PC的仿真。finger 和who 用于查询用户登录信息。ps 与kill 显示或删除用户进程。chmod 改变用户(u, g, w, a)的权限(r, w, x)。script/exit 可以记录运行过程。

许多Unix命令有相应的DOS仿真程序，如：tar、cp、mv、rm等；带星号* 表示实用程序。Unix的ln命令建立文件与目录间的动态链接，文件在显示时后缀一个位置符”@”。标准的Unix编辑是vi，其功能强大(emacs与pico也很流行)。过滤是指对输入流的提取、插入或重

表 3.4 DOS与Unix命令对照表

PC	Unix	功能
attrib	- chmod	改变文件属性
chkdsk	- du,df,quota -v	磁盘空间
cls	- clear	清屏幕
command	- csh/tcsh	系统外壳
copy	- cp	文件拷贝
dir	- ls	列目录
del	- rm	文件删除
deltree	- rm -r	删目录树
date/time	- date	日期、时间
echo	- echo	显示
edit,edlin	- vi	编辑
help	- man	帮助
more	- more	显示控制
move	- mv	移动
md,cd,rd	- mkdir,cd,rmdir	建目录、换目录、删目录
comp/fc	- diff	文件比较
print	- lp, lpr[lpq, lprm]	打印
ren	- mv	文件改名
set	- set/setenv	环境设置
call	- source	系统命令调用
type	- cat,head,tail	显示文件内容
lharc*	- xlharc*	文件压缩
tar*	- tar	存取目录树
uuencode,	uudecode	文件编码
^Z	- ^D	文件结束符

排，最常用的有sort、sed和awk。sort的选项有-m(合并未排序的文件)、-d(只有字母/数字/空格参加排序)、-f(忽略大池写)、-r(反序)、-b(忽略前导空格)、+/-(排序关键字起始)。sed是一个流编缉器，它接受一个文件作为输入，并对文件的每行使用编辑命令。sed 使用ed 的指令集，如sed 's/DOS/Unix/g' chapter3 将文件chapter3中所有Unix替换为DOS；多个ed 编辑指令可以存放于一个专门文件，使用sed -f 启用这些命令。awk是一个模式匹配/ 操作处理器，它检查输入文件的每一行是否有给定的模式，从而执行特定的操作，如：awk '$1 == *SUBROUTINE* (print $1, $2) ' fortran 寻找文件fortran 中SUBROUTINE模式并把它打印出来。

§3.6 Internet 简介

Internet 是国际间计算机间的网络互连。使用特殊的语言协议(TCP/IP)，提供电子邮件、远程登录和文件传输服务。环球信息系统(WWW，万维网) 是一种客户——服务器软件包，在Internet使用超文本来组织和提供信息服务。下面是几个典型的地址：

```
http://www.sas.com
ftp://ftp.sas.com
gopher://gopher.gdb.org
E-mail: websupport@spss.edu
```

前两行分别是SAS公司的homepage和ftp地址，需要用支持超文本协议(http)的软件netscape、mosaic、lynx以及ftp访问。第三行是Johns Hopkins大学的gopher 基因数据库；第四行是SPSS的E-mail地址。

它们可以用通用的格式(URL)来描述，第一部分表示了访问的方法，如上面的http、ftp、gopher；第二部分则是Internet地址，有时这里包括了有关的文件指示，即文件的路径的名字。在E-mail地址中，位置符号@(读做英文的at) 以前的内容表示用户名，有时还包括字符!、识别并转换各种通讯协议。网络的节点圆点区分。DECnet节点名表示方法略有不同，要用::。从E-mail地址可以大致区分其含义：如国名：cn(中国)、ca(加拿大) 、au(澳大利亚)、jp(日本)、uk(英国)；类别：edu(教育)、com(商业)、org(各种组织)、mil(军事)。以下针对有关的问题集中整理，分常用工具、实例进行介绍。

§3.6.1 常用工具

1. E-mail：即电子邮件。是人们使用最频繁因而也最为熟悉的工具。在Unix、VAX系统中都有相应的mail命令管理邮件的收发，其它实用程序如：pine、elm 、eudora 等也可以用于邮件管理。这里提醒注意的是象DOS 执行程序、图形、WordPerfect和Word、Excel、中文WPS一类有格式文件的的发送。发送前需要事先编码。DOS执行文件是二进制文件，国标汉字其高位字节非空，其它有格式文件则包含了有关的格式定义信息，它们在直接发送时则收方无法得到这些信息，从而无法读取。编码格式以uudecode和mime比较常用。Unix上实现的命令是uuencode，收方需用uudecode进行解码。pine的“捎带(attach，即在发送邮件时在pine 中指定同时发送的文件)”功能采用的是mime格式，若收发双方均有pine，则使用其“捎带”功能不编码直接收发。否则若发方用了“捎带”方式，收方须有mime格式的解码程序(如munpack，其相应的编码程序是mpack)。这里提及的大部分软件都有Unix 和DOS两种版本，或者VAX的版本等，如DOS有实用程序进行uuencode/uudecode操作。

使用E-mail可以加入电子邮件转发系统(LISTSERV)。如HEALTH-L的加入方法：是向地址listerv@irlearn.bitnet发一个E- mail ， 其第一行内容是subscribe HEALTH-L zhao jinghua。则将会定期收到有关的信息。可以用unsubscribe 或leave一类的命令退出LISTSERV。国内影响较大的例子是通过HEALTH-L 广泛的征询获取北京大学学生朱铃铊中毒的诊断。

2. TELNET：即远程上网。允许用户在Internet 上从本地计算机建立与远程计算机间交互式登录。该命令后面所跟随的参量可以是域名或IP 地址。例如：命令telnet hollis.harvard.edu使用户与Harvard实时图书馆系统连接。使用IP地址的相应命令是：telnet 128.103.60.31。域名服务器(DNS)中提明系统命名(hollis. harvard.edu)与IP(128.103.60.31)间的对应关系。退出远程登录可用exit、bye等。

   telnet的另一种功能是核实Internet上某人的E-mail地址：telnet gwh. bmi. ac.cn 25

   使用命令：vrfy yaoc

   系统则提示该用户的真实E-mail地址，有时系统将这一功能取消，您会得到“被拒绝”的信息。这种查询方式以quit命令退出。

   Archie：即查挡案。是一个特殊的服务器/数据库，保存了Internet 上不记名FTP(见下面小节的说明)地址的文件目录，其名字与地址，其信息按月更新。不记名ftp到archie.ans.net, cd pub/archie/doc, 索取whatis.archie。一些有用的命令形象地罗列如下：

archie> help	帮助
archie> done	结束帮助
archie> prog <string>	搜索字串
archie> mail <email address>	将结果发往指定地址
archie> help set	设置帮助
archie> manpage	显示用法指南
archie> bye or quit	退出
archie> list	archie 名表

   使用archie.internic.net/archie，键入命令：prog ¡pkunzip¿则给出实用程序pkunzip所在地址，该软件用于打开.zip形式压缩的文件。第三部分给出了世界上不同地区的archie地址，列出尽可能多的地址可以在某个地址使用用户太多时更换。即便如此，其进入方式也经常更换，如使用：ds.internic.net，系统则提示使用ds1.internic.net。

   其它的如WAIS是一个分布式文本查寻系统，也可经telnet进入。

3. FTP：即文件传输。该工具允许Internet两个计算机(PC 与网络节点机之间或Internet上的两台计算机)间的文件交换。

   不记名FTP允许用户不记名或作为客人方式与远程计算机连接，从而传输公用文件如软件、文本等。多数系统上使用的命令如下：

ftp> help/?	ftp 命令目录
ftp> open/close	连接和关闭ftp 主机或ip地址
ftp> ls/dir/ldir	文件目录
ftp> cd/lcd	转换目录
	(Unix/VMS与分别用../[-] 表示上级目录)
ftp> prompt	交互回话开关
ftp> get/put	收发文件文件
ftp> mget/mput	索取或发送多个文件
ftp> pwd/lpwd	显示工作目录
ftp> ascii/binary/image	传输状态
ftp> !	系统外壳，如列本地目录：!dir
ftp> bye/quit	退出

如要索取美国疾病控制中心(CDC)的EPI INFO 6.0，可以直接使用：

ftp://ftp.cdc.gov

打入账户名anonymouse和您的E-mail地址做为口令，进入/pub/epi/epiinfo子目录，打get命令索取，别忘记预先使用binary。

根据系统的不同，命令略有不同，如最早FTP到york.cpmc.columbia.edu 索取有关程序，因为系统是VAX/VMS所以用cd [pub]命令，用cd [.dirname] 到下级目录，用cd [-] 返回到上级目录；后来因为换成Unix ，地址为：linkage. cpmc. columbia.edu/pub, 则用cd /pub。若熟悉远程计算机系统，则可得到更多的信息，如对方计算机为Unix，ls */*则可列出其下一层目录的文件。有时ftp 索取的目录名加上.zip或.tar可以直接取得按目录压缩或合并后的文件。

不同系统文件压缩/存档的说明ftp.cso.uiuc.edu/doc/pcnet/compression得到。

4. gopher：即信息鼠。是一个提供层次式驱动菜单进行信息检索的计算机程序，交互式的特点大大方便了Internet计算机系统间的访问，如经美国密尼苏达大学查询国际上有关机构、人员的电话号码、E-mail地址等。在gopher菜单中使用等号键(”=”)可以显示您所漫游(navigate)到的位置，很有用。如果使用gopher 索取软件，则较ftp方便。

   使用(”a”) 命令可以设置自己喜欢的地址从而形成新的菜单。Veronica 和Jughead是两个程序，使用关键字搜索所有gopher上的菜单标题、文件名。

5. USENET，用户信息网。是各类专题的“兴趣小组”(discussion group) 。如果您的系统提供了该功能，则使用rn、trn、tin、pine，netscape 等软件申请加入或退出。

   TRN 是rn(readnews) 的”threaded”(文章以回答的次序连接) 版本。每个discussion thread 是一个文章树，所有的派生信息(child)从原来的(parent) 文章中分支。常用命令有：

h	帮助
a ¡pattern¿	模式匹配
c	设“读完”标记
=	列出科目
_ a	逐条列表，使一系列命令生效
s/w	存为文件
q	退出

如：comp.soft-sys.sas、comp.soft-sys.spss是有关SAS和SPSS 的讨论组。sci.math.symbolic则与符号计算如mathematica有关。

上面几种工具，对于初学者来说有些繁琐，但它们最终可以简单地用单个软件来实现，它们称为browser，如：netscape、mosaic和lynx等，以netscape 最为流行。这些软件只消借助于鼠标或光标键进行菜单操作。

在netscape下，可以直接使用http、ftp、gopher、telnet等命令，也可以直接读取NEWSGROUP信息，这样做的好处是只需打鼠标就可以了，而且转换地址极为方便。netscape中按类别搜索常用如Yahoo(http://www.yahoo.com)、Lycos等，也可形成自己喜欢的标记(bookmark)，真正做到“随叫随到”。

§3.6.2 应用举例

Internet最主要的功能是信息的传递与检索，从而成为“信息高速公路”的支柱。信息的内容有如时事性新闻、广告、人物、最新的动向、研究成果、出版、电子杂志等；它们称之为“多媒体”，即文本、声音、图象与电影等各种媒介的集合体，文本包含了与其它文档的联接。

计算机系统维护的信息、实用工具程序，解决硬软件系统问题，可以通过以上工具来“开发”。如http://wuarchive.wustl.edu，或者通过archie寻找。如运行C++教科书的程序书中可以指明从何处可以得到这些标准的程序。

Internet可以成为一个“图书馆”：如进行分子生物学研究要了解某种疾病如囊性纤维化(cystic fibrosis)的概况，可以获得特别详细的综述、所涉及文献的MEDLINE摘要等，您还可以得到最新的基因标记图。

从SAS公司可以得到SAS样本程序。ftp://lib. stat. cmu. edu 可以得到英国Applied Statistics 上的标准算法、Splus、Minitab等的宏程序等。从http: //econwpa.wustl.edu/limdep.html 可以得到计量经济学分析软件LimDep第七版的完整说明书。Numerical Recipes 亦可以经cfatab.harvard.edu/nr/bookf.htr取。

各种电子邮件转发系统和“兴趣小组”也提供了卫生方面研究的课题，如上面提到的HEALTH-L。

以上我们简单介绍了Internet常用的功能，许多新工具如JAVA限于篇幅不能一一介绍。进入Internet需要在Internet的计算机上拥有账户，可以经调制解调器(Modem)、Ethernet/PCMCIA组合卡上网，软件有Kermit、PC/TCP等。经Kermit和电话线(dial-in)联接，可以使用大部分Internet功能包括lynx，但需要PPP/SLIP 方可以使用netscape与mosaic。象packet 驱动程序3c503、美国高速计算机研究协会(NCSA)的FTP与TELNET软件也是免费提供的，如果感兴趣，您可以研究一下它们的源程序。目前网络服务器多采用Unix操作系统。Internet还赖于局网的有效使用。

【附】计算机应用问题与解答(Q&A)
下面结合MS-DOS 下常碰到的问题，给出尝试性的答案：

Q . ASCII 码表控制缩写字符(如EOT, LF)的含义是什么？

A . 这里给出其英文全称，请见下表。

Q . 请给出计算机数制符号表？

A . 见下表：

Q . DOS 5.0 较DOS3.30增加了哪些功能？

A . DOS 3.30 相对于它以前的版本，已相当完善，故目前用户仍然很多；而DOS 5.0 是经DOS 4.x的过渡又一次飞跃。新增加的命令有：DOSKEY、DOSSHELL 、EDIT、EMM386、EXPAND、LOADFIX、LOADH
这还不包括CONFIG配置方面所做的改进。EMM386.EXE，是与VCPI相兼容的EMS管理程序，用于将XMS转为高存块UMB，LOADHIGH 及DEVICEHIGH可将TSR 及设备驱动程序装入UMB。另外，DOSSHELL的OPTIONS 下的ENABLE TASK SWAPPER 可激活多任务切换功能，使PC进行前后台处理；可支持高达2GB的硬盘分区；DOS支持/? 询问命令所需参数，使用MIRROR对分区表进行务备份以防病毒等。

Q . 如何制做一个系统盘？

A . 使用微机时，通常应备有一个与硬盘DOS系统版本相同的系统盘。对于新盘可用DOS的格式化命令FORMAT加参数/S生成系统。或者用/B 参数预留空间后，用SYS d: 命令结合COPY命令把COMMAND.COM命令拷到目标盘上。新版DOS改进了SYS 的功能。使用SYS命令制作系统盘时，常常显示No room on destination disk for the system 是由于系统文件所在引导区被占用而致，有时即使显示SYSTEM TRANSFERRED，仍不能启动，在软硬盘均可以出现这种情况。建议用NORTON 等软件进行处理。如利用NORTON 4.5 中的NDD，择Common Solution，Make a Disk Bootable。手上无软件、软盘上无其他内容，则可用FORMAT [d:] /S 做系统盘。还有一种情况是由于某种原因导致了硬盘分区顺序错误，可用PARTED软件纠正，它在长城DOS3.2版中用PFC.BAT文件调用。

Q . 使用计算机时硬盘系统不能自举，如何处理？

A . 常见是由于软件原因。若是系统文件损坏，则应重新拷贝COMMAND. COM 和两个隐含文件(用SYS [d:])。若由于病毒感染使分区信息丢失，则可用软件PARTED.EXE，只要操作一下，问题往往就可以解决。若问题较大，则建议做一下低级格式化。有时磁盘信息丢失很多，也可以用NORTON NDD 中的RECOVER功能找回子目录，目录命名为DIRXXXX，通过其中的文件确定真实的目录名，用PCTOOLS或DOS6.0的MOVE功能将子目录改回原名。

Q . DOS 的命令很多，难于记忆，有简便的方法吗？

A . 比较好的方法是使用DOS 5.0它提供了HELP 命令，在系统提示下打入一个命令，系统给出一个DOS 命令的清单或表，打入命令HELP 加上特定命令则给出相应命令的用法，

表 3.5 ASCII 码表[0-127]

ASCII 字符	Hex 码	控制字符	ASCII 字符	Hex 码	控制字符	ASCII 字符	Hex 码	控制字符
NUL	00	Null	+	2B		V	56	
SOH	01	Start Heading	’	2C		W	57	
STX	02	Start text	-	2D		X	58	
ETX	03	End text	.	2E		Y	59	
EOT	04	End transmission	/	2F		Z	5A	
ENQ	05	Inquery	0	30		[	5B	
ACK	06	Acknowledgment	1	31		\	5C	
BEL	07	Bell	2	32		]	5D	
BS	08	Backspace	3	33		^	5E	
HT	09	Horizontal tab	4	34		_	5F	
LF	0A	Line feed	5	35		‘	60	
VT	0B	Vertical tab	6	36		a	61	
FF	0C	Form feed	7	37		b	62	
CR	0D	Carriage return	8	38		c	63	
SO	0E	Shift out	9	39		d	64	
SI	0F	Shift in	:	3A		e	65	
DLE	10	Data link escape	;	3B		f	66	
DC1	11	Device control 1	<	3C		g	67	
DC2	12	Device control 2	=	3D		h	68	
DC3	13	Device control 3	>	3E		i	69	
DC4	14	Device control 4	?	3F		j	6A	
NAK	15	Neg.acknowledge	@	40		k	6B	
SYN	16	Synchronous/Idle	A	41		l	6C	
ETB	17	End trans. block	B	42		m	6D	
CAN	18	Cancel data	C	43		n	6E	
EM	19	End of medium	D	44		o	6F	
SUB	1A	Start special seq.	E	45		p	70	
ESC	1B	Escape	F	46		q	71	
FS	1C	File separator	G	47		r	72	
GS	1D	Group separator	H	48		s	73	
RS	1E	Record separator	I	49		t	74	
US	1F	Unit separator	J	4A		u	75	
SP	20	Space	K	4B		v	76	
!	21		L	4C		w	77	
”	22		M	4D		x	78	
#	23		N	4E		y	79	
$	24		O	4F		z	7A	
%	25		P	50		{	7B	
&	26		Q	51		—	7C	
’	27		R	52		}	7D	
(	28		S	53		~	7E	
)	29		T	54		DEL	7F	Delete

表 3.6 计算机的数制表

数目	前缀	简写	数值	科学记数法
Quinillion	Exa	E	1,000,000,000,000,000,000	1E+18
Quadrllion	Peta	P	1,000,000,000,000,000	1E+15
Trillion	Tera	T	1,000,000,000,000	1E+12
Billion	Giga	G	1,000,000,000	1E+9
Million	Mega	M	1,000,000	1E+6
Thousand	Kilo	K	1,000	1E+3
Hundred	Hecto	H	100	1E+2
Ten	Deka	Da	10	1E+1
One			1	
Tenth	Deci	d	0.1	1E-1
Hundredth	Mili	m	0.001	1E-3
Millionth	micro	μ	0.000 001	1E-6
Billionth	nano	n	0.000 000 001	1E-9
Trillionth	pico	p	0.000 000 000 001	1E-12
Quadrillionth	Femto	f	0.000 000 000 000 001	1E-15
Quantillionth	atto	a	0.000 000 000 000 000 001	1E-18

如：HELP DIR，或DIR/?。不熟悉英文时，不妨作为引子去查说明书。与DOS 6.0又使这种帮助更加完善。另外，DOS 的命令尽管很多，但经常使用也就熟悉了。

Q . Abort, Retry, Ignore, Fail ? 是何意思?

A . DOS 运行错误信息，放弃，重试，忽略或无效。以DIR 为例，用它显示另一驱动器上盘的目录，但该驱动器盘并未准备好，则可以打A(bort)，插好盘后再打R(etry)，拷文件时，用I(gnore) 则可以跳过出错的部分。在列未格式化盘的目录时，显示General Failure Error，则打F(ail)，之后输入有效驱动器的名字。

Q . 什么是DOS 文件指示?

A . 包括文件所在的磁盘、路径和文件名，如C:\DOS\ FORMAT. COM 表示C: 盘上DOS 目录下的文件FORMAT.COM，它用于磁盘格式化。

Q . 什么是文件属性，如何改变?

A . DOS 文件的属性规定了文件的几种特性，如A (Archive) 表示修改记录，R(Read-only) 表示只读，H (Hide) 表示隐藏，可以使用ATTRIB 命令来改变，方法是在文件指示前用这些记号，加上前缀设置(+) 或取消(-)，如DOS 5.0 下使用命令ATTRIB +H CONFIG.SYS 将使用系统配置文件为隐藏属性，执行后使用DIR 命令就不可见了，以前版本的DOS可用CHKDSK /V 来观察或使用PCTOOLS 等工具。ATTRIB 命令可以对整个目录进行操作，当不设定属性而只用文件指示时将显示现有文件的属性。DOS 6.0 的ATTRIB命令增加了对隐含属性的操作，也可以在DIR 命令中使用DIR /A[属性] 来显示，如：DIR /AH。

Q . 如何删除BACKUP.XXX 文件？

A . 在使用较高版本的DOS 进行BACKUP 时，软件内容为CONTROL.### 和BACKUP.###，要删这些文件时，计算机提示Access denied，实际上是对它们进行了保护，可用ATTRIB -R <文件名> 先把文件属性改成读写，再删除之。

Q . 由于操作不慎，删除了需要的文件，该怎么办？

A . 可以使用PCTOOLS 和NORTON 等工具软件提供的UNDELETE 功能，在DOS 5.0 或6.0下用UNDELETE.EXE 对误删文件进行恢复。删除文件实际上只是在文件上打了删除标记，要进行恢复，指示文件名的首字母，其余文件名、扩展名均保留。注意若被删文件的空间已被占用，则原删文件不能恢复。要彻底删除文件，用NORTON 等的WIPE 功能，这时无法UNDELETE。有时盘也可能误格式化，使用DOS 5.0 和6.0的UNFORMAT.COM 工具也可以恢复到格式化前的状态。

Q . 软件生成的数据或结果太大，保存到一张软盘上时放不下怎么办？

A . 原始数据太大，可以首先考虑文件的压缩。如PKARC.COM程序对dBASE Ⅲ数据库文件甚至可以压缩到只有原来的90的，压缩对于网络传输也很有必要。使用PAK.EXE 等软件可直接生成可执行文件，在打开原来文件时，可以不用原来的压缩程序，对于软件散发很便利。不太常用的数据在存放时，建议也进行压缩以节省磁盘空间。

有时候压缩后的文件仍然太大，而且压缩方法未必奏效，这时备份应当使用BACKUP命令，该命令由DOS 提供，其含义是盘的备份，命令格式为BACKUP [d:] path\filename.exe [/s] [/a] [...] [d:] 命令，即把源盘硬盘上的文件连续备份到软件盘上，当一张软件盘的空间不够时，系统提示换下一张盘，直到备份结束。这样的好处是充分利用了软盘，缺点是当备份后一段时间后，其中一张软盘出现问题相关的内容会受影响。注意这样生成的文件经常返回不到硬盘，其原因在于没有使用[/s] 参数，指示文件连同所在路径一起拷贝。又这样拷贝时，软盘上根目录的文件会丢失，当软盘没有格式化时，可以指定[/f] 参数指定格式化，这样在DOS 5.0 下可以边格式化边拷贝，还可以使用[/a] 参数，这样可以原有备份的最末继续备份。另外，该命令也有其它的用法，可参考DOS 说明书。不同DOS 版本间BACKUP 文件常不兼容，应当与同版本的RESTORE 文件相匹配，或者使用其他工具如FASTBACK，PC-BACKUP ，SAS 有自己的备份工具。DOS6. 0使MSBACKUP命令进行备份，其功能类似于FASTBACK，同时DOS6.0的RESTORE.EXE兼容其它版本的备份文件。

Q . 使用硬盘时，空间忽然不够，是怎么一回事？

A . 往往是由于簇(Cluster) 的丢失，这时可以用DOS 的CHKDSK [d:]/F ，提示您是否把丢失的簇做成文件时，回答No。建议此命令经常使用。在DOS6.0 中提供了DBLSPACE实用程序，可使所使用的盘的可用空间增大一倍。一些碎片也可经相应的程序除去。注意丢失和簇恰可以用来找回丢失的文件，如四通经常出现坏盘现象，用CHKDSK或RECOVER命令很有用。

Q . 计算机运行时，显示环境空间不够怎么办？

A . 这在网络环境下就会碰到，应对系统设置文件CONFIG.SYS 内容进行改动，增加命令如：SHELL=C:\DOS\COMMAND.COM /P /E:256。

Q . 如何加快软盘的拷贝速度?

A . 刚接触计算机的用户，往往是先格式化一张空盘，然后用命令如COPY A:*.* B:，这样很费时，而且容易漏掉子目录的内容，建议使用DOS 系统提供的DISKCOPY 命令，该命令用于整盘拷贝，特别是系统盘的拷贝。但在高密盘的拷贝时往往仍然要换三次盘。实际上，可以使用PCTOOLS 工具的磁盘操作功能。条件是预先设定扩展内存EMS，若计算机的RAM 较大，一般一次即告完成。文献上有实用的方法，不妨一试。

Q . 如何拷贝子目录的内容?

A . 可以使用DOS 提供的COPY命令，但这往往会遗漏子目录的内容，若是整盘拷贝则用DISKCOPY 命令，若仅拷子目录，建议用XCOPY /S [/E]命令。它可以直接生成子目录，子目录的内容也不易遗漏。利用XCOPY 可以用更复杂的选择项，如指定拷贝特定日期以后的文件等，可参考DOS 说明书。

Q . 如何尽快删除嵌套子目录的内容?

A . 介绍一种用PCTOOLS 的方法。首先在它的磁盘功能屏下使用查找(FIND)，此时按照提示给出盘名和要寻找的内容(目录名)，当显示目录区后选择编辑(EDIT) 参照其目录属性标记,把目录属性变为文件，修改后写盘，退出PCTOOLS，删除文件，再运行DOS的CHKDSK/F，在提示下回答N把磁盘中的丢失的簇修好即可。

DOS 6.0 提供了一个新的命令DELTREE，可用于删除目录树。该命令不受DOS 版本的限制。

Q . 要编辑二进制文件怎么办?

A . 这一用法不普遍，万一使用，应当使用EDLIN /B 的方式。通常的文件结束符是^Z，一般软件读取文件到此为止，而EDLIN 的的这一用法则不然，其它的如PCTOOLS 的文件操作也可以做到。

Q . 如何在高密驱动器上格式化低密盘?

A . 要启用Format 命令行参数：若是5.25盘，则应用Format [d:] /4，若是3.5 盘，则用Format [d:]/n:9。此二参数的结果是格式化成360K 和720K 盘，其他情况可做类似处理。在DOS 5.0 和6.0下，有一个方便的处理办法，即指定[/F:size] 参数。预先使用FORMAR /? 命令有相应的提示。

Q . 一些软件的屏幕图形怎样打印?

A . 建议在运行软件之前，预先调入GRAPHICS，然后使用屏幕打印。如BASIC 的图形和今后将介绍的Stata 软件的图形就可以这样印出，但最好还是利用软件原有的功能，如Stata 使用GPHDOT 和GPHPEN。

Q . 在西文状态运行中文软件，退出软件后计算机屏幕光标看不到怎么办?

A . 许多人采用热启动的方法，实际上不必这么做。只要使用下述命令即可：MODE 80 或MODE CO80。一些汉字系统如联想下，使用MODE命令效果不明显，可试以WordPerfect 5.1 中的CURSOR工具。

Q．使用WordStar 时，不能调其他目录的文件，怎样才能调用呢？

A．这是指用WordStar 3.x 时的情况。这时可以用DOS提供的工具，把子目录与驱动器联系起来，如设子目录C:\USER 为G盘，则可用SUBST G: C:\USER，但在config.sys 中应加上Lastdrive=G。那么在编辑时用G:文件名即可。亦可用前面介绍过的APPEND 和PATH 的组合。注意在WordStar 4.0 以后可不必这么做。

Q．在Wordstar 中，如何制做矩形块？

A．应当用^KN 把块定义改成列方式，仍然用^ KB 矩形块的块的左上角，用^KK 定义块的右下角。矩形块的操作同行快。块操作以PEII 软件最为灵活。

Q．软件包软件的结果文件太大，编辑软件调不进怎么办？

A．在文件太长太宽时容易出现这种情况。最好是避免此类问题的出现，碰到时可用实用的高级语言程序或软件包进行简单的读写，从而分成小文件进行处理。如下面的小程序把HUGE.LST分成PART1.LST与PART2.LST文件。

```
100 REM 分割文件程序—1993.11.12
110 OPEN "I",#1,"HUGE.LST"          '打开1号通道作为输入
120 OPEN "O",#2,"PART1.LST"         '打开2、3号通道作为输出
130 OPEN "O",#2,"PART2.LST"
140 I=0                             '计数开始
150 WHILE NOT EOF(1)                '文件结束则停止
160    LINE INPUT #1,LINE$          '读取一行
170    I=I+1                        '标记本行
180    IF I>6000 THEN 200           '写哪一个文件?
170        PRINT #2,LINE$           '写第2号文件PART1.LST
180        GOTO 200
190    PRINT #3,LINE$               '写第3号文件PART2.LST
200 WEND
210 PRINT CHR$(7);"END"             '振铃并结束程序
END
```

源程序采用GWBASIC写成，结果是生成的第一个文件有6000行，软件可以接受。

Q．BASIC 源程序和结果如何打印？

A．用LLIST 即可，BASIC 运算结果用PRINT，也可以用PRINT 代替LPRINT，方法是在IBM PC机上用^*或SysRq。

Q．常见计算机中颜色的英文名，怎么译法？

A．有下面的表供参考，如在FoxBASE+中各种颜色简单的缩写，用SET COLOR 设定。在DOS下启用颜色可用ANSI.SYS驱动程序。

Q．SAS 的OUTPUT 窗口存贮的文件，在打印时往往疯狂走纸，该怎么办？

表 3.7 颜色的名称

编号	英文名	译 名	编号	英文名	译 名
0	Black	黑色	8	DarGray	深灰色
1	Blue	兰色	9	LightBlue	浅兰色
2	Green	绿色	10	LightGreen	浅绿色
3	Cyan	深兰色	11	LightCyan	浅青色
4	Red	红色	12	LightRed	粉红色
5	Magenta	深红色	13	LightMagenta	浅红色
6	Brown	棕色	14	Yellow	黄色
7	LightGray	浅灰色	15	White	白色

A . 这是由于SAS 为了结果的显示方便，把纸的页长设成21，而通常的打印机纸长是66 行。而且SAS 在每页的开始处加了打印走纸符^L，所以为避免这个问题，可在SAS 运行时就利用OPTIONS 语句，把页长设得大些，同时在结果存贮后，不直接打印，先用编辑软件如PEII 把走纸符去掉。SAS 的页长、行宽选择对报告的生成过程如PROC TABULATE 的影响是很大的。在SAS 本身也可以去掉，方法是，预先把结果文件输出到一个外部文件，然后在程序编辑窗口调入，置光标到窗口命令行，并使用命令CHANGE '^L' ' ' ALL 。如对SPSS/PC+ 等的输出文件处理方法类似。注意这些操作中^L 的录入方法是先按住键盘的ALT 键，再打小键盘上的数字12，然后使手离开键盘，其它ASCII 码的录入与此相仿。

Q . SAS 的运行结果可以边运行边存贮或打印吗？

A . 在SAS 可以用PROC PRINTTO 过程，建议在运行结果比较多时使用该过程，因为OUTPUT 窗口贮存的行数是有限的。此过程的格式是PROC PRINTTO PRINT='文件名' LOG='文件名';RUN;如要把运行信息打印出来，则择LOG='PRN:' 即可。若仍要在LOG 窗口内显示，则复用PROC PRINTTO; RUN;

Q . 什么是高速缓冲，怎么使用？

A . 高速缓冲(cache) 是RAM 中的一个区域，用于存贮经常使用的信息，从减少访问磁盘的次数。在WINDOWS 3.X、DOS 5.0 及DOS 6.X 中的SMARTDRV 能够真正使用cache。DOS 用于缓冲的方法还有在CONFIG.SYS 中设置BUFFERS= 选项，在CONFIG.SYS 或命令启用FASTOPEN 等等。现在常用磁盘120 MB 以上的计算机，可以设置BUFFERS=50 以上。PCTOOLS 中的PC-CACHE 工具及NORTON 中的NCACHE 都可以用于高速缓冲。

Q . 如何设置不同的运行环境？

A . DOS 6.0 以下的系统，不同的配置可以保存在不同的文件CONFIG. SAS，CONFIG.WPS，CONFIG.NET 等。由于系统使用CONFIG.SYS，则可用不同配置文件替换，使用时先用命令：COPY CONFIG.XXX CONFIG.SYS 进行热启动则配置有效。

在MS-DOS 6.0已对这种不经济的方法进行了改进，不仅使CONFIG.SYS使用的命令增多，而且允许在多种配置中进行选择。一面是一个用例：

```
[menu]
menuitem=English
menuitem=Japanese
menucolor=15,1
numlock=off

[common]
FILES=60
buffers=20,0
FCBS=4,0
shell=c:\command.com /e:512 /p

[English]
DEVICE=C:\DOS\himem.sys
DEVICE=D:\DOS6\EMM386.EXE HIGHSCAN 1536 RAM
DOS=high,UMB
DEVICEHIGH /L:1,3136 =C:\LADDRV.SYS /D:5
lastdrive=Z

[Japanese]
country=081,932,c:\dos\country.sys
device=d:\dos\$font.sys
device=d:\dos\$disp.sys
device=c:\dos\limsim.sys 1024
device=d:\dos\$ias.sys /x=1
device=d:\dos\$iaeskk.sys /x=1 /h=1
device=d:\dos\$prnuser.sys
install=c:\dos\keyb.com jp,932,c:\dos\keyboard.sys
INSTALL=D:\DOS\IBMMKK.EXE /T /K /M=J /S=D:\DOS\$MULTDCT.PRO
```

在[menu]选项下定义供选择的配置，用menuitem具体指定配置命令。这样系统自举后系统首先给出提示供选择。DOS6.0的CONFIG.SYS文件也允许对数字键的锁定与否进行说明。

Q．文献中常出现一些英文缩写，能进行一下解释吗？

A．这里列出一部分，可以查阅计算机词典，一些软件包的缩写集中到本书附录中介绍。

IBM	国际商用机器公司
PC	个人计算机
EBCDIC	扩展的二进制编码的十进制交换码
ANSI	美国国家标准协会
ASCII	美国标准信息交换码
RAM	随机访问存贮
ROM	只读存贮
PROM	可编程只读存贮
EPROM	可去可编程只读存贮
OS	操作系统
DOS	磁盘操作系统
DEC	数字设备公司
MIPS	每秒百万次指令
MVS	多虚拟存贮
MVS/TSO	多虚拟存贮/分时选择
SQL	结构查询语言
VAX	虚拟地址扩展
VAX/CMS	VAX 代码管理系统
VAX/VMS	VAX 虚拟内存系统
VM/CMS	虚拟机/对话监控系统
FAT	文件分配表
CGA	图形适配器
EGA	增强图形适配器
VGA	视频图形适配器
EMS	扩充内存管理
XMS	护展内存管理
API	应用程序接口
GDI	图形设备接口
CD-ROM	光盘
DBMS	数据库管理系统
DB2	IBM 数据库系统
NFS	网络文件系统
DCA	文档内容体系
RFS	远程文件系统
LAN	局部网
VISCALC	电子报表VISCALC (可视计算器)
HPGL	惠普图形语言
CGM	计算机图形中间文件
TIFF	标记影象文件格式
TSR	内存驻留
OEM	设备生产厂家

与Internet有关的简写如：

AIX	- IBM 开发的Unix操作系统
anonymous FTP	- 不记名FTP
BITNET	- Because It's Time Network 一些教育网络
CERN	- European Centre for Particle Physics 欧洲粒子物理中心，瑞士
Client	- 通过网络向服务器请求服务的程序
Domain	- DNS的命名类
DNS	- Domain Name System 层次式系统命名系统
Driver	- 控制外设或输入输出(I/O)口的程序
EARN	- European Academic Research Network 欧洲教育研究网络
EMBnet	- European Molecular Biology Network
Ethernet	- 网络物理和数据链接层的标准
FTP	- File Transfer Protocol
HTML	- Hypertext Markup Language
HTTP	- Hypertext Transfer Protocol 超文本转换协议
Hypermedia	- 多媒体超文本
Hyperlink	- 在网络(Web)上两个对象的关联
Hypertext	- 包含与其它文本链接指针的特殊文本
Interrupt	- 中断，请示注意的信号
IP	- 标准Internet协议
IRIX	- SGI所开发的Unix系统
LAN	- Local Area Network 局网
MIME	- Multipurpose Internet Mail Extensions 多用途邮件扩展
NFS	- Network File System 网络文件系统
NNTP	- News Network Transfer Protocol 传递消息文本的方法
Packet	- 一条网络信息，包括头、访问信息和数据
Protocol	- 两个计算机交换信息必须遵守的规则
Server	- 提供客户服务的程序
SLIP	- Serial Line Internet Protocol 串行线协议

SMTP	- Simple Mail Transfer Protocol 简单邮件传递协议
Subnet	- 是在internetwork内的局部网络
TCP/IP	- Transmission Control Protocol/Internet Protocol 传输控制协议/Internet协议
TELNET	- Internet 远程终端连接的标准协议
ULTRIX	- DEC所开发的Unix操作系统
Unix	- Bell 实验室开发的操作系统
URL	- Uniform Resource Locator 统一资源定位
VM	- IBM开发的操作系统
VMS	- DEC开的操作系统
WAIS	- Wide Area Information Server 广域信息服务器
WWW	- World Wide Web 环球信息系统

Q . 请问活动的archie服务器?

A . 多数地址可以由Internet系统给定，但有时需要用户指定，它们是：

archie.au	139.130.4.6	Australia/New Zealand
archie.edvz.uni-linz.ac.at	140.78.3.8	Austria
archie.univie.ac.at	131.130.1.23	Austria
archie.uqam.ca	132.208.250.10	Canada
archie.mcgill.ca	192.77.55.2	Canada
archie.funet.fi	128.214.6.102	Finland/Mainland Europe
archie.univ-rennes1.fr	129.20.128.38	France
archie.th-darmstadt.de	130.83.128.118	Germany
archie.ac.il	132.65.16.18	Israel
archie.unipi.it	131.114.21.10	Italy
archie.wide.ad.jp	133.4.3.6	Japan
archie.hana.nm.kr	128.134.1.1	Korea
archie.sogang.ac.kr	163.239.1.11	Korea
archie.uninett.no	128.39.2.20	Norway
archie.rediris.es	130.206.1.2	Spain
archie.luth.se	130.240.12.30	Sweden
archie.switch.ch	130.59.1.40	Switzerland
archie.nctuccca.edu.tw		Taiwan
archie.ncu.edu.tw	192.83.166.12	Taiwan
archie.doc.ic.ac.uk	146.169.11.3	United Kingdom
archie.hensa.ac.uk	129.12.21.25	United Kingdom
archie.unl.edu	129.93.1.14	USA (NE)
archie.internic.net	198.49.45.10	USA (NJ)
archie.rutgers.edu	128.6.18.15	USA (NJ)
archie.ans.net	147.225.1.10	USA (NY)
archie.sura.net	128.167.254.179	USA (MD)

第二部分

通用统计分析软件包

第四章　SAS

§4.1　SAS 系统导引

§4.1.1　简介

SAS 最早由美国北卡(North Carolina) 的A.J. Barr 和J.H. Goodnight 等于六十年代设计。目前，它已成为集多种功能于一身的完善的应用系统，并且公认为第四代计算机语言的代表。SAS 公司于1976 年组成，总部设在美国北卡CARY，在世界各地都有相应的办事机构。

- SAS 的基本功能主要有以下几个方面。
- 数据的录入(data entry and retrieval)
- 报告撰写(report writing)
- 图形(graphics)
- 统计分析(statistical analysis)
- 商业计划与预测(business planning and forecasting)
- 应用开发(applications development).

VAX/VMS SAS 6.07 功能的描述很好地说明了这一点(此处仅仅突出了数据分析)，它们是：

建模与分析工具(Modeling & Analysis Tools)
项目管理(project management)
质量控制(quality improvement)
计量经济学与时间序列(econometrics and time series)
数据分析(data analysis)
　回归分析(regression)
　方差分析(analysis of variance)
　分类数据分析(categorical data analysis)
　基础统计分析(elementary statistics)
　多元分析(multivariate data analysis)
　实用程序(utility)
　时间序列(time series)
　聚类分析(clustering)
　生存分析(survival analysis)
　判别分析(discriminant analysis)
　方程组(systems)
　控制图(control charting)
预测(forecasting)
实验设计(experimental design)

商务应用(financial application)
运筹学(operations research)
交互式矩阵语言(interactive matrix language)

SAS 的功能由其相应的产品(products) 或模块完成，较主要的模块有：
SAS/BASE：基础模块，是一个通用的数据管理、录入和报告书写工具。
SAS/STAT：统计模块。SAS 的大部分统计功能经此模块实现。
SAS/AF：全屏幕交互式应用开发工具，用于制作菜单、设计教学和热线帮助。
SAS/FSP：数据录入、查询、书信、报告模块
SAS/GRAPH：绘图模块。
SAS/IML：矩阵程序语言(Interac-tive Matrix Language)。
SAS/ETS：时间序列分析、预测和计量经济学分析
SAS/OR：运筹学和项目管理。
SAS/QC：质量控制。
其中基础模块是运行SAS系统所必需的。

SAS/RTERM 模块，是一个终端仿真程序，能使PC机产生高分辨硬拷图形，可使微机方便地与VAX 等小型机联用。其它的模块有：

SAS/ACCESS 提供不同系统及文件间的透明接口，如：MVS、CMS、VSE、OPEN VMS VAX、OPEN VMS AXP、Solaris、HP-UX、RS/6000 AIX 、OS/2 、Windows 、Windows NT。

SAS/ASSIST	菜单驱动的用户接口
SAS/CALC	电子报表
SAS/CONNECT	分布处理软件
SAS/CPE	计算机系统评价、功能规划和网络管理
SAS/EIS	决策信息系统开发工具
SAS/ENGLISH	查询和报告的自然语言接口
SAS/GIS	空间地理数据分析
SAS/IMAGE	数字和扫描图象处理
SAS/INSIGHT	交互式图形程序
SAS/LAB	交互式导引程序
SAS/NVISION	三维图形程序
SAS/PH-Clinical	药物和生物技术行业用程序
SAS/SHARE	网络数据共享软件
SAS/TOOLKIT	程序开发工具箱
SAS/TUTOR	教学

SAS 强大的功能是通过SAS 程序完成的。SAS 程序是针对SAS 系统的指令序列。其程序编写非常简便，主要由大量的数据步(DATA STEP) 和过程步(PROC STEP)组合而成。数据步是SAS 程序中产生数据集的部分，SAS 数据集是一个由SAS 系统产生的具有特殊组织的文件。SAS 过程(procedures) 是预先写好的程序，可以被用于对SAS 数据集进行各种操作。过程步是SAS 程序是调用SAS 过程的部分。

SAS 软件进行统计分析并非仅仅基于统计模块。其实验设计可用SAS/STAT 和SAS/QC 完成，其时间序列分析则使用SAS/ETS 。数据的操作与描述主要用SAS/BASE。使用SAS/BASE 和SAS/STAT 进行描述常常结合SAS/GRAPH。本章对这些方面略做介绍。SAS 软件对这些

产品都提供了丰富的实例，本章也列举出来。

§4.1.2 SAS 的运行

SAS 有三种常用的运行方式：

1. SAS -NODMS (不使用显示管理系统)，在问号(?) 提示下依次打入SAS 命令。命令之间以分号(;) 分开。其特点是边执行边可查看执行结果，也较节省内存。此工作状态以ENDSAS; 语句退出。

【例4.1】产生1-200 的均匀分布伪随机数，设系统存放文件名是RAND.SAS，其内容如下：

```
data rand;
do i=1 to 10;
   y=int(200*ranuni(123)+1); output;
end;
proc print; var y;
run;
```

键入指令，D:\SAS>sas -nodms

系统显示：

```
NOTE: Copyright(c) 1985,86,87 SAS Institute Inc., Cary, NC 27512-8000, U.S.A.
NOTE: SAS (r) Proprietary Software Release 6.04
      Licensed to MINISTRY OF PUBLIC HEALTH, Site 15670003.

NOTE: Additional user information:
      the Ministry of Public Health, P.R.C.

NOTE: AUTOEXEC processing completed.

    1? %include rand;

NOTE: The data set WORK.RAND has 10 observations and 2 variables.
NOTE: The DATA statement used 10.00 seconds.
                                    SAS     15:52 Wednesday, June 3, 1992   1

                                  OBS     Y

                                    1    151
                                    2     65
                                    3     36
                                    4    182
                                    5     72
```

```
                                    6     45
                                    7    158
                                    8     80
                                    9     25
                                   10     38
NOTE: The PROCEDURE PRINT used 4.00 seconds.
    2? endsas;
NOTE: SAS Institute Inc., SAS Circle, PO Box 8000, Cary, NC 27512-8000


D:\SAS>
```

程序中使用%include 语句调用rand.sas 而运行。

2. SAS <SAS 程序名> 类似于第一种工作方式，尤适于使用中文系统时。运行结束，SAS 自动退出。执行情况可以由.LOG 而看出，根据运行结果，随时调用中文编辑软件，修改源程序。系统运行的结果存于.LST 文件中。现运行程序RAND.SAS，使用命令D:\ >SAS RAND <Enter>，结果存放在RAND.LST，运行信息记录在文件RAND.LOG。

```
D:\SAS>sas rand


NOTE: Copyright(c) 1985,86,87 SAS Institute Inc., Cary, NC 27512-8000, U.S.A.
NOTE: Source statements read from file D:\SAS\RAND.SAS,
      log listing written to file D:\SAS\RAND.LOG,
      procedure output, if any, written to file D:\SAS\RAND.LST.
```

运行结果存放在文件RAND.LST 中，RAND.LOG 存放着系统运行信息。

3. SAS 或者SAS -DMS (默认方式)，进入SAS的显示管理系统(DMS) 下。以窗口命令行或运行程序内的ENDSAS 语句退出或在窗口命令行上打BYE、ENDS 退出。在VAX 机上，若使用终端VT382，则使用命令SAS /FSD=VT382 进入此方式。现运行RAND.SAS，可以在程序窗命令行上使用命令INClude 'RAND.SAS'，然后使用提交命令SUBMIT。也可在程序区使用%INCLUDE 'RAND.SAS'; 并提交运行。此方式里还可单步(SUBTOP)、菜单式(AF 或MENU) 和一次性提交(SUBMIT)。填充式执行PROC UNIVARIATE 的显示如下：

```
UNIVARIATE: DESCRIPTIVE STATISTICS
Command ===>

 PROC UNIVARIATE   DATA  =
                          -----------------
    Printed output options:
    NOPRINT  [ _ ]    No printed output.
    PLOT     [ _ ]    Stem and leaf and other plots
    FREQ     [ _ ]    Frequency table
```

```
    NORMAL   [ _ ]    Test statistic for normal distribution

    VARDEF  =  ________   Divisor for calculation of variances.
               Choices: DF N WEIGHT WDF

    ROUND   =  ___________________  Units to round variable values;

 VAR    ..................................................... ;
 BY     ..................................................... ;
 ID     ..................................................... ;
 FREQ   ........ ;
 WEIGHT ........ ;
```

图4.1(a) SAS 填充式运行用例

第二屏的内容为：

```
Output data set and output options:

 OUTPUT OUT =
              _________________
 Enter the output options as: 'keyword' = 'varname(s)', where
 'keyword' represents a statistic and 'varname(s)' are the names of
 the variable or variables to contain the statistic.
 Some keywords are:

 N NMISS NOBS MEAN SUM STD VAR SKEWNESS KURTOSIS SUMWGT MAX MIN RANGE
 Q3 MEDIAN Q1 QRANGE P1 P5 P10 P90 P95 P99 MODE SIGNRANK NORMAL.

 Enter the output options:

 ..............................................................
 ..............................................................  ;
```

图4.1(b) SAS 填充式运行用例(续)

管理系统有若干个窗口，但通常只有OUTPUT (输出窗口)、LOG (登录窗口)、和PGM (程序窗口)被激活。OUTPUT 存放运行的结果，LOG 存放运行信息，PGM则用于程序编辑。有关的窗口还有：HELP (帮助窗口)、AF (辅助窗口)、MENU (填充式执行)、CATALOG (目录文件)、LIBNAME (库窗口)、DIR (目录窗口)、VAR (变量窗口)、TITLE (标题窗口)、FOOTNOTE (脚注窗口)、NOTEPAD (记事窗口)、KEYS (功能键窗口)、OPTIONS (选择窗口)、SETINIT (初始化窗口)。它们有一些共同的命令，对其功能亦可大致归类，象“显示”一类窗口之间有其特有的层次关系。

这些窗口可执行：

- 文件管理命令(copy, delete, file, formname, free, include, lock, print,prtfile, save, sprint, wpopup)
- 窗口管理(autopop, bye, cancel, clear, command, details, end, endsas, home, icon, keydef, zoom, next, pmenu, prevcmd, prewind, purge, reshow, scrollbar, status, update, x, zoom)
- 窗口大小与位置控制(cascade, resize, tile, wdef, wgrow, wmove, wsave, wshrink)
- 颜色(color)
- 滚动(backward, forward, hscroll, left, n, right, top, vscroll)
- 文本存贮与剪贴(mark, pclear, plist, smark, store, unmark, cut, paste)
- 搜寻(bfind, change, find, rchange, rfind)

其中少数命令不能在PC上使用，而在PC SAS 窗口命令还有clock，显示时，各窗口所用命令略有差异。

在一个窗口的命令行打入窗口名则进入相应的窗口。如在PGM 的命令行上打入KEYS，然后回车，则进入KEYS 窗口，显示当前的功能键定义或改变功能键定义，退出时可以用CANCEL 或CAN/QCAN 来放弃或者用END 存贮功能键。在命令行上还可以用分号间隔许多命令，连续执行。

在SAS 程序内执行窗口命令可用DM 命令。进入帮助窗口时有以下提示，可以一览整个SAS 系统的功能。如程序dm"log;clear;output;clear;pgm"; 可以存放文件CLEAR.SAS 中，新程序前面加上%INCLUDE CLEAR; 语句，则先清除激活的三个窗口已有的内容，调试程序时很有用。

在窗口的命令行上打入HELP <过程名>，可立即显示特定过程的语法。当按照窗口提示进入多层的帮助时，可以用=x 退出或用F10/END 返回至上级帮助屏幕。

PGM 窗口常用于程序编辑，有一套完整的行编辑命令，这些命令可以预先在KEYS 窗口存贮起来。在命令行上常用命令有：CAPS, FILL, RESET, NUMBERS，数字区使用的命令有：M,MM(移动)、C,CC (拷贝)、D,DD (删除)、R,RR (复制)、A (拷贝、移动到本行后)、B (拷贝、移动到本行前)、I (本行后插(行前插入可用IB)。其中两个字母则需要在不同行号用两次。也可以用D[行数]、C[行数]、M[行数] 指定被操作的行数。NUMS、TABS、COLS，分别切换程序行的显示、标尺及列标度，在格式输入时有用。另外还有一些文本移动、改换大小写等命令。把一行某处分开可用TS 命令，连接则用TF 命令。在KEYS 窗口进行热键定义时，应注意在这些行命令前缀以冒号(:)，然后以END 进行存贮，定义成功时立即为SAS 系统采用。行命令可以通过在PGM 命令行上打RESET 而放弃。外部程序的调入，是在PGM 窗口打入INCLUDE'文件指示' (或INC '文件指示'，文件指示包括驱动器、路径和文件名)，编辑的文件则以FILE'文件指示' 存贮。END 在PGM 窗口时则意味着提交执行(SUBMIT)。

简单的示例：现有一个销售情况的原始数据列表，你可以用INFILE 把它调入。注意SAS 的语句以一个关键字开头以分号结尾。

```
data sales;
   infile 'sales.dat' pad;
   input salesrep $ 1-7 sales 8-12 region $ 14-18
         machine $ 19-20;
```

```
run;
```

Stafer	9664	east SM
...		
Ryan	32915	west SM
Tomas	42109	west SM
Thalman	94320	southC

程序中以DATA 关键字开始的部分为数据步，用于创建数据文件，即SAS 的数据集。下述程序进行打印、绘立体条图和制频数表。

```
proc print data=sales;
proc chart data=sales;
   block region / type=mean sumvar=sales;
proc freq data=sales;
   tables machine*region;
run;
```

以PROC 开始的部分为过程步，接在PROC 后的关键字为过程名，用于进行大部分的分析。SAS 允许连续使用多个数据步和过程步。

§4.1.3 微机系统SAS 的配置

为了保证SAS 系统有效地运行，在安装时应对软件进行恰当的配置。此处以PC SAS 为例，介绍与使用有关的注意点。

在MS-DOS 下，CONFIG.SYS 应存于引导盘的根目录中，文件内设FILES 选择项，其值一般不低于50，即FILEs=50。此外，还可以在AUTOEXEC.BAT 文件中设置计算机的运行环境，该文件在计算机开机后自动执行。如果要使用SAS/GRAPH 和SAS/QC，一般要安装扩展内存(expanded memory specification EMS)驱动程序，其版本最好不低于Lotus-Intel-Microsoft 标准LIM 3.0。例如EMS 驱动程序为LIMSIM.SYS，则CONFIG.SYS 内容为：

```
FILES=60
DEVICE=LIMSIM.SYS 1024
BUFFERS=15
```

其中BUFFERS 选项用于指示DOS 的缓冲区。有关内存管理的知识详细内容请参阅第3章。

类似地，SAS 系统有自己的软件的自动执行与环境配置文件，即CONFIG.SAS 与AUTOEXEC.SAS，可与DOS 相应。SAS 有一些默认值，如CONFIG.SAS 内容可以是(/**/内为注释)：

```
-PATH           C:\SAS\SASEXE\CORE   /* 执行文件定义路径和搜索次序*/
-PATH           C:\SAS\SASEXE\BASE
-PATH           C:\SAS\SASEXE\STAT
-CONFIG         C:\SAS\CONFIG.SAS
-FSDEVICE       SASXDICA             /* 显示管理全屏幕设备驱动程序*/
-FILEBUFFERS    5 512                /* 文件缓冲为5个，大小为512K */
-VERBOSE        ON                   /* 显示配置信息*/
-SET            SASROOT C:\SAS       /* 定义SAS 的根目录，是必选项
                                        此处SAS 所在的路径为C:\SAS */
-DMS                                 /* 在显示管理系统下运行SAS */
```

若要使用SAS 在根目录下运行，则与路径指示有关的量都应做相应改动。各选项的含义在CONFIG.HLP 文件内有较详细的说明，掌握其意义后，重新进行有关设置，可以避免系统重装。

1. -CONFIG

   句法：-CONFIG 文件名

   如：-CONFIG myconfig.sas。这一选择定义一个不同于CONFIG.SAS 的配置文件。使用该选项时，应当使用文件的全名。这一选项仅当SAS 启动时使用，在一个配置文件内使用-CONFIG选择时则被忽略。

2. -DMS (-NODMS)

   句法：-DMS或-NODMS

   指示SAS的一次运行是否应该使用显示管理系统。-DMS 指示使用而-NODMS指示不使用。该项不写时SAS 将启用显示管理系统，即屏幕上的多窗口功能，这时SAS 系统要多用大约111K 内存。

3. -ECHO

   句法：-ECHO ”字串” | CLS

   如：-ECHO CLS。指示SAS 在启动时屏幕显示一个或多个信息，可用于对用户提示重要信息，使用多个-echo 语句可充满整个屏幕。-ECHO CLS 是清屏的特例。

4. -EMS

   句法： -EMS num_16k_pages — ALL

   如： -EMS 128。指示 SAS 系统使用LIM 扩展内存(EMS)，该项默认时不选。-EMS ALL 指示SAS使用直至两兆的所有EMS 存贮。否则，-EMS 数字指示SAS 使用大小为16k的EMS页数。如-EMS 16 指示SAS使用16个大小为16K的EMS页，也即256k。仅仅指示-EMS 或-EMS 0 时，EMS 存贮不被使用。使用EMS 可以增强SAS 的处理能力，然而使用cpu 较多而磁盘访问较少时将减低SAS 的能力。

5. -FILEBUFFERS

   句法：-FILEBUFFERS 缓冲区数目缓冲区大小

   如：-FILEBUFFERS 5 512。允许SAS 对少量的磁盘读写进行缓冲，以减少磁盘访问增强能力。在速度与存贮方面有一个折衷，即filebuffers 的值越大，则SAS越能有效地使用磁盘而过程可用的内存将减少。

6. -FILECACHE

   句法：-FILECACHE path num_files

   如：-FILECACHE !sasroot\sasexe\core 15。指示SAS 对特定路径或目录的文件进行高速缓冲，即一旦这些文件被调用，则再次调用时速度加快。注意：对包含SAS执行文件和信息文件的目录设定高速缓冲时，SAS的运行达到最佳。

7. -FSDEVICE

句法：-FSDEVICE driver_name options

如：-FSDEVICE sasxdiea lines43 typeslow mode=co80。指定显示驱动设备。

机型与显示设备	建议-fsdevice 的选择
IBM AT CGA	-fsdevice SASXDICA
IBM XT CGA	-fsdevice SASXDICX
IBM AT Monochrome	-fsdevice SASXDIMA
IBM XT Monochrome	-fsdevice SASXDIMX
IBM PC 3270 AT	-fsdevice SASXDNCA
IBM PC 3270 XT	-fsdevice SASXDNCX
IBM AT EGA	-fsdevice SASXDIEA
IBM XT EGA	-fsdevice SASXDIEX
IBM PS/2 VGA	-fsdevice SASXDIVA
Wang Color	-fsdevice SASXDWGC
Wang Monochrome	-fsdevice SASXDWGM
Compaq AT Color	-fsdevice SASXDICA NOWAIT
Leading Edge XT	-fsdevice SASXDICX
AT&T Color	-fsdevice SASXDICX
AT&T Monochrome	-fsdevice SASXDICX MODE=BW80 GRAY=BLACK BLACK=WHITE
Non-IBM compatible supporting ANSI.SYS	-fsdevice SASXDASY

Monochrome 为单色显示器、CGA即图形适配器、EGA为增强图形适配器、VGA为视频图形适配器。在其他驱动程序无效时，应当使用SASXDASY，这时应当在DOS的CONFIG.SYS文件中指示DEVICE=ANSI.SYS。在安装EMS 和中文时，可使用SAS在中文系统下工作。

8. -NEWS 句法：-NEWS 文件名

如：-NEWS mynotes.dat。指示在SAS 启动后，该文件内容将被显示于SAS的登录窗口。

9. -PATH

句法：-PATH 路径名

如：-PATH !sasroot\sasexe\core。-path 选项指示SAS 定位系统可执行文件(.EXE 和.EMS)的搜寻次序和路径。CORE 与BASE 的目录应于第一个和第二个路径。

10. -set SASROOT

句法：-set SASROOT 路径名

如：-set SASROOT \sas。-set SASROOT 指示SAS 的根目录。指示错误时，将出现文件找不到的信息。

11. -VERBOSE

句法：-VERBOSE

在SAS 启动前显示所有配置信息，可用于检查与配置有关的问题。如SAS 系统配置错误不能运行时，用于显示实际设定的设置，便于确实。

12. 注释：/* comments */

句法: /* comment */

如：/* This is a comment line */。注释可用于配置文件的任何地方。

AUTOEXEC.SAS 中可以放一段程序，如设自己的OPTIONS，设描述文件(SCRIPT FILE)等。

```
OPTIONS RLINK 'C:\SAS\SASLINK\DECNET.SCR';
OPTIONS PS=300 LS=132 nodate nonumber;
TITLE 'SAS ANALYSIS -- CHSI/MOPH DEC.10.1991';
```

此处指出SAS系统的备份方法。微机系统的备份可使用SASBACK.EXE文件来完成。该文件存放于SAS系统目录的按装目录SASINST内。其语法是：

SASBACK -BACKUP 源目录目标盘

SASBACK -RESTORE 源盘目标盘

使用时可用SASBACK -USAGE BACKUP/RESTORE 细读其语法。

§4.1.4 SAS/STAT

㈠模快功能分类

SAS 的统计分析是放在“数据分析”的框架之下的，下图说明了这一点。图中对各种统计分析进行了分类。

HELP: SAS System Help
Command ===>
SAS SYSTEM HELP: Data Analysis

Regression	Analysis of	Categorical	Elementary	Multivariate
CALIS	Variance	CATMOD	CAPABILITY	CALIS
GLM	ANOVA	CORRESP	CORR	CANCORR
LIFEREG	GLM	FREQ	FREQ	CORRESP
LOGISTIC	LATTICE	LOGISTIC	MEANS	FACTOR
NLIN	MIXED	PRINQUAL	SUMMARY	GLM
ORTHOREG	NESTED	PROBIT	TABULATE	MDS
PROBIT	NPAR1WAY		UNIVARIATE	MULTTEST
REG	PLAN	Utility		PRINCOMP
RSREG	TTEST	INBREED	Time Series	PRINQUAL
TRANSREG	VARCOMP	RANK	ARIMA	REG
		SCORE	AUTOREG	TRANSREG
	Survival	STANDARD	STATESPACE	
Clustering	Analysis			
ACECLUS	LIFEREG	Discriminant		Control
CLUSTER	LIFETEST	CANDISC	Systems	Charting
FASTCLUS	LOGISTIC	DISCRIM	MODEL	CUSUM
TREE	PHREG	STEPDISC	SIMLIN	MACONTROL
VARCLUS	PROBIT		SYSLIN	SHEWHART

图4.2 VAX/VMS SAS 6.07 系统帮助(数据分析功能)

即回归、方差分析、分类资料分析、基础统计、多元分析、聚类分析、生存分析、判别分析、方程组、控制图，以及实用程序，各分类间有重叠。

现以PC SAS为例，其各部分功能特点略做介绍如下：

1. 回归分析. 可采用过程CATMOD，GLM，LIFEREG，NLIN，ORTHOREG，REG，RSREG，LOGISTIC，PROBIT。通常的回归分析用REG，其它的过程是针对特定类型问题的。

   (a) CATMOD 尤适于可排成列联表形式的数据如对数线性模型、LOGISTIC 回归. 重复测量分析等，LOGISTIC 过程尚可用于多分类LOGISTIC 分析和变量的筛选。

   (b) GLM 用于配合一般线性模型，如方差分析。

   (c) LIFEREG 可以对左、右、区间截尾的失效时间数据配合参数模型。

   (d) NLIN 进行非线性回归分析，采用梯度法、牛顿法、修正牛顿法、麦夸特(Marquardt)法及弦截法求解，可以得到加权最小二乘解。

   (e) ORTHOREG 对于病态的资料回归效果较佳。

   (f) REG 提供了丰富的回归诊断功能，可用多种方法选择模型。

   (g) RSREG 建造二次响应曲面模型。

   过程STEPWISE、RSQUARE 分别用于逐步回归和最优子集回归，其内容已纳入PROC REG 中(由语句MODEL 语句选项SELECTION= 指定)。

2. 方差分析. 可用ANOVA、CATMOD、GLM、NESTED、NPAR1WAY、PLAN、TTEST和VARCOMP 进行。

(a) ANOVA 主要针对平衡设计。进行方差分析、多因素方差分析、重复测量的方差分析。能用多种方法进行两两比较。

(b) NESTED 对完全钳套的随机模型进行方差和协方差分析。

(c) TTEST 进行成组t-检验分析。

(d) VARCOMP 对于随机或混合模型估计方差分量。

3. 分类资料分析. 除了CATMOD 外，SAS 可用基础模块中的FREQ。FREQ 可产出频数表，进行检验和关联性度量如两维表卡方、比数比（odds ratio）、相关统计量、Fisher 精确概率法检验。另外可以进行分层分析、计算Cochran- Mantel-Haenszel 统计量和相对危险度。

   过程FUNCAT 的功能也由CATMOD 过程实现。

4. 多元分析. 这里多元的涵义是指探讨各变量的关系而不指明哪些是因变量，｜些是自变量。有PRINCOMP、FACTOR 和CANCORR，分别进行主成分分析、因子分析和典型相关分析。

5. 判别分析. 有DISCRIM、CANDISC 和STEPDISC。可用线性或二次函数作为判别函数，进行典型分析和逐步判别。分布的假设可以不是多元正态。

6. 聚类分析. 可用CLUSTER、FASTCLUS、VARCLUS 和TREE。FASTCLUS 用于分解法聚类，尤适于大批量数据的处理，最多可容纳10万个观察；TREE 过程用于画谱系图(dendrogram 或phenogram)。ACECLUS、PRINCOMP、STANDARD 则可为聚类分析进行数据预处理。

7. 计分计算。有STANDARD、RANK 和SCORE。STANDARD 对于给定的均值和标准差标化变量；RANK 产生变量的秩次；SCORE 根据FACTOR 等过程产生的因子负荷和有关计分值，形成线性组合。

8. 生存分析. 用过程LIFEREG 和LIFETEST。LIFEREG 主要用于拟合参数模型，LIFETEST 则进行一些检验。

SAS/STAT 对以上各部分中各过程功能特点进行了详细的比较。

微机SAS/STAT 6.04 较6.03 增加了CALIS 和LOGISTIC，用于结构方程模型分析和LOGISTIC 回归分析。

㈡SAS/STAT 说明书

在SAS/STAT 用户指南中，过程描述包括以下的内容：

ABSTRACT 简单说明该过程的用途是什么。

INTRODUCTION 介绍和背景材料，包括一些定义和用例。

SPECIFICATIONS 语句写法。

DETAILS 特点说明、内部操作、输出、缺失值处理、计算方法、资源占用情况和使用注解。

EXAMPLES 分析实例，包括数据、SAS 语句和打印输出。

REFERENCES 部分参考文献. 它能帮助您掌握更多的背景知识以便使用。

NOTES 软件包在各操作系统下的异同。

表 4.1 SAS 的运算符号的优先级

优先级	计算方法	符号	等价表示	定义
0 组	由内向外	()		括号
1 组	自左至右	**		指数
		+,-		正数/负数
		^	NOT	逻辑非
		><	MIN	最小
		<>	MAX	最大
II 组	自左至右	*,/		乘/除
III 组	自左至右	+,-		加/减
IV 组	自左至右	||		字串并置
		可随系	统而变	
V 组	自左至右	<	LT	小于
		<=	LE	小于等于
		=<	LE	小于等于
		=	EQ	等于
		^=	NE	不等于
		>=	GE	大于等于
		=>	GE	大于等于
		>	GT	大于
			IN	集合操作
VI 组	自左至右	&	AND	逻辑与
VII 组	自左向右	|	OR	逻辑或

§4.2 SAS 语言

§4.2.1 有关概念

SAS 的表达式是操作符和操作数的序列，序列的结果是SAS 的常数。

SAS 的常数是一个数字、用引号引起来的字符串，或其他指示一个固定值的特殊记号。

数值常数可以含有小数点、正负号及科学记数法中的E格式。缺失值一般用小圆点(.)来表示。用16 进制格式表示时，通常以0 引导，后缀以字母X。

字符常数长度为1-200，若字串含有单引号(')，则应用双引号(")括起来。缺失值用引号中的空格来表示。字符常数可以采用16 进制记号，此时奇数个字符用引号括起来，后缀以字母X。

字符型常数用于赋值、计算和比较时可自动转为数值常数。

日期时间及日期时间常数是用引号把日期或时间括起来，后缀以D(日期)、T(时间)或DT(日期时间)。

根据操作符号的多少，表达式有简单表达式和复合表达式之分。亦分算术、比较、逻辑和其他操作符号。其形式、等价写法与优先级列表如下。

§4.2.2 SAS 语句

SAS 的语句可分为DATA 步语句、PROC 步语句和全程语句。

在DATA 步中，有大量的语句对变量进行操作，赋值语句可用于产生新的变量和改变变量属性，而PROC 步中的语句除PROC NLIN 等少数过程外，比较简单和固定，DATA 步有关语句的句法如下。

ABORT <ABEND|RETURN>< n >;

ARRAY,explicit: ARRAY 数组名{下标} < $ ><长度><<数组元素><(初值)>>; 用于显式地定义数组。

ARRAY,implicit ARRAY 数组名<(指示变量)>< $ ><长度> 数组元素。用于隐式地定义数组。

数组元素是SAS的变量，引用时可用隐式格式或显式格式，前者为数组名(下标) ，后者为单纯的数组名。

ATTRIB 变量表1 属性表1 <...变量表n 属性表n>; 用于改变SAS变量的属性如标签、格式等。

BY <DESCENDING><GROUPFORMAT> 变量1 <... <DESCENDING < GROUPFORMAT> 变量n><NOTSORTED>; 用于指示排序变量列表。

CALL 程序(<参数<, ... >>); 用于SAS程序中的子程序调用。

CARDS; 用于引导读取一个数据列表，数据列表以分号(;)结束。

CARDS4; 用于代替CARDS;引导读取有分号(;)的数据列表，此时列表应以四个分号结束。

DATA <数据集<(选项)>><...数据集<<选项)>></VIEW= 视图名|PGM=程序名>; PGM=程序名; 用于引导数据步的开始。

DELETE; 用于删除记录。

DISPLAY 窗口<.组名><NOINPUT><BLANK><BELL>; 用于显示定义的窗口。

DO; 与END结合使用形成复合语句。

DO,iterative: DO 指示变量=指示1 <,...指示n>; 用于迭代式循环。

DO OVER (数组名); 用于对数组进行循环操作。

DO UNTIL (表达式); 循环控制语句。

DO WHILE (表达式); 循环控制语句。

DROP 变量列表; 用于删除数据集中的变量。

END;

ERROR <指示信息>; 给出错误信息。

FILE 文件指示<选项><系统选项>; 用于存贮非SAS格式的文件。

FORMAT 变量<格式><DEFAULT=默认内部格式>...; 用于对变量进行格式化。

GOTO 标号; 程序转向语句。

IF 表达式; 用于进行观察的筛选。

IF 表达式THEN 语句;

IF 表达式THEN 语句; ... ELSE 语句。

INFILE 文件指示<选项><系统选项>; 用于调入外部文件。

INFORMAT 变量<内部格式><DEFAULT=默认内部格式>; 用于指示内部格式。

INPUT <指示1><...指示n><@|@@>;

INPUT,column: INPUT 变量< $ > 开始列<-终止列><.小数位数><@|@@>;

INPUT,formatted: INPUT <指针控制> (变量列表) (内部格式表) <@|@@>;

INPUT <指针控制> (变量列表) (<n*> 内部格式) <@|@@>;

INPUT,list: INPUT <指针控制> 变量<: |&|~ ><内部格式><@|@@>;

INPUT,named:INPUT <指针控制> 变量= < $ ><内部格式><@|@@>;

KEEP 变量列表; 用于保持变量。

LABEL 变量1='标签1' <...变量n='标签n'>; 用于给变量加标签。

Lables,statement: 标号:语句; 用于给语句增加标号。

LENGTH <变量指示1 <...变量指示>><DEFAULT=n>; 指示变量的长度。

LINK 标签; 用于调用子程序。

LIST;

LOSTCARD;

MERGE 数据集1 <(数据集选项)> 数据集2 <(数据集选项)><... 数据集n <(数据集选项)>><END=变量名>;用于合并数据集。

Null; 即空语句，为一个分号。语句之间用分号(;) 隔开，注意书写程序时不要遗漏。

OUTPUT <数据集1 <..数据集n>>; 指示被输出的数据集。

PUT <指针控制><指示><...指示><@>;

PUT,column: PUT <指针控制><变量>< $ > 开始列<-结束列><. 小数点位置>;

PUT,formatted: PUT <指针控制> 变量格式<@;>

PUT <指针控制> (变量列表) (格式列表) <@>;

PUT <指针控制> (变量列表) (<n*> 格式) <@>;

PUT,list: PUT <指针控制> 变量< $ ><@>;

PUT <指针控制> <n*> '字符串' <@>;

PUT <指针控制> 变量<:> 格式<@>;

PUT,named: PUT <指针控制> 变量= <@>;

PUT <指针控制> 变量= <格式><@>;

PUT 变量= <> 开始列<-结束列><.小数位数><@>;

RENAME 旧名1=新名1 <...旧名n=新名n>;进行变量更名。

RETAIN <变量名表1 <初值1—(初值1)—(初值表1)><..变量名表n < 初值n|(初值n)|(初值表n)>>>;用于保持变量的值。

RETURN;

SELECT <(选择表达式)>; WHEN (表达式) 语句; <OTHERWISE 语句;> END; 用于选择方式进行某操作。

SET <数据集1<(选项)>>< ... <(数据集n <选项)>><< POINT= 变量名><KEY=索引名>><NOBS=变量名><END=变量名>; 用于读入数据。

STOP;

Sum 变量+表达式; 把表达式的值累积到变量中。

UPDATE 主数据<(选项IN=变量1)> 转换数据集<(选项IN=变量)> <END= 变量>;用于数据更新。

WHERE 逻辑表达式; 执行条件选择。

WINDOW 窗口名<选项><字段><GROUP=组名<字段>> ...; 用于定义窗口。

Stored Program Facility 允许编译并存贮数据步的程序然后于其它的时间执行，其步骤如下：

DATA 数据集;
源程序语句;
RUN PGM=存贮程序名;
DATA PGM=存贮程序名;
REDIRECT INPUT|OUTPUT 旧名1=新名11 <...旧名n=新名n>;
RUN;
以下命令是全程命令，能用于SAS 程序的任何地方。
注释/*信息*/|*信息;
DM < window > '命令-1<;...命令-n>' <window>; < CONTINUE >;
ENDSAS; 结束SAS运行
FILENAME fileref <设备类型> '外部文件' <主机选项>;
fileref CLEAR;
fileref 设备类型<主机选项> ;
fileref|_ALL_ LIST;
FOOTNOTE < n ><'文本'|"文本">;
%INCLUDE 程序-1 <...程序-n></<SOURCE2><S2=长度>>;
%LET 宏变量=变量列表;
LIBNAME libref < engine > < 'SAS-data-library' >
< SAS-选项> < engine/ 主机选项> ;
libref|_ALL_ CLEAR;
libref|_ALL_ LIST;
%LIST <n <:m— -m>>;
LOCK libref<.成员名<.成员.类型| .进入名.进入类型>> < LIST | CLEAR >;
MISSING 字符-1 <...字符-n>; 定义缺失值，在PROC TABULATE很有用。
OPTIONS 选项-1 <...选项-n>; 指定SAS系统选项。
PAGE;
%PUT <信息>; 用于打印一些信息。
RUN <CANCEL>;
%RUN;
SKIP < n >;
TITLE < n ><'文本'|"文本">;
X <'命令'>; 进入系统外壳，与x'command' 相当；在UNIX X- windows下使用x'csh' 或x'tcsh'。

§4.2.3 SAS 函数

SAS 的函数是一个例行程序，根据一个或几个给定参数返回相应的值。调用的格式是函数名(表达式<,表达式>) 或函数名(OF 变量列表)，列表的方式如x1-x 10、a b c d、A–Z等。下面分类列出。其中的argument 表示参数，其他的依英文名类推。

1. 算术函数
   ABS(argument) 绝对值函数。

DIMn(arrayname) 返回一维或多维数组中指定维数的元素。
DIM(arrayname,arraybound) 同上。
HBOUNDn(arrayname) 返回数组上界。
HBOUND(arrayname,boundn) 同上。
LBOUNDn(arayname) 返回数组下界。
LBOUND(arrayname,boundn) 同上。
MAX(argument,...) 返回最大值。
MIN(argument,...) 返回最小值。
MOD(arguemtn,argument2) 取模。
SIGN(x) 返回符号或零。
SQRT(argument) 平方根。

2. 四舍五入函数
CEIL(argument) 大于等于参量数的最小整数。
FLOOR(argument) 小于等于参量的最大整数。
FUZZ(argument) 若参量小于1E-12则返回整数。
INT(argument) 返回整数值。
ROUND(argument,roundoffunit) 据四舍五入单位返回一个值。
TRUNC(number,length) 对于指定的长度返回一个截断数值。

3. 数学函数
DIGAMMA(x) 伽马函数导数。
ERF(x) 误差函数。
ERFC(argument) 误差函数的补。
EXP(argument) 自然指数。
GAMMA(x) 伽马函数。
LGAMMA(argument) 伽马函数的对数。
LOG(arguemnt) 自然对数。
LOG2(argument) 以2为底的对数。
LOG10(argument) 常用对数。
TRIGAMMA(argument) 对数伽马函数的二阶导数。

4. 三角函数
ARCOS(argument) 反余弦函数。
ARSIN(argument) 反正弦函数。
ATAN(argument) 反正切函数。
COS(argument) 余弦函数。
COSH(argument) 超余弦函数。
SIN(argument) 正弦函数。
SINH(argument) 超正弦函数。
TAN(argument) 正切函数。

TANH(argument) 超正切函数。

5. 概率函数

POISSON(lambda,n) 泊松分布函数。
PROBBETA(x,a,b) 贝塔分布函数。
PROBBNML(p,n,m) 二项分布函数。
PROBCHI(x,df<,nc>) 卡方分布函数。
PROBF(x,ndf,ddf<,nc>) F 分布函数。
PROBGAM(x,a) 伽马分布函数。
PROBHYPR(nn,k,x<,or>) 超几何分布函数。
PROBNEGB(p,n,m) 负二项分布函数。PROBNORM(x) 标准正态分布函数。
PROBT(x,df<,nc>) t-分布函数。

6. 分位点函数

BETAINV(p,a,b) 贝塔分布逆函数。
CINV(p,df<,nc>) 卡方分布分位点。
FINV(p,ndf,ddf<,nc>) F 分布分位点。
GAMINV(p,a) 逆伽马分布分位点。
PROBIT(argument) 逆正态分布函数。
TINV(p,df<,nc>) t-分布分位点。

7. 简单统计函数

CSS(argument,...) 校正平方和。
CV(argument,...) 变异系数。
KURTOSIS(argument,...) 峰度系数。
MAX(argument,...) 最大值。
MIN(argument,...) 最小值。
MEAN(argument,...) 均值。
N(argument,...) 非缺失值数目。
NMISS(argument,...) 缺失值的数目。
ORDINAL(count,argument,argument,...) 给出第一个计数参量的最大者。
RANGE(argument,...) 极差。
SKEWNESS(argument,...) 偏度。
STD(argument,...) 标准差。
STDERR(argument,...) 标准误。
SUM(argument,...) 和。
USS(argument,...) 未校正和。
VAR(argument,...) 方差。

8. 随机数函数

NORMAL(seed) 返回一个正态变量。

RANBIM(seed,n,p) 返回二项分布的一个量。
RANCAU(seed) 返回一个柯西分布变量。
RANEXP(seed) 返回一个指数分布变量。
RANGAM(seed,alpha) 返回伽马分布的一个量。
RANNOR(seed) 返回一个正态变量。
RANPOI(seed,lambda) 返回一个泊松分布的量。
RANTBL(seed,p1,..,pi,..pn) 返回表格形式密度函数的变量。
RANTRI(seed,h) 返回一个三角分布的观察。
RANUNI(seed) 返回一个均匀分布变量。
UNIFORM(seed) 返回一个均匀分布变量。

9. 商用函数
COMPOUND(amount,future,rate,number) 复利。
DACCDB(period,value,years,rate) 累积递减平衡折旧值。
DACCDBSL(period,value,years,rate) 转化为直线折旧。
DACCSL(period,value,years) 累积直线折旧值。
DACCSYD(period,value,years) 累积sum-of-years'-digits 折旧。
DACCTAB(period,value,tab1,...,tabn) 从指定表中的累积折旧值。
DEPDB(period,value,years,rate) 递减平衡折旧值。
DEPDBSL(period,value,years,rate) 转为直线的抵减平衡。
DEPSL(period,value,years) 直线折旧。
DEPSYD(period,value,years) sum-of-years 折旧值。
DEPTAB(period,value,tab1,...,tabn) 从指定的表中返回折旧值。
INTRR(period cash0,cash1,...) 返回内部率。
IRR(period,cash10,cash2,...) 返回用百分比表示的内部率。
MORT(argument,patmet,rate.number) 返回抵押损失。
NETPV(raet,period,cash0,cah1,..) 返回率为分数时的净现值。
NPV(rate,payment,rate,number) 返回率为百分比时的净现值。
SAVING(future,payment,rate,number) 定期存款的未来值。
这一类函数中，有许多是关于折旧计算的，列表如下：

参数说明：value 是折旧前资产的值，years 是recovering period，period 是recovering period 中的年份。rate 是折旧率。其它的如：
对COMPOUND(a,f,r,n), f=a*(1+r)**n;
对MORT(a,p,r,n), p=r*a*(1+r)**n/((1+r)**n-1);
对SAVING(f,p,r,n), f=p*(1+r)*((1+r)**n-1)/r;

10. 字符函数
Byte(n) 返回ASCII 码或EBCDIC 序列的值。
COLLATE(n,m,l) 返回按collating 序列的字符。
COMPRESS(argument) 返回空格被压缩的字符。

表 4.2 几种商用函数的换算关系

折旧方法	周期折旧函数	累积折旧函数
年份数位和 (sum of years digits)	DEPPSTD	DACCSYD
直线 (stright line)	DEPSL	DACCSL
递减平衡 (decline balance)	DEPBB	DACCDB
递减平衡换为直线 (decline balance to straight line)	DEPDBSL	DACCDBSL
表 (table)	DEPTAB	DACCTAB

INDEX(argument...) 字符模式。
INDEXC(argument...) 指示字符出现的第一个数。
LEFT(argument) 字符左齐。
LENGTH(argument) 字符长度。
RANK(x) 返回ASCII 或EBCDIC 序列中的字符位置。
REPEAT(argument,n) 重复字符。
REVERSE(argument) 反转字符。
RIGHT(argument) 字符右齐。
SCAN(argument,n,delimiters) 寻找字。
SUBSRE(argument,to,from,...) 抽取字符。
TRIM(argument) 舍弃尾部空格。
UPCASE(argument) 转为大写。
VERIFY(argument1,argument2,...) 确认字符的取值。

11. 日期与时间函数(date and time):
DATE() 返回当天的SAS 日期。
DATEJUL(juliandate) 把西历转为SAS 日期值。
DATEPART(datetimre) 从SAS 日期时间值或literal 返回日期部分。
DATETIME() 返回当天的日期和时间。
DAY(date) 返回SAS日期值中的日数。
DHMS(date,hour,minute,second) 对于给定的日、时、分、秒返回一个SAS 日期时间值。
HMS(hour,minute,second) 对给定的时、分秒返回一个SAS 时间值。
HOUR(time) 返回SAS 日期时间或时间或literal 的小时数。
INTCK(interval,from,to) 返回时间间隔数。
INTNX(interval,from,number) 对于给定的间隔向前推算一个时间。
JULDATE(date) 从SAS 日期或literal 中返回西历值。

表 4.3 州与ZIP 码函数的关系

	返	回	值	
参数(argument)	FIPS 码	大写州名	大小写州名	邮政编码
FIPS		FIPNAME	FIPNAMEL	FIPSNAME
邮政编码	STFIPS	STNAME	STNAMEL	
ZIP 码	ZIPFIPS	ZIPNAME	ZIPNAMEL	ZIPSTATE

MDY(month,day,year) 从月、日和年中返回一个SAS日期值。
MINUTE(time) 或MINUTE(datetime) 从SAS 日期、时间日期或literal中返回分钟数。
MONTH(date) 从SAS 日期值或literal 中返回月份值。
QTR(date) 从SAS 日期值或literal 中返回季度值。
SECOND(time) 从SAS 时间或日期时间值或literal 中返回秒数。
TIME() 返回当天的时间。
TIMEPART(datetime) 从SAS 日期时间值或literal 中抽出时间部分。
TODAY() 返回当天的SAS 日期值。
WEEKDAY(date) 从SAS 日期值或literal 中返回星期数。
YEAR(date) 从SAS 日期值中返回年份值。
YQQ(year,quarter) 从年份和委度值中返回SAS日期值。

12. 州与ZIP (Zone Improvement Plan)码函数
这一类函数使用的参数有FIPS 州码、两个字母的邮政编码(postal code)、ZIP 码。FIPS 用于人口普查资料、ZIP 的长度为5，邮政编码是地址中常用的两字母的缩写。这些函数涉及的地区包括了美国50 个州、波多黎哥、哥伦比亚地区、和Guam。这类函数使用较少。
FIPNAME(fips) 把FIPS 转成州名(所有均大写)。
FIPNAMEL(fips) 把FIPS 码转为州名(大写或小写)。
FIPSTATE(fips) 把FIPS 码转为两字符邮码。
STFIPS(fips) 把邮码转为FIPS 州码。
STNAME(postalcode) 把邮码转为州名(所有均大写)。
STNAMEL(postalcode) 把邮码转为州名(大写或汪写)。
ZIPFIPS(zipcode) 把ZIP 码转为FPIS 州码。
ZIPNAME(zipcode) 把ZIP 码转为州名(所有均大写)。
ZIPNAMEL(zipcode) 把ZIP 码转为州名(大写或小写)。
ZIPSTATE(zipcode) 把ZIP 码转为两字母州名。

13. 特殊函数
DIFn(argument) 返回延迟为n的一阶差分。
INPUT(argument,informat) 用指定的内部格式返回一个值。

LAGn(argument) 返回第n个延迟的值。

PUT(argument,format) 用指定的格式返回一个值。

SYMGET(argument) 返回宏变量的值。

SAS CALL Routines 产生特定分布的随机变量，同时进行种子更新。在使用CALL语句调用这些过程之前，应首先对种子进行初始化。这些程序调用很方便，主要是对调用参数进行恰当的匹配。

CALL RANBIN(seed,n,p,x) 并产生均值为np，方差为np(1-p)的二项分布变量x。

CALL RANCOU(seed,x) 产生一个柯西分分布的变量x，其位置参数为1而尺度参数为1。

CALL RANEXP(seed,x) 产生一个指数分布的变量x，其参数为1。

CALL RANGAM(seed,a) 产生一个参数为a 的伽马分布变量x。

CALL RANNOR(seed,x) 产生一个均值为0 方差为1 的正态变量x。

CALL RANPOI(seed,m,x) 产生一个均值为m的泊松分布变量x。

CALL RANTBL(seed,p1,...,pi,...,pn,x) 产生一个以p1,...,pn 为概率密度的变量x。

CALL RANTRI(seed,h,x) 产生参数为h 的三角分布的变量x。

CALL RANUNI(seed,x) 产生以(0,1) 区间上均匀分布的变量x。产生的方法是素数模乘法。据Fishman and Moore 1982，模数是2*31-1，因子为397204094。

CALL SOUND(freq<,dur>) 产生声音。

㈣数据步选项

SAS 能够在数据步或过程步进行一定的数据控制，常用的如：

DROP=变量列表控制不包括这些变量。

FIRSTOBS=n 指示处理从第n个记录开始。

IN=变量用于SET、MERGE 与UPDATE 语句中，指明数据集是否对观察有所贡献。

KEEP=变量列表指示保留处理的变量。

OBS=n 用于读数据时，指示读入的记录数。

RENAME=(旧名1=新名1<...旧名n=新名n>) 指示变量改名。

REPLACE=用于指示数据集是否被替换。

TYPE=CORR—DATA—COV—EST—SSCP 常用于统计过程，指示数据的类型。

WHERE (表达式) 用于进行数据的条件选择。

用例：以下程序控制仅仅打印数据集的头二十个记录。

PROC PRINT DATA=original(obs=20);RUN;

在数据库转换时把名为A、B、C 的变量分别换成X、Y、Z。

PROC DBF DB3=MYFILE OUT=MYFILE (rename=(a=x b=y c=z)); RUN;

变量引用如: X1-X100、_ALL_、_CHAR_ 或_CHARACTER_、_NUMERIC_、A–B、A _CHARACTER_ B、A _NUMERIC_ B 等等。

㈤SAS 宏定义

SAS 提供了丰富的宏调用函数，灵活应用，可以大大提高编程能力。宏定义的格式为：

%MACRO 宏定义名(参数表);

宏语句;

宏语句可以是DATA步语句如%DO,...,%END，也可是SAS 过程。

%MEND;

以后即可用%宏名(参数表); 的方式调用了。可以用OPTIONS MPRINT; 打出宏实际执行的的语句。

SAS可在程序中运行显示管理系统语句，语句为DM。

【例4.2】下面是一个假想的数据，使用宏结合TABULATE过程制表。

```
options nocenter ps=66 ls=115 missing=' ' mprint;
data test;
input x1 x2 x3 count city $19.;
cards;
1 1 2 10 beijing
2 2 1 5 tianjin
2 1 2 4 shanghai
1 2 1 7 guangzhou
1 3 2 8 harbin
2 2 1 2 wuhan
2 1 2 8 chengdu
1 3 1 23 xian
proc print; id city; var x1-x3; run;
proc format;
value $city 'beijing'='北京' 'harbin'='哈尔滨'
'tianjin'='天津' 'wuhan'='武汉'
'shanghai'='上海' 'chengdu'='成都'
'guangzhou'='广州' 'xian'='西安';
run;
proc datasets;
modify test;
label city='城市名';
format city $20.;
run;
%macro tab(a,b,c);
proc tabulate f=6. noseps fc=' ————-';
freq count;
class &a &b &c;
table &a all,all &b*(n pctn<&a all>='列%'*f=5.2
pctn<&b all>='行%'*f=5.2
pctn<all*&b &a*&b>='keylabel n=' ' all='合计';
format city $city.;
run;
%mend;
/*宏调用，在记录文件中打印真实程序*/
options mprint;
%tab(x3,x1,x2);
```

```
%tab(city,x1,x2);
/*类似PROC FREQ 的表格*/
proc tabulate f=6. formchar=' ————-';
class x1 x2 x3;
freq count;
keylabel n='计数' all='合计' pctn='%';
table x1*(x2 all ) all,x3*(n pctn*f=6.2) all pctn*f=6.2
/rts=18;
run;
```

PRINT过程的ID在大数据集中标识很有用。使用MPRINT可以了解SAS系统实际运行的程序。本例使用tab宏时，变量的顺序不同，则产出不同的交叉表；第二部分程序产出了类似FREQ那样的交叉表，程序使用了选项FORMCHAR ，它也可以经OPTIONS 语句或窗口进行仁全程定义。TABULATE 过程的优点是制表可用一些修饰，但产出检验统计量。程序输出结果如下。

		X1	
		1	2
	合计	列% 行% %	列% 行% %
X3			
1	37	30 62.50 81.08 44.78	7 36.84 18.92 10.45
2	30	18 37.50 60.00 26.87	12 63.16 40.00 17.91
合计	67	48 100.0 71.64 71.64	19 100.0 28.36 28.36

		X1	
		1	2
	合计	列% 行% %	列% 行% %
城市名			
北京	10	10 20.83 100.0 14.93	
成都	8		8 42.11 100.0 11.94
广州	7	7 14.58 100.0 10.45	
哈尔滨	8	8 16.67 100.0 11.94	
上海	4		4 21.05 100.0 5.97
天津	5		5 26.32 100.0 7.46
武汉	2		2 10.53 100.0 2.99
西安	23	23 47.92 100.0 34.33	
合计	67	48 100.0 71.64 71.64	19 100.0 28.36 28.36

		X3		
		1	2	合计
		计数%	计数%	计数%
X1	X2			
1	1		10 14.93	10 14. 93
	2	7 10.45		7 10.45
	3	23 34.33	8 11.94	31 46.27
	合计	30 44.78	18 26.87	48 71.64
2	X2			
	1		12 17.91	12 17.91
	2	7 10.45		7 10.45
	合计	7 10.45	12 17.91	19 28.36
合计		37 55.22	30 44.78	67 100.00

掌握了SAS 的语言后，最主要的还是掌握其众多过程的使用，这一方面可经其检测程序来得到，另一方面则是其实例分析(SAMPLES)。这些用例也有其自身的归类方法。如后缀以EX者为使用手册上提供过的。有时则是直接给出样本所在的章节。

【例4.3】下面程序用循环和函数产生二项分布和泊松分布的概率和累积概率。

```
/* B */
data binom;
 do y=0 to 8;
    cum=probBNML(0.35,8,y);
    if y=0 then p=cum;
    else do; prev_cum=probBNML(0.35,8,y-1);p=cum-prev_cum;end;
    output;
 end;
 keep y p cum;
 proc print noobs; var y p cum;
 run;
 /* P */
 data poisson;
 y=0;p=1;
 do until (p<0.0001);
    cum=poisson(7.4,y);
    if y=0 then p=cum;
    else do; prev_cum=poisson(7.4,y-1);p=cum-prev_cum;end;
    if p$>$0.0001 then output;
    y=y+1;
 end;
 keep y p cum;
 proc print noobs; var y p cum;
 run;
```

这样可以造出一般统计书上难以见到的统计表，这些程序可以做成SAS 带有参数的宏过程供调用。

㈥过程简介

SAS 过程指南将基础过程分为三类，报告输出、计分和工具过程。

第一类，包括PRINT, FORMS, CHART, PLOT, CALENDAR 和TIMEPLOT。

第二类，包括STANDARD 和RANK。

第三类，包括APPEND, COMPARE, CONTENTS, COPY, DATASETS, DBF, DIF 、DOWNLOAD, FORMAT、SORT, TRANSPOSE, UPLOAD。

它们的使用比较简单，在以后的介绍中基本上涉及到了，这里不多介绍。

SAS 一系列功能主要由各模块的提供的过程来完成，各过程的选项、语句的细节可参考其说明书，SAS 的过程调用有一个基本的格式，如近交分析的格式如下。尽管对分析还没有熟悉，但一览便知其要点所在，其中大写字母为关键字，/* */ 内为注释。

```
PROC INBREED options; /* 选项*/
VAR variables; /* 分析变量*/
CLASSES variables; /* 分类变量*/
ID variables; /* 标识变量*/
MATINGS individual-list1 mate-list1 * ...; /* 指示模型*/
BY variables; /* 分析用的分组变量*/
RUN
```

对大多数用户来说，掌握上述用法一般没有困难，主要问题是结果的判读费功夫，这需要对统计过程所涉及的理论知识有足够的了解，同时也要掌握SAS 处理统计问题的习惯。在SAS手册中，重要的输出结果用圆圈罩住的数字来标注，其中的数字与DETAILS 节中的Printed Output 中的序号相应。

§4.2.4 微机SAS 系统示范程序

PC SAS/STAT 6.02-6.04 及SAS/QC 的样本程序，其分类是按照SAS提供的描述文件.BLS。列表时忽略的文件扩展名.SAS。统计检验如Bartlett和FRIEDMAN 检验无专门的过程，以样本程序方式给出。在Windows版SAS 6.11 中样本程序在帮助菜单下可以据模块调用所需样本程序，利用剪贴功能调入PROGRAM EDITOR从而提交运行。因此，用户可以根据自己需要进行一些小的修正。

§4.3 基础统计分析

§4.3.1 统计描述

描述统计包括原始数据的列表、图示以及综合性统计量的计算，可以理解为对统计数据的一种综合的表达方式。SAS 还提供一专门的过程用于数据的转换。

原始数据的列表，有过程PRINT 用于数据打印、FORMS 用于产生标签、FREQ 和TABULATE 用于产生交叉表，FREQ 产生有关的列联表统计量。

统计指标计算采用过程UNIVARIATE, SUMMARY, MEANS, CORR。在SAS/QC 中拥用CAPABILITY过程，其统计指标的产出与UNIVARIATE 相仿。利用过程TABULATE，可以产生连续变量的均值、方差等统计量。

SAS 统计数据的图示有两种方式，第一种是字符类型的绘图过程，产生的图不需要图形输出设备就输出，这一功能由过程PLOT 和CHART 来实现，PLOT 主要用于散点的绘制，SAS 给用户提供了很大的灵活性，如据页长改变图的纵轴长短、在同一坐标系下重叠绘制不同变量对的散点图，限定图轴的起止点、标度、绘图符号等。CHART 则可以绘制直方图、圆图、直条图和星形图等。第二种图示方法需要SAS/GRAPH 装入，相应的过程为GPLOT 和GCHART，它需要标准的图形输出设备如计算机图形显示器或图形打印机，其用法与PLOT 和CHART 相仿。

SAS 对三种描述方法可以结合在一个过程中，如茎叶图可以与统计指标同时给出。

箱尾图在SAS/IML 的说明书和样本程序库中有示范的写法。

现对数据集MYFILE 中的变量X1 到X10 计算综合统计量，变量COUNT 代表每一种观察组合下出现的次数，可以使用以下程序：

```
PROC MEANS DATA=MYFILE N MEAN STD MIN MAX RANGE SUM VAR MAXDEC=4;
FREQ COUNT;
VAR X1-X10;
RUN;
```

重要的是搞清楚各个量的含义。SAS 提供了大量专用统计函数如第二节所述，使用时应加以注意，如SUM其求和是仅对非缺失值进行的，与SPSS/PC+ 设为缺失值有所不同，N 给出非缺失值的变量个数。

交叉表格的制做使用PROC TABULATE 和PROC FREQ，后者常用于计算一些列联表检验统计量。TABULATE 能使用格式对数据进行格式化、使用频数变量，关键字、变量属性，以及对关键字进行重新命名，如：table ms=' ',x1=' '*x=' '; 及attrib age label='年龄' 等。

列表描述控制：OPTIONS、TITILE、BY 和FOOTNOTE，这些设定将影响到产出的页长、行宽、标题、脚注等，PROC TABULATE 受其影响最为明显。特别有用的是FORMAT语句，在数据描述时适当使用可以使结果更为直观，有时还相当于数据的转换功能。在此设定下，使用专门的过程如PROC PRINT;BY VARS; PAGEBY VAR; SUMBY; PROC FORMS用于产生格式标签。利用PROC PRINTTO 过程可以直接把结果输入到ASCII 文件或打印机(如LIST='LPT1')。在指示DATA _NULL_ 时，结合PUT 语句将把结果在LOG 窗口内输出。若拥有SAS/GRAPH 软件，标题、脚注等内容有更复杂的控制，详见 1 5 章。

数据转换，最简单的情形如常用对数转换，只需在DATA 步使用LOG10(.) 函数即可，也可用SAS 的函数来构造新量，常用的Box-Cox 转换可在SAS/QC 的ADX 宏定义中实现，见本章 §6。

【例4.4】下面程序对系统安装的教学数据CLASS.SSD 进行描述、分析。

```
data;
    N1='SAS INSTITUTE INC.';
    N2='SAS CIRCLE';
    N3='P.O. BOX 8000';
    N4='CARY, NC   27512-8000';
```

```
    N5='U.S.A.';
run;
proc forms copies=2 indent=10;
   line 1 n1; line 2 n2; line 3 n3; line 4 n4; line 5 n5;
run;
libname user 'sasinst';
options _last_=user.class;
proc contents;
run;
proc print;
   var age name sex height weight;
run;
proc univariate normal;
   var age height weight;
proc capability normaltest;
   var age height weight;
   cdfplot / normal;
   histogram /normal;
   qqplot ;
proc format;
   value $sexfmt 'F'='female'  'M'='male';
proc tabulate;
   class sex;
   var age height weight;
   table sex all, (age height)*(mean std);
   format sex $sexfmt.;
   keylabel all='Total';
proc sort;
   by sex;
proc summary noprint;
   by sex;
   var age;
   output out=test1 mean=m var=var;
proc print;
options _last_=user.class;
proc means noprint;
   by sex;
   var weight;
   output out=test2 mean=ubar var=variance;
proc plot;
   plot variance*ubar;
options _last_=user.class;
```

```
proc chart;
   hbar sex;
proc plot;
   plot (height weight)*age;
proc freq order=formatted;
   table sex*age /nopercent nocol norow expected chisq;
   format sex $sexfmt.;
run;
```

相应的产出如下，FORMS 的产出结果由于在过程步语句中指示两个拷贝，故有两个标签。

```
SAS INSTITUTE INC.
SAS CIRCLE
P.O. BOX 8000
CARY, NC   27512-8000
U.S.A.

SAS INSTITUTE INC.
SAS CIRCLE
P.O. BOX 8000
CARY, NC   27512-8000
U.S.A.
```

过程CONTENTS 和PRINT 的结果，可见第16章 § 2。

UNIVARIATE 对不同名称及标号所标识的变量产出三个部分的统计量，第一部分，第一栏依次为矩统计量，即数目、均值、标准差、偏度、未校正平方和、变异系数、总体均值为零的t-检验，符号秩和、不为零的数目、正态W-检验统计量。第二栏依次为权重总和、和、方差、峰度、校正平方和、标准误、第一栏t-检验的概率、符号秩次检验的概率、小于W-统计量的概率；第二部分为分位点统计量；第三部分为其极值及其对应的记录号。

[(19+1)19]/4=95 即符号秩和。由于程序中没有指定各观察的权，系统默认各记录的权重为 1 ，故权重的和是19。从t-检验的结果可以看到，据年龄的数据，应以0.0001 的概率拒绝总体均值为0 的假设。从W-检验的结果看，不能拒绝服从正态分布的假设。

过程CAPABILITY 的输出结果包括了UNIVARIATE 的输出内容，这里仅仅给出其特有的内容。对年龄来说，正态性拟合结果表明，数据来自正态总体的假设未被拒绝。除了UNIVARIATE 外，SAS 使用MEANS 和SUMMARY 输出综合统计量，MEANS 与UNIVARIATE 类似。分组数据处理以前，一般要先排序。本例结合了过程PLOT 的图示。

TABULATE 的输出内容是按性别给出年龄、身高、体重的均值与标准差，更详细的统计量可经SAS 的DMS 下运行HELP TABULATE 给出，或据SAS 系统说明书。SAS 的这一用法与Stata 类似(带有summarize 选项的table 命令)。

输出报表的样式与系统设置有很大关系，当分类较多时，应对OPTIONS 中的pagesize—P= 和linesize—L= 进行适当设置；分类多而页长过短时，指定合计(ALL) 时，会产出一些小表。这对PLOT 过程也适用，如设一个很大的页长，系统要画一个很大的纵轴。

	Age in years		Height in inches	
	MEAN	STD	MEAN	STD
Gender				
female	13.22	1.39	60.59	5.02
male	13.40	1.65	63.91	4.94
Total	13.32	1.49	62.34	5.13

SUMMARY 的产出结果，数据集TEST1.SSD 存贮了不同性别年龄的均值和方差。有两个特殊的量，_TYPE_ 表示计算统计量的类型，_FREQ_ 存贮了每一分组的例数。

PLOT 的产出似乎没有什么意义，考虑身高与体重的关系或许会好些。

TABLE OF SEX BY AGE

SEX(Gender) Frequency| Expected	AGE(Age in years) 11	12	13	14	15	16	Total
female	1	2	2	2	2	0	9
	0.9474	2.3684	1.4211	1.8947	1.8947	0.4737	
male	1	3	1	2	2	1	10
	1.0526	2.6316	1.5789	2.1053	2.1053	0.5263	
Total	2	5	3	4	4	1	19

TTEST 的产出结果，同样给出了分组下的样本例数、标准差(误)、极值。对体重来说，方差是齐的，启用通常的分组t-检验结果，体重在男女生之间没有差别。

§4.3.2 统计推断

SAS提供了密度函数、包括非中心分布在内的分布函数、分位点函数及随机函数。分布的拟合在SAS/QC 的过程CAPABILITY 内可以完成。

PROC TTEST用于t-检验, UNIVARIATE给出SHAPIRO-WILKS统计量，NPAR1WAY 进行非参分析, FREQ提供了许多列联表统计量。方差齐性的Bartlett 检验在SAS样本程序中提供。RANK过程用于生成秩次变量。

在SAS 过程中，一般拥有WEIGHT 或FREQ 语句指示计算与分析使用的权变量，这样可以考虑比较复杂的情形，如平均数是加权平均。多数情形下，用户需要借助SAS 给出的参数估计量和标准误来进行计算。

多元假设检验：多变量的均值检验在GLM 中提供了Hotelling-Lawley 迹和Wilks 统计量、Pallai 迹等，几个方差协方差阵的检验可经过程DISCRIM来完成。其计数与计算的概

念也是很明显的。如GLM 与CATMOD 则是结合计数与计量分析的典型。

现举一个使用PROC IML 的例子。

```
proc iml;reset print;
   y={ 1 2,3 4,5 6};
   g=ginv(y);
   z={1.0676 0.1848,0.1848 1.130};
   call svd(p,d,q,z);
   x={1 2  3  4  5,
      2 4  7  8  9,
      3 7 10 15 20,
      4 8 15 30 20,
      5 9 20 20 40};
   g=ginv(x);
   e=eigval(x);
   d=eigvec(x);
quit;
```

第一句调用PROC IML，第二句控制每步都输出结果。矩阵的赋值很简单，只消使用大括号()把元素括起来，矩阵的每行用逗号分开。GINV 是一个函数，用于求矩阵的广义逆；SVD 是一个过程，被CALL调用进行矩阵的奇异值分解，这个分解有重要的理论与实际意义，许多文献详有讨论。EIGVAL 与EIGVEC 给出矩阵的特征值与特征向量，注意IML许多运算是针对对称矩阵进行的，这更适应统计问题的处理。PROC IML 最后以QUIT 语句退出。PROC IML 的前身是PROC MATRIX ，SAS/IML 手册详细讨论了两者语句的转换的例子。IML 的多数例子在样本程序中出现，便于应用。结果如下：

下面是SAS/IML样本程序REG.SAS的部分内容，包括了一系列回归分析的过程，REGTEST1.SAS和REGTEST2.SAS 演示它的用法。程序仅仅是一个示范，不支持缺失值处理和共线性处理。

```
proc iml worksize=60;
/*-----REGEST: Regression Parameter Estimation-----
 *arguments:
 * x    the regressors, design matrix
 * y    the response, dependent variable
 * names the names of the regressors
 */
start regest;
  n=nrow(x);                  /* number of observations   */
  k=ncol(x);                  /* number of variables      */
  xpx=x‘*x;                   /* cross-products           */
  xpy=x‘*y;
  xpxi=inv(xpx);              /* inverse crossproducts    */
  beta=xpxi*xpy;              /* parameter estimates      */
  sse = y‘*y-xpy‘*beta;       /* sum of squares error     */
```

```
  dfe = n-k;                  /* degrees of freedom error */
  mse = sse/dfe;              /* mean square error      */
  rmse = sqrt(mse);           /* root mean square error */
  rsquare = 1-sse/((y-y[:])[##]);
  print ,,'Regression Analysis',,'Residual Error:'
           sse dfe mse rmse rsquare;
  stderr = sqrt(vecdiag(xpxi)#mse); /* std error of estimates  */
  tratio = beta/stderr;             /* test for parameter=0    */
  probt=1-probf(tratio##2,1,dfe);   /* significance probability */
  print ,,'Regression Parameter Estimates ',,
       names beta stderr tratio probt;
  covb=xpxi#mse;              /* covariance of estimates  */
  s=1/stderr;
  corrb=s#covb#s`;            /* correlation of estimates */
  print ,"Covariance of estimates",  covb[r=names c=names],
         "Correlation of estimates",corrb[r=names c=names];
finish;
```

【例4.5】两组t-检验，继续用第三节的数据，比较不同性别的学生体重相同吗?

程序为：proc ttest; class sex; var weight;

可见与通常的演算不同，程序是指定一个分组变量，不必输入两组排好的数据。程序运行结果如下：

性别	数目	均值	标准差	标准误	最小值	最大值
F	9	90.1111111	19.38391372	6.46130457	50.50	112.50
M	10	108.9500000	22.72718636	7.18696737	83.00	150.00

方差	t-值	自由度	P 值
不等	-1.9493	17.0	0.0680
相等	-1.9322	17.0	0.0702

针对检验H0: 方差相等, F' = 1.37, DF = (9,8), P = 0.6645

可认为方差是相等的，故使用表中第二行的t-值，经检验两组无差别。

【例4.6】非参检验，格式与参数检验相仿，下面给出相应程序，输出结果从略。

```
PROC NPAR1WAY WILCOXON;
   CLASS SEX;
   VAR WEIGHT;
RUN;
```

§4.4 多元统计分析

SAS 的多元分析过程格式如下：

```
PROC 过程名DATA= OUT= OUTSTAT= 其它过程选项; /* 必选项*/
  VAR 变量表;
  ID 变量;
  MODEL 模型/选项;
  OUTPUT OUT= 选项;
  BY 变量表;
  WEIGHT 变量;
  FREQ 变量;
  WHERE 条件;
      ....
RUN;
```

DATA= 指示的数据集指示为TYPE=CORR, COV 或SSCP 等，这时一些需要用原始数据的选项就不能产生结果。过程选项OUT= 生成的数据集多含有原始数据，用OUTSTAT=生成的数据集含有模型及有关参数。若要生成永久性数据集，则应使用由“库名.文件名”组成的两水平文件名。

ID 语句对OUT=中的原始变量进行标识，BY 语句指示按变量的不同取值分组分析，分别计算，这种分组可以是格式定义，BY 语句隐含数据是按升序排列的，使用NOTSORTED或DESCENDING指示观察未排序或按降序排列。WEIGHT 指示每个记录使用的权重。当FREQ语句出现时，表示输入数据集符合特定条件的记录不至一个。WHERE 用于对数据集进行筛选，如：WHERE AGE¿5;表示过程仅对年龄大于5 岁的对象进行分析。OUTPUT OUT=生成由原始数据生成的新变量。

许多SAS/STAT过程可以交互式运行，如PROC CATMOD和GLM等都是交互式过程，执行RUN;语句出现光标后，在窗口右下角标志行仍有R提示，表示过程并没有退出运行，仅当执行了QUIT命令以后才算过程结束。要先退出交互式过程，然后才能退出SAS系统。

其它的语句也可以结合使用，这里给出一个使用FORMAT 语句的例子。CATMOD 过程进行LOGISTIC 分析时，为了便于解释，对连续变量要进行一些分组。通常分组变量可以在数据步产生，但使用FORMAT语句后就没有必要这么做。注意这种格式通常是在FORMAT过程中定义的。程序如下：

```
title2 'Logistic 多因素分析';
proc catmod;
response clogit;
model pass27=streptas single duration asa mf/ml nogls;
format streptas str. single single. duration dur. asa asa. mf $sex.;
run;
```

上述程序用于一个心肌梗塞药物的疗效分析。其中streptas表示发病时间，是一个连续变量，但使用STR格式后定义为分组变量，原始数据集并未做任何改动。

本节结合几种统计分析，简介几个统计过程的用法，过程的选项很多，但只需对具体过程的用法有一个概念，结合其手册和有关文献能得到进一步的理解。

§4.4.1 回归分析

这里介绍REG 过程的使用方法。REG 过程用最小二乘方法拟合线性回归模型。对因变量能够最佳拟合的自变量子集可由多种模型选择方法确定。REG 可以交互式使用。

REG 是一个通用的回归过程，SAS 中其它的回归过程进行更特殊的回归。REG 过程有九种模型选择方法，能够对线性假设的多变量假设进行检验，产生数据和各种统计量的散点图，计算共线性诊断和影响统计量，产生偏回归图，并且把预测值、残差、岭回归估计和可信限等统计量输出到SAS数据集。

语句格式及说明如下：

PROC REG 过程选项;
标号: MODEL 因变量= 自变量表/ <选项>;
BY 变量;
FREQ 变量; ID 变量;
VAR 变量表; ADD 变量表;
DELETE 变量表; WEIGHT 变量;
REWEIGHT <条件|ALLOBS></选项> | <STATUS|UNDO>;
标号: MTEST <方程1, ... 方程k / 选项>;
OUTPUT OUT=SAS 数据集关键字=存贮名...;
PAINT <条件|ALLOBS></选项> | <STATUS|UNDO>;
PLOT <y1*x1><=符号1>, ... <yk*xk><=符号k></选项>;
PRINT <选项ANOVA MODELDATA>;
REFIT;
RESTRICT 方程1, ... 方程k;
标号: TEST 方程1, ... 方程k / 选项;

其中的标号是可选的，PROC REG 是必选项，若要拟合模型，则MODEL语句也是必选的。若只用PROC REG 的过程选项，则MODEL语句非必需，但须有VAR语句。

1. PROC 语句启用回归过程，选项包括数据集选项、打印及其它信息。DATA= 指示REG操作的SAS数据集，OUTEST=指示存放参数的数据集，OUTSSCP= 指示TYPE= SSCP类型的数据集。这些数据集的命名规则同一般SAS数据集相同，如使用两水平的名字user.mydata。

   ALL 与MODEL语句中的ALL相当，包括了SIMPLE, USSCP, CORR的结果。

   不产生输入则使用NOPRINT，SIMPLE 用于打印简单的描述统计量，COVOUT指示生成协方差阵的数据集，检验奇异性的准则使用SINGULAR=指示。ALL 打印所有统计量，USSCP是未修正的矩阵。

2. MODEL 语句选项：MODEL 语句指示分析的模型，模型选择方法由SELECTION 指定，如SELECTION=FORWARD(F,向前), BACKWARD(B, 向后), STEPWISE(逐步) , MAXR,

MINR, RSQUARE, ADJRSQ, CP或NONE。筛选的细节可由DETAILS给出。变量的进入或删除可以成组进行，这通过大括号指定。每组变量用GROUPNAMES='名字1' ' 名字2'... 指示，用于FORWARD，BACKWARD或STEPWISE。如：model y={ht wgt age} bodyfat/selection=stepwise groupnames='hwa' 'f'; INCLUDE=n 指示模型的头n个变量一直保留在方程中。
SLENTRY|SLE=值及SLSTART—SLS=值表示进入或删除变量的显著性水平。
选项I和XPX指示$(X'X)^{-1}$及$X'X$矩阵。

ps 假设方差不齐，ACOV 打印渐近协方差阵，COLLIN 指示多重共线性分析。COLLINOINT指示没有截距项的多重共线性分析。CORRB 打印估计量相关阵。COVB打印估计量的估计协方差阵。PCORR1 打印平方偏相关系数，即 I 型平方SS与SS+ SSE 的比值，SSE是误差平方和。PCORR2 使用 II 型平方和进行与PCORR1 类似的计算。SCORE1使用 I 型平方和计算半偏相关系数SS/SST，SST 是修正的总平方和。指定NOINT时使用未修正总平方和。SCORE2与SCORE1类似，但用 II 型平方和进行计算。SEQ表示当一个变量进入模型时，打印出由一行一行估计量组成的矩阵。SPEC 指示关于模型的一阶和二阶矩的检验。SS1指示 I 型平方和。SS2指示 II 型平方和。STB 表示标准偏回归系数。TOL 打印估计量的容许值，即$1-R^2, R^2$是该变量与模型中其它变量回归时的复相关系数。VIF 即方差澎涨因子，它是容许值的倒数。

用于预测和残差分析的统计量通常可以由MODEL语句使用相应的选项得到，但在输入为TYPE=CORR, COV, SSCP 几种特殊类型的数据集时不能进行。这些选项有CLI(个体预测值的95%可信限)、CLM(因变量的95%可信限)、DW(Durbin-Watson统计量)、INFLUENCE(影响统计量)、P(预测值)、PARTIAL(偏回归杠杆图)、R(残差)。

NOPRINT 将不打印回归结果。ALL 选项的功能与使用众多的选项相当。这些选项是：ACOV、CLI、CLM、CORRB、COVB、I、P、PCORR1、PCORR2、R 、SCORE1 、SCORE2、SEQB、SPEC、SS1、SS2、STB、TOL、VIF、XPX。

仅仅用于RSQUARE, ADJRSQ, CP中的选项：RDJRSQ 是调整了自由度的复相关系数。AIC 计算每个模型的Akaike信息准则。B 计算回归系数。BIC 计算Bayes 信息准则。CP 计算Mallows的Cp统计量。GMSEP 假设回归自变量与因变量均符合多元正态分布并计算预测均方误差。假设回归自变量是固定的，JP 指示计算预测均方误差。MSE 指示计算均方误差。PC 指示计算Amemiya预测准则。RMSE 打印均方误差的方根。SBC 计算每个模型的SBC统计量。SIGMA=n 指示计算CP及BIC准则所使用的误差项标准差。SP 计算Hocking的Sp统计量。SSE 计算每个模型的误差平方和。

3. OUTPUT 语句产生一个输出数据集，其中的统计量针对每个记录。对于每个统计量，指定一个关键字，一个等号，以及统计量在输出数据集中对应的变量名。若不指定OUT=，则产生的输出数据集按DATAn 的习惯命名法。

   可以用做关键字的输出统计量有：

   predicted|p= 预测值
   residual|r= 残差
   L95M=，U95M= 因变量预测值(均值) 95%可信上下限
   L95=，U95= 个体预测值95%可信上下限

STDP=平均预测值标准误
STDR=残差标准误
STDI=个体预测值标准误
STUDENT=学生化残差(标准化残差)
COOKD=库克氏距离
H=杠杆
PRESS=预测均方误差
RSTUDENT=删除本记录后的学生化残差
DIFFITS=本观察删除后对预测值的影响
COVRATIO=本观察对回归系数协方差的影响

4. PLOT 语句PLOT 语句用y和x做散点图，点的符号用引号括起或是输入数据集中的变量名。y 变量和x 变量可以是在第一个RUN语句前的VAR 或MODEL 语句包含的任意变量，也可以是OUTPUT语句中的统计量，或者OBS(记录号)。
PLOT 选项为：CLEAR，COLLECT，HPLOTS=，NOCOLLECT，OVERLAY，SYMBOL=，VPLOTS=。

5. RESTRICT 语句RESTRICT 语句用于对MODEL 语句中的参数施加约束。可以用RESTRICT 指定几个约束，约束之间用逗号分隔；几个约束语句也是允许的。指定的约束在下一个MODEL 语句指定前一直有效。

PROC REG 是一个交互式过程，用RUN; 分隔各次运行。如针对指定的模型，用ADD/DELETE 增加/减少变量。使用PAINT语句可使符合特定条件的观察在散点上列出，这对模型的分析和考核很有用。如：PAINT name='Henry'—name='Mary'; 及PAINT obs.¿=11 and residual.¡=20; 等。PAINT的选项包括NOLIST和RESET，用于记录号和图示符号的变动。语句PLOT的用法与PROC PLOT类似。RESTRICT 用于有约束回归分析。REWEIGHT 用于改变参与计算时的各观察的权重。如：REWEIGHT name='Alan'; ...; reweight /weight=0.5。

PROC REG 一次运行若变量集相同，也可指定多个模型，利用这个特点，进行基于线性回归的通径分析很方便。

6. 输入和输出数据集

OUTEST=选项产生一个TYPE=EST的数据集。其内容有：
BY 变量。
MODEL 字符变量，默认为MODELn，包含MODEL语句的标号。
TYPE 字符变量，对每个记录指示'PARMS'。
DEPVAR 因变量名。
RMSE 均方误差的方根，也是误差项标准标的估计。
INTERCEP 估计截距。
MODEL 语句中指定的所有变量，其值是回归系数，不在模型中的自变量为缺失值，因变量为-1。
若指定COVOUT，则输出估计的协方差阵，_TYPE_取值为'COV'而每行用八个

字符的变量_NAME_ 标记。
对于RSQUARE，ADJRSQ，和CP 方法，REG 对每个子集模型输出一条记录。附加的变量有：
_IN_ 模型中自变量的数目，不包括截距。
_P_ 模型中参数的数目，包括截距。
EDF 误差自由度。
SSE 误差平方和。
MSE 均方误差。
RSQ 复相关平方统计量。
ADJRSQ 调整复相关平方。
_CP_ Mallows C_p 统计量。
其它指定时产生的统计量有：_SP_、_JP_、_PC_、_GMSEP_、_AIC_、_BIC_ 、_SBC_。

【例4.7】通径分析(path analysis) 是利用专业与统计学的知识，描述变量间联系的结构关系进行定量分析。它把因素间的相互影响用图示的方法表达出来并且把它们区分为几种情况，一个因素可以完全受另一个因素影响，也可以受几个因素的共同影响，也可以反过来，两个因素受一个共同的因素影响。在一些假设下，可以得出循路径的方法，称作通径分析法。现有美国41个城市平均气温(X1)、企业数(X2)、人口数(X3)、平均风速(X4)、平均降水量(X5)、平均降水天数(X6)对大气SO_2(Y)的影响(《中国医学百科全书》第一卷，预防医学，上海科学技术出版社，1991.12)。

```
data path (type=CORR);
infile cards missover;
input _type_ $ _name_ $ x1-x6 y ;
cards;
CORR x1 1.000
CORR x2 -.188 1.000
CORR x3 -.063  .955 1.000
CORR x4 -.350  .237  .213 1.000
CORR x5  .424  .029  .017  .005 1.000
CORR x6 -.430  .131  .042  .164  .443 1.000
CORR  y -.434  .645  .494  .095  .015  .370 1.000
   N  .    41    41    41    41    41    41    41
proc reg data=path;
m1:model x5=x1 x6/stb;
m2:model y=x1-x6/stb;
m3:model y=x1-x2/stb;
run;
```

自变量的筛选是在MODEL 语句中的选项SELECTION= 中指示向前、向后、逐步法。

【例4.8】汽车流量、风速对NO_2 的影响[9]，进行有关的回归诊断计算，原始数据和使用λ =0.6 的Box-Cox 转换同时计算。

x1: 交通点汽车流量(辆/小时) x2: 风速(米/秒) y: 大气NO2 含量

```
data guo;
%put NOTE: A transportation data.;
input x1 x2 y @@;
format x1 x2 y 24.4;
ty=(y**0.6-1)/0.6;
cards;
 1300        .45       .066       948          2       .005
 1444         .5       .076      1440        2.4       .011
  736        1.5       .001      1080          3       .003
 1652         .4        .17      1844          1        .14
 1736         .8       .156      1116        2.8       .039
 1754         .8        .12      1656       1.45       .059
 1200        1.8        .04      1536        1.5       .087
 1500         .6        .12       960        1.5       .039
 1200        1.7         .1      1784         .9       .222
 1476        .65       .129      1496        .65       .145
 1820         .4       .135      1060       1.83       .029
 1436          2       .099
proc reg data=guo;
var y ty x1 x2;
   model1:model y =x1 x2;
   output out=a1 p=yhat r=e h=h student=s rstudent=r
          cookd=c press=p covratio=c dffits=d;
   run;
   model2:model ty=x1 x2;
   output out=a2 p=yhat r=e h=h student=s rstudent=r
          cookd=c press=p covratio=c dffits=d;
quit;
proc print data=a1;
proc plot data=a1; plot e*yhat;
run;
proc print data=a2;
proc plot data=a2; plot e*yhat;
run;
```

上述程序还使用OUTPUT语句输出预测值(predict=yhat)、残差(residual=e)、杠杆(h=h)等，接下来使用PLOT过程绘制残差对预测值的图。

原始数据和转换数据同时进行上述分析以供比较，SAS 提供了PRINQUAL 和TRANSREG 过程可进行更为复杂的转换，如本例：

```
proc transreg data=guo method=morals;
model power(y /parameter=0.6) =linear(x1 x2);
output out=a;
```

```
run;
```

运行结果：

$$\hat{y} = \underset{(0.0489)}{-0.038839} + \underset{(0.0003)}{0.000116}x1 - \underset{(0.0111)}{0.027813}x2$$

R**2=0.7329, F=27.439，P<0.001

t0=-0.739, P¿0.4，t1=4.229，P¡0.001，t2=-2.502，P<0.03

原回归方程残差对预测值的图示表明，存在着一定程序的方差不稳。

```
          Plot of E*YHAT.  Legend: A = 1 obs, B = 2 obs, etc.
    0.10 +
         |
R        |                                          A
e        |
s   0.05 +                 A
i        |
d        |      A               A            A      A
u        |           A                      A      A
a   0.00 +   B       A                       A
l        |      A         A            A             A
         |                                         A    A
         |                                A   A
   -0.05 +                   A              A
         ---+-------------+-------------+-------------+-------------+--
          0.00          0.05          0.10          0.15          0.20
                                Predicted Value of Y
```

有关的诊断统计量列表如下：

记录	残差	STUDENT	库克矩离	帽子矩阵	PRESS	RSTUDENT	DFBETA
1	-0.033608	-1.11933	0.08401	0.16747	-0.040368	-1.12685	-0.50541
2	-0.038939	-1.25541	0.06596	0.11155	-0.043828	-1.27488	-0.45173
3	-0.003910	-0.14368	0.00318	0.31619	-0.005718	-0.14012	-0.09528
4	0.028126	0.90602	0.03383	0.11003	0.031603	0.90178	0.31708
5	0.015496	0.49537	0.00871	0.09627	0.017147	0.48582	0.15856
6	-0.022594	-0.72448	0.01983	0.10182	-0.025155	-0.71559	-0.24093
7	-0.010448	-0.32846	0.00252	0.06559	-0.011181	-0.32101	-0.08505
8	0.001339	0.04261	0.00006	0.08749	0.001468	0.04153	0.01286
9	0.046771	1.46797	0.04793	0.06255	0.049891	1.51473	0.39126
10	0.014517	0.46081	0.00644	0.08345	0.015839	0.45154	0.13625
11	-0.026383	-0.86342	0.03969	0.13774	-0.030597	-0.85770	-0.34280
12	0.026709	0.85724	0.02828	0.10349	0.029792	0.85132	0.28925
13	-0.010622	-0.34757	0.00642	0.13747	-0.012315	-0.33979	-0.13566
14	-0.050630	-1.70873	0.22712	0.18920	-0.062445	-1.80220	-0.87059
15	-0.000138	-0.00486	0.00000	0.25788	-0.000186	-0.00474	-0.00279
16	-0.007482	-0.24602	0.00344	0.14572	-0.008759	-0.24015	-0.09919
17	0.026119	0.89267	0.07033	0.20936	0.033035	0.88793	0.45691
18	-0.054135	-1.73513	0.11281	0.10105	-0.060221	-1.83493	-0.61520
19	-0.010810	-0.34015	0.00278	0.06734	-0.011590	-0.33250	-0.08935
20	0.008078	0.26732	0.00443	0.15668	0.009579	0.26102	0.11251
21	0.078704	2.54048	0.27589	0.11366	0.088797	3.00875	1.07746
22	0.028195	0.89398	0.02361	0.08142	0.030694	0.88929	0.26476
23	-0.004356	-0.13927	0.00069	0.09656	-0.004822	-0.13581	-0.04440

对y 使用Box-Cox 转换，λ =0.6，回归效果有所改善，现将结果列如下。

$$\hat{y} = -1.692660 + 0.000353x1 - 0.085791x2$$
$$(0.1365) \quad (0.00008) \quad (0.0310)$$

```
    R**2=0.7675, F=33.018, P<0.001
    t0=-12.399, P<0.01, t1=4.617, P<0.001, t2=-2.771, P<0.02
             Plot of E*YHAT.  Legend: A = 1 obs, B = 2 obs, etc.
   0.2 +
       |                    A
R      |                                          A
e      |        A                   A
s      |              A                          A       A
i      |                                        A       A
d  0.0 +               A    A               A      A
u      |    A                                            A
a      |     A   A                           A           A      A
l      |                                        A
       |                                       A
```

```
      |                        A
 -0.2 +
      --+----------+----------+----------+----------+----------+----------+--
      -1.6       -1.5       -1.4       -1.3       -1.2       -1.1       -1.0
                              Predicted Value of TY
```

残差与YHAT 的图示亦表明效果有所改善。

§4.4.2 方差分析

ANOVA 过程用于分析各种平衡试验设计的方差分析，若不平衡，则宜用GLM过程，GLM的用法与ANOVA 与之相似。ANOVA 可以交互式使用。

语句格式及其用法说明如下：

PROC ANOVA DATA= MANOVA MULTIPASS OUTSTAT=;
  CLASS 分组变量表; /* 必选*/
  MODEL 因变量=效应/ 选项; /* 必选*/
  ABSORB 变量表;
  BY 变量表;
  FREQ 变量;
  MANOVA H= 效应E= 效应M= 方程...
  MNAMES= PREFIX= / 选项;
  MEANS 效应/ 选项;
  REPEATED 因素名水平(水平取值) 转换<,...> / 选项;
  TEST H= 多个效应E= 效应;

ANOVA 过程可以指定主效应、交互效应和区套效应。主效应由自变量名指定，如：a b c；交互效应由星号联结两个变量指定，如：a*c b*d；区套效应是在主效应或交互效应后的括号内列出，如：a(b d) c*f(d)。对于高阶交互，为了书写方便，用竖线分隔这些效应并用@符号跟随一个整数指定最高次交互，如模型MODEL Y=A B C A*B A*C B*C 可以简单地写成MODEL Y=A|B|C|@2;。

ANOVA 过程是交互式的，也就是说用一个CLASS 和MODEL 语句可以进行个分析，请求进行的分析之间用RUN 语句分隔。在使用QUIT 语句或碰到下一个DATA 或PROC 语句出现时，交互式的运行结束。若指定ABSORB 和FREQ 语句，则它们应在第一个RUN 语句之前出现并在以后的运行中有效。BY 语句不能用于交互式运行，因而指定了BY 后只能有一次运行。

过程GLM、NESTED 的用法与ANOVA类似。

1. PROC 语句DATA= 指示数据集，MANOVA 指示以变量态进行有缺失值记录的删除，即若记录中任何一个自变量缺失，则舍弃这个记录。MULTIPASS 指示在必要的时候重新读取数据集而不是把自变量写入暂存文件。这样做会节省磁盘空间，但一般说来，程序的执行时间要长。OUTSTAT=数据集包括平方和、F值、模型各效应的概率值。若在MANOVA语句中指示CANONICAL 但没有使用M=指示，数据集也包括了典型分析的结果。

2. ABSORB 语句对于某些类型的模型节省时间和存贮，使用该语句要求数据集(每个BY组)按ABSORB 变量排序，在CLASS 或MODEL 语句中使用ABSORB变量可以产生错误的均方。在交互式运行时，ABSORB 语句应在第一个RUN语句之前出现。

3. MANOVA 语句若MODEL语句中包含不至一个因变量，使用MANOVA语句可以获得多元统计量。一旦指定该语句，ANOVA 便把分析变量有缺失的记录将被忽略，MANOVA 作为过程选项也能做到这一点。H= 指示假设的矩阵，E=指示误差项，M=指示因变量的转换矩阵，PREFIX=用于指示M=所生成变量的前缀。CANONICAL 不打印特征根而进行H矩阵和E矩阵的典型分析。ORTH 要求M= 的转换阵行正交规格化。PRINTE 要求打印误差SSCP矩阵E。PRINTH 要求打印H矩阵，SUMMARY要求对每个因变量打印方差分析表。用例：

```
proc anova;
class a b;
model y1-y5=a b(a);
manova h=a e=b(a) /printh printe;
manova h=b(a) /printe;
manova h=a e=b(a) m=y1-y2,y2-y3,y3-y4,y4-y5; prefix=diff;
manova h=a e=b(a) m=(1 -1  0  0  0,
                     0  1 -1  0  0,
                     0  0  1 -1  0,
                     0  0  0  1 -1,
                     0  0  0  0  1) prefix=diff;
```

第一个MANOVA语句指示A是假设效应，B(A)是误差项，选项PRINTH要求打印与A 有关的矩阵，PRINTE要求打印与B(A)有关的误差阵。第二个MANOVA语句指示B(A)为效应矩阵，PRINTE要求打印误差矩阵。第三个MANOVA语句进行的分析与第一个相同，但分析是针对连续的变量差值而进行的，经过转换的变量名为DIFF1,DIFF2,DIFF3, DIFF4。第四个MANOVA语句使用了M=选项，其作用与第三句相同。

4. MEANS 语句对于MODEL右端出现的任何效应计算均值，MEANS 语句只能在MODEL语句后出现。可以用星号联结自变量得到组合的水平。在一个MEANS 语句中可以列出一个或多个效应。MEANS 语句有许多选项进行多重比较，选项如：

   BON Bonferroni t-检验

   DUNCAN Duncan 的multiple-range test

   DUNNETT 进行Dunnett 双尾检验，如means a /DUNNETT('CONTROL') ; means a b c d/DUNNETT('CNTLA' 'CNTLB' 'CNTLC' 'CNTLD'); 括号内是对照水平的取值，默认是第一组为对照。

   DUNNETTL Dunnett 单侧检验，检验处理是否比对照低。

   DUNNETTU Dunnett 单侧检验，检验处理是否比对照高。

   GABRIEL 进行Gabriel 两两比较。

REGWF 进行Ryan-Eliot-Gabriel-Welsch 多F test。

REGWQ 进行Ryan-Eliot-Gabriel-Welsch 多均数比较test。

SCHEFFE 进行Scheffe 两两检验。

SIDAK 进行Sidak 两两检验。

SMM, GT2 进行两两比较，样本不等时即Hochberg 的GT2方法。

SNK Student-Newman-Keuls 多均数test。

T,LSD 两两 t 一检验，所有格子数相同时即Fisher的LSD法。

TUKEY 进行Tukey 的HSD。

WALLER 进行Waller-Duncan k-ratio 检验。

以下选项用于指定多重比较的细节：

ALPHA=指示均值检验的显著性水平。CLDIFF 要求BON、GABRIEL、SHEFFE 、SIDAK、SMM、GT2、T、LS可信区间形式给出。CLM 指示对于MEANS 的各个水平，BON、GABRIEL、SCHEFFE、DIDAK、SMM、T以及LSD 选项以可信区间的形式给出。E=指示用于两两比较的误差均方，KRATIO=可以指示50,100,500用于Waller -Duncan检验。LINES指示仅不能用于Dunnett 的三种检验，它把均值以从小到大的形式排列并指出均值的差别情况。

5. MODEL 语句INT—INTERCEPT 要求模型打印模型中的截距效应。NONUI 不打印一元分析的结果。

6. REPEATED 语句若MODEL 语句中的因变量表示对同一实验单元的重复测量，则检验测量因素以及它们与MODEL 语句中自变量间的交互可用REPEATED 语句。在REPEATED 语句中指定以下信息。

因素名　　水平　　水平取值　　转换　　选项

选项有CONTRAST、POLYNOMIAL、HELMERT、MEAN和PROFILE。在斜线后可以使用NOM、NOU、PRINTE、PRINTH、PRINTM、PRINTRV和SUMMARY。NOM与NOU分别控制一元与多元分析结果的输出。

如进行析因设计方差分析，使用语句：

```
proc glm;
    classes a b c;
    model y1-y3=a|b|c;
    manova h=a|b|c /printe printh;
```

方差分析中的效应变量可使用DATA 步中的循环来构造，现具两例。

【例4.9 】下表数据记录了一个石油公司下属四个油田分别采用三种方法采油后，每两口井的出油桶数，现要看方法的有效性及在油田间的差异。

	产油量			
方法	油田一	油田二	油田三	油田四
法一	2,1	4,2	3,1	1,1
法二	4,5	3,3	6,7	6,5
法三	6,4	8,8	7,8	5,6

采用析因设计方差分析，其程序如下：

```
data oil;
   do method=1 to 3;
      do field=1 to 4;
         do reps=1 to 2;
            input barrels @@;output;
         end;
      end;
   end;
   cards;
2 1 4 2 3 1 1 1
4 5 3 3 6 7 6 5
6 4 8 8 7 8 5 6
proc anova;
   class method field;
   model barrels=method field method*field;
   test h=method e=method*field;
quit;
```

程序将自动产生效应变量method, field 用于后续分析。

产出结果：F=14.48，P<0.001。

R^2 变异系数均方误差方根BARRELS 均值0.929596 19.60812 0.866025 4.41666667

来源	自由度	平方和	均方	F 值	P
模型	11	118.8333333	10.8030303	14.40	0.0001
方法	2	88.08333333	44.04166667	58.72	0.0001
油田	3	9.83333333	3.27777778	4.37	0.0268
方法*油田	6	20.91666667	3.48611111	4.65	0.0115
误差	12	9.0000000	0.7500000		
校正平方和	23	127.8333333			

使用METHOD*FIELD 的均方做误差项进行检验：

来源	自由度	Anova SS	均方	F 值	Pr ¿ F
方法	2	88.08333333	44.04166667	12.63	0.0071

【例4.10】一个公司拟选择车型，共有五种，其价格与维修差不多，现看其耗油量与里数的关系，使用嵌套设计的方差分析，在程序中，嵌套效应放在括号内。

```
data taxis;
   do type=1 to 5;
      do car=1 to 2;
         do rep=1 to 3;
            input miles@@;output;
         end;
      end;
   end;
```

```
   cards;
15.8 15.6 16.0 13.9 14.2 13.5
18.5 18.0 18.4 17.9 18.1 17.4
12.3 13.0 12.7 14.0 13.1 13.5
19.5 17.5 19.1 18.7 19.0 18.8
16.0 15.7 16.1 15.8 15.6 16.3
proc print;
proc anova;
   class type car rep;
   model miles=type car(type);
   test h=type e=car(type);
run;
```

计算结果如下，应当使用22.64 的F 值解释。

R^2	变异系数	均方误差方根	BARRELS 均值
0.971420	2.776601	0.447958	16.1333333

来源	自由度	平方和	均方	F 值	P
模型	9	136.4133333	15.1570370	75.53	0.0001
类型	4	129.2766667	32.3191667	161.06	0.0001
汽车(类型)	5	7.1366667	1.4273333	7.11	0.0006
误差	20	4.0133333	0.2006667		
校正平方和	29	140.4266667			

使用汽车(类型) 的均方做误差项进行检验：

来源	自由度	Anova SS	均方	F 值	Pr> F
类型	4	129.2766667	32.3191667	22.64	0.0021

Cody, R.P. 与Smith, J.K. (1991) 介绍了许多SAS 用于重复测量数据分析的用例。

【例4.11】协方差分析用例[17]：27只老鼠分成三组，第一组为控制组，第二组有甲状腺素，第三组在其饮用水中加入硫尿嘧啶，起始重量与1,2,3,4 周后增加情况。现在欲分析三种实验处理其体重的增加是否相同，而将起始体重对体重增加的作用一并考虑。程序如下：

```
title 'MANCOVA';
data;
input x0-x4 group@@;
cards;
57 29 28 25 33 1 59 26 36 35 35 2 61 25 23 11  9 3
60 33 30 23 35 1 54 17 19 20 28 2 59 21 21 10 11 3
52 25 34 33 41 1 56 19 33 43 38 2 53 26 21  6 27 3
49 18 33 29 35 1 59 26 31 32 29 2 59 29 12 11 11 3
56 25 23 17 30 1 57 15 25 23 24 2 51 24 26 22 17 3
46 24 32 29 22 1 52 21 24 19 24 2 51 24 17  8 19 3
51 20 23 16 31 1 52 18 35 33 33 2 56 22 17  8  5 3
63 28 21 18 24 1                  58 11 24 21 24 3
49 18 23 22 28 1                  46 15 17 12 17 3
```

```
57 25 28 29 30 1                  53 19 17 15 18 3
proc sort; by group;
proc means; var x0-x4; classes group;
proc plot; by group; plot (x1-x4)*x0;
proc reg; model x1-x4=x0; mtest;
proc glm;
   classes group;
   model x1-x4=x0 group x0*group;
   manova h=x0*group;
proc glm;
   classes group;
   model x1-x4=x0 group;
   manova h=group /printe printh;
   means group;
   lsmeans group/stderr pdiff;
quit;
```

GROUP 是有3个水平的处理变量，协变量为x0。分析思路：首先对数据排序，计算基础统计量和进行图示。然后看x1-x4 与x0 间存在直线关系吗？若存在直线关系，观察协变量与分组效应的交互作用是否有意义。最后检验调整总体均值向量是否相同和获得调整均值。MANOVA 语句检验处理平均值的多变量检验，PRINTE 与PRINTH 选择项印出误差阵与假设平方和矩阵。MEANS GROUP 是求出所有实验处理的平均值而LSMEANS 则求出调整均值，STDERR为其标准误，PDIFF 给出调整均值相等检验的的概率值。

基础统计量如下：

GROUP	N Obs	Variable	N	Mean	Std Dev
1	10	X0	10	54.0000000	5.4365021
		X1	10	24.5000000	4.8362060
		X2	10	27.5000000	4.7434165
		X3	10	24.1000000	5.8774522
		X4	10	30.9000000	5.5467708
2	7	X0	7	55.5714286	2.9920530
		X1	7	20.2857143	4.3094580
		X2	7	29.0000000	6.4031242
		X3	7	29.2857143	8.8828352
		X4	7	30.1428571	5.3984125
3	10	X0	10	54.7000000	4.6916001
		X1	10	21.6000000	5.3789714
		X2	10	19.5000000	4.2229532
		X3	10	12.4000000	5.3995885
		X4	10	15.8000000	6.8280467

多变量检验F=1.5282，P=0.2286，不显著但为示范继续分析。

带交互项的模型X0*GROUP 效应Wilks's 近似F=0.9430, P=0.4943 故多变量协方差分析有意义。继续做协方差分析X1，F=2.94, P=0.0727。X2，F=9.05 ，P=0.0013。X3，F=14.58，P=0.0001。X4，F=18.49，P=0.000
未调整协变量的Wilks' F=5.0694，P=0.0002。各组间有显著差异。调整均值如下：

```
GROUP          X1     Std Err      Pr > |T|    Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
           LSMEAN      LSMEAN   H0:LSMEAN=0   i/j      1       2       3
1      24.8536109  1.3837468         0.0001   1    .      0.0291  0.1075
2      19.8058138  1.6559283         0.0001   2  0.0291   .      0.4179
3      21.5823195  1.3778612         0.0001   3  0.1075  0.4179   .
GROUP          X2     Std Err      Pr > |T|    Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
           LSMEAN      LSMEAN   H0:LSMEAN=0   i/j      1       2       3
1      27.4555825  1.6310956         0.0001   1    .      0.5363  0.0022
2      29.0602809  1.9519304         0.0001   2  0.5363   .      0.0010
3      19.5022209  1.6241579         0.0001   3  0.0022  0.0010   .
GROUP          X3     Std Err      Pr > |T|    Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
           LSMEAN      LSMEAN   H0:LSMEAN=0   i/j      1       2       3
1      24.0318380  2.1368946         0.0001   1    .      0.1240  0.0008
2      29.3782198  2.5572195         0.0001   2  0.1240   .      0.0001
3      12.4034081  2.1278054         0.0001   3  0.0008  0.0001   .
GROUP          X4     Std Err      Pr > |T|    Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
           LSMEAN      LSMEAN   H0:LSMEAN=0   i/j      1       2       3
1      30.7851951  1.9374437         0.0001   1    .      0.8741  0.0001
2      30.2986638  2.3185369         0.0001   2  0.8741   .      0.0001
3      15.8057402  1.9292030         0.0001   3  0.0001  0.0001   .
```

据Huitema, B (1980) The Analysis of Covariance and Alternatives, New York: Wiley 中的建议，协变量的数目应满足公式：C+(J-1)/N¡0.10，C 为协变量数目本例为27，J 是组数本例为3，代入此式C+(3-1)/27¡0.10 即C¡0.7，本例不宜带有协变量。

【例4.12】轮廓分析用例：45 名受试者接受放射性处理之后三天，测精神运动分数。使用PROC GLM 进行轮廓分析[17]。

原始数据已含在SAS分析程序中，原始数据的均值与标准差：

组别	样本	均值	标准差	均值	标准差	均值	标准差
对照组	6	133.00	73.66	159.33	86.34	165.50	95.70
37.5r	14	105.07	73.10	129.57	75.12	138.21	82.71
87.5r	15	151.80	62.86	169.60	57.67	178.73	72.57
187.5r	10	169.50	65.40	191.80	65.62	193.40	69.67

经初步分析，不能拒绝轮廓相同的假设。对轮廓间等条件进行检验，表明应拒绝等条件的假设，即精神运动分数每天都改变。最后是等水平的检验，结果是四组的平均精神运动分数有显著的差异。相应的程序：

```
data;
input y1-y3 a @@;
u1=y1-y2;u2=y2-y3;z=(y1+y2+y3)/3;
```

```
b=1;
cards;
223 242 248 1  53 102 104 2 206 199 237 3 202 229 232 4
 72  81  66 1  45  50  54 2 208 222 237 3 126 159 157 4
172 214 239 1  47  45  34 2 224 224 261 3  54  75  75 4
171 191 203 1 167 188 209 2 119 149 196 3 158 168 175 4
138 204 213 1 183 206 210 2 144 169 164 3 175 217 235 4
 22  24  24 1  91 154 152 2 170 202 181 3 147 183 181 4
              115 133 136 2  93 122 145 3 105 107  92 4
               32  97  86 2 237 243 281 3 213 263 260 4
               38  37  40 2 208 235 249 3 258 248 257 4
               66 131 148 2 187 199 205 3 257 269 270 4
              210 221 251 2  95 102  96 3
              167 172 212 2  46  67  28 3
               23  18  30 2  95 137  99 3
              234 260 269 2  59  76 101 3
                          186 198 201 3
proc glm; /* equality of three means */
   class a;
   model y1-y3=a/nouni;
   means a;
   manova h=a/printe;
proc glm; /* equality of profiles */
   class a;
   model u1 u2=a/nouni;
   manova h=a;
proc glm;
   class a;
   model z=u1 u2 a;
   lsmeans a /stderr pdiff;
proc glm; /* equality of profile means */
   class b;
   model u1 u2=b /noint;
   manova h=b;
quit;
```

若轮廓图相等，语句LSMEANS A /STDERR PDIFF 产生水平和估计、水平间差异的均值等。以上程序也可以考虑带有将放射性处理前的分数作为协变量的情形，此处从略。

第一部分结果表明“均值相同”。

```
                          S=3    M=-0.5    N=18.5
Statistic                    Value          F      Num DF    Den DF  Pr > F
Wilks' Lambda             0.77325240     1.1772         9  95.06636  0.3185
```

```
Pillai's Trace             0.23285996     1.1501         9       123  0.3332
Hotelling-Lawley Trace     0.28534495     1.1942         9       113  0.3056
Roy's Greatest Root        0.25454432     3.4788         3        41  0.0243
```

第二部分结果：

```
                                  Sum of           Mean
Source                    DF        Squares         Square   F Value     Pr > F
Model                      5    78498.76606    15699.75321      4.02     0.0049
Error                     39   152150.00925     3901.28229
Corrected Total           44   230648.77531
Dependent Variable: Z
Source                    DF       Type I SS    Mean Square   F Value     Pr > F
U1                         1       482.59138      482.59138      0.12     0.7269
U2                         1     43597.82098    43597.82098     11.18     0.0018
A                          3     34418.35370    11472.78457      2.94     0.0449
Source                    DF     Type III SS    Mean Square   F Value     Pr > F
U1                         1      5730.40926     5730.40926      1.47     0.2328
U2                         1     52989.04377    52989.04377     13.58     0.0007
A                          3     34418.35370    11472.78457      2.94     0.0449
             A              Z       Std Err     Pr > |T|   LSMEAN
                       LSMEAN        LSMEAN   H0:LSMEAN=0   Number
             1    151.515482     25.582478        0.0001     1
             2    119.473451     16.772678        0.0001     2
             3    164.892380     16.273133        0.0001     3
             4    195.022642     19.945033        0.0001     4
                  Pr > |T| H0: LSMEAN(i)=LSMEAN(j)
                   i/j     1       2       3       4
                   1   .      0.3001  0.6629  0.1875
                   2  0.3001   .      0.0599  0.0063
                   3  0.6629  0.0599   .      0.2507
                   4  0.1875  0.0063  0.2507   .
```

第 3 、4 组间调整均值有所不同。　　最后的检验表明各均值不同。

```
                            S=1    M=0    N=20.5
Statistic                     Value          F      Num DF    Den DF  Pr > F
Wilks' Lambda              0.38631074    34.1547         2        43  0.0001
Pillai's Trace             0.61368926    34.1547         2        43  0.0001
Hotelling-Lawley Trace     1.58858969    34.1547         2        43  0.0001
Roy's Greatest Root        1.58858969    34.1547         2        43  0.0001
```

【例4.13】多元线性模型问题(multivariate general linear model)，PROC IML 程序如下：

```
/* Maindonald, J.H.(1984) Statistical Computations, Wiley */
```

```
libname user '.';
data mlm;
    x1=1;
    input x2-x4 y1 y2;
    cards;
     7  5  6   7  1
     2 -1  6  -5  4
     7  3  5   6 10
    -3  1  4   5  5
     2 -1  0   5 -2
     2  1  7  -2  4
    -3 -1  3   0 -6
     2  1  1   8  2
     2  1  4   3  0
proc iml;
    reset print;
    use mlm;
    read all into xy;
    x=xy[,1:4]; y=xy[,5:6];
    n=nrow(y);  q=ncol(x);
    beta=y'*x*inv(x'*x);
    sigma=1/(n-q)#(y'-beta*x')*y;
    z=xy'*xy; yy=css(z,n);
    t=half(z);
proc glm;
    model y1 y2=x1-x3;
    manova h=x1-x3 /printh printe;
run;
```

据多元线性模型理论，算得回归系数(BETA)：$\begin{pmatrix} 7.7333333 & -0.2 & 2.3333333 & -1.666667 \\ -1.633333 & 0.4 & 0.1666667 & 0.6666667 \end{pmatrix}$

SIGMA 是协方差矩阵：$\begin{pmatrix} 0.8 & 4 \\ 4 & 22 \end{pmatrix}$

YY 是6x6阶CSSP矩阵。$\begin{pmatrix} 9 & 18 & 9 & 36 & 27 & 18 \\ 18 & 136 & 58 & 92 & 94 & 96 \\ 9 & 58 & 41 & 52 & 67 & 50 \\ 36 & 92 & 52 & 188 & 68 & 112 \\ 27 & 94 & 67 & 68 & 237 & 70 \\ 18 & 96 & 50 & 112 & 70 & 202 \end{pmatrix}$

T 是CSSP矩阵的Cholesky分解，故其阶数也是6x6。$\left(\begin{array}{cccccc} 3 & 6 & 3 & 12 & 9 & 6 \\ 0 & 10 & 4 & 2 & 4 & 6 \\ 0 & 0 & 4 & 2 & 6 & 2 \\ 0 & 0 & 0 & 6 & -10 & 4 \\ \hline 0 & 0 & 0 & 0 & 2 & 10 \\ 0 & 0 & 0 & 0 & 0 & 3.1622777 \end{array}\right)$

据Maindonald, J.H. 多变量线性模型的许多统计量均可经T而得，矩阵最后两列应予特别注意，其第一行与其转置的乘积对应常数项的平方和，第二行开始则依次对应x1, x2, x3的平方和。顺序的含义是指x2 是调整了x1的平方和，x3 是调整x1, x2的平方和，右下角2x2矩阵的乘积则对应剩余平方和。这对于理解SAS的I 类平方和也是很有帮助的。PROC GLM回归的结果与分别进行一元计算时是相同的，列出部分计算结果，完整的结果可由运行上面程序而获。

Y1 与X1-X3的回归方程模型F=63.33, P=0.0002，R-平方=0.974359。

I 类平方和及参数估计值：

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	16.0000000	16.0000000	20.00	0.0066
X2	1	36.0000000	36.0000000	45.00	0.0011
X3	1	100.0000000	100.0000000	125.00	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	7.733333333	12.30	0.0001	0.62857864
X1	-0.200000000	-1.58	0.1747	0.12649111
X2	2.333333333	9.90	0.0002	0.23570226
X3	-1.666666667	-11.18	0.0001	0.14907120

Y2 与X1-X3的回归方程模型F=0.85, P=0.5240，R-平方=0.337349。I 类平方和及参数估计值：

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	36.00000000	36.00000000	1.64	0.2570
X2	1	4.00000000	4.00000000	0.18	0.6875
X3	1	16.00000000	16.00000000	0.73	0.4327

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-1.633333333	-0.50	0.6413	3.29629422
X1	0.400000000	0.60	0.5728	0.66332496
X2	0.166666667	0.13	0.8980	1.23603308
X3	0.666666667	0.85	0.4327	0.78173596

多元方差分析结果是根据每个x的假设矩阵与误差矩阵分析，做出其效应的判断。误差阵是：

$$\begin{array}{cc} 4 & 20 \\ 20 & 105.55555556 \end{array}$$

X1-X3 的假设矩阵是：$\begin{array}{cccccc} 2 & -4 & 78.4 & 5.6 & 100 & -40 \\ -4 & 8 & 5.6 & 0.4 & -40 & 16 \end{array}$

多变量检验，三个检验均有S=1, M=0, N=1，自由度2,4，使用Wilks' Lambda，其值在X1为：0.08849558, F=20.60, P=0.0078。X2为：0.00473844, F=420.08 ，P=0.0001。X3为：0.00314861，F=633.20, P=0.0001。

§4.4.3 分类数据分析

表格数据的描述可以采用PROC TABULATE，它的产出格式很灵活但不能得到列联表统计量。一些常用的列联表统计量可以由PROC FREQ 得到，这里主要介绍CATMOD 过程。SAS 过程CATMOD 是为数不多的用于分类数据以及分类与连续变量混合类型数据分析的优秀软件之一，可以拟合对数线性模型。

CATMOD 进行各种分类数据的分析的许多方法是连续数据分析方法的推广。如方差分析在传统的意义上是均值的分析，以及把均值间的变异按来源分隔。而这里的方差分析则是指反应函数的分析，以及对反应函数的变异分为不同的来源。

反应函数可以是因变量为有序时的平均分数，也可以是边缘概率，累积的logits 或其它函数。

CATMOD也是一个交互式过程。

语句格式及其说明如下：

```
PROC CATMOD DATA= ORDER=DATA;
  DIRECT 变量表;
  MODEL 反应=设计效应表/选项; /* 必选*/
  CONTRAST '标号' 对比的描述, 对比的描述,...;
  BY 变量表;
  FACTORS 因素描述,... / 选项;
  LOGLIN 效应/ 选项;
  POPULATION 变量表;
  REPEATED 因素描述,... / 选项;
  RESPONSE 函数/选项;
  RESTRICT 参数=取值<...参数=取值>;
  WEIGHT 变量;
RUN;
```

1. PROC 语句ORDER= DATA 指示变量水平的排序是按照输入数据的顺序；否则，变量水平根据内部排序如数据值的次序或字母顺序。

2. DIRECT 语句指示用作边续变量处理的数值变量名，应先于MODEL 语句指示。

3. MODEL 语句指定自变量和因变量以及模型效应。

   反应效应指示决定反应类别的因变量(隐含列联表的列)，反应效应可以是单一的变量或交互效应。设计效应是变异的来源，如主效应和交互。

MODEL 语句的选项中，(1)计算与打印方面的选项：CORRB(参数的相关阵) 、COV(每个总体的反应函数矩阵)、COVB(估计协方差阵)、FREQ(反应与总体的两维频数表)、ML(使用极大似估计)、ONEWAY(对于参与分析的每个变量产生一个频数分布) 、PREDICT打印每个总体的观测值与期望值，以及它们的标准误和残差。若反应函数是标准的广义logits，则PRED=FREQ计算和打印预测格子频数，而PRED= PROB 或PREDICT 则计算和打印相应的格子概率。PROB 打印反应与总体的两维交叉表，TITLE=”指示相应于MODEL语句的标题，XPX指示打印正规方程的交叉乘乘积矩阵。(2)节约计算与打印的选项：NODESIGN(不打印设计矩阵)、NOGLS(不使用广义加权最小二乘法)、NOINT(不打印截距项)、NOITER(不打印极大似然估计每步上的有关信息)、NOPARM(不打印估计参数及其是否为零的检验)、NOPROFILE(不打印总体和反应的分类情况)、NORESPONSE(不打印对数线性模型的_RESPONSE_矩阵)。(3)控制计算与打印细节的选项有：ADDCELL=(增加到每格子的数)、AVERAGED(指示自变量主效应在总体反应函数中取均值)、EPSILON=(参数极大似然估计的收敛精度) 、MAXITER=(极大似然法的最大迭代次数)。(4)MODEL语句可以直接接受输入的设计矩阵，如：

```
model r=(1  1  0  0, 1  1  0  0, 1  1  0  2,
         1  0  1  0, 1  0  1  1, 1  0  1  2,
         1 -1 -1  1, 1 -1 -1  1, 1 -1 -2 2)
         ( 1='Intercept', 2 3 ='Group Main Effect',
                    4='Linear Effect of Time');
```

CONTRAST 一定紧跟在MODEL 或LOGLIN 语句后，它能构造和检验由MODEL 的模型参数或LOGLIN 语句所列效应构成的线性函数。’标号’是必选的，最长24个字符，用于标记。每行的描述指示矩阵C的一行，C用于CB=0的假设检验，行描述之间用逗号分开。

在MODEL, POPULATION, 或WEIGHT 语句中出现的的任何变量若缺失，则该记录被忽略。

4. CONTRAST 语句

对于参数的线性函数进行检验，与GLM有所不同，因此在使用时应当谨慎。设变量A有四个水平，相应于四个参数$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ ，$\alpha_4 = -\alpha_1 - \alpha_2 - \alpha_3$。因此检验$\alpha_1 = \alpha_4$就相当于$2\alpha_1 + \alpha_2 + \alpha_3 = 0$。相应的CONTRAST 语句就是：CONTRAST ’1 vs. 4’ a 2 1 1; 所有的效应可以用关键字ALL_PARMS代替，过程认为效应参数与设计矩阵的列数相同。这在直接输入设计矩阵的情况下是很有用的，如：

```
model y=(1 0 0 0, 1 0 1 0, 1 1 0 0, 1 1 1 1);
contrast 'Main Effect of B' all_parms 0 1 0 0;
contrast 'Main Effect of C' all_parms 0 0 1 0;
contrast 'B*C Interaction ' all_parms 0 0 0 1;
```

5. FACTORS 语句

通过指示因素名和水平数区分同一个总体中不同的反应函数，若因素名是字符型的，则加$后缀。其斜线后的选项有三个，即PROFILE、_RESPONSE_和TITLE。PROFILE指

示每个反应函数的因素的取值。设一个数据集含有十个函数及其协方差矩阵的估计值，关联自变量a,b对这些函数进行分析。

```
proc catmod;
   response read b1-b10;
   model _f_=_response_;
   factors a $ 2, b $ 5 /_response_=a b;
quit;
```

6. LOGLIN 语句

定义对数线性模型，它与MODEL中的_RESPONSE_是对应的。它可以在斜线后用TITLE=”选项标识进行的分析。

7. POPULATION 语句

指示总体根据指定变量的交叉情况形成，否则是用MODEL语句形成。因此，直接输入设计矩阵时，必须使用POPULATION语句。POPULATION 的第二个用途是当模型一些项需要约化时，仍保持原来的总体。

8. REPEATED 语句

用于处理重复测量因素，当多于一个自变量和MODEL语句中出现_RESPONSE_ 的情形。其选项与FACTORS类似。

9. RESPONSE 语句

指示反应概率的函数，若不加指示，CATMOD隐含地使用广义logits。函数的指示可为CLOGIT|CLOGITS, JOINT, LOGIT|LOGITS, MARGINAL| MARGINALS, MEAN|MEANS, READ 变量。

RESPONSE 的选项有OUT=, OUTEST=, 指示输出数据集和TITLE=”指示标题。下面分析因变量r1, r2和自变量a, b的关系，使用边缘概率和主效应模型：

```
proc catmod;
   weight wt;
   response marginals;
   model r1*r2=a b;
quit;
```

用对数线性模型分析因变量r1,r2,r3，模型包括主效应和r1*r2的交互：

```
proc catmod;
   weight wt;
   model r1*r2*r3=_response_ /ml nogls pred=freq;
   loglin r1|r2 r3;
quit;
```

极大似然法进行顺变量r与自变量x1,x2的logistic模型分析：

```
proc catmod;
   weight wt;
   direct x1 x2;
   model r1=x1 x2/ml nogls;
quit;
```

因变量r1,r2,r3表示三个不同时间的同一类测量，分析因变量、时间及自变量a的关系，用重复测量分析：

```
proc catmod;
   weight wt;
   response marginals;
   model r1*r2*r3=_response_ a;
   repeated time 3 /_response_=time;
quit;
```

分析因变量r与自变量a,b的关系，使用方差分析：

```
proc catmod;
   weight wt;
   response mean;
   model r=a|b;
quit;
```

因变量r1,r2与自变量x1,x2的关系，使用线性回归分析因变量的边缘概率：

```
proc catmod;
   weight wt;
   direct x1 x2;
   response marginals;
   model r1*r2=x1 x2;
quit;
```

分析有序分类变量r及自变量a的关系，使用累积logit考虑因变量的特性：

```
proc catmod;
   weight wt;
   response clogits;
   model r=_response_ a;
quit;
```

【例4.14】下面给出重复测量分析的两个例子。在分类数据分析中，边缘是一个未有调整其它量的合计，详见A. Agresti (1990) 的讨论。

第一个是SAS/STAT PROC CATMOD 的一个样本程序。检查7477 名30-39岁的妇女左右眼视力，使用重复测量因子SIDE 的主效应检验边缘的一致性(MARGINAL HOMOGENEITY)。

由于有四个水平，RESPONSE 语句对每个反应变量进行三个边缘概率，则分析共有六个反应函数。重复测量因子SIDE 有LEFT 和RIGHT 两个水平，CATMOD 把这些函数分成三组，有三个自由度，即寻每上边缘概率有一个自由度，因而进行边缘一致性检验是合适的。

```
title 'VISION SYMMETRY';
data vision;
   input right left count @@;
cards;
1 1 1520    1 2  266    1 3  124    1 4  66
2 1  234    2 2 1512    2 3  432    2 4  78
3 1  117    3 2  362    3 3 1772    3 4 205
4 1   36    4 2   82    4 3  179    4 4 492
proc catmod;
   weight count;
   response marginals;
   model right*left=_response_ / freq;
   repeated side 2;
   title2 'TEST OF MARGINAL HOMOGENEITY';
quit;
```

结果如下：

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	3	78744.17	0.0000
SIDE	3	11.98	0.0075
RESIDUAL	0	.	.

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	0.2597	0.00468	3073.03	0.0000
	2	0.2995	0.00464	4160.17	0.0000
	3	0.3319	0.00483	4725.25	0.0000
SIDE	4	0.00461	0.00194	5.65	0.0174
	5	0.00227	0.00255	0.80	0.3726
	6	-0.00341	0.00252	1.83	0.1757

方差分析表显示，SIDE效应是显著的，即左右眼之间不存在边比一致，或者说，受试者两眼视力的分布差别有显著意义。

【例4.15】下面也是SAS/STAT 的样本程序。在一个随访研究中，两种不同诊断(mild, severe)的病人接受两种治疗(std, new)，在三个不同时间(1,2,3周) 测量受试者对治疗的反应(n=正常, a=不正常) 。分析的目的是评价重复测量因子(TIME) 及因变量诊断(DIAG)和治疗(TRTMENT)的

效果。RESPONSE语句用于计算边缘概率的对数比数比(logits)，设计矩阵中使用的时间值(0,1,2)与实际值(1,2,4)的以2为底的对数相应。共有四个POPULATION (2 诊断x 2 治疗)，八个反应(2 WEEK1 x 2 WEEK2 x 2 WEEK3)。

```
title 'GROWTH CURVE ANALYSIS';
data growth2;
   input diag $ trt $ week1 $ week2 $ week4 $ count @@;
cards;
mild std n n n 16    severe std n n n  2
mild std n n a 13    severe std n n a  2
mild std n a n  9    severe std n a n  8
mild std n a a  3    severe std n a a  9
mild std a n n 14    severe std a n n  9
mild std a n a  4    severe std a n a 15
mild std a a n 15    severe std a a n 27
mild std a a a  6    severe std a a a 28
mild new n n n 31    severe new n n n  7
mild new n n a  0    severe new n n a  2
mild new n a n  6    severe new n a n  5
mild new n a a  0    severe new n a a  2
mild new a n n 22    severe new a n n 31
mild new a n a  2    severe new a n a  5
mild new a a n  9    severe new a a n 32
mild new a a a  0    severe new a a a  6
proc catmod order=data;
   title2 'REDUCED LOGISTIC MODEL';
   weight count;
   population diag trt;
   response logit;
   model week1*week2*week4=(1 0 0 0 ,1 0 1 0 ,
                            1 0 2 0 ,1 0 0 0 ,
                            1 0 0 1 ,1 0 0 2 ,
                            0 1 0 0 ,0 1 1 0 ,
                            0 1 2 0 ,0 1 0 0 ,
                            0 1 0 1 ,0 1 0 2 )
                               (1='Mild diagnosis, week 1',
                                2='Severe diagnosis, week 1',
                                3='Time effect for std trt',
                                4='Time effect for new trt') /freq;
   contrast 'Diagnosis effect, week 1' all_parms 1 -1  0  0;
   contrast 'Equal time effects' all_parms 0  0  1 -1;
quit;
```

结果如下：

ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
Mild diagnosis, week 1	1	0.28	0.5955
Severe diagnosis, week 1	1	100.48	0.0000
Time effect for std trt	1	26.35	0.0000
Time effect for new trt	1	125.09	0.0000
RESIDUAL	8	4.20	0.8387

ANALYSIS OF WEIGHTED-LEAST-SQUARES ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
MODEL	1	-0.0716	0.1348	0.28	0.5955
	2	-1.3529	0.1350	100.48	0.0000
	3	0.4944	0.0963	26.35	0.0000
	4	1.4552	0.1301	125.09	0.0000

ANALYSIS OF CONTRASTS

Contrast	DF	Chi-Square	Prob
Diagnosis effect, week 1	1	77.02	0.0000
Equal time effects	1	59.12	0.0000

程序用指定的设计矩阵进行分析，分析程序使用了POPULATION 语句指示分类量的组合，方差分析表显示，这批数据用给定的参数能很好地表达。对比分析表明，第一周的诊断效应高度显著，由于重症患者logit 的估计(参数2)较轻者更小，表明这些患者第一周出现异常反应的概率要大。同时也显示，标准疗法的时间效应与新疗法不同；参数表显示，新法的时间效应比标准疗法要强得多。

在CATMOD过程省略POPULATION 语句时，得到的将是一个合并了MODEL 所指示的分类效应以外所有效应的边缘结果。

§4.4.4 LOGISTIC 回归分析

【例4.16】Hosmer, D.W. and S. Lemeshow (1989) 中假想的资料，四种民族共100 人患冠心病的情况。民族(RACE)有：黑人(black)、西班牙人(hispanic)、白人(white) 及其他(other)；冠心病(STATUS)有有(present)、无(absent)两种。

STATUS 与RACE 交叉表

STATUS / Frequency / Percent	black	hispanic	other	white	Total
absent	10	10	10	20	50
	10.00	10.00	10.00	20.00	50.00

```
---------+--------+--------+--------+--------+
present  |     20 |     15 |     10 |      5 |     50
         |  20.00 |  15.00 |  10.00 |   5.00 |  50.00
---------+--------+--------+--------+--------+
Total          30        25        20        25      100
              30.00     25.00     20.00     25.00   100.00
```

关于白人的比数比是8.0, 6.0, 4.0，如对于西班牙人(15x20)/(5x10)=6.0。SAS 的分析程序如下：

```
/* Applied logistic regression */
 ** David W. Hosmer & Stanley Lemeshow (1989). Wiley;
 options ps=60;
 data hosmer;
 format race $8.;
 do status='present','absent';
   do race='black','hispanic','ords;
 20 15 10  5
 10 10 10 20
 proc freq;
    weight count;
    table status*race/chisq;
 run;
 proc catmod;
    weight count;
    response clogit;
    model status=race/ml nogls;
 quit;
```

FREQ 过程产出了结局和民族(RACE) 有关的卡方统计量：

统计量	自由度	卡方值	P
卡方	3	13.333	0.004
似然比卡方	3	14.042	0.003
Mantel-Haenszel 卡方	1	11.821	0.001
Φ 系数		0.365	
列联表系数		0.343	
Cramer's V		0.365	

CATMOD 过程使用极大似然估计法估计模型参数，四次迭代后似然值为124. 58744，收敛精度为2.197E-11。参数迭代初值为0，四迭代后分别为-0.0719、0. 7651、0.4774 和0.0719。极大似然估计卡方检验：模型截距(INTERCEPT) 卡方=0. 11, P =0.7425，民族(RACE) 卡方=11.77，P=0.0082。

效应	参数	估计值	标准误	卡方	P 值
INTERCEPT	1	-0.0719	0.2189	0.11	0.7425
RACE	2	0.7651	0.3506	4.76	0.0291
	3	0.4774	0.3623	1.74	0.1876
	4	0.0719	0.3846	0.03	0.8517

这里指出的是，SAS 采用最高一组做为对照。

【例4.16】此处用例是较早出现的LOGISTIC 分析软件LOGRESS 所引用的Framingham Heart Study 的数据。该数据是一个成组的logistic 资料，描述年龄(age)、性别(sex)、糖尿病(diabetes mellitus)、性别和糖尿病交互影响对冠心病发病的影响，下面是它的分析结果：

```
DEPENDENT VARIABLE:  CORONARY HEART DISEASE (0=NO,  1=YES)
TOTAL POPULATION:   26746     NUMBER OF CASES=   498
INDEPENDENT VARIABLE COEFFICIENT STD. ERROR     Z
AGE IN YEARS .0907991 .009451     9.61
SEX (0=FEMALE, 1=MALE) .9688146 .098534     9.83
DIABETES MELLITUS (0=NO, 1=YES)        1.1267654 .275625     4.09
SEX*DIABETES (INTERACTION TERM)        -.7464846 .366532    -2.04
CONSTANT       -9.5404578 .539102    *****
LIKELIHOOD RATIO STATISTIC ( 4) D.F.:  213.4489
```

给出估计的logistic 系数/估计标准误/ Wald 检验，即系数除以标准误，看特定的系数是否为零实际也就是看协变量与二分类结果之间有否显著的关联，可按Z 值查标准正态统计量表。如年龄的Z 值是9.61，指示年龄与冠心病发生之间有显著的关联。上表也给出了检验是否所有系数皆零的似然比检验LRT，该统计量服从卡方分而，自由度与模型中的因变量数目相同。它有两个用途：其一是看该值是否显著，表明至少一个系数非零；其二是判断一个因变量追加到模型时，是否有意义，因而可用于变量的筛选。

95 % 可信限(COEFFICIENTS AND 95% CONFIDENCE INTERVALS)：

```
DEPENDENT VARIABLE COEFFICIENT  LOWER   UPPER
AGE IN YEARS   .0908   .0723    .1093
SEX (0=FEMALE, 1=MALE)   .9688   .7757   1.1619
DIABETES MELLITUS (0=NO, 1=YES)   1.1268   .5865   1.6670
SEX*DIABETES (INTERACTION TERM)   -.7465        -1.4649   -.0281
CONSTANT -9.5405       -10.5971  -8.4838
```

比数比及相应的可信限(ODDS RATIOS AND 95% CONFIDENCE INTERVALS)：

```
DEPENDENT VARIABLE ODDS RATIO  LOWER    UPPER
AGE IN YEARS   1.0950  1.0750    1.1155
SEX (0=FEMALE, 1=MALE)   2.6348  2.1721    3.1961
DIABETES MELLITUS  (0=NO, 1=YES)   3.0857  1.7978    5.2962
SEX*DIABETES  (INTERACTION TERM)    .4740   .2311     .9723
```

若p1 是某人患病的概率，q1 = 1 - p1，则p1/q1 是他发病的比数。设另一个人有类似的定义p2和q2，则两个人发病的比数是比数比(odds ratio) (p1* q2)/(p2*q1)。应用于logistic 函数，第一种情况就是下述方程：

Odds Ratio = exp [b1 *(x11-x12) + ... + bp * (x1p - x2p)]

若两人其他方面相同，只有一个特征不同，则相诮的公式就是：

Odds Ratio = exp [bi * (xi1-xi2)]

当研究的特征仅仅取两个值0 和1, 则有：Odds Ratio = exp [β]

Beta 表示两人不相同特征的系数值，利用它来计算比数比。若感兴趣的变量是二分类的(通常是0 和1)，比数比可理解作仅一个结果变动造成的影响。本例吸烟的比数比是1.4。若变量是计量的，比数比可理解作一个单位变动时的相对比数，本例中的年龄用年数表示，比数比是1.10，表明每一年将增加10% 的发病危险。比数比的上下可信限算式如下：$L = \exp(\hat{\beta} - 1.96 \times se(\hat{\beta}))$, U=$\exp(\hat{\beta} + 1.96 \times se(\hat{\beta}))$, PROC CATMOD 程序和结果：

```
data logress;
  input age sex DM SD CHD freq @@;
  label age = 'AGE IN YEARS'
        sex = 'SEX (0=FEMALE, 1=MALE)'
        DM  = 'DIABETES MELLITUS (0=NO, 1=YES)'
        SD  = 'SEX*DIABETES  (INTERACTION TERM) '
        CHD = 'CORONARY HEART DISEASE (0=NO, 1=YES) '
        freq= 'NUMBER OF OBSERVATIONS';
cards;
50 1 0 0 0 6434  50 0 0 0 0 8519
50 1 0 0 1  124  50 0 0 0 1   45
50 1 1 1 0  193  50 0 1 0 0  159
50 1 1 1 1    6  50 0 1 0 1    5
60 1 0 0 0 4298  60 0 0 0 0 6199
60 1 0 0 1  179  60 0 0 0 1  116
60 1 1 1 0  218  60 0 1 0 0  228
60 1 1 1 1   13  60 0 1 0 1   10
proc fsprint;run;
proc catmod data=logress;
   weight freq;
   response clogit;
   direct age sex DM SD;
   model CHD=age sex DM SD/ML NOGLS NOITER;
quit;
```

CATMOD 过程的RESPONSE 有几种情况，默认的是RESPONSE LOGIT，但是对于正常编码的二分类数据的分析，程序得到的回归系数与其它程序符号相反，故指定RESPONSE CLOGIT; 同时也应使用极大似然方法估计(ML)，这要抑制广义最小二乘法的实施(NOGLS)。上面程序保持与LOGRESS.EXE 结果的一致，使用DIRECT AGE SEX DM SD 语句使用这些变量以连续变量的形式参与计算。程序运算结果包括：一些说明、总体反应的轮廓、极大似然估计方差分析表以及各参数、标准误的估计值。

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-9.5405	0.5392	313.11	0.0000
AGE	2	0.0908	0.00945	92.27	0.0000
SEX	3	0.9688	0.0985	96.67	0.0000
DM	4	1.1268	0.2758	16.69	0.0000
SD	5	-0.7465	0.3672	4.13	0.0420

在PC SAS 6.04 中，也可以使用LOGISTIC 过程进行分析，它使用极大似然法对拟合线性logistic 回归模型，同时提供几种模型选择方法对自变进行筛选。对二分类变量的模型产出加归诊断情况也是可能的，在logistic 模型中的logit 链接函可以换成normit 函数或complementary log-log 函数。

二分类资料如成功、失败以及有序反应资料如无、轻微、严重在许多研究中产生。Logistic 回归分析常用于研究反应概率与自变量的关系，最主要的是二分类反应模型，有序反应模型的最简模型是关于某些选定尺子下平行线的构造，对于对数比数比(log-odds)尺度，平行线回归模型常称做比例比数比模型(proportion -al odds model)，LOGISTIC 过程的语法是：

PROC LOGISTIC 过程选项;
MODEL 反应量=自变量/ 选项; /* 必选*/
WEIGHT 变量表;
FREQ 变量表;
OUTPUT ¡OUT= SAS 数据集名¿ ¡关键字= 名...关键字=名¿ ¡/ALPHA=值¿ ;
BY 变量表;

对二分类资料，模型中使用INFLUENCE 选项可以指示Pregibon (1981) 的回归诊断。数据集OUTEST= 包含有回归系数的估计值，若指定COVOUT 选项，该数据集也包含了估计参数的协方差矩阵。数据集对每个截距参数以及MODEL 语句中的每一个自变量有一个变量，第一个记录是回归系数的极大似然估计值，若指定COVOUT= 选项，数据集还包含了估计协方差矩阵的各行。

使用LOGISTIC 过程进行上例的分析，程序如下：

```
proc logistic data=logress;
   weight freq;
   model CHD=age sex DM SD;
quit;
```

其输出结果同样包括了一些说明、反应的轮廓、因变量的均值、标准差、最大最小值、评价模型拟合好坏的几种准则，以及实际结果和模型结果的比较。利用过程输出的结果，可以对数据进行检查。

所估计的系数符号是与LOGRESS 和CATMOD 不同，可以使用PROC 选项DESCEND 进行调整。

§4.4.5 生存分析

LIFEREG 过程对失效时间拟合参数模型，失效可以是左截、右截或区间截尾。反应变量的模型包括一个协变量和随机项组成的线性效应。

随机项的分布可以由极值分布、正态分布、logistic分布、及使用对数转换后所对应的指数、威布尔、对数正态、对数logistic和伽马分布。

语句格式及说明如下：

```
PROC LIFEREG DATA= COVOUT NOPRINT ORDER= OUTEST= ;
标号: MODEL 反应=变量/ 选项; /* 必选*/
    CLASS 变量表;
      WEIGHT 变量;
    OUTPUT OUT= 选项;
      BY 变量表;
```

1. PROC 语句

   OUTEST= 有以下变量： _MODEL_ 长度为8的模型标号，MODEL 语句不指定时为空。_NAME_ 长度为8的因变量名。_TYPE_ 记录的类型，参数为PARM，协方差阵为COV。_DIST_ 长度为8分布名。_LNLIKE_ 对数似然值。INTERCEP 模型常数项和协方差。_SCALE_ 尺度参数及其协方差。_SHAPE1_ 形状参数和协方差。数据集在模型中出现CLASS 语句时不产生。COVOUT 指示OUTEST=数据集含有估计协方差阵和参数值。

2. MODEL 语句

   MODEL 语句指示模型回归部分所使用的变量，反应变量的分布，以及与每个记录的截尾类型。其格式如下：

   标号: MODEL (下界,上界)=变量表/ 选项; /* 必选*/

   标号: MODEL 变量¡*截尾(数据表)¿=变量表/ 选项;

   标号: MODEL 事件/ 试验=变量/ 选项;

   MODEL 语句选项：

   DISTRIBUTION= DISTRIBUTION—DIST—D= 指定分布类型，有效的分布类型为：

   WEIBULL Weibull 分布(也是默认分布)

   EXPONENTIAL 指数分布

   LNORMAL 对数正态分布

   LLOGISTIC 对数logistic分布

   GAMMA 伽马分布

   NORMAL 正态分布

   LOGISTIC logistic 分布

   NOLOG 指示对反应变量不取对数。COVB 输出观察信息矩阵的逆。CORRB 输出参数间的相关阵。NOINT/INTERCEPT=指示常数项固定或给出其初值。NOSCALE /

SCALE= 指示尺度参数固定或给出其初值。NOSHAPE1/SHAPE1= 指示形状参数固定或给出其初值。当回归发生困难时，可以用INITIAL=指示参数的初值。MAXIT= 指示最大迭代次数。ITPRINT 指示详细的迭代过程，最终的梯度、海森阵。CONVERGE= 指示收敛准则，即每步上参数变动值小于此值时收敛。默认值为0.001，当参数大于0.01时则是一个相对的变动准则。SINGULAR=指示信息矩阵奇异的准则，默认为1E-12。

3. CLASS 语句指示变量为离散变量，然而只能指示主效应模型。

4. OUTPUT 语句关键字有Q，CONTROL，P，XBETA，STD，SURVIVAL，CENSORED。OUT=的数据集中包含输入数据集的变量和_PROB_即分位点估计的概率值。

现对例6.3 白血病数据进行分析。

```
data life;
input group time ind @@;
cards;
1 6   1  1 17  1    2  1  0  2  8  0
1 6   0  1 19  1    2  1  0  2  8  0
1 6   0  1 20  1    2  2  0  2  11 0
1 6   0  1 22  0    2  2  0  2  11 0
1 7   0  1 23  0    2  3  0  2  12 0
1 9   1  1 25  1    2  4  0  2  12 0
1 10  1  1 32  1    2  4  0  2  15 0
1 10  0  1 32  1    2  5  0  2  17 0
1 11  1  1 34  1    2  5  0  2  22 0
1 13  0  1 35  1    2  8  0  2  23 0
1 16  0             2  8  0
proc lifereg;
     class group;
     model time*ind(1)=group/dist=weibull;
run;
```

结果如下：

```
     Data Set          =WORK.LIFE
     Dependent Variable=Log(TIME)
     Censoring Variable=IND
     Censoring Value(s)=     1
     Noncensored Values=    30  Right Censored Values=     12
     Left Censored Values=   0  Interval Censored Values=   0
     Log Likelihood for WEIBULL -47.06410176
                      L I F E R E G  P R O C E D U R E
Variable  DF   Estimate  Std Err ChiSquare  Pr>Chi Label/Value
INTERCPT   1 2.24835236 0.165972  183.5102  0.0001 Intercept
GROUP      1                      16.64439  0.0001
```

```
           1 1.26733459  0.31064  16.64439  0.0001            1
           0          0        0         .    .               2
SCALE      1  0.7321944 0.107846                   Extreme value scale paramet
```

LIFETEST 过程用右截尾的数据计算生存函数的非参估计，各层间生存分布相同的检验，计算反应变量和其它变量关联的秩统计量。

一个记录的失效时间或截尾变量有缺失值时，该记录不用于分析。除非指示MISSING选项，STRATA 变量为缺失值的记录不能于计算。若TEST 语句中的变量具有缺失值，则它对应的记录不用于计算秩统计量。

语句格式及说明如下：

PROC LIFETEST 过程选项;
  TIME 变量<*截尾(数值列表)>; /* 必选*/
  STRATA 变量<(数值列表)><...变量<(数值列表)>>;
  TEST 变量表;
  ID 变量表;
  FREQ 变量;
  BY 变量表;

1. PROC 语句在默认情况下或指定METHOD=PL—KM时，为积限估计，否则是寿命表方法。NOTABLE 指示不打印生存函数估计，因而仅输出图示和检验结果。MISSING 指示缺失值在参加分组时为有效。PLOTS 指示哪种估计量与时间的图示，如：S、LS、LLS、H和P。INTERVALS=/NINTE示寿命表的时间分组界值/区间数/区间宽度。ALPHA= 指示的是0.0001-0.9999之间的数，默认为0.05 ，设定可信区间概率水平。若指示了GRAPHICS 选项，则由图形设备输出PLOT指定的图，ANNOTATE=则指定附注数据集，对于BY指示的不同的组，可用LANNOTATE= 数据集来标记。
2. TIME 语句用于指示失效的时间变量，以及一个可选的截尾变量，必须是数值型。
3. STRATA 语句定义分层的变量名和分层水平数据。
4. ID 变量指示标识变量名，标记各记录的积限生存函数估计量。
5. FREQ 语句指示每个记录重复的次数。

```
proc lifetest plots=(s,ls,lls);
     time time*ind(1);
     strata group;
run;
```

程序输出分组的积限生存估计、生存函数、对数生存函数、双对数生存函数估计等结果。数据大致情况如下：

GROUP	总数	失效	截尾	%截尾
1	21	9	12	57.1429
2	21	21	0	0.0000
Total	42	30	12	28.5714

两组生存率差异的检验：

Test	Chi-Square	DF	Pr $>$Chi-Square
Log-Rank	16.7929	1	0.0001
Wilcoxon	13.4579	1	0.0002
-2Log(LR)	16.4852	1	0.0001

PROC PHREG 过程语句格式：

PRCO PHREG 选项;

MODEL 反应<*截尾(数据列表)>=变量选项;

其它程序语句;

STRATA 变量<(列表)><...变量<(列表)>></选项>;

标号:TEST 方程1<,...,方程k>< /选项>;

FREQ 变量;

ID 变量;

OUTPUT <OUT=SAS数据集><关键字=命名变量...关键字=命名变量>< /选项>;

BASELINE <OUT=SAS数据集><COVARIATES=SAS数据集><关键字=命名变量..
.关键字=命名变量>< /选项>;

BY 变量;

只有MODEL语句是必选的，括号(¡¿)内的项目是可选的。MODEL语句指示哪些变量是时间变量，哪些变量是截尾变量，哪些变量是解释变量。STRATA语句指示分层分析，TEST语句仍然是关于模型参数线性函数的检验。ID语句指示用于标识输出数据集的变量名。OUTPUT与BASELINE语句产生包含生存估计的数据集。DATA步语句可以用来产生时间协变量。

1. PROC PHREG 语句

DATA=指示分析的数据集。MULTIPASS 选项要求在每步牛顿—拉弗森迭代重新计算程序语句所定义的变量值，在模型含有时变协变量时有用。NOPRINT 选项不输出结果，NOSUMMARY不输出截尾和失效的频数。SIMPLE 打印协变量的简单统计量。

OUTEST=指示存放估计参数的数据集，此时COVOUT选项可输出参数的协方差阵。内容包括BY变量、_TIES_、_TYPE_、_NAME_、_LNLIKE_即重复失效的处理方法(BRESLOW,DISCRETE,EFRON,
估计量类型(PARMS,COV)、名称和对数似然值。

2. MODEL 语句

(1)重复失效的处理方法：TIES=BRESLOW, DISCRETE, EFRON, EXACT。

(2)模型指示方法：BEST=n与SELECTION=SCORE共用，指示打印具有最高的计分χ^2统计量的n个模型。NOFIT进行总的计分检验。SELECTION=指示模型选择方法，如：BACKWARD—B(向后), FORWARD—F(向前),NONE—N(不筛选), SCORE(最优子集)，STEPWISE—S(逐步法)。

(3)模型建立：DETAILS(详细输出每步结果)、INCLUDE=n(MODEL语句中的前n个变量进入模型)、MAXSTEP=n(指示逐步法最多的步数)、SEQUENTIAL(强迫以MODEL语句的顺序挑选变量)、SLENTRY—SLE=(指示进入模型的概率值)、SLSTAY—SLS=(指示删除的概率值)。START=n(从前n个变量开始筛选)、STOP=n(指示模型中最终的变量数)、STOPRES—SR(指示变量的增删是根据未选入模型变量的联合似然比检验的显著性)。

(4)牛顿—拉弗森迭代：CONVERGE=值(收敛准则，默认10^{-6}) 。CONVERGEPARM=值(参数收敛的准则)、MAXITER= n(最大迭代次数，默认25) 、SINGULAR=值(协变量间线性相关的奇异性准则，默认为10^{-12})。

(5)打印：ALPHA=值指示条件风险比的显著性水平，与RISKLIMITS—RL(参数取自然指数)联用有效。CORRB与COVB打印参数的相关阵和协方差阵。ITPRINT 打印迭代过程。

3. 可编程语句

包括ABORT, ARRAY, 赋值语句, CALL, DO, DO/END, GOTO, IF- THEN/ELSE, LINK/RETURN, SELECT, SUM等语句。

4. STRATA 语句

定义分层，如语句STRATA AGE (5,10 TO 40 BY 10) SEX;定义了<5, 5-,10-, 20-,30-及性别的交叉共12层。MISSING选项指示缺失值参与分层有效。

5. TEST 语句

用例：proc phreg; model time=a1 a2 a3 a4; test1:test a1,a2; test2:test a1=a2=a3; run;

PRINT选项打印中间计算结果。

6. OUTPUT 语句OUT=(输出数据集)、LOGLOG(SURVIVAL的重对数)、LOGSURV(SURVIVAL的对数)、NUM_LEFT(处于危险的对象数) 、RESDEV(离真度残差) 、RESMART(martingale残差)、STDXBETA(线性预测因子的标准误)、SURVIVAL(生存函数估计)、XBETA(线性预测因子)。选项ORDER=DATA—SORTED指示OUTPUT 数据集中记录的顺序。

7. BASELINE 语句关键字LOGLOGS, LOGSURV, STDXBETA,SURVIVAL, XBETA 与OUTPUT语句情形类似。选项NOMEAN不包括相应于协变量样本均值的生存函数估计量。OUT=指示生成的数据集，COVARIATES=指示含有协变量的数据集。

【例4.17】SAS 样本程序库PHRE0 的数据集RATS 来自Kalbfleisch and Prentice(1980)，两组大白鼠接受不同的预处理(GROUP)，然后接触一种致癌因子，鼠从接触到死于阴道癌的生存天数为DAYS，由于四只鼠死于其它原因而出现截尾，变量STATUS 是截尾指示变量(0=截尾；1=未截尾)，现比较两组生存曲线是否相同。

```
DATA rats;
   label days  ='Days from Exposure to Death';
   input days status group @@;
   cards;
143 1 0   164 1 0   188 1 0   188 1 0
```

```
190 1 0   192 1 0   206 1 0   209 1 0
213 1 0   216 1 0   220 1 0   227 1 0
230 1 0   234 1 0   246 1 0   265 1 0
304 1 0   216 0 0   244 0 0   142 1 1
156 1 1   163 1 1   198 1 1   205 1 1
232 1 1   232 1 1   233 1 1   233 1 1
233 1 1   233 1 1   239 1 1   240 1 1
261 1 1   280 1 1   280 1 1   296 1 1
296 1 1   323 1 1   204 0 1   344 0 1
proc phreg data=rats;
   model days*status(0)=group;
run;
```

比较两组预处理的比例风险模型是

$$h(t) = \begin{cases} h0(t) & \text{if GROUP=0} \\ h0(t)exp(b1) & \text{if GROUP=1} \end{cases}$$

风险比值是exp(b1), 并不依赖于时间，若风险比值随时间而变，则比例风险模型不成立，数据对比例风险模型简单的变化是如下依赖于时间的变量x=x(t):

$$x(t) = \begin{cases} 0 & \text{if GROUP=0} \\ log(t) & \text{if GROUP=1} \end{cases}$$

模型为h(t)=h0(t) exp[b1 GROUP + b2 x]，x=LOG(T)。风险比值成为(exp (b1) t ** b2), b2 是时间协变量x 的回归参数，其符号的正负表示风险比随时间增减的趋势。

分析的程序如下，MODEL 语句包括了变量X，它由模型中的编程语句它义，在每个发生事件的时刻，危险集中的对象的X 值都相应地变动。

```
proc phreg data=rats;
   model days*status(0)=group x;
   x=group*(log(days));
run;
```

程序输出截尾比例为4/40 x 100 % = 10.00 %

H0: BETA=0 回归系数为零的检验

```
                Without      With
   Criterion  Covariates  Covariates    Model Chi-Square
   -2 LOG L      204.317     201.438     2.878 with 1 DF (p=0.0898)
   Score               .           .     3.000 with 1 DF (p=0.0833)
   Wald                .           .     2.925 with 1 DF (p=0.0872)
```

极大似然估计(MLE) 分析

```
               Parameter    Standard     Wald          Pr >        Risk
Variable  DF    Estimate      Error   Chi-Square    Chi-Square    Ratio
GROUP      1   -0.595896     0.34840     2.92532        0.0872    0.551
```

后一部分的结果如下：模型的似然比检验、计分检验、Wald 检验统计量值分别为2.890 (0.2353)，3.051 (0.2176)，2.965(0.2271)。

```
                Without        With
Criterion     Covariates    Covariates    Model Chi-Square
-2 LOG L         204.317       201.423    2.894 with 2 DF (p=0.2353)
Score               .             .       3.051 with 2 DF (p=0.2176)
Wald                .             .       2.965 with 2 DF (p=0.2271)
                Parameter   Standard     Wald        Pr >       Risk
Variable  DF     Estimate     Error   Chi-Square  Chi-Square   Ratio
GROUP      1     0.639657    9.82972    0.00423      0.9481    1.896
X          1    -0.229521    1.82489    0.01582      0.8999    0.795
```

两个生存曲线的比较，基本上同log-rank(Mantel-Haenszel)，事实上若生存时间无重复，似然比检验与log-rank 检验相同。但Cox 模型能够调整其它变量的影响。

§4.4.6 主成分分析

PRINCOMP 过程用于主成分分析。语句格式及说明如下：

PROC PRINCOMP options;
VAR 变量;
PARTIAL 变量;
FREQ 某个变量;
WEIGHT 某个变量;
BY 分类变量;

1. PROC 语句各选项含义解释如下：

DATA= 指出被分析的SAS数据集的名称，若缺省，则使用最新创建的SAS 数据集，该数据集可以是原始数据集，也可以是TYPE=CORR, COV, EST, SSCP, UCORR 或UCOV 的数据集。

OUT= 输出一个数据集，它包含原始数据和主成分得分数据，但当DATA= 的数据集为特殊结构的数据集(TYPE=CORR 或COV 或SSCP), 则不能生成OUT=的数据集。

OUTSTAT= 可产生一个新的数据集，它可以包含均值，标准差，观测个数，相关阵或协差阵(若规定选择项)，特征根和特征向量等等，详细内容可参见后面的内容。

N=k 用此选择项，用户可以自己确定所需主成分的个数，例如: PROC PRINCO MP DATA=a N=4; 语句说明，对数据集a作主成分分析，并取前4个主成分，计算机在输出输出时，只打印4个主成分。若缺省，则主成分的个数为变量的个数。

PREFIX= 规定主成分名字的前缀，若缺省，则主成分的前缀为PRIN，并用PRIN1，PRIN2，...,PRINK 来表示主成分的名字，若规定PREFIX=Z，则主成分的名字为Z1,Z2,... 。

VARDEF= 规定用计算方差和协方差的除数，当输入数据集为TYPE=SSC P时，该项选择是必须的，因为在此数据集中，观测个数并不能反映出来，的可能值为N，DF，WEIGHT,WGT

或WDF，VARDEF=N表明要求用观测个数n作除数; VARDE F=DF 表明要求用误差自由度n-i(偏现变量前)或n-i-p(偏出变量后)，其中p是在PARTIAL语句中这些变量的自由度，而i值为0(当规定NOINT时)或1，VARDEF=WEIGH T 或WGT 表明要求用权数和W; VARDEF=WDF 表明要求用W-i(偏出变量前)或W-i-p(偏出变量后)。缺省时，用DF。

COVARIANCE—COV 要求从协差阵出发计算主成分，如果省略，则从相关阵出发进行分析，一般地，为了消除量纲的影响，把原变量进行标准化，由于标准化后的协差阵即为原变量的相关阵，一般情况下，不使用此项选择。

NOINT 要求不使用截距项，这时协方差和相关系数没有对均值做修正。

STANDARD—STD 要求在OUT=数据集里，主成分得分标准化为单位方差，若缺省，则主成分的方差等于相应的特征根。

NOPRINT 限制打印输出，只限制所在步的打印输出，对于其它的PROC 步并无影响。

2. VAR 语句列出被分析的数值变量，若缺省，则分析所有没有在基它语句中规定的数值变量，此语句比较常用。

3. PARTIAL 语句若用户想基于偏相关阵或偏协差阵进行主成分分析，则使用该语句规定被偏出去的变量。

4. BY 语句得到由BY变量定义的几组观测分别分析。

5. FREQ 语句指出频数变量，频数指的是观测出现的次数。

6. WEIGHT 语句指出数据集中的权变量，当同每个观测有联系的方差不相等时经常使用这个语句，而且权数变量的值是和方差的例数成比例。

输入数据集可以是由原始数据组成的数据集,也可以是TYPE=CORR或COV或SSCP的特殊数据集，原始数据可直接在步创建，而上述三种特殊的数据集可用其它过程创建，或者，也可以在PROC步创建，以下我们作简要的说明。

若在PROC步创建TYPE=SSCP的数据集，可用过程REG，例如：

PROC REG DATA=a1 OUTSSCP=b1;

PROC PRINCOMP DATA=b1;

第一句用REG过程创建名为b1的TYPE=SSCP的数据集，它包含变量的平方及叉积和。第二句: 用b1作为PRINCOMP过程的输入数据集作主成分分析。

7. 输入与输出数据集PRINCOMP 可产生两个输出数据集，一是由选择项OUT=产生，二是由OUTSTAT= 产生，例如; PROC PRINCOMP OUT=b1 OUTSTAT=b2; 该语句执行后可产生两个输出数据集，但是并不在OUTPUT窗口显示数据集的内容，用户可用PROC PRINT DATA=b1; PROC PRINT DATA=b2;RUN; 浏览b1与b2的内容。

OUT=生成数据集的内容有：观测序号、原始数据集中的所有变量、主成分得分的新变量，选择项N=k确定了新变量的个数，新变量的名字PRIN1，PRIN2,. . . ，PRINK(若缺省PREFIX=选择项)。新变量的均值为 0 ，方差等于相应的特征根，如果规定STD选择项，则新变量为标准化变量(均值为0，方差为1)。若规定PARTIAL语句，则还有用PARTIAL变量预测变量的残差，残差变量的名字由词头R_和变量VAR 的名字形成。

OUTSTAT=生成数据集的内容有：观测序号、字符变量_TYPE_和_NAME_、由VAR语句确定的被分析的变量，若无此语句，则为没有列在其它语句中的所有数值变量。如果规定PARTIAL语句，还有在OUT= 产生的数据集中描述过的残差变量。若有BY语句，则还包含BY变量。_TYPE_的内容如下：

MEAN 被分析变量的均值，若规定或则无此项观测。

STD 被分析变量的标准差，若规定，则无此项观测，若规定语句，则变量的标准差用变时预测的均方根计算。

N 观测样本的个数。

CORR 变量间的相关系数，若规定语句，则输出偏相关。

EIGENVAL 变量的特征根，特征根的个数由来确定，其余的特征根用缺失值代替。

SCORE 特征向量它的个数也有来确定，一般情况下，特征向量是正则化特征向量，若规定选择项，这时的特征向量要除以特征根的平方根，以使得到的得分具有单位标准差。

COV 变量间的协方差，只有当规定选择项时才产生，若使用语句，则输出偏协方差而不是原始的协方差。

SUMWGT 观测的权数和，这值对每个变量都相等，如规定语句和选择项，则权数和少减了变量的自由度，仅当这个值同的观测不同时才输出这个观测。

这里值得注意的是，若输入数据集是特殊结构的数据集，则不能生成产生的数据集，另外，由产生的数据集可以用来作回归，聚类等的输入数据集，由产生的数据集可以用于过程来计算主成分得分或者作为过程的输入数据集。

【例4.18】主成分回归，其中y为进口总额，x1为国内产值，x2为储存量，x3 为国愉消费量，可根据最小二乘法求出y与x1,x2,x3之间的回归方程为:

y=-10.130-0.051x1+0.578x2+0.287x3

从中可观察到x1的特号小于0，与实际不符，因此，我们可得用主成分分析，先对原始数据进行”预处理”。

下面的程序中，原始数据是存放chst.dat文件中，程序：

```
DATA chst;
  INFILE 'chst.dat';
  INPUT x1-x3 y @@;
RUN;
PROC PRINCOMP OUT=a1 OUTSTAT=a2;
  VAR x1-x3;
PROC PRINT DATA=a1;
PROC PRINT DATA=a2;
PROC REG DATA=a1;
  MODEL y=prin1-prin2;
```

§4.4.7 因子分析

FACTOR 过程进行几种类型的因子分析，可以使用正交或斜交旋转。输入数据可以是原始多变量数据，也可以是相关阵、协方差阵、因子载荷阵(模式阵) 或得分系数矩阵。

FACTOR过程的PROC步由以下语句组成：

```
PROC FACTOR 过程选项; /* 必选*/
  VAR 变量表;
  PRIORS 先验公因子方差表;
  FREQ 变量;
  WEIGHT 变量;
  BY 变量表;
  PARTIAL 变量表;
RUN;
```

1. PROC 语句

该语句后面的选择项共有40几个，这几只介绍几个常用的选择项，其大约可分为数据集选项、因子抽取、旋转、输出四类。

OUT=包括被分析数据的全部数据，及名为FACTOR1,...,FACTORk 的新变量，其中k为公因子的个数，它可以由选择项规定，若输入数据集为特殊结构的数据集，则无此项输出。OUTSTAT= 它包含因子分析的大部分统计结果，具体内容将在下面介绍。

因子提取方法选择项包括采用何种方法提取公因子，先验公因子方差初始值的估计是什么，因子分析是从相关阵还是从协差阵进行分析，确定公因子的个数等等。METHOD=| M= 该语句规定提取公共因子的方法，缺省时M=P 即用主成份分析法提取公共因子的数据集类型TYPE=FACTOR时除外，当输入数据集类型为TYPE=FACTOR 时缺省值为M=PATTERN。FACTOR 公共因子提取的方法有：

M=PRINCIPAL|PRIN|P 进行主成分分析，若规定PRIORS语句或者PRIORS不等于ONE，则进行主因子分析。

M=ML|M 进行极大似然法分析，该方法要求协差阵或相关阵是非奇异的。

M=PRINIT 进行迭代主因子分析。

M=ULS|U 进行没有加权的最小二乘因子分析。

M=ALPHA|A 进行 α 因子分析。

M=IMAGE|I 进行映象分量分析。

M=HARRIS|H 进行Harris分量分析，该方法要求|R|≠0。

M=PATTERN 从TYPE=FACTOR，CORR 或COV 的数据集中读取因子模型。

M=SCORE 从TYPE=FACTOR、CORR 或COV 的数据集中读取得分系数(_TYPE_='SCORE')。

PRIORS=name 该语句规定先验公因子方差初始值的估计方法即规定h_i^2 的取值方法，用户可以从以下几种方法中选择一种。

PRIORS=ONE|O 令$h_i^2 = 1, i = 1, ..., p$

PRIORS=SMC|S 取h_i^2 为第i个变量与其它所有变量的复相关系数的平方。

PRIORS=ASMC|A 取h_i^2 正比于未校正的复相关系数的平方。

PRIORS=MAX|M 取h_i^2为第个i变量与基它变量中最大绝对相关系数。

PRIORS=RANDOM|R 取h_i^2为在0-1之间豚从均匀分布的伪随机数。

PRIORS=INPUT|I 在DATA=的数据集中(TYPE=FACTOR)从_TYPE_= 'PRIORS' 或_TYPE_='COMMUNAL
第一个观测读取h_i^2。

若缺省PRIORS=name，则当M=P或M=PRINIT时，PRIORS=ONE，当M=A、M=U或M =ML 时，PRIORS=SMC。

COVARIANCE|COV 要求从协差阵出发作因子分析，该选择项只能同M=P, M= PRINIT，M=U或M=I MAGE 一起使用。

NFACTORS=k|NFACT=k|N=k 规定被提取公共因子的最大数目，缺省值为变量的个数在因子分析中，若缺省则初始公共因子的个数根据相关阵的特征根值而定，系统内部自动取特征根值大于1的个数为初始公共因子的个数，或者用户自已规定。

PROPORTION=n|PERCENT=n|P=n 对被保留因子规定使用先验公因子方差估计的公共方差所的比例。PROPORTIO N=0.85 和PECENT=85 是等价的。当该值大于1，被认为是百分数并用100除，用户不能在M=PATTERN或M=SCORE下规定此选择项，当该项缺省时，若被估的公因子方差超过1 ，则M=U，M=A 或ML 停止迭代并令因子个数为0，这时，下面的选择，允许迭代继续进行。

HEYWOOD|HEY 公因子方差大于1时令其为1。

ULTRAHEYWOOD|ULTRA 允许公因子方差超过1。

MINEIGEN=n|MIN=n 规定被保留因子的最小特征值，当M=PATTERN，M= SCORE时, 不能用此选择项，缺省值为0，除非规定NFACTOR=或者PROPORTION=，当对未加权的相关阵进行因子分析时，这个值为1

旋转方法选择项

ROTATE= |R= 规定公因子旋转的方法，缺省时，R=NONE，即不进行旋转，R=的后面共有七种填入方法

R=VARIMAX|V 规定方差最大旋转法。

R=OTHONAX 规定正交最大方差旋转法。

R=PROMAX 规定在正交最大方差旋转基础上进行斜交旋转。

R=EQUAMAX|E 规定均方最大旋转。

R=QURTIMAX 规定4 次方差最大旋转。

R=HK 规定Harris-Kaiser 情况 II 的斜正交旋转。

R=NONE|N 规定不进行旋转。

ALL 打印除图形之外的所有可选择的输输，当输入数据集为TYPE=CORR，COV或FACTOR 时，不能输入简单统计量，相关和MSA。

SIMPLE|S 打印均值和标准差。

CORR|C 打印相关阵。

MSA 打印抽样适当的Kaiser测度和负反映象相关阵。

NPLOT=n 规定作图个数，缺省值是所有因子，$2 \leq n \leq q$ (q为公因子总数)，若规定NPLOT=n，则对作n 个公因子组成的所有因子对作载荷图，共可作C_n^2张图。

PLOT 作旋转后的因子模型图。

PREPLOT 作旋转前的因子模型图。

RESIDUALS|RES 打印残差相关阵和有关的贪偏相关阵。

SCORE 打印因子得分系数，每个因子同这些变量的平方多重相关也被输出，但没有旋转的主成分分析情况除外。

输出选择项还有许多，由于不常用，故不在这儿一一列出。

2. VAR 语句

列出被分析的数值变量，缺省时，表示分析在其它语句中没有列出有所有数值变量。

3. PRIORS 语句

格式$h_1^2, h_2^2, h_3^2, ...$; 对语句中每个变量规定一个0-1 之间的数值作为先验公因子方差的初始估计，顺序必须语句相对应. 例：

PROC FACTOR; VAR x1 x2 x3; PRIORS 0.90 0.93 0.95;

若在PROC FACTOR 语句中已使用PRIORS选择项，则此句可省略。

4. FREQ 语句作用同PRINCOMP过程中的FREQ语句。

5. WEIGHT 语句如果用户对输入数据集中每个观测使用相对权数时，用语句规定一个包含权数的变量，当同每个观测有联系有方差不相同时，经常使用这个语句，而且权数变量的值与方差的倒数成比例。

6. BY 语句对由变量定义的几个观测组进行独立的分析。

7. PARTIAL 语句规定被偏出的变量名字。

8. 输入与输出数据集

(1). 输入数据集最简单的输入数据集是由原始数据丢失，只有原始数据的相关阵或协差阵等等，这时，可以用前节介绍的方法，在步建立具有特殊结构的数据集，过程的输入数据集有特殊形式有以下四种：TYPE=CORR，TYPE= COV ，TYPE= FACTOR(这个数据集必须包含_TYPE_='PATTERN'的那些观测，若这些因子相关，还要求输入因子间的相关系数_TYPE_='FCORR')和TYPE=CORR(该数据集与前不同，它必须包含相关阵从及因子的得分系数_TYPE_='SCORE') 它们都可以在步或步创建，前两种的创建方法已在前节介绍，下面介绍后两种数据集的创建方法。

TYPE=FACTOR 的创建方法

这个数据集必须包含_TYPE_='PATTERN'的那些观测，若这些因子相关，还要输入因子间的相关系数(_TYPE_='FCORR')。例如：

```
DATA a1(TYPE=FACTOR);
  INPUT _TYPE_ $ _NAME_ $ x1 x2 x3;
CARDS;
  PATTERN FACTOR1 -0.079  0.98   0.048
  PATTERN FACTOR2  1.002 -0.094  0.975
  PATTERN FACTOR1  1.000  0.202  .
  PATTERN FACTOR2  0.202  1.000  .
```

用上述语句，就在DATA步创建了一个名为a1的TYPE=FACTOR的数据集，它可作为FACTOR 过程的输入数据集，并用METHOD=PATTERN读入因子模型。

由于因子分析的方法较多，因此，有时必须多次调用FACTOR过程，若每次都用原始数据集作为输入数据集，则计算的时间较长，占用的内存也较多，因此可以在步创建TYPE=FACTOR的数据集作为FACTOR过程的输入数据集。例：

```
PROC FACTOR DATA=a1 OUTSTAT=a2 M=ML;
PROC FACTOR DATA=A2 ROTATE=P;
PROC FACTOR DATA=A2 M=PRIN;
```

第一句: 用FACTOR过程创建一个输出数据集类型A2，A2的类型为TYPE=FACTOR，A1为原始SAS数据集，公因子提取方法为极大似然法。

第二句: 用作为输入数据集作因子分析，由于M缺省，按过程规定M= PATTERN, 并执行主因子分析，旋转方法为PROMAX。

第三句: 用A2作为输入数据集,用主成分分析法提取公因子。

(2). 输出数据集过程可产生两个输出数据集，一个由选择项OUT=产生，一个由OUTSTAT=产生. 例如: PROC FACTOR OUT=b1 OUTSTAT=B2。可用PRINT过程看其内容：PROC PRINT DATA=B1;PROC PRINT DATA=B2;RUN;

以下的程序直接读入类型为相关阵的数据，进行因子分析。文献中往往不给出原始数据却给出了相关系数矩阵，这种方法特别适用。数据指定的变量名_NAME_以及_TYPE_是SAS 系统的保留名，表示变量的名称和类型。这里的类型为CORR，在SAS 样本程序中对例数未加指定则系统设定一个大的值，在例2.7和第十五章中都给出了读取该数据的另一种简捷方法。

【例4.19】对某项研究的数据使用相关阵进行因子分析[24]

```
data p341(type=corr);
input _name_$ x1-x3 _type_$;
cards;
x1 1.0000000  -.3333333  0.6666667 corr
x2 -.3333333  1.0000000  0.0000000 corr
x3 0.6666667  0.0000000  1.0000000 corr
.  5          5          5         n
proc factor data=p341 nfactors=3 r=varimax corr;
run;
```

使用主成分分析方法(Initial Factor Method: Principal Components)：

先验的公因子方差估计为1 (Prior Communality Estimates: ONE)：

相关阵的特征值：

	1	2	3
特征值	1.745356	1.000000	0.254644
差值	0.745356	0.745356	
贡献率	0.5818	0.3333	0.0849
累计贡献率	0.5818	0.9151	1.0000

因子载荷(Factor Pattern，Λ)：$R = \Lambda\Lambda' + D$：

	因子1	因子2	因子3
X1	0.93417	0.00000	0.35682
X2	-0.41777	0.89443	0.15958
X3	0.83555	0.44721	-0.31915

每因子解释的方差是各个分量的平方和：1.745356、1.000000、0.254644。

使用方差极大(VARIMAX) 旋转，正交变换矩阵：

	1	2	3
1	0.65098	-0.32887	0.68416
2	0.43802	0.89884	0.01529
3	-0.61998	0.28972	0.72917

旋转后的因子结构：

	因子1	因子2	因子3
X1	0.38690	-0.20384	0.89931
X2	0.02088	0.98757	-0.15579
X3	0.93768	0.03472	0.34577

每个因子解释的方差为：1.029366、1.018051、0.952582。

FACTOR 提供了许多方法，如近似方法中的METHOD= PRIN (PRIORS= SMC，Squared Multiple Correlations)、METHOD=Harris，METHOD=Image；最优方法中的METHOD=Prinit，METHOD=Alpha，MET 在因子分析的输出结果中都给予相应的提示。

§4.4.8 典型相关分析

CANCORR 过程用于典型相关分析、偏典型相关分析以及输出各种结果，还用于典则冗余分析(canonical redundancy) 分析。CANCORR 对每个典型相关以更小的相关为零的假设进行系列检验。为了使检验的概率值有效，两组变量中至少一组应具有近似多元正态分布。

CANCORR 过程还提供了多重回归分析选项，以帮助使用者解释典型相关分析的结果。你可以检查一组变量中的每个变量与另一组变量的线性回归。CANCORR 使用线性回归中的最小二乘准则。语句格式及说明如下：PROC CANCORR 过程选项; /* 必选*/ VAR 变量表; WITH 变量表; /* 必选*/ PARTIAL 变量表; FREQ 变量; WEIGHT 变量; BY 变量表; 因为要明确分析名称及其变量，故PROC CANCORR 与WITH 语句都是必选语句。

1. PROC 语句

选择项较多，用户可根据需要选择几个。DATA= 被分析数据集名，它既可以是原始的数据集，也可以是TYPE=CORR 或COV的特殊数据集，若缺省，则使用最新建立的数据集。OUT= 输出数据集，它包括原始数据和典型变量得分，当的类型为TYPE=CORR或COV时，没有输出。OUTSTAT= 输出数据集，它包括过程产生的各种统计量，由选择项的不同，数据集包含的内容出不尽相同。

输出选择项

ALL 打印所有选择的输出。

NOPRINT 限制打印输出。

SHORT 除典型相关和多元统计列表外，限帛所有缺省时的输出。

SIMPLE|S 打印均值和标准差。

CORR 打印原变量间的相关系数。

NCAN=n 规定要求输出的典型变量的个数。

VPREFIX|VP= 规定来处语句中的典型变量各字的前缀，例如，VP=S P 则典型变量的各字为SP1，SP2 等等，若缺省，则典型变量的名字为V1，V2 等等，注意: 典型变量各字的字符个数不能超过8个。

VNAME|VN='label' 在打印输出时对VAR语句中的变量规定最多40个字符长的字符常数作为变量的标记，必须用单引号反字符常数括起来，若省略，这些变量称为VAR变量。

WPREFIX|WP= 规定来处语句中的典型变量各字的前缀，缺省时，典型变量各字为W1，W2等等。

WNAME|WN='label' 在印输出时对语句中的变量规定最多个字符长的字符常数作为该变量的标记，必须用单引号把字符常数括起来，若省略，则统称为WITH变量。

RDF=回归自由度. 若输入的观测数据是回归分析的残差，它用于规定回归自由度，观测的有效个数是实际值减EDF=值，截距项的自由度没有包含在RDF=的选择项中。

EDF=误差自由度. 若输入的观测数据是跨归的残差，此项选择用于规定回归分析的误差自由度，观测的有效个数为EDF=的值加1，如输入数据集(在DATA 步) 为TYPE=CORR或COV等时，过程中没有合适的选择项可以将原始数据的样本含量n 准确地输入，因此，一般用选择项EDF=n-1，为典型相关分析提供一个计算误差自由度的参考值，若缺省，此时，系统内部指定n=10000 作为样本含量参与有关计算加统计检验，不是很合适。

2. VAR 语句

该语句用来列出被分析的第一组数值变量，若缺省时第一组变量为在其它语句中没有提到的所有数值变量。

3. WITH 语句

列出被分析的第二组数值变量，不能缺省。

4. PARTIAL 语句

用于在偏相关基础上进行典型相关分析并列出从VAR变量和WITH 变量中偏出去的变量。

5. FREQ 语句

指示频数变量名。如果FREQ变量的值小于1，这个观测在分析中不使用。当CANCORR过程计算显著性概率时，观测的总数取为变量FREQ的和。

6. WEIGHT 语句

给出权数变量的名字，WEIGHT语句和FREQ语句的作用类似，差别在于WEIGHT语句不能必改变自由度或观测的个数，仅当WEIGHT变量值大于0 时这个观测才能用于分析计算。

7. BY 语句

得到由BY变量定义分组的独立分析。

§4.4.9 结构方程模型分析

结构方程模型常用于社会学和计量经济学分析，详见第 1 2 章LISREL，除LISREL 外，SAS PROC CALIS 也可用于分析。

计量经济学(econometrics 或经济计量学) 是经济理论、数理经济的一种综合，它是以经济变量为出发点，论述各种经济变量之间关系的计量方法。SAS 有专门的模块SAS/ETS 来处理这一类问题，如时间序列分析，在SAS/STAT 部分中主要可借助过程CALIS 对描述经济变量相互关系的方程组进行分析。在CALIS 框架中纳入了较为广泛的模型，如COSAN、LISREL、RAM等。

CALIS 过程使用协方差结构分析估计和检验线性结构方程模型的适度，结构方程建模是计量经济学以及行为科学中重要的统计学工具，它表达了几个变量的关系，这些变量既可能是直接测量的，也可能是虚拟的变量(latent variables)。CALIS 使用广义COSAN 模型方法，模型的参数可以具有线性或非线性约束。

CALIS 过程能用于协方差结构分析、拟合线性结构方程组以及通路分析。这些名称或多或少可以互用，但却强调了分析的不同的方面。协方差结构分析指对一系列变量的方差协方差构造一个模型并且使用观察协方差阵来拟合它；在线性结构方程模型中，模型是一个方程组，把几个随机变量联系起来，也对随机变量的方差协方差进行假设；在通路分析中，模型的构造是以通路图的形式出现，一些箭头连接各变量。通路模型和线性结构方程模型可以转化为协方差阵模型，并因而使用协方差结构分析进行拟合，三种模型均允许使用隐含变量和测量误差。具体的模型如：

- 多重与多元线性回归
- 有测量误差的模型
- 带有隐变量的结构方程
- 通路分析和因果建模
- 有相互因果关系的模型
- 任意阶的探索性与确证型因子分析
- Three-mode 因子分析
- 典型相关
- 一系列其它隐变量模型。

有几种方法指定CALIS 的模型，如：

- 通过FACTOR 语句结合可选的MATRIX 和VARNAMES 语句进行约束一阶因子分析或分量分析
- 使用列表形式的RAM 语句结合可选的VARNAMES 语句指示简单的路径分析模型(McArdle 的RAM 模型)
- 使用方程类型的LINEQS 语句结合STD 和可选的COV 语句指示结构方程。
- 使用COSAN 和MATRIX 语句以及可选的VARNAMES 语句分析一簇矩阵模型(与McDonald 和Fraser 的COSAN 程序类似)。
- 使用INRAM= 指示模型，INRAM= 通常由前一次的CALIS 运行产生，或者由数据步产生。

必须给CALIS 中的每一个参数一个至多八个字符的名称，第一个字符应是下线或字母，若使用编程语句，应当避免与CALIS 的语句冲突。变量名用于结果输出和OUTRAM= 及OUTEST= 数据集、施加等式约束以及使用程序语句施加复合约束。变量名可以使用施前缀来产生。

限制参数为一个常数，在MATRIX，RAM，LINEQS，STD，COV 语句或INRAM= 指示这个值；要限制两个参数相等，给它们使用相同的名称；要限制一个参数大于等于或小于等于一个常数，使用BOUNDS 语句，这对保证方差非负很有用；更复杂的约束可以通过程序语句来实现，此时一些参数不是模型矩阵的元素，却在PARAMETERS 语句中定义，模型矩阵的元可使用PARAMETERS 语句的参数编程来计算，函数的导数不需要指定，过程自动计算解析导数。

参数的估计准则有，不加权最小二乘(ULS)，广义最小二乘(GLS)，多元正态资料的极大似然(ML)，加权最小二乘(WLS 包括权矩阵输入以及Browne's 不依赖于特定分布的渐近方法)，对角元加权最小二乘(DWLS 包括权矩阵输入)。

估计方法通过METHOD=，ASYCOV=，INWGT=，NODIAG，WPENALTY=，和WRIDGE= 选项指定。CALIS 在默认情况下使用相关矩阵进行拟合。参COV，UCORR，UCOV，和AUGMENT 的有关选项。CALIS 提供了几种最优化算法，它们是：Levenberg-Marquardt 算法，修正牛顿法，各种拟牛顿法，各种共轭梯度法，拟牛顿法和共轭梯度法可以因各种一维搜索方案而变化。

CALIS 语法

PROC CALIS < 过程选项>;
模型是以下五种情况的一种：
RAM 对通路分析的语句，加上：
  VARNAMES 名称指示;
LINEQS 线性结构方程语句，加上：
  STD 方差指示;
  COV 协方差指示;
COSAN 矩阵模型语句，加上：
  MATRIX 矩阵元素定义;
  VARNAMES 名称指示;

FACTOR 一阶因子模型语句，加上：
MATRIX 矩阵元素定义;
VARNAMES 名称指示;
INRAM= 数据集模型指示;

以COSAN语句为例，常用的记号有两种，一种是矩阵的形状，一种是逆阵信息。第一种有：IDE(单位矩阵)、ZID(单位矩阵)、DIA(对角阵)、ZDI(对角阵)、LOW(下三角阵)、UPP(上三角阵)、SYM(对矩阵)、GEN(方阵)，第二种有IND(逆矩阵)、IMI(单位阵减去矩阵后的逆)。

以下语句对所有模型适用。

BOUNDS 边界约束;
BY 变量表;
FREQ 频率变量;
PARAMETERS 参数名称;
PARTIAL 偏去变量;
VAR 分析变量;
WEIGHT 权变量;

可以使用编程语句对参数施加约束。语句有ABORT，ARRAY,CALL，DELETE，DO ,GOTO,IF/IF-THEN-ELSE，LINK,PUT,RETURN,SELECT,STOP,SUBSTR,WHEN。

过程及输入输出指示涉及以下有关信息，数据集如输入数据集DATA=、INRAM=、INWGT=，输出数据集OUTSTAT=、OUTRAM=、OUTWGT=、OUTEST=；缺失值；估计准则；标准误；相关阵的拟合；自动变量筛选；外生变量；最优化技术；迭代过程，等等。

PROC CALIS 选项索引。
ALL 请求所有输出，其部分内容见下例。
ALPHARMS= α , 0$\leqslant \alpha \leqslant$1。默认值为0.1。打印Steiger与Lind的均方误差系数。
ASYCOV|ASC=BIASED,UNBIASED,CORR 渐近协方差阵公式。
AUGMENT 给协方差阵增加一列，分析增广矩阵。
BIASKUR 计算未校正偏差的偏度和峰度。
CORR|PCORR 打印参与分析和估计的修正或未修正的协方差阵或相关阵。
COV 分析协方差矩阵。
DATA=数据集输入数据集。
DEMPHAS|DE=r 增强中心模型矩阵对角元的影响。
DFREDUCE|DFRED=i 使χ^2降低的自由度数目。
EDF|EDF=n 设定有效观察数目为n+i，当选定NOINT,UCORR或UCOV时i=0 。NOBS也可用于指定观察数目。
FCONV|FTOL=r 指示函数的相对收敛准则。
GCONV|GTOL=r 指示绝对梯度收敛准则，精度越高，耗机时越长。
G4=i Hessian阵奇异时的标准误算法，默认值为60。
HESSALG|HA=1|2|3|4|5|6|11 指定LEVMAR和NEWRAP优化算法的海森矩阵。解析

法用1,2,3,4,11指定，有限差分法用4,5指定，密集存贮用1,2,3,4,5,6，稀疏存贮用11。
INRAM=数据集含模型描述的输入数据集。
INWGT=数据集含权矩阵的输入数据集。
KURTOSIS|KU 打印单变量和多变量峰度。
MAXFUNC|MAXFU=i 最多的函数调用次数。
MAXITER|MAXIT=i 最大迭代次数。
METHOD|MET=名称估计方法ML|M|MAX,GLS|G,WLS|W|ADF,DWLS|D,ULS| LS| U ,LSML|LSM|LSMAX，LSGLS|LSG，LSWLS|LSW|LSADF,LSDWLS|LSD,NONE|NO|N。
MODIFICATION|MOD 打印Lagrange乘子检验指标或修正指数。
NOBS=观察数设观察数.
NOINT 分析不包含常数项的协方差阵和相关阵。
NODIAG|NODI 拟合中不使用对角元。
NOMOD 不打印修正指数。
NOPRINT|NOP 不打印结果。
NOSTDERR|NOS 不计算标准误。
OMETHOD|OM|TECHNIQUE|TECH=名称指定最优化方法，内容有：LEVMAR| LM |MARQUARDT，NEWRAP|NR|NEWTON,QUANEW|QN，CONGRA|CG，NONE|NO。
OUTEST=数据集输出参数数据集。
OUTRAM=数据集模型和估计量的输出数据集。
OUTSTAT=数据集统计量输出数据集。
OUTWGT=数据集含权矩阵的输出数据集。
PCOVES|PCE 打印信息矩阵和估计协方差阵。
PDETERM|PDE 打印决定系数。
PESTIM|PES 打印参数估计值。
PINITIAL|PIN 打印模型矩阵和参数初值。
PJACPAT|PJP 打印Jacobi矩阵的结构与常数元素。
PLATCOV|PLC 打印内生变量之间以及内生与外生变量之间的协方差、内生变量得分回归系数。
PREDET|PRE 分析模型所定义的预测乘积矩阵的形式与常数元素。
PRIMAT|PMAT RAM或LINQUES语句模型参数矩阵形式的输出。
PRINT|PRI 增加KURTOSIS,RESIDUAL,PLATCOV及TOTEFF的内容到打印输出。
PRIVEC|PVEC 参数估计、标准误、梯度及t-值用向量形式输出。
PUNDOC|PUND 打印手册未加说明的一些信息如内存使用量等。
PWEIGHT|PW 打印权矩阵。
RADIUS=r 在Levenberg-Marquardt中的初始置信区域半径。
RANDOM=i 随机产生参数的初值。
DFR|RDF=n 设定有效观察数目为实际观察数-n，常数项自由度不应算在n 之内。使用PROC CALIS计算回归模型时，可以设定RDF=自变量数来得到PROC REG算得的那种通常的标准误。
RESIDUAL|RES 打印绝对和规格化剩余矩阵及有关信息。
RIDGE=r 岭因子。
SALPHA=r 前五次迭代的不维搜索初始步长上界，默认值为1。

SHORT|PSH 不包括PINITIAL,SIMPLE及STDERR的打印内容。
SIMPLE|S 打印单变量均值、标准差、偏度和峰度统计量。
SINGULAR|SING=r 0¡r¡1，奇异性准则。
SLMW=r 逐步多变量Wald 检验的概率极限，默认为0.05，概率值小于该值时停止计算。
SMETHOD|SM|LINESEARCH|LIS=1|2|3 指定一维搜索方法，1 表示三次内插和外插所需函数和梯度调用相同，2指示函数调用较梯度调用多一些,这在协方差结构分析中是可取的，因为函数调用的代价相对便宜。3含义与1相同，但它可以经SPRECISION=选项修正。
SPRECISION|SP=r 一维搜索精度，由0.06到0.4不等。
START=r 常数初值，多数情况下过程自行定义。
STDERR 打印近似标准误。
SUMMARY|PSUM 打印拟合情况、误差、警告等信息。
TOTEFF|TE 打印总效应和间接效应。
UCORR 分析未修正CORR 矩阵。
UCOV 分析未修正COV 矩阵。
UPDATE|UPD=名称拟牛顿法或共轭梯度法的修正技术。对于QUANEW 内容有BFGS,DFP,DBFGS,DDFP;对于CONGRA内容有PB,FR,FR。
VARDEF=DF,N,WDF,WEIGHT,WGT 指定方差除数。
WPENALTY|WPEN=r 增加一个相关阵对角元的惩罚权重，约束对角元为1.0。
WRIDGE=r 对于GLS,WLS,DWLS估计的权矩阵的指定岭因子。

【例4.20】下面是一个食品消费和价格的模型[6]
需求方程：y=a0 + a1 x1 + a2 x2 + u1
供给方程：y=b0 + b1 x1 + b3 x3 + b4 x4 + u2
各变量的含义是：y：每人的食品消费；x1 ：食品价格与日用品价格的比率；x2: 价格稳定下的可自由支配的收入；x3：农产品价格与日用品价格的比率；x4：年份时间，u1、u2 是方程中的误差项。其中变量x2、x3、x4 是外生变量(exo -genous variable)，它们的值可影响食品市场，y 与x1 是内生变量(endo ge -nous variable)。模型可整理成与LISREL 类似的形式：

$$\begin{pmatrix} y \\ x_1 \end{pmatrix} = \begin{pmatrix} 0 & a_1 \\ -1/b_1 & 0 \end{pmatrix} \begin{pmatrix} y \\ x_1 \end{pmatrix} + \begin{pmatrix} a_0 & a_2 & 0 & 0 \\ -b_0 & & -b_3 & -b_4 \\ -- & 0 & -- & -- \\ b_1 & & b_1 & b_1 \end{pmatrix} \begin{pmatrix} 1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

上述形式可直接引入CALIS 过程，LINEQS 语句要求每个内生变量恰好出现在一个方程的左边，在SAS的样本程序中用下面的形式：
Q = α 1 INTERCEP + α 2 P + α 3 D + E1,
P = γ 1 INTERCEP + γ 2 Q + γ 3 F + γ 4 Y + E2;
其中x3 的含义与上面略有不同。程序如下：

```
DATA FOOD;
TITLE 'Food example of KMENTA(1971, p.565 & 582)';
```

```
* Kmenta, J.(1971) Elements of Econometric New York: MacMillan;
TITLE2 'Compare CALIS with SYSLIN estimates';
   INPUT Q P D F Y;
   LABEL Q='Food Consumption per Head'
         P='Ratio of Food Prices to General Price'
         D='Disposable Income in Constant Prices'
         F='Ratio of Preceding Years Prices'
         Y='Time in Years 1922-1941';
  CARDS;
  98.485  100.323   87.4   98.0   1
  99.187  104.264   97.6   99.1   2
 102.163  103.435   96.7   99.1   3
 101.504  104.506   98.2   98.1   4
 104.240   98.001   99.8  110.8   5
 103.243   99.456  100.5  108.2   6
 103.993  101.066  103.2  105.6   7
  99.900  104.763  107.8  109.8   8
 100.350   96.446   96.6  108.7   9
 102.820   91.228   88.9  100.6  10
  95.435   93.085   75.1   81.0  11
  92.424   98.801   76.9   68.6  12
  94.535  102.908   84.6   70.9  13
  98.757   98.756   90.6   81.4  14
 105.797   95.119  103.1  102.3  15
 100.225   98.451  105.1  105.0  16
 103.522   86.498   96.4  110.5  17
  99.929  104.016  104.4   92.5  18
 105.223  105.769  110.7   89.3  19
 106.232  113.490  127.1   93.0  20
PROC CALIS UCOV AUG DATA=FOOD ALL;
TITLE3 'Compute ML estimates with intercept';
 LINEQS  Q = ALF1 INTERCEP + ALF2 P + ALF3 D + E1,
         P = GAM1 INTERCEP + GAM2 Q + GAM3 F + GAM4 Y + E2;
 STD  E1-E2 = EPS1-EPS2;
 COV  E1-E2 = EPS3;
 BOUNDS EPS1-EPS2 >= 0. ;
 RUN;
```

方程组的常数项系数的求解可经过程选项UCOV和AUGMENT实现。样本程序也给出了还原成原来系数的相应语句，此处从略。模型的估计结果：

		真实参数	OLS	TSLS	ML
需求方程	常数	96.5	99.90	94.63	93.62
	x1	-0.25	-0.32	-0.24	-0.23
	x2	0.30	0.33	0.31	0.31
供给方程	常数	62.5	58.28	49.53	49.53
	x1	0.15	0.16	0.24	0.24
	x3	0.20	0.25	0.26	0.26
	x4	0.36	0.25	0.25	0.25

CALIS 的输出很详细，列出内容很多，只能择其部分加以解释。

第一部分：模式与初值

	Matrix	行与列	矩阵类型	
1	_SEL_	6 8	SELECTION	
2	_BETA_	8 8	EQSBETA	IMINUSINV
3	_GAMMA_	8 6	EQSGAMMA	
4	_PHI_	6 6	SYMMETRIC	

内生变量数目为 2 。用P、Q 表示。外生变量有六个，分别用D、F 、Y 、INTERCEP、E1、E2 表示，其中E1、E2 是误差。以下是各变量的均值、标准差、偏度峰度、一系列统计指标。

一系列评价拟合指标如下：

```
Fit criterion . . . . . .     0.1603
Goodness of Fit Index (GFI) . . . . . .     0.9530
GFI Adjusted for Degrees of Freedom (AGFI)  . . .     0.0120
Root Mean Square Residual (RMR) . . . . . .     2.0653
Chi-square = 3.0458        df = 1       Prob>chi**2 = 0.0809
Null Model Chi-square:     df = 15                  534.2738
```

拟合方程为：Q = - 0.2295*P + 0.3100*D + 93.6196*INTERCEP + E1 标准误0.0923 α 2 0.0448 α 3 7.5742 α 1 t 值-2.4857 6.9187 12.3603 P = 4.2140*Q - 0.9305*F -1.5580*Y - 218.8971*INTERCEP +E2 标准误1.7540 γ 2 0.3960 γ 3 0.6650 γ 4 137.6989 γ 1 t 值2.4025 -2.3500 -2.3429 -1.5897

标化方程：

Q = - 0.2278*P + 0.3016*D + 0.9273* INTERCEP+ 0.0181 E1
α 2 α 3 α 1

P = 4.2468*Q - 0.9048*F - 0.1863*Y - 2.1849*INTERCEP + 0.0997 E2
γ 2 γ 3 γ 4 γ 1

§4.4.10 多维尺度变换

MDS 拟合两维或三维模型，具有ALSCAL 和MLSCALE等许多过程的优点，SUGI 补充程序库中包含了ALSCAL和MLSCALE。MDS 使用非线性最小二乘估计下列参数：

配置(configuration) 每个对象在一维或多维欧氏空间上的坐标

分维上的影射系数(dimension coefficients)

转换参数(transformation parameters) 指关联距离和数据的有关参数

根据LEVEL= 的不同，MDS 拟合下面两个模型之一：

fit(datum)=fit(trans(distance))+error

fit(trans(datum))=fit(distance)+error

其中，fit 是由FIT=选项事先指定的指数或对数转换，trans 是一个估计的最优转换(线性,affine,指数或单调)，datum 是现个对象相似或不相似的度量，distance 是算得的两对象在一维或多维空间中的距离，指定COEF=IDENTITY时，是未加权欧氏距离; 若COEF=DIAGONAL，则它是加权欧氏距离，权重是影射系数的平方，error 是误差项并假定独立同分布，分布为正态。PROC MDS 过程的格式如下：

PROC MDS <选项>;
  VAR 变量表;
  INVAR 变量表;
  ID—OBJECT 变量;
  MATRIX—SUBJECT 变量;
  WEIGHT 变量;
  BY 变量表;

一般来说，MDS只打印迭代过程，因而总需要一些选项。MDS 也需要PLOT 或GPLOT过程进行图示。BY 语句指示按照指定的变量分组分析。ID 语句指示记录标号。INVAR 语句指示INITIAL=数据集中的数据变量，第一个变量相应于第一维，第二个变量相应于第二维，等等，语句省略时为DIM1，DIM2,...,等。MATRIX 语句指示DATA=数据集中针对数据矩阵或对象的标号，标号将用于打印及在OUT=和OUTRES= 数据集中使用，语句省略时用标号1,2,...,等。VAR 语句指示DATA= 数据集中包含对象间相似或不相似的度量。每变量相应于一个对象，语句省略则表示使用所有未被其它语句使用的变量。WEIGHT 表示公变量。MDS的细节可参文献[]。

程序中就可以使用LEVEL=ABSOLUTE选项。结果给出的不适合度指标(Badness- of-fit)提示模型拟合非常之好。对结果进行图时，图轴上应有相同的单元，可以利用PLOT中的VTOH来指示纵轴和横轴的比例，同时，也应该指示VAXIS=和HAXIS 有相同的刻度。

【例4.21】美国十城市飞行距离的数据，是欧氏距离很好的近似，因而不需要任何转换。程序及运行结果如下：

```
DATA   CITY;
TITLE 'INTERCITY FLYING MILEAGES';
INPUT  (ATLANTA CHICAGO DENVER HOUSTON LOSANGEL
       MIAMI NEWYORK SANFRAN SEATTLE WASHDC) (5.)
       @56 CITY $15.;
CARDS;
  0                                              ATLANTA
587    0                                          CHICAGO
1212  920    0                                    DENVER
 701  940  879    0                              HOUSTON
1936 1745  831 1374    0                         LOS ANGELES
 604 1188 1726  968 2339    0                    MIAMI
 748  713 1631 1420 2451 1092    0               NEW YORK
2139 1858  949 1645  347 2594 2571    0          SAN FRANCISCO
```

```
   2182 1737 1021 1891  959 2734 2408  678    0          SEATTLE
    543  597 1494 1220 2300  923  205 2442 2329    0     WASHINGTON D.C.
   PROC MDS DATA=CITY FIT=2 LEVEL=ABSOLUTE OUT=OUT OUTRES=RES;
      ID CITY;
   TITLE2 'ABSOLUTE LEVEL, GOOD START';
   RUN;
   PROC PLOT DATA=OUT; PLOT DIM2 * DIM1 $ CITY; WHERE _TYPE_='CONFIG';
   RUN;
              Plot of DIM2*DIM1$CITY.  Symbol points to label.
       |
       |
   750 +
       |
       |
       |                                      > HOUSTON            > MIAMI
   500 +
       |
D      |        > LOS ANGELES
i      |
m  250 +
e      |
n      |   > SAN FRANCISCO                                 > ATLANTA
s      |
i    0 +                      > DENVER
o      |
n      |
       |
2 -250 +                                                  WASHINGTON D.C.
       |                                          > CHICAGO     ^
       |
       |
  -500 +                                                        > NEW YORK
       |     > SEATTLE
       |
       |
  -750 +
       |
       --+----------+----------+----------+----------+----------+----------+-
       -1500      -1000      -500         0         500       1000       1500

                                  Dimension 1
```

上图形象地说明了多维尺度变换的用处。

§4.5 统计实验设计

§4.5.1 简述

SAS/STAT 可以用用PLAN 过程进行随机化实验，而大部分实验设计是在SAS/QC 中，SAS/QC 模块的内容有：实验设计、统计过程控制(CUSUM、MACONTROL、SHEWCHART和一系列函数)。过程能力分析(CAPABILITY)。抽样方案评价。

§4.5.2 SAS/QC 实验设计功能

过程FACTEX 可进行析因设计、部分析因设计、混合析因设计。这三类设计均可以出现区组。另外，象不完全区组设计也可以经FACTEX 与DATA 步结合而生成。FACTEX 是交互式运行的，初步实验设计后，可以追加其他的语句，但不必用PROC 语句重新开始启动。在FACTEX 中可以：①打印设计点；②检查设计的结构或生成设计的规则；③修正这个设计的大小、区组的指示方法或重新指定模型；④把设计输出到一个数据集；⑤对设计进行随机化；⑥重复这个设计；⑦把设计中表示因水平的标准编码换成适当的值，如-1和+1 换成low和high；⑧寻找其它的设计。

过程OPTEX 用于一个标准的设计如析因或部分析因不适用的情况，包括因子的某些水平不能实验、资源的限制了实验的次数、以及非标准的线性可非线性模型。该过程使用DETMAX，sequential,exchange 或Federov 方法，基于A-最优(极小化信息矩阵X'X 逆阵的迹) 或D-最优(极大化设计信息阵的行列式—X'X—) 产生设计。OPTEX 也是一个交互式过程，初步设计后可以继续：①检查这个设计；②输出到数据集；③改变模型并寻找另外的设计；④改变寻找的方式。

ADX 宏系统包括一系列宏调用，需要使用SAS/BASE、SAS/STAT、SAS/QCS 、AS/GRAPH 模块，用于构造：①2-水平析因或部分析因设计。可多达128 次实验和11 种处理，可以有区组。②多达47 个因素的2-水平筛选设计(screening 或Plackett-Burman 设计)。③8 因素的正交和旋转中心复合设计(central composite 或Box-wilson 设计)，有区组或无区组。④有或无约束组分的混合设计，因素数目不受限制。这包括中心或网格单纯形(simplex- centroid 或simplex-lattice) 及McLean-Anderson 设计。

ADX 菜单系统是完全交互式，适于初学者，使用SAS/AF 、SAS/STAT 与SAS/GRAPH 进行ADX 宏中的大部分设计、回归分析并行变量筛选、估计效应、极大似然指数转换以及拟合响应的轮廓图及立体图象，有时还需要使用SAS/FSP。设计内容如2-水平析因或部分析因设计，有或无区组效应、中心复合(Box-Wilson) 与Box-Behnken 设计、有或无约束组分的混合设计，因素数目不受限制。这包括中心或网格单纯形(simplex-centroid 或simplex- lattice) 及McLean- Anderson 设计。

1. ADXGEN.SAS (general) 含宏定义adxcode、adxdcode、adxinit、adxqmod、adxprt、adxtrans。
2. ADXFF.SAS (部分析因) 含宏定义adxalias、adxffa、adxffd、adxpbd 、adxpff。
3. ADXCC.SAS (中心复合) 含宏定义adxadcen、adxpcc。
4. ADXMIX.SAS (混合设计) 含宏定义adxmamd、adxscd、adxsld、adxxvert。

ADX 菜单系统的使用应有扩充内存的存在，特别地，应有LIM EMS 3.2 或以后的版本。

调用ADX 宏方法：是在SAS 的程序文件中使用%INCLUDE '!SASROOT\ SASMACRO\FILENAME'; 第一个文件名应是ADXGEN.SAS，并且在新在设计开始时，使用ADXINIT 宏。

§4.5.3 用例

完全2^5 析因设计，没有区组：

PROC FACTEX; FACTORS x1 x2 x3 x4 x5; RUN;

完全2^5 析因设计，使用区组：

```
PROC FACTEX;
   FACTORS x1 x2 x3 x4 x5;
   BLOCKS SIZE=16;
   MODEL est=(x1|x2|x3|x4|x5@2);
RUN;
```

MODEL 语句中指示所有主效应和两因子交互是可估的，忽略其它效应。

除了FACTORS 语句，在程序中指定SIZE 和MODEL 语句进行部分析因设计，

```
PROC FACTEX;
   FACTORS x1 x2 x3 x4 x5;
   MODEL RES=4;
   SIZE FRACTION=2;
RUN;
```

本例是一个五个因素各两个水平的二分之一析因设计，各主效应的估计与其它效应及两因素的交互无关。

下例是一个带有区组的部分析因设计：

```
PROC FACTEX;
   FACTORS x1 x2 x3 x4 x5;
   SIZE FRACTION=2;
   BLOCKS SIZE=MINIMUM;
   MODEL est={x1 x2 x3 x4 x5} nonneg=(x1|x2|x3|x4|x5@2);
RUN;
```

混合设计，以下产生一个4×2^3 的设计，即四个设计因素，一个拥用四个水平，三个具有两水平：

```
PROC FACTEX;
   FACTORS a1 a2 b c d;
   MODEL estimate=(b c d a1|a2)
            nonneg=(b|c|d@2 a1|a2|b a1|a2|c a1|a2|d);
   SIZE DESIGN=16;
   OUTPUT OUT=mixed [a1 a2]=a cvals={'A' 'B' 'C' 'D'};
RUN;
```

使用设计因素A1、A2 来构造导出因素A。

随机区组设计：

```
PROC FACTEX;
   FACTORS blocks /nlev=3;
   OUTPUT OUT=genblok blocks nvals=(1 2 3) randomize;
RUN;
   FACTORS trt/nlev=10;
   OUTPUT OUT=rcbd trt
          cvals=('A' 'B' 'C' 'D' 'E' 'F' 'G' 'H' 'J' 'K')
          designrep=genblok randomize;
RUN;
```

第一个FACTORS 语句产生用于设计的区组和包含水平编码的数据集GENBLOK。第二个FACTORS 产生“处理”因素，有十个水平，第二个OUTPUT 语句对GENBLOK 的设计进行重复。

拉丁方设计，下面是一个3 x 3 拉丁方的例子：

```
PROC FACTEX;
   FACTORS row col trt /nlev=3;
   SIZE design=9;
   MODEL res=3;
   OUTPUT OUT=latinsq ROW nvals=(1 2 3)
                      COL nvals=(1 2 3)
                      TRT cvals=('A' 'B' 'C');
RUN;
```

SAS 样本库程序ADXEG7.SAS，系一个纺织问题的研究，据Box, G.E.P., and Cox, D.R. "An Analysis of Transformations". JRSS B-26, pp. 211-243.

Box, G.E.P. and N.R.,Draper(1987)也引用了这个例子，转换步骤：1. 计算被转换数据的几何均值；2. 计算转换值；3.针对每个λ，用最小二乘法拟合最简捷模型$y = g(\xi, \beta) + \varepsilon$并记录剩余均方和S($\lambda$)；4 . 利用ln S($\lambda$) 与$\lambda$ 的图，使ln S(λ) 最小的λ值即是转换值；5. 求取λ的$100(1-\alpha)\%$可信区间。

现在，27个数据的几何均值是562.34，对于任何给定 λ 的转换公式是下式：

$$Y(\lambda) = \begin{cases} \lambda^{-1}(562.34)^{1-\lambda}(Y^{\lambda} - 1), & \text{if } \lambda \neq 0, \\ (562.34) lnY, & \text{if } \lambda = 0 \end{cases}$$

要拟合的模型是$g(\beta, \xi) = \beta_0 + \beta_1\xi_1 + \beta_2\xi_2 + \beta_3\xi_3$，剩余均方及其自然对数的取值如下：

```
   L    -1.0    -0.8    -0.6    -0.4    -0.2     0.0     0.2
 S(L)    3.9955  2.1396  1.1035  0.5478  0.2920  0.2519  0.4115
ln S(L) 1.3852  0.7606  0.0985 -0.6018 -1.2310 -1.3787 -0.8897
          0.4     0.6     0.8     1.0
          0.8178  1.5986  2.9978  5.4810
         -0.2011  0.4680  1.0979  1.7013
```

$\ln S(\lambda)$对λ的图示显示约在$\lambda = -0.06$ 时产生极小值。λ的95%可信区间由下式算得：

$\chi^2_{(1);0.05}$/剩余均方自由度=3.84/23=0.167

λ的取值范围是$-0.20 \sim 0.08$。

实验是一个3**3 设计，SAS 处理时首先调用ADXGEN.SAS 初始化，然后直接使用FACTEX构造设计并且打印出来。其次，对实验数据进行编码，产生二阶模型的交叉乘积及平方项，并且进行数据转换。

```
%inc 'sasmacro\adxgen.sas';
%adxinit
proc factex;
   factors len amp load / nlev=3;
   output out=yarn len  nvals=(250 300 350)
                   amp  nvals=(  8   9  10)
                   load nvals=( 40  45  50);
run;
%adxrprt(yarn,failcyc)
data yarn; set yarn;
   label len='length of specimins of yarn'
         amp='amplitude of loading cycle'
         load='load'
         failcyc='number of cycles to failure';
   format len amp load 20.4;
   input failcyc @@;
   output;
cards;
 674  370  292  338  266 210  170 118  90
1414 1198  634 1022  620 438  442 332 220
3636 3184 2000 1568 1070 566 1140 884 360
;
%adxcode(yarn,yarn,len amp load)
%adxqmod(yarn,yarn,len amp load,1)
%adxtrans(yarn,tranyarn,failcyc)
```

输出的设计结果、转换 λ 、均方误差、可信限。

OBS	LEN	AMP	LOAD	FAILCYC	ADXLAM	_RMSE_	ADXCONF
1	350	9	40	_____	-2.0	2713.50	
2	250	9	40	_____	-1.8	2125.08	
3	350	8	45	_____	-1.6	1684.89	
4	300	10	50	_____	-1.4	1355.22	
5	350	10	50	_____	-1.2	1108.55	
6	250	10	50	_____	-1.0	924.81	
7	250	8	45	_____	-0.8	789.48	
8	250	9	45	_____	-0.6	692.14	
9	300	10	40	_____	-0.4	625.52	

```
10    350     10     45     ________   -0.2      584.77       *
11    300      9     40     ________    0.0      566.99       *
12    250      8     40     ________    0.2      571.00       *
13    300     10     45     ________    0.4      597.18       *
14    250     10     45     ________    0.6      647.59
15    300      8     45     ________    0.8      726.19
16    300      9     50     ________    1.0      839.25
17    300      8     50     ________    1.2      996.03
18    300      9     45     ________    1.4     1209.70
19    300      8     40     ________    1.6     1498.71
20    350      8     40     ________    1.8     1888.63
21    250      9     50     ________    2.0     2414.89
22    350      9     45     ________ ------------------------------
23    350      8     50     ________
24    250      8     50     ________
25    350     10     40     ________
26    250     10     40     ________
27    350      9     50     ________
-----------------------------------
```

转换的结果，λ= -0.2，但由于其95 % 的可信区间中包含了$\lambda = 0$的情况，故使用对数转换。

§4.6 其它

SAS/OR 提供了运筹学工具，这里只给出NLP的用例，SAS/IML 有NLP函数，实现也很方便。

```
data lp(type=est);
input _type_ $ x1-x3 _rhs_;
cards;
PARMS   0.   0.  0.    .
LE     12.   5. 30. 120.
LE      2.  10. 30.  95.
LOWERBD 0.   0.  0.    .
UPPERBD 90. 90.  2.    .
;
PROC NLP TECH=TR INEST=LP OUTMOD=MODEL ALL;
   MAX Y;
   PARMS X1-X3;
   Y = x1 + 3. * x2 + 10. * x3;
 RUN;
```

```
/*
IML NLP: Rosenbrock Function as an Optimization Problem
The two-dimensional Rosenbrock function is defined as:
f(x) = 1/2 { 100 (x[2] - x[1]**2)**2 + (1 - x[1])**2 }
*/
   proc iml;
   start F_ROSEN(x);
      y1 = 10. * (x[2] - x[1] * x[1]);
      y2 = 1. - x[1];
      f  = .5 * (y1 * y1 + y2 * y2);
      return(f);
   finish F_ROSEN;
   start G_ROSEN(x);
      g = j(1,2,0.);
      g[1] = -200.*x[1]*(x[2]-x[1]*x[1]) - (1.-x[1]);
      g[2] =  100.*(x[2]-x[1]*x[1]);
      return(g);
   finish G_ROSEN;

/*
The minimum function value
f* = f(x*) = 0 is at the point x* =  (1,1).
The trust region algorithm NLPTR is shown in this  example,
but other subroutines can be used for the  minimization:
*/

   x = {-1.2 1.};
   optn = {0 2};
   CALL NLPTR(rc,xres,"F_ROSEN",x,optn, , , , ,"G_ROSEN");
   quit;
```

第五章　SPSS

§5.1　SPSS/PC+ 导引

§5.1.1　简介

SPSS 由美国斯坦福大学1965 年开始研究并于1970 年推出。SPSS- X 用于IBM CMS、MVS/TSO、UNIX 和DEC VAX/VMS 系统，允许用户以批处理方式运行。微机版SPSS/PC+V2.0，由Chicago 为基础的SPSS 公司于1987 年推出。SPSS 广泛用于商务、政府部门、教学与科研单位进行调查分析、市场研究、产品检测、人事管理与决策、卫生服务分析以及统计质量控制等。

本章主要介绍SPSS/PC+，其功能概要如下：

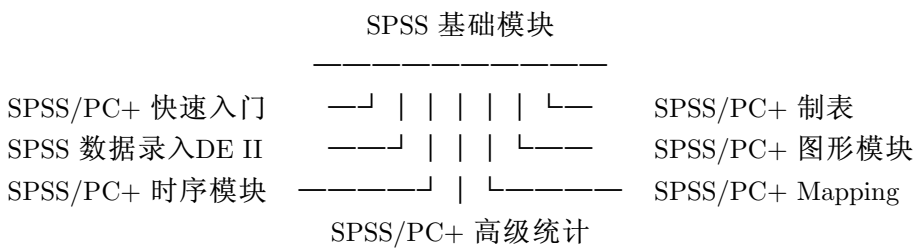


图 5.1　SPSS/PC+ 功能示意图

它支持的数据格式有ASCII、dBASE II-IV、Lotus 1- 2- 3 、symphoney 、mutiplan、及SPSS-X 传输格式。SPSS/PC+提供与其它图形软件的接口。其graph -in-the-box 允许用户在SPSS/PC+ 环境下产生、浏览和修正图形；SPSS/PC+ Mapping 支持Ashton-Tate MAP-MASTER。

SPSS/PC+ 高级统计拥有多元方差分析(MANOVA)、判别分析、因子分析、聚类分析、对数线性模型、非线性回归、logistic回归分析以及可靠性分析(、对应分析等。SPSS/PC+的多分类分析(Multiple Classification analysis, MCA) 一般统计分析软件不专门具备。

时间序列分析是利用它的Trend部分，其功能有指数平滑、曲线拟合、特定形式的回归、ARIMA 模型(Box-Jenkins)、谱分析。

§5.1.2　SPSS/PC+ 工作方式

在DOS 系统下，设软件存放于目录SPSS，打入命令：CD SPSS ¡Enter¿，这时可用三种方式运行SPSS/PC+。

㈠SPSS/PC+ 菜单方式

执行命令：C:\SPSS>SPSSPC <Enter>

屏幕显示：

这时打入菜单上的英文，则光标移至相应的项目，<Enter>键选择后，进入下一层菜单，再打入子菜单英文名，用<Enter>继续选择。

利用光标键移动时，每项下有相应的解释，右箭头由上级菜单向下级菜单推进，到达待选命令或选项时，以回车选定项目，如果不满意可以使用Alt-D 删除；左箭头令光标返回上级菜单。在菜单方式下，使用Alt-E 键进行编辑态，就可以编辑程序或插入外部程序(F3)，打入SPSS/PC+的关键字后再用Esc 键则系统立即调出键入命令有关信息。打入F10，系统提示下一步的运行方式，同第7章将介绍的SYSTAT 一样，SPSS/PC+ 可以从光标位置开始运行。

orientation	入门
read or write data	读写数据
modify data or files	修改数据或文件
graph data	数据绘图
analyze data	分析数据
session control& inf	运行控制和信息
run DOS or other pgms	运行DOS或其他程序
extended menus	扩展菜单
SPSS/PC+ options	选项
FINISH	完成
	F1=帮助Alt-E=编辑Alt-M=菜单开/关

图 5.2 SPSS/PC+ 主菜单

㈡行命令方式

在菜单方式下，使用Alt/SHIFT-F10 进入对话方式，系统使用提示SPSS/PC:，此时每输入一条命令，都被立即执行，执行结束后仍然返回提示下。每行命令以圆点(.) 结束，当一行写不完时，可用<Enter> 在下一行继续输入。这种运行方式适于程序的调试。

㈢批处理方式

在MS-DOS 系统下相应的批处理方式是：SPSSPC ¡程序文件名¿ ¡Enter¿。除非特别指定(SET LOG/LISTING)，程序的运行情况和结果分别存于SPSS. LOG 和SPSS.LIS 中。

现运行系统提供的基础模块检验程序BASETEST.INC：使用命令为：

C:\SPSS>SPSSPC BASETEST.INC <Enter>

同时处理多个程序，可把它依次写在SPSSPC 后面即可，程序之间用空格分开。

在SPSS/PC+ 提示下使用INC ”程序名”. ¡Enter¿，即交互下的批处理方式。也可以用@程序名.

SPSS/PC+ 通过执行命令FINISH、STOP 或BYE、EXIT 返回DOS 系统。使用DOS ¡DOS 命令¿ 运行DOS 命令或仅仅使用DOS. 命令时进入DOS 内核，返回外壳SPSS/PC+ 仍然用EXIT 命令。

在行命令方式下，使用REVIEW SCRATCH. 或REVIEW. 命令返回系统菜单，使用REVIEW LOG, REVIEW LISTING, REVIEW BOTH, REVIEW FILENAME 也都合法。SPSS/PC+ 菜单方式下，有关的功能键列表如5.3(a)-(b)：

编辑态的的几种功能可单独使用功能键或通过Ctrl、Alt 与功能键或字母的组合来完成，有些功能与不同的编辑状态有关，如F3仅当插入态能用。SPSS/PC+ 可以标记系统命令，这种做法很有特色，另外，SPSS/PC+可以对输出结果的小数位进行四舍五入。做法是先做标记，然后打Ctrl-F7 系统就要提示需要的小数位数。SPSS/PC+关键字的在线帮助可经Alt-G来完成，可在使用Alt-E并移动光标至命令字处，此时使用Alt-G则调出相应的字汇解释，Alt键与一些字母组合的功能如下：

信息	F1	Review 帮助和菜单，变量和文件列表，专用词汇表
窗口	F2	切换，大小，缩放
输入文件	F3	插入文件，编辑其它文件
行	F4	插入，删除，恢复
搜寻替换	F5	文本查找，替换文本
跳转	F6	区域，输出页，错误行，最未执行行
定义区域	F7	标记/取消行标记，矩形标记或命令标记
区域操作	F8	拷贝，移动，删除，数据操作，拷贝专用词项
输出文件	F9	写标记区或文件，文件删除
运行	F10	从光标或标记区运行，退至命令行提示下

图 5.3 review 功能键

ENTER	剪贴选择以及在菜单下移一个水平
TAB 或→	暂时剪贴选择以及菜单下移一个水平
ESC 或←	最未一个暂存剪贴上移一个水平
Alt-ESC	到主菜单(同Ctrl-ESC)
Alt-K	删除所有剪贴暂存区
Alt-T	进入录入窗口
Alt-E	编辑态切换
Alt-M	关闭/启用菜单
Alt-V	进入变量窗口
Alt-C	自光处标运行

图 5.4 主要菜单命令

Alt-B 向前插入一行
Alt-D 删除一行，不论是否为编辑态
Alt-F 磁盘文件列表
Alt-G 专用词汇表
Alt-H 开启/关闭帮助窗口
Alt-I 向后插入一行
Alt-P 块写文件
Alt-R REVIEW 帮助
Alt-S 窗口切换
Alt-U 删除恢复(UNDELETE)
Alt-W 写文件
Alt-X 标准菜单与扩展菜单切换，后者专门存放一些不太常用的信息
Alt-Z 窗口缩放切换，放大时可进行全屏幕编辑
如对系统命令比较熟悉，则进入SPSS/PC+后直接用Alt-Z进行全屏幕编辑。

§5.1.3 系统装卸

SPSS/PC+ 的“运行控制和信息”项下的SPSS MANAGER 命令完成。STATUS 用于显示当前的安装情况，INSTALL 指定安装，REMOVE 指定删除。

§5.2 SPSS/PC+ 语言

§5.2.1 语言要素

1. 表达式。常用操作符：
算术运算符+、-、*、/、** 分别对应加、减、乘、除、乘方等运算。逻辑操作符=、<>, <=, <, >, >=运算符，及Fortran中的EQ、GT等。关系运算符：all、by、and、not、or、to、with。

2. 函数。
ABS(绝对值) RND(四舍五入) TRUNC(取整)
MOD10(对10 取模) SQRT(平方根) LG(常用对数)
LN(自然对数) SIN(正弦) COS(余弦)
ARTAN(反正切) UNIFORM(0～x 间的均匀分布随机函数)
NORMAL(均值为零、标准差为x 的正态分布随机数)
LAG(函数取前一个变量的值赋给命名量)
YRMODA 是一个时间函数，把年月日转成天数。

在SPSS for Windows中函数的种类大大增多。

3. 语句。SPSS/PC+ 命令由关键字和说明部分组成，命令关键字告诉系统进行哪一种操作，说明部分也就是命令参数(命令对象和选择项)，即：

命令关键字＋命令对象＋命令选项＋命令结束符(.)

命令对象明确对变量表、表达式或文件进行操作。命令选项一般应予使用，分必选项和可选项两种。

如：GET /FILE='NEW.SYS' /KEEP=AGE. 关键字是GET，其余为说明部分。

约定文件名、变量名不超过8个字符，不能有空格，首字母必须是字母，文件名应用引号括起来。定义变量名时可以使用关键字TO，如定义变量X1,..., X10 可用X1 TO X10 表示。命令说明部分的关键字一般是保留关键字，如TO、LT 、NOT、LOWEST、THRU、HIGHEST等。变量值可以是数值型或字符型，数值型值可以是整数或小数，字符型必须用引号括起来。标号是对变量或变量值的说明。两个元素或变量之间一般用空格或逗号分隔，函数符号与自变量用圆括号分隔、子命令关键字与参数用等号分隔、子命令之间用斜杠分隔。

SPSS/PC+ 命令大致归类：系统命令，包括安装、参数设定、显示、列表等命令；数据定义命令，包括变量的产生、修改、分组调整等命令；文件管理、合并等命令；统计分析命令。频数表、列联表、方差分析、多元分析等；其它命令。SPSS/PC+命令可以(组合)简写：如：COM(COMPUTE)、REC(RECODE)、DATLIS(DATA LIST)、详见文件SPSSV4.TBL。

§5.2.2 数据和文件管理

SPSS/PC+ 的文件有活动文件(active file)、数据文件(data file)、系统文件(system file) 、引用文件(include file)、列表文件(list file)与工作文件(working file)几种。

活动文件是系统运行时所使用的文件，包括数据和数据结构，它们在退出系统后消失。数据文件一般是ASCII 文件，由DATA LIST 读取。系统文件是SPSS/PC + 内部文件，可以使用GET、JOIN、SAVE、AGGREGATE 命令进行操作，扩展名一般取为.SYS。引用文件使用扩展名.INC 及.LOG。由INCLUDE 命令引用。列表文件使用扩展名.LIS，存放执行的结果，如SPSS.LIS。工作文件由SPSS/PC+ 运行时使用的暂存文件，使用扩展名.SY1、.SY2。

活动数据文件经命令DATA LIST、IMPORT 或GET 生成，可使用ALL 通配文件中的所有变量。SPSS/PC+ 最多可用变量数为200，变量名不超于8个字符。

SPSS/PC+ 对数据的管理命令可大致分为数据的定义命令、数据转换命令、记录操作、变量操作(MODIFY VARS) 和文件操作命令。在文件读写和转换中，许多参数是相同的，/KEEP 表示保留变量, /DROP 表示删除变量，/RENAME() 中含有重新命名的变量名表，有关命令及格式可详见第16章。

§5.2.3 运行控制

SPSS/PC+ 的系统设置命令为SET，有关的设置内容：如SET LISTING='CHINA. LIS'. 设置运行结果存于文件CHINA.LIS 中。当前的设置状态可用命令SHOW 来显示。运行其它软件系统命令使用DOS 和EXECUTE，如：DOS DIR。

与SAS 软件件相比，SPSS/PC+ 没有严格区分出数据步和过程步，故它的运行控制、数据管理、数据分析功能可在同一程序中灵活出现，只有少数例外。

【例5.1】现对系统检验程序BASETEST.INC的数据处理过程进行简单说明，所分析的数据是某公司雇员情况的一个调查。使用命令INC '\basetest.inc' 运行。

```
SHOW.
DATA LIST /MOHIRED YRHIRED 12-15 DEPT79 TO DEPT82 SEX 16-20
      /SALARY79 TO SALARY82 6-25 AGE 54-55 RAISE80 TO RAISE82 56-70
      /JOBCAT 6 EMPNAME 25-48 (A).
DISPLAY.
```

```
MISSING VALUES DEPT79 TO SALARY82 AGE (0) RAISE80 TO RAISE82 (-999)
               JOBCAT (9).
VAR LABELS AGE 'Age in years'.
VALUE LABELS SEX 1 'Male' 2 'Female'/
             JOBCAT 1 'Stk Clk' 2 'Admin' 3 'Sales' 4 'Mgr'.
COMPUTE GRPAGE = AGE.
RECODE GRPAGE (low THRU 25=1) (26 THRU 30=2) (31  THRU  39= 3)
              (40 THRU 49=4) (50 THRU HI=5).
VALUE LABELS GRPAGE 1 'Low - 25' 2 '26 - 30' 3 '31 - 39'
                    4 '40 - 49' 5 '50-High'.
DISPLAY GRPAGE JOBCAT.
FREQUENCIES  VARIABLES=AGE GRPAGE /FORMAT=LIMIT(10) /HBAR NORMAL
             INCREMENT(4).
DES SALARY79 TO SALARY82.
 CROSSTABS DEPT82 BY GRPAGE BY SEX/ CELLS = COLUMN NONE
            /STATISTICS = CHISQ GAMMA.
SORT CASES BY EMPNAME.
PLOT HORIZONTAL='Raise in 1982' MIN(0)/VERTICAL=MIN(0)
                  /SYMBOLS=' '/PLOT RAISE81 WITH RAISE82.
PROCESS IF (GRPAGE = 5).
LIST VARIABLES = EMPNAME SALARY80 RAISE80.
SAVE FILE='TEST.SYS'.
SORT CASES BY grpage.
TRANSLATE to DBASE4.DBF/ type=DB4/map /REPLACE.

REPORT /VARIABLES salary79 to salary82 (label)
    /BREAK grpage '年龄分组' (LABEL)
    /SUMMARY MEAN '平均值' /summary STDEV '标准差'
    /summary MINIMUM '最小值' /summary MAXIMUM '最大值'
    /summary KURTOSIS '峰度'
    /title='SPSS/PC+ BASETEST.INC 运行结果'.
```

SHOW命令显示运行环境，用DATA LIST 命令准备数据，用DISPLAY 命令显示数据结构。使用MISSING VALUE 进行缺失值定义。利用RECODE命令对年龄分组。用FREQUENCIES显示数据的频数分布。利用DESCRIPTIVES命令获得描述数据的综合统计量。使用CROSSTABS命令计算列联表统计量，从而与SAS的PROC FREQ过程对应，与SAS PROC TABULATE 对应的命令是REPORT。

使用DATA LIST命令创建的原始数据，必须放在BEGIN DATA/END DATA. 命令之间，这与SAS的"CARDS;"类似，所有与DATA LIST 有关的变量定义和说明都应放在BEGIN DATA之前。其它操作有排序、用PLOT命令图示，使用PROCESS IF 命令选择部分数据处理。结果存贮。用REPORT 命令产生报表。

利用系统提供的TRANSLATE命令，把系统文件转换成.DBF格式，为了保持数据的格式，使用FORMAT命令，如：FORMAT engcc (comma7.1).

最后，以FINISH程序结束。

§5.3 描述统计

指统计指标的计算、频数表和直方图、列联表分析，对应命令DESCTIPTIVES、FREQUENCIES和CROSSTABLES 命令，此处略作介绍。

§5.3.1 DESCRIPTIVES

格式：DESCRIPTIVE VARIABLES=变量表/STATISTICS=N /OPTION=N.

意义：显示不带频数表的描述统计量，其中的STATISTICS 和OPTION 子命令在许多命令中出现，可借助其菜单提示选择，故在下面介绍中，一般不再列出。

/VARIABLES 指示分析变量名。分析变量后加括号，可引入记录Z-值的新变量。

/STATISTICS 后加号码指示输出的统计量，省略时，SPSS/PC+ 输出均值、标准差、最小值和最大值。指示了/STATISTICS 时，仅仅得到所要求的统计量，其内容如下：

1 均数(MEAN)

2 标准误(SEMEAN)

5 标准差(STDDEV)

6 方差(VARIANCE)

7 峰度(KURTOSIS)及其标准误(SEKURT)

8 偏度(SKEWNESS)及其标准误(SESKEW)

9 全距(RANGE)

10 最小值(MINIMUM)

11 最大值(MAXIMUM)

12 观测值之和(SUM)

13 默认(DEFAULTS)统计量：均值、标准差、最小值和最大值选择13 和其它项的组合得到默认统计量和其它统计量。

ALL 上述所有统计量。

/OPTION 指示缺失值的处理方法，默认情况下使用所有有效的记录。/OPTION 的几种选项含义说明如下：

1 包含了用户所指定缺失值的记录也参加计算，命令仅当使用了MISSING 命令后方有效。

5 使用"listwise"方法排除含有缺失值的记录，在DESCRIPTIVE 命令中的任何一个变量的值出现缺失时，这个记录就废弃不用。

3 对于/VARIABLES 中所有变量在活动文件中增加Z-值，新变量取名为Z 和原始变量名中的头七个字符。

用例：

DESCRIPTIVES /VARIABLES score1 score2 score3.

DESCRIPTIVES /VARIABLES age var1 to var5 income /OPTIONS 3 5 /STATISTICS ALL.

一般说来，缺失值处理方法中LISTWISE 指示仅仅使用各个分析变量均有效的哪些记录；PAIRWISE 删除一对变量有缺失值的哪些记录；INCLUDE 括入含有缺失值的哪些记录；MEANSUBSTITUTION 用均值代替含有缺失的资料。

§5.3.2 FREQUENCIES

显示频数表、统计量、条图和直方图，其格式为：

FREQUENCIES /VARIABLES=变量列表

/FORMAT=LIMIT(N)—ONEPAGE表格输出格式

/HISTOGRAM=INCREMENT NORMAL... 直方图

/BARCHART 条图

/HBAR=INCREMENT NORMAL... 直条图

/STATISTICS 统计量选择.

其它子命令有：/GROUPED、/PERCENTILES、/NTILES、/STATISTICS、/MISSING =INCLUDE。

/FORMAT 子命令中的LIMIT(N)表示当分组数超于N时不显示表格，ONEPAGE指示将一个大的表压缩到一责内显示。/HISTOGRAM 子命令中的INCREMENT(N) 表示纵轴的间隔尺度，NORMAL表示根据本变量的均值与标准差画正态曲线。

STATISTICS 可指示均值、中位数、标准差、偏度、极差、峰度、最大值、最小值、标准误、众数、方差、偏度标准误、峰度标准误及总和，默认内容为均值、标准差、标准误，ALL将显示所有统计量。

例：FREQUENCIES /VARIABLES sex race dept.

FREQUENCIES /VARIABLES systolic diastol hemoglob

/STATISTICS MEAN SEMEAN MEDIAN MINIMUM MAXIMUM.

FREQUENCIES /VARIABLES height weight /FORMAT NOTABLE

/HISTOGRAM NORMAL.

§5.3.3 CROSSTABS

用交叉制表的方式显示变量的分布，进行关联度量，其格式为：

CROSSTABS 变量表BY 变量表...

/TABLES=变量表/OPTIONS=选项表

/FORMAT=格式定义

/CELLS=格子统计量

/STATISTICS=列联表统计量表

/MISSING=TABLE—INCLUDE—REPORT 指定缺失值处理方式

/WRITE=NONE—CELLS—ALL 结果文件.

其中的BY 可最多有十层。

格子统计量内容有：COUNT、ROW、COLUMN、TOTAL、EXPECTED、RESID、SRESID、ASRESID、ALL、NONE

列联表统计量有χ^2值、列联表系数、λ 统计量、Kendall相关等，其关键字为：CHISQ、PHI、CC、LAMBDA、UC、B
、CORR 、KAPPA、RISK、ALL、NONE。

/FORMAT 子命令选项有：AVALUE、DVALUE、LABELS、NOLABELS、NOVALLABS
、INDEX、NOINDEX、TABLES、NOTABLES、BOX、NOBOX、

CROSSTABS /TABLES= vote BY sex /STATISTICS= CHI LAMBDA.

CROSSTABS /TABLES= educatn test BY sex agegroup race.

§5.3.4 PLOT

统计做图，其命令格式为：

PLOT /FORMAT /TITLE ' ' /VERTICAL /HORIZONTAL /VSIZE /HSIZE /MISSING
~/PLOT. 如：

PLOT /PLOT Y WITH X.

PLOT /FORMAT REGRESSION /PLOT sales WITH advertis.

PLOT /PLOT income WITH age BY sex.

经GRAPH 绘图程序制记扇形图、直条图、线图、直方图和散点图并把它们传递到绘图软件中，默认是Harvard Graphics。

§5.3.5 其它命令

使用以下命令进行描述报告和报表(reports and tables)：

1. LIST 给出数据列表。

2. REPORT 产生综合统计量的报告和观察列表，它有/FORMAT、/MISSING 、/TITLE、/FOOTNOTE、/BREAK
' '几个子命令。

若指定/FORMAT ，则它应在程序中应首先出现，它有几个函数：AUTOMATIC 提供格式化选项的默认值，LIST 指示按记录列表。其它关键字控制报告各部分间的留空，大部分的扩充菜单中出现。

/VARIABLES 是必选项，对“报告变量”进行命名，每个变量的报告中定义一列。可用(VALUE)、(LABEL)、(DUMMY) 等选项指示显示的内容。注意：含有缺失值的记录在记录列表中出现，但综合统计量计算时不被包括。

/MISSING 子命令指示用户包括定义的缺失值、若超过某个指定的界值，则记录彻底被剔除。

/TITLE 子命令给输出的每页显示一个标题。可以指定LEFT、CENTER或RIGHT 使之左齐、居中、右齐或不指定。

/FOOTNOTE 子命令则用于指定每页的脚注，其选项与/TITLE 相同。

/BREAK 子命令指定一个或多个分组变量，其选项有(NOBREAK)、(TOTAL) 、(VALUE)、(LABEL)
等。

/SUMMARY 子命令指定统计量的名称，当指定/FORMAT LIST时不使用。

/OUTFILE 子命令指定结果到其它的文件中。

3. EXAMINE 提供了茎叶图、盒式图、位置的稳健估计、正态性检验以及其它描述统计量和图形，分组分析也是可能的。

其较为重要的子命令是/MESTIMATOR，可以计算M-估计量，也即位置的稳健极大似然估计。SPSS/PC+ 计算的有四个：Huber's M-估计量、Andrew's wave 估计、Hampel's M-估计、Tukey's biweight 估计量。

例：EXAMINE /VARIABLES=engsize, cost.

EXAMINE /VARIABLES=mipergal BY prototyp,prototyp BY pistons.

EXAMINE /VARIABLES=yield weevils BY field

| /COMPARE=GROUPS/PLOT=SPREADLEVEL(.5).

其图形分析包括：STEMLEAF、BOXPLOT、NPPLOT、SPREADLEVEL、HISTOGRAM、ALL、NONE，其义自显。

4. TABLES 产生高质的stub-and-banner 表，SPSS/PC+程序TBLTEST.INC 是一个用例，此处不介绍。

5. PRINT TABLES, 把TABLES 的输出在各种打印机打出。

命令使产生的表格得到尽可能高的质量的打印效果，子命令/DEVICE 指定输出设备名，使用/PORTRAIT、/LANDSCAPE、/PICA、/ELITE、COMPRESSED 定义打印特性。支持的设备如：GIBM(IBM 图形打印机)、OTHER、PIBM、HPLASER、FXEPSON、RXEPSON、LQEPSON、93OKIDATA、92OKI、TEKTRONIX。

【例5.2】不同学历的调查对象对总统能力的评价[1]，变量EDUC 表示调查对象受教育的程度，取值1,2,3,4对应高中以下、高中、大学、研究生；变量RATING 表示对总统能力的打分，取值1,2,3,4对应差、尚可、好、很好。

```
data list free/educ rating count.
TITLE 'The Performance of President'.
variable labels educ 'Education' rating 'Rating scores'.
value labels educ 1 'Less than HS' 2 'HS degree'
                  3 'College' 4 'Post graduate' /
           rating 1 'Poor' 2 'Fair' 3 'Good' 4 'Excellent'.
begin data.
1 1 4   1 2 9   1 3 10  1 4 7
2 1 5   2 2 8   2 3 22  2 4 14
3 1 20  3 2 31  3 3 11  3 4 12
4 1 9   4 2 6   4 3 8   4 4 4
end data.
WEIGHT BY count.
CROSSTABS /TABLES= educ BY rating /STATISTICS BTAU CTAU CORR GAMMA.
```

Statistic	Value	ASE1	T-value	Approximate Significance
Kendall's Tau-b	-.20098	.06158	-3.26920	
Kendall's Tau-c	-.19416	.05939	-3.26920	
Gamma	-.27112	.08218	-3.26919	
Pearson's R	-.21816	.07180	-2.98249	.00326
Spearman Correlation	-.23747	.07206	-3.26149	.00133

本例数据属于有序的情形，针对H_0: ρ =0即相关为零的假设，首先检查一致(concordance, C)与不一致(discordance, D)的观察对子，本例C= 3171 ，D= 5530，C-D 反映了两变量关联的方

向，gamma(ɣ)=(C-D)/(C+D)=(3171- 5530) /(3171+5530)=-0.27，然后计算统计量z=ɣ/SE(ɣ)=-0.27/0.082=-3.29，有显著意义。由第二章的介绍，Kendall 相关系数的含义与此相仿。

SPSS/PC+ 的CROSSTABS 语句提供了Kappa统计量，它用于比较两种评判方法的一致性，因此当列联表的行列数相同时方可得到。

设有30名受试者，其行为分为无问题、内向、外向，评判两次，现考察评判者的一致性。一致的数目是(15+3+3)=21，占21/30x100题”的方向偏。第一次有：16/30=0.53，第二次有：20/30=0.67。若两次评判是独立的，就有0.53x0.67=0.36，即应有30x0.36=10.67个“无问题”。

		JUDGE2			
	Count				
	Exp Val				
					Row
		1.00	2.00	3.00	Total
JUDGE1					
	1.00	15	2	3	20
		10.7	4.0	5.3	66.7%
	2.00	1	3	2	6
		3.2	1.2	1.6	20.0%
	3.00	0	1	3	4
		2.1	.8	1.1	13.3%
	Column	16	6	8	30
	Total	53.3%	20.0%	26.7%	100.0%

Kappa修正的思想是：假设评判独立，则对角元可以用通常的期望卡方。公式为：Kappa=(ΣO-ΣE)/[N-ΣE), 其中O表示观察数目，E是理论数目。现在ΣO=21，ΣE=10.67+1.2+1.07=12.94，Kappa=0.4。修正是分子分母各除去只是偶然引起的一致，Cohen已经造了界值表，若一致性足够低，判断就值得怀疑。SPSS/PC+ 程序如下：

```
data list free /judge1 judge2 count.
begin data
1 1 15   2 3  2
1 2  2   3 1  0
1 3  3   3 2  1
2 1  1   3 3  3
2 2  3
end data.
weight by count.
CROSSTABS /table=judge1 by judge2 /STATISTICS=CHISQ KAPPA
          /CELLS= COUNT EXPECTED.
```

计算结果：Kappa=.47266, ASE1=.13615, T-value=3.68100。

进一步的讨论可以参考：Cohen, J.(1960). A Coefficient of agreement for nominal scales. Educational and Psychological Meansurement, 100,37 -46.

§5.4 统计检验

§5.4.1 t-TEST

1.成组t-检验

T-TEST GROUPS=分类变量(k1, k2) /VARIABLES=被检验变量名(K) /OPTIONS=N.

仅指示一个K时将按照K 分成两类，指示K1, K2 则是按K1 和K2 分成两类。OPTION=1 时括入缺失值的记录，OPTION=2 删除缺失值记录，OPTION=3 不显示变量说明。

2.配对t-检验

T-TEST PAIRS=变量1 [WITH 变量2] [/PAIRS=] 变量...] [/OPTION=N].

OPTION=1,2,3 时含义与T-TEST 相仿，OPTION=4 时表示WITH 前后的量一对一比较。

3. 样本与总体的检验

现进行总体均值(POPM) 为200 的检验，则可使用语句：

COMPUTE POPM=200.

T-TEST PAIRS=A M.

§5.4.2 MEANS

MEANS 计算均值与标准差、变量总和、方差，进行单因素方差分析，格式为：

MEAN 变量表[BY 变量表] [/STATISTICS N] [/OPTION N].

在STATISTICS=1 时进行完全随机的方差分析。结果中的eta 统计量主要适于因变量为等级变量而自变量为连续性的资料。eta 平方值反映了通过已知自变量值所解释的因变量总变异的比例大小。

§5.4.3 ONEWAY

进行单因子的方差分析，其格式为：

ONEWAY 变量BY 变量[/RANGE=范围] [/OPTIONS=选项] [/STATISTICS] [/CONTRAST] [/POLYNOMIAL].

/OPTIONS 指示缺失值的处理方式、变量标签等。

/STATISTICS 指示各组的描述统计量、固定效应和随机效应统计量以及方差齐性检验。

/POLYNOMIAL 子命令是将组间平方和分解成多项式。

例：ONEWAY Y BY X(1,2) /POLYNOMIAL=2.

将平方和分解成2次多项式。

均值间的两两比较方法有SNK, Scheffe, LSD, Duncan, BTurkey, Turkey, MODLSD。POLYNOMIAL 表示平方和对应的多项式的次数

方差分析命令ANOVA 的用法与ONEWAY 类似。

【例5.3】五种方法(method)对延迟黄油变质(spoilage)的作用[1]，前两种方法属一类，后两种方法属一类，最后一种作为对照。记五种方法下，效果的均值分别为 μ 1, μ 2, μ 3, μ 4, μ 5，现比较方法的作用以及两类方法效果一样吗？H_{01}:(μ 1+ μ 2)/2=(μ 3+ μ 4)/2, 即 μ 1+ μ 2+ μ 2+ μ 4=0,

第二个检验是前四种方法与最后的对照方法进行比较, 即H_{02}:(μ 1+ μ 2+ μ 3+ μ 4)= μ 5 或0.25(μ 1+ μ 2+ μ 3+ μ 4)- μ 5=0。

现使用单因素方差分析和Kruskal-Wallis 非参检验。

```
set width=80.
title 'Comparison of five methods to retard spoilage of magarine'.
data list free/method spoilage.
begin data.
1 28   2 30   3  7   4  23   5  52
1 37   2 19   3 16   4  23   5  42
1 43   2 20   3 23   4  30   5  38
1 31   2 18   3 11   4  20   5  54
end data.
oneway spoilage by method(1,5)
       /range=LSD /range=Tukey /range=Duncan
       /contrast=.5 .5 -.5 -.5 0
       /contrast=.25 .25 .25 .25 -1.
npar tests k-w spoilage by method(1,5).
finish
```

其中NPAR TESTS K-W 是用非参数方法进行比较，运行结果如下：

```
                    Analysis of Variance
                         Sum of         Mean          F      F
        Source   D.F.    Squares       Squares      Ratio   Prob.
Between Groups    4     2526.5000      631.6250   15.7578  .0000
Within Groups    15      601.2500       40.0833
Total            19     3127.7500
```

对比系数和方差估计：

```
Contrast  1    .5    .5   -.5   -.5    .0
Contrast  2    .3    .3    .3    .3  -1.0
                            Pooled Variance Estimate
                 Value   S. Error   T Value    D.F.       T Prob.
Contrast  1     9.1250     3.1656     2.883    15.0          .011
Contrast  2   -22.8125     3.5392    -6.446    15.0          .000
                            Separate Variance Estimate
                 Value   S. Error   T Value    D.F.       T Prob.
Contrast  1     9.1250     2.9660     3.077    10.8          .011
Contrast  2   -22.8125     4.1371    -5.514     3.9          .006
```

两两比较结果，程序给出在0.05水平下三种检验的界值：

```
 LSD    3.01   3.01   3.01   3.01
```

```
HSD    4.37   4.37   4.37   4.37
Duncan 3.01   3.16   3.26   3.31
```

第J个均值Mean(J) 与第I个均值Mean(I)的差是与下面的量比较：4.4768 * Range * Sqrt(1/N(I) + 1/N(J)) 星号(*)表示两组在0.05水平上有显著差异。

```
                        LSD         HSD          Duncan
Mean      Group      3 2 4 1 5   3 2 4 1 5    3 2 4 1 5
14.2500   Grp 3
21.7500   Grp 2
24.0000   Grp 4      *
34.7500   Grp 1      * * *       *            * * *
46.5000   Grp 5      * * * *     * * *        * * * *
```

均值为一致的子集(Homogeneous Subsets)划为一组，则最高与最低均值的差不超于相应样本下的最短距离。三种检验的结果用组号表示是：

LSD法　子集一：3,2；　子集二：2,4；　子集三：1；　子集四：5。

HSD法　子集一：3,2,4；　子集二：2,4,1；　子集三：1,5。

Duncan法　子集一：3,2,4；　子集二：1；　子集三：5。

可见的确HSD法不容易出现显著。

下面是Kruskal-Wallis 检验结果：

```
Mean Rank    Cases
    14.50        4   METHOD =    1
     7.00        4   METHOD =    2
     3.75        4   METHOD =    3
     9.25        4   METHOD =    4
    18.00        4   METHOD =    5
                --
                20   Total

                                          Corrected for Ties
   CASES    Chi-Square  Significance    Chi-Square  Significance
      20       15.0429          .0046       15.1110          .0045
```

方差分析与非参检验证明五种保存方法之间的差别有统计意义，非参检验的效率要较通常的F-检验略低。

【例5.4】活产数与初婚年龄及教育程度的关系分析。

	文化程度 (例数)		
初婚年龄	高	中	低
15-19	4.17 (518)	3.65 (888)	3.27 (24)
20-22	3.70 (231)	2.95 (643)	2.88 (322)
23-24	3.60 (21)	2.12 (300)	2.68 (309)
25-34	3.15 (10)	2.68 (134)	2.45 (476)

其分析程序如下：

```
set more off length=100.
```

```
data list free /agegrp educat children count.
title '活产数与初婚年龄及文化程度的关系'.
var labels agegrp   '初婚年龄' educat '文化程度'.
val labels agegrp 1 '15-19' 2 '20-22'
                  3 '23-24' 4 '25-34'/
          educat 1 '低'  2 '中'  3 '高'.
begin data.
1 1  4.17  518  1 2  3.65  888  1 3  3.27  24
2 1  3.70  231  2 2  2.95  643  2 3  2.88  322
3 1  3.60  21   3 2  2.12  300  3 3  2.68  309
4 1  3.15  10   4 2  2.68  134  4 3  2.45  476
end data.
weight by count.
anova children by agegrp(1,4) educat(1,3) /statistics 1.
```

计算结果：

活产数与初婚年龄及文化程度的关系

* * * A N A L Y S I S O F V A R I A N C E * * *

CHILDREN

BY AGEGRP 初婚年龄

EDUCAT 文化程度

Source of Variation	Sum of Squares	DF	Mean Square
Main Effects	1463.284	5	292.657
AGEGRP	615.731	3	205.244
EDUCAT	225.004	2	112.502
2-way Interactions	70.140	6	11.690
AGEGRP EDUCAT	70.140	6	11.690
Explained	1533.424	11	139.402
Residual	.000	3864	.000
Total	1533.424	3875	.396

CHILDREN

By AGEGRP 初婚年龄

EDUCAT 文化程度

```
Grand Mean =  3.162

                                          Adjusted for
                              Unadjusted Independents
Variable + Category      N    Dev'n  Eta Dev'n   Beta
AGEGRP
    1 15-19           1430      .67        .58
```

```
    2 20-22          1196     -.09       -.08
    3 23-24           630     -.72       -.62
    4 25-34           620     -.65       -.55
                                    .90          .77
EDUCAT
    1 低              780      .84        .50
    2 中             1965     -.04       -.14
    3 高             1131     -.51       -.10
                                    .74          .40
Multiple R Squared                               .954
Multiple R                                       .977
```

总均值(3.162) 加上分组变量和协变量调整值就是所得的调整结果，随着年龄的增加或教育程度的增加，平均子女数向下调整(原表23-24岁年龄组中、高文化的影响略为不同)。eta是自变量对因变量变异的解释程度，即自变量引起的方差占原方差的百分比；beta 是控制其它因素下的影响，其值越大则自变量对因变量的影响越大，本例年龄的影响要大一些。

§5.4.4 CORRELATIONS

计算Pearson 相关系数，进行相关分析，其格式为：

CORRELATION VARIABLES= 变量表1 [WITH 变量表2] [/OPTION N] [STATISTICS N]

/OPTION=4 把相关阵以及有效记录的个数写入结果文件，供其它程序使用，这时省略WITH 语句，/OPTION=3 显示双侧概率。

/STATISTICS 1,2 分别指明单变量的均值、标准差、样本例数，以及协方差。

回归分析使用REGRESSION 命令，其用法说明详见第五节。

§5.4.5 NPAR TESTS

使用NPAR TESTS命令进行非参数分析。能够实现第二节提到的所有分析方法。

【例5.5】现针对第2章中的例2.8，SPSS/PC+程序为：

```
data list free /class scores.
begin data.
1 2.87  2 3.23  3 2.25
1 2.16  2 3.45  3 3.13
1 3.14  2 2.76  3 2.44
1 2.51  2 3.77  3 3.27
1 1.80  2 2.97  3 2.81
1 3.01  2 3.53  3 1.36
1 2.16  2 3.01
end data.
npar tests k-w=scores by class(1,3).
finish
```

【例5.6】第2章例2.9 的实现程序如下：

```
data list free/subjects coffe judge.
begin data.
1  1 1  1 2 3  1 3 2
2  1 2  2 2 3  2 3 1
3  1 1  3 2 2  3 3 3
4  1 1  4 2 3  4 3 2
5  1 2  5 2 3  5 3 1
6  1 1  6 2 3  6 3 2
end data.
npar tests friedman =judge coffe subjects.
```

运行结果：

```
 Mean Rank    Cases
      7.64        7   CLASS =    1
     14.93        7   CLASS =    2
      8.67        6   CLASS =    3
                  --
                  20   Total

                                          Corrected for Ties
     CASES    Chi-Square  Significance    Chi-Square  Significance
       20        6.1313         .0466        6.1405         .0464
```

§5.4.6 其它

第一个例子是拟合优度检验，取20个均匀分布随机数，用K-S 检验与理论相符吗？程序如下：

```
data list free /i.
begin data.
 1  2  3  4  5  6  7  8  9 10
11 12 13 14 15 16 17 18 19 20
end data.
compute x=uniform(1).
list.
npar tests k-s(uniform,0,1)=x.
```

程序首先用UNIFORM函数产生20个随机数，数据中的变量i 是给这20个随机数计数的，原始数据可见LIST命令产生的列表，检验是非参的，在括号内指定分布类型，本例还包括均匀分布的区间，因为是0-1，故也可以省略。

```
      I        X             I        X
-------------------------------------
1.00      .49           11.00      .46
2.00      .75           12.00      .67
```

```
        3.00      .16          13.00      .22
        4.00      .85          14.00      .94
        5.00      .89          15.00      .11
        6.00      .20          16.00      .79
        7.00      .39          17.00      .33
        8.00      .68          18.00      .14
        9.00      .86          19.00      .66
       10.00      .69          20.00      .60
    - - - - - Kolmogorov - Smirnov Goodness of Fit Test
         Test Distribution  -  Uniform       Range:   .00 To 1.00
         Most Extreme Differences
  Absolute       Positive       Negative          K-S Z      2-tailed P
    .15892         .06366        -.15892           .711           .693
```

结果表明双尾的P值=0.693，20 个伪随机数的分布的确与理论相符合。NPAR TESTS 命令除了K-S 检验外，SPSS/PC+可以用/EXPECTED子命令指定期望值进行分布的拟合优度检验。

下面的例子采自Stevens J.(1992). Applied Multivariate Statistics for the Social Sciences, 2nd Ed. Lawrence Erlbaum Associates, Inc. 中的例子。资料是研究儿童今后五年阅读困难的可能性。儿童按单词识别(WI)、单词理解(WC)及段落理解(PC)各五种等级打分。有两组儿童，第一组26名，是低风险的，第二组有12 名儿童，是高风险的。分析时要看两者协方差阵是否相同，可以用SPSS/PC+的MANOVA 语句，其程序如下：

```
set length=200.
title  'Check for equal covariance matrices'.
data list free /wi wc pc treats.
begin data.
5.8   9.7  8.9 1 6.2   3.0 4.3 1 5.7 10.3 5.5 1 2.4 2.1 2.4 2
10.6 10.9 11.0 1 4.2   5.3 4.2 1 6.0  5.7 5.4 1 3.5 1.8 3.9 2
8.6   7.2  8.7 1 6.9   9.7 7.2 1 5.2  7.7 6.9 1 6.7 3.6 5.9 2
4.8   4.6  6.2 1 5.6   4.1 4.3 1 7.2  5.8 6.7 1 5.3 3.3 6.1 2
8.3  10.6  7.8 1 4.8   3.8 5.3 1 8.1  7.1 8.1 1 5.2 4.1 6.4 2
4.6   3.3  4.7 1 2.9   3.7 4.2 1 3.3  3.0 4.9 1 3.2 2.7 4.0 2
4.8   3.7  6.4 1 6.1   7.1 8.1 1 7.6  7.7 6.2 1 4.5 4.9 5.7 2
6.7   6.0  7.2 1 12.5 11.2 8.9 1 7.7  9.7 8.9 1 3.9 4.7 4.7 2
7.1   8.4  8.4 1 5.9   9.3 6.2 1                4.0 3.6 2.9 2
                                                5.7 5.5 6.2 2
                                                2.4 2.9 3.2 2
                                                2.7 2.6 4.1 2
end data.
list.
manova wi wc pc by treats(1,2)/
       print=cellinfo(means,cov,cor) homogeneity(cochran,boxm).
```

程序首先印出单变量的分组均值和标准差，其次是一元方差齐性检验，方差- 协方差阵和相

关阵。矩阵齐性的Box 检验:

```
Multivariate test for Homogeneity of Dispersion matrices
Boxs M =                          14.12135
F WITH (6,2993) DF =               2.08589, P =   .052 (Approx.)
Chi-Square with 6 DF =            12.54363, P =   .051 (Approx.)
```

可见P=0.052，是近似齐性的。下面是第 4 章重复测量分析的相应程序。

```
data list free/y1 y2 y3 group.
begin data.
223 242 248 1  53 102 104 2 206 199 237 3 202 229 232 4
 72  81  66 1  45  50  54 2 208 222 237 3 126 159 157 4
172 214 239 1  47  45  34 2 224 224 261 3  54  75  75 4
171 191 203 1 167 188 209 2 119 149 196 3 158 168 175 4
138 204 213 1 183 206 210 2 144 169 164 3 175 217 235 4
 22  24  24 1  91 154 152 2 170 202 181 3 147 183 181 4
             115 133 136 2  93 122 145 3 105 107  92 4
              32  97  86 2 237 243 281 3 213 263 260 4
              38  37  40 2 208 235 249 3 258 248 257 4
              66 131 148 2 187 199 205 3 257 269 270 4
             210 221 251 2  95 102  96 3
             167 172 212 2  46  67  28 3
              23  18  30 2  95 137  99 3
             234 260 269 2  59  76 101 3
                           186 198 201 3
end data.
sort by group.
report vars=y1 to y3 /break=group
       /summary=mean /summary=STDEV.
manova y1 to y3 by group(1,4) /transform=repeated
       /rename=average dif2and1 dif3and2
       /print=transform
       /analysis=(dif2and1 dif3and2/average)
       /design.
```

REPORT 命令求出各组的均值与标准差，REPORT命令在本章中已经使用。MANOVA 结果首先是转换矩阵的转置(从略)。第一部分是轮廓平行的检验。

```
EFFECT .. GROUP
Multivariate Tests of Significance (S = 2, M = 0, N = 19 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .05251     .36850       6.00      82.00       .897
Hotellings       .05394     .35061       6.00      78.00       .908
Wilks            .94817     .35956       6.00      80.00       .902
```

```
Roys             .02867
```

接受轮廓平行假设。第二部分是轮廓水平一致的检验，拒绝假设。

```
EFFECT .. CONSTANT
Multivariate Tests of Significance (S = 1, M = 0, N = 19 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .59687   29.61176       2.00      40.00       .000
Hotellings      1.48059   29.61176       2.00      40.00       .000
Wilks            .40313   29.61176       2.00      40.00       .000
Roys             .59687
```

第三部分是轮廓重合检验(变量AVERAGE)。

```
Tests of Significance for AVERAGE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F
WITHIN CELLS          206430.49      41   5034.89
CONSTANT              975937.26       1 975937.26    193.83      .000
GROUP                  24218.29       3   8072.76      1.60      .203
P>0.05
```

§5.5 多元统计分析

§5.5.1 回归及其残差分析

命令格式：

REGRESSION
~/VARIABLES 指定分析回归分析的变量
/DESCRIPTIVE 均值、标准差、相关矩阵等统计量。
/SELECT 选择分析所用的记录。
/MISSING 指定缺失值的处理办法。
/STATISTICS 计算统计量。
/CRITERIA 指定回归分析准则。
/REGWGT 指示回归中的权。
/ORIGIN 使回归线通过原点。
/NOORIGIN 关闭ORIGIN，使回归不通过原点。
~/DEPENDENT 指定回归分析的因变量。
~/METHOD 指示变量的筛选准则
/RESIDUALS 回归残差。
/CASEWISE 按记录给出的统计量。
/SCATTERPLOT 产生一个或几个散点图。
/PARTIALPLOT 偏回归残差图。
/SAVE 存贮残差分析结果。
命令规则：(1) 必选项为VARIABLES、DEPENDENT 和METHOD 子命令；(2) VARIABLES 只能使用一次且应置于程序开始，(3) DEPENDENT 子命令可使用多次，

对每个DEPENDENT 子命令，估计一个方程；(4) DEPENDENT必须紧接一个或多个METHOD 子命令；(5) MISSING、DESCRIPTIVE 和SELECT 子命令在满足(1)、(4)的条件下可在任意位置出现；(6) CRITERIA、STATISTICS 和ORIGIN 子命令在被替代之前对其后所有方程有效；(7) 所有子命令以斜杠分开。

主要子命令及其用法：

1. VARIABLES

指定参与分析的变量的变量名，默认值为/VARIABLES (COLLECT) , 即/DEPENDENT与/METHOD 子命令中所有的变量，使用了/VARIABLES (COLLECT) , 必须在ENTER之后给出一些自变量，如：REGRESSION /VARIABLES Y x1 to X10 /dependent y /method enter /method remove x8 x10.

2. DEPENDENT

指定因变量名。可多次使用，且每次必须在其后紧接一个以上的METHOD 子命令，所指定的因变量名必须是在VARIABLES 中已定义过的。

例：REGRESSION VARIABLES=X1 TO X5，Y /DEPENDENT=Y. 表示以Y 为因变量，X1 互X5 为自变量参与回归分析。

3. METHOD

指定一个变量选择方法，BACKWARD 、FORWARD、STEPWISE、ENTER 表示后退、前进、逐步及全部变量入选。

前进法(FORWARD) 的实施步骤：首先选取与因变量相关系数绝对值最大的变量作为备选变量。同时，还应当对该备选变量的系数为零的假设作F 检验以决定该变量是否的确应入选。REGRESSION提供了两种准则：

准则一(FIN)：一个变量入选则最小应达到的F统计量值(小于FIN 值不入选)。在REGRESSION中，其关键字为FIN，默认值为3.84。

准则二(PIN): 一个变量入选则最大不能超过超过的(该F统计量相应的) 概率值(大于PIN 值不入选，小于才能入选)，其关键字为PIN，默认值为0.05。

对准则一、二在程序中首先指定其一，若没有变量满足入选准则，则所有变量均不进入方程。

若第一个变量已进入方程，则前进法继续进行变量选择。此时，首先计算尚未进入方程的那些自变量与因变量的偏相关系数，绝对值大者作为下一个候选者，并考察是否满足指定的准则，若满足则入选。

前进法进行到到没有变量可进入方程为止(变量进入方程时，还必须考察其容许性)。

⑵后退法(BACKWARD)与前进法相反，后退法首先将所有变量包括进方程，然后逐个将不合要求者删除。REGRESSION 提供了两种变量删除准则供选择使用。

准则一(FOUT)：一个变量要留在方程中最小应达到的F 值(小于FOUT值则删除，大于则留下)，其关键字为FOUT，默认值为2.71。

准则二(POUT)：一个变量要留在方程中最大能具有的概率值(大于POUT值则删除，小于则留下),其关键字为POUT，默认值为0.10。

⑶逐步选择法(STEPWISE)是前进法后退法的组合，其实施步骤如下：

第一个变量入选方法与前进法相同，若无一变量满足入选准则，则终止；第二变量入选与前进法相同；每次有一个新的变量入选后，都应考察前面已入选变量是否满足删除准则而被删除；继续考察是否有方程之外的变量应入选；当既无变量入选又无变量可删除时，终止。

注意，为防止同一变量反复入选——删除，应保证PIN ¡POUT 或FIN¿FOUT。
4. STATISTICS
控制关于方程和自变量统计量的输出显示，当STATISTICS 子命令缺损时或未指定任何关键字果的输出，包括R、ANOVA、COEFF 和OUTS。
ALL 输出除F，LINE 和END 外的全部统计量。
R 包括R^2，校正R^2 和这些估计的标准差。
ANOVA 方差分析表，包括回归平方和，残差平方和，F 及其概率等。
CHA R^2 的改变量。
BCOV 未标准化回归系数的方差——协方差矩阵。
XTX 矩阵X'X。
COND 条件数的界，包括已入选方程变量构成有矩阵X'X 的条件数的上下界。
COEFF 回归系数，包括回归系数及其标准误差、标准化回归系数、t 一值等。
OUTS 关于尚未入选变量的统计量。
ZPP 零阶、部分和偏相关。
CI 未标准化系数的95SES 标准化回归系数的近似标准误差。
TOL 容许性，包括已入选变量和未入选变量的容许性，以及下一个即将入选变量的容许性。
F 回归系数的F 值及其概率。
5. CRITERIA
控制建立回归方程时的统计准则，关键字：DEFAULT。当CRITERIA 子命令缺损时的默认值，当准则已被改变时，可用DEFAULT 恢复默认值。
PIN(值)，变量入选的F 概率，默认取值0.05。
FIN(值)，变量入选的F 值，其默认值为3.84，PIN 和FIN 只须指定一个。
POUT(值)，变量删除的概率值，其默认值为0.01。
FOUT(值)，变量删除的F 值，默认值为2.71，POUT 和FOUT 只能指定一个。
TOLERANCE (值)，容许性，默认值为0.0001。
MAXSTEPS(N) 最大步数，其默认值:
后退法或前进法: 满足PIN/POUT或FIN/FOUT的变量个数。
逐步选择法: 自变量个数的两倍。
例: REGRESSION VARIANLES=X1 TO X5，Y
/CRITERIA=PIN(.1) POUT(.15) TOL(.001)
/DEPENDENT=Y /METHOD=BACKWARD.
表示变量进入和删除标准都比默认值松。
6. ORIGIN 和NOORIGIN
控制是否对数据作中心化(即方程中是否包括常数项) 。必须放在其修饰的DEPENDENT和METHOD 之前，其默认值为NOORIGIN–表示方程中包括常数项。
例：REGRESSION VAR=V1 TO V3, Y, Z /DEPENDENT=Y /METHOD=FORWARD
/ORIGIN /DEP=Z /METHOD=FORWARD。
表示第一个和第二个回归方程分别不作中心化变换和作中心化变换。
7. DESCRIPTIVES
输出显示变量的描述统计量，关键字为：
NONE 不作任何输出，也是DESCRIPTIVES 省略时的对应项。

DEFAULTS 输出MEAN,STDDEN 和CORR。
MEAN 变量均值。
STDDEN 变量标准差。
VARIANCE 变量方差。
CORR 相关矩阵。
SIG 相关系数的单边概率。
BADCORR 当某些系数不能计算时，输出相关矩阵。
COV 协方差矩阵。
XPROD 对均值离差的交叉积。
N 用于计算相关系数的观测个数。
ALL 显示所有描述统计量。
8. MISSING
指示缺失值的处理办法，关键字为：
LISTISE 默认，删除在/VARIABLES 中出现缺失时的任何记录。
PAIRWISE 分别计算相关系数，但这样做有时会出现一些不太可能的结果。
MEANSUBSITUTION 缺失值用变量均值代替，这样会影响相关系数和预测值。
INCLUDE 指定缺失值为有效。
程序用例：

```
REGRESSION VARIABLES=X1 TO X5, Y /DEPENDENT=Y /METHOD=ENTER X1
X2 X3.
REGRESSION VARIABLES=X1 TO X5, Y /DEPENDENT=Y /METHOD=STEPWISE.
```

第一行程序为以变量X1,X2 和X3 为自变量的回归分析。其余所有子命令均取其默认值,程序将输出，方差分析表，回归系数等有关统计量；第二行程序为逐步回归，采用逐步选择法选择变量。
在输出结果中，B 所在列为回归系数，SEB 所在列为回归系数的标准误差，Beta 所在列为标准化回归系数，T 所在列为相应的t 值，SigT 为t 的双边概率。
残差分析：(1) 作为可选项的子命令RESIDUALS、CASEWISE、SCATTERPLOT 和PARTIALPLOT必须紧接着某个方程的最后一个METHOD 子命令，当拟合多个方程时, 可对每一方程作一次残差分析。(2) 残差子命令可以以任意顺序设置。(3) 残差子命令只影响它们紧接着的方程。(4) 采用矩阵输入时，不得使用残差命令。
分析中，REGRESSION 计算12 个临时变量如PRED、RESID，等。
主要子命令简述如下：
1. RESIDUAL
控制异常点信息的显示和标记，对临时变量输出Durbin-Watson 统计量，直方图和正态概率图。
关键字：DEFAULTS 也是RESIDUALS 不选任何关键字时的默认值，在各量后的括号内示意出来，包括SIZE(LARGE)DURBIN，NORMPROB(2RESID) ，HISTOGRAM (2RESID)，OUTLIER(2RESID)。
SIZE() 指示图的尺寸，取值为LARGE 或SMALL。
HISTOGRAM() 标准化临时变量的直方图。
NORMPROB() 标准化值的正态概率图(P-P 图)。
OUTLIER() 指定的临时变量最为显著的10个异常点。

Durbin-Watson 检验统计量。
ID(变量名) 异常点图上的观测个体的标识。
用例：/RESID=DEFAULT 表示对变量ZRESID 作正态概率图、直方图，给出Durbin-Watson 统计量和异常点图。
2. CASEWISE
对所指定的临时变量产生残差逐点图，关键字为：
DEFAULTS 关于标准化残差绝对值大于3.0 时的记录的图示，如果显示宽度足够，标准化因变量及其预测值随图列出。DEFAULT 包括OUTLIERS(3) , PLOT(ZRESID), DEPENDENT, PRED和RESID。
PLOT(临时变量) 绘制除标准化残差以外的记录图示，可供的选择有删除残差、学生化残差和学生化删除残差。
OUTLIERS(值) 给定记录图示新的异常值界值，默认值为3.0。
ALL 对所有的记录绘图，并不仅仅是残差超于界值的记录。
DEPENDENT 因变量。
PRED 预测值。
RESID 残差。
ZPRED 标准预测值。
ADJPRED 调整的预测值。
ZRESID 标准残差。
DRESID 删除残差，其计算方法是因变量减去其预测值，预测值由除去本记录以外的记录算得。
SRESID 学生化残差，即因变量减去其预测值，并除以预测值的标准差，标准差依自变量而变。
SDRESID 学生化删除残差，即除去本记录后回归方程的学生化残差。
SEPRED 预测值标准误。
MAHALANOBIS 反映观察影响回归的程度，用观察与自变量平均值的距离来度量。
COOK 反映观察对回归的影响，用观察不参加回归时所有残差的改变来表示。
LEVER 杠杆值，也能反映记录对回归的影响，与Mahalanobis' 距离相关。
DFBETA 第i 个观察删除后回归系数的改变情况。
SDBETA 标准化的DFBETA。
DFFIT 第i 个记录删除后模型拟合上变化情况。
SDFIT 标准化DFFIT。
COVRATIO 第i 个记录删除后协方差行列式的改变。
MCIN 因变量平均响应的预测下界LMCIN 和上界UMCIN。
ICIN 单个观察的预测区间下界LICIN 和上界UCIN。
3. SCATTERPLOT
指定一对变量并输出其散点图。对于每对变量名，前者为纵坐标，后者为横坐标，对临时变量名前应加上* 号，所有散点图中的变量均应为标准化的(所以指定*RESID与指定*ZRESID 一样)
例：/SCATTERPLOT (*RES,*PRE) (*RES, V1) 产生两个散点图,一个是残差——预测值图,一个是残差——变量V1 图。
4. PARTIALPLOT

产生偏残差图。若PARTIAL 使用时不作任何变量指定，则对方程中每个自变量均产生一个偏残差图，也可在PARTIALPLOT 之后指定欲作图的自变量，则此时只对指定变量作偏残差图。

5. SAVE

存贮生成的临时变量，由紧随关键字后的括号指定生成的变量名，如：RESID() 和ZRESID()。另外，FITS() 用于存贮DFFIT，SDFIT,DFBETA,SDBETA 和COVRATIO。

典型程序：

```
REGRESSION VARIABLES V1 TO V3 Y
/DEPENDENT=Y METHOD=ENTER RESIDUALS
/SCATTERPLOT (*RESID, *2PRED) /PARTIALPLOT.
```

【例5.7 】第 4 章的汽车数据分析

```
* NOTE: A transportation data.
DATA LIST FREE/ X1 X2 Y.
* FORMATS Y (F6.3).
BEGIN DATA.
 1300        .45       .066    948          2       .005
 1444         .5       .076   1440        2.4       .011
  736        1.5       .001   1080          3       .003
 1652         .4        .17   1844          1        .14
 1736         .8       .156   1116        2.8       .039
 1754         .8        .12   1656       1.45       .059
 1200        1.8        .04   1536        1.5       .087
 1500         .6        .12    960        1.5       .039
 1200        1.7         .1   1784         .9       .222
 1476        .65       .129   1496        .65       .145
 1820         .4       .135   1060       1.83       .029
 1436          2       .099
END DATA.
LIST.
regression /variables y x1 x2 /dependent y /method=enter
           /CASEWISE DEPENDENT PRED RESID ZPRED DFBETA COVRATIO.
compute ty=(y**0.6-1.0)/0.6.
regression /variables ty x1 x2 /dependent ty /method=enter
           /CASEWISE DEPENDENT PRED RESID ZPRED DFBETA COVRATIO.
```

§5.5.2 对数线性模型

命令格式：

```
HILOGLINEAR
~variable list 指定分析变量。
/PRINT 指定打印结果。
```

/PLOT 残差图。
/MAXORDER 限定最高交互项。
/CRITERIA 改变收敛准则和最大迭代次数。
/METHOD BACKWARD 改变默认的向前法。
/MISSING INCLUDE 使MISSING VALUE 引入的缺失值纳入分析。
/DESIGN 指定模型。
命令规则: (1) 程序的必选项为定义具有至少两个变量的变量列表，每个变量之后给出其最小和最大取值，其余子命令均为可选项。(2) 变量必须在程序最开始定义。(3) METHOD、PRINT、PLOT、CRITERIA、MAXORDER 和CWEIGHT 子命令必须置于它们要修饰的DESIGN 子命令之前。(4) 可指定多个METHOD 子命令，但每个仅影响下一个DESIGN 子命令。(5) 子命令之间应以斜杠分开。
下面将主要子命令简述如下：
1. VARIABLE LIST (变量列表)
定义参与分析的变量。变量必须取整数值。
例：HILOGLINEAR V1(1,2) V2(1,3) V3(1,4) 表示分析由变量V1、V2 和V3 构成的2 x 3 x 4 列联表。
2. METHOD
对其后的DESIGN 子命令，指定采用后退删除进行模型选择。METHOD 缺损时，所有变量均进入模型。然后把P值小于0.05的去掉。关键字为BACKWARD，只影响下一个DESIGN。
3. MAXORDER
控制在其后的DESIGN 中模型的最高阶次，例：
HILOGLINEAR v1 v2 v3 (1,3) /MAXORDER=2 表示对于变量V1、V2 和V3 构成的列联表，拟合的模型的最高项为两两交互项。
4. CRITERIEA
对其后的DESIGN，改变迭代拟合模型选择的迭代停止规则。关键字为：
CONVERGE(n) 收敛准则，其默认值为0.25，当被拟合的频数的改变量小于指定值时停止迭代。
ITERATE(n) 迭代最大次数，其默认值为20。
P(prob) 模型的卡方概率，默认值为0.05，仅当指定BACKWARD 方法时有效。
MAXSTEPS(n) 最大步数，默认值为10，仅当指定BACKWARD 方法时才有效。
DEFAULT 用来把CRITERIA 中的关键字的参数改变为其默认值。
5. CWEIGHT
含义：对一个模型指定各格点的权重，通常被用于指定列联表中的结构零。
用法：有下列三种方法指定权重。
. 指定一个变量名，以该变量的取值为格点权重。
. 直接提供一个格点权重矩阵，权重按变量列的顺序由左至右取值。
. 在方法2 中，可使用n*CW 表示权重CW 重复n 次。
例HILOGLINEAR V1(1,2) V2(1, 3) /CWEIGTH= CELLWGT 表示权重由变量CELLWGT 的取值给出。
HILOGLINEAR V1(1,2) V2(1,3) /CWEIGHT=(1 1 1 1 0 1 1 1 0) 或等价地使用/CWEIGHT=(0 3*1 0 3*1 0) 表示了对角元结构零，即：

	V2		
V1	0	1	1
	1	0	1
	1	1	0

6. PRINT

对其后的DESIGN 控制输出显示，关键字为：

FREQ 频数，给出观测和期望格点频数。

RESID 残差，给出原有和标准化残差。

ESTIM 饱和模型的参数估计(对其它模型该选择无效)。

ASSOCIATION 饱和模型效应的偏相关。

DEFAULT PRINT 缺损时的默认显示，包括显示FREQ、RESID 和所有模型以及饱和模型的ESTIM。

ALL 全部显示。

/PRINT 影响到后面的模型输出。

7. PLOT

含义：对其后的DESIGN，给出残差图，关键字为：

RESID 观测和期望频数的标准化残差。

NORMPLOT 调整后残差的正态概率图。

NONE 不做任何图。

DEFAULT PLOT 缺损时的默认图形显示，包括DESIGN 和NORMPLOT。

ALL 给出全部图形显示。

8. DESIGN

其缺损值时计算包括变量列表中所有变量在内的饱和模型，使用DESIGN 指定该饱和模型不同的生成类。将最高阶效应项列出(使用变量名和* 号表示交互效应项)。一个DESIGN 只估计一个模型，可多次使用DESIGN 子命令。

例：HILOGLINEAR V1(1,2) V2(1,2) V3(1,3) /DESIGN=V1*V2, V3

表示将对变量V1、V2 和V3 建立一个2 x 2 x 3 列联表。按照DESIGN 子命令将产生一个包括全部主效应和包括V1 和V2 交互项的模型。

【例5.8】下面是例6.2的(DV,DP,VP)模型的程序：

```
data list free/ p v d count.
value labels p 1 'yes'   2 'no'/
             d 1 'white' 2 'black'/
             v 1 'white' 2 'black'.
begin data
1 1 1 19   1 1 2 0  1 2 1 11  1 2 2 6
2 1 1 132  2 1 2 9  2 2 1 52  2 2 2 97
end data.
weight by count.
hiloglinear p(1,2) d(1,2) v(1,2) /design d*v d*p v*p.
```

部分结果：观察频数、期望频数及残差：

```
Factor    Code         OBS count  EXP count  Residual  Std Resid
  P       yes
   D       white
    V       white          19.0       18.7       .33        .08
    V       black          11.0       11.3      -.32       -.09
   D       black
    V       white            .0         .3      -.33       -.57
    V       black           6.0        5.7       .32        .13
  P       no
   D       white
    V       white         132.0      132.3      -.32       -.03
    V       black          52.0       51.7       .30        .04
   D       black
    V       white           9.0        8.7       .32        .11
    V       black          97.0       97.3      -.30       -.03
```

拟合优度统计量：

```
Likelihood ratio chi square =      .70080    DF = 1  P =  .403
         Pearson chi square =      .37446    DF = 1  P =  .541
```

LOGLINEAR 的句法与HILOGLINEAR类似，注意非层次模型没有层次模型那样的包含关系。其子命令简介如下：BY 指示模型中的主效应变量。WITH 指示模型所用的非表格形式的协变量。/CONTRAST () 指示对照的方法，括号内为对照的因素名。DEVIATION 与总效应比较。DIFFERECE 是各水平与其前水平的均值比较。HELMERT 是各水平与后面水的平均比较。SIMPLE 用每效应最末的水平作为标准。REPEATED 指示相邻水平间的比较。PLOYNOMIAL 指示为多项式，在平衡设计中是正交多项式。SPECIAL 和BASE SPECIAL用户自定义的对照。/CRITERIA 收敛控制。CONVERT()是收敛的精度，默认值为0.001。ITERATRE()为最大迭代次数，默认为20。DELTA()指示迭时每格子加上的值，默认值为0.5 。DEFAULT为默认值。/PLOT 结果的图示，残差、去趋势正态图。/PRINT—NOPRINT指示ESTIM、COR、RESID、FREQ、FREQ、DES设计的主效应。James Stevens 的例：数据是关于电视网络好坏的调查。年份为1959、1971，对象为白人或黑人，结果有“好”、“一般”、“差”三种。

```
data list free/year color response freq.
value labels year     1 '1959'  2 '1971'  /
             color    1 'black' 2 'white' /
             response 1 'good'  2 'fair'  3 'poor'.
begin data.
1     1     1    81    2     1     1   224
1     1     2    23    2     1     2   144
1     1     3     4    2     1     3    24
1     2     1   325    2     2     1   600
```

```
1     2     2   253    2     2     2   636
1     2     3    54    2     2     3   158
end data.
weight by freq.
loglinear response(1,3)  by color(1,2) year(1,2)/
          criteria=delta(0)/
          print=default estim/
          contrast(response)=special(1 1 1 1 -0.5 0.5 0 1 -1)/
          design.
```

结果：

```
Analysis of Dispersion
                                  Dispersion
  Source of Variation       Entropy  Concentration      DF
     Due to Model            32.854         24.316
     Due to Residual       2338.172       1438.487
     Total                 2371.026       1462.803    5050
Measures of Association
         Entropy =   .013857
   Concentration =   .016623
Estimates for Parameters
RESPONSE
 Parameter      Coeff.   Std. Err.     Z-Value Lower 95 CI Upper 95 CI
       1  1.2829846562      .08466    15.15509     1.11706     1.44891
       2  1.4512383652      .10809    13.42659     1.23939     1.66309
RESPONSE BY COLOR
       3   .4526490203      .08466     5.34686      .28672      .61858
       4   .3018183041      .10809     2.79237      .08997      .51367
RESPONSE BY YEAR
       5   .2951119842      .08466     3.48597      .12918      .46104
       6   .1612112793      .10809     1.49150     -.05064      .37306
RESPONSE BY COLOR BY YEAR
       7   .1028092054      .08466     1.21442     -.06312      .26874
       8   .0271094119      .10809      .25081     -.18474      .23896
```

§5.5.3 LOGISTIC 回归

命令格式：

LOGISTIC REGRESSION
~/VARIABLES 回归因变量和自变量。
/CATEGORICAL 指定名义的或有序的自变量。
/CONTRAST 在/CATEGORICAL 子命令指定为分类变量的对比类型。
/METHOD 选择变量的方法。

/SELECT 选择部分记录进行分析。
/ORIGIN 强迫回归线过原点。
/PRINT 选项DEFAULT 打印变量的分类表及统计量。
/CRITERIA 决定估计何时停止。
/CLASSPLOT 每步因变量实际值与预测值的分类图示。
/MISSING INCLUDE 指定包括缺值。
/CASEWISE 按照记录列出预测值、残差和其它暂存量。
/ID 对记录列表时，标识记录的变量。
/SAVE 对活动数据集增加预测变量。
/EXTERNAL 分析时将结果存放在外部暂存文件，节省内存的使用。
该命令的用法与线性回归命令REGRESSION类似，主要子命令说明如下：
1. VARIABLES
指定模型中的变量，为必选项。
2. CONTRAST
指定因变量对比类型，关键字为：
DEVIATION 即各效应与变量最后分类的偏差，亦是默认值。
DIFFERENCE 因素的每个水平与其前面的水平的平均效应比较。
HELMERT 因素每个水平与其后面水平的平均效应相比较。
SIMPLE 因素每水平与省略的或“参考”的水平相比，比较不是正交的。
REPEATED 是因素相邻各水平之间的比较，除第一个水平外，因变量的每个分类与其前面的分类相比较。
POLYNOMIAL 第一自由度包含因子水平间的线性效应，第二自由度包含二次效应，等等。
INDICATOR 指示分类成员的出现或不出现。
SPECIAL 用户指定，紧跟的可以是一个(k-1)xk方阵，k是因变量的水平。
3. CRITERIA
控制收敛的准则。关键字为：
BCON() 回归系数的改变, ITERATE 迭代次数，LCON()对数似然的改变，PIN()，POUT()，EPS()是进出的概率值和redundancy 检查。
4. METHOD
建模方法，关键字为：
FSTEP() 括号内指示WALD或LR，向前法逐步回归。删除变量使用Wald 统计量或似然比统计量。BSTEP() 用法与FSTEP类似，可以指示WALD或LR，向后法删除变量。
5. SELECT
对子集分析，变量间可使用关系运算符：EQ,NE,LT,LE,GT,GE。
6. PRINT
关键字为：default 对每个/METHOD 子命令，显示分类表、进入方程变量的统计量或虽虽然未进入方程但在/METHOD 或/VARIABLES 子命令指定过的变量。summary 与DEFAULT效果相同，只是在最后一步给出结果。corr 给出估计量间的近似相关。iter 每迭代步上参数估计值。all 给出所有的输出。
7. CRITERIA

控制收敛的准则，其关键字为：
BCON() 回归系数B 的改变量。ITERATE() 最大迭代次数，默认值为20 。LCON() 对数似然下降的百分比，默认值为0.0001。PIN() 变量入选计分统计量的概率，默认值为0.05。POUT() 变量删除的概率值，默认值为0.1。EPS() 冗余检验(redundancy checking)的精度，取值范围为大于10E-12到小于等于0.05，默认为10E-8，该检验避免变量的线性组合入选方程。
8. CASEWISE
PRED 预测概率。PGROUP 预测分组残差。观察组编码为0 到1, 该值等于观察组别减去在第二组的预测概率。RESID 即残差。DEV 为离真度。LRESID 为logit 残差。SRESID 为学生化残差。ZRESID 标准化残差。LEVER 杠杆。COOK 为COOK氏距离。DFBETA 为删除该记录后回归系数的改变。OUTLIER() 是控制标准化残差SRESID 大于某个数值时方显示。
9. SAVE
其关键字与CASEWISE 类似，每个关键字后可紧随一括号指出存贮量新的名称。

【例5.9】第 4 章的LOGISTIC 回归分析用例

```
data list free / age sex DM SD CHD freq.
variable labels
         age  'AGE IN YEARS'
         sex  'SEX (0=FEMALE, 1=MALE)'
         DM   'DIABETES MELLITUS (0=NO, 1=YES)'
         SD   'SEX*DIABETES  (INTERACTION TERM) '
         CHD  'CORONARY HEART DISEASE (0=NO, 1=YES) '
         freq 'NUMBER OF OBSERVATIONS'.
begin data.
50 1 0 0 0 6434  50 0 0 0 0 8519  60 1 0 0 0 4298  60 0 0 0 0 6199
50 1 0 0 1  124  50 0 0 0 1   45  60 1 0 0 1  179  60 0 0 0 1  116
50 1 1 1 0  193  50 0 1 0 0  159  60 1 1 1 0  218  60 0 1 0 0  228
50 1 1 1 1    6  50 0 1 0 1    5  60 1 1 1 1   13  60 0 1 0 1   10
end data.
TITLE 'LOGISTIC REGRESSION example from Chapter four'.
weight by freq.
logistic regression /variables CHD with age sex DM SD
                    /METHOD ENTER.
```

§5.5.4 因子分析

命令格式：

FACTOR
~/VARIABLE 列出FACTOR 命令所要求的因子分析的所有变量。
/MISSING 缺失值的处理方法。
/ANLYSIS 指示部分变量分析。

/PRINT 输出显示控制。
/PLOT 抽取因子的图示。
/FORMAT 因子矩阵的显示格式。
/DIAGONAL 指定相关阵对角元的初始共因子方差估计。/CRITERIA 指定因子提取和旋转的准则。
/EXTRACTION 因子抽取方法。
/ROTATION 因子旋转方法。
/SAVE 存贮因子得分到活动文件。
规则：(1) 只有VARIABLES 子命令为必选项；(2) VARIABLES, MISSING 和WIDTH 为全局子命令，对整个FACTOR 程序有效并且只能使用一次，VARIABLE和MISSING必须最先设置。
子命令用法如下：
1. VARIABLES
含义：指定参与因子分析的变量。
用例：FACTOR VARIABLES = V1 TO V10 指示变量V1～V10 参与分析。
2. ANALYSIS
含义：指定VARIABLE 变量集内的一个子集作为分析之用。
用法：可建立多个ANALYSIS 模块进行多次分析，在ANALYSIS 之后列出该模块的变量名，每一个ANALYSIS 作为一个模块开始，待下一个ANALYSIS 出现或FACTOR的结束作为其结束。
例：FACTOR VARIABLES =V1 TO V5 /ANALYSIS=V1 TO V3 /ANALYSIS=V3 TO V5。
指定V1～V3 和V4～V5 分别作为模块进行因子分析。
3. FORMAT
含义：重新设置因子载荷矩阵，关键字为：
SORT 以因子载荷递减方式排列因子载荷矩阵；BLANK(n) 删除因子载荷矩阵中绝对值小于指定值的系数；DEFAULT 取消SORT 和BLANK(n) 的设置。
例：FACTOR VARIABLES=V1 TO V5 /FORMAT=SORT BLANK(0.2) 表示因子载荷矩阵中以载荷递减顺序排列，且删除所有绝对值小于0.2 的载荷。
4. PRINT
控制一个分析模块中的统计输出显示，关键字为：
UNIVARIATE 显示有效观测个体数、均值、标准差；INITIAL 输出每个变量的初始公共因子方差，每个因子的相关矩阵特征根、方差百分比；CORRELATION 输出显示相关矩阵；SIG 输出相关系数的显著性水平；DET 输出相关阵的行列式值；INV 输出相关矩阵的逆矩阵；AIC 输出反象相关阵；KMO 输出Kaiser- Meyer -Olkin 指数和Bartlett 检验；EXTRACTION 输出因子载荷矩阵，每个因子的特征根和方差百分比；ROTATION 输出旋转矩阵的因子载荷矩阵和变换矩阵；FSCORE 输出因子得分系数矩阵(用回归方法获得的) ；ALL 输出所有得到的统计量；DEFAULT 相当于输出INITIAL, EXTRACTION 和ROTATION 所包含的显示内容。
例：FACTOR VARS=V1 TO V6 /PRINT=DET FSCORE 表示既要输出DEFAULT 的内容，还要输出因子得分系数。
5. PLOT

控制图形显示，关键字：
EIGEN scree 图，以降序方式输出特征根分布图；ROTATION 对每个旋转给出图示空间中变量位置分布图，括号内指定图中因子轴对应的因子。如：
FACTOR ... /PLOT ROTATION (1,2) (1,3) (2,3) ... 其中的数字表示图轴所使用的因子。
6. CRITERIA
指定因子提取和旋转的准则，关键字：
FACTOR (因子数) 提取的因子个数，其默认值是特征根大于等于MINEIGEN 的个数；MINEIGEN(值) 控制因子提取的最小特征值(只提取其对应的特征值大于给定值的因子)，其默认值为1；ECONVERGE(值) 采用迭代法提取因子时的迭代收敛准则，其默认值为0.001；ITERATE(迭代次数) 因子提取或旋转求解过程中的迭代次数，其默认值为25；RCONVERGE(值) 旋转迭代的收敛准则，其默认值为0 .0001；Kaiser 旋转时的Kaiser 正规化，这也是其默认值，由NONKAISER 废弃；DELTA(值) 斜交旋转的 δ，仅当指示了/ROTATE OBLIMIN 以后使用，其默认值为0。；DEFAULT 将所有准则恢复为默认值。
例：FACTOR VARS=V1 TO V6 /CRITERIA=FACTORS(3) 表示提取三个因子。
7. EXTRACTION
指定因子提取的方法，关键字：
PC 主成分方法(默认方法)；PAF 主轴因子法；ALPHA α 方法；IMAGE 象因子法；ULS 不加权最小二乘法；GLS 广义最小二乘法；ML 极大似然法。
DIAGONAL 子命令指定/EXTRACTION PAF 中相关阵对角元的初始共因子方差估计，默认初始共因子方差估计值为复相关平方和(SMC)。
8. ROTATION
指定旋转方法，关键字：
VARIMAX 方差极大旋转法(是默认值) ; EQUMAX 使用equamax 旋转; QUARTIMAX 使用quartimax 旋转法；OBLIMIN 斜交旋转；NOROTATE 不作旋转。
例：FACTOR VARS= V1 TO V5 /EXTRACTION= GLS /ROTATION /ROTATION OBLIMIN 表示采用加权最二乘法提取因子，第一次旋转用VARIMAX 法，第二次使用斜交旋转。
9. SAVE
指定因子得分的计算方法，并将因子得分以新变量形式存贮到当前文件，关键字(计算因子得分的方法)：
REG 回归方法；BART 使用Bartlett 方法；AR 使用Anderson-Rubin 方法；DEFAULT 默认值(回归方法)，在关键字后面，紧接着置于括号内的要存放的因子得分个数和根名字，其中所指定的数字不能超过所能获得的因子个数，也可用ALL 代替该数字。必须紧随/ROTATE, 存贮多次是允许的。
例：FACTOR VARS=V1 TO V12 /SAVE REG(ALL,FACT) 表示建立FACT1, FACT2, ... 等名字存放因子得分。
10. MISSING
DEFAULT 和LISTWISE 是等价的, PAIRWISE, MEANSUB, INCLUDE 分别表示用变量对判定、用均值代替、包含有缺失值的记录。
典型程序及结果说明：

FACTOR VARIABLES= V1 TO V5.

使用因子分析唯一的变量选项VARIABLES，其余子命令内容采用默认值，即主成分方法提取因子、方差极大法旋转。

【例5.10】Linden 数据因子分析，采自Johnson, R.A.。数据是关于二次世界大战以来奥林匹克十项全能得分，共160组数据，对每项得分施以标准化，对样本相关矩阵做主成份和极大似然因子分析。十项运动为：百米(x1)、跳远(x2)、铅球(x3)、跳高(x4)、400米栏(x5)、110米栏(x6)、铁饼(x7)、撑杆跳(x8)、标枪(x9)、1500米(x10)，样本相关矩阵为：

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.00									
x2	0.59	1.00								
x3	0.35	0.42	1.00							
x4	0.34	0.51	0.38	1.00						
x5	0.63	0.49	0.19	0.29	1.00					
x6	0.40	0.52	0.36	0.46	0.34	1.00				
x7	0.28	0.31	0.73	0.27	0.17	0.32	1.00			
x8	0.20	0.36	0.24	0.39	0.23	0.33	0.24	1.00		
x9	0.11	0.21	0.44	0.17	0.13	0.18	0.34	0.24	1.00	
x10	-0.07	0.09	-0.08	0.18	0.39	0.00	-0.02	0.17	-0.00	1.00

相关矩阵的前四个特征值分别为3.78, 1.52, 1.11 和0.91，累积频率为73 .3%，取个3或4个主成分，程序及结果如下：

```
DATA LIST MATRIX FREE/ X1 TO X10.
N 160.
BEGIN DATA.
1.00
0.59  1.00
...
-0.07  0.09 -0.08  0.18  0.39  0.00 -0.02  0.17 -0.00 1.00
END DATA.
FACTOR READ=COR TRIANGLE /VARIABLES=X1 to X10/CRITERIA FACTORS (4)
      /ROTATION VARIMAX.
FACTOR READ=COR TRIANGLE /VARIABLES=X1 to X10/CRITERIA FACTORS (4)
      /EXTRACTION ML /ROTATION VARIMAX.
```

未旋转时的结果：

	主成分法				极大似然法			
	因子1	因子2	因子3	因子4	因子1	因子2	因子3	因子4
x1	.69052	.21701 -	.52025	.20603	-.07027	.34879	.82887	-.16853
x2	.78854	.18360 -	.19260	-.09249	.08966	.43140	.59312	.27456
x3	.70187	-.53462	.04699	.17534	-.08079	.99618	-.00394	-.00075
x4	.67366	.13401	.13875	-.39590	.17969	.39761	.33440	.44513
x5	.61965	.55112	-.08376	.41873	.38983	.22492	.67031	-.13721
x6	.68689	.04206	-.16102	-.34462	-.00028	.36337	.42341	.38776
x7	.62121	-.52112	.10946	.23437	-.02058	.73125	.02676	.01819
x8	.53848	.08698	.41090	-.43955	.16980	.25601	.22761	.39371
x9	.43405	-.43903	.37191	.23451	-.00035	.44169	-.0115	.09714
x10	.14660	.59611	.65812	.27866	.99999	.00080	-.00001	.00000

旋转后的结果：

	主成分法				极大似然法			
	因子1	因子2	因子3	因子4	因子1	因子2	因子3	因子4
X1	.88383	.13651	.15619	-.11324	.16675	.85723	.24576	-.13773
X2	.63130	.19420	.51465	-.00557	.23951	.47650	.58033	.01101
X3	.24462	.82467	.22272	-.14791	.96530	.15373	.20015	-.05852
X4	.23934	.15046	.74966	.07647	.24192	.17289	.63175	.11320
X5	.79687	.07452	.10159	.46816	.05489	.70923	.23635	.32988
X6	.40381	.15319	.63466	-.17019	.20509	.26105	.58863	-.07061
X7	.18583	.81365	.14698	-.07890	.69726	.13288	.17967	-.00937
X8	-.03626	.17578	.76179	.21688	.13709	.07797	.51264	.11624
X9	-.04775	.73493	.10988	.14135	.41667	.01854	.17521	.00213
X10	.04467	-.04090	.11167	.93353	-.05520	.05572	.11333	.99045

极大似然法，Chi-square Statistic:10.5626, D.F.:11, P=.4806

公共因子方差：

主成分法：.83702,.70115,.81140,.64776,.87005,.61828,.72438,. 65957, .57446,.88761；极大似然法：.84202,.62132,.99892,.50037,.67069, . 46166, .53655,.30115,.20477,.99999。

由此可以得到特殊因子方差估计。

使用两种方法所得结论很不相同，主成分解中，除1500长跑外，所有项目在第一因子上有较大的正载荷，这个因子可称为一般运动能力，但是其余因子不能很好解释。而在极大似然解中，1500米跑是唯一在第一因子上有较大载荷的变量，所以这个因子可称为长跑耐力因子，因子2似乎是臂力因素(铅球与铁饼有较大载荷)，因子3是短跑速度(100米和400米跑载荷较大)。

进行旋转后，容易看出，铅球、铁饼和标枪都在同一因子上有较大载荷，因子可称为爆发性臂力。跳高、100 米栏和撑杆跳在某种程度上是包括跳远都在另一因子上有较大载荷，可称为爆发性腿力，100米和400米跑在一定程度上包括跳远都在第三个因子上有较大载荷，可称之为短跑速度。最后，1500米长跑在第四因子上有较大载荷，400米跑有中等载荷，可称为长跑耐力因子。用因子分析法得到的结论与田径运动中传统分类基本一致。

§5.5.5 判别分析

命令格式：

DSCRIMINANT
~/GROUPS 指示分组变量。
~/VARIABLES 指示参加判别的分析变量。
/SELECT 对具有指定的某一变量的指定值的子样进行判别分析。
/ANALYSIS 对VARIABLES 中的不同变量进行不同的判别分析。
/METHOD 提供变量筛选方法。
/TOLERANCE 改变变量进入的容许性。
/FUNCTIONS 限制判别函数的数目。
/PRIORS 指定先验概率。
/SAVE 存贮新变量，包括分组记号，判别得分，分组概率。
/OPTIONS 选择输出和控制。
/STATISTICS 输出统计量。
规则：(1) 必选项为GROUPS 和VARIABLES，其余为可选项；(2) GROUPS 、VARIABLES 和SELECT 应该顺序放在其余子命令之前；(3) 每一个ANALYSIS 命令指定其单独分析中所使用的预测变量，且这些变量应为VARIABLES 中变量的子集；(4) 所有其他命令可以以任何顺序安排，且仅影响其紧接着的ANALYSIS；(5) OPTIONS 和STATISTICS 控制输出选择；(6) 子命令以斜杠分开。
各子命令简要介绍如下：
1. GROUPS= 分类变量名(min, max)
指定分类变量名称及其取范围，min 和max 是变量的最小值最大值。
2. VARAIABLES
指定(用以对观测个体进行分类的) 预测变量名。
例：DSCRIMINANT GROUPS=AB(1,3) /VARIABLES= V1 V2 V3.
表示分类变量名为AB，它只有三个取值1, 2, 3，故构成三类判别。参与判别的预测变量为V1、V2 和V3。
3. ANALYSIS
用法：(i) 可对VARIABLES 中的不同变量指定不同的判别分析；(ii) 在逐步判别分析中来控制变量的入选方式；(iii) 其默认值对应于将VARIABLES 中所有变量进行分析；如：
DSCRIMINANT GROUPS=G(0,1) /VARIABLES= X1 TO X9
/ANALYSIS= X6 TO X9 /ANALYSIS=ALL.
第一次分析将只使用变量X6, X7 X8, X9 进行判别分析，第二次将使用X1 到X9 的全部变量进行判别分析。
4. INCLUSION LEVELS
在判别分析中控制变量进入或删除顺序，用法：①在ANALYSIS 中各变量之后附上一个0 到99 之间的整数表示其入选水平，默认值为1。变量入选按入选水平高低而先后不同。②具有偶数水平的变量按组同时进入模型，而具有水平为1 的变量则单独进入模型；③只有水平为1 的变量才可能删除；④ 0 水平变量永远不入选，但参与入选准则计算；⑤不论水平高低，不满足TOLERANCE (容许值)，则该变量不入

选。

常用的入选方法有：①DIRECT (全部入选，即逐步判别) ANALYSIS=ALL(2)表示将全部变量同时进入方程；②STEPWISE (逐步选择法) ANALYSIS=ALL(1)表示按逐步选择法增删方程中的变量；③FORWARD (前进法) ANALYSIS=ALL(3) 表示按向前法进入变量(不做变量删除)；④BACKWARD (后退法) ANALYSIS=ALL(2) ALL(1) ⑤表示按后退法选择变量(首先将全部变量选入方程，然合将满足删除原则的变量删除)。如：

DSCRIMINANT GROUPS=G(1,2) /VARIABLES= V1 TO V3

/ANALYSIS=V1 TO V3 (2) V4 V5(1) /METHOD=WILKS.

表示当V1、V2、V3 满足容许限时，同时也进入方程V4 和V5 按照逐步选择法进入。

5. SELECT

确定训练样本，待判样本(或验证样本)。用法：(i) SELECT 变量是数值型的，不必在VARIABLES 中；(ii) 若使用OPTION 9，则只对未选择的个体判别(即对训练样本不再判别)，例：

DSC GRO=A(1,2) /VAR=V1 TO V5 /SEL =V0(1) /OPT=9.

表示带有变量V0=1 的个体构成训练样本，其每个体构成待判样本，且只对V0¡¿1 的个体进行判别。

6. METHOD

指定变量选择的准则。关键字为：

DIRECT (默认值) 对所有通过容许限的变量同时进入；WILKS 表示Wilks λ 极小者进入，MAHAL 使两组间Mahalanobis 距离最大者进入，MAXMINF 使组间最小F 比最大者进入，MINRESID 对所有组对未解释变异的和极小的变量。，RAO 使Rao 的V 统计量最大者进入。

7. FUNCTIONS

确定判别函数的个数，其默认值为给出所有线性判别函数。该命令有三个参数：nf 函数的最大个数；cp 累积待征值的百分比。sig 函数的显著水平(默认为1.0) 只需使用一个参数即可控制函数的个数，三个参数应以nf、cp、sig 的顺序出现，如：

DSC GRO=A(1,4) /VAR = 1 TO 9 /FUNCTIONS=3,100,0.9

注意此时为 4 类判判，预测变量为 9 个，nf 的默认值为 9 个，所以该例前面两项指定为默认值，第三项说明若显著怍水平大于0.9，将产生相应的先概率。

8. PRIORS

含义：为各母体指定相应的先验概率。

关键字为：EQUAL 表示各母体有相等先验概率，为PRIORS 的默认值；SIZE 表示采用样本中各母体个体比例为先验概率的估计，如：

DSC GRO=A(1,3) /VAR= V1 TO V10 /PRIORS =.1 .5 .3. 表示母体1, 2,3 分别具有先验概率0.1, 0.5 和0.3。

9. SAVE

含义：存贮每个体的判别结果信息。

关键字：CLASS 后给出一个变量名，用来存贮每个个体的分类信息；SCORES 后指定一个根名字来存贮判别得分；PROBS 后指定一个根名字存贮各个体归属母体的概率，如：

DSC GRO=A(1,2) /VAR= V1 TO V10 /SAVE CLASS=P SCORES=Q PROBS=R

对上述的两类判别，在对每个个体判别之后，将在每个个体的信息中附加如下变量。P (个体被判定所属母体)；Q(判别得分)；R1(属于母体1 的概率)；R2 (属于母体2 的概率)。

10. 判别结果输出

/OPTIONS 取值为：

6 模式矩阵(pattern matrix) 的VARIMAX 旋转。

7 结构矩阵(structure matrix) 的VARIMAX 旋转。

9 仅对/SELECT 未选的记录分类。

10 仅对分组变量范围以外的记录分类

11 使用各组的协方差阵分类而非合并组内协方差阵

1 使用用户缺损值。

8 分类过程中用均值替换缺损值。

/STATISTICS 取值为：

1 判别变量的总均值及各组均值

2 判别变量的总均值及各组均值和标准差

10 区域图，使用头两个判别函数为轴做图，组均值用星号表示。

13 分类结果表，显示正确分类的比例。

14 每个记录的分类信息。

15 所有组的散点图，组号做记号，头两个判别函数为图轴。

16 每组散点图或直方图。

常用输出有：STATISTICS 10 区域图，该图以前两个判别函数为坐标轴，将各类判别的区域显示出来；STATISTICS 11 给出非标准化判别函数；STATISTICS 13 给出分类结果表；STATISTICS 14 给出每个个体的分类信息。

非逐步判别：DSC GRO=G(1,2) /VAR=V1 TO V3 /STATISTICS=11,13,14.

说明：G(1,2) 表示分类变量为G 的两类判别，参与判别的变量为V1, V2 和V3，显示结果见(图1-图3)。

逐步判别：DSC GRO=G(1,2) /VAR=V1 TO V3 /METHOD=WILKS /STATISTICS=11 , 13,14。

说明：采用Wilks 的 λ 准则作为变量选择准则，变量选择方法的默认值表示采用逐步选择法。

结果1: 非标准化典型判别函数系数

Unstandardized canonical discriminant function coefficients

	FUNC
V1	.4253184
V2	-.1196977
V3	.5354328
(CONSTANT)	.3255612

由此，非标准化典型判别函数为：

D=0.3255612+0.4253184 V1 - 0.1196977 V2 + 0.5354528 V3

结果2、分类结果表(Classificaiton results)

实际类别	观察个体数	判别结果	
G1	40	37	3
G2	30	2	28

在由40 个体组成的G1 中，有37 个被正确判归G1，3 个误判属于G2，由30 个个体组成的G2 中，有28 个正确判归G2，2 个误判给G1。

结果3. 每个个体的详细判别信息

个体序号	缺失值	训练样本个体标识	实际类别		最可能组 P(D|G)	P(G|D)	次可能组	判别得分
1		yes	1	1	0.4328	0.9811	2 0.0189	2.3015
2		yes	1**	2	0.4882	0.9773	1 0.0227	0.0366
.		.	.	.	.	.	. .	.
.		.	.	.	.	.	. .	.
.		.	.	.	.	.	. .	.

上表给出的是每个观测个体的判别信息，判别并非直接根据判别得分｜大小或正负进行，而是按最大后验概率(highest probability) P(D|G) 对应的类别进行，如：对个体1，实际类别为1，最大后验概率P(G|D)=0.9811所对应的类别为1，故判为类别1；对个体2，实际类别为1，最大后验概率P(G| D) = 0. 9773 所对应的类别为2，故判为属于类别2 (此时为错判，以** 标记)。

§5.5.6 聚类分析

命令格式：

CLUSTER

~variable list 指示参与聚类的变量。

/PRINT 控制打印输出。

/PLOT 显示成类过程图。

/MEASURE 度量准则。

/SAVE 保存成类结果。

/METHOD 聚类方法。

/MISSING 缺失值处理方法。

/ID 指示一个聚类成员表的标志变量，默认情况下使用记录号。

说明：CLUSTER 为系统聚类法程序，当所有可选项均缺损时，观测个体之间距离采用平方欧氏距离，各类之间采用类间平均法。输出显示包括：用于分析的观察个体数量，并类过程表和纵向逐步并类图。将具有缺损观测分量的观测个体略去不参与分析。

规则：(1) 程序必选项为指定一个变量列表，其余的为可选项；(2) 变量列表必须首先在其它子命令之前给出；(3) 变量表和子命令可各被指定一次；(4) 对同一矩阵可使用多种聚类方法。

下面将各种子命令简述如下：

1. VARIABLE LIST (变量列表) 指定每观测个体所包含的变量指标。为必选项且必须在其他子命令前定义。例：CLUSTER V1 V2 V3 表示共三个变量指标V1、V2 和V3 参与分析。CLUSTER ALL 在当前文件中用户定义的所有变量均参与分析。

2. MEASURE 子命令指定观测之间个体距离的度量方式(缺损或DEFAULT 时为SEUCLID)。关键字为：SEUCLID(平方欧氏距离)、EUCLID(欧氏距离)、COSINE(变量间的夹角余弦) 、BLOCK(城区距离)、CHEBYCHEV(Chebyshev 距离)、POWER(p,r) 中的(p,r)取为，(2, 1)、(2,2)、(1,1) 表示平方欧氏距离、欧氏距离和城区距离。例：CLUSTER ALL /MEASURE=BLOCK。表示采用绝对距离作为观测个体这间距离的度量方式。

3. METHOD 子命令指定一个或多个类间的距离度量方法，只能使用一次该命令，但可指定多个方法，缺损为BAVERAGE。关键字为：

BAVERAGE 类间平均距离法，WAVERAGE(类内平均法)，SINGLE(最短距离法)，COMPLETE(最长距离法)，CENTROID(重心法)，MEDIAN(中间距离法)，WARD 法。例：CLUSTER V1 V2 V3 /METHOD=BAVERAGE CENTROID 表示对变量V1 V2 和V3，分别采用类间平均法和重心法进行聚类。

4. PRINT 子命令

控制除图形之外的其它聚类结果显示，缺损值为SCHEDULE。关键字为：

SCHEDULE 显示并类过程信息(agglomeration schedule)。CLUSTER() 类别成员表，min 和max 分别指在聚类解中最小和最大的类别数，在括号内可以打入一个数，表示聚成的类数；打入两个数则是成类数的范围。DISTANCE 距离或相似系数矩阵，其类型因度量而定。DEFAULT 同SCHEDULE。NONE 取消上述选择，当希望用SAVE 存贮时，可用该选项。

例：CLUSTER V1 V2 V3 /PRINT=CLUSTER(3,5) 将显示聚为3、4 和5 类时各观测个体类别归属。

5. PLOT 子命令控制图形输出。该命令缺损值为VICICLE。关键字为：

VICICLE(min,max,inc) 纵向逐步并类图(或称为冰棱图)。范围的指定是可选项。若指定范围时应采用整数。min 与max 是开始显示和结束显示的聚类解的个数，inc 为增量，其缺损值为1。

HICICLE(min,max,inc) 水平逐步聚类图，用法仿上。若同时指定VICICLE 和HICICLE，则最后指定者有效。

DENDROGRAM 树形图，以合并距离按比例缩放为坐标尺。

NONE 无图形输出。

例：CLUSTER V1 V2 V3 /PLOT=HICICLE(4,10,2) 表示产生4 类，6 类，8 类和10 类水平逐步聚类图。

6. SAVE 子命令

对所指定的聚类解的水平，将类别成员(即所包含的观测个体) 以新的变量存放到当前文件。该命令的唯一指定是关键字CLUSTER，并在其后按括号内的数字表示聚类解的个数，或用(min,max) 指定为聚类解的范围，与/PRINT CLUSTER() 对应。注意它们为SAVE 子命令中的必选项。

例：CLUSTER V1 V2 V3 /METHOD=BAVERAGE(CLSUMEM) /SAVE=CLUSTER(4,5) 表示首先产生两个新变量CLUSMEM4 和CLUSMEM5，分别包含了当为为4 类和5 类时各观测个体的类别。

对每个要存贮其聚类信息的聚类METHOD，应在其后指定根名(rootname)。因此，当使用SAVE 时，METHOD 成为必选项。

7. MISSING 子命令

控制对带有缺损观测分量的个体的处理，该命令的缺损值为LISTWISE。关键字为：

LISTWISE 略去所有带有缺损值分量的个体。

INCLUDE 将带有用户缺损分量的个体包括进来参加分析。

8. ID 子命令

含义：在类别成员表，逐步分类图和树形图中命名一个字符串变量作为每个个体的标识符。该命令缺损时，以个体序号为标识符。

此外，还有直接按读写数据矩阵的WRITE 和READ 子命令。

程序范例：

CLUSTER V1 V2 V3 /PLOT=DENDROGRAM, VICICLE /PRINT= CLUSTER(2, 4) , SCHEDULE.

程序说明：⑴参与分析的变量为V1、V2 和V3。⑵聚类采用平方欧氏距离和类间平均值。⑶显示树形图和纵向合并图。⑷显示在分为2 类、3 类和4 类时的类别成员。

输出结果说明：

⑴类别成员表(cluster membership of cases) 第一列(CASE) 为按序号排列的观测个体；以后各列分别为聚成不同类时，各个体的所属类别。

⑵聚类过程表(Agglomeration schedule for clustering) 第一列为聚类步(stage)，第二、三列(clusters combined)给出每步所合并的为哪两类；第四列(coefficient) 为该两类的距离；第五、六列表示该两类分别在哪一步形成(stage cluster 1st appears)；第七列(next stage) 表示现在形成的类在哪一步又被合并。

⑶纵向逐步并类图(纵向冰棱图vertical icicle plot) 横向表示各个体的序号(case number)，纵向表示分成几类(number of clusters)，各类之间用空格分开。

⑷树形图(DENDROGRAM) 横坐标表示各类之间距离按比例缩放后的尺度数，纵坐标为个体标号。

【例5.11】从21个工厂抽取同类产品，每个产品测两个指标，欲将各厂的质量情况进行分类[2]，其程序如下：title '杨维权，刘兰亭，林鸿洲:《多元统计分析》，第218页'.

```
set /more off /LENGTH 35.
data list free /x1 x2.
begin data.
 0 6  0  5  2 5  2 3  4 4 4 3 5  1 6  2  6  1  7  0
-4 3 -2  2 -3 2 -3 0 -5 2 1 1 0 -1 0 -2 -1 -1 -1 -3 -3 -5
end data.
list.
CLUSTER X1 X2 /METHOD SINGLE COMPLETE CENTROID MEDIAN WARD
              /PLOT DENDROGRAM VICICLE.
```

LIST 命令给出数据列表，CLUSTER 命令使用不同的聚类方法对样品进行聚类，输出结果包括聚类过程(Agglomeration Schedule)、树形图(Dendrogram)和垂直冰棱图(Vertical Icicle Plot)，详细输出结果从略。

§5.5.7 生存分析

SPSS 和生存分析命令有三组，即KM、Survival和COXREG。用例：

```
KM survt BY treat
   /STATUS=cens event(0)
   /TEST=logrank,breslow, tarone.
```

```
SURVIVAL /TABLES=studytim BY drug(1,3)
         /STATUS=died(0)
         /INTERVALS= THRU 30 BY 3.9
         /PLOT(logsurv).
COXREG survt WITH treat age
   /STATUS=cens event(0)
   /CATEGORICAL=treat
   /PRINT=all.
```

附：SPSS/PC+ 4.0 运行菜单

由于SPSS/PC+ 强大的菜单提示，实际上没有必要将其完整的语法都列出来，少数不明确的地方可以注明。在菜单方式下，用右光标键→进入下一级菜单，用回车键选定即可，若不选定，则用左光标键←返回上级菜单。现将SPSS/PC+ 4.0 的运行控制菜单列表如下：

★入门是SPSS/PC+ 总的说明。
★读写数据
DE 数据录入工具。
GET /FILE ' ' 读取SPSS/PC+ 系统文件。
SAVE /OUTFILE ' ' 存贮SPSS/PC+ 系统文件，默认文件名为SPSS.SYS。
TRANSLATE FROM " 转入外部文件。
TRANSLATE TO " 转贮外部文件。
对电子报表，/FIELDNAMES 和/RANGE 指示区域名。
DATA LIST [FILE "] FIXED/、TABLE/、FREE/ 读取列表数据并为系统设定活动文件。变量表包括变量名和格式(N—A)，FREE 表示自由格式读取数据，N 表示小数点位置，A 表示字符类型变量。
BEGIN DATA 列表数据开始。
标号与格式化
VARIABLE LABELS 变量标号。
VARIATE LEBELS 变量名'说明' [/...] 指示变量的标号。
VALUE LABELS 变量名值'说明' 值'说明' ... 指示变量各取值的标号。
ADD VALUE LABELS like VALUE LABELS
FORMATS 格式化变量，常见格式如：(F) (COMMA) (DOLLAR)。
MISSING VALUES 变量名(缺失值) ... 指示系统的缺失值(SYSMIS)。
DATE (Trend 选项) 产生一个规则间隔的时序资料。
IMPORT /FILE ' ' 读入SPSSX 格式文件。
EXPORT 生成SPSS/PC+ 格式文件，选项与IMPORT 类似。
MODIFY VARS 修改变量的属性。
WRITE 将活动文件写于ASCII 文件，/CASES和/VARIABLES指示记录数和变量。
★修改数据或文件
1. 修改数据的值
COMPUTE 命令生成SPSS/PC+ 新变量。

CREATE (Trend 选项) leads, lags, 差分，移动平均，及类似的运算。
IF (条件) 变量=值条件语句。
RECODE 变量表(取值列表=新值) [...]，对数据重新编码。
COUNT 新计数变量=旧变量(值)，依条件计数。
RMV (Trend 选项) 去掉时序资料中的缺失值。
AUTORECODE 将变量的值重编码为连续的整数。
RANK 产生秩次，正态性得分，Savage 分和分位点。
2. 选择或数据加权
SELECT IF 变量条件 值. 做为以后处理的永久性选择条件。
PROCESS IF 变量条件 值. 为下一个统计命令进行有条件的选择。
N 数. 选择前N 个数量的记录进行处理。
SAMPLE 样本比例 样本大小[FROM 文件大小]. 随机选择一定比例的量。
WEIGHT BY 变量. 指示分析所用的加权变量。
PREDICT (Trend 选项) 指示Trends 命令的预测范围。
USE (Trend 选项) 指示命令分析的记录。

3. 文件操作
SORT CASE [BY] 变量名[A/D] [变量名]. 对活动文件的记录进行排序。
JOIN 合并两个或多个SPSS/PC+ 文件。
AGGREGATE 把亚组合并成单一记录。
FLIP 将活动文件的行列转置。
★数据作图
1. PLOT 产生两变量的散点图，并且可以同时计算回归统计量。
2. GRAPH 计算统计量并把它们传给绘图软件包(Harvard Graphics等)。
3. CASEPLOT Trends 选项。时间序列的示意图，图中时间轴是垂直的，与TSPLOT 不同，结合其他绘图软件可以产生高分辨图形。
4. TSPLOT Trends 选项，产生一个或多个时间序列的图示，时间轴是水平的。
5. NPPLOT 是Trends 中的选项，产生一个或多个时间序列的正态概率图。
6. FASTGRAF 调Graph-in-the-Box 至内存，并且返回至REVIEW。
7. MAP 是Mapping 中的选项，计算统计量并且传递到由MapInfo 而来的SPSS/PC+ Map (或Ashton-Tate 的Map-Master) 显示地图。
★分析数据
1. 描述统计(DESCRIPTIVE STATISTICS)
FREQUENCIES 产生频数表、综合统计量、直条图和直方图。
DESCRIPTIVES 产生描述统计量。
CROSSTABS 使用交叉表形式显示两个变量的分布，可以计算列联表统计量。MEANS 显示分组的均值。
EXAMINE 进行探索性数据分析。
2. 报告和制表(REPORTS and TABLES)
LIST VARIABLES=变量表/CASE=FROM 值TO 值BY 值数据快速、简单的列表。
REPORT 产生综合统计量报告，和记录列表。
TABLES 产生高质量的表格。

PRINT TABLES 在多种打印机上进行表格打印。
3. 相关与回归(CORRELATION and REGRESSION)
CORRELATIONS pearson 相关计算。
REGRESSION 多元线性模型的估计、假设检验和残差分析。
Trends 中的选项有
CURVFIT 趋势回归模型。
AREG 出现一阶相关误差时的回归分析。
WLS 加权最小二乘回归。
2SLS 两阶段最小二乘回归。
Advanced Statistics 中的选项
NLR 非线性回归。
LOGISTIC REGRESSION LOGISTIC 回归分析
PROBIT probit 或logit 分析。
4. 均数比较(COMPARING GROUP MEANS)
T-TEST 两组均值相等的检验。
ANOVA 多因素方差协方差分析。

ONEWAY 单因素方差分析，进行两两比较。
5. 高级统计(ADVANCED STATISTICS)
MANOVA 处理包括协变和重复测量量在内的多元方差分析。
6. 分类与聚类(CLASSIFICATION and CLUSTERING)
FACTOR 因子分析。
QUICK CLUSTER 当类数已知时的高效聚类分析。
CLUSTER 一般的系统聚类分析。
Advanced Statistics 中的选项。
DSCRIMINANT 判别分析。
7. 时间序列分析(TIME SERIES)
EXSMOOTH 指数平滑模型
SEASON 季节模型
ACF 自相关函数
PACF 偏自相关函数
CCF 互相关函数
ARIMA Box-Jenkins ARIMA 模型分析
FIT 评价模型的拟合情况
SPECTRA 周期的谱分析
X11ARIMA Census Method II X-11 季节调节模型
8. 分类数据分析(CATEGORIES)
ANACOR 对应分析(correspondence analysis)
HOMALS 使用交错最小二乘(ALS) 的一致性分析(HOMogeneity analysis)。过程对名义尺度的分类数据进行分析，把观察分成相一致的子集。
PRINCALS 使用ALS 的主成分分析(PRINcipal Components analysis)。过程对一组变量进行分析，确定它们变动的维数。与普通主成分分析不同，命令不要求变量用区

间尺度测量，只假设变量间的关系是线性的。
OVERALS 使用ALS 技术进行两个或多个变量集的非线性典型相关分析。与普通典型相关不同，OVERALS 并不需要变量用区间尺度测量，也不假定变量间的关系为线性。
OTHOPLAN 为conjoint 分析准备正交的设计。设计能用较少的几种选择实现实验对象到各因子水平组合的分配，它为PLANCARDS 和CONJOINT 准备设计文件(”plan file”)。
PLANCARDS 打印CONJOINT 的设计内容或给实验对象的几种选择。
CONJOINT conjoint 分析研究的结果，它既使用ORTHOPLAN 的设计文件，又使用数据文件，数据文件包含了实验对象对几种选择所排的秩次或打分。
9. 其它(OTHER)
NPAR TESTS 非参检验
一个样本的Binomial, chi-square, Kolmogorov-Smirnov 和runs 检验。两个样本的McNemar, sign, Wilcoxon 检验。
K 个相关样本的Cochran, Friedman, Kendall 检验。
两个独立样本的Man-Whitney, Kolmogorov-Smirnov, Wald- Wolfowitz 和Moses 检验。
k 个相关样本的Kruskal-Wallis 和median 检验。
RELIABILITY 在可加性的尺度上进行item analysis，计算一系列常用的可靠性指标，如Cronbach’s alpha。RELIABILITY 并不是把这些标度直接施于分析数据，若分析的结果很好，使用COMPUTE 命令产生包括这个标度的新的变量值。
HILOGLINEAR n-维交叉表层次log-linear 分析，检验模型中所有效应的显著性，估计饱和模型的参数，进行模型的选择。
LOGLINEAR 进行log-linear 和logit 分析，采用Newton-Raphson 算法估计饱和及非饱和模型，检验模型中所指定的效应，使用极大似然方法估计参数。
SURVIVAL 利用寿命表，图示及有关统计量，考察两个事件时间的长度，记录可以分组分析和比较。时间间隔可以使用SPSS/PC+ 的日期转换函数YRMODA。
★运行控制及信息
1. Set 命令改变由show 命令所报告的所有设置。

⑴菜单控制
/AUTOMENU ON/OFF 控制菜单的自动出现。
/HELPWINDOWS ON/OFF 控制菜单旁边的帮助窗口。
/MENUS STANDARD/EXTENDED 设置窗口为标准/扩展。
⑵输出控制
控制屏幕、打印机和文件的输出。
/SCREEN ON/OFF 开启/关闭屏幕输出。
/PRINTER OFF/ON 关闭/开启打印机，开启时运行速度减慢。
/LENGTH 页长，默认为屏幕24 行和打印机59 行。
/WIDTH 页宽，默认为79 字符。
/EJECT 打印机或输出文件中的回车控制，当/SCREEN ON 时关闭，关闭时以折线分页。
/INCLUDE ON/OFF 显示INCLUDE命令文件中的命令。

/ECHO ON/OFF 将命令复制到结果文件。
/LISTING 改变默认的输出文件SPSS.LIS。
/LOG 拷贝执行的命令到磁盘文件，默认文件是SPSS.LOG。
/RESULTS ” 矩阵(CORRELATIONS, FACTOR, MANOVA, 等产生) 和WRITE 输出的文件名，默认为SPSS.PRC。
⑶操作控制
/RUNVIEW ON/(OFF 或MANUAL), 返回REVIEW，可以在SPSSPROF.INI 中改动。
/PROMPT ” 行命令方式的提示，默认为SPSS/PC:
/CPROMPT ” 行命令方式下的续行提示。
/MORE ON/OFF 输出时的暂停。
/BEEP ON/OFF 系统振铃控制。
/COLOR ON/OFF 开启/关闭颜色。
/RCOLOR() REVIEW 的颜色，用括号中的三个整数表示。
/VIEWLENGTH 屏幕显示行数，默认为25。
/ERRORBREAK ON/OFF 终止一组命令的执行。
⑷工作文件控制
/COMPRESS ON/OFF 指示文件是否被压缩。
⑸其它
/SEED 随机数的种子，默认用系统时钟。
/BLANK 默认是数值变量中的空格设为系统缺失值，可设如/BLANK -99999。
⑹categories plots
调整ANACOR, HOMALS, PRINCALS, OVERALS 几个命令中图轴的刻度。
/CPI 横轴的每英寸字符数，默认为10。
/LPI 纵轴的每英寸字符数，默认为6。
2. SHOW 报告当前的设置情况。
3. DISPLAY 显示活动文件的变量名和标号
4. SYSFILE INFOR 用于检查非活动系统文件的内容。
5. SPSS MANAGER 内容包括STATUS, INSTALL, REMOVE，如：

SPSS MANAGER INSTALL REGRESSION /FROM ’d:\SPSSBACK’.
6. TITLE 和SUBTITLE 指示输出标题，COMMENT 或* 指示程序与注释。
7. TIME SERIES UTILITIES 包括：
TSETS, 设置DEFAULT, /PRINT, /NEWVAR, /GRAPHICS, /GOUT, /GINVOKE, /MX-AUTO, /MXCROSS, /MXNEWVARS, /MXPREDICT, /MISSING。
TSHOW 显示当前的设置。
MODEL NAME, TDISPLAY, SAVE MODEL, and READ MODEL
save model / OUTFILE=” /KEEP /DROP /TYPE
read model /FILE ” /KEEP /DROP /TYPE
VERIFY 检查日期与记录中的一致性，变量由/VARIABLES 指定。
8. 图形设置(graphics setup)
设置与SPSS/PC+ 相连的绘图软件，包括：HARVARD, CHART, CMASTER, 3GTALK, 4GTALK, DA，如：

GSET PACKAGE HARVARD
GSHOW 显示GSET 参数的当前值。
★运行DOS 或其它程序
使用DOS 或EXECUTE 命令运行DOS 或其它命令，如：
DOS DIR.
EXECUTE '\FORMAT.COM ' 'A:'.
★扩展菜单(extended menus)
在菜单与帮助系统中的某些菜单包括了一些看不到的内容，需要借助于扩展菜单，它们是一些较高级的或不太常用的特色，使用Alt-X 或者SET 命令进行标准菜单与扩展菜单的切换，在REVIEW 屏幕的右下角状态行上显示当前值。
★SPSS/PC+ 选项(SPSS/PC+ options)
SPSS/PC+ 软件是一系列功能的组合，通过安装特定的程序实现这些功能。
★退出
用FINISH 退出SPSS/PC+ 至DOS 系统，注意数据、程序和结果的保存。
系统菜单用例，现欲使用数据录入工具DE建立系统文件，进入SPSS/PC+ 后，选择read or write data，打右箭头出现DE，因光标恰好在DE位置，打¡Enter¿，则编辑窗内出现关键字DE，用Alt-C(或F10, 择run from cursor) 即进入DE。一般的命令可先使用Alt-E，然后打入关键字，打ESC，则提示窗口出现有关的子命令，然后用左右箭头进行选定。

第六章　BMDP

§6.1　概要

§6.1.1　简介

BMDP 由美国加州大学研制，为生物医学领域的数据分析而设计的经典统计分析软件包，用于各种计算机系统。在PC、VAX 和UNIX 系统均有菜单驱动的界面，但它们都有一个MENTOR，帮助用户自动产生所需的程序；一个编辑器，用于产生或修改命令文件；数据输入和产出功能。部分版本还有特定的在线帮助、高分辩图形程序、数据录入的电子报表。

BMDP的程序用两个字符标识，统计功能分为八类：

D 系列数据描述系列，如数据描述、t-检验、缺失值处理及简单图表。

F 系列列联表分析，如对数线性模型分析。

L 系列寿命表和生存分析，如Cox 回归分析。

M 系列多元分析，如聚类分析、因子分析、典型相半分析及逐步判别分析等。

R 系列回归分析，如多元回归分析、逐步回归分析、多项式回归、非线性回归及logistic 回归等。

S 系列非参数统计。

T 系列时间序列分析。包括一维与二维谱分析及Box-Jenkins 模型。

V 系列各种方差分析过程。

§6.1.2　运行

有批处理和交互两种运行方式，前者运行预先编好的程序，后者是在程序开始后，通过系统提示而运行。不带参数直接打入命令BMDP 将进入菜单控制下的运行方式，设软件安装于D:\BMDP>，先用SET设置运行环境(如SET DNEWS=D:\BMDP)，则命令：

D:\BMDP> BMDP <Enter>

启动交互式操作。BMDP 当光带落在相应功能上时，打<Enter> 则进入相应的选择，以<ESC> 返回上一级菜单或由主菜单返回DOS 操作系统。若选择RUN，系统首先提示运行的模块名，然后询问是交互式还是批处理方式。批处理的输入输出文件名隐含以.INP 和.OUT 作为文件名。

每个BMDP 程序都是为特定类型的统计问题而设计的，所有BMDP 的程序都使用BMDP.EXE 文件运行，例：

D:\BMDP> BMDP xx <Enter>

D:\BMDP> BMDP xx OUT=文件名<Enter>

D:\BMDP> BMDP xx IN=文件名<Enter>

D:\BMDP> BMDP xx IN=文件1 OUT=文件2 <Enter>

第一行表示交互方式启动程序"xx" (1D, 2D, 等)。第二行表示交互方式启动程序"xx"，所有输出存入文件中，同时在屏幕上显示运行结果。第三行表示批处理方式启动程序"xx" 从文件中读取指令。第四行表示以批处理方式启动程序"xx"，从文件1 中读取BMDP 指令，结果存入文件2。

表 6.1 收入情况与工作满意的程度

收入	满意度(satisf)				
(income)	很不满意	不满意	一般	很满意	小计
<6	20	24	80	82	206
6-15	22	38	104	125	289
15-25	13	28	81	113	235
>25	7	18	54	92	171
小计	62	108	319	412	901

批处理和交互式处理的不同点：1. 错误处理。BMDP 指令出错时，批处理将返回DOS，而交互式回到编辑态。2. 运行：交互式运行时，一些逐步过程允许覆盖每步的变量筛选。3. 高分辨率绘图仅在交互式有效。字符图形及图形文件仅在批处理方式有效。4. 指令执行：在交互式中，以下程序在读入数据后可以解释方式运行。3D (t 检验), 1T (谱分析), 2T (Box-Jenkins 时序分析), 4V and 5V (方差分析)。DM 是一个特例。它总是一步一步地执行。

§6.1.3 BMDP 有关概念和编辑工具

§6.1.3.1 BMDP 文件

命令文件包含BMDP的指令，用其它软件或BMDP 的全屏幕编辑产生。数据文件，含有待分析的数据。数据也可以列在指令文件END的后面，对于较大的问题，指令和数据最好分开。所有程序输出、分析结果输出至屏幕(CON:)。你可在BMDP中使用DOS的再定向或屏幕打印命令(”CTRL/PrtSc”)。屏幕输出可以写入列表文件，屏幕菜单和高分辨图形除外。系统存贮文件是BMDP产生和使用的一种二进制文件。PLOTFILE 是能由BMDPLOT使用的图形文件。

BMDP最多能使用3个暂存中间文件，多数情况下只用1个，用户可以不去考虑它，在磁盘空间不够或分析一个大的数据集时会碰到问题。

BMDP 也使用一个文件，它在系统安装了BMDP后产生。全屏幕编辑把指令拷贝到文件SCRATCH.EDT，而文件BMDP.LOG包括运行程序的所有指令。

§6.1.3.2 命令文件

㈠文件格式

【例6.1】下表是一个社会调查的例子[2]，变量是收入、满意度。收入单位为千美元，满意度分级为：很不满意、不满意、一般、很满意。

分析程序4F.INP 的内容如下：

```
/input variables are 2.
       format is free.
       table is 4,4.
/variable names are satisf,income.
/category names(1) are 'v. dissat','lit sat',
                       'mod sat','v. sat'.
          codes(1) are 1,2,3,4.
```

```
          names(2) are '<6','6-15','15-25','>25'.
          codes(2) are 1,2,3,4.
/tables columns are satisf.
       rows are income.
/statistics chisq.
           gamma.
/end
20 24 80  82 22 38 104 125
13 28 81 113  7 18  54  92
/end
```

BMDP 程序好象通常的英语文章，由段落(paragraph)和句子语组成(sentences)。尽管每个程序有多种选择，但BMDP 能用一些预先指定的值或默认值，因而BMDP 运行并不需要太复杂的指令。

各段用斜杠(/)引导。此处INPUT 段指定变量数目为2、自由格式数据、4x4 表格数据；VARIABLE段指示变量名为SATISF 和INCODE；CATEGORY 段指定各变量的分类代码；TABLES 段指示将构成的表样；STATISTICS 段指示计算χ^2和伽马系数。

句子也就是命令，用关键字开头，再用合适的语法规则选择IS, ARE 或= 之一连接一个项目表，以句号(.) 结束。一组相关的句子就构成一个段。每个段有一个名字(如：INPUT、VARIABLE、GROUP、END)并且以反斜杠(/) 分开。自由格式书写的程序也是允许的，但不提倡。

多数段落和句子可以简写，如INP表示INPUT，VAR表示VARIABLE；一个单句可以分成任意数目的行，每段内只放属于自己的句子；一般来说，段的次顺和句子的顺序是任意的，仅少数例外；使用井号(#)引导程序的注释，但不要把它放在引号中；出错的句子在警告信息中显示。

通常，BMDP 要求所有指令被分成段落，以EMD 段结束，程序把它们作为一个单元处理，而数据管理(DM) 则是一次一个段落，称为解释式执行，少数其它程序也这样执行，只是在给定读取数据的指令之后，如：3D (t 检验), 1T (谱分析), 2T (Box-Jenkins 时序分析), 4V及5V (方差分析)。

在交互式运行方式下，BMDP 用反斜杠(\) 结束每一段的指令，它必须是本行最后一个非空字符，END 段结束一个运行。

BMDP 使用TRANSFORM 段中的USE 选择分析用的记录。TRANSFORM 段有几个保留的名字，当每个记录都使用时，USE 的值设为+1，而KASE 是记录的顺序号，若设置USE 的值为零，则相关的记录在分析中不被使用，如：USE = AGE >= 30 AND AGE <= 50 仅使用AGE 从30 到50 的数据。另外，省略第17、23 及37可以使用：OMIT = 17, 23, 37. BMDP 使用VARIABLE 段中的USE 句子选用变量，省略时处理所有变量。

使用TRANSFORM 段转换和生成变量，如：

```
LOGAGE = LOG(AGE).
AVE = (V1+V2+V3+V4)/4.0.
AVE = MEAN(V1,V2,V3,V4).
NEWVAR = OLDVAR.
IF(OLDVAR > 0 AND KASE < 10) THEN NEWVAR = 1/OLDVAR.
```

数据管理程序(DM)提供了更复杂的数据处理功能。

建议：1.使用BMDP 的文件将简化BMDP 指令数目和改善系统运行速度，2.在VARIABLE段中使用USE，若文件较大，生成关于这些变量一个专用的BMDP 文件，由SAVE 段中的KEEP完成。3.读取有格式数据时，省略INPUT 段中的CASE，可用程序算得的记录做比较。

实际分析时，使用一个专门的文件存放数据，开始使用简单的描述分析了解数据的性态并且生成后续分析的BMDP系统文件，最后使用BMDP进一步分析。

数据文件可以是文本(ASCII)文件或特殊的BMDP二进制文件。BMDP 也能产生不同计算机系统间传输的格式，参见SAVE段的有关说明。BMDP 等待使用的文件是一矩形的格式，可以想象成一个表。每行是一个case，每列是一个变量。但不要把case、observation与record或文件中的一行相混淆。BMDP的一个case 允许包含任何数目的物理记录或行，反过来一行也可以有几个case 。数据格式有自由格式(FREE), STREAM, SLASH, BINARY 及固定格式”fixed.”。

星号(”*”) 表示缺失值，对于固定格式空格作为缺失值。AM 程序对于缺失值的处理很有用。

㈡段落选项

1. PROBlem TITLE=’text’. 标题，最长160字符。LENgth=#. 贮数组的最大长度。INTeractive. 交互式运行，默认时非交互式。ERRlev=STRICT或NORMAL. 指示BMDP检查程序的方法。

2. INPUT 段选项

TITLE=’标题’.	(选项) 程序输出标题。
FILE=’文件名’.	数据文件名。
CASES=#.	读入的记录数，省略时读至数据尾。
VARiables=#.	每记录的变量数。
FORMAT=FREE,STREAM,	自由格式、连续格式等数据格式
SLASH,BINARY.	自由格式由空格或逗号分隔，每个case 在新的记录
FORMAT=’指示’(fixed).	或行开始。STREAM中行的界
	限被取消。SLASH 同STREAM 一样，只是case
	之间用斜线(/) 分隔。固定格式是FORTRAN 格式
CODE=name.	”名字” 是长度1-8 的字母代码，
	CODE=’ ’ 表示读取文件中的第一个数据集。
CASE=#.	指定在许多问题中，本分析读入的例数。
REWIND.	重绕命令，用于磁盘或磁带的重读。
DIRectory.	打印BMDP文件的CODE, CONTENT,LABEL等信息。
CONTent=参数名.	指示BMDP文件的格式，如DATA,CORR,MEAN,FREQ。
LABEL=’标号’.	用于两个或多个BMDP文件读取时。
RECLEN=#.	记录长度，默认为80或72。
MCHAR=’字符’.	指示缺失，缺省时为’*’。
ERRMAX=#.	最大允许出错的记录数。
MULTiple=#.	允许一行读入多个有格式记录。
BLEVEL=#.	空格的处理方法，默认取值为2。

3. VARIABLE 段选项

NAMES=变量列表.	变量名表，省略时命名为X(1), X(2), ...。
LABEL=变量列表.	记录标号。
USE=变量列表.	使用的变量，参阅HELP 及SELECT的有关内容。
MINIMUM=#-列表.	
MAXIMUM=#-列表.	每个变量取值的上下界及缺失值。
MISSING=#-列表	其中的# 表示与变量表相应。
BLANK=ZERO|MISSing.	空格的处理方法。
ADD=NEW|#.	命名经转换的变量或数目。
BREore|AFTer|CHECK.	变量转换前后的检查，默认为BEFORE。
GROUPING=变量.	分组变量。
RETAIN.	经转换的变量保持上一个记录的值，可定义滞后变量。

4. GROUP 段选项

RESET. 删除所有分组信息。
CODE(变量表)=#-列表. 为变量定义有效码表(数据值)
CUTPoint(变量表)=#-列表. 为变量定义分组间隔。
NAME(V-List)=Name-list. 每个码值或范围的名字。
BMDP 使用CODE 或者CUTP 与NAME 进行变量分组，如：
CODE(STYLE) = 1, 2, 4, 8.
NAME(STYLE) = NONE, SOME, OK, AMAZING.
CUTP(AGE) = 25, 35, 45.
NAME(AGE) = KID, YOUNG, MIDDLE, OLD.

设定STYLE 的有效码值为1,2,4,8; 其它任何码将作为“坏码”处理。AGE 被分成四个区间：¡=25, ¿25 及¡=35, ¿35 及¡=45, 及¿45。名字与码或区间是相配的相同名字下给予码值将使用不同的组合并，如：”NAME(AGE)=OUT, LOW, HI, OUT.” 把小于25 及大于45 的两组合并单一的组，其标记为”OUT”。

有必要对GROUP 段与TRANSFORM 段的操作做一区分，前者并不改变数据的值，后者则不同，以下句子改变AGE 的值。

```
IF(AGE <= 25) THEN TEMP = 1.
IF(AGE > 25 AND AGE <= 35) THEN TEMP = 2.
IF(AGE > 35 AND AGE <= 45) THEN TEMP = 3.
IF(AGE > 44) THEN TEMP = 4.
AGE = TEMP.
```

5. TRANSFORM 段选项

句式为：变量= 表达式。
IF (关系) THEN 句子.
IF (关系) THEN (句子-1，句子-2 ...).
OMIT = #-列表.

DELETE = #列表.
USE = 表达式.
表达式：+, -, *, /, MOD, <, <=, >, >=, ==, <>, OR, AND, NOT
系统保留字
USE 记录使用变量：+1 使用记录；0 不用；
-1 从数据中删除记录；-100 停止数据输入.
KASE 正处理的记录数.
XMIS 内部缺失值标记.
TOOLARGE 数据超出范围的内部标志- 太大.
TOOSMALL 数据超出范围的内部标志- 太小.
函数
LOG、LN、SQRT、EXP、ABS、SIN、COS、TAN、ATAN、ASIN、ACOS、INT、SIGN、CHAR。
综合统计函数
N、NMIS、MIN、MAX、SUM、SUMC、MEAN、MED、SD、SEM、T等的含义与SAS
相同。
TRIM(i,a1,...,an) 关于第i个最大值和第i个最小值的截尾均值
TT(i,a1,...,an) 关于第i个水平截尾均值的t-值
IQR() 四分差
RHO(y1,...,yn) (1,...,n) 与(y1,...,yn) 的相关
B(x1,y1,...,xn,yn) y=a+bx 的直线回归系数B
A(x1,y1,...,xn,yn) y=a+bx 的直线回归截距A
B(x1,y1,...,xn,yn) (x,y) 的直线相关系数R
TRND(y1,y2,...,yn) 是(y1,...,yn) 在(1,...,n) 上的趋势
TCON(y1,y2,...,yn) 是(y1,...,yn) 在(1,...,n) 上的回归截距
AREA(y1,...,yn) y 下的面积
TRAP(x1,y1,...,xn,yn) y 下的面积，采用梯形法则
LINT(x0,x1,y1,x2,y2) 线性插值y1+(x0-x1)*(y2-y1)/(x2-x1)
LIND() 最末一个可用值的位置
LVAL() 最末一个可用值
INDEX(a,b1,...,bn) 第一个等于a 的b 的位置
REC(a,b1,c1,...,bn,cn) 换码函数，当bj的值等于a时存于cj
日期函数
DAYS(mm,dd,yy) 1960年1月1日起的天数
DAYS(mmddyy) 1960年1月1日起的天数
MDY(#of days) 六位日期作为一个变量
MM(#of days) 月份
DD(#of days) 日期
YY(#of days) 年份
JULN(#of days) 五位数西历(Julian date)
DAYJ(Julian) 1960年1月以来的天数
命令
REPL(y1,...,yn) 使用这些值的线性插值替换缺失值

FILL(x1,y1,...,xn,yn) 用周围的(x,y) 变量对替换缺失的y值
TEXT('message') 打印记录号和标号的讯息
SHOW() 使用a1,...,an的值打印记录号和标号

6. SAVE 段选项

存贮段产生一个数据文件，所有读入或TRANSFORM 产生的变量可被存贮，许多程序产生的特殊变量如回归中的预测变量及残差也可以存贮。生成的文件亦有BMDP系统文件及ASCII 文件两种。两种文件都需要指示：

FILE='文件名'.	文件名
KEEP=变量表.	(可选) 要存贮的变量，隐含为自动行成的任何量
DELete=变量表.	不保存的变量列表
CODE=名字.	长度为1-8 的数据集标识名
COMPlete.	仅保存那些不缺失的记录
LABEL='说明'.	(可选) 至多为40 个字符组成的数据集说明
CONTENT=名称.	(可选)
NEW.	生成一个新的系统文件；若文件已存在，则数据集被追加到文件中，(即一个BMDP 文件可包括几个数据集，经其CODE 来区分)
PORT.	BMDP 传输文件，由其他计算机系统如VAX/VMS 使用
FORMAT=F.	FORTRAN 10F8.3 浮点格式
FORMAT=G.	FORTRAN 5G16.6 的可变格式
FORMAT='specifier'.	FORTRAN 的格式指示
MISSING=标志表.	(可选) 对内部缺失数据标志重编码为指定的值

7. PRINT 段选项

多数程序具有专门的打印段选项；以下的说明适于所有程序。

NEWS.	打印程序信息(同NEWS 菜单选项)。
LEVEL=xxx.	MINIMAL, BRIEF, NORMAL, VERBOSE 控制程序输出的量
LINESIZE=#.	程序输出宽度80 或132
PAGESIZE=#.	每项行数，指定PAGE=0则不换页，默认为59
VNAME.	指定NO VNAME可以省略输入变量有关信息的打印
GNAME.	指定NO GNAME可以省略有关分类信息的打印
VUSE.	指定NO VUSE.可以省略/VARIABLE USE表的信息
DEBUG=xxx.	NONE,TEST,INFO,ALL指定调试的方法

§6.1.4 BMDP 模块用例

BMDP 的统计分析基本功能与SAS 及SPSS 相仿，现对它的列联表分析、生存分析和回归分析有关内容略作介绍，使用的有关记号如下，注意数据输入等段落总是必要的。

/ —指示段落	# —数字
. —句子结束	'c' —字符
—对于特定过程为必选项	VT —转换后的变量数
v —变量(名称或下标)	list —多于一个项目的列表
g —分组(名称或下标)	大写字母表示BMDP 的关键字

§6.2 4F、1L、2L

4F 语句格式：

```
/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/CATEGORY
/TABLE       ───┐
/PRINT          │  对子问题进行重复
/STATISTICS     │
/FIT         ───┘
/END
```

/INPUT 有选项TABLE=# list指示多维列联表的各个维数，CONTent= DATA 或TABLE 指定从BMDP文件中读入数据或表格，隐含是DATA类型。若使用TABLE，而且该表不是文件中的第一个，LABEL应指明表的顺序号，如：LABEL IS TABLE2。

/TABLE 是必选的，对两维表或多维表使用ROW=v list，COLumn=v list. 定义行列分类；对于多维列联表应使用INDices=v list或CATvar=v list. 若变量的分类的数目超过10个，则应在CATEGORY段指示CODES或CUTPOINTS。出现多个CATvar 时，第一个CATvar变化较第二个快，依次类推。PAIR 或CROSS 指示行列是配对或交叉。CONDition=v list. 指示对每个条件变量的每水平形成一个表。STACK=list. 作为单一指标，包含变量的所有可能的组合。COUNT=v list. 当输入为格子指标及频数时，指示频数若多于一个，则对每个变量分别进行。DELTA=#. 分析前每个格子加上的数，默认为 0 。EMPTY(#)=# list. 指示作为结构零处理的格子指标。INIT(#)=# list. 拟合矩阵的元素，默认值皆为1，当指示为零时即为结构零。SYMBols=c list. 指定模型识别各分类指标的符号，默认值为各变量的第一个字符。

/PRINT OBServed 打印观察频数表，除非指示NO OBS。EXCluded 删除某值的表，LIST=# 列出未进入分析的记录，LAMBda 对数线性模型参数估计值与其标准误，VARiance 对数线性模型参数间的方差协方差阵，PRECent=NONE,ROW,COL, TOT 即百分比，EXPected 为期望值，STANdardized. 打印标准化偏差。DIFFerence 指示观察值与期望值的差，FReeman为Freeman-Tukey 偏差统计量，CHISQuare. 对每个两维表或模型打印Pearson χ^2统计量，LRCHI 对每个两维列联表或模型打印似然比统计量。ADJusted 打印调整标准化偏差，MARGinal=# 打印#阶的边缘合计表。LAMBda. 打印对数线性模型的估计参数，BETA. 打印相乘参数的估计即估计值取自然指数，VARIance. 打印参数的相关和协方差阵。BAR='c'. 指示两维或多维表中的竖分界线。

/STATISTICS 有CHISQuare 即列联表χ^2，CONTingency 即列联表统计量，LRCHI 即似然比χ^2，FISHer 即Fisher 精确概率(2x2表)，TETRAchoric 即四格表相关，CORRelation 积矩相

表 6.2 美国Florida 州1976-77 年度死刑的数据

Defendent	受害者的	死刑(P)		小计
种族(D)	种族(V)	是	否	
白人	白人	19	132	151
	黑人	0	9	9
黑人	白人	11	52	63
	黑人	6	97	103

关，SPEARman 即Spearman相关，GAMma 即Γ、Kendall, Sommer's D 等统计量，LAMBda 即λ，TAUS 即Goodman 和Kruskal τ，UNCertainty 是不确定系数，MCNemar 即McNemar 对称检验、kappa 可靠性检验，LINear 即2xC 或Rx2 表，进行趋势性检验，ALL 打印所有统计量，NO ALL 选项可以由其它选项覆盖，MINmum=# 指示表中期望值小于该值时则行合并。

/FIT ALL 对两维或三维表拟合所有的层次模型。SIMULtaneous 打印给定阶数所有效应的同时检验。MODEL=list 如：MODEL=vp,dp. 没有指定符号时使用名字的第一个字母。ALL. 拟合两维或三维表所有层次模型。ADD=SIMPle,MULT. 从特定的模型开始逐步法拟合模型，每步增加一个或多个效应，DELete=SIMPle, MULT. 逐步法每步剔除一个或多个效应。STEP=# 指示从用户指定的模型增加或剔除的最大步数或在逐步法中可以认为是极端值的格子数。INCLude=list. 拟合模型时必定包括的效应，用符号表示。CELL=NO,STAN,FR 指示在逐法中使用最大标准化偏差或最大FREEMAN偏差。PROBability=#. 决定模型拟合显著程度的概率水准，省略时为0. 05。STRATA= all 或list 依次删除每个指标的层，对仅有两层的变量无效。CONVERGE=#,#. 指示允许的最大绝对误差和偏差，省略时默认为0.01和0.00001。ITERation=# 指示最大迭代次数。

/SAVE CONTent=DATA 或/和TABLE 存贮频数表存于BMDP文件。

【例6.2】下面是Agrest, A.(1990) 分析Radelet(1981) 的数据，见表6.2。

```
/PROBLEM    TITLE IS 'Death Penalty'.
/INPUT      VARIATES ARE 3.
            FORMAT IS free.
            TABLE IS 2,2,2.
/VARIABLE   NAMES ARE penalty,victim,defendent.
/TABLE      INDICES ARE penalty,victim,defendent.
            SYMBOLS ARE p,v,d.
/CATEGORIES CODES(1) ARE 1,2.
            NAMES(1) ARE yes,no.
            CODES(2) ARE 1,2.
            NAMES(2) ARE white,black.
            CODES(3) ARE 1,2.
            NAMES(3) ARE white,black.
 /PRINT     MARGINAL IS 2.
            EXPECTED.
            STANDARDIZED.
```

```
          LAMBDA.
/FIT      ALL.
/FIT      MODEL IS vd,p.
/FIT      MODEL IS vd,vp.
/FIT      MODEL IS vd,vp,dp.
/FIT      MODEL IS pvd.
/END
```

标题是'Death Penalty'，读入变量数为3，转换增加的变量是0，变量总数为3。使用的变量是penalty，victim，defenden，格式是自由格式。表的形式是penalty (p) x victim (v) x defenden (d)。检验所有的模型，最大迭代次数为20，收敛准则是0.01000，0.000010000，显著水平为0.0500。

结果包括观测值、排除的值、期望值、对数线性模型估计参数及标准化偏差、二阶边缘表。

所有模型的结果。

MODEL	DF	LIKELIHOOD-RATIO CHISQ	PROB.	PEARSON CHISQ	PROB.	ITERATIONS
p.	6	170.50	0.0000	140.08	0.0000	1
v.	6	363.46	0.0000	399.45	0.0000	1
d.	6	395.80	0.0000	416.31	0.0000	1
p,v.	5	138.04	0.0000	122.58	0.0000	1
v,d.	5	363.35	0.0000	398.68	0.0000	1
d,p.	5	170.39	0.0000	142.02	0.0000	1
p,v,d.	4	137.93	0.0000	122.40	0.0000	1
pv.	4	131.79	0.0000	115.97	0.0000	1
pd.	4	170.16	0.0000	141.36	0.0000	1
vd.	4	233.55	0.0000	200.64	0.0000	1
p,vd.	3	8.13	0.0434	6.98	0.0726	1
v,pd.	3	137.71	0.0000	121.32	0.0000	1
d,pv.	3	131.68	0.0000	115.90	0.0000	1
pv,pd.	2	131.46	0.0000	115.75	0.0000	1
pd,vd.	2	7.91	0.0192	7.04	0.0296	1
vd,pv.	2	1.88	0.3903	1.43	0.4889	1
pv,pd,vd.	1	0.70	0.4025	0.38	0.5402	6

故模型(VD,P)、(VD,PV)与(PV,PD,VD)值得特别注意，从最后三个模型可见PD 没有什么影响，而PV的影响很大而以模型(VD,PV)较佳。模型(VD,P) 估计参数如下，其标准误估计使用 δ 方法直接估计，对数线性模型估计THETA(MEAN)= 2.8379。

下表是模型(VD,PV) 的期望值，括号内为标准化偏差，即：(观察值- 期望值)/$\sqrt{\text{期望值}}$。

defenden	victim	penalty

		yes	no	TOTAL
white	white	21.2	129.8	151.0
		(-0.5)	(0.2)	
	black	0.5	8.5	9.0
		(-0.7)	(0.2)	
	TOTAL	21.7	138.3	160.0
black	white	8.8	54.2	63.0
		(0.7)	(-0.3)	
	black	5.5	97.5	103.0
		(0.2)	(-0.0)	
	TOTAL	14.3	151.7	166.0

对数线性模型参数为：THETA(MEAN) = 2.7237，λ 的估计值如下，括号内数据为估计值与标准误的比值。

penalty		victim		defenden	
yes	no	white	black	white	black
-1.171	1.171	0.799	-0.799	-0.391	0.391
(-10.108)	(10.108)	(5.796)	(-5.796)	(-4.130)	(4.130)

victim	penalty		defenden	victim	
	yes	no		white	black
white	0.264	-0.264	white	0.828	-0.828
	(2.282)	(-2.282)		(8.748)	(-8.748)
black	-0.264	0.264	black	-0.828	0.828
	(-2.282)	(2.282)		(-8.748)	(8.748)

$\lambda_{vp} = 0.264$表示受害者是白人时，所接受的处罚要重些，$\exp(4 * \lambda_{vp})$是D各水平上的估计比数比。

模型(VD,PV,PD) 参数标准误用信息矩阵的逆而获。参数估计THETA(MEAN) = 2.6922，其它结果从略。

表 6.3 两种实验条件下的生存时间(Gehan白血病数据)

病人分组	生存时间(有加号为截尾)
6-MP	6+,6,6,6,7,9+,10+,10,11+,13,16,17+,19+,20+,22,23
	25+,32+,32+,34+,35+
Control	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23

1L 语句格式：

/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/FORM
/GROUP
/ESTIMATE ┐ 对子问题重复 ┘
/PRINT
/END

/FORM 必选，指定生存变量的结构。共有三类即时间型、日期型和寿命表型。UNIT=c. c可以是DAY,WEEK,MONTH,YEAR；省略时使用MONTH。对时间型有TIME= v. 指示每记录的时间变量名或下标；STATus=v. 用作RESPonse 或LOSS 的示性变量，当所用的代码不出现时为截尾；RESPonse=# list，指示反应的代码，如RESP=1,6 表示1,6为失效，省略时取最小的代码；LOSS=# list 指示截尾的取值如LOSS= 2 TO 5。对日期型，输入包括ENTRY=v,v,v. 包含月、日、年的三个变量名或下标。TERMination=v,v,v. 指示截止的月、日、年变量，省略时系统处理为缺失。STATus=v.、RESPonse=# list.、LOSS=# list.与时间型类似，CENSOR=#,#,#. 用于研究或分析结束时的月、日、年标志，仅用于截尾观察的截止日期不清时。对寿命表类型的输入，每个记录包括了一个时间区间内的事件，记录应按时间排序。NENTer=v. 进入区间的数目。NDEAD=v. 区间内死亡数。NLOST=v. 区间内丢失数。NWITHdrawn=v. 区间退出数。INTERval=v. 每区间的下限。

/ESTimate 用于指示分析的方法、统计量和图示方法，为可选项并且可以重复。METHod=c. c 可以是LIFE或PROD，对于METHOD=LIFE.，PERIOD=# 是期望相等的区间数，WIDTH=#. 是时间间隔宽度，CUTPoint=# list. 划分间隔的时间点。对于METHOD=PROD，VARiable=v list. 指示与PL 估计一起打印的变量名表；PRINT. 打印生存分布的估计。PLOT=c list. 打印图示，包括SURV,LOG,CUM,HAZ,DEN 对应不同的生存估计。SIZE=#,#. 即横轴与纵轴的宽度与高度。GROUPing=v. 分组变量名。STATistic=MANTEL或BRESLOW 是可选项，指示生存曲线相等的检验。

【例6.3】生存分析。BMDP1L 用于寿命表分析，下面是Gehan [3] 有关白血病的数据，第一组用六腺嘌呤(6-MP) 和对照组(control) 的情况，数据列于下表：

相应的程序如下：

```
/INPUT     TITLE IS 'Kaplan-Meier test example'.
```

```
            VARIABLE=3.
            FORMAT=STREAM.
 /VARIABLE  NAMES=group,time,indica.
 /FORM      TIME=time.
            UNIT=weeks.
            STATUS=indica.
            RESPONSE=0.
 /GROUP     CODES(indica)=1,0.
            NAMES(group)='6-mp','control'.
 /ESTIMATE  METHOD=life.
            GROUPING=group.
            STATISTICS=Mantel,BR,TAR.
            PRINT.
 /END
  1 6   1  1 17  1    2  1  0  2  8  0
  1 6   0  1 19  1    2  1  0  2  8  0
  1 6   0  1 20  1    2  2  0  2  11 0
  1 6   0  1 22  0    2  2  0  2  11 0
  1 7   0  1 23  0    2  3  0  2  12 0
  1 9   1  1 25  1    2  4  0  2  12 0
  1 10  1  1 32  1    2  4  0  2  15 0
  1 10  0  1 32  1    2  5  0  2  17 0
  1 11  1  1 34  1    2  5  0  2  22 0
  1 13  0  1 35  1    2  8  0  2  23 0
  1 16  0             2  8  0
  /END
```

反应编码：0 (DEAD)，截尾编码：1(CENSORED)

检验统计量：

	STATISTIC	D.F.	P-VALUE
GENERALIZED SAVAGE (MANTEL-COX)	16.793	1	0.0000
TARONE-WARE	15.124	1	0.0001
GENERALIZED WILCOXON (BRESLOW)	13.458	1	0.0002

详细的寿命表，共有十个间隔，风险函数在每个时间间隔的中点计算 λ i= 2qi/[hi(1+pi)]，死亡密度函数为$p_iq_i/h_i, h_i$是i个区间的宽度。

表 6.4 与生存时间有关的描述统计量

	分位点	估计值	标准误
处理组	75TH	10.40	4.05
	MEDIAN (50TH)	23.40	3.13
对照组	75TH	3.72	1.57
	MEDIAN (50TH)	8.31	2.00
	25TH	12.91	1.92

6-MP 组

区间 weeks 等于 小于	进入	退出	失访	死亡	暴露	死亡比例	生存比例	区间开始时的累积生存频率	风险函数	死亡密度
0.00- 3.50	21	0	0	0	21.0	0.0000	1.0000	1.0000	0.0000	0.0000
3.50- 7.00	21	1	0	3	20.5	0.1463	0.8537	1.0000	0.0451	0.0418
7.00-10.50	17	2	0	2	16.0	0.1250	0.8750	0.8537	0.0381	0.0305
10.50-14.00	13	1	0	1	12.5	0.0800	0.9200	0.7470	0.0238	0.0171
14.00-17.50	11	1	0	1	10.5	0.0952	0.9048	0.6872	0.0286	0.0187
17.50-21.00	9	2	0	0	8.0	0.0000	1.0000	0.6217	0.0000	0.0000
21.00-24.50	7	0	0	2	7.0	0.2857	0.7143	0.6217	0.0952	0.0508
24.50-28.00	5	1	0	0	4.5	0.0000	1.0000	0.4441	0.0000	0.0000
28.00-31.50	4	0	0	0	4.0	0.0000	1.0000	0.4441	0.0000	0.0000
31.50-35.00	4	4	0	0	2.0	0.0000	1.0000	0.4441	0.0000	0.0000
对照组										
0.00- 3.50	21	0	0	5	21.0	0.2381	0.7619	1.0000	0.0772	0.0680
3.50- 7.00	16	0	0	4	16.0	0.2500	0.7500	0.7619	0.0816	0.0544
7.00-10.50	12	0	0	4	12.0	0.3333	0.6667	0.5714	0.1143	0.0544
10.50-14.00	8	0	0	4	8.0	0.5000	0.5000	0.3810	0.1905	0.0544
14.00-17.50	4	0	0	2	4.0	0.5000	0.5000	0.1905	0.1905	0.0272
17.50-21.00	2	0	0	0	2.0	0.0000	1.0000	0.0952	0.0000	0.0000
21.00-24.50	2	0	0	2	2.0	1.0000	0.0000	0.0952	0.5714	0.0272

2L 语句格式

/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/GROUP ┐
/PRINT │
/FORM │
/REGRESS │ 对子问题重复
/FUNCTION │
/TEST │
/PLOT │
/END ┘

在重复的子问题前应指定新的数据。FUNCTION 段对于时间协变量的情形是必需的，它可以使用TRANSFORM 段的句子，时间协变量应赋一个值，在FUNCTION 中只使用在REGRESSION 段句子COVARIATE, ADD 或AUXILIARY 中出现的变量名，TIME 是本段的保留字。

/FORM 与1L相仿，有时间型和日期型两种输入。

/REGRESS 指示回归模型，如COVAriates=v list. 指示固定协变量名称和下标。STATA=v. 指示分层变量名或下标。与变量筛选有关的选项有：STEPwise= MPLR 或PHH 指示最大偏似然比检验或Peduzzi-Hardi-Holford 统计量；REMOVE= #. 及ENTER=#. 指示逐步筛选变量p值的大小；START=IN,OUT 指示一个协变量在第一步是否在回归方程中；MOVE=# list. 指示一个协变量最多能被删除的次数，省咯为两次。RISK=LOGLINear, LINEAR,COMBINATION,USER 指示风险函数的形式为对数线性、线性、组合型及自定义，指定了LOGLIN以后对于每个固定协变量减去均值；与牛顿—拉弗森算法有关的选项有：CONVergence=#.与ITERation=#.指定收敛准则和迭代次数，默认值分别为0.00001和15；HALVing=#.指示每步最多使用的两分法次数，默认为5；TOLerance=# 矩阵求逆的容许限值，默认值为0.00001；INITial =# list. 指示对应于COVARIATES变量的初值。时间协变量的选项有：ADD=v list. 指示FUNCTION段中的时变协变量名；AUXIlliary=v list. 定义FUNCTION 中的时变协变量；PASS=# 指示时变协变量的层数。

/TEST 段是可选的，当变量筛选时忽略。ELIMinate=v list. 指示待检验的协变量回归系数；STATistics=c list. 指定计算WALD、LRATIO或SCORE 统计量，默认为WALD。

/PLOT 是可选的，指示STATA时对每层进行。TYPE=c list.对SURV,LOG或FIT 绘图。PATtern=# list. 定义生存函数的协变量值；SIZE=#,# 是横轴与纵轴字符的数目，默认值为100和50。

/PRINT CASE=#. 打印转换后原始数据的数目，默认为10。SURVIVAL. 打印排序的生存时间、Kaplan-Meier估计量、风险值、生存函数。CORRelation. 打印近似相关。COVAriance. 打印近似协方差。ITERations. 打印每步Newton-Raphson迭代的对数似然及参数估计值。

【例6.4】比例模式的检验。是1983年版BMDP引用Pike关于两组鼠接触某种致癌物数据进行生存分析的例子，拟合模型是$h_0(t)\exp(\beta_1 z_1+\beta_2 z_2(t))$，其中有一个时间协变量，可详见Kalbfleisch and Prentice(1980)，程序如下：

```
problem   title is 'checking the proportionality assumption'./
input     variables are 3. format is stream./
```

```
variable  names are survival, followup,group./
form      time=survival. status=followup. response=1. /
print     covariance./
regress   covariate=group. add=z2./
function  z2=group*(ln(time)-5.4)./
end
143 1 0  156 1 1  220 1 0  239 1 1
164 1 0  163 1 1  227 1 0  240 1 1
188 1 0  198 1 1  230 1 0  261 1 1
188 1 0  205 1 1  234 1 0  280 1 1
190 1 0  232 1 1  246 1 0  280 1 1
192 1 0  232 1 1  265 1 0  296 1 1
206 1 0  233 1 1  304 1 0  296 1 1
209 1 0  233 1 1  216 0 0  323 1 1
213 1 0  233 1 1  244 0 0  204 0 1
216 1 0  233 1 1  142 1 1  344 0 1
/end
```

共有36名失效，4名截尾，截尾占10%。自由变量名称及编码：3 group，4 z2。对数似然比：LOG LIKELIHOOD = -100.7113，χ^2：CHI-SQUARE = 3. 05，D.F.= 2，P-VALUE =0.2176。

参数估计结果如下，z2的回归系数相对于其标准误的值-0.1258很小，表明用一个固定协变量GROUP的模型不恰当。

VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3 group	-0.5998	0.3484	-1.7216	0.5489
4 z2	-0.2295	1.8249	-0.1258	0.7949

渐近协方差阵：

```
ESTIMATED ASYMPTOTIC COVARIANCE MATRIX
---------------------------------------
                  group          z2
                      3             4
group      3         0.1214
z2         4         0.0563         3.3302
```

【例6.5】临床研究中，病人的预后可因为治疗过程中的一些事件而改变，我们可以把这些事件作为时间协变量引入。下面是著名的Stanford心脏移植数据，共有99例，数据放在文件HEART.DAT变量是生存天数(survival)，是否截尾(status)，等待移植的时间(waittime)，移植时的年龄(age)，以及排斥打分(mismatch)。当一个或多个协变量是生存时间的函数时，还需要更多的运行控制，下面程序也是上述BMDP 手册上的例子，固定协变量仍然用COVARIATE指示，时变协变量必须用REGRESSION段中的ADD指示，有必要用协变量以外的变量定义时变协变量时，应在AUXILIARY中引入。

```
problem   title is 'Heart Transplant Data with
          Time-dependent Covariates'./
input     variables are 6. file is '2L.DAT'. mult=4.
          format is '(4(F3.0,F5.0,F2.0,2F3.0,F5.2))'./
variable  names are id,survival,followup,waittime,age,mismatch.
          blanks are missing./
form      time=survival. status=followup. response=1. /
regress   add is xplant,xplntage,score.
          auxiliary=waittime,age,mismatch. /
function  xplant=0.0. xplntage=0.0. score=0.0.
          if (time GE waittime) then xplant=1.0.
          if (time GE waittime) then xplntage=age.
          if (time GE waittime) then score=mismatch./
print     cases are 99. covariance./
end
```

参与分析变量：1 id，2 survival，3 followup，4 waittime，5 age ，6 mismatch，格式为：(4(F3.0,F5.0,F2.0,2F3.0,F5.2))。时变协变量为：7 xplant，8 xplntage，9 score。失效为71，截尾28，占总例数的28.28%。自由变量：7 xplant，8 xplntage，9 score。

对数似然比：LOG LIKELIHOOD = -275.9557，χ^2 ：GLOBAL CHI- SQUARE=9.01，D.F.=3，P-VALUE =0.0291。

参数估计：

VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
7 xplant	-3.1780	1.1861	-2.6793	0.0417
8 xplntage	0.0552	0.0226	2.4423	1.0567
9 score	0.4442	0.2803	1.5851	1.5593

协方差阵：

		xplant	xplntage	score
		7	8	9
xplant	7	1.4069		
xplntage	8	-0.0246	0.0005	
score	9	-0.0870	-0.0003	0.0785

【例6.6】是Collett, D. (1991)[4] 的例子，资料是Smith, W.(1932) 关于一种保护血清对肺炎球菌的影响，第七天仍存活的鼠为生存，血清单位为cc 。利用LR 进行LOGIT 分析，程序如下：

```
/PROBLEM        TITLE = 'SERUM'.
/INPUT          VARIABLES = 3.
                FORMAT=FREE.
```

```
    /VARIABLE       NAMES = DOSE, Y, N.
    /TRANSFORMATION LOGDOSE=LN(DOSE).
    /REGRESS        COUNT=N.
                    SCOUNT=Y.
                    INTERVAL=LOGDOSE.
                    MODEL=LOGDOSE.
    /PRINT          CELLS=MODEL.
                    COVA.
    /END
    0.0028 35 40
    0.0056 21 40
    0.0112 9  40
    0.0225 6  40
    0.0450 1  40
```

结果：logit(p)=-9.19-1.83 log(dose)

ED50=(0.0054,0.0081)，LOG(ED50)=-5.021±1.96 x 0.1056

```
                        LOG LIKELIHOOD =   -87.062
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E))  =     2.809  D.F.=   3  P-VALUE= 0.422
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)=     2.917  D.F.=   3  P-VALUE= 0.405
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )  =     1.871  D.F.=   2  P-VALUE= 0.392

                                      STANDARD
     TERM               COEFFICIENT       ERROR   COEFF/S.E.  EXP(COEFFICIENT)
LOGDOSE                   -1.8296      0.2545      -7.188      0.1605
CONSTANT                  -9.1894       1.255      -7.322      0.1021E-03

SUMMARY DESCRIPTION OF CELLS.
      CELLS ARE FORMED BY ALL COMBINATIONS OF VALUES OF VARIABLES IN THE MODEL.
-------------------------------------------------------------------------
                OBSERVED PREDICTED  S.E. OF  OBS-PRED  PRED.                    HAT
NUMBER NUMBER  PROPORTION PROB.OF  PREDICTED --------   LOG                   MATRIX
Y      FAILURE  Y         Y          PROB.   S.E.RES.  ODDS    CHI    DEVIANCE DIAGONAL INFLUENCE LOGDOSE
    1     39     0.0250   0.0289    0.0137   -0.1710  -3.5156 -0.1463  -0.1496   0.2686    0.011    -3.10
    6     34     0.1500   0.0956    0.0288    1.4926  -2.2474  1.1707   1.0901   0.3848    1.393    -3.79
    9     31     0.2250   0.2747    0.0423   -0.8797  -0.9710 -0.7039  -0.7186   0.3598    0.435    -4.49
   21     19     0.5250   0.5738    0.0501   -0.8116   0.2972 -0.6235  -0.6210   0.4098    0.457    -5.18
   35      5     0.8750   0.8271    0.0454    1.2314   1.5654  0.8008   0.8344   0.5771    2.069    -5.88

MINIMUM EXPECTED CELL FREQUENCY          =      1.15
NUMBER OF EXPECTED VALUES LESS THAN 5.0  =     2
```

与SAS 的程序进行比较:

```
    DATA SERUM1;
```

```
   INPUT DOSE Y N;
   LOGDOSE=LOG(DOSE);
CARDS;
PROC PROBIT;
   MODEL Y/N=LOGDOSE/D=LOGISTIC;
   OUTPUT OUT=SERUM2 PROB=PHAT;
PROC PRINT;
DATA SERUM3;
   SET SERUM2;
   YHAT=N*PHAT;
PROC PRINT;
```

§6.3 各系列模块功能概要

为了应用的方便，此处在上节的基础上，介绍一下BMDP各模块的功能。

§6.3.1 D 系列

1D 简单数据描述

对所有或部分记录提供常用的描述统计量，数据列表、排序等。

2D 详细的数据描述

计算许多描述统计量，并对每个变量绘直方图。2D 用于识别异常值，研究分布的形状，以及对样本数据初步描述。输出内容包括：均值，中位数，众数，标准差和均值和中位数的标准误，偏度与峰度，极值，Shapiro 与Wilks' W，截尾均值, Hampel 和biweight 估计量。

3D T 检验

提供三种不同的t 检验，结果输出包括直方图和描述统计量。TWOGROUP 是两组t 检验，方差等或不等，包括方差齐性Levene 检验及截尾t 检验(trimmed t)、非参Mann-Whitney 秩和检验、Hotelling's T-方及Mahalanobis D-方，同时给出每组内的变量相关值。MATCHED 进行配对t 检验，输出包括配对t 统计量和Pearson 相关系数，也能给出trimmed t，非参符号和Wilcoxon 符号秩次检验，Spearman 相关，Hotelling T-方和Mahalanobis D-方。ONEGROUP 提供单样本t 检验，类似于配对t 检验，不打印Spearman 相关。

4D 字符频率—数字的和非数字的

计算单列字段每个字符(数字、字母或符号) 的频率，产生的数据可以是原来数据列表或者用指定符号替换过的数据，这样易于找出不需要的符号或字母。当数据以固定格式排齐时，4D 可用于初步检查数据的类型，计量单列数据的频率并发现数据错误，所有数据均以A1 格式(宽度为一个字符) 读取，4D 不接受BMDP 文件输入。输出为一个频数表，显示每列中不同字符的频率。

5D 直方图和单变量图

多组数据可以画在同一图上，也可分组画在几个图上。图的大小、标度以及每个区间的名称都可以控制，5D 的直方图较其他程序如2D 详细，输出内容：直方图与累积直方图，

正态概率图，去趋势正态图，半正态图，图的标度和大小设定，用于数据分组的标号和符号，分组或未分组资料的描述统计量

6D 双变量(散点) 图

产生一个变量对另一个变量的散点图，并且计算最佳拟合直线。可把记录分组、使用符号区分不同的组，按组别绘制、控制图的大小和所用数据范围等。输出内容：参加绘图的点数，Pearson 相关系数及其p-值，均值和标准差，最小二乘回归曲线及其截距，剩余均方

7D 方差分析和数据筛检

7D 用于数据的筛检和方差分析，第一项功能有：分组变量不同水平上的复合直方图，分组及不分组描述统计量，分组方差相等的Levene 检验，选择方差稳定性转换的Box-Cox 图。第二项功能有：完全随机单方式和双方式方差分析，平衡或不平衡固定效应模型，对比和条件对比，多重比较(Bonferroni, Duncan,Dunnett,Newman-Keuls,Tukey, Scheffe 法)，分组方差不等时的稳健性检验(Welch 与Brown-Forsythe)，使用截尾均值的ANOVA。

8D 相关，不完全资料

当资料有缺失时，利用四种方法计算方差与协方差。ALLVALUE: 对每个变量所有的数据计算均值，使用均值的偏差用于计算方差和协方差。COVPAIR: 使用两个变量都可接受的值计算协方差，而使用每个变量所有可接受的值计算方差。CORPAIR: 仅对两个变量都可以接受的记录计算方差和协方差。COMPLETE: 仅使用完全的样本计算。输出包括：均值，标准差，方差，变异系数；变量对的频数表，权重和，均值和方差矩阵；相关的估计；与不完全资料有关的两两t 检验。

9D 分组的多方式描述

根据一个或多个分组变量计算每个类别的均值和描述统计量。9D 用于产生下述统计量的图示：两个或多个因素析因设计的格点均值，重复测量设计的均值，同时计算两个或多个变量的均值。本程序对评价各组的一致性，在方差分析中观察数据格子均值间存在的趋势和交互有用。其它的如：单向方差分析，卡方检验，各组间均值的变动情况，边缘合计图示(Plots of marginal subsets)。

AM 缺失资料的描述与估计

对于多变量资料描述缺失值的模式，并利用三种处理方法获得协方差阵或相关阵。把缺失值替换成均值或者关于被估变量和其它变量的回归。其它变量是与被估变量的相关最大的一个或者一组高度相关的变量，或者是所有其它变量。输出内容：单变量综合统计量及头五例数据列表，样本对变量的图示，显示缺失值和极值的位置或模式，相关阵和特征值，每变量与其它变量的复相关平方，回归显著性检验，对具有缺失值或超范围值估计的样本列表，估计变量与被估变量间的R-平方，样本到均值的Mahalanobis D-平方，完整样本与带有估计样本的图示。

§6.3.2 F 系列

4F 两维与多维频数表

使用4F 构造、分析及存贮两维、多维或多维表的分表的情况，使用4F 来产生两个或多个分类变量的log-linear 模型。分析内容有：两维表独立性检验：卡方，似然比，Fisher 精确

检验，列联表系数，phi, Cramer's V，Yule's Q 与Y, 交叉乘积比，Yates' 校正卡方，关联情况(Kendall's tau, Somers' D, 及其它)，预测情况(Goodman 与Kruskal's tau 及lambda, 不确定系数)，McNemar's 对称检验，kappa，比例的线性趋势检验；关于对数线性模型，4F 可以拟合：所有可能的模型(两维或三维)，所有饱和模型，每个交互的边缘和偏性关联检验，用户指定的模型，从用户指定的模型中增加或剃除效应；对于每个指定的模型，4F 提供：模型适合度检验，指定模型的预测频数，对数线性模型及其标准误，Freeman-Tukey 量，标化统计量，χ^2 统计量的组分。

§6.3.3 L 系列

1L 寿命表和生存函数

提供了两种估计生存率的方法，即寿命表法和积限方法。进行两组生存曲线的比较，将给定资料列成寿命表的形式。输出内容：Kaplan-Meier统计量，寿命表(Cutler-Ederer)，生存函数图，对数生存函数及累积风险函数，寿命表法的风险和密度函数图，检验生存曲线是否相同的Mantel-Cox, Breslow, 及Tarone-Ware 统计量。

2L 带协变量的生存分析—COX 模型

分析影响生存时间的其它测量变量，分析使用Cox 的比例风险回归模型，模型假设风险函数能用协变量的线性函来表达。量化生存时间与协变量之间的关系，估计回归系数从而给出每个变量对风险函数的相对作用，可以检验表示处理效应的回归系数显著性。这些回归系数以比例风险模型中病人的基线特征为条件。程序提供了逐步方法，处理分组资料。输出包括：每个协变量的回归系数、渐近标准误、标化回归系数；对于极大偏对数似然函数的卡方显著性检验；生存函数图、对数累积风险函数图。

§6.3.4 M 系列

1M 变量的聚类分析

提供了四种相似性的测量方法，三种聚类方法。输出内容：所有类一览表，每步形成的聚类树，聚类过程解释，影子相关阵(Shaded correlation matrix)，相关阵。

2M 记录的聚类分析

根据几种距离的测量方法，进行样本的聚类。输出内容：树形聚类图，amalgamation 距离、每变量均值和新类均值列表，影子相关矩阵，原始标度或标化后的数据列表，样本距离矩阵。

3M 分块聚类(BLOCK CLUSTERING)

对于分类资料形成类别的块，结果把数据矩阵排成分块矩阵。输出内容：识别子矩阵的分块记号，分块的符号表，计数及其值，每种码的频数。

4M 因子分析

四种据相关或协方差阵抽取因子的方法，几种旋转方法。输入数据可以是相关阵或协方差阵、因子载荷或因子得分系数。输出内容：单变量综合统计量，旋转和未旋转因子载荷及其图示，排序和旋转后的因子载荷，因子得分系数及其图示，原始数据的Mahalanobis 距离，因子得分和差值，复相关平方，特征值，排序和影子相关阵，标准得分，协方差阵，相关阵的逆，偏相关，剩余相关。

6M 典型相关分析

两组变量的典型相关，及Bartlett 关于剩余特征值的检验。输入可以是协方差阵或相关阵。输出内容：一元综合统计量和头五个样本数据列表，各变量与其它变量复相关平方，典型相关及相应的特征值和典型变量载荷，典型变量得分和系数，任意变量或典型变量对其它变量或典型变量的双变量图。

7M 逐步判别分析

对于两组或多组资料进行判别分析，可以交互式地每一步指示那一个变量进入或剃除。每组采用刀切法和交叉识别(jackknife-validation) 方法减少偏性。输出内容：每步F 统计量，Wilks' Lambda 或U 统计量(具有近似F 值)和马氏距离，分类函数，矩阵，刀切法分类，正确分类比例，分类一览表，每样本分到各个组时的后验概率及马氏距离，典型判别函数系数，特征值，每个样本的典型得分，头两个典型变量图。

8M BOOLEAN 因子分析

对于二分类资料估计布尔型因子，与传统的方法不同，它在矩阵相乘时采用逻辑算法，因而得分和因子载荷是二分类的。输出内容：每步上偏差为正(# 乘以观察得分为1，估计值为0) 和为负(# 乘以观察得分为0，估计值为1) 的数目、每次循环的总偏差，因子得分，数据矩阵和偏差的Compact 显示。

9M PREFERENCE PAIRS 资料的线性得分

计算每个观察的得分，对观测变量按其评价的重要性进行加权。输出内容：进入线性函数中的变量系数及其t 值，每个样本对的preference matrix, 原始评价值和预期值、误差，在每步结束时每个样本的得分，多种评判下的得分及其相关，变量或得分的散点图。

KM K-MEANS 记录聚类

使用欧氏距离来度量每个样本与每类中心的距离。输出内容：每类中各变量的描述统计量，直方图示类均值到类内和类外样本的距离，样本到三个最大类形成平面的正交投影散点图，类间均方与类内均方的方差分析及F 比，类的轮廓，合并类内协方差和相关阵，类中心距离，类和用户指定变量的交叉表。

§6.3.5 R 系列

1R 多元线性回归

对全部样本、部分样本或多组样本估计多元线性回归方程，检验各组回归线是否相等。输出内容：单变量综合统计量；复相关系数及估计标准误；回归方差分析表；回归系数及其标准误，t 值及标准偏回归系数；相关和协方差阵；残差，预测值及针对每个记录的统计量；残差的散点图、正态概率图及偏残差图。

2R 逐步回归

用逐步方法估计多元线性回归方程，每次进入或剃除的变量可以是一个或者一组，进行向前法或向后法筛选。有四种准则进行逐步筛选，强迫一些变量留在方程中。提供回归诊断功能。输出内容：每步上的：R-平方，调整R-平方, 及估计标准误；回归方差分析表；回归系数及其标准误标准回归系数，容许值，选入方程的F 值；偏回归，容许值，选入F 值(对尚未选入的变量)。同时有：各步一览表；回归系数表；数据、预测值、残差；回归诊断结果；编相关一览表，进入与剃除变量的F 值。

3R 非线性回归

非线性函数的最小二乘参数估计，六种函数及其导数是系统提供的，在FUN 段落中可以指示其它的函数。对于参数可以施加上下界约束和线性等式约束。使用迭代重加权最小二乘法(iteratively reweighted least squares) 获得极大似然解。输出内容：描述统计量；每步上的参数估计，剩余平方和，incremental halvings 数；渐近相关阵及参数标准误，残差的序列相关；每记录因变量的观测值和预测值，残差，权；散点图及正态概率图。

4R 主成分回归

关于因变量和一组主成分进行回归分析，主成分逐个引入，回归系数用主成分或原始或标化变量的形式报告。进入的次序是因变量与主成分的相关大小，能进行岭回归计算。输出内容：特征值、特征向量、累积方差贡献；主成分与因变量间的相关；主成分的回归系数；每步：进入的主成分，剩余平方和，F 比，R-平方，回归系数；每例的主成分值；散点图和正态概率图；在ridge 选项下：R-平方及其剩余平方和，每组岭因子的回归系数；岭迹，R-平方和残差平方和图示。

5R 多项式回归

对因变量拟合关于一个自变量的多项式，采用正交项式计算方法。每个样本可以有自已的权。输出内容：每个正交多项式的t 值、回归系数及其标准误，剩余均方自变量每个幂次的回归系数和剩余平方和，拟合度统计量一览表，回归残差、拟合值、正交多项式的值及对应每记录的有关统计量，散点图和正态概率图。

6R 偏回归与多元回归

计算剃除一组变量的线性效应以后，另一组变量的偏回归。6R 还特别用于多个因变量下的回归，分析使用原始数据、协方差阵或相关阵。输出内容：单变量综合统计量，相关阵，每个自变量与所有其它自变量的R-平方，每个因变量与自变量的R-平方，除去自变量效应后因变量间的偏相关，协方差阵及偏协方差，对应每个因变量的偏回归系数，散点图和正态概率图。

9R 所有子集回归

对于预测变量计算最优子集回归，子集的数目可以指定。子集选择有三种准则：1. 样本R-平方，2. 调整R-平方，3. Mallows 氏Cp。输出内容：对每种子集：小于十个子集的R-平方、调整R-平方及Mallows 氏Cp；对于最优子集：R-平方，调整R-平方，Mallows 氏Cp，估计标准误，回归系数F 检验；对最优子集中的每个变量：回归系数及其标准误，标准回归系数，t 统计量及其p 值、容许值，标准化、删除、加权残差及预测值，散点图和正态概率图；Mahalanobis 距离和Cook 距离；学生化残差的直方图；Durbin-Watson 统计量和序列相关。

AR 不用导数的非线性回归

利用拟高斯—牛顿最小二乘法估计非线性函数的参数，AR 内存六种函数，使用FUN 指定其它函数，参数可以有上下界约束和线性等式约束。AR 能用极大似然法估计参数的函数及其标准误，用于差分方程组的参数估计。输出内容：描述统计量；每次迭代的剩余平方和，参数估计，对分的数目；渐近相关阵估计；渐近标准误估计；每个记录的残差、预测值用其标准误、权、自变量与因变量值；残差和预测值、变量的散点图；加权残差的正态概率图。

LR 逐步LOGISTIC 回归

用逐步法估计线性logistic模型的参数向量。对于分类变量及其交互产生设计变量，在逐步过程中视做一组。在逐步法中，连续变量或一组设计变量同时进入或。剃除。其层次规则是仅当低阶效应和主效应在模型时，高阶交互才进入模型。程序使用的数据既可以是每种不同协变量取值下的表格式数据，也可以是每个对象或样本的单一记录。输出内容：描述统计量(区间尺度的变量) ；不同的取值及其频数(分类变量)；每步的对数似然值及其改变量，拟合度卡方，Hosmer 和C. C. Brown拟合度卡方检验；回归系数及其标准误，它们的比值，系数的渐近相关阵；每步上的进入及剃除统计量；所有步骤的一览表；每组预测概率的直方图；正分与误分表；对分析变量的不同组合提供：成功与失败的频数，预测概率，观察比例，对数比数比，标准残差；第一组比率对其预测概率和对数比数的散点图。对分析变量的不同组合提供：综合描述和图示。

PR 多分类Logistic回归

多项和有序资料的处理，系数的逐步极大似然比估计和近似方差估计，对数似然值和拟合优度。

§6.3.6 S 系列

3S 非参统计

计算下面一个或几个非参统计量：符号检验，Wilcoxon 符号秩次检验，Kruskal-Wallis 单方式方差分析，Kendall 一致性系数，Friedman 两方式方差分析，Mann-Whitney 秩和检验，Kendall 和Spearman 秩次相关系数。

§6.3.7 T 系列

1T 一维与二维谱分析

图示、描述统计及单一或序列对的分析，1T 计算谱分解，绘谱密度图，显示每个频带对时间序列总方差的相对贡献。其选项与特点有：基于协方差的谱估计，谱分析之前的数据处理(Tapering and padding)，带宽和其它指示，Y 关于X 的滞后，协方差或周期图的权，部分时点的谱分析，指示分析的频率范围，可信带，缺失值处理，预滤波和再染色(Prefiltering and recoloring)。

2T BOX—JENKINS 时序分析

使用Box-Jenkins 自回归—积分移动平均方法建立时序模型和传递函数模型。估计模型参数，进行诊断检查或残差分析。提供的方法：时间序列图示，识别(自相关，偏自相关及互相关函数)，季节组分建模，估计(包括参数估计，t 检验，缺失值估计)，残差诊断(包括Ljung-Box Q 统计量)，预测，干预分析，多输入传递函数模型，包括“白化”(”prewhitening” 除去自相关的滤波)。

§6.3.8 V 系列

1V 单方式方差分析及协方差分析

各组间协变量回归系数的平行检验。指示组间或调整均值间的线性对照，并对每个对比进行t 检验。输出内容：每组均值及合并均值；方差分析表和两两t 检验；可选组内统计量：极值、协方差阵和相关阵。指示协变量时有：回归系数，标准误及t 值；各组均值，调整均值及其标准误；检验斜率为零及斜率相等的方差分析表；每个协变量的组内斜率；调整各组均值的两两t 检验；可选的散点图；回归系数及调整均值间的相关。

2V 重复测量资料的方差协方差分析

对各种固定效应和重复测设计进行方差协方差分析，每格子数等或不等。固定效应设计包括完全和不完全析因设计如拉丁方、不完全区组设计、部分析因设计，输出内容：变量的格子均值及标准差，方差分析，每个记录的的预测值及其剩余，调整协变量的格子均值。重复测量设计允许组合重复因素和分组因素，但必须交叉不能嵌套。每个对象必须在重复测量因素的任何组合下有一个反应的取值。输出包括：格子均值，标准差、方差分析表，平方和以及正交组分的相关阵，球型条件的检验，组内因素的正交分解，重复测量因素的保守检验。

3V 一般的混合模型方差分析使用极大似然或约束极大似然方法估计固定和随机模型。混合模型可以很任意，而不需要2V 或8V 那样需要平衡。3V 允许针对特定假设进行检验。输出内容：单变量统计量；模型参数估计及其渐近标准误，t 统计量和p-值；2* 对数似然函数值；参数协方差阵估计；哑变量；固定效应所定义的格子均值、预测均值及其标准误；剩余；指定假设的检验，对数似然值及对数似然比检验，相应的自由度和概率。

4V 重复测量数据的单变量和多变量方差协方差分析

是一个通用程序，处理平衡或不平衡设计和重复测量，裂区和交叉(change over) 设计。输出包括：因素水平上的权或格子的权，格子为空时的分析，多变量分析，同时单变量、重复测量以及多变量分析，针对性处理某些形式的数据缺失，检验用户指定的关于因子水平或格子均值间的对比，用户指定的正交化效应。

5V 有结构协方差阵的不平衡重复测量模型

针对一大类实验设计和模型进行重复测量分析，包括那些协方差阵为特定形式的设计和不完全资料。使用ML 或REML 方法得到回归和协方差估计。实验设计：Longitudinal studies 以及重复测量实验；平衡或不平衡设计，包括由于缺失观察引起的不平衡，时变协变量。协方差结构，完全的指定包括：复合对称(Compound symmetry)，一阶自回归，Banded 或一般自回归，结构未定义(完全参数化)。需要附加输入的有：因子分析，随机效应，线性模型，用户定义的FORTRAN 子程序。

8V 一般混合模型方差分析– 格子大小相同

对任何格子大小相同的完整设计进行方差分析，如区套、交叉或部分嵌套和交叉设计进行方差分析，效应可以是固定的、混合的(包括重复测量) 和随机的。8V 不使用分组变量区分分组，不存在GROUP 段。输出内容：含期望均值的方差分析表，方差分量估计，格子均值、边缘均值、剩余，以及其它可选的统计量。

CA 对应分析

是一个多变量探索性数据分析程序，用于把频数表转成图示。CA 对频数表关联度的分分解类似于连续变量的主成分分析。CA 使用的数据格式可以是记录、标记格子频数、频数表。

DM 数据管理

交互式数据处理程序，与BMDP 各个程序是兼容的，DM 使用BMDP 文件和ASCII 文件，读取多记录类型和层次文件。其功能有：三个过程合并文件，二十四种函数抽取数据信息，压缩和不压缩单记录和多记录数据互换过程，排序、转换。打印、数据存贮，显示记录结构，计算综合统计量。

【附】例6.5的数据。

```
 1   49 1             26 1400 0             54    2 1             79   95 1 66 54 1.08
 2    5 1             27  262 1             55   60 1  9 52 1.51 80  481 0 25 46 1.41
 3   15 1  0 54 1.11 28   71 1 70 54 0.47 56  941 0 66 38 0.98 81  444 0  5 52 1.94
 4   38 1 35 40 1.66 29   34 1             57  148 1             82  427 0
 5   17 1             30  851 1 15 44 1.58 58  342 1 20 48 1.82 83   79 1 31 53 3.05
 6    2 1             31   15 1             59  915 0 77 41 0.19 84  333 1 36 42 0.60
 7  674 1 50 51 1.32 32   76 1 16 64 0.69 60   52 1  2 49 0.66 85    4 1
 8   39 1             33 1586 0 50 49 0.91 61    1 1             86  396 0  7 48 1.44
 9   84 1             34 1571 0 22 40 0.38 62   68 1             87  109 1 59 46 2.25
10   57 1 11 42 0.61 35   11 1             63  841 0 26 32 1.93 88  369 0 30 54 0.68
11  152 1 25 48 0.36 36   99 1 45 49 2.09 64  583 1 32 48 0.12 89  206 1138 51 1.33
12    7 1             37   65 1 18 61 0.87 65   77 1 11 51 1.12 90  185 1159 52 0.82
13   80 1 16 54 1.89 38    4 1  4 41 0.87 66   31 1             91  339 1
14 1386 1 36 54 0.87 40 1407 0 40 48 0.75 67  284 1 56 19 1.02 92  339 0309 45 0.16
15    0 1             41 1321 0 57 45 0.98 68   67 1  2 45 1.68 93  264 0 27 47 0.33
16  307 1 27 49 1.12 42    2 1             69  669 0  9 48 1.20 94  164 1  3 43 1.20
17   35 1             43    1 1             70   29 1  4 53 1.68 96  179 0 12 26 0.46
18   42 1 19 56 2.05 44   39 1             71  619 0 30 47 0.97 97  130 0 20 23 1.78
19   36 1             45   44 1  0 36 0.0  72  595 0  3 26 1.46 98  108 0 95 28 0.77
20   27 1 17 55 2.76 46  995 1  1 48 0.81 73   89 1 26 56 2.16 99   20 1
21 1031 1  7 43 1.13 47   71 1 20 47 1.38 74   16 1  4 29 0.61100   38 0 37 35 0.67
22   50 1 11 42 1.38 48    8 1             75    1 1           101   30 0
23  732 1  2 58 0.96 49 1141 0 35 36 1.35 76  544 0 45 52 1.70102   10 0
24  218 1 82 52 1.62 51  284 1 31 48 1.08 77   20 1           103    5 1
25 1799 0 24 33 1.06 52  101 1             78  514 0209 49 0.81
```

第七章　SYSTAT

§7.1　SYSTAT 应用概要

1983年Leland Wilkinson 就已经拥有微机上SYSTAT版本。历经CP/M、MS-DOS、VAX/VMS, UNIX, DATA General, NCR Tower, IBM PC兼容机及Apple Macintosh 系统。SYSTAT 3.0 和4.1 分别是SYSTAT 公司1986 年和1989 年推出的产品。它们的特征完全相仿。SYSTAT 最显著的特点是模块化功能，目前SYSTAT 也有Windows下的产品，如SYSTAT for Windows 5.0基本上保持原有的模块化特征。

SYSTAT 3.0 系统由12 个相对独立的功能模块组成，这些模块可分成数据处理模块和统计模块。SYSTAT 的数据管理模块是DATA，它用于SYSTAT 的数据预处理并把外部数据文件如ASCII、dBASE、Lotus 格式，用于其系统文件。DATA 模块以外的模块是用做统计分析的。在统计处理过程中，有一些统计模块也能产生存储计算结果的SYSTAT 数据文件，对同一个SYSTAT 文件的数据，可以在多个统计模块中使用，进行不同的统计处理。

SYSTAT 其它的产品有Probit、Logit和Score等。PROBIT 使用累积正态分布估计二分类变量的反应函数，它是一个极大似然程序，可以自动产生哑变量和MGLH 中其它的特征。LOGIT 对二分类数据或多分类数据进行多项logit 模型分析，可以处理更大的模型和数据。SCORE 提供一些检验综合统计量，可靠性系数以及item analysis、Rasch 模型、多项选择或两极尺度(bipolar scales) 问题。REPORT WRITER 是为科学和商务报告而设计的，有格式输出、标题居中、综合统计量以及打印机字型控制。另外，有每页的边界、页长、行宽控制。数字和字符串可以分别定格式而打印。STAT/TRANSFER 模块提供了一种方便的方法，能够在SYSTAT, LOTUS, SPSS/PC，和STATA 之间转换数据，具体操作不过是简单的菜单选择。LAZERTE EDITOR 是一个高速数据编辑器，需要8087 或80287 数学协处理器。Macintosh 版包括下拉式菜单、窗口、以及剪贴板等与其它Macintosh 软件的接口。Mainframe 版: DEC VAX 11/780、MicroVAX 和Hewlett-Packard 9000 大型机上有相应的产品，IBM 大型机版本于1986 年秋问世。

§7.1.1　运行

应注意检查CONFIG.SYS 文件和一个引导文件(名为SYSTAT)，另外还在AUTOEXEC.BAT 文件中设SYSTAT 程序的路径说明，如：PATH C:\DOS;C:\SYSTAT;D:\wp51 在硬盘上运行SYSTAT 很简单，在DOS 引导之后，于根目录下键入：C:\ >SYSTAT 这时可看到由“#”字符组成的SYSTAT 字样出现，按回车后即进入SYSTAT 菜单。只要键入该模块编号或名称后按回车即可。例如要调用STATS模块则键入：

> 3 或

> STATS

这时屏幕清屏，于屏幕底部的左则再次出现SYSTAT 提示符“>”，即表明已进入程序模块。在菜单系统中可以使用HELP 和SUBMIT 命令寻求帮助和运行命令文件。从菜单或程序模块返回DOS 系统，用QUIT 命令。

SYSTAT的各个模块可以分别运行，如：

A>DATA

进入数据管理模块。注意单个模块的运行应有DATA.DEF 文件存在。

程序调入内存后，屏幕上出现由“#”字符组成的SYSTAT 字样，随后下方出现箭头“>”提示符，此时说明已进入SYSTAT 系统，即可开始工作。从一个程序模块转入另一个程序模块时，一般要先退出当前模块(用QUIT 命令)，然后再调入另一模块。如果所使用的程序不在同一张软盘上，则必须更换B 驱动器上的软盘。

在SYSTAT 的执行菜单上追加其它程序也是可能的，这时应编辑文件SYSTAT. DEF. 指定的内容包括命令名、有效文件名、一行或多行的帮助信息。命令名前导以@，必须用大写，文件名应从第十列开始，下面是一个例子。

@WORDPERF \WP\WP.EXE

命令WORDPERF 启用WordPerfect 进行文字处理，使用它来生成SYSTAT 的命令文件或其它文本。文件名应是DOS 的可执行程序。

SYSTAT 作图使用的是IBM 扩展图形字符集。一般的打印机不能打印这个图形集。因此，要从打印机上输出图形，必须改变隐含图形设置。即恢复使用标准ASCII 字符集。则需将DATA 盘中的名为DATA.DEF 文件与名为GENERIC.DBF 文件名互换。

SYSCROLL 可以浏览系统进行过的操作。程序允许在内存中保存最多九屏的输出，由于未采用直接视频显示，所以运行速度较慢，也不与DOS 的其它命令冲突。使用DOS 通常的办法运行SYSCROLL. EXE ，默认保存四屏内容，可以使用SYSCROLL 2 或SYSCROLL 9 等等来调节屏数。

程序驻留后，使用Ctrl- 健来激活SYSCROLL，其时功能键有：

PgUp - 上滚一屏　Up arrow - 上卷一行　Home - 窗口顶部

PgDn - 下翻一屏　Dn arrow - 下卷一行　End - 窗口的底

再次使用Ctrl- 则返回运行的程序。

§7.2 SYSTAT 命令和模块

§7.2.1 SYSTAT 命令

进入SYSTAT 模块后，显示器屏幕上会出现SYSTAT 字样，并显示'>'。这时，就可以打入该模块的命令，命令计算机进行相应的操作。若命令较长，一行打不下，可在这行结束处打一个逗号，表明命令未输入完，然后转下一行继续输入。

㈠. 命令特点

SYSTAT 的命令有冷热之分。打入热命令，计算机立即执行，并给出执行结果。而打入冷命令时，计算机并不立即执行。实际上打入冷命令，只是作了某种选择，或指定了某个条件。一般地说，冷命令的次序可以任意，但必须在打入一个热命令之前，把所需要的冷命令全部打完。对于SYSTAT 的命令，计算机只辨认其前两个字母。因此，输入命令时，可以只打入命令的前两个字母。

在SYSTAT 系统中，所有模块都能用USE 命令读取SYSTAT 文件中数据，而用SAVE 命令来存写数据，但只有数据模块能够读取来自其他途径的数据。

㈡. 公用命令

在SYSTAT 中，有些命令可以在所有的模块中使用，这些命令称为公用命令。常用的公用命令有：

1. BY：指定一个或多个分组变量，命令格式为：

   BY <变量> [,<变量>,<...>] . 如：

BY AGE (指定数值变量AGE 为分组变量)

BY SEX,NAME$ (指定变量SEX 和变量NAME$ 为分组变量)

BY 命令后面的变量必须是经过排序的。重新用USE 命令打开文件或打入一个后面无变量的BY 命令都可取消以前指定的分组变量。

2. FORMAT：规定输出数据小数点后的位数，命令格式为：

FORMAT=5 (规定保留5位小数)

FORMAT=3 (规定保留3位小数，即默认值)

FORMAT 命令后面的数字不得小于0，不得大于9。

3. HELP：输出帮助信息，命令格式为：

HELP [<命令>]，如：

HELP (显示有关模块的帮助信息)

HELP BY (显示有关BY命令的帮助信息)

4. OUTPUT：指定输出装置，命令格式为：

OUTPUT * (指定显式器为输出装置)

OUTPUT @ (指定打印机为输出装置)

OUTPUT <文件>，如：

OUTPUT RESULT (建立ASCII 码的磁盘文件RESULT.DAT，存储输出结果)

一旦打入OUTPUT 命令，此后的输出结果将从指定装置输出。

5. QUIT：终止SYSTAT 的运行。命令格式为：

QUIT

6. SAVE：建立一个新的SYSTAT 数据文件，格式为：

SAVE <文件>，如：

SAVE MANOVA (建立名为MANOVA.SYS 的数据文件)

不需输入文件的扩展名'.SYS'，SAVE 命令会自动地加上。

7. SELECT：规定一项或多项选择数据的标准，命令格式为：

SELECT <变量>=<数值><字符变量>=<字符串>< ... >

SELECT STATE$='NY'

SELECT REGION=4 STATUS=2

打入后面无选择标准的SELECT 命令可取消以前所作的规定。

8. SUBMIT：从扩展名为'.CMD'的命令文件中取出命令并执行。命令格式

SUBMIT <文件>，如：

SUBMIT MYFILE

打入这个命令后，计算机就寻找文件MYFILE.CMD，并依次执行文件中的所有命令。

9. USE：打开一个已存在的SYSTAT数据文件，命令格式为：

   USE <文件>，如：

   USE OLDFILE (打开文件OLDFILE.SYS，准备读取数据)

   执行USE命令将显示被打开的数据文件的所有变量名称。

10. WEIGHT：指定一个权数变量，命令格式为：

    WEIGHT=<变量>，如：

    WEIGHT=NUMBER (指定变量NUMBER 为权数变量)：

只有QUIT 命令是热命令，其余均为冷命令。

㈡. SYSTAT 文件和变量

SYSTAT 系统有四种文件：SYSTAT 数据文件，ASCII 码的字符文件，SYSTAT 命令文件和SYSTAT 的临时文件。四种文件的扩展名分别为.SYS .DAT .CMD 和.TMP。SYSTAT 中的文件名由1–8个字母或数字构成，但必须是字母打头。为了说明文件盘所在的驱动器，可以在字母和数字组成的名字前面加上驱动器符。在SYSTAT 中不需输入文件的扩展名，计算机会自动加上。

SYSTAT 中的变量有两类：数值型变量和字符型变量。数值变量其取值均为数值，而字符变量的取值均为字符。数值变量名的构成与文件名相同，由1–8个字母或数字组成，不同之处是在变量名中可以使用字符'_'。字符变量名是在数值变量名之后再加上一个字符'$'。

在SYSTAT 中，无论数值变量或是字符变量都可以加上最多两位数的下标，并且在一些命令中还可以指定下标的范围。

§7.2.2 数据和统计模块

㈠DATA 模块

DATA 模块是SYSTAT 软件包中唯一的数据模块。它能够接受来自键盘，ASCII 码文件和已存在的SYSTAT 文件的数据，经过整理加工，生成新的SYSTAT 数据文件。

1. DATA 模块的命令

   (a) APPEND：将两个具有相同变量的SYSTAT 文件串连，串连生成数据文件例数为两文件数据例数之和，命令格式：
   APPEND <文件> <文件>，如：
   APPEND FILE1 FILE2 把文件FILE2.SYS 追加到文件FILE1.SYS。

   (b) DELETE：删除当前例数据。

   (c) DROP：删除指定的变量，命令格式：
   DROP <变量> [,<变量>,< ... >]，如：
   DROP SEX,NAME$ (删除数值变量SEX 和字符变量NAME$)

   (d) EDIT：进入全屏幕编辑器，命令格式：
   EDIT [<文件>]，如：
   EDIT (进入编辑器，编辑一组新数据)
   EDIT AFILE (进入编辑器，编辑文件AFILE.SYS 的数据)

(e) GET：打开一个扩展名为.DAT 的ASCII 码文件，命令格式：

GET <文件>，如：

GET ASCFILE (打开ASCII 码文件ASCFILE.DAT)

(f) IF：规定一个比较条件，并判断其是否满足，命令格式：

IF <条件> THEN <命令>，如：

IF CASE=4 THEN LET AGE=39

(g) INPUT：规定输入数据的变量个数，以及各变量的名称和性质，命令格式：

INPUT <变量>[,<变量>, < ... >]，如：

INPUT AGE NAME$

(h) LET:计算表达式值，赋给变量，命令格式：

LET <变量>=<表达式>，如：

LET LAGE=LOG(AGE) (计算变量AGE 的对数赋给变量LAGE)

(i) LIST:输出文件中所有或部分变量的数据，命令格式：

LIST [<变量>, < ... >]，如：

LIST (输出文件中所有变量的数据)

LIST AGE,NAME$ (输出文件中变量AGE 和NAME$ 的内容)

(j) LRECL:规定各类输入数据的读取长度，命令格式：

LRECL=<数值>，如：

LRECL=256 (对每例数据，读前256 个子符，后面的不读)

此命令的缺省值为80，即每例数据只读前80 个子符。

(k) PUT:建立一个扩展名为.DAT 的ASCII 码文件，命令格式：

PUT <文件>，如：

PUT ASCFIL (在磁盘上建立文件ASCFIL.DAT)

(l) RUN: 执行此命令之前的冷命令所规定的任务，命令格式：

RUN

(m) SORT:将文件中所有或部分变量的数据，按其值的大小，从小到大排序。命令格式：

SORT[<变量>, < ... >]，如：

SORT (将文件中所有变量的数据排序)

SORT AGE (将文件中变量AGE的数据排序)

(n) USE: 打开SYSTAT 文件，指定可读取数据的变量；并连两个SYSTAT 文件中的所有或部分变量，命令格式：

USE <文件>[<变量>, < ... >][<文件>[<变量>, < ... >]]，如：

USE DATAFILE (打开文件DATAFILE.SYS)

USE DATAFILE(AGE,NAME$) (指定文件DATAFILE.SYS 中AGE，NAME$ 为可读取数据的变量)

USE FILE1,FILE2 (并连文件FILE1.SYS和文件FILE2.SYS的所有变量)

USE FILE1(AGE) FILE(NAME) (将文件FILE1 中的变量AGE 和文件FILE2 .SYS 中的变量NAME 并连起来)

(o) NEW: 取消在此之前的所有命令，并清除数据空间，命令格式：

NEW

以上15 条命令中，APPEND,RUN 和NEW 命令是热命令，其余均为冷命令。

2. 从键盘输入数据

从键盘输入的数据，有两种方法：

(a) 在DATA 模块中直接输入。

在操作系统状态下，进入DATA 模块。然后打入：

SAVE MYFILE

INPUT AGE NAME$

RUN

这时屏幕显示：

INPUT DATA ONE CASE AT A TIME AFTER PROMPT ARROW

现在可以输入数据。注意按例输入，一行输入一例，每行末尾应按一下回车键。不要打'>'。下面是屏幕上看到的输入内容：

```
>33 YANGHONG
>22 WANGWEI
>24 LIMING
>36 ZHANGJIE
>26 YUANPING
>37 LIZHIQIANG
>42 WANGHONG
```

数据输入完时，应在下一个'>'出现之后，打入'~'。这时应看到：

```
     7 CASES AND 2 VARIABLES PROCESSED
SYSTAT FILE CREATED.
WORKSPACE CLEAR FOR CREATING NEW DATASET
```

至此，一个内含2 个变量7 例数据的SYSTAT文件就被建立在磁盘上，其名字为MYFILE.SYS。如果想看一下输入的数据，输入命令：

USE MYFILE

屏幕上显示：

SYSTAT FILE VARIABLES AVAILABLE TO YOU ARE:

AGE NAME$

因为只是简单地看一下数据，所以只需再打入如下命令：

LIST

RUN

这时屏幕上出现：

```
                    AGE          NAME$
CASE   1         33.000         YANGHONG
CASE   2         22.000          WANGWEI
CASE   3         24.000           LIMING
CASE   4         36.000         ZHANGJIE
CASE   5         26.000         YUANPING
CASE   6         37.000       LIZHIQIANG
CASE   7         42.000         WANGHONG
```

前面打入的命令，RUN 是热命令，其余均为冷命令。

(b) 利用全屏幕编辑器输入。

在DATA 模块中有一个全屏幕编辑器，利用它可以很方便地建立SYSTAT 文件。

在DATA 模块中，打入EDIT 命令即可进入全屏幕编辑器。进入编辑器，屏幕上会出现一个数据表格，光标在变量名行上。首先按下面顺序输入所有变量名：

'AGE <Enter> 'NAME$ <Enter>

然后按<Home>键使光标转到数据区左上角。再开始输入数据，按例输入，每输入一个数据就应按一下回车键。实际输入顺序为：

```
33    'YANGHONG   <Enter>
22    'WANGWEI    <Enter>
24    'LIMING     <Enter>
36    'ZHANGJIE   <Enter>
26    'YUANPING   <Enter>
37    'LIZHIQIANG <Enter>
42    'WANGHONG   <Enter>
```

数据输入完了应按<Esc>键，使光标转到命令行，然后打入SAVE MYFILE 将输入的数据存入文件MYFILE.SYS。最后打入QUIT 命令退出全屏幕编辑器。

3. ASCII 码文件数据的转换

为了实现与其他软件的数据交换，达到数据共享的目的，DATA 模块提供了ASCII 码文件与SYSTAT 文件相互转换的功能。

(a) ASCII 码文件转换成SYSTAT 文件。首先，检查待转换的磁盘文件是否为ASCII 码文件。此外，还应保证文件扩展名为.DAT 。然后，在操作系统状态下打数据模块的名字，进入DATA 模块，打入如下命令：

```
GET ASCFILE
INPUT AGE NAME$
SAVE MYFILE
RUN
```

这样，一个名为ASCFILE.DAT 的ASCII 码文件就被转换成SYSTAT 文件MYFILE.SYS。若ASCII 码文件有些数据的列数超过80 列(比如最多是236 列)，则应在热命令RUN 之前输入：LRECL=236

(b) SYSTAT 文件转换成ASCII 码文件。

这一转换的方法很简单，只需在DATA 模块中打入如下命令：

```
USE MYFILE
PUT ASCFILE
RUN
```

就可将SYSTAT 文件MYFILE.SYS 转换成ASCII 码文件ASCFILE.DAT。

4. SYSTAT 文件的再加工

在实际统计分析过程中，常常需要对已存在的SYSTAT 文件的数据重新整理，经过取舍组合，加工成新的SYSTAT文件。

(a) 对一个SYSTAT 文件的再加工。

在DATA 模块中可对一个SYSTAT 文件中的数据，实施排序，转换，删除变量或数据等操作。

①排序：在DATA 模块中打入如下命令：

```
USE MYFILE
SORT AGE
SAVE AGESORT
RUN
```

就可将文件MYFILE.SYS 的数据，按变量AGE 数值大小，从小到大排序并存入新的SYSTAT 文件AGESORT.SYS。

②转换：利用转换命令LET 可将文件中的数值变量X 转换成它的某种函数f(X)，函数f(X) 指用运算符将变量和标准函数连接起来的表达式。下面的命令：

```
USE DATAFILE
LET SAGE=SQR(AGE)+.5
SAVE AGESQR
RUN
```

将文件DATAFILE.SYS 包含的变量AGE，求其平方根加上0.5 作为新的变量SAGE，并存入新的SYSTAT文件AGESQR.SYS。

SYSTAT 中的标准函数及运算符有：

	标准函数		运算符
SQR(X)	平方根函数	+	加号
LOG(X)	自然对数函数	-	减号
EXP(X)	指数函数	*	乘号
ABS(X)	绝对值函数	/	除号
SIN(X)	正弦函数	^	乘方号
COS(X)	余弦函数	<	小于号
TAN(X)	正切函数	=	等于号
ASN(X)	反正弦函数	>	大于号
ACS(X)	反余弦函数	<>	不等于号
ATN(X)	反正切函数	<=	小于或等于号
INT(X)	取整函数	=>	等于或大于号

③删除变量：如：

```
USE DATAFILE
SAVE AGEFILE
DROP NAME$
RUN
```

即可删除文件DATAFILE. SYS 中的字符变量NAME$，并将结果存入文件AGEFILE.SYS 中。

④删除部分数据：利用条件命令IF 规定删除数据的条件，然后用删除命令DELETE 把符合条件的数据删除掉。输入命令：

```
USE DATAFILE
SAVE AFILE
IF CASE>5 THEN DELETE
RUN
```

其执行结果：删掉了文件DATAFILE.SYS 中的第6，7 例数据，并把未被删除的数据存入文件AFILE.SYS 中。

(b) 对两个SYSTAT 文件的再加工。

在DATA 模块中，可以将两个SYSTAT 文件的数据并连或串连。

①合并文件：将两个文件的变量并列，合成一个包括所有变量的新文件。若两文件的变量不相同，则新文件的变量个数是两个文件变量个数之和。并连文件的命令如下：

```
USE FILE1 FILE2
SAVE ALLFILE
RUN
```

此命令序列可将文件FILE1.SYS 和FILE2.SYS 的数据并连起来，存入新文件ALLFILE.SYS 中。

②追加文件：将两个具有相同变量的文件顺序衔接，形成一个新文件。新文件的数据例数是两个文件数据例数之和。顺接文件的命令如下：

```
SAVE ALLFILE
APPEND MANAGE1 MANAGE2
RUN
```

此命令序列可将文件MANAGE1.SYS 和MANAGE2.SYS 的数据顺接起来，存入文件ALLFILE0.SYS中。

5. 数据的修改

对于已建立的SYSTAT 文件中的错误数据，可以用下面两种方法修改：

(a) 利用全屏幕编辑器修改

在DATA 模块，打入跟有文件名的EDIT 命令，计算机把文件中的数据读入全屏幕编辑器。这时就可以使用光标移动键，把光标移到一个待修改数据的位置上，输入正确的数据。这样一个错误的数据就修改完了。依次重复上述步骤，直到文件中所有错误数据都修改完毕。打入SAVE 命令，把修改后的数据存入一个新文件。最后用QUIT 命令退出全屏幕编辑器。

(b) 利用条件和转换命令修改

在DATA 模块中，打开待修改的文件，用条件命令IF 和转换命令LET 组合使用，修改错误数据。比如，文件DATAFILE.SYS 中第4 例变量AGE 的值应是39，变量NAME$ 的'ZHANGJIE' 应为'GAODA'。欲以改正，打入如下命令

```
USE DATAFILE
IF CASE=4 THEN LET AGE=39
IF NAME$='ZHANGJIE' THEN LET NAME$='GAODA'
RUN
```

㈡GRAPH 模块

GRAPH 模块是一个统计模块，它主要功能是根据SYSTAT 文件中的数据，按照使用者的绘图命令绘制各种统计图，并通过指定输出装置输出。GRAPH 模块既可以绘制常用的统计图如：直方图，条图，散点图和概率图；也可以绘制一些较少使勇的新型统计图，如：茎叶图和盒式图。此外，若数据是多组的，还可使用BY命令指定分组变量，绘制各组的统计图。

BAR：绘制所有或部分指定的数值变量，字符变量的条图。命令格式：

BAR [<变量><变量>< ... >][/CUM LOW=¡数值¿ WIDTH=¡数值¿] . 如

BAR (对文件中所有变量各绘制一个条图)

BAR TYPE,MONTH$ (绘制指定变量TYPE, MONTH$的条图)

BAR TYPE/CUM (绘制数值变量TYPE的累计条图)

我们用一组关于医院工作质量的数据来说明绘制统计图的具体步骤。数据已经存入名为HOSPITAL.SYS 的数据文件中，它含六个变量：有效率(X1)，病死率(X2) ，平均住院日(X3)，病床周转率(X4)，病床使用率(X5)和分组变量(GROUP)。下面就是在DATA 模块中列出的这组数据，一共是12 例。

```
                X1       X2       X3       X4       X5     GROUP
CASE    1     94.270    2.020   15.990   17.750   84.190       1
```

CASE	2	94.060	2.080	14.230	16.480	82.680	1
CASE	3	95.080	1.570	13.240	20.090	81.680	2
CASE	4	94.480	2.010	15.360	16.390	80.160	2
CASE	5	94.740	1.850	15.810	17.770	83.510	2
CASE	6	94.940	1.810	16.580	16.960	83.570	2
CASE	7	95.250	1.820	16.800	16.630	83.900	2
CASE	8	93.430	2.410	15.860	17.240	81.570	1
CASE	9	94.120	2.000	16.000	16.120	82.240	1
CASE	10	93.360	2.080	16.120	17.150	83.360	1
CASE	11	94.090	2.120	16.240	16.030	83.340	1
CASE	12	96.000	1.800	16.120	16.220	82.420	2

实际上，在GRAPH 模块中，绘制统计图的方法非常简单。我们只要用USE 命令打开包含待绘图变量的文件，就可以根据统计分析的需要，打入不同的绘图命令，计算机将立即执行打入的每条绘图命令，对指定的变量绘制相应的统计图，并把结果从指定的输出装置上输出出来。

下面用HOSPITAL.SYS 文件中的数据绘制统计图：

进入GRAPH 模块，打入命令USE HOSPITAL，打开待处理的文件，输入绘图命令绘制相应的统计图。如要绘制变量X1的直方图，只需打入命令HISTOGRAM X1。

如果我们要进一步考察变量X1的分布是否为正态分布，可以绘制这个变量的正态概率图。即打入命令PPLOT X1，执行结果将出现在显示器上。

利用BAR 命令绘制条图。如绘制X1 的条图，命令为：

BAR X1/LOW=93,WIDTH=1 (X1的最小值取93, 组距取1)

其它绘图命令的使用方法与之类似。

㈢STATS 模块

STATS 模块是一个基本统计模块。它的主要功能是计算各种统计量，如：均值(MEAN)，标准差(SD)，偏度系数(SKEWNESS)，峰度系数(KURTOSIS)，极大值(MAX)，极小值(MIN)，全距(RANGE)，方差(VARIANCE)，标准误(SEM) 和数据和(SUM)。对于分组数据，STATS 模块不仅可以计算各组的统计量，还能对各组的均值作t-检验或方差分析。

1. STATISTICS：计算所有或部分指定变量的统计量。可输出的统计量有：均值，标准差，标准误，偏度系数，峰度系数，最大值，最小值，全矩，方差和总和。命令格式为：

   STATISTICS [<变量>< ... >] [/MEAN SD SEM SKEWNESS KURTOSIS MAX MIN RANGE VARIANCE SUM]，如：

   STATISTICS (计算文件中所有数值变量的均值，标准差，极大值和极小值)

   STATISTICS TREAT/MEAN SEM VARIANCE (计算变量TREAT 的均值，标准误，方差)

2. TTEST：对指定变量作配对t检验或分组t检验。格式如下

   TTEST <变量> [<变量>< ... >][* <变量>]，如：

   TTEST X1 X2 (将变量X1和X2配成对，作配对t检验)

   TTEST X*SEX (根据变量SEX分组，对变量X 作分组t检验)

TTEST X1 X2 X3 (将三个变量X1，X2，X3 两两配对，分别作配对t检验)

TTEST X Y*SEX (根据变量SEX 分组，分别对变量X，Y作分组t检验)

上面两条统计命令均为热命令。

3. PRINT：规定结果输出的等级。其命令格式如下：

PRINT=SHORT (规定仅仅输出基本的计算结果)

PRINT=LONG (规定除基本结果之外，还输出更详细的信息)

这是一条冷命令。除在STATS模块中可使用外，它还可以在后面介绍的几个模块中使用。

以下用文件HOSPITAL.SYS说明使用STATS 模块的统计命令。进入STATS 模块，打入命令FORMAT=5,规定输出结果保留五位小数；尔后就可以作如下的统计处理：

假若我们要计算文件HOSPITAL.SYS 中的三个变量X1，X2，X3 的均值，标准差，偏度和峰度系数，打入命令：

USE HOSPITAL

STATISTICS X1 X2 X3/MEAN SD SKEWNESS KURTOSIS

计算结果:

```
TOTAL OBSERVATIONS:   12

                          X1          X2          X3
 N OF CASES               12          12          12
 MEAN                 94.485       1.964      15.696
 STANDARD DEV          0.763       0.212       1.008
 SKEWNESS              0.310       0.218      -1.447
 KURTOSIS             -0.452       0.240       1.163
```

STATS模块可以作分组t-检验和配对t-检验。下面以医院工作质量为例，进行分组t-检验：

USE HOSPITAL

TTEST X1*GROUP

在TTEST 命令中分组的变量必须在星号后面。分组t-检验的结果如下：

```
INDEPENDENT SAMPLES T-TEST ON       X1     GROUPED BY    GROUP

     GROUP         N     MEAN           SD
       1.000       8       94.069        0.474
       2.000       4       95.318        0.472

  SEPARATE VARIANCES T =       4.312 DF =  10.0 PROB = .002
   POOLED VARIANCES T =       4.306 DF =    10 PROB = .002
```

现在，我们看一下两组医院工作质量X1,X2 X3的均值，标准差和它们的极值，可打入下面命令：

```
USE HOSPITAL
BY GROUP
STATISTICS X1 X2 X3
```

BY 命令指定以变量GROUP的值分组，前提条件是GROUP 已排序。

如果在命令STATISTICS 之前，还打入命令PRINT=LONG，除了得到所计算的统计量外，得到均数差别的显著性检验。

其中BARTLETT TEST 是两组间方差齐性检验，大样本时计算的是Bartlett 卡方值；若例数小于10 时，则改用计算小样本的近似F 值(APPROXIMATE F) 。OVERALL MEAN 是两组合并的均数。最后一行的T STATISTICS 是按POOLED T检验计算的统计量。

如果有一个文件是按一个或多个分组变量分组的，而且我们想再建立一个文件把各组的统计量存起来，那么只要将命令SAVE、BY和STATISTICS配合起来使用，就能够完成这个任务。下面的命令序列

```
USE HOSPITAL
BY GROUP
SAVE MEANHOS
STATISTICS X1, X2, X3/MEAN
SAVE SDHOS
STATISTICS X1, X2, X3/SD
```

就可将所计算的三个变量的各组均值存入文件MEANHOS.SYS，而把计算的标准差存入文件SDHOS.SYS。注意：由BY命令所规定的每一组在新文件中只是一例数据。

㈡TABLES 模块

TABLES 模块也是一个基本统计模块，它可以产生各种维数的表，并能用对数线性模型加以拟合，还可以对其拟合结果作卡方拟合检验。TABULATE: 产生一个一维的或多维的表。其格式为

TABULATE <变量> [* <变量> * < ... >] [/FREQUENCY PERCENT ROWPCT COLPCT LIST]，如：

TABULATE AGE (产生一个按变量AGE 分组的一维频数表)

TABULATE AGE*SEX (产生一个按变量AGE，SEX 交叉分组的二维频数表)

TABULATE AGE*SEX*STATE$ (产生一个按变量AGE，SEX 和字符变量 STATE$ 交叉分组的三维频数表)

TABULATE AGE，SEX*STATE$ (产生两个二维频数表)

TABULATE AGE*SEX/PERCENT (产生一个以总计数为分母的二维百分数表)

TABULATE AGE*SEX/ROWPCT (产生一个以行合计数为分母的二维百分数表)

TABULATE AGE*SEX/COLPCT (产生一个以列合计数为分母的二维百分数表)

TABULATE AGE/LIST (以清单形式列出频数表)

MODEL：指定一个对数线性模型，用以拟合前面TABULATE命令产生的表，并对其结果作拟合检验。

MODEL <变量> + <变量> + < ... > + <变量> * <变量> + < ... > [/FITTED DIFFERENCES RESIDUALS]，如：

表 7.1 护理工作评分比较资料

患者评分	科主任评分		
	差	一般	好
差	20	15	15
一般	30	30	20
好	25	30	15

MODEL AGE+SEX+AGE*SEX (规定一个二维模型，对指定的频数表拟合，并输出拟合值)
MODEL AGE*SEX (规定一个二维模型，与前面命令相同)
MODEL AGE+SEX (规定一个忽略交互作用的二维模型)
MODEL AGE*SEX+AGE*STATE+SEX*STATE (规定一个忽略变量AGE，SEX 和STATE的交互作用的三维模型)
MODEL AGE*SEX/RESIDUALS (规定一个二维模型，对指定的频数表拟合，并输出剩余值)
MODEL AGE*SEX/FITTED RESIDUALS (规定一个二维模型，对指定的频数表拟合，并输出拟合值和剩余值)

所谓n 维表就是依据n 个定性或等级变量将考察对象分组，数各组的考察对象个数而获得的频数表。

按一个定性或等级变量分组而得到的频数表称一维表，也就是通常意义上的频数表。按两个定性或等级变量交叉分组而得到的频数表称二维表，也就是通常所说的R x C 表。若依据分组的定性或等级变量多于两个，则称获得的频数表为多维表。

以下面的二维表为例说明在DATA 模块中建立产生n 维表的数据文件的方法。这是一个关于某医院评价护理人员工作质量的调查资料。在评价时，由患者和科主任同时给护士评分，评分结果分为好、一般和差三个等级。见表 7.1

首先，为两个等级变量取名，并将其值进行编码。即

科主任评分	GRADE1	GRADE1=1	差
		GRADE2=2	一般
		GRADE3=3	好
患者评分	GRADE2	GRADE2=1	差
		GRADE2=2	一般
		GRADE2=3	好

然后，进入DATA 模块，输入如下命令：

```
SAVE TABLE
INPUT GRADE1 GRADE2 FREQUEN
RUN
1 1 20
1 2 30
```

```
1 3 25
2 1 15
2 2 30
2 3 30
3 1 15
3 2 20
3 3 15
```

上述命令将在磁盘上建立一个名为TABLE.SYS的SYSTAT数据文件。注意文件中除两个等级变量GRADE1和GRADE2之外，还有一个数值变量FREQUEN，它是为存储二维表格内的频数而设立的。

在我们的例子中，各格子里的频数是按一种特殊的顺序输入的，也可以不这样做。事实上，在输入数据时，格子可以重复，TABULATE命令会自动地把相同格内的频数加在一起。

产生N 维表的具体步骤是：先用USE命令打开数据文件，然后打入WEIGHT命令指定频数变量，最后输入TABULATE命令产生所期望的N 维表。下面以前面建立的TABLE.SYS 文件为例，打入命令：

```
USE TABLE
WEIGHT=FREQUEN
TABULATE GRADE2*GRADE1
```

显示下面结果：

```
TABLE OF   GRADE2     (ROWS) BY    GRADE1    (COLUMNS)
FREQUENCIES                1          2          3     TOTAL
                    ---------------------------------
                  1 |     20         15         15   |     50
                  2 |     30         30         20   |     80
                  3 |     25         30         15   |     70
                    ---------------------------------
           TOTAL          75         75         50        200
```

如果TABULATE 命令加上相应的选择项，还可以得到以行合计，列合计或总合计为分母的百分数表。

对于二维表来说，如果在上面命令序列中的TABULATE命令之前，还打入了PRINT=LONG命令，那么，我们除了可以得到产生的表之外，还可以得到这个表的假设检验的结果。本例有：

```
TEST STATISTIC                      VALUE          DF          PROB
  PEARSON CHI-SQUARE                2.286           4          .683
  LIKELIHOOD RATIO CHI-SQUARE       2.305           4          .680
  MCNEMAR SYMMETRY CHI-SQUARE       9.500           4          .023

COEFFICIENT                         VALUE       ASYMPTOTIC STD ERROR
  PHI                               .1069
```

```
CRAMER V                          .0756
CONTINGENCY                       .1063
GOODMAN-KRUSKAL GAMMA            -.0203            .09709
KENDALL TAU-B                    -.0133            .06388
STUART TAU-C                     -.0131            .06283
COHEN KAPPA                      -.0093            .04845
SPEARMAN RHO                     -.0149            .07120
SOMERS D    (COLUMN DEPENDENT)   -.0134            .06395
LAMBDA      (COLUMN DEPENDENT)    .0400            .05813
UNCERTAINTY (COLUMN DEPENDENT)    .0053            .00697
```

一般表的假设检验则主要靠MODEL 命令进行的，即在产生表之后，打入MODEL 命令对指定模型进行假设检验，检验的结果将输出。我们还以前面二维表为例，说明一般表假设检验的步骤，命令序列为：

```
USE TABLE
WEIGHT=FREQUEN
TABULATE GRADE2*GRADE1
MODEL GRADE1+GRADE2
```

结果：

```
MODEL WAS FIT AFTER  2 ITERATIONS.
TEST OF FIT OF MODEL
DEGREES OF FREEDOM =   4
PEARSON CHI-SQUARE =       2.29    PROBABILITY =  .683
LIKELIHOOD RATIO CHI-SQUARE =       2.30    PROBABILITY =  .680
```

㈣CORR 模块

CORR 模块它是一个专门用来计算数据集合内变量间的相关或相似系数矩阵的统计模块。通过该模块的计算结果可以考察各变量的关联程度，为进一步的统计处理作准备。现介绍五个命令：

1. SSCP：计算文件中所有或指定数值变量的离均差平方和及各变量间的离均差叉积和，命令格式：

   SSCP [<变量>, <变量>, < ... >]，如：SSCP (计算文件中所有数值变量的离均差平方和及叉积和) SSCP HEIGHT，WEIGHT (计算变量HEIGHT和WEIGHT的离均差平方和及叉积和)

2. COVARIANCE：计算文件中所有或指定数值变量的方差及各变量间的协方差，命令格式：

   COVARIANCE [<变量>, <变量>, < ... >]，如：

   COVARIANCE (计算文件中所有变量的方差及协方差)

   COVARIANCE HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的方差及协方差)

3. PEARSON：计算文件中所有或指定数指变量间的PEARSON 乘积矩阵相关系数，命令格式：

   PEARSON [<变量>,<变量>,< ... >]，如：

   PEARSON (计算文件中所有数值变量的相关系数)

   PEARSON HEIGHT，WEIGHT (计算指定变量HEIGHT 和WEIGHT 的相关系数)

4. SPEARMAN：将文件中所有或指定数值变量排序编秩，计算SPEARMAN 等级相关系数，命令格式：

   SPEARMAN [<变量>,<变量>,< ... >]，如：

   SPEARMAN (计算文件中所有数值变量的等级相关系数)

   SPEARMAN HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的等级相关系数)

5. EUCLIDEAN：计算文件中所有或指定数值变量间的欧氏距离，并用样本含量N 相除得其平均距离，命令格式：

   EUCLIDEAN [<变量>,<变量>,< ... >]，如：

   EUCLIDEAN (计算文件中所有数值变量的平均距离)

   EUCLIDEAN HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的平均距离)

在CORR 模块中，统计命令的使用很简单.只要打开数据文件，就可打入上述统计命令中的任意一条。计算各变量间的相应统计量。下面以计算PEARSON 相关系数为例进行讨论。以前面的医院工作质量数据为例，计算X1，X2 ，X3 之间的PEARSON相关系数，命令如下：

```
USE HOSPITAL
FORMAT=5
PEARSON X1 X2 X3
```

命令序列中FORMAT=5 是用来规定输出结果的小数位数，计算结果：

```
PEARSON CORRELATION MATRIX
                    X1          X2          X3
     X1            1.00000
     X2           -0.80204     1.00000
     X3            0.00503     0.27146     1.00000
NUMBER OF OBSERVATIONS:   12
```

如果打算在CORR 模块存储某个统计命令的计算结果的话，只需在相应统计命令之前打入形如：SAVE <文件名> 的命令。这样做就可以把后面统计命令的计算结果(如相关系数)，存储在以SAVE 命令中的¡文件名¿ 为名的SYSTAT 数据文件中，利用SYSTAT 中的其他模块，可以对这个文件的数据进行各种统计处理。下面的命令是计算文件HOSPITAL 中三个变量的相关系数，并将其存入指定的名为CORRHOS 文件中，命令如下：

```
USE HOSPITAL
SAVE CORRHOS
PEARSON X1 X2 X3
```

㈤MGLH 模块

MGLH模块是一个高级统计模块。它不仅能够估计各种单变量或多变量的一般线性模型的参数，而且能够对其参数的线性假设进行检验。因此，MGLH 模块的功能比大多数的回归程序要强得多。利用它可以很容易地进行简单回归(一个因变量对一个自变量的回归)，多元回归(一个因变量对多个自变量的回归) 和多因变量的多元回归(多个因变量对多个自变量的回归)；各种试验设计的单元或多元方差分析；以及其他一些多元统计分析方法，如多变量断面分析，线性判别分析，典型相关分析等等。这里介绍其中8 条命令的功能及格式。

1. CATEGORY：指定一个或多个数值变量为分组变量。并限定分组的个数。命令格式：

   CATEGORY <变量>=<数值> [<变量>=<数值>, < ... >]，如：

   CATEGORY GROUP=3 (指定变量GROUP为分组变量，并限定其组数为三组)

   CATEGORY GROUP=3 SEX=2 (指定变量GROUP 和SEX 为分组变量，并限定其组数分别为三组和两组)

2. MODEL：规定一个线性模型，命令格式：

   MODEL <变量> [, <变量>, < ... >] = [CONSTANT+]<变量> [+ <变量> + < ... >][+ <变量> ∗ <变量> + < ... >]，如：

   MODEL Y=CONSTANT+X (指定一个因变量Y，自变量X 的简单线性模型)

   MODEL Y=CONSTANT+X1+X2+X3 (指定一个因变量Y，自变量X1,X2,X3 的线性模型) MODEL Y1,Y2,Y3=CONSTANT+X1+X2+X3 (规定多个变量Y1,Y2,Y3 为因变量，变量X1，X2，X3 为自变量的简单线性模型)

3. ESTIMATE：对MODEL 命令所规定的线性模型作最小二乘估计。

4. STEP：对MODEL命令所规定的线性模型作逐步回归，命令格式：

   STEP [/ENTER=<数值>，REMOVE=<数值>]

   STEP (按标准阈值ENTER=.15,REMOVE=.15 作逐步回归)

   STEP /ENTER=.3，REMOVE=.3

   注意：引进变量(ENTER) 的阈值必须小于等于剔出变量(REMOVE) 的阈值。

5. HYPOTHESIS：进入假设检验程序，标志假设检验的开始，命令格式：HYPOTHESIS

6. EFFECT：指定线性模型中一个自变量或一个自变量组合，以便对其系数或系数组合进行假设检验，命令格式：

   EFFECT=X (规定对变量X的系数进行假设检验)

   EFFECT=X*GROUP (规定对变量X,GROUP的系数组合进行假设检验)

7. CONTRAST 规定一个多重比较假设，以便对EFFECT 命令所指定的系数或系数组合进行多重假设检验，命令格式：

   CONTRAST <数值>, <数值>, <数值> [, <数值>, <数值>, < ... >]. 如

   CONTRAST

   1,-1,0,0

   注：输入的数值之和必须为零。

表 7.2 工作人员能力和生产效率打分

能力(X)	生产效率(Y)	能力(X)	生产效率(Y)
41	32	38	29
35	20	38	33
34	35	46	36
40	24	36	23
33	27	32	22
42	28	43	38
37	31	42	26
42	33	30	20
30	26	41	30
43	41	45	30

8. TEST 按所规定假设进行假设检验，命令格式：TEST

以上命令，其中ESTIMATE,STEP,TEST命令是热命令，其余均为冷命令。

关于MGLH 模块的应用，我们将介绍如何使用MGLH 模块中的统计命令去进行回归分析。1. 简单回归

例：工作人员能力测定与其生产效率(表 7.2)

假设我们已经建立了包含这些数据的SYSTAT 文件，其文件名为LEVE.SYS 文件中的两个数值变量分别为X 和Y。则做简单回归的步骤为：首先进入MGLH 模块，打入命令：

```
USE LEVE
MODEL Y=CONSTANT+X
ESTIMATE
```

显示器显示下面的结果

```
DEP VAR:     Y   N:  20   MULTIPLE R: .609   SQUARED MULTIPLE R: .371
ADJUSTED SQUARED MULTIPLE R: .336  STANDARD ERROR OF ESTIMATE:  4.769

VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T   P(2 TAIL)
CONSTANT      1.016       8.710       0.000 1.0000000   0.117   0.908
       X      0.734       0.225       0.609 1.0000000   3.260   0.004

                        ANALYSIS OF VARIANCE
  SOURCE   SUM-OF-SQUARES     DF  MEAN-SQUARE     F-RATIO       P
REGRESSION        241.766      1       241.766      10.629       0.004
  RESIDUAL        409.434     18        22.746
-------------------------------------------------------------------------
```

2.多元回归

表 7.3 13个疾病观察点的发病水平及病因学因素

观察点编号	疾病学因素				发病水平
NO	X1	X2	X3	X4	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

例：设调查了13 个疾病观察点的某病发病水平(Y) 及一组病因学观察指标(X1,X2,X3,X4)，资料见表 7.3。

并假设数据已存入MULREG.SYS 文件中。则多元回归的操作步骤为：进入MGLH 模块，然后键入如下命令

```
USE MULREG
MODE Y=CONSTANT+X1+X2+X3+X4
ESTIMATE
```

运算结果如下：

```
DEP VAR:  Y    N: 13   MULTIPLE R: .991   SQUARED MULTIPLE R:   .982
ADJUSTED SQUARED MULTIPLE R: .974  STANDARD ERROR OF ESTIMATE: 2.446

VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T   P(2 TAIL)
CONSTANT    62.405       70.071      0.000  1.0000000  0.891    0.399
      X1     1.551        0.745      0.607   .0259766  2.083    0.071
      X2     0.510        0.724      0.528   .0039305  0.705    0.501
      X3     0.102        0.755      0.043   .0213363  0.135    0.896
      X4    -0.144        0.709     -0.160   .0035397 -0.203    0.844
 REGRESSION       2667.899     4      666.975     111.479       0.000
   RESIDUAL         47.864     8        5.983
-------------------------------------------------------------------------
```

3.逐步回归

我们仍以上面的疾病病因学研究为例。则逐步回归的运算步骤为：进入MGLH 模块，然后使用命令：

```
USE MULREG
MODEL Y=CONSTANT+X1+X2+X3+X4
STEP/ENTER=.2,REMOVE=.2
```

输出结果:

```
STEPWISE REGRESSION WITH ALPHA-TO-ENTER= .200 AND ALPHA-TO-REMOVE= .200

  STEP=  1     ENTER       X4     R=  .821     RSQUARE=  .675
  STEP=  2     ENTER       X1     R=  .986     RSQUARE=  .972
  STEP=  3     ENTER       X2     R=  .991     RSQUARE=  .982
  STEP=  4    REMOVE       X4     R=  .989     RSQUARE=  .979

  THE SUBSET MODEL INCLUDES THE FOLLOWING PREDICTORS:
          CONSTANT
              X1
              X2
USE  THESE  PREDICTORS  IN  A  NEW  MODEL  SENTENCE  TO  ESTIMATE  THE
COEFFICIENTS.
```

STEP 命令仅仅输出了筛选步骤，并给出了每一步的复相关系数和决定系数。最后得到的是在给定的显著性水平下保留的预测因子(即自变量)，程序称之为子集(SUBSET)。本例只有X1和X2在0.2水平下有显著意义。要求这个回归方程只要用MODEL 语句加上这些预测因子，然后键入ESTIMATE 命令即可。

```
USE MULREG
MODEL Y=CONSTANT+X1+X2
ESTIMATE
```

输出结果:

```
  DEP VAR:   Y   N:  13   MULTIPLE R: .989  SQUARED MULTIPLE R:  .979
  ADJUSTED SQUARED MULTIPLE R: .974  STANDARD ERROR OF ESTIMATE:2.406

  VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T  P(2 TAIL)
  CONSTANT    52.577       2.286      0.000   1.0000000  22.998  0.000
        X1     1.468       0.121      0.574    .9477514  12.105  0.000
        X2     0.662       0.046      0.685    .9477514  14.442  0.000

                          ANALYSIS OF VARIANCE

     SOURCE   SUM-OF-SQUARES    DF  MEAN-SQUARE     F-RATIO       P
```

表 7.4 甘蓝叶中核黄素之浓度

高锰酸钾			高锰酸钾		
样本重量	处理	未处理	样本重量	处理	未处理
	27.2	39.5		24.6	38.6
0.25g	23.2	43.1	1.00g	24.2	39.5
	24.8	45.2		22.2	33.0

表 7.5 甘蓝叶中核黄素之浓度($\mu g/g$)

测定次数	经高锰酸钾处理		未经高锰酸钾处理	
	0.25g 样本	1.00g 样本	0.25g 样本	1.00g 样本
1	27.2	24.6	39.5	38.6
2	23.2	24.2	43.1	39.5
3	24.8	22.2	45.2	33.0

```
REGRESSION       2657.859     2     1328.929      229.504       0.000
  RESIDUAL         57.904    10        5.790
-------------------------------------------------------------------------
```

从输出的参数来看，取 x 1和 x 2两个因子来预测的效果较好。其回归方程为: Y=52.577+1.468X1+0.662X2

4. 析因设计的方差分析(1)2 x 2 析因设计的方差分析

2 x 2 设计是指有两个因素，每个因素有两个水平的实验设计，共有2 x 2=4 个组，各因素各水平均相遇一次。这是析因设最简单的一种行式。如以a1 表示a 因素1 水平，a2 表示a 因素2 水平，b1 表示b 因素1 水平，b2 表示b 因素2 水平，则2 x 2 设计的模型为：

a_1b_1	a_1b_2
a_2b_1	a_2b_2

表 7.4是甘蓝叶中核黄素含量($\mu g/g$)的荧光测定结果。所用的甘蓝叶有经过二氧化氢高锰酸钾处理的，也有未经处理的，甘蓝叶的样本有0. 25g 与1g两种。试问高锰酸钾处理与否样本量不同对甘蓝叶中核黄素含量的测定结果有无显著差别。

本题甘蓝叶的处理方法是一个因素，分处理与否两个水平；样本重量为另一个因素，分为0.25g 与1g 两个水平；每种组合都经三次测定。为了便于对应于2×2 析因设计模型建立数据文件，现将表 7.4整理成以下形式：

和两因素方差分析一样，建立数据文件对每个因素各设一个变量，水平取值用1,2,3,...表示。现设处理因素为TREAT，取值1 表示用高锰酸钾处理，2 表示未处理，重量因素为WEIGHT，取值 1 表示0.25g，2 表示1.00g。因考虑到测定次数可能引起的误差，我们把测定次数看作区组，用变量BLOCK 表示，取值1，2，3，count 为测量值，RESULT 作为测量变量。

设数据存于文件EXAM111，分析步骤如下：

```
C:\SYSTAT>MGLH
```

```
>USE EXAM111
>CATEGORY TREAT=2,WEIGHT=2,BLOCK=3
>MODE RESULT=CONSTANT+TREAT+WEIGHT+BLOCK+TREAT*WEIGHT
>ESTIMATE
```

这里MODEL 语句从形式上看是三因素方差分析，但我们主要研究的因素是TREAT 和WEIGHT，BLOCK只是为控制不同时间测量的影响而设计的变量。TREAT *WEIGHT 就是指定求两因素的交互作用。

输出结果：

```
DEP VAR:RESULT  N:  12  MULTIPLE R:  .970  SQUARED MULTIPLE R: .940

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
 TREAT         716.108     1      716.108    87.547     0.000
WEIGHT          36.401     1       36.401     4.450     0.079
 BLOCK           3.762     2        1.881     0.230     0.801
TREAT*
WEIGHT          13.021     1       13.021     1.592     0.254
ERROR           49.078     6        8.180
```

结果表明，只有TREAT 有极显著性差异。即样品用高锰酸钾处理与否对测定结果有极显著的影响。

对析因设计资料作方差分析时，一般来说，如果计算出来的交互作用没有显著意义的话，可以把这部分的差异与误差项合并，然后重新计算 F 值。用程序计算时，只要从MODEL语句中去掉交互项即可。下面是本题去掉交互项后的方差分析结果。

```
DEP VAR:RESULT  N:  12  MULTIPLE R:  .961  SQUARED MULTIPLE R: .924

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
 TREAT         716.108     1      716.108    80.722     0.000
WEIGHT          36.401     1       36.401     4.103     0.082
 BLOCK           3.762     2        1.881     0.212     0.814
ERROR           62.099     7        8.871
```

去掉交互项后，结论仍和前面一样。但是由于误差项的离均差平方和增大，各因素的F 值变小了。

⑵$3\times2\times2$ 析因设计的方差分析

$3\times2\times2$ 设计是指有三个因素，其中一个因素有3个水平，另两个因素各有2 个水平的实验设计。这些因素的水平在一个设计中相互组合一次。

表 7.6是一个钩端螺旋体的资料，血清种类有兔血清与胎盘血清两种，每种血清有5%与8% 两种浓度，所有基础液有三种，即缓冲剂、蒸馏水与自来水。试分析钩端螺旋体计数学清种类、血清浓度及基础液种类的关系。为了便于说明，令A表示基础液，B 表示血清种类，C 表示浓度，按各因素的水平数，构成$3\times2\times2$ 实验。试作三因素析因设计方差分析。

表 7.6 $3 \times 2 \times 2$ **析因实验结果(钩端螺旋体计数)**

基础液 (A)	血清种类(A)			
	兔血清		胎盘血清	
	血清浓度(C)		血清浓度(C)	
	5%	8%	5%	8%
缓冲剂	648	1144	830	578
	1246	1877	853	669
	1398	1671	441	643
	909	1845	1030	1002
蒸馏水	1763	1447	920	933
	1241	1883	709	1024
	1381	1896	848	1092
	2421	1962	574	742
自来水	508	1789	1126	685
	1026	1215	1176	546
	1026	1434	1280	595
	830	1651	1212	566

建立数据文件，各因素的变量名称和各水平的取值意义如下：

变量名	取值意义
基础液A	1－缓冲剂，2－蒸馏水，3－自来水
血清种类B	1－兔血清，2－胎盘血清
血清浓度C	1%～5%，2%～8%
测量值COUNT	

设数据文件存于EXAM112，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXAM112
 >CATEGORY A=3,B=2,C=2
 >MODE COUNT=CONSTANT+A+B+C+A*B+A*C+B*C+A*B*C
 >ESTIMATE
```

本题有三个因素，可以构成高阶交互项。一般这种检验都从高阶交互项开始，如果高阶交互项不显著，则把它从模型中去掉，然后再检验低阶交互项如果都不显著，则构造一个只有主效应的模型。上面MODEL 语句包含了所有交互项，故称“饱和”模型，输出结果：

```
DEP VAR:COUNT   N:  48  MULTIPLE R:  .872  SQUARED MULTIPLE R: .761

             ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
    A      692513.375     2   346256.688     4.978      0.012
    B     4142462.521     1  4142462.521    59.551      0.000
```

```
      C      248976.021     1   248976.021     3.579      0.067
     A*
      B      726923.042     2   363461.521     5.225      0.010
     A*
      C       99091.542     2    49545.771     0.712      0.497
     B*
      C     1111729.688     1  1111729.688    15.982      0.000
     A*
     B*
      C      946085.375     2   473042.688     6.800      0.003
  ERROR     2504233.750    36    69562.049
```

检验的结果表明，除主效应C和A*C交互项外，其他各项均有显著意义。现将A*C 从MODE语句中去掉，重新检验，结果如下:

```
DEP VAR:COUNT   N:  48  MULTIPLE R:  .867  SQUARED MULTIPLE R: .751

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P

      A      692513.375     2   346256.688     5.054      0.011
      B     4142462.521     1  4142462.521    60.466      0.000
      C      248976.021     1   248976.021     3.634      0.064
     A*
      B      726923.042     2   363461.521     5.305      0.009
     B*
      C     1111729.688     1  1111729.688    16.228      0.000
     A*
     B*
      C      946085.375     2   473042.688     6.905      0.003
  ERROR     2603325.292    38    68508.560
```

去掉A*C交互项后结论不变。

由于本题三个因素ABC 的交互作用显著，主效应诸均数比较时，就有可能被交互作用所掩盖。在这种情况下，可将一个因素固定在一定水平上，用Duncan 多重极差检验来比较在该水平下处于另一因素的诸均数。

首先进入STATS 模块，用分组统计命令打印小组(共12 组) 的均数：

```
C:\SYSTAT>STATS
 >USE EXAM112
 >BY A,B,C
 >STATISTICS COUNT/MEAN
```

为了便于分析，现将各组均数从以上输出中整理于下表。

(注:括号中的数字对应于各因素变量的取值)

表 7.7 各小组均数

血清种类	浓度	基础液(A)		
(B)	(C)	缓冲剂(1)	蒸馏水(2)	自来水(3)
兔血清	(1) 5%(1)	1050.25	1701.50	847.50
	8%(2)	1634.25	1788.00	1522.25
胎盘血清	(2) 5%(1)	788.50	762.75	1198.50
	8%(2)	723.00	947.75	598.00

计算Duncan 多重极差检验的显著性界值(仍在STATS 模块下)。这一步运算需要用到前方差分析的结果，并通过键盘输入。命令格式如下：

DUNCAN/K=组数，MSE=误差的均方

ALPHA=显著性水平，DFE=误差的自由度

N=每组观察例数

本题的计算操作为:

>DUNCAN/K=3，MSE=68508.56

ALPHA=0.05; DFE=38; N=4

因本题最多只有3个组对比，所以K=3。结果输出:

```
DUNCAN  MULTIPLE  RANGE  TESTS
ORDERED MEANS DIFFER AT ALPHA=  .050 F THEY EXCEED FOLLOWING GAPS
GAP ORDER DIFFERENCE
      1      374.850
      2      393.968
THIS TEST ASSUMES THE COUNTS PER GROUP ARE EQUAL
```

结果中的差值(DIFFERENCE)即为显著性检验的极差值。这里我们给定ALPHA=0 .05，这就意为着，如果排序后的两均数之差大于相应间隔(GAP)的差值，则在此0 .05水平上拒绝无效假设。

下面计算每两均数的相差与间隔的差值比较确定显著性。加“*”号的表示P¡0. 05。

①在B x C 同一水平上比较A (三种基础液)

B x C 兔学清浓度5对比组	差值
蒸馏水与缓冲剂	651.25
蒸馏水与自来水	854.5*
缓冲剂与自来水	202.75
B x C 兔血清浓度8%	
对比组	差值
蒸馏水与缓冲剂	153.75
缓冲剂与自来水	112.00
蒸馏水与自来水	265.75
B x C 胎盘血清浓度5 %	
对比组	差值
自来水与缓冲剂	410.00
蒸馏水与蒸馏水	25.75
自来水与蒸馏水	435.75
B x C 胎盘血清浓度8%	
对比组	差值
蒸馏水与缓冲剂	224.75
缓冲剂与自来水	125.00
蒸馏水与自来水	349.75

②在A x C 同一水平上比较B (两种血清)

基础液	浓度	兔血清	胎盘血清	差值
缓冲剂	5 %	1050.25	788.50	261.75
缓冲剂	8 %	1634.25	723.00	911.25*
蒸馏水	5 %	1701.50	762.75	938.75*
蒸馏水	8 %	1788.00	947.75	840.25*
自来水	5 %	847.50	1198.50	-351.00
自来水	5 %	1522.25	598.00	924.25*

③在A x B 同一水平上比较C（两种血清浓度）

基础液	血清种类	浓度8%	浓度5%	差值
缓冲剂	兔血清	1634.25	1050.25	584.00*
蒸馏水	兔血清	1788.00	1701.50	85.50
自来水	兔血清	1522.25	847.50	674.75*
缓冲剂	胎盘血清	723.00	788.50	-65.50
蒸馏水	胎盘血清	947.75	762.75	185.00
自来水	胎盘血清	598.00	1198.50	-600.50*

从方差分析来看，主效应C (浓度间) 差别不显著，当在A x B 同一水平比较时，则有三对均数的差别有显著意义，说明前者由于交互作用显著掩盖了浓度均数间的显著性。通过上述比较不难得出结论：用兔血清浓度为8馏水为基础液时，钩端螺旋体计数较高。

⑶正交试验设计的方差分析

当研究的因素超过三个时，并且因素间又有可能存在交互作用，可用正交试验设计。正交试验是将各试验因素，各水平进行合理组合，均匀搭配，由此大大减少试验次数而又能得到较多的信息。正交试验设计的分析可采用方差分析，它把总变异的离均差平方和及其自

表 7.8 过氧乙酸稳定性试验的因素分析及水平

试验因素	水平	
	1	2
A：稳定剂	加磷酸0.3%	不加磷酸
B: 水溶温度	25~30℃	35 40℃
C: 浸泡口表	浸泡口表10支	不浸泡口表
D: 加盖与否	加盖	不加盖

表 7.9 过氧乙酸定性试验安排及其结果

试验号	不同因素的水平号		24 小时过氧乙酸		残存量(mg/3ml)	
	A	B	C	D	1	2
1	1	1	1	1	7.00	4.11
2	1	1	2	2	6.05	3.50
3	1	2	1	2	1.10	0.80
4	1	2	2	1	1.90	0.96
5	2	1	1	2	2.40	1.65
6	2	1	2	1	4.00	1.50
7	2	2	1	1	0.35	0.30
8	2	2	2	2	0.30	0.90

由度分为各因素的各水平间，因素间的交互作用及误差几部分，因此能经确地说明各因素诸水平间的差别，确切地判断各因素间的交互作用。

过氧乙酸是广泛应用的一种杀灭病毒性肝炎病毒的主要消毒剂，但其有效成份极不稳定，以致影响其消毒效果。现欲通过实验找出有关因素对其稳定性的影响，并指出哪个是主要的，哪个是次要的，哪个起交互作用，尔后选出各因素中的一个最佳水平，组成保持过氧乙酸稳定性的最优条件。

本例试验因素及其水平数如表 7.8：

本题有4 个因素，每个因素各有2 个水平，此外还要观察稳定剂与温度(A *B)、稳定剂与加盖与否(A*D)的交互作用。试验选用L8(27)正交表，试验安排及测定结果见下表。

用SYSTAT 程序作用多因素方差分析，交互作用只要在模型中指定就能自动计算，所以正交表中的交互列在建立数据文件时不必考虑，误差列也是如此，上面的正交表每列设一个变量。每个变量都按水平号取值，另设一个测量值变量(本题设为VAL)。

设数据存于文件EXAM13，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXAM113
 >CATEGORY A=2,B=2,C=2,D=2
 >MODE VAL=CONSTANT+A+B+C+D+A*B+A*D
 >ESTIMATE
```

结果输出:

```
DEP VAR      X    N:  16  MULTIPLE R:  .896  SQUARED MULTIPLE R: .803

              ANALYSIS   OF   VARLANCE
SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
    A          12.205     1       12.285     9.708     0.011
    B          34.810     1       34.810    27.507     0.000
    C           0.123     1        0.123     0.097     0.762
   A*
    B           4.202     1        4.202     3.321     0.098
   A*
    D           0.164     1        0.164     0.130     0.726
ERROR          12.655    10        1.266
```

方差分析结果表明，稳定剂(A)与水浴温度(B)是影响过氧乙酸稳定性两个最主要因素，A与B及A与D的交互作用不显著。为了便于选择最优水平，我们可以在SYSTAT 模型下用分组统计命令计算出 4 个因素各水平的合计数，结果如下：

	A	B	C	D
水平1	25.42	30.21	17.71	20.12
水平2	11.40	6.61	19.11	16.70

数值大者说明过氧乙酸残存量多，更有利于过氧乙酸的稳定。通过比较不难看出，当A、B因素取 1 水平时效果较好。C、D 两因素的作用不显著，可以任选一水平，不过D 因素的1 水平从数值上还是明显大于2 水平，故选1 水平为宜。最后结论：加0.3件。

5. 协方差分析

是把直线回归与方差分析结合起来的一种统计方法。它利用回归的关系把与因变量Y值呈直线关系的自变量X值化成相等后，再进行方差分析。它比较的是调正均数间的差异。通过协方差分析，能够校正由于各组X值的不同所引起的偏差，更恰当地评价各种处理的优劣。

⑴完全随机设计的协方差分析

男性运动员和大学生的平均肺活量分别为4399cm，3667cm，经假设检验有差别。但我们已知肺活量与身高有一定关系(一般来说，肺活量随身高增大而增大)。本例中运动员的身高高于大学生，因此在比较肺活量时必须对身高作校正，这就需用协方差分析进行处理。

协方差分析的前提条件是，各组资料（样本）都来自方差相同的正态分布总体；各组的回归系数本身有显著性意义，但各个回归系数间无显著性差别，即斜率是齐性的。关于资料的方差齐性检验可用STATS 模块的STATISTICS 命令去完成。协方差分析的数据格式与方差分析基本相似，也要设置一个分组变量来标识对比的各组，所不同的是增加了一个自变量。本假设分组变量为GROUP；运动员取值为1，大学生取值为2。在作协方差分析之前，我们先检验一下两组的斜率是否相同。所谓斜率齐性检验就是对协变量X和分组变量GROUP的交互项作假设检验。

设数据存于文件EXAM121，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXMA121
```

表 7.10 20 运动员及大学生的身高(X,cm)与肺活量(Y,cm^3)

运动员		大学生		运动员		大学生	
X1	Y1	X2	Y2	X1	Y1	X2	Y2
184.9	4300	168.7	3450	169.0	4500	173.8	4150
167.9	3850	170.8	4100	188.0	4780	174.0	3450
171.0	4100	165.0	3800	176.7	3700	170.5	3250
171.0	4300	169.7	3300	179.0	5250	176.0	4100
188.0	4800	171.5	3450	183.0	4250	169.5	3650
179.0	4000	166.5	3250	180.5	4800	176.3	3950
177.0	5400	165.0	3600	179.0	5000	163.0	3500
179.5	4000	165.0	3200	178.0	3700	172.5	3900
187.0	4800	173.0	3950	164.0	3600	177.0	3450
187.0	4800	169.0	4000	174.0	4050	173.0	3850

```
>CATEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X+GROUP*X
>ESTIMATE
```

因为交互项GROUP*X 的作用不显著(P=0.859)，故认为两组的回归斜率无显著差异。从上面的结果可以看出，自变量X的作用有显著意义，这说明总体直线回归系数不为0。如果X无显著性，则作协方差分析就无意义了。在这种情况下可以不考虑X 的影响，直接作方差分析即可。

现在我们拟合一个上面资料的协方差分析模型：

```
>CAEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X
>ESTIMATE
```

如果熟悉前面的方差分析和线性回归模型就会发现，协方差模型正是两者的组合。

```
                ANALYSIS  OF  VARIANCE

SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP     1407847.095   1  1407847.095    9.220   0.004
     X     1630762.635   1  1630762.635   10.679   0.002
 ERROR     5649992.365  37   152702.496
```

结果表明，均衡了协变量的影响后，两组间的肺活量仍有极显著差别。

协方差分析是对修正均数的差别进行显著性检验，程序并没有输出修正均数。若要显示修正均数，必须在ESTIMATE命令之前打入一条存SYSTAT文件的命令，并加上选择项ADJUSTED,如SAVE MYFILE/ADJUSTED。这样程序就会在MYFILE 中产生一个名为ESTIMATE的变量用于存放修正均数。由于这个文件没有分组变量，所以用户要记住每一组在文件中的

表 7.11 三组大鼠的进食量(X，g) 与所增体重(Y，g)

窝别	(1)核黄素缺乏组		(2)限食量组		(3)不限食量组	
	X	Y	X	Y	X	Y
1	256.9	27.0	260.3	32.0	544.7	160.3
2	271.6	41.7	271.1	47.7	481.2	91.6
3	210.2	25.0	214.7	36.7	418.9	114.6
4	300.1	52.0	300.1	65.0	556.6	134.8
5	262.2	14.5	269.7	39.0	394.5	76.3
6	304.4	48.8	307.5	37.9	426.6	72.8
7	272.4	48.0	278.9	51.5	416.1	99.4
8	248.2	9.5	256.2	26.7	549.9	133.7
9	242.8	37.0	240.8	41.0	580.5	147.0
10	342.9	56.5	340.7	61.3	608.3	165.8
11	365.9	76.0	365.3	102.1	559.6	169.8
12	198.2	9.2	199.2	8.1	371.9	54.3

起始记录号。当然，也可以用下面的方法把原文件的分组变量与ESTIMATE变量连结后形成一个新文件,便于结果的阅读。

```
C:\SYSTAT>DATA
 >USE EXAM121(GROUP)MYFILE(ESTIMATE)
 >SAVE NEWFILE
 >RUN
```

新产生的NEWFILE 文件包含GROUP 和ESTIMATE 两个变量，用LIST 命令即可看到GROUP 的不同取值对应不同的修正均数为3805.836。由此作出结论，两组肺活量均数在消除身高因素的影响后仍有极显著差别，运动员的肺活量大于大学生。

⑵随机区组设计的协方差分析

如果实验中包含两因素，其中一个因素的记录具有依存关系(直线关系) 的成对(X，Y) 数值，也可用协方差分析。

在“核黄素缺乏对于蛋白质利用的影响之研究”中，将体重相近(30 38g),出生三周的大鼠36只,按照窝别、性别等条件分成12窝,每窝3只,随机分到三个不同饲料组进行喂养。观察记录列于表 7.11，观察黄素缺乏对体重增长的影响。

本题作协方差分析要设置两个分组变量。处理组变量设为GROUP，取值1 表示核黄素缺乏组，2 表示限食量组，3 表示不限食量组;窝别变量设为BLOCK，取值1 12，表示12 区组;进食量X 为协变量; 所增体重Y 为因变量。

设数据存于EXAM122，分析步骤如下：

先作斜率的齐性检验。

```
C:\SYSTAT>MGLH
 >USE EXAM122
 >CATEGORY GROUP=3,BLOCK=12
```

```
 >MODE Y=CONSTANT+GROUP+BLOCK+X+GROUP*X
 >ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  36  MULIPLE R: .986  SQUARED MULIPLE R:  .971

               ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP         105.208   2       52.604    0.461   0.637
 BLOCK        3765.619  11      342.329    3.001   0.017
     X        2827.148   1     2827.148   24.788   0.000
GROUP*
     X          66.105   2       33.052    0.290   0.752
 ERROR        2167.034  19      114.054
```

交互项的F值的概率=0.752,说明各组回归的斜率无显著性差异。下面作协方差分析:

```
>MODEY=CONSTANT+GROUP+BLOCK+X
>SAVE MODY/ADJUSTED
>ESTIMATE
```

在配合模型时用SAVE命令指定存贮修正均数,文件名为MODY。结果输出;

```
DEP VAR:    Y    N:  36  MULIPLE R: .985  SQUARED MULIPLE R:  .971

               ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP         469.157   2      234.578    2.206   0.135
 BLOCK        3761.319  11      341.938    3.216   0.010
     X        6175.031   1     6175.031   58.069   0.000
 ERROR        2233.139  21      106.340
```

协方差分析结果表明，当消耗食物量相同时，各不同饲料组大鼠平均增重没有明显不同。本题如果不考虑进食量，仅作随机区组的方差分析，则处理间比较的概率小于0.01,结论恰恰相反。

由于协方差模型包含区组变量，所以MODY文件中同一处理组的各例的修正均数不尽相同。要求各处理组的修正均数，可在STATS模块下按GROUP分组计算ESTIMATE 均数,操作方法如下:

```
C:\SYSTAT>DATA
 >USE EXAM122(GROUP)MODY(ESTIMATE)
 >SAVE NEWFILE
 >RUN
C:\SYSTAT>STATS
 >USE NEWFILE
```

表 7.12 六组公鼠的食物消耗量(X,10cal)及所增体重(Y,g)

高蛋白						低蛋白					
牛	肉	谷	类	猪	肉	牛	肉	谷	类	猪	肉
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
108	73	99	98	194	94	165	90	124	107	140	49
136	102	117	74	198	79	164	76	95	95	177	82
138	118	90	56	196	96	161	90	116	97	189	73
159	104	141	111	198	98	159	64	112	80	142	86
146	81	106	95	210	102	175	86	123	98	216	81
141	107	112	88	196	102	135	51	110	74	200	97
175	100	110	82	230	108	132	72	137	74	255	106
149	87	117	77	222	91	190	90	105	67	173	70
174	117	111	86	220	120	145	95	135	89	153	61
176	111	122	92	228	105	142	78	126	58	160	82

```
>BY GROUP
>STATISTICS ESTIMATE/MEAN
```

⑶析因设计的协方差分析

在析因设计时，如果对比的各水平要考虑协变量的影响就应用析因设计的协方差分析。

将60只公鼠随机分成六组,分别饲以不同来源及成分的蛋白质，并记录食物消耗(X,10cal),所增体重(Y,g)于表 7.12，试作分析。

本题主要分析的因素有两个，蛋白质含量的高低和蛋白质的食物来源。前者用变量 A 来表示，取值1 表示高蛋白组，2 表示低蛋白组；后者用变量 B 表示，取值1 为牛肉，2 为谷类，3 为猪肉。分析的目的是要了解公鼠所增体重是否与蛋白质含量高低有关，是否与蛋白质的食物来源有关以及蛋白质含量高低与食物来源间对体重增加有无交互作用。分析时公鼠的食物消耗作为协变量考虑。

设数据存于文件EXAM123，分析步骤如下：

统计分析就是把析因设计的方差分析与协变量配合在一个模型里。

```
C:\SYSTAT>MGLH
 >USE EXAM123
 >CATEGORY A=2,B=3
 >MODE Y=CONSTANT+A+B+X+A*B
 >SAVE MODY/ADJUSTED
 >ESTIMATE
```

结果输出:

```
DEP VAR:     Y     N:  60  MULIPLE R: .685  SQUARED MULIPLE R:  .469

                    ANALYSIS  OF  VARIANCE
```

表 7.13 各小组修正均数

高蛋白组			低蛋白组		
牛肉	谷类	猪肉	牛肉	谷类	猪肉
101.55	100.77	80.21	78.42	96.72	69.55

```
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
     A        2343.463   1     2343.463   14.450   0.000
     B        1673.305   2      836.653    5.159   0.009
     X        2990.626   1     2990.626   18.441   0.000
    A*
     B         933.812   2      466.906    2.879   0.065
 ERROR        2233.139  21      106.340
```

从结果中看出，当均衡了进食量影响后，蛋白质含量高低间有极显著差别，蛋白质食物来源间也有极显著差别，而两者的交互作用无意义。由于交互作用的不显著，分析主效应就可以单独比较修正均数。按前述方法，在STATS 模块下统计出各小组的修正均数(分组命令用BY A,B)，整理于表 7.13。

总的来看，高蛋白组的体重增加大于低蛋白组。这样只要在高蛋白组中比较食物来源就能得出结论。小组间均数的两两比较可采用DUNCAN多重极差检验。

在STATS模块健入如下命令,求出显著界值:

```
C:\SYSTAT>STATS
>DUNCAN/K=3,MSE=162.177,ALPHA=0.05,DFE=53,N=10
```

结果输出：

```
DUNCAN MULTIPLE RANGE TESTS
ORDERED MEANS DIFFER AT ALPHA=0.050 IF THEY EXCEED FOLLOWING GAPS
GAP  ORDER       DIFFERENCE
       1              11.427
       2              12.018
THIS TEST ASSUMES THE COUNTS PER GROUP ARE EQUAL
```

经比较，在高蛋白组中，牛肉组与谷类组的差别无显著意义，而二者分别与猪肉组比较差别有显著意义。有此我们得到的结论是，当公鼠的摄入量（按热量计）相同时，进食高蛋白牛肉或谷类食物的体重增加较快。

⑷多元协方差分析在比较两组或多组因变量时，如果因变量与多个自变量间存在着一定的线性关系，就应考虑这些自变量的影响。多元协方差分析就是将各个自变量调整到相同的水平，再对因变量的均数作比较。

某地30名初生至三周岁儿童的身高、体重和体表面积，记录于下表。考虑男、女两组的体表面积与身高、体重的关系是否相同，能否合并为一个推算方程。

建立数据文件EXAM124，设性别变量为GROUP，取值1 为男性，2为女性。

首先进行斜率的齐性检验。因为有两个协变量，所以模型中要包含两个交互项。

表 7.14 30 名婴儿身高(X1,cm)体重(X2,kg)及体表面积(Y,cm2)

男			女		
X1	X2	Y	X1	X2	Y
54	3	2446.2	54	3	2117.3
50.5	2.25	1928.4	53	2.25	2200.2
51	2.5	2094.5	51.5	2.5	1906.2
56.5	3.5	2506.7	51	3	1850.3
52	3	2121.0	51	3	1632.5
76	9.5	3845.9	77	7.5	3934.0
80	9	4380.8	77	10	4180.4
74	9.5	4314.2	77	9.5	4246.1
80	9	4078.4	74	9	3358.8
76	8	4134.5	73	7.5	3809.7
96	13.5	5830.2	91	12	5358.4
97	14	6013.6	91	13	5601.7
99	16	6410.6	94	15	6074.9
92	11	5283.3	92	12	5299.4
94	15	6101.6	91	12.5	5291.5

```
C:\SYSTAT>MGLH
>USE EXAM124
>CATEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X1+X2+GROUP*X1+GROUP*X2
>ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  30  MULIPLE R: .993  SQUARED MULIPLE R:  .986

                ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP       88053.569   1    88053.569    2.161   0.155
    X1      976354.460   1   976354.460   23.965   0.000
    X2      331960.960   1   331960.960    8.148   0.009
GROUP*
    X1       62455.886   1    62455.886    1.533   0.228
GROUP*
    X2       46356.884   1    46356.884    1.138   0.297
 ERROR      977776.305  24    40740.679
```

结果表明两组回归的斜率无显著差别。接着作二元协方差分析。

```
>MODE Y=CONSTANT+GROUP+X1+X2
```

```
>ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  30  MULIPLE R: .992  SQUARED MULIPLE R:  .985
                ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP      139769.340   1   139769.340    3.411   0.076
    X1      938153.704   1   938153.704   22.895   0.000
    X2      368954.790   1   368954.790    9.004   0.006
 ERROR     1065399.759  26    40976.914
```

均衡了身高、体重后，男、女间的体表面积仍无显著差别，故认为两者可合并建立推算方程。

㈥FACTOR 模块

FACTOR 模块用于进行主成分分析和因子分析，它的基本命令格式为:

FACTOR <变量1>,<变量2>……

这个命令可输出相关矩阵、特征根、特征根的贡献率及因子负荷。

某单位研究儿童生长发育情况，测量了30 名三岁男童的六项基本体格指标：体重(X1)，身高(X2)，胸围(X3)，上臂围(X4)，三头肌(X5)，肩胛下角(X6)。设数据文件为EXAM161，试作主成分分析。

```
C:\SYSTAT>FACTOR
 >USE EXAM161
 >NUMBER=2
 >FACTOR X1,X2,X3,X4,X5,X6
```

上述一组操作命令中出现了二个FACTOR,第一个FACTOR是从SYSTAT进入FACTOR 模块；第二个FACTOR 后跟六个变量名，表示对该六个原指标作主成分分析。NUMBER=2 表示只取前二个主成分，即P=2。其输出结果如下：

```
MATRIX TO BE FACTORED
      X1       X2       X3       X4       X5       X6
 X1   1.000
 X2   0.609   1.000
 X3   0.716   0.487    1.000
 X4   0.771   0.401    0.451   1.000
 X5   0.312  -0.222    0.256   0.350    1.000
 X6   0.391   0.032    0.273   0.555    0.432    1.000

 TATENT ROOTS(EIGENVALUES)
    1          2          3          4          5          6
 3.086      1.432      0.654      0.427      0.276      0.126
 COMPONENT LOADINGS
                          1              2
```

表 7.15 30 名三岁男童六项体格指标测量结果

X1	X2	X3	X4	X5	X6
13.500	95.000	52.500	15.500	10.000	6.000
14.500	102.000	49.000	16.000	8.000	7.000
13.000	97.600	49.000	15.000	8.000	6.000
15.400	100.000	53.500	15.500	8.000	5.000
16.500	100.000	54.000	17.000	9.000	8.000
13.100	93.500	51.000	15.000	9.000	8.000
14.700	97.500	50.000	15.500	9.000	7.000
14.300	95.100	51.400	15.700	9.000	6.000
13.850	95.600	52.000	14.500	10.000	6.000
11.250	99.000	51.000	13.700	7.000	5.000
15.000	100.000	52.000	15.500	10.000	6.000
15.300	100.000	53.000	16.000	9.000	7.000
11.700	93.400	45.500	14.000	7.000	6.000
12.500	93.300	48.500	15.500	8.000	6.000
14.250	92.800	52.500	16.000	11.000	9.000
14.750	100.000	51.500	15.300	6.000	7.000
14.750	98.500	51.500	16.000	7.000	5.000
13.300	92.600	48.000	15.300	7.000	6.000
13.500	93.500	49.500	16.000	12.000	7.000
12.500	93.000	49.000	15.900	8.000	7.000
13.250	95.800	51.400	14.000	6.000	5.000
14.100	95.400	50.000	15.000	9.000	6.000
13.100	94.900	50.500	14.000	9.000	6.000
12.700	93.300	51.200	13.500	8.000	6.000
17.300	97.600	54.500	17.000	12.000	7.000
14.700	99.500	49.400	15.800	8.000	6.000
11.350	90.400	46.500	14.000	10.000	6.000
12.550	93.000	49.500	14.500	8.000	5.000
13.700	95.300	49.000	14.500	10.000	6.000
14.200	92.700	50.000	15.000	10.000	6.000

```
          X1              0.929             0.167
          X2              0.579             0.720
          X3              0.772             0.209
          X4              0.856            -0.100
          X5              0.441            -0.740
          X6              0.603            -0.533
 VARIANCE EXPLAINED BY COMPONENTS
                            1                 2
                        3.086             1.432
 PERCENT OF TOTAL VARIANCE EXPLAINED
                            1                 2
                       51.426            23.860
```

结果最前面的部分是PEARSON相关系数矩阵，接着是特征根(LATENT ROOTS),它体现了某一主成分对所有指标的总贡献。COMPONENT LOADINGS为主成分负荷量，它反映的是某一主成分对某个指标的贡献（即主成分所包含原来某个指标的信息量）。从结果中看出，第一主成分主要包含X1,X3,X4中的信息量; 而第二主成分主要包含X5,X2,X6的信息量;应该指出X6 在两个主成分中的作用基本是相等的。VARIANCE EXPLAINED BY COMPLAINED表示主成分的总贡献率。此处，二个主成分的贡献率已达到75.3需要可取三个或四个主成分，以便增加总贡献率，一般达到85

应该注意：要指定保留主成分的个数的方法，除了上面的NUMBER命令外，还有指定最小特征根的方式,其命令格式为:

EIGEN=<要保留的最小特征根>

不用这个命令，相当于EIGEN=0。

即保留所有特征根大于0的主成分。

如果既用了EIGRN 命令,又用了NUMBER 命令, 则根据两种标准所得的主成分数，哪个少就按哪个输出。

FACTOR 命令加上PLOT 可选项,则可打出指标Xi(i=1,2,⋯,m) 与主成分Zj (j=1,2,⋯,p) 之间的关系。其命令格式为：

FACTOR/PLOT 或

FACTOR <变量1>, <变量2>,... /PLOT

上面的例子加上PLOT 选择项，则输出散点图。图上每一个字母代表一个指标, 字母顺序与因子负荷的指标顺序相对应。很明显，FACTOR 1 主要反映了A,D, C 三个指标: FACTOR 2 主要反映了B，E 二个指标; 而对于F 指标，FACTOR1 和2 都有较大的反映。

FACTOR 模块还提供了一些可选择的命令:

⑴、因子负荷排序

如果在FACTOR 命令之前,键入SORT 命令, 就能使每个变量的负荷量按高到低顺序输出。这条命令仅仅改变负荷量的输出顺序，对结果的其他方面无任何影响。

⑵、因子旋转

最常用的是最大方差旋转。其命令格式为：

ROTATE=VARIMAX

用了这个命令后,先输出未旋转的结果,然后紧接着输出旋转后的结果。如果加上PLOT选择项,那么因子负荷图也是旋转后的。

经过最大方差旋转后,因子负荷量有所改变,也就是说,各指标在主成分中的作用大小改变了。在实际应用中,应根据指标的专业意义来决定是否需要旋转。理想的情况是,要使得每个主成分能突出反映所观察指标的某一部分的特征。

⑶、将部分结果存入SYSTAT 文件中

在每次键入FACTOR 命令之前,可以用SAVE 命令把算出的因子得分,负荷量或因子得分系数作为SYSTAT 文件存入盘中,以进一步分析。

SAVE <文件名>/SCORES(存因子得分) 或

SAVE <文件名>/LOADINGS(存负荷量) 或

SAVE <文件名>/COEF(存因子得分系数)

每次只能存一种结果。如果文件名后面不跟选择项,程序默认存入因子得分。若存入了得分,则文件中的变量名为FACTOR(i)(i=1,2,⋯,n)。存入的得分都经过了标准化,均值为0。如果计算用的是相关矩阵,则方差为 1 ,如果用的是协方差阵,而且未旋转,则因子得分不进行标准化,方差之和与原始数据算得的相同。

FACTOR 在默认状态下采用相关矩阵进行因子分解。但是,还允许用户选择协方差阵进行因子分解。

命令格式为:

TYPE=COVARIANCE

而不用TYPE 命令就相当于键入了

TYPE=CORRELATION (采用相关矩阵)

现在我们将选择的命令联合起来处理。拟选取三个主成分,按主成分各指标贡献的大小顺序输出,并作最大方差旋转,将因子得分存入SYSTAT 文件。操作如下:

```
C:\SYSTAT>FACTOR
 >USE EXAM161
 >SAVE SCORE
 >NUMBER=3
 >ROTATE=VARIMAX
 >SORT
 >FACTOR X1,X2,X3,X4,X5,X6/PLOT
```

取三个主要成份总的贡献率已达86.2 %。

在结果输出之后随即将因子得分存入名为SCORE 的文件中。这个文件中的变量名为FACTOR(1)和FACTOR(2)⋯。这些因子都经过了标准化,在必要时可利用因子得分作进一步的统计分析。例如可对SCORE 文件直接作聚类分析,把体格发育特征相似的儿童进行分类等等。

此外,取三个主成分,使用PLOT 选择项。所以,有了三个图(FACTOR 1 FACTOR 2,FACTOR 1 FACTOR 3,FACTOR 2 FACTOR 3)。其上的A、B、C、D、E、F 都是在这些平面上的投射,实质上表示了因子量的大小。

FACTOR 模块使用的注意事项:

FACTOR 命令对丢失数据的处理采用成行消除的原则。如果想用成对消除原则,在TYPE语句后面加上PAIRWISE 选择项:

TYPE=COVARIANCE/PAIRWISE 或TYPE=CORRELATION/PAIRWISE

程序在输出负荷时经过判断,如果主成分的负负荷大于正负荷,则取它的反方向, 这就避免了输出的因子带有很多负号。

㈦MDS 模块进行多维尺度变换。SYSTAT MDS 所使用的数据类型可以是协方差阵、相关阵、相似阵或不相似阵，建立时需要在DATA块中进行说明。现第四章的例子程序如下：

```
save CITY
NOTE 'INTERCITY FLYING MILEAGES'
TYPE= SIMILARITY
INPUT  ATLANTA,CHICAGO,DENVER,HOUSTON,LOSANGEL,MIAMI,NEWYORK,
       SANFRAN,SEATTLE,WASHDC
RUN
    0
  587    0
 1212  920    0
  701  940  879    0
 1936 1745  831 1374    0
  604 1188 1726  968 2339    0
  748  713 1631 1420 2451 1092    0
 2139 1858  949 1645  347 2594 2571    0
 2182 1737 1021 1891  959 2734 2408  678    0
  543  597 1494 1220 2300  923  205 2442 2329    0
RUN
SWITCHTO MDS
CHARSET GENERIC
METHOD=KRUSKAL
DIMENSION=2
SCALE
```

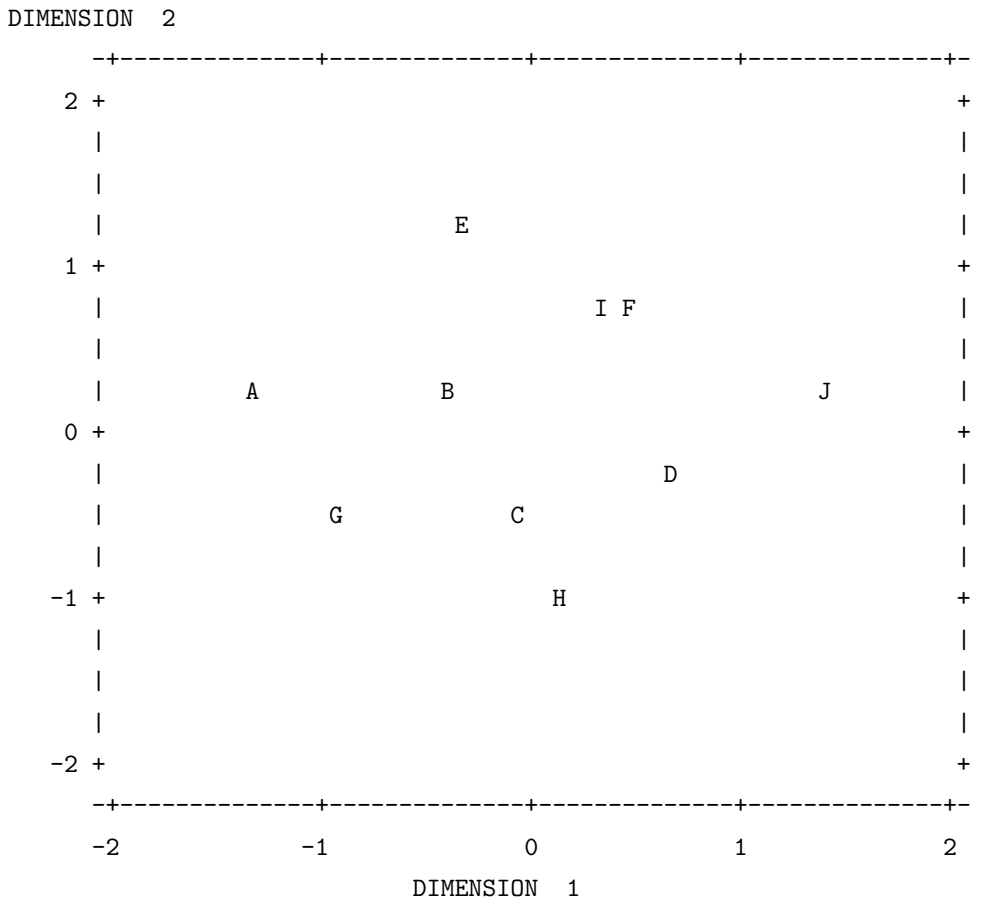
运行结果：

```
                          INTERCITY FLYING MILEAGES

 MONOTONIC MULTIDIMENSIONAL SCALING
 MINIMIZING KRUSKAL STRESS (FORM 1) IN 2 DIMENSIONS
 ITERATION  STRESS          ITERATION  STRESS
 ---------  ------          ---------  ------
     0       .419               8       .262
     1       .316               9       .260
     2       .297              10       .260
     3       .287              11       .260
     4       .280              12       .259
     5       .275              13       .259
     6       .269              14       .259
     7       .264
```

```
    STRESS OF FINAL CONFIGURATION IS:      .25909
                                  SHEPARD DIAGRAM
DISTANCES
      -+---------+---------+---------+---------+---------+---------+-
  3.0 +                                                             +
      |           +                                                 |
      |                                                             |
  2.5 +    +  +                                                     +
      |                                                             |
      |              +                            +                 |
  2.0 +              +            +                  +  +          +
      |            2         +                    +   ++  +        |
      |                 +          + +                              |
  1.5 +                       +            +                       +
      |                   ++              +   +       +   +         |
      |                   +                   +                     |
  1.0 +            + ++   +    2        2 2           +            +
      |                  22                                         |
      |                                                             |
  0.5 +                                                             +
      |                                                       +     |
      |                                                             |
  0.0 +                                                             +
      -+---------+---------+---------+---------+---------+---------+-
       0       500      1000      1500      2000      2500      3000
                              SIMILARITIES

    COORDINATES IN 2 DIMENSIONS
   VARIABLE    PLOT   DIMENSION
   --------    ----   ---------
                        1     2
    ATLANTA      A   -1.32   .24
    CHICAGO      B    -.39   .13
     DENVER      C    -.09  -.52
    HOUSTON      D     .69  -.47
   LOSANGEL      E    -.32  1.11
      MIAMI      F     .47   .51
    NEWYORK      G    -.95  -.54
    SANFRAN      H     .14 -1.23
    SEATTLE      I     .35   .63
     WASHDC      J    1.41   .14
```



§7.3 SYSTAT 4.1 简介

SYSTAT 4.1 模块共有17个，即DATA、EDIT、SSORT、MACPC、MACRO、GRAPH、STATS、TABLES、CORR、MGL、SYGRAPH，可通过执行其相应的.EXE文件而调用。文件SYSTAT.DEF 对各个模块进行了说明。

EDIT 模块允许交互式输入、编辑和转换数据。DATA 模块提供许多工具，用于产生和转换SYSTAT 数据文件。DATA 拥有文件管理工具、转换命令、BASIC 程序语言、Lotus, dBASE和DIF 数据文件转贮。SSORT 工具程序允许对SYSTAT文件快速排序，可以使用10个关键变量。MACPC 工具程序为Macintosh 和MS- DOS/PC- DOS 机间互换二进制SYSTAT 系统文件。MACRO 用于宏定义，即产生完成特定功能的程序。SYGRAPH 包可产生一系列2-维和3-维高分辨率图形，如散点图、矩阵图、平面和轮廓图、直方图、茎叶图、箱式图、Chernov脸谱图、概率分布和分位点图、圆图和条图。CLUSTER 模块提供原始数据或对称数据阵的聚类分析。相关分析模块用于计算对称相关或相似三角阵。FACTOR 进行主成分分析，进行旋转和计算因子得分。GRAPH 模块产生字符统计图形，如散点图、直方图、茎叶图、箱式图、概率分布图和分位点图。MDS 在1-5 维空间对相似或不相似阵进行非度量的多维尺度变换。MGLH 过程用于估计和检验一元或多变元线性模型。NONLIN 模块实现拟牛顿(Quasi-Newton)和单纯形法非线性估计，可对极大似然和相关方法指定损失函数。NPAR 模块进行非参数统计。SERIES 模块用于时序分析。三条基本的命令实现一系列时序分析模型，包括Box-Jenkins ARIMA, Fourier 分析以及线性和非线性滤波。STATS 计算综合统计量。TABLES 模

块用于产主多维列联表并拟合对数线性模型。SYSTAT 4.1支持数学协处理器。SYSTAT 4.1的SURVIVAL 和LOGIT 能够进行Cox 回归和多分类LOGIT 分析。基本的模块需要九张360 KB 软盘，有专门的安装程序INSTALL，其安装大致与第一节中介绍的一样。各模块的调用风格也保持了原来的特色。

SYSTAT 4.1 增强了3.0 的共用命令和分模块命令，这里介绍几个。

1. SWITCHTO 'module'[<file>/ECHO]

   切换到另一个SYSTAT 模块。原来的版本中，不同的分析要用不同的模块，当进行模块切换时，需要重新调用数据，提供这个命令以后，就不必这么做了。

2. DATA 模块命令IMPORT/EXPORT 命令可用于转贮外部文件的数据，用法详见第16章。

   CASELIST 命令在热命令RUN 运行之后进行记录的列表，格式：

   CASELIST [<变量1>, <变量2>, < ... >]

   用例：

   CASELIST (列出整个文件)

   CASELIST MURDER,ROBBERY (列出所有记录的MURDER 和ROBBERY)

   同样可以用REPEAT N 命令列出前面N 个记录。

3. FEDIT 命令启动SYSTAT 文件编辑，可对任何ASCII 文本文件，包括SYSTAT 命令文件和输出文件，在文件编辑时可使用块标记。

   语法：FEDIT <file> | * |#

   用例：

   FEDIT 文件名(编辑新文件或旧文件，永久保留改动)

   FEDIT * (浏览最近一次屏幕输出，可倒数256 行)

   FEDIT > (浏览命令记录文件，编辑和重新提交命令)

   FEDIT # (编辑当前SYSTAT 输出文件)

4. FPATH 命令指定一个自动前缀给SYSTAT 文件，用于和特定的目录或设备联系，有七种文件可以分别加上前缀。

   GET 指GET 命令的ASCII 输入文件(.DAT)

   OUTPUT 指PUT 和OUTPUT 命令的ASCII 输出文件(.DAT)

   SAVE 指SYSTAT 输出文件(.SYS)

   SUBMIT 指SYSTAT 命令文件(.CMD)

   USE 指所有SYSTAT 输入数据文件(.SYS)

   FEDIT 指所有由FEDIT 存取的文件

   TRANSFER 指DATA 模块中所有由IMPORT 或EXPORT 命令存取的文件

   语法：FPATH 'prefix' / GET OUTPUT SAVE SUBMIT USE FEDIT TRANSFER

   用例：

FPATH 'D:' / SAVE (指示所有SYSTAT 输出数据文件到D:)

FPATH '\MYDATA\'/USE GET SAVE (.DAT 和.SYS 文件在\MYDATA)

FPATH 'C:\USR\SYSTAT\' /SUBMIT (.CMD 文件在C: 的目录)

5. CHARSET 命令选择IBM 屏幕/打印机图形字符或通用字符。

语法：CHARSET GRAPHICS | GENERIC

用例：

CHARSET GRAPHICS

CHAR GENERIC (使用通用字符)

第八章　Stata

§8.1　应用概要

§8.1.1　简介

Stata 是一个非常实用的统计软件包，具有数据显示与管理、统计分析和统计作图等功能，与电子报表程序、数据库管理程序、统计软件包、图形软件包和程序语言有着共同点。它的设计注重实用，系统小从而不但使用方便，而且功能强大。从3.0版以后功能大大增强。从4.0开始，有了Windows产品，可以全屏幕浏览数据，仅仅使用鼠标器便从review窗口和变量窗口挑选用过的命令和变量，十分方便，输出结果可以将命令醒目地标示出来。另外，其宏定义ADO对用户是透明的，特别适于研究者实现新的分析方法。5.0中增加了长期数据分析和复杂调查分析等内容。

Stata 可用于各种计算机，如IBM PC、PS/2，80386、68000 、VAX/VMS 和UNIX 工作站，但各种机型的文件可以自动转换。

§8.1.2　系统运行

㈠、进入和退出

建议DOS 的CONFIG.SYS 文件中增加语句DEVICE=[d:][path]ANSI.SYS，这样在软件使用时，将显示以彩色。在DOS 提示符下，打入

D:\Stata>stata

出现系统提示符(.)。

Windows版需要运行wstata。使用命令行参数/kxxx可以指示所占用内存的大小。stata 支持数学协处理器和EMS。

Stata 只识别以小写字母打入的命令。

退出仅需打入exit 命令(放弃改动过的数据时用exit,clear 命令)。

㈡、启用帮助和菜单

在提示符下打入help (F1+Enter) 命令，软件显示Stata 的全部命令，这一点与FoxBASE+有些类似。在新版本Stata中，使用help contents给出详细的命令分类，使用lookup可以查找命令所关联的产品。

编辑键：Ctrl-字母键的组合：W(End)、E(Insert)、R(estore)、U(Esc)、J(Enter)、M(Enter)、D(elete)、G(Begin of Word)、O(Being of line)、P(End of line)。在SUN 计算机上尚有R1-R15键。其它的如↑、↓、←、→移动，Ins插入、Del 删除。行首(Home)、行尾(End)、上一行(PgUp)、下一行(PgDn)、跳格(Tab)、寻找(Ctrl-Home 加行号)。可用#review [n] 显示已入过的命令。Windows 版功能键略有不同。

在安装了stata/mnu时，可用F10 进入菜单系统。菜单项以光标键和跳格键结合回车键进行选择。在Windows版本中使用edit和browse编辑和浏览数据工作表。

㈢、语言特色

Stata 的数据存贮类型有五种，即整型、单精度或双精度浮点数。5、-5、5.2、.5、5.2E+2、5.2D+2均是合法的表示。有以下的表格供参考：

字符串存贮类型有str2,str4,...,str80，如"Example"、"2.1"、""。

Stata 的名称是1-8 个字母、数字、下线，对大小写是敏感的。Stata保留使用以下变量：

_all(所有变量)、float(指示浮点类型)、_rc(返回码)、_skip(跳记录)、_b(回归系数)、if(条件)、_n(当前记录)、using(使用文件名)、_coef(系数)、in(观察范围)、_N(总观察)、_weight(体重)、_cons(常数)、int(整型指示)、_pi(π)、with(与)、double(双精度)、long(长整数)、_pred(预测值)。

Stata 的原始数据是一个矩形的表格，不同的变量以变量名表示观测值由1至N。Stata 同时使用交叉乘积类型的资料。设原始数据矩阵为X，第一列由系统变量_cons 组成，其每个元为 0，X'X称为XP形式，可由命令convert 形成。set contents xp 可以把第一个量改名为_cons。

Stata 的使用以下的文件指示：

.dct ASCII 数据字典
.do 运行文件
.dta Stata 格式的数据文件
.gph 图形文件
.log 运行记录文件
.raw ASCII 格式的数据集
.XP Stata 格式的交叉乘积数据集

文件操作具有基本格式：<文件操作关键字> [变量列表] using < 文件名>，如：

```
infile x1 x2 str10 (x3 x4) y1-y10 using myin
outfile using myout
graph using mygph
log using mylog
```

第一句和第二句分别读入文件myin 和写出文件myout，第三句调入并显示图形文件mygph，第四句用mylog 文件存放结果，打开以后，以LOG ON 和LOG OFF 作记录运行信息的切换开关。

除少数例外，Stata 使用的语句格式为：

[by varlist] 命令[varlist] [=exp] [if exp] [in range] [,options]。其中varlist 为变量列表。exp 为施加的权，exp 是一个容纳条件的表达式，range 表示范围，options 是有关的选项，一般来说options 与nooptions 对应。某些选项含有参量应该放在括号内。

Stata 把开头为星号(*)的行视为注释，在.DO文件中可用/* */ 表示。

Stata 的表达式与其它软件的用法并没有区别，与C 语言类似。

数学函数：abs(x)、acos(x)、asin(x)、atan(x)、comb(x)、cos(x)、exp(x)、ln(x)、lnfact(x)、lngamma(x)、log(x)、log10(x)、

统计函数：

Binomial(n,k,π) 在n次观察中k次以上的概率
binorm(h,k,ρ) 相关系数为 ρ 的两维正态累积分布
chiprob(df,x)、invchi(df,p) 自由度为x 的累积卡方。
fprob(df1,df2,f)、invfprob(df1,df2,p) F-分布。
gammap(a,x)、invgammap(a,p) 不完全伽马分布
ibeta(a,b,x) 不完全贝塔分布
invbinomial(n,k,p) 逆二项分布

invnchi(df,l,p)、nchi(df,l,x)、npnchi(df,x,p) 非中心卡方分布。
invnorm(p)、normd(z)、normprob(z) 正态分布。
invt(df,p)、tprob(df,t) t分布。
uniform()均匀伪随机数

日期函数：date(s1,s2)、day(e)、dow(e)、mdy(m,d,y)、month(e)、year(e)。
字符串函数：index(s1,s2)、length(s)、lower(s)、ltrim(x)、real(s) 、rtrim(x)、string(n)、substr(s,n1,n2)、trim(x)、upper(s
其他函数：autocode(x,ng,xmin,xmax)、cond(x,a,b)、float(x)、group(x)、int(x)、max(x1,...,xn)、min(x1,...,xn)、sum(x)、
, xn) 、round(x,y)、sign(x)、sum(x)。

Stata 编程首先要了解参数的宏传递的，所以可用命令'help macros' 获得一些信息，如程序名后第一个参数是%_1，第二个是%_2，等等。设数据集中的缺失值记为-9，要把它们改为Stata 的缺失值，相当于命令：

replace varname=. if varname==-9

但若有十个这样的变量，打很多这样的命令是很烦琐的。借助程序

```
program define fix
  replace %_1=. if %_1===9
end
```

就产生了名为fix 的程序名，此后仅仅使用'fix var1', 就相当于执行：'replace var1=. if var1==9'，其余的情形依次类推。操作如下：

```
. program define fix
 1. replace %_1=. if %_1==9
 2. end
```

设原始数据变量为z，则命令fix z 启用fix的操作。

㈣、运行实例

Stata 提供了一套较完整的教学程序，如：intro.tut,graphics.tut,tables .tut,anova.tut,regress.tut,probit.tut survive.tut, statkit.tut, graphkit. tut, datakit.tut, qckit.tut, ourdata.tut, yourdata.tut.使用do或tutorial 命令。如intro.tut的分析：

```
. use c:\stata\auto
. describe
. drop mpg hdroom runk length turn displ gratio
. list in 1
. summarize rep78
. tabulate rep78, plot
. list if rep78==1
. list if rep78==5
. tabulate rep78 foreign
. tabulate rep78 foreign, column
. tabulate rep78 foreign, column nofreq chi2
. tabulate foreign, summarize(rep78)
```

```
. graph weight price
. plot weight price
. correlate weight price
. regress weight price
. predict resid, residual
by foreign: summarize weight price resid
. generate weightd = weight if ~foreign
. generate weightf = weight if foreign
. graph weightd weightf price
. plot weightd weightf price
. graph weightd weightf price
. graph price weight, by(foreign) total
. plot weightd weightf price
. regress weight price foreign
```

变量rep78 是消费者的报告维修排名，用summarize 来观察，其次是最坏的车型和最好的车型。结果好的车型似以外国车为主，平均来说是这样的吗？变量'foreign' 是一个示性量，若车为国产，其值为0 否则为1，借助它对两种类型行制表。使用绝对数和百分比，并计算国产车与外国车之间的卡方统计量，而且比较两者之间的平均维修记录。另外看车重与车的价格，关系如何？一种方式是通过绘图来观察。图形显示，随着价格的上升，车重也增加了。没有图形设备时，使用plot 图示效果也一样。可以算得车重与价格的相关系数，预计会是一个正值。进一步，有车重与价格的回归分析。回归的残差用一个名为'resid' 的量来表示。回归分析之后，Stata 存贮了这些变量以及系数，predict 则使用了这些信息，结合原始数据来计算predictions, residuals, 以及influence statistics。这样可以观察，国产与进口车维修记录有所不同，在车重/价格关系上也是这样的吗？图形显示，同一价格下的国产车几乎均比进口车要重。

§8.2 统计分析

§8.2.1 统计制表

命令为：tabulate

58 个病人的材料，它们患有三种病，服用四种药，记录他们血压的变化。原始数据为：

	Disease 1	Disease 2	Disease 3
Drug 1	42,44,36	33,26,33	31,-3,25
	13,19,22	21	25,24
Drug 2	28,23,34	34,33,31	3,26,28
	42,13	36	32,4,16
Drug 3	1,29,19	11,9,7	21,1,9
		1,-6	3
Drug 4	24,9,22	27,12,12	22,7,25
	-2,15	-5,16,15	5,12

```
. use systolic, clear
```

```
. tabulate drug
. tabulate drug disease
```

使用命令tabulate 产生列表。在两维表中，Stata 可以有许多选择项，如nofreq (不打印频数)、column (报告列百分比)、row (报告行百分比) 及cell (格子百分比)。下表显示了disease与drug 交叉表有关统计量，命令、选项和变量名都略写了。

```
. tab dis dr, col row cell
. tab disease drug, chi2
  Patient's| Drug Used
    Disease|         1          2          3          4      Total
-----------+--------------------------------------------+----------
         1 |         6          5          3          5 |         19
         2 |         4          4          5          6 |         19
         3 |         5          6          4          5 |         20
-----------+--------------------------------------------+----------
     Total |        15         15         12         16 |         58

           chi2(6) =   1.4048  Prob>chi2 = 0.966
```

对于多维表格也类似，如关于三个量var1、var2 及var3, 使用：

```
. sort var1
. by var1:  tabulate var2 var3
```

每种药物与收缩压的关系列表。

```
. tabulate drug, summarize(systolic)
           | Summary of Increment in Systolic B.P.
  Drug Used|        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |       26.07       11.68          15
         2 |       25.53       11.62          15
         3 |        8.75       10.02          12
         4 |       13.50        9.32          16
-----------+------------------------------------
     Total |       18.88       12.80          58
. tab disease drug, sum(systolic)
. tab disease drug, sum(systolic) mean
. sort var1
. by var1: tab var2 var3, sum(contvar)
```

产生分组统计量，结果包括了均值、标准差及收缩压增加频数或者仅仅显示其中某个统计量(mean, standard, freq)，多维也类似。

§8.2.2 方差与协方差分析

进行方差分析使用命令oneway、anova 和test。现仍然采用制表部分所提供的58 个病人的材料。oneway 命令估计one-way ANOVA 模型并能做两两比较，其语法为：oneway 反应变量控制变量，此处

```
. oneway systolic drug
. oneway systolic drug, tabulate
. oneway systolic drug, bonferroni
. test drug
. test drug drug*disease
. test drug, error(drug*disease)
. anova systolic drug disease
. regress
. anova systolic drug disease
. anova systolic disease drug, sequential
. anova systolic drug disease drug*disease
. test _coef[ drug[2] ] = _coef[ drug[3] ]
. test _coef[disease[2]] + 3*_coef[disease[3]] = 6 + _coef[disease[3]]
. quietly anova systolic drug disease drug*disease
. test, symbolic
. test drug, symbolic
. use sysage, clear
. summarize age
. anova systolic drug disease age disease*age, continuous(age)
```

程序结果如下：

```
                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3133.23851      3   1044.41284      9.09     0.0001
 Within groups      6206.91667     54   114.942901
------------------------------------------------------------------------
    Total           9340.15517     57   163.862371

Bartlett's test for equal variances:  chi2(3) =   1.0063  Prob>chi2 = 0.800
```

Summary of Increment in Systolic B.P.			
Drug Used	Mean	Std. Dev.	Freq.
1	26.07	11.68	15
2	25.53	11.62	15
3	8.75	10.02	12

```
           4 |       13.50        9.32          16
-------------+------------------------------------
       Total |       18.88       12.80          58

Bartlett's test for equal variances:  chi2(3) =   1.0063  Prob>chi2 = 0.800
           Comparison of Increment in Systolic B.P. by Drug Used
                                (Bonferroni)
Row Mean-|
Col Mean |          1          2          3
---------|---------------------------------
       2 |      -0.53
         |      1.000
         |
       3 |     -17.32     -16.78
         |      0.001      0.001
         |
       4 |     -12.57     -12.03       4.75
         |      0.012      0.017      1.000

                    Number of obs =      58      R-square     =  0.4560
                    Root MSE      = 10.5096      Adj R-square =  0.3259
           Source |  Partial SS    df       MS           F     Prob > F
------------------+----------------------------------------------------
            Model |  4259.33851    11  387.212591       3.51     0.0013
                  |
             drug |  2997.47186     3  999.157287       9.05     0.0001
          disease |  415.873046     2  207.936523       1.88     0.1637
     drug*disease |  707.266259     6   117.87771       1.07     0.3958
                  |
         Residual |  5080.81667    46  110.452536
------------------+----------------------------------------------------
            Total |  9340.15517    57  163.862371

           Source |  Partial SS    df       MS           F     Prob > F
------------------+----------------------------------------------------
             drug |  2997.47186     3  999.157287       9.05     0.0001
         Residual |  5080.81667    46  110.452536
           Source |  Partial SS    df       MS           F     Prob > F
------------------+----------------------------------------------------
drug drug*disease |  3770.69912     9  418.966569       3.79     0.0012
         Residual |  5080.81667    46  110.452536
           Source |  Partial SS    df       MS           F     Prob > F
```

```
           -----------+----------------------------------------------------
                 drug |  2997.47186     3  999.157287       8.48     0.0141
        drug*disease |  707.266259     6   117.87771

  Source |       SS         df       MS           Number of obs =      58
---------+------------------------------          F(  5,    52) =    6.38
   Model |  3552.07225      5  710.414449          Prob > F      =  0.0001
Residual |  5788.08293     52  111.309287          R-square      =  0.3803
---------+------------------------------          Adj R-square  =  0.3207
   Total |  9340.15517     57  163.862371          Root MSE      =   10.55

                      Response variable is systolic
------------------------------------------------------------------------------
Variable    Coefficient    Std. Error       t    Prob > |t|        Mean
------------------------------------------------------------------------------
_cons          9.878751      3.317476     2.978     0.004             1
drug
       1       12.46897      3.807342     3.275     0.002      .2586207
       2       12.36457       3.80698     3.248     0.002      .2586207
       3      -4.526825      4.033439    -1.122     0.267      .2068966
       4      (dropped)
disease
       1       6.434081      3.388837     1.899     0.063      .3275862
       2       4.294931      3.400646     1.263     0.212      .3275862
       3      (dropped)
------------------------------------------------------------------------------

( 1)  drug[2] - drug[3] = 0.0, F(1, 52) = 16.89, Prob > F = 0.0001
( 1)  disease[2] + 2.0 disease[3] = 6.0, F(1,52) = 0.25, Prob > F= 0.6182

                      Number of obs =      58      R-square     =  0.6221
                      Root MSE      = 8.48737      Adj R-square =  0.5604

             Source |  Partial SS    df       MS           F     Prob > F
           ---------+----------------------------------------------------
              Model |  5810.41855     8  726.302318      10.08     0.0000
                    |
                drug |  2791.94475     3  930.648251      12.92     0.0000
            disease |  129.092158     2  64.5460789       0.90     0.4148
                age |  1817.80067     1  1817.80067      25.23     0.0000
        disease*age |  43.4069507     2  21.7034754       0.30     0.7412
                    |
```

```
    Residual |  3529.73663    49  72.0354414
-------------+------------------------------------------------
       Total |  9340.15517    57  163.862371
```

算得的组间方差提示，接受不同药物的病人之间收缩压的改变是显著的。使用tabulate 选项时，Stata 报告控制变量每个水平上的均值，其余结果相同。本例用Bonferroni方法进行两两比较的方法有。首先指示一个两因素、没有交互的模型，反应变量是血压，控制变量是药物和疾病种类。检验中用的平方和是偏平方和。使用test语句检验药物的作用以及交互作用：除非特别指定，Stata 使用的是用剩余均方作为误差项，否则要进行特指。方差分析之后，又能得到回归。test 语句也可用于检验回归系数，此处检验第二、第三种药物的系数是否相同。Stata 还可以给出用记号表示的函数和特定检验的可估计函数。在明确那些变量是连续的之后，可进行协方差分析。以上采用的协变量是age。

§8.2.3 回归分析

使用语句：regress、test、predict 和stepwise。现用人口普查的资料进行说明。运行的程序为：

```
    . describe
    . summarize
    . regress drate medage medagesq pcturban
    . test medage medagesq
    . test medage=2*medagesq
. test 2*(medage-medagesq)-(medage-medagesq)/2=(medage-medagesq)/2+medagesq
    . test pcturban=medage, accumulate
    . correlate, _coef
    . stepwise drate medage medagesq pcturban
    . predict dhat
    . summarize drate dhat
    . predict influ, cooksd
    . summarize influ, detail
    . list state if influ>1
    . regress drate medage medagesq pcturban if influ<1
    . summarize pop
    . regress drate medage medagesq pcturban =pop
    . use hsng, clear
    . keep hsngval faminc rent pcturban region
    . describe
    . regress rent hsngval pcturban (faminc reg2-reg4 pcturban)
```

其结果为：

```
  Source |       SS       df       MS              Number of obs =      50
---------+------------------------------           F(  3,    46) =   31.47
```

```
   Model |   .00005593     3  .000018643        Prob > F      =  0.0000
Residual |  .000027249    46  5.9236e-07        R-square      =  0.6724
---------+------------------------------        Adj R-square  =  0.6510
   Total |  .000083179    49  1.6975e-06        Root MSE      =  .00077

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+--------------------------------------------------------------
   drate |                                                      .008436
---------+--------------------------------------------------------------
  medage |     .0004851       .001207     0.402     0.690         29.54
medagesq |     2.37e-06      .0000206     0.115     0.909       875.422
pcturban |    -.0000353      8.29e-06    -4.262     0.000      66.94913
   _cons |     -.005598      .0178979    -0.313     0.756             1
---------+--------------------------------------------------------------

     ( 1)  medage = 0.0
     ( 2)  medagesq = 0.0
           F(  2,    46) =   44.03, Prob > F =    0.0000
     ( 1)  medage - 2.0 medagesq = 0.0
           F(  1,    46) =    0.15, Prob > F =    0.7021
     ( 1)  medage - 2.0 medagesq = 0.0
           F(  1,    46) =    0.15, Prob > F =    0.7021
     ( 1)  medage - 2.0 medagesq = 0.0
     ( 2) - medage + pcturban = 0.0
           F(  2,    46) =   21.85, Prob > F =    0.0000

             |   medage medagesq pcturban    _cons
    ---------+------------------------------------
      medage|   1.0000
    medagesq|  -0.9985   1.0000
    pcturban|   0.3235  -0.3352   1.0000
       _cons|  -0.9984   0.9942  -0.3385   1.0000

Variable |   Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
   drate |    50     .008436    .0013029   .0039915   .0106902
    dhat |    50     .008436    .0010684   .0045433   .0111047

  Source |     SS         df       MS           Number of obs =      50
---------+------------------------------        F(  3,    46) =   40.83
   Model |  .000038537     3  .000012846        Prob > F      =  0.0000
Residual |  .000014471    46  3.1458e-07        R-square      =  0.7270
```

```
---------+------------------------------        Adj R-square  =  0.7092
   Total |  .000053008    49  1.0818e-06        Root MSE      =  .00056

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+------------------------------------------------------------------
   drate |                                                       .0087368
---------+------------------------------------------------------------------
  medage |     .0005844       .000994     0.588     0.559      30.11047
medagesq |    -4.36e-07      .0000163    -0.027     0.979      909.3716
pcturban |    -.0000285      6.53e-06    -4.368     0.000      73.66408
   _cons |    -.0063644      .0152071    -0.419     0.678             1
---------+------------------------------------------------------------------

Dropping: medagesq F=      0.0133
                                                              (stepwise)
  Source |     SS         df       MS           Number of obs =      50
---------+------------------------------        F(  2,    47) =   48.22
   Model |  .000055922     2  .000027961        Prob > F      =  0.0000
Residual |  .000027256    47  5.7993e-07        R-square      =  0.6723
---------+------------------------------        Adj R-square  =  0.6584
   Total |  .000083179    49  1.6975e-06        Root MSE      =  .00076

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+------------------------------------------------------------------
   drate |                                                        .008436
---------+------------------------------------------------------------------
  medage |     .0006238      .0000658     9.483     0.000         29.54
pcturban |     -.000035      7.73e-06    -4.531     0.000      66.94913
   _cons |    -.0076466      .0019034    -4.017     0.000             1
---------+------------------------------------------------------------------
                                                                  (2SLS)
  Source |     SS         df       MS           Number of obs =      50
---------+------------------------------        F(  2,    47) =   42.66
   Model |  36677.4033     2  18338.7017        Prob > F      =  0.0000
Residual |  24565.7167    47  522.674823        R-square      =  0.6448
---------+------------------------------        Adj R-square  =  0.6297
   Total |    61243.12    49  1249.85959        Root MSE      =  22.862

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+------------------------------------------------------------------
    rent |                                                        234.76
---------+------------------------------------------------------------------
```

```
 hsngval |     .0022398      .0003388     6.612     0.000         48484
pcturban |      .081516      .3081528     0.265     0.793      66.94913
   _cons |     120.7065      15.70688     7.685     0.000             1
---------+------------------------------------------------------------
```

本例用regress 命令进行线性回归。回归方程为：

drate $= b_0 + b_1$ medage $+ b_2$ medagesq $+ b_3$ pcturban

drate 是州死亡率，medage 是州人口龄中位数，medagesq 是年龄平方中位数，pcturban 是生活在城区的人口比例。回归结束后，使用regress 命令可以重显回归的结果。使用test 命令进行有关的检验，若指定accumulate 选项后，可以检验多个假设。估计量的协方差阵可以使用'correlate, _coef'而得。逐步回归方法有向后(backward)、向前(forward)、逐步(stepwise)，进入和剔除的F-值由fenter(#) 和fstay(#) 指定，默认值分别为0.5 和0.1。

其它的量有：预测值、残差、标化残差、学生化残差、预测标准误、影响统计量(Cook's 距离、投影阵对角元和DF-Betas。一个观察是高影响点，现把它找出来。除掉该点继续估计。regress 命令也能进行加权回归，现在数据中含变量pop 表示每处的人口数，最后一个是工具变量法和两阶段最小二乘解。语句格式为regress lhsvar rhsvar1 rhsvar2 ... (exogvar1 exogvar2 ...)，之后进行假设检验或计算预测值了。

数据hsng 记录了1980 人口普查数据分析(median housing values 及rents)，估计的模型是：

hsngval $= a_0 + a_1$ faminc $+ a_2$ reg2 $+ a_3$ reg3 $+ a_4$ reg4

rent $= b_0 + b_1$ hsngval $+ b_2$ pcturban

§8.2.4 logit/probit 分析

命令probit 和logit，仍然采用汽车数据示范。

```
. use auto, clear
. keep make foreign mpg weight
. describe
. inspect foreign
. logit foreign weight mpg
. logit, tabulate nocoef
. probit foreign weight mpg
. predict probhat
. summarize probhat
. list in 1/13
. predict zhat, index
. summarize probhat zhat
. correlate, _coef
. cor, _c cov
. test (weight-mpg)/2 - (mpg-weight) = mpg - (weight+mpg)*2
. test mpg=0, accumulate
. probit foreign weight mpg if mpg>18
. tabulate foreign repair
. probit foreign repair1 repair2
```

运行结果：

```
foreign:  Car type                          Number of Observations
------------------                                                 Non-
                                            Total   Integers    Integers
|  #                          Negative          -          -           -
|  #                          Zero             52         52           -
|  #                          Positive         22         22           -
|  #                                          -----      -----       -----
|  #   #                      Total            74         74           -
|  #   #                      Missing           -
+----------------------                      -----
0                     1                       74
   (2 unique values)

Logit Estimates                               Number of obs =     74
                                              chi2(2)       =  35.72
Log Likelihood =-27.175156                    Prob > chi2   = 0.0000

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+--------------------------------------------------------------
 foreign |                                                       .2972973
---------+--------------------------------------------------------------
  weight |    -.0039067      .0010116    -3.862     0.000      3019.459
     mpg |    -.1685869      .0919175    -1.834     0.071       21.2973
   _cons |     13.70837      4.518709     3.034     0.003             1
---------+--------------------------------------------------------------

        Comparison of Outcomes and Probabilities

     Outcome |    Pr < .5    Pr >= .5 |      Total
    ---------+-------------------------+-----------
     Failure |         46           6 |         52
     Success |          9          13 |         22
    ---------+-------------------------+-----------
       Total |         55          19 |         74

Probit Estimates                              Number of obs =     74
                                              chi2(2)       =  36.38
Log Likelihood =-26.844189                    Prob > chi2   = 0.0000

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
```

```
---------+--------------------------------------------------------------
 foreign |                                                     .2972973
---------+--------------------------------------------------------------
  weight |    -.0023355      .0005661    -4.126     0.000      3019.459
     mpg |    -.1039503      .0515689    -2.016     0.048       21.2973
   _cons |     8.275464      2.554142     3.240     0.002             1
---------+--------------------------------------------------------------

    Variable |   Obs        Mean    Std. Dev.       Min        Max
    ---------+-----------------------------------------------------
     probhat |    74     .294487    .3074146   9.52e-06   .9029781
        zhat |    74   -.9904795    1.376307  -4.275976   1.298709

             |   weight      mpg    _cons
    ---------+---------------------------
      weight|   1.0000
         mpg|   0.7809   1.0000
       _cons|  -0.9504  -0.9305   1.0000
             |   weight      mpg    _cons
    ---------+---------------------------
      weight|  3.2e-07
         mpg|  .000023  .002659
       _cons| -.001374 -.122554  6.52364

   ( 1)  3.5 weight - .5 mpg = 0.0, F(1, 71) = 3.26, Prob > F = 0.0753
   ( 1)  3.5 weight - .5 mpg = 0.0
   ( 2)  mpg = 0.0
           F(  2,    71) =   10.38
                Prob > F =    0.0001

Probit Estimates                                  Number of obs =     47
                                                  chi2(2)       =  23.10
Log Likelihood =-19.203993                        Prob > chi2   = 0.0000

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+--------------------------------------------------------------
 foreign |                                                     .3617021
---------+--------------------------------------------------------------
  weight |    -.0023289      .0007581    -3.072     0.004      2642.766
     mpg |    -.0302598       .060018    -0.504     0.617      24.34043
   _cons |      6.14531       3.02718     2.030     0.048             1
---------+--------------------------------------------------------------
```

```
Probit Estimates                                  Number of obs =     48
                                                  chi2(1)       =   9.53
Log Likelihood =-22.229138                        Prob > chi2   = 0.0020

Variable |  Coefficient    Std. Error       t    Prob > |t|        Mean
---------+--------------------------------------------------------------
 foreign |                                                          .25
---------+--------------------------------------------------------------
 repair2 |    -1.281552      .4297326    -2.982     0.005          .625
   _cons |    -3.75e-16       .295409    -0.000     1.000             1
---------+--------------------------------------------------------------
```

logit 和probit 用于分析属性因变量的情形，这里用是否国产车或进口车作为分析变量，从inspect我们明确分析变量具体的取值。foreign 加了标号，所有取值存在标号里。取协变量为车重和里数做回归，计算开始时对数似然值为- 45. 03321，迭代5次后收敛于-27.175156，继续用logit 可以重显计算结果。probit 和logit 命令都可以产生一个原始分类与计算分类的比较表，这由tabulate选项完成。probit 计算迭代初值为亦为-45.03321，5次后的结果为-26.844189。probit 和logit 均由迭代技术求取非线性估计量，收敛性可由三个选项控制。iterate(#) 指最大迭代次数，tolerance(#) 指示系数容许性，ltolerance(#) 指示对数似然函数的容许性。predict 命令同样可以用于预测，这里的预测结果是概率值。predict 也能计算index 函数的预期值，对于probit 分析，预测概率是probhat = F(zhat)，F() 是累积正态分布。Stata 的probit 和logit 命令与其它命令的特性一样，模型分析之后，可以得到估计量基于信息矩阵的协方差阵，即信息阵的逆阵。根据这些信息进行检验，accumulate 选项允许检验的累积，因而是联合检验。其它的特征，如指示子集分析，对数似然函数迭代初值为-30.756389，5 次收敛于-19.203993。特别重要的是，当模型中的因变量完全决定了结果的成功与否，那么许多程序会算出无穷大的系数，或者产出不过是四舍五入误差的解，Stata 能及时以现这种情形。汽车数据增加了一个名为repair 的量，取值为1, 2, 3, 表示差、平均和好。由于数据中维修记录最差的均是国产车，若要用维修记录来预计外国厂商，则repair==1 的概率为0，即probit 或logit的系数会无穷小。而Stata 的结果正确。迭代自-26.992087 始， 4 次收敛于-22.229138。

§8.2.5 生存分析

使用命令kapmeier, gwood, mantel, wilcoxon, logrank, survcurv, loglogs, cox, coxvar, coxbase, boxhaz. cox 命令能对有关时间、截尾和时变协变量资料估计比例风险模型，现用一个发电机的资料，比较新旧类型的轴承，记录这些发电机在超负荷下运行直至损坏的小时数。

```
. use kva, clear
. describe
. cox failtime load bearings
. correlate, _coef
. correlate, _coef covariance
```

运行结果：

```
Cox regression                                  Number of obs =     12
                                                chi2(2)       =  23.39
Log Likelihood = -8.577853                      Prob > chi2   = 0.0000

Variable |  Coefficient    Std. Error      t     Prob > |t|      Mean
---------+------------------------------------------------------------
 failtime|                                                    74.66667
---------+------------------------------------------------------------
     load|     .4229578      .1433485     2.951     0.015          27.5
 bearings|    -2.754461      1.173115    -2.348     0.041            .5
---------+------------------------------------------------------------

             |     load bearings
    ---------+------------------
        load|  .020549
    bearings| -.121008   1.3762
```

对数似然迭代初值为-20.274897，经5次迭代收敛至-8.577853。load 与bearings 间的相关为-0.7196。不带参数执行cox 命令时，也会重显回归的结果。

估计带有截尾观察的数据，使用的是一个药物实验的数据。

```
. use cancer, clear
. describe
. tabulate died, summarize(studytim)
. tabulate drug, summarize(studytim), if died
. tabulate drug, summarize(died)
. quietly tabulate drug, gen(drug)
. cox studytim age drug2 drug3, dead(died)
. test drug2=drug3

1 if patient| Summary of Months to death or end of exp.
        died|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   21.117647   10.652948          17
          1 |   12.419355   8.7512825          31
------------+------------------------------------
      Total |        15.5    10.25629          48

   Drug type| Summary of Months to death or end of exp.
 (1=placebo)|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   9.0526316   6.6204539          19
          2 |        14.5   7.2318739           6
```

```
          3 |          21   10.620734           6
------------+------------------------------------
      Total |   12.419355   8.7512825          31
  Drug type|     Summary of 1 if patient died
 (1=placebo)|        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |         .95    .2236068          20
          2 |   .42857143   .51355259          14
          3 |   .42857143   .51355259          14
------------+------------------------------------
      Total |   .64583333   .48332111          48

Cox regression                                  Number of obs =     48
                                                chi2(3)       =  36.52
Log Likelihood =-81.652567                      Prob > chi2   = 0.0000

Variable |  Coefficient    Std. Error      t     Prob > |t|      Mean
---------+--------------------------------------------------------------
 studytim|                                                         15.5
     died|                                                       .6458333
---------+--------------------------------------------------------------
      age|       .11184      .0365789     3.058     0.004        55.875
    drug2|     -1.71156      .4943639    -3.462     0.001       .2916667
    drug3|    -2.956384      .6557432    -4.508     0.000       .2916667
---------+--------------------------------------------------------------

     ( 1)  drug2 - drug3 = 0.0, F(1, 45) = 3.31, Prob > F = 0.0757
```

17 个病人仍然存活，他们的生存时间自然要长。死亡的病人中，编号为 2 的药好于 1 ，而编号 3 的好于 2 。似乎编号为 1 的药最差，使用模型$h(t) = h_0(t)exp(b_1age + b_2drug_2 + b_3drug_3)$并考虑到截尾，使用dead( ) 选项，先对编号2和3的药生成指示变量，就可以使用cox 命令了。对数似然值由-99.911448，经四次迭代为-81.652567。最后检验与第一种药是有区别的，第二、第三两种药是否有所不同。

Survive.Kit 是一组Stata 用于生存分析的程序，使用命令run Survive. Kit 装入。在Stata 3.0 以后以ADO文件格式调用。kapmeier 和gwood 用于绘制Kaplan-Meier 生存曲线和给出根据Greenwood 公式给出的置信带。loglogs 绘制log(-log(S(t))) 对log(t) 的图，其中S(t) 为由Kaplan-Meier 积矩统计量定义的生存函数。若曲线为直线，则数据服从威布尔(Weibul) 分布。survsum 和survcurv 用于显示生存分析的综合统计量和生成生存变量。logrank 计算两组或多组log-rank 统计量，用于比较两个或多个组的生存曲线。mantel和wilcoxon计算两组时的Mantel-Haenszel 检验及Wilcoxon-Gehan 检验。另外，coxhaz 、coxbase和coxvar 给出cox 回归分析后基线风险函数、生存曲线和基线生存变量。这些命令通用的格式是：

<命令> 时间变量死亡指示变量[, by(分组变量)]

死亡指示变量取值为 1 时表示死亡，程序及结果如下：

```
. kapmeier studytim died
. kapmeier studytim died, by(drug)
. gwood studytim died
. gwood studytim died, by(drug)
. survsum studytim died
. survsum studytim died, by(drug)
. loglogs studytim died
. loglogs studytim died, by(drug)
. survcurv studytim died
. gen loglogs = log(-log(_surv))
. gen logt = log(studytim)
. regress loglogs logt
. logrank studytim died, by(drug)
```

```
  Source |       SS       df       MS              Number of obs =      48
---------+------------------------------           F(  1,    46) = 4439.69
   Model |  41.886338     1  41.886338             Prob > F      =  0.0000
Residual |  .433987958    46  .009434521           R-square      =  0.9897
---------+------------------------------           Adj R-square  =  0.9895
   Total |  42.320326     47  .900432467           Root MSE      =  .09713

Variable |  Coefficient    Std. Error      t    Prob > |t|        Mean
---------+--------------------------------------------------------------
 loglogs |                                                     -.7396131
---------+--------------------------------------------------------------
    logt |     1.075852      .0161464    66.631     0.000      2.449985
   _cons |    -3.375435      .0419694   -80.426     0.000             1
---------+--------------------------------------------------------------
```

```
    Group            Events           Predicted
    ------------------------------------------
    1                19               7.2459559
    2                6                8.1984653
    3                6                15.555579
    Chi2( 2 ) =  25.526241 , P =  2.864e-06
```

结果表明三组的确不同，P-值小于0.1%，注意到有19+6+6=31 个死亡以生(标记为) 预测值为31。对于对照组，(第一组1)，观察到19 个死亡，但若所有病人具有相同的生存率。只会有7.25 个亡发生。

§8.2.6 Stat.Kit

是一组程序，用于一系列统计检验。通过命令run Stat.Kit 调入。它提供了十五个新命令。

dbeta	计算影响统计量DF-Betas
genrank	产生变量的秩次，并考虑到相同秩次
genstd	产生标化变量(均值 0 ，方差1)
glogit	分组logit
gprobit	分组probit
ksmirnov	Kolmogorov-Smirnov 分布相等检验
kwallis	Kruskal-Wallis 单因素方差分析
means	算术平均、几何平均和调和平均
ranksum	Wilcoxon 秩和(Mann-Whitney 两样本) 统计量
regdw	回归并产出Durbin-Watson 统计量
signrank	Wilcoxon 配对符号秩次检验
signtest	配对检验中位数检验
spearman	Spearman 秩和相关系数
teststd	检验方差是否相同或者是否为某个常数
ttest	保种类型的t-检验

现要检验一种新的燃料添加剂的有效性，用12辆汽车实验，进行有无添加剂的里程的比较，结果如下：

Without Treatment	With Treatment	Without Treatment	With Treatment
20	24	18	17
23	25	24	28
21	21	20	24
25	22	24	27
18	23	23	21
17	18	19	23

```
. summarize
Variable |   Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------
    mpg0 |    12          21   2.730301         17         25
    mpg1 |    12       22.75   3.250874         17         28

. means
         |      Arithmetic        Geometric          Harmonic
         |      Mean   Obs        Mean   Obs        Mean   Obs
---------+--------------------------------------------------
mpg0     |        21    12    20.83629    12    20.67272    12
```

```
mpg1     |     22.75    12    22.52909    12    22.30015    12

. ttest mpg0=20
. ttest mpg0=mpg1, paired
. ttest mpg0=mpg1
. signtest mpg0=mpg1
. signrank mpg0=mpg1
```

以上命令进行均值是否20的检验，配对检验和成组t-检验。使用命令'ttest mpg0=mpg1, unequal' 不做方差齐性的假设。其次用signtest 命令检验中位数是否相同，运用Wilcoxon 配对符号秩次检验检验分布是否相同。

接下来使用人口普查数据census.dta，检验结婚率，首先生成变量，用结婚数除以18岁以上的人数，然后考察结婚率与年龄的相关，可以使用原有的pearson 相关命令或spearman 命令。

```
. use census, clear
. generate mrgrate = marriage/pop18p
. summarize mrgrate
. correlate mrgrate medage
. spearman mrgrate medage
. summarize mrgrate, detail
```

§8.3 高分辨统计制图

§8.3.1 图形命令

Stata 1.0 的高分辨统计制图功能是靠Stata/GRAPH 完成的，Stata 2.0 的制图功能被集成在命令graph，其语法为：

[by varlist:] graph 变量列表[=权] [in 范围] [if 条件表达式] [, 选项] 或: graph using 文件名[文件名[..]] [, 选项]

使用graph 命令可以绘制八种图形，即histogram (直方图)、matrix(二维散点图距阵，最多可指30 个变量)、box(多达六个变量的Box-whisker 图)、bar(直条图)、twoway(二维散点图)、oneway(一维条形图，可与box 结合使用) 、star(星形图，最多有16 个变量) 和pie(圆图)，图形可以重显和拼联。

graph using 文件名[文件名[..]] [, 选择项]

选项t|b|l|r1|2title("标题") margin(#) saving(文件名[,replace])

用例：

```
. graph x, saving(hist)
. graph y, saving(hist2)
. graph y x, saving(twoway)
. graph y x, by(region) saving(many)
. graph using hist hist2 twoway many
```

注意同时用using 和saving() 同样的文件名有时会破坏原有的文件。

共用的选择如：

1. saving(文件名[,replace]) 图形文件名，replace 指示覆盖原有文件。
2. by(变量名) 指示分组产生图形，分组变量应按升序排列。
3. total 结合by() 指示产生整个数据的图形。
4. bsize(#) 指示分组变量标号的字体大小。
5. Rescale 仅结合by() 使用，使每个分组量用不同的尺度。
6. title("标题") 在图形的下端加标题。多数不包含特殊字符的情况下，引号可以省略。在图形的上下左右分别可以有两个标题，简记t1()、b1() 等。
7. x|y|r|t<n>title("标题") 使用set textsize # 来控制标题文字的大小。
8. gap(#) 显示数轴时gap() 设定左标题和图轴数值的大小，默认为8。
9. x|y|r|tlable([#,...,#]) 和x|y|r|ttick([#,..,#]) 与标题相仿，用于指示指示图轴标号和图的标度。x|y|r|tline([#,.., #]) 指示在图中加上纵横分隔线。x|y|rscale(#,#) 指示用新的较宽的尺度绘图。
10. noaxis 指示不显示图轴。
11. [no]border 用于控制图的边框。
12. log 指示直方图用对数尺度。在两维图中用rlog、ylog、xlog 来指示。
13. pen(#..#) 号码由一到九，Stata 用一号笔标记图的标号。
14. shading(#..#) 阴影深度，号码越大越深，用于直方图、直条图和圆图。
15. symbol(c..c) 指示twoway 和matrix 图中的符号，可简写为s()。有O 大圆、o 小圆、S 大方块、d 小菱形、T 大三角、p 小加号[变量名] 把变量名用做标记、. 点、[_n] 使用记录号、i 隐含。
16. trim(#) 当使用变量名做符号时，图形中最多存放的字符数，默认为8 。名称"California", 因此graph y x, s(变量名) trim(2) 中符号变成为Ca。
17. psize(#) 指示[变量名] 符号的大小，默认为100。
18. connect(c..c) 指示twoway 和matrix 中点间的连接方法，可简写为c()。. 表示不连，l 在点间画直线，m 连接median bands，s 表示用三次样条连接。
19. bands(#) 指示x-轴上bands 的数目，此时计算x 与y 的中位数。
20. density(#) 指示在样条间的带子数，默认为5。
21. bin(#) 指示区间数目，默认为5。
22. freq 指示纵轴用频数单位代替分数单位。

23. normal[(#,#)] 指示在直方图上方画一个正态密度曲线。

24. density(#) 仅与normal 默认为100。

25. jitter(#) 增加一个随机噪声。# 表示噪声占图形区的比例，最大为5。

26. rescale 指示每个y-变量分别规格化。

27. rbox 指示图上显示一个box 图。

28. y|x|rreverse 指示尺度是从高到底。

29. jitter(#) 指示# 个点在科轴上显示出来。

30. half 指示仅用下三角阵。

31. jitter(#) 与twoway 相仿。

32. [no]alt 指示在箱式图下的标号是否要交错。

33. vwidth 产生不同宽度的箱式图，图的宽度与观察数成比例。

34. root 指示vwidth 和root 选项后，图的宽度与观察数目方根成正比。

35. [no]alt 与box 相仿。

36. means 指示用均值而非用变量的和规格化。

37. stack 指示图形是叠式的而非毗邻的。

38. label(变量名) 指示对每个记录用给定变量名的内容做标号。

39. select(#,#) 用于放大数据中某些星形图。

21-24用于直方图，25-28用于二维图，29用于一维图，30-31用于矩阵图，32 -34用于箱式图，35-37用于条形图，38和39用于星形图。

用例：

graph y x =pop，绘x 与y 的图，以圆的大小指示pop 的大小。

graph y x =pop, psize(150)，与上同，但所有符号增大50%。

graph y x =pop, s([name])，使用变量name 的内容作标记，字体大小与pop 成比例。

graph y x =pop, symbol(.) jitter(4) 随机地指示圆点表示pop 的大小。

graph y hat x =pop, s(Oi) c(.l) sort 绘制y hat 关于x 的图，用直线连接各点。hat 是由regress y x =pop 回归并由predict hat 语句而来的预测值。

使用汽车数据的例子，用Stata 绘制直方图、双向图、矩阵图、箱式图、星形图、直条图和圆图。

```
. graph mpg, histogram normal saving(hist)
. graph price mpg, twoway saving(twoway)
. graph price mpg weight length, matrix saving(matrix)
. graph weight, oneway by(foreign) total saving(onwway)
. graph weight, box by(foreign) total saving(box)
```

```
. graph price mpg displ wwight in 1/9, star saving(star)
. graph rep1-rep5, bar saving(bar)
. graph rep1-rep5, pie saving(pie)
. graph using hist twoway matrix oneway box  star  bar  pie  saving
(combined) title(The Combined Graphic Analysis)
```

§8.3.2 图形打印

Stata 2.0 使用程序gphdot 与gphpen 进行图形硬拷贝。它们是独立的程序，虽然可以用在Stata 下使用SHELL 来做，但最好在DOS 系统下来做。

gphdot 支持象素输出(pixel-oriented) 的设备如点阵打印机和HP laser 打印机。而gphpen 支持向量输出(vector-oriented) 设备如使用绘图笔的绘图仪和PostScript 打印机。

语法：gphdot|gphpen 文件名[/option /option ...]

文件名即图形(.gph) 文件，选项有：

/Lp|l (仅对gphdot) 指示portrait/landscape 状态。

/Dfilename 指示打印机描述文件，默认值是default.gdi/default.pen.

/Odevice: 输出设备。gphdot 通常送往并行口，如PRN:。/Ocom1 将送往

/Ofilename COM1:。/Omyfile.bin 输出至myfile.bin (可以用DOS 的COPY/B 命令送往打印机，因为文件是二进制格式)。gphpen 一般送往COM1:，但/Oprn: 则是送往PRN: /Omyfile.asc 是送往myfile.asc，除非是.pic 文件，gphpen 的输出总是ASCII 类型。用DOS 的COPY 或PRINT 命令可送往打印机。

/- 图形打印后不换页

/+ 开始打印时先换页

/C# 指示拷贝份数，不加指示时为1

/N 不打印Stata 标志

/R# 改变图形的大小，默认为/R100

/RX# 图形水平方向大小调整

/RY# 图形纵向调整

/S# 指示所有Stata 符号的大小，默认是/S100

/SO|S|T|o|d|p|.# 指示特定符号的大小，如/SO50

/I文件名对PostScript 使用ps.plf，对.pic 文件用pic.plf。

/X (仅对gphpen) 使用绘图笔放慢速度以防墨溅到透明薄膜上。

/P#..# 改动默认绘图笔号/P123456789。P122 指示3 号笔同2号笔。

/T#..# 指示笔的厚度。默认为1。使用300-dpi 的设备时，指示为4可使图形好看一些。/T144 将使一号笔为1, 2 号笔与3 号笔为4。

打印上面的图形，在DOS 系统下，运行命令：

C>gphdot combined /r52 /Dhplphr

图形文件combined 将在HP laserJet+ 上印出。

§8.3.3 记录文件输出

Stata 的运行记录文件虽然可以在DOS 下利用PRINT 命令印出，但使用系统提供的格式化程序FSL 并结合gdf 和profile.fsl 可以把Stata 的关键字醒目地显示出来。与其它软件相比，这一功能使人耳目一新。

运行方法：Dos: fsl i文件名[.log] [/选项]
Unix: fsl [-选项] i文件名[.log]

选项是：

l设计名	格式化运行记录的设计名字
lfx	Epson FX, LX, RX
llq	Epson LQ, SQ
lmx	Epson MX
lhp	HP LaserJet
lhpfonts	HP LaserJet with prestige elite fonts
libmgr	IBM Graphics printer
lhispeed	IBM Mainframe
lnecp67	NEC P6 or P7
lnec8023	NEC 8023A-C
lokidata	Okidata 84, 92, 93, 182, 183
ltoshiba	Toshiba 3-in-1
lelite	Other, elite font or 90+ characters across
ldefault	Other, less than 90 characters across (default)

p文件名	包含设计的文件名，默认为pprofile.fsl。
o文件名	输出文件名，默认为o文件名.prn
h[”]标题[”]	页标题
u[”]用户名[”]	用户的名字
a[”]帐号[”]	帐号/代码
sc|p	压缩输出类别
q	迅速执行

§8.3.4 graph.kit 与qc.kit

在stata 2.0中graph.kit与qc.kit定义一组绘图和质量控制图的命令，用run 命令引用，使用命令help graph.kit或help qc.kit 将获得它们用法的说明。在后期生动本中已经被ado文件取代。

graph.kit 的命令有:

grebar	error-bar chart
leverage	偏回归杠杆图
quantile	分位点图
qnorm	分位点—正态图
qqplot	分位点—分位点图
symplot	对称图

若没有图形显示设备，则只能在图形打印设备上输出这些图形，使用Sunview 下的Sun时，预先设置’window define tut’效果会好一些。使用命令query 可以得到计算机的配置情况。命令是：

. use auto, clear

. keep make price mpg weight displ foreign

. hilite mpg displ, hilite(foreign) ylabel xlabel

画一个两维散点图，并且使部分数据醒目。

可以使用regress mpg displ weight foreign 或regress mpg foreign weight displ 命令进行一般回归分析，右边变量的次序与分析无关，对应于车重的偏回归杠杆图则不然，必需使weight成为方程右边第一个变量。相应于指示变量foreign,可以使用命令

. leverage mpg foreign displ weight

有时我们对一个变量的分布是否对称感兴趣，可以使用对称图示法。

汽车数据中最高的车价为$15,906，最低的为$3,291，现在将它俩与数据中典型的车型比较，若分布对称，价格的差别应是雷同的。车价的中位数为$5,006. 50，因此最贵的车多花$10,899.50 而最便宜的车少花$1,715.50，由此看它们是对称的吗？对其它的数据重复此做法，就是对称图的思想，现只消用命令。

. symplot price

qc.Kit 的命令是：cchart，pchart，plotebar，shewhart。现用Acheson J . Duncan 著Quality Control and Industrial Statistics 中的数据，由每天流出的50个railway frames 样本所预计的railway frames数目，这些数据是五月份头28天的数据。

```
. describe
. summarize rejects, detail
. pchart rejects day ssize, ylabel xlabel
. pchart rejects day ssize, ylabel xlabel
. pchart rejects day ssize, xlabel ylabel stabilized title(May Production)
```

为演示命令cchart，使用新的数据，它记录了每组 5 台，共 2 5 组收音机的发现问题的数目。

```
. cchart defects sample, ylabel xlabel title(c-Chart for Radio Subassemblies)
. list
         date       mean        std
  1.        8   192.2194   3.937047
  2.        9   192.6444   2.833564
  3.       10   192.3667   4.578077
  4.       13   194.7625   3.250883
  5.       14   192.6889   2.889829
  6.       15   195.0182   1.730357
  7.       16   193.4028    2.61576

. plotebar mean std date, yline(195) title(Weight Variation) ylabel xlabel
```

最后是cusum 图，用graph 命令就可以了。

. graph sum top bot unit, yline(0) ylabel xlabel c(.ll) s(oii)

第九章　Splus

§9.1　简介

Splus由MathSoft统计科学部开发，供统计、应用数学和科学研究者使用的通用工具包。其前身是AT&T的Bell实验室Becker, Chambers和Willks研制的S 语言。1988年S系统被彻底改写从而称为"New S"，其3.0、3.1、3.2版分别于1991、1992 和1994年引入。其设计目的是向用户提供：动态、交互和高品质的图形，探索性数据分析方法，统计方法，进行数学计算。

Splus引导用户进行探索性、数据驱动和面向图形的分析。它允许用户编程和与其它语言的接口对进行功能扩展。Splus也有专用工具箱，如S+BOX用于工业设计。

§9.2　操作使用

运行环境：为运行图形用户接口(GUI)的工作站(如openLook 、motif) 和Microsoft Windows，或X-终端。在Splus中可以用命令?Devices列出软件所支持的设备名。典型的环境下，用户可以录入Splus表达式、阅读帮助文件、显示图形并与操作系统交互。Splus也可以在非图形终端运行。

§9.2.1　开始与结束

以Microsoft Windows系统为例，运行时需要home和shome两个环境变量。可以在系统配置文件config.sys或自动批处理文件autoexec.bat中使用命令：

```
set home=c:\splus
set shome=c:\splus
```

随后在Windows内就可以启用SPLUS.EXE了。

又以UNIX系统为例，在系统提示后键入：Splus

稍微停顿后，出现提示：>

使用命令q()退出S-Plus。

在系统下使用命令：setenv S_CLEDITOR emacs，则使用：Splus - e 进入Splus时将启用与emacs相容的命令行编辑功能：

^P(Previous，上次命令)、^N(Next，下一次命令)、^ K(Kill，删行)、^A(Begin，开始)、^E(End，行末)、^D(Delete，删字符)、^X(Insert，插入)等。命令history()用于显示曾经键入的命令，而again("dta",ed=T)则使包含"dta"的命令再次显示从而进行编辑。

S-Plus也可以运行命令文件：Splus BATCH <命令文件名><输出文件名>

在交互方式下也可用source()命令调用外部命令文件，类似SAS的include。

Splus具有解释语言的特征，系统执行读入后的每个命令；Splus又有功能性语言的特征，录入的表达式是系统功能调入，数据是调用的参量，同时系统又返回数据；最重要的是Splus是一个面向目标的程序语言，一个通用的功能可以由多种类型的目标调用。目标是Splus的数据、结果及函数，它们以系统文件的形式贮存，用命令objects()得到它们的列表，用rm()命令进行删除。

如上所述，用户用表达式与Splus进行交互，表达式多种多样，但在交互式环境下最主要的是命名和功能调用。打入Splus函数名就会得到相应的定义。一个功能调用通常是函数名及其参数(一般放在括号内)。所有的Splus表达式返回一个值，通常这个值会打印出来。

Splus对命令字符大小写是敏感的，因此Age与age是不同的。当表达式做为命令键入时，则其值被计算、打印并被舍弃(存于隐含变量.Last.value)。赋值语句(用<-引导)能够计算表达式的值却不自动打印，如：

```
> 12+3
[1] 15
>x<-c(1,2,3,4,5,6,7,8,9)
>m<-mean(x);v<-var(x)
>m/sqrt(v)
```

多个命令用分号分隔，若命令在一行未有打完，则系统以"+"提示。

Splus中的线性统计模型采用通用的Wilkinson-Rogers记号，操作符有+,-,*,/，%和:，:表示交互项。. 是Splus有效命名的一部分，如car.weight；同时它又可以表示公式默认的左端或右端，如update(model,.~.-Age) 。lm(skip ~.^2, data= solder. balance) 表示使用solder. balance中所有变量的主效应和二阶交互。

sink("文件名")可以把运行过程存贮于文件，sink()将关闭文件。

!用于引导系统外壳命令(与Stata相同)，如:!csh进入Unix系统外壳。

§9.2.2 取得帮助

>help(mean) 或?mean

列出具体命令的语法，在UNIX系统下若事先使用setenv EDITOR < 编辑软件名称> 则可在使用v命令后进入相应的编辑，存为文本文件。

在UNIX Windows状态下，用help.start()和help.off()进入和退出帮助系统。前者可以带有自己的参量，如：help.start(gui="openlook")。在Windows下结合Winhelp的编辑剪贴功能可以将帮助内容以文件形式保存下来。

§9.2.3 数据

数据集存放在.Data目录(Unix系统)或__DATA目录(Microsoft Windows)，它们是永久性的，使用ls()命令可以看到。每个数据集可以经attach(目标数据集) 与detach()引用，数据目标可以经rm()命令删除。数据类型：有向量(vector)、数组(array)、矩阵(matrix)、列表(list)、数据框(data.frame)。

> x<-c(1:9,10) 生成一个向量，其中1:9表示1 到9 之间的数字，有如Pascal语言中的枚举类型。seq(-5,5,by=0.5)则生成[-5,5]之间以0.5等分的数列，类似地，req(x,times=5)则将x重复5次。

向量运算规则与其它软件相仿，如z <- 5 * x + y表示z的每个元素是x的每个元素乘5并与y相应的元素相加的结果。z <- y > x 是将y与x 间的逻辑运算结果存放在z中。x[1:5]是指x向量的第1至第5个值，其下标表示方式与Fortran 中的字符操作类似；但是x[-(1:5)则表示把1:5的数据排除在外]。

数组可以想象成有下标的同样类型的数据的集合，设z是有900个元素的向量，>dim(z)<-c(3,2,150)则使z作为3 x 2 x 150阶的数组。又如：x<-array(1:20, dim=c(4,5))。

列表是由各个部分组成的目标的有序集合，其各个组分用$加以区分并可以拥有自己的命令，如：>name$component，各组分的命令可以用>names(name) 给出，这样做不需要打印出具体的数据。[[]]用于捡出列表的一个元素而[]用于向量的下标，两者的作用是不相同的。

数据框架是可以包含字符的矩阵，也可看成紧密排列的列表，行列均可以拥有自己的标号，其列可以做为列的的组分来处理。数据框架可以用命令data.frame生成。Splus中缺失值用NA表示，常用操作是na.action=na.omit，如用于lm()指令。

函数sort()用于对数据进行排序，其最简单的形式是用一个参量，如：sort(age)。更灵活的方法是使用sort.list产生一个索引。因此x[sort. list(x)] 与sort(x)的结果相同而x[sort.list(-x)]是降序结果。进一步有order()，它可以取任意数目的参量。

§9.2.4 读入转贮外部数据

可以通过赋值语句、scan()和read.table()来进行。

> counts <- scan()

将等待用户从键盘读入数据，直至文件结束符如UNIX的^D结束。diet<-scan(,"")则是读入字符类型数据。scan()与read.table() 中可以指示文件名，如

auto<-read.table("auto.dat")。

read.table()主要用于读取数据框架，

外部数据的转贮可以用write.table()完成，write()函数写出向量或矩阵。

利用专用程序，可以读取SAS数据集。假设在UNIX上用户name有一个SAS文件是test.ssd，则使用以下命令读至data。

data<-sas.get("/usr2/user2/name/",mem="test")

names<-sas.contents(lib=unix("echo /home/sparc6a/zhao"),mem="test")

§9.2.5 图形

Microsoft Windows下使用win.graph激活图形显示，在UNIX的motif 下使用命令motif()激活图形显示设备，graphics.off()，dev.off()关闭设备。其它的设备如：x11()、openlook()、sunview()等，通讯软件Kermit或NCSA的telnet 均可以使用tek4014()进行仿真。S-Plus 并不需要象SAS那样繁复的options语句，绘图存贮时只需导以postscript()，则自动生成PostScript格式文件。以(-3.14, 3. 14) 内cos(x)的做图为例，其命令只需要两条：

> angle <- seq(-pi, pi, len=100)

> plot (angle, cos(angle), type="1")

绘图函数大致分类如下：

单变量数据：	barplot	条图
	hist	直方图
	dotchart	点图
	pie	圆图
	stem	枝叶图

双变量数据	plot	散点图
	boxplot	盒式图
	qqnorm	单样本正态概率图
	qqplot	两样本分位点图
	plot.surv.fit	生存曲线图
	shewhart	Shewhart质量控制图
	cusum	cusum质量控制图
三维图形	contour	轮廓图
	persp	透视或mesh图
	image	影象图
多变量数据	coplot	条件图
	faces	脸谱图
	maplot	maplot
	pairs	两两散点图
	stars	星形图
	symbols	绘图符号
时序数据	tsplot	一元或多元时序图
	acf	自相关函数图
	spectrum	图谱
动态图象	brush	链接散点矩阵
	spin	可旋转三维图

绘图命令有大量的选择项，如：lty=n指示线型，pch=”c”指示画点用的符号。

points	向当前图形增加点
lines	向当前图形增加线
text	向当前图形增加文字
abline	画点斜式直线

交互式命令有identify、locator、legend。

§9.2.6 概率和统计

主要内容有：概率分布、综合统计量、统计检验、统计模型。与概率分布有关的函数有d分布名、p分布名、q分布名、r分布名。其分布有：

```
beta            binomial        Cauchy          chi-squared
exponential     F               gamma           gemetric
hypergeometric  log-normal      logistic        negative binomial
normal          Poisson         stable          Student's t
uniform         Weibull         Wilcoxon rank sum
```

ppoints()函数用于产生0-1间上的等分点。summary()给出描述统计量。下列语句画一个贝塔分布的直方图和分布图。

```
> sample.data <-rbeta(100,2,9)
> hist(sample.data, den=-1, prob=T)
> p<-ppoints(100)
```

```
> lines(qbeta(p,2,9),dbeta(qbeta(p,2,9),2,9)
> title(main="Data from Beta(2,9) distribution")
```

综合统计量用summary函数得到，根据参数的类别不同，它可以给出相应的结果。综合统计函数有：

mean	算术均值
median	中位数
var	向量的方差、矩阵的协方差阵
cor	向量或矩阵的相关
quantile	经验分位点
location.m	M-估计
mad	平均绝对偏差
scale.m	Bisquare A 方差估计
scale.tau	Huber的方差估计
robloc	M估计和Huber方差估计
cov.mve	多变量数据的稳健位置和方差估计

常用的统计检验有：

t.test	t-检验：单样本、两样本、配对、方差等或不等。
wilcoxon.test	秩和与符号秩次检验
var.test	两方差齐性检验
kruskal.test	单向设计的Kruskal-Wallis检验
friedman.test	无重复区组设计的Friedman秩和检验
cor.test	零相关检验，包括Pearson、Kendall和Spearman相关
binom.test	单个率的精确检验
prop.test	率相等的检验
chisq.test	两维列联表Pearson卡方检验
fisher.test	两维列联表Fisher精确检验
mcnemar.test	两维列联表的McNemar卡方检验
mantelhaen.test	三维列联表的Mantel-Haenszel检验

下面指令演示了两样本t-检验：

```
> x<-rnorm(10)
> y<-rnorm(5, mean=1)
> t.test(x,y)
```

实验设计：fac.design(析因设计)、oa.design(正交设计)、alias()给出实验设计混杂的结构(完全或部分)。

统计模型：Splus中的许多模型是一个统一的框架，数据是一个数据框架，待拟合的模型用一个公式表示出来。公式用符号～引导，如：yield～ Temp+ Conc，log(Mileage)～Weight+ploy(HP,2)。

crosstabs	从一系列因素生成多维列联表
aov, manova	拟合一元和多元方差分析模型
lm	线性模型
glm	拟合广义线性模型
gam	拟合广义加性模型
loess	拟合局部回归模型
tree	拟合分类或回归树模型
nls,ms	拟合参数非线性模型
factanal	因子分析
princomp	主成分分析

回归分析还有l1fit进行L1回归、rreg进行稳健回归，ltsreg 进行最小截尾均方回归等。

生存分析	
surv.fit	拟合Kaplan-Meier生存模型
coxreg	拟合Cox比例风险模型
agreg	拟合Anderson-Gill推广Cox模型
时间序列分析	
ar	一元或多元自回归模型
arima.mle	ARIMA模型
spectrum	时间序列谱估计
质量控制图	
shewchart	Shewchart(xbar、s、R、p、np、u和c型控制图)
cusum	cusum图(xbar、s、R、p、np、u和c)

§9.2.7 数学计算

Splus可以进行积分和导数、插值、逼近和最优化，如：

```
>integrate(sin,0,pi) [1:2]
>D(expression(3+x^2),"x")
>approx(spline(1:10,(1:10)^2),xout=1.5:3.5)
>polyroot(c(6,-5,1))
```

最优化的函数有：	ployroot	复杂多项式的根
	imorppt	给定区间内一元函数的根
	peaks	一系列离散点的局部最大值
	soptimize	给定区间内一元函数的极值
	ms	多元函数的局部极值
	minib	多元函数的极值，变量有上下界约束
	nls	一个或多个多元函数平方和的极小值
	nlregb	一个或多个多元函数平方和的有界约束极小值
	nnls	系数非负的最小二乘解

§9.2.8 用例：图形、经典分析、生存分析、局部回归

Splus系统提供了许多数据，用户用以直接引用，如乙醇数据ethanol。

```
>summary(ethanol)
>pairs(ethanol)
>attach(ethanol)
>loess(NOx~C*E,span=1/2,degree=2,parametric="c",drop.square="c")
>E.Intervals <-co.intervals(E, number=9,overlap=1/4)
>coplot(NOx~C|E, given=E.intervals, panel=function(x,y) panel.
+smooth(x,y,degree=1, span=1))
>m1<-lm(NOx~C+poly(E,2),data=ethanol)
>summary(m1)
>par(mfrow=c(2,2))
>plot(m1)
>plot.gam(m1,resid=T,rug=F)
>m2<-gam(NOx~C+lo(E,degree=2),data=ethanol)
>plot(m2,resid=T,rug=T)
>anova(m1,m2,test="F")
>m3<-gam(NOx~lo(C,E,span=1/4,degree=2),data=ethanol)
>anova(m2,m3,test="F")
```

fitted(), residuals(), summary(), predict(), family(), deviance(), formula()可用于获得gam目标的相应结果。

脊柱侧弯数据Kyphosis:

```
> attach(Kyphosis)
> kyph.gam1 <-gam(Kyphosis~s(Age)+s(Number)+s(Start), family=binomial)
> class(kyph.gam1)
> plot(kyph.gam1, residuals=T, rug=F)
> summary(kyph.gam1)
> kyph.gam2 <-gam(kyph.gam1, ~ . -s(Number))
> summary(kyph.gam2)
> plot(kyph.gam2, se=T)
> anova(kyph.gam1,kyph.gam2, test="Chi")
```

s()表示非参的平滑项，bs()与ns()则是B-样条和自然样条，lo()表示loess() 平滑，它们可以用df选项指示自由度。

最后的语句对两个模型进行比较，结果表明省略的项并不显著。step.gam()和predict.gam()可用于逐步模型选择和预测。

书写用户自定义函数一般的格式是：name <- function (arguments) body 。函数的参数在括号内给出，用逗号分开，使用等号可以给参数设默认值。函数体含在大括号内，函数返回的值是函数最后的计算值。函数体内的定义对于函数本身来说是局部的。以下是一个函数绘图的例子：

```
> f.plot
function(f,minx,maxx,nx=100, type="1",...)
{
```

```
        x <- seq(minx, maxx, length.out=nx)
        y <- f(x)
        plot(x, y, type = type, ...)
}
```

参数是minx与maxx指示x的取值范围，nx指示等分点数，默认各点用线是连接起来。可以用指令：

```
>f.plot(cos, -pi, pi)
>f.plot(function(x) {1+2*x+x^2},-10,10)
```

设要在UNIX系统下使用pico编辑，可以进行如下操作：

```
> !pico myeditor
function (data,file,editor="pico")
{
   if (missing(data))
      ed(editor=editor)
   else if (missing(file))
      ed(data,editor=editor)
   else ed(data,file,editor=editor)
}
>pico <-source("myeditor")
>pico (d)
>pico (d,"e")
```

首先启用系统的pico，然后进行函数定义并存于文件myeditor ，最后作为Splus目标存起来，以后就可以在Splus内不用系统外壳直接使用pico了。

if/else语句用于控制转向，for 用于迭代。

Splus主要有五种方法调用C或Fortran函数，它们是：

1. 静态调用. 产生Splus函数的用户拷贝，包括所有子程序。它启动编译程序并且产生执行文件local.Sqpe(Microsoft Windows 下为nsplus.exe)如调用gee 进行广义估计方程程序：

   $ Splus LOAD gee.c

2. 动态调用(dyn.load). 每个文件用通常的方式编译，如果多于一个文件，它们应累积为单一的可重新定位的目标文件：

   $ ld -r -d objects.o chi.o lgamma.o other.o

   在SunOS用-d而在Sun Solaris用-dn。之后在Splus内启用dyn.load("objects .o")动态调用。Unix Splus拥用COMPILE工具，Microsoft Windows用Watcom编译。

3. 共用库(dyn.load.shared). 在SGI和DEC Alpha这是唯一的方法，如：

   dyn.load.shared("./shlib.so")

   其参数必须是绝对路径。共享库用Splus SHLIB -objects.o chi.c lgamma.c other.c 产生。

4. 增强动态调用(dyn.load2). 基本上与dyn.load类似。

5. 动态链接库(dll.load). 在Microsoft Windows 3.2引入。

函数is.loaded可用以测试某个函数是否已被调入。

library() 引用用户自定义库。比较著名的如survival用于生存分析和oswald 用于长期数据分析。若在用户级安装这些库，要用lib.loc参量，如：

```
> assign(where=0, "lib.loc","/home/sphajiz/oswald")
> library()
```

第十章　Minitab

§10.1　简介

Minitab于1972年由美国Pennsylvania 州立大学开发的供教学使用的软件包，已成为主要统计软件包之一。它在DOS、MicroSoft Windows、Macintosh、VAX/VMS 以及Unix都有相应的产品。数据分析功能包括探索性数据分析、基本统计量、实验设计、回归分析、方差分析、多元分析、非参统计、时序分析、模拟、图形和质量控制。另外，它的宏功能供用户进行功能扩展。不同系统的用户界面不同，但其基本的特征，即工作表及有关命令是一致的。

§10.2　操作使用

§10.2.1　作业表

Minitab的中心是作业表，它包括C1，C2，...等列，K1，K2，...等常数，M1，M2，...等矩阵。其大小随系统而变，如DOS上的第8版允许16,714个数据元素，包括100列、100个常数和15个矩阵。Windows第9版允许100, 000 个数据元素，包括1000列、1000个常数和100个矩阵。所有版本中，π =3.141592..和 e =2.71828.. . 均是一个常数，* 是缺失值的符号。列、常数和矩阵均可以直接引用或用别名引用。

§10.2.2　命令

所有命令具有相同的结构，即命令名与相应的参量。它们独立于系统菜单。命令由命令名引导，与Systat类似，仅其前四个字母有效，命令中大小写英文字母均可以采用；参量指示相应的列、常数或矩阵。此外，许多子命令具有与命令同样的结构。在系统帮助和说明书中描述的命令有这样的格式：C、K、E、M、FILENAME和方括号。C指示列号或相应的别名；K指示只能接受常数值或存贮常数(如K9)；E 表示接受列、数字或常数；M表示矩阵；方括号表示为可选项。其它对命令的解释用小写文本给出。一般说来，可选项有相应的默认值。如：

```
TTEST [of mu=K] on data in C,...,C
ALTERNATIVE=K
```

说明对所给的列单样本t-检验。ALTERNATIVE=1, -1 分别给出上侧和下侧检验。使用子命令时，要在主命令后缀以分号(";")，在交互式运行方式下将有SUBC¿提示出现，最后一个子命令用圆点(".")结束。这样上面的命令就变成：

```
MTB>TTEST of mu= 50 on data C1, C4, C7;
SUBC>ALTERNATIVE=1.
```

即对C1, C4 和C7列上的数据进行均值为50的检验。由上所述，简捷的格式是：

```
MTB>TTES 50 C1, C4, C7;
SUBC>ALTE 1.
```

再看样本均值的语法：MEAN of the values in C [put into K]

MEAN C1 K3 将把第2列的均值放在K3。

FILENAME 用于读取或存贮作业表、数据、命令和Minitab 运行过程。如读取ASCII文件：

READ data [from 'FILENAME'] into C,...,C

§10.2.3 使用帮助

```
MTB > help commands
To get a list of the Minitab commands in one of the categories
below, type HELP COMMANDS followed by the appropriate number,
for example, HELP COMMANDS 1 for General Information.

 1 General Information          10 Tables
 2 Input and Output of Data     11 Time Series
 3 Editing and Manipulating     12 Statistical Process Control
   Data                         13 Distributions & Random Data
 4 Arithmetic                   14 Sorting
 5 Plotting Data                15 Miscellaneous
 6 Basic Statistics             16 Stored Commands and Loops
 7 Regression                   17 How Commands are Explained
 8 Analysis of Variance            in Help
 9 Nonparametrics
```

在通常的求助命令后写上子命令名，可以得到相应子命令的信息，这与VAX/VMS类似。

MTB> help COMMANDS 5

MTB> help REGRESS rmatrix

与SAS一样，高分辨图形命令要在标准命令前冠以G。通常，Minitab 将图形输出到打印机，通过MSETUP命令指示特定的输出。输出到绘图仪也可以用命令：

MTB>GOPTIONS;

SUBC>PLOTTER.

§10.2.4 录入和保存数据

READ、SET、INSERT、END、NAME及RETRIEVE用于把数据存放到作业表。NAME用于提定列、常数和矩阵的别名，END用以SET和READ命令中指示数据的结束。

数据可以用SET、READ和INSERT命令直接读入，RETRIEVE 用以读取先前生成的作业表，若未指定扩展名，系统默认为.MTW，扩展名为.MTP的文件应用PORTABLE子命令调用。用例：

```
MTB> read c1,c2,c3
DATA> 1 2 3
DATA> 2 3 1
DATA> 3 1 2
DATA> 1 1 1
DATA> 2 2 2
DATA> end
```

这里每行相当于一个记录。使用SET命令需要三次来读取：

```
MTB> SET c1
DATA> 1 2 3 1 2
DATA> end
MTB> SET c2
DATA> 2 3 1 1 2
DATA> end
MTB> SET c3
DATA> 3 1 2 1 2
DATA> end
```

若上述数据存放在test.dat中，则可以read 'test.dat' c1, c2, c3 读取。

SET 命令中的数字可进行一些简化：

(a) 连续整数：6:10 即6, 7, 8, 9, 10；4:-1 即4, 3, 2, 1, 0, -1。

(b) 使用一个增量：0:10/3 即0, 3, 6, 9；1:3/.5 即1, 1.5, 2, 2.5, 3。

(c) 重复因子：

2(1,2,4) 即1, 2, 4, 1, 2, 4；

(1,2,4)2 即1, 1, 2, 2, 4, 4。

进行插入：

```
INSERT BETWEEN ROWS 2,3 OF C1-C3
 62 105 0.4
 63 120 0.7
END
```

插入前				插入后			
C1	C2	C3	C4	C1	C2	C3	C4
61	96	0.5	14	61	96	0.5	14
65	115	0.3	12	65	115	0.3	12
67	131	0.8	13	62	105	0.4	13
64	125	0.5	17	63	120	0.7	17
				67	131	0.8	
				64	125	0.5	

浏览数据：

```
MTB > print c1-c2.
MTB > NAME C2 = 'SEX' C4 = 'HT 79' C5 = '1/TEMP'
MTB > READ C1 'SEX' 'HT 79'
MTB > TABLE C1 BY C2
```

使用write命令写入数据集：

```
MTB>  write 'cc' c1-c2.
MTB>  type cc.dat
```

或者传输格式或Lotus工作表。

```
MTB>  save 'cc';
SUBC> portable.
MTB > save 'cc';
SUBC> lotus.
```

事实上，DOS 下的Minitab 8和Windows下的Minitab 9以及Macintosh下，数据可以在作业表中直接输入。

使用PRINT 命令可以观察存贮的数据。数据的读写可以按照Fortran 格式，WRITE或SAVE命令分别用以ASCII和.MTW格式存贮数据。

运行过程的存贮使用命令OUTFILE 'FILENAME' 记录，使用NOOUTFILE 关闭。其文件的隐含扩展名是.LIS。用命令JOURNAL'FILENAME' 用于记录录入的命令和数据，隐含文件扩展名为.MTJ，记录同样可以用NOJOURNAL命令中止。

§10.2.5 编辑和管理数据

有许多命令，这里只给出最一般的命令：LET、DELETE、ERASE和COPY，如：

```
MTB>ERASE E,...,E
MTB>DELETE rows K, ..., K of columns C,...,C
MTB>LET C(K)=K
```

最后一行命令的第一个K指示行号而第二个K表示要替换的值。

作业表上列的数据可以用命令COPY、CODE、CONVERT 、STACK 、UNSTACK 、CONCATENATE完成。COPY可以进行全列或部分列拷贝，CODE 允许给列中的某些范围内的值赋值，CONVERT用于数字、字符类型的互换，STACK与UNSTACK 用于对数据进行重新配置，CONCATENATE用于把许多字符数据合并到一列。

数值变量有许多转换方式，除绝对值、符号、指数函数、对数、三角反三角函数、数据处理函数和逻辑函数AND、OR、NOT外，还有大量的统计函数如：MEAN、MEDIAN、STDEV、SSQ、SORT、RANK。MTB
C11=(C1-MEAN(C1))/STDEV(C1) 结果是把C1的标化值放在C11列。

§10.2.6 统计过程

基础统计量

DESCRIBE	计算某列的标准描述统计量
ZINTERVAL	方差已知时均值的可信区间
ZTEST	单样本均值检验
TINTERVAL	单样本 t 一分布的可信区间
TTEST	单样本t-检验
TWOSAMPLE	两列上的两样本t-检验和可信区间
TWOT	一列上的两样本t-检验和可信区间
CORRELATION	计算样关系数及其矩阵
COVARIANCE	计算样协方差及其矩阵
CENTER	数据中心化和标准化

绘图

HISTOGRAM	产生某列上的直方图
STEM-AND-LEAF	某列数据的茎叶图
DOTPLOT	画点图
BOXPLOT	画盒式图
PLOT	y对x画图
MPLOT	在同样数轴上画几个变量
LPLOT	y对x做图, 用字母区分组别
TPLOT	伪三维图，符号表示z-值

回归

REGRESS	线性回归和多项式回归
STEPWISE	逐步回归
BREGRESS	最大R平方准则的最优子集回归
RREGRESS	稳健回归

方差分析

AOVONEWAY	单向方差分析，各组存放于不同列
ONEWAYAOV	单向方差分析，各组存放于一列，组别放于另外一列
TWOWAYAOV	双向平衡设计
ANOVA	多向、多因素平衡设计
ANCOVA	固定效应的正交设计分析
GLM	拟合一般线性模型，包括不平衡设计

多元分析

PCA	主成分分析
DISCRIMINANT	线性和二次判别函数
FACTOR	因子分析

非参统计

RUNS	随机游程检验
STEST	符号检验
SINTERVAL	根据符号检验计算中位数的可信区间
WTEST	单样本Wilcoxon符号秩次检验
WINTERVAL	根据Wilcoxon符号秩次检验计算中位数的可信区间
MANN-WHITNEY	两样本Mann-Whitney-Wilcoxon秩和检验和可信区间
KRUSKAL-WALLIS	k个中位数相等的Kruskal-Wallis检验
MOOD	Mood中位数检验
FRIDEMAN	随机区组的Frideman检验
WALSH	所有对子的Walsh平均
WDIFF	计算两两差值
WSLOPE	计算两两斜率

列联表

TABLE	显示列联表及有关统计量
TALLY	单向表计数和百分比
CHISQUARE	列联表卡方检验

时间序列

TSPLOT	时间序列做图
MTSPLOT	对几个时间序列做图
ACF	时间序列自相关函数
PACF	时间序列偏自相关函数
CCF	互相关函数
DIFFERENCE	时间序列差分
LAG	序列滞后
ARIMA	拟合Box-Jenkins的ARIMA模型

统计过程控制图

XBARCHART	样本均值控制图
MACHART	移动平均图
NPCHART	不相容图
RCHART	样本极差图
EWMACHART	指数加权移动平均图
CCHART	Poisson计数图
SCHART	标准差控制图
MRCHART	移动极差图
UCHART	单元Poisson计数图
ICHART	单个观察控制图
PCHART	不相容比例图

探索性数据分析

STEM-AND-LEAF	某列数据的茎叶图
BOXPLOT	Box-and-Whisker图
GBOXPLOT	高分辨Box-and-Whisker图
LVALS	字母数值图
CPLOT	凝聚散点图
RLINE	拟合稳健回归线
RSMOOTH	平滑数据
CTABLE	格式化两维列联表
MPOLISH	两维设计的中位数平滑化
ROOTORAM	悬浮根(suspended rootogram)图

概率分布和随机数

RANDOM	产生随机数
PDF	离散分布概率计算和连续分布密度函数
CDF	累积分布函数
INVCDF	累积分布逆函数
SAMPLE	有放回或不放回取样

实验设计

FFDESIGN	两水平的全部或部分析因设计
PBDESIGN	Blackett-Burman设计
FFACTORIAL	正交和非正交两水平设计

质量控制宏

ANOM	均值的单向或双向分析
CAPA	过程能力直方图和统计量
CUSUM	累积和控制图
PARETO	Pareto控制图
RSDESIGN	2-6因素中心复合设计和3-6因素Box-Behnken设计
RSMODEL	对RSDESIGN所产生的设计拟合二次模型

与SAS一样，Minitab的宏是一个命令文件。其好处是可以避免命令的重复，特别适合模拟、对某列的特殊操作以及功能的扩展。所有版本的Minitab 可以用EXECUTE命令执行宏定义，宏命令文件一般用.MTB做扩展名。此外，Minitab还提供了全局宏，一般用.MAC做为扩展名。

MTB>EXECUTE 'FILENAME' [K 次]

【例10.1】太阳黑子数据分析：使用AR(2)模型。

```
MTB>  Dir
MTB>  System
MTB>  Retrieve 'c:\sunspot'.
MTB>  Gplot c2 c1;
SUBC>  Symbol 'x';
SUBC>  Line 0 1 c2 c1.
MTB> ACF c2.
MTB> PACF c2.
MTB> Differences 15 c2 c3.
MTB> Arima 2 0 0 c2.
```

第十一章　Genstat

§11.1　Genstat 简介

Genstat 由Rothamsted Experimental Station六十年代开发。用于VAX/VMS 系统、许多大型机、小型机、Unix 工作站和IBM 兼容微机。与众多的统计软件包相比，它为通用的目的而设计。虽然使用它的指令可进行通常的统计分析，但它不仅仅是从有限种预先写好的程序的集合中挑选，而是一个灵活的命令语言，需要产出标准方法没有提供的结果、需要进行新方法的研究，则Genstat 是一个优良的选择。

Genstat 的统计功能分为四个部分，回归分析、实验设计分析、多元和聚类分析以及时间序列分析。回归分析包括一元和多元线性回归、非线性回归、广义线性模型如生物检测(bioasay) 的probit 和logit 分析以及列联表的对数线性模型。方差分析可以包括几乎所有的标准实验设计进行，如完全随机化正交设计、随机区组设计、裂区设计、拉丁方和希腊拉丁方(Graeco-Latin)、重复测量数据分析、平衡不完全区组设计以及平衡混杂因素的其它设计如Youden 方，可对这些设计进行协变量分析并处理缺失值。多元分析包括主成分分析、典型变量分析、因子旋转、主坐标分析以及Procrustes 旋转。使用矩阵、向量的操作可以进行对应分析和典型相关分析等。亦有多种系统聚类方法，可以产生最小支撑树。时间序列分析包括Box-Jenkins 的ARIMA 及其季节模型。序列之间的关系可以用传递函数模型进行研究，进行模型的选择与检查、估计与预测。计算富里叶转换，进行谱分析等。

没有任何计算机程序能完成用户所需要的一切，Genstat 可以使用Fortran 77 来增强自身的功能，进行与其它软件的数据交换等。OWN 语句对于所有版本的Genstat 都有效、PASS 指示使用户不必把它们直接连进来，有时软件并不包括此功能。

§11.2　Genstat 语言

Genstat 功能和语言特色类似GLIM(见第11章)，如下例：

```
VARIATE [NVALUES=10] X
READ X
24.3 25.6 57.3 43.8 45.3
46.5 47.9 97.0 77.5 64.3 :
CALCULATE Xbar=MEAN(X)
PRINT STRUCTURE=Xbar; DECIMALS=2
PRINT STRUCTURE=X; DECIMALS=1
STOP
```

第一句指示X 长度为10，即包含10 个值，接下去三条语句给X 赋值，在数据的末尾用冒号(:) 结束。接下去算得X 的均值，最后的PRINT 语句把数据打印出来。

与Fortran 等高级语言一样，Genstat 的说明(declaration) 用于指示一种数据结构的类型和标识，这种说明可以是显式的或是隐式的，上例中头三句是显式的，而CALCULATE 语句中计算一个均值是隐式说明了一种常数类型数据结构，隐式的定义常称为默认的或缺损的(default definitions)。Genstat 的语句有相同的语法，首先是指令的名称，其次是语句的选项(options)、参数(parameters)。选项放在方括号内，参数放在选项方括号的外面，参数项

之间用分号(;) 隔开。Genstat 的续行符号是反斜杠(\)，注释是用一对双引号括起来的字串。Genstat的字符集是ASCII 码的子集，与多数软件是相同的，需要特别指出的是一些特殊的用法，他们有：

& 表示重复最近一次的命令或过程及其相应的选项设置；* 表示缺失值；$ 表示某些结构中数据的子集，此时多后随一个由方括号括起来的数据表；！引入一个未命名数据结构，有的计算机上使用符号(|)，# 用于引入一个数据结构并嵌入到当前的程序中，在有的计算机上使用英镑符号。

Genstat 的字符组成了六种项(items)，它们是数、字串、标识(identifier) 系统专用字(system words)、缺失值、操作符。它的算术以及关系和逻辑运算符与Fortran 也类似，特别指出的是以下操作符：

== 与.EQ. 等价，/= 与.NE. 等价；字串相等使用.EQS.，不等使用.NES.；标识相等使用.IS.，不等使用.ISNT.；包含使用.IN.，不包含使用.NI.；.EOR. 表示异或(exclusive disjunction)；矩阵相乘使用*+。另外，Genstat 有一些公式操作符：加(+)、点积(.)、交叉积(*)、嵌套积(/)、删除(-)、交叉删除(-*)、嵌套删除(-/)、项目间连接(//)。

列表(链表, lists) 是一些项的集合，有数字列表、字符串列表及标识列表，列表中各项通常用逗号分开，可以省略符(...) 助记，如：

-2,-1.5...0.4 相当于-2, -1.5, -1, -0.5, 0，1 到10 的数表可记为1.. .10，2(A,B,C) 相当于A,A,B,B,C,C，而('a','b')2 相当于'a','b','a','b'。

Genstat 的宏也是一段Genstat 程序，定义之后，可以使用一对替换符号进行调用。设文件ALG.DAT 存放一段程序，用于迭代计算一个数的平方根，可用以下的程序调用。

```
SET [IPRINT=statements,macros]
SCALAR IDENTIFIER=X,Root; VALUE=48;
TEXT [NVALUES=3] Estsqrt
OPEN NAME='ALG.DAT'; CHANNEL=2
READ [CHANNEL=2] STRUCTURE=Estsqrt
##Estsqrt
##Estsqrt
##Estsqrt
PRINT [IPRINT=*] '3 Iterations to calculate sqrt(48) as',Root
STOP
```

文件ALG.DAT 的内容为：

```
'CALCULATE Previous=Root'
'        & Root=(X/Previous+Previous)/2'
'PRINT STRUCTURE=Root,Previous; DECIMALS=4':
```

Genstat 也有交互和批处理两种类型的工作方式，运行结束时使用STOP 命令返回操作系统。

Genstat 的工作控制：包括循环、选择、过程的形成。

Genstat 的程序是标准指令或过程组成的一系列语句。程序可用于检查不同的数据集，同时进行几种分析等。指令JOB/ENDJOB 把Genstat 程序分为功能上相互独立的工作。许多

环境在不同的工作之间保持不变，除非使用了JOB 语句。在一项工作结束后，所有过程和数据结构的值和标识均被删除。

用例：

```
1 JOB 'Example of ENDJOB messages'
2 PRINT 'This job just prints this message.'
3 ENDJOB
```

STOP 指令用于结束一个Genstat 程序。可见Genstat 的JOB/ENDJOB/STOP 与GLIM 软件中的SUBFILE/STOP 相应。

下面的循环用于求平方根：

```
FOR [NTIMES=3]
    CALCULATE Previous=Root
  & Root=(X/Previous + Previous)/
    PRINT root,Previous; DECIMALS=4
ENDFOR
```

其中的NTIMES 是循环的次数，除此以外，还可用COMPILE 指示语句为编译状态执行。语句允许更为复杂的循环指示变量，如：

```
FOR Ind=x1,x2,x3; Dir='descending','ascending'
    SORT [INDEX=Ind; DIRECTION=#Dir] x1,x2,x3
    PRINT x1,x2,x3
ENDFOR
```

相当于下列语句：

```
SORT [INDEX=x1; DIRECTION='descending'] x1,x2,x3
PRINT x1,x2,x3
SORT [INDEX=x2; DIRECTION='ascending'] x1,x2,x3
PRINT x1,x2,x3
SORT [INDEX=x3; DIRECTION='descending'] x1,x2,x3
PRINT x1,x2,x3
```

选择结构中，有块IF 语句，其格式为：IF ... ELSEIF ... ENDIF，用法与Fortran 相似。第二种格式为：CASE ... OR ... ELSE ... ENDCASE 与Foxbase+ 中的DO CASE 语句相类似。这些控制都用EXIT 指令退出，其选项为：NTIMES (控制结构的数目)、CONTROL (控制类型，for, if, case, procedure)、REPEAT (在FOR 中是否转到后续参数控制)。

指令PROCEDURE 用于指示一个Genstat 过程，指令OPTION 与PARAMETER 用于定义过程的选项和参数，两者均可以带名称选项，过程以ENDPROCEDURE 指令结束。可以使用用户自己定义的过程库，如：

```
OPEN 'graphicslib'; CHANNEL=2; FILETYPE=procedurelibrary;
```

存贮的方法是语句STORE，如：

```
STORE [CHANNEL=1;SUBFILE=Jacknife; PROCEDURE=yes] Jacknife
```

指令CATALOG 可用于显示一个库及其了文件的内容。
程序运行过程中，可以执行中断调试，用BREAK 和DEBUG 完成，如：

```
1 PROCEDURE 'polar'
2    PARAMETER 'x','y','r','theta'
3    CALCULATE R=SQRT(X*X+Y*Y)
4    CALCULATE THETA=ARCOS(X/R)
5    CALCULATE THETA=THETA+2*(3.14159-THETA)*(Y<0)
6 ENDPROCEDURE
7 SCALAR Xpos,Ypos; VALUES=3,4;
8 DEBUG
9 POLAR Xpos; Y=Ypos; R=Radius; THETA=Angle
10 ENDBREAK
11 PRINT R
12 ENDBREAK
13 PRINT THETA
14 ENDBREAK
15 CALCULATE Deg=THETA*180/3.14159
16 PRINT Deg
17 ENDDEBUG
18 PRINT Xpos,Ypos,Radius,Angle
```

下面列出Genstat 的几类函数，其中的x、y 可以是常量、变量、因素、表、矩阵、对角阵或对称阵，s 表示常量，f 表示因素，v 表示变量，t 表示表，d 表示哑元。

通用和数学函数

- ABS(x) 绝对值函数。
- ANG(p)或ANGULAR(p) 角度转换，对于$0 < p < 100, x = (180/\pi)\arcsin(\sqrt{p/100})$
- ARCCOS(x), $-1 \le x \le 1$ 反余弦函数，结果为孤度值。
- ARCSIN(x), $-1 \le x \le 1$ 反正弦函数。
- CIRCULATE(x;s) 把x 向左(s¡0)右循环移动s 个位置，s 的默认值为1。
- COS(x) 余弦函数。
- CUMULATE(x) 或CUM(x) 累积和。
- DIFFERENCE(x;s) s 阶差分。
- EXP(x) 自然指数。
- INTEGER(x) 取整函数。
- LOG(x) 自然对数。
- LOG10(x) 常用对数。
- MVREPLACE(x;y) 用y 中相应的值替换x 中的缺失值，当x 与y 均为缺失值给以警告信息。

- NEWLEVELS(f;x) 从因素f 构造一个变量，x 存放相应于各水平的值。
- REVERSE(x) 反转数据。
- ROUND(x) 四舍五入到最近的整数。
- SHIFT(x;s) 把x 的值向左移或右移s 个位置，移动将产生一些缺失值。
- SIN(x) 正弦函数。
- SORT(x;y) 按y 的值的升序对x 排序。
- SQRT(x) 平方根函数。

常量函数

- MAXIMUM(x) 或MAX(x) 最大值函数。
- MEAN(x) 均值函数。
- MEDIAN(x) 或MED(x) 中位数函数。
- MINIMUM(x) 或MIN(x) 最小值函数。
- NCOLUMNS(x) 矩阵的列的数目。
- NLEVELS(f) 因素的水平数。
- NMV(x) x 中未缺失值的数目。
- NOBSERVATIONS(x) 非缺失值的数目。
- NROWS(m) 矩阵的行数。
- NVALUES(x) 包括缺失值在内的x 的长度。
- SUM(x) 或TOTAL (x) 求和。
- VARIANCE(x) 或VAR(x) 方差函数。
  以上函数除了NMV(x) 和NVALUES(x) 以外，均不对缺失值操作。

变量函数

有VMAXIMA(p)、VMEDIANS(x)、VMINIMA(p)、VNMV(p)、VNOBSERVATIONS(p) 、VNVALUES(p)、VSUMS(p)
们与常量函数类似，只是各个函数前缀以V，只有一个指针形式的参量。

矩阵函数

- CORRMAT(x) 从对称阵x 中形成一个相关阵。
- CHOLESKI(x) 进行cholesky 分解。
- DETERMINANT(x) 或DET(x) 或D(x) 求对称阵行列式的值。
- INVERSE(x) 或INV(x) 或I(x) 对称阵求逆。
- LTPRODUCT(x;y) 即x 转置与y 的积。
- PRODUCT(x;y) x 与y 的积。
- QPRODUCT(x;y) y 关于x 的二次型，即x *+ y *+ TRANSPOSE(x)。
- RTPRODUCT(x;y) 即x 与y 转置的积。

- SOLUTION(x;y) 求齐次线性方程组的解。
- TRACE(x) 求矩阵的积。
- TRANSPOSE(x) 或T(x) 求方阵的转置。

表函数

这些函数由表形式的数据组成边缘值，它们有TMAXIMA(t)、TMEDIANS(x) 、TMEANS(t)、TMINIMA(p)、T
或TTOTAL(p)、TVARIANCES(p)。它们与常量函数类似，只是各个函数前缀以T，只有一个指针形式的参量。

哑元函数

UNSET(d) 用于检验哑元是否已定义，返回逻辑值0、1。

元素操作函数

- ELEMENTS(x;e1,e2) 指定一个元素的集合，e1 与e2 是表达式。
- EXPAND(x;s) 从x 的值形成一个0 和1 变量，s 是结果的长度。
- RESTRICTION(x) 相应于x 当前的限制构造一个值为1 的变量。

统计函数

- ANGULAR(%p) 角度转换，$\%p$ 是百分比，结果是$(180/\pi)arcsin(\sqrt{\%p/100})$。
- CED(p;s) 卡方分布变量，给定自由度为s，概率为p，返回相应的卡方值。
- CHISQ(x;s) 自由度为s 的χ^2分布下，小于x 的概率值。
- FED(p;s1;s2) 给定概率值、自由度时F-分布的变量。
- FRATIO(x;s1;s2) 或$FPROBABILITY(x;s1;s2)$ 小于x 的概率。
- LLBINOMIAL(x;n;p) 或LLB(x;n;p) 二项分布$B(n;p)$ 对数似然值。
  $\Sigma xlog(np/x)) + (n-x)log(n(1-p)/(n-x))$，样本量为n，比例为p。
- LLGAMMA(x;m;d) 或LLG(x;m;d) 伽马分布对数似然值。
  $\Sigma d(log(dx/m) - x/m) - LOGGAMMA(d)$，均值为m，index 为d。
- LLNORMAL(x;m;v) 正态分布对数似然值。
  $-0.5\Sigma LOG(v) + (x-m)(x-m)/v$，均值为m，方差为v。
- LLPOISON(x;m) 泊松分布对数似然值。
  $\Sigma xlog(m/x) + x - m, m$ 是样本大小。
- NED(p) 与概率p 相应的正态变量。
- NORMAL(x) 正态分布小于x 的概率。
- URAND(s1;s2) 产生均匀分布伪随机数，s1 是随机数的种子，s2 是长度。

下面是一个Genstat 进行奇异值分解(SVD) 的例子：它把一个矩阵表成左右两个正交阵与一个对角阵的乘积，即 m A n = m U p *+ p S p *+ p V n 。这一分解对于线性方程组的求解、求矩阵的广义逆等都很有意义。

```
1. matrix [rows=6;columns=4]A;
     values=(15,5,9,16,3,20,7,12,22,17,10,11,\
2  13,8,1,23,2,4,6,14,18,21,24,19)
3  SVD[print=left,singular,right]A
```

奇异值分解结果(Singular Value Decomposition)

奇异值矩阵=$diag$(65.30　17.75　14.29　10.82)

左右奇异值向量

$$\begin{pmatrix} .35066 & -.33717 & -.30338 & .26324 \\ .32462 & .30654 & .69925 & -.39495 \\ .45861 & .18847 & -.51086 & -.52922 \\ .37075 & -.71706 & .15091 & -.29662 \\ .20711 & -.27069 & .36641 & .39876 \\ .61629 & .41157 & -.03151 & .49765 \end{pmatrix} \quad \begin{pmatrix} .50011 & -.13783 & -.80932 & -.27549 \\ .50254 & .53368 & .40553 & -.54607 \\ .40479 & .48070 & -.09456 & .77210 \\ .57749 & -.68199 & .41425 & .17258 \end{pmatrix}$$

下面程序利用SVD 的结果得到Moore-Penrose 广义逆或伪逆：

```
4  MATRIX [ROWS=6;COLUMNS=4] Uda
5  & [ROWS=4;COLUMNS=4] Vda
6  DIAGONALMATRIX [ROWS=4] Sda
7  SVD A; LEFT=Uda; SINGULAR=Sda; RIGHT=Vda
8  CALCULATE [ZDZ=zero] Splus=Sda/Sda/Sda
9  & Aplus=Vda *+ Splus *+ TRANSPOSE(Uda)
10 & Aa,Aap=A,Aplus *+ Aplus, A *+ A, Aplus
11 PRINT A; FELDWIDTH=9; DECIMALS=3
12 & Aa; FIELDWIDTH=9; DECIMALS=3
13 & Aplus; FIELDWIDTH=9; DECIMALS=3
14 & Aap; FIELDWIDTH=9; DECIMALS=3
15 CALCULATE Asa, Asapa=A, Aplus *+ Aplus,A
16 PRINT Asa; FIELDWIDTH=9; DECIMALS=3
12 & Aspa; FIELDWIDTH=9; DECIMALS=3
```

其中的CALCULATE 语句是得到奇异值的逆，伪逆为：

$$\begin{pmatrix} 0.016 & -0.029 & 0.044 & 0.007 & -0.027 & -0.009 \\ -0.029 & 0.052 & 0.021 & 0.001 & -0.016 & -0.009 \\ 0.014 & -0.022 & -0.026 & -0.039 & 0.020 & 0.051 \\ 0.011 & 0.005 & -0.026 & 0.030 & 0.029 & -0.003 \end{pmatrix}$$

EDIT 指示提供了一系列行编辑命令。

§11.3　统计图形

Genstat 提供两种基本的绘图功能，一种是为行打印机准备的字符类型的，如直方图、散点图和线图及等值线图，分别用指令HISTOGRAM、GRAPH 和CONTOUR 来实现；另一种是为图形输出设备如图形打印机、绘图仪准备的高分辨图形。

第一类指令说明如下：

HISTOGRAM 的选项有：CHANNEL= 输出文件的通道号，默认值为当前文件；TITLE=总标题；LIMITS=分组组界；NGROUPS=在没有指定LIMITS 时，设定分组数；LABELS=各组标题；SCALE=每个星号所表示的单元数。参数有：DATA=绘直方图的数据；NOBSEVATIONS=存贮每组例数的一维表；GROUPS=存贮变量所指定的分组信息；SYMBOLS=每个直方图条的符号；DESCRIPTION=关键字脚注。

```
TEXT Title
READ [CHANNEL=2;SERIAL=yes;SETNVALUES=yes] Title,Data
HISTOGRAM [TITLE=Title] Data
```

GRAPH 的选项有：CHANNEL=功能同HISTOGRAM；TITLE=总标题；YTITLE=y 轴标题；XTITLE=x 轴标题；YLOWER=y 轴下界；YUPPER=y 轴上界；XLOWER=x 轴下界；XUPPER=x 轴上界；MULTIPLE=每个帧(FRAME) 中的图；JOIN= 连接点的次序(ascending,given)；EQUAL=边界设置(no, scale, lower, upper)；NROWS= 每个帧中的行数；NCOLUMNS=每个帧中的列数；YINTEGER=轴是否一致XINTEGER 是否一致；参数有Y=y 坐标；X=x 坐标；METHOD=每个图的类型(point, line, curve , text)；SYMBOLS=每个单元的标号；DESCRPITION=关键字脚注，如：

```
1 VARIATE [VALUES=-16,-7,9,16,7,-8,-12,-5,0,10,4,-4,-3,3,16] X
2   & [VALUES=0,-14,-12,5,0,14,0,12,0,-10,-9,5,6,-6,-1,5,16] Y
3 GRAPH Y;X
```

CONTOUR 的选项有：CHANNEL=输出文件号；INTERVAL=等值区间；TITLE= 总标题；YTITLE=y 轴标题；XTITLE=x 轴标题；YLOWER=y 轴下界；YUPPER=y 轴上界；XLOWER=x 轴下界；XUPPER=x 轴上界；YINTEGER=y 轴标号的一致；XINTEGER=x 轴标号的一致；LOWERCUTOFF=数组的最小值；UPPERCUTOFF=数组的最大值；参数有GRID=数据指针；DESCRIPTION=关键字脚注，如：

```
1 MATRIX [ROWS=5; COLUMNS=7] Xval,Yval; VALUES=!((1...7)5),!(7(1...5))
2 CALCULATE Z=(Xval-2.5)*(Xval-6)*Xval-10*(Yval-3)(Yval-3)
3 TEXT [VALUES='Z(x;y)=x*(xl-2.5)*(x-6)-10*(y-3)**2'] Top
4 TEXT [VALUES='X Values'] Botttom
5 TEXT [VALUES='Y Values'] Side
6 CONTOUR [TITLE=Top; YTITLE=Side; XTITLE=Bottom] Zval
```

第二类高分辨率图形的有四种即为DHISTOGRAM、DGRAPH、DCONTOUR 及DPIE。

高分辨率图形的绘制，与有关的设置是很有关系的。其设置有：AXES、PEN 、DEVICE、FRAME 分别用于定义图轴、笔、设备及窗口位置。

功　　能	AXIS 图轴	PEN 笔的特性	DEVICE 设备切换	FRAME 视窗位置
选项：				
EQUAL	图轴的等同 (no,scale,lower,upper)			
参数：				
NUMBER			指示设备号	
WINDOWS	窗口数目			窗口号码
YTITLE	y-轴标题			
XTITLE	x-轴标题			
YLOWER	y-轴下界			同左
YUPPER	y-轴上界			同左
XLOWER	x-轴下界			同左
XUPPER	x-轴上界			同左
YINTEGER	y-轴标题一致(yes/no)			
XINTEGER	x-轴标题一致(yes/no)			
YMARKS	y-轴标度			
XMARKS	x-轴标度			
YLABELS	y-轴标号			
XLABELS	x-轴标号			
YORIGIN	y-轴原点			
XORIGIN	x-轴原点			
STYLE	轴的类型 (none,x,y,xy,box,grid)			
NUMBER		笔的数目		
COLOUR		每笔所用的颜色		
LINESTYLE		线型		
METHOD		定点方法(point, line, monotonic,closed,open)		
SYMBOLS		点的记号		
JOIN		连点的次序(ascending,given)		
BRUSH		指示填充的区域号		

例：

```
AXIS WINDOW=3; YLOWER=10; YUPPER=0; XLOWER=0; XUPPER=10; \
     XMARK=Xval; XLABEL=!T(NORTH,EAST,SOUTH,WEST); \
     YTITLE='Y AXIS'; XTITLE='X AXIS'
FACTOR [LEVELS=4; VALUES=1...4] F1
PEN NUMBER=1,2; COLOUR=1; LINESTYLE=1,2; METHOD=line,monotonic; \
    SYMBOLS=F1,3; JOIN=given,ascending;
```

四种指令的选项与参数列如下表，选项为TITLE、WINDOW、KEYWINDOW 、SCREEN和参数PEN 和DESCRIPTION。TITLE 是横标题，WINDOW 与KEYWINDOW 取值范围为0

到8，后者为窗口准备一个指示关键字。CLEAR 取值为'clear' 或'keep'，指示绘图之前当前屏幕是否保存。PEN 指示绘图用的笔，可以通过上述专用的指令设置默认值。DESCRIPTION 指示在key 的位置显示一段文本。

功　　能	DHISTOGRAM 绘直方图	DGRAPH 散点图和线图	DCONTOUR 绘等值线图	DPIE 绘圆图
选项:				
TITLE	标题	同左	同左	同左
WINDOW	图形窗口	同左	同左	同左
KEYWINDOW	关键字窗口号	同左	同左	同左
LIMITS	组界变量			
NGROUPS	分组数			
LABELS	各组的标号			
APPEND	y/n 直方图连接			
SCREEN	c/k 清屏/保留	同左	同左	同左
INTERVAL			等值间隔	
LOWERCUTOFF			数组的下界	
UPPERCUTOFF			数组的上界	
参数:				
DATA	绘图数据			
NOBSERVATIONS	存贮每组数目的表			
GROUPS	从变量定义的因子			
PEN	每个直方图的笔号	笔号	笔号	笔号
DESCRIPTION	关键字附注	同左	同左	说明
Y		纵坐标		
X		横坐标		
YLOWER		纵条的下界		
YUPPER		纵条的上界		
XLOWER		横条的下界		
XUPPER		横条的上界		
GRID			数据指针	
SLICE				扇面大小

图形的输出：OPEN NAME='文件名'; CHANNEL=1; FILETYPE='graphics'; 将产生一个转贮文件(metafile)，最常见的是GHOST、GINO 和GKS。

下面程序产生一系列伪随机数，然后绘直方图。

```
CALCULATE Var[1...3]=URAND(1237,0,0;30)
        & Var[1...3]=10,11,12+NED(Var[1...3])*1,1.2,1.3
"Default histogram with single colour pen"
PEN 1...3; COLOUR=1
DHISTOGRAM [TITLE='Default'] Var[]; PEN=1...3
"Repeat setting the APPEND option and different brush styles"
```

```
PEN 1...3; BRUSH=4,5,9
DHISTOGRAM [TITLE='Appending & new brush style'; APPEND=YES] \
           Var[]; DESCRIPTION='First','Second','Last'
VARIATE [VALUES=6,9,12,15] Limits
AXES WINDOW=1; YUPPER=30
DHISTOGRAM [TITLE='YUPPER set & limits'; LIMITS=Limies] \
           DATA=Var[]; PEN=1,2,3; DESCRITION='First','Second','Last'
STOP
```

下面程序画一个三个扇面的的圆图：

```
PEN 1...3; COLOR=1; BRUSH=1,6,11
DPIE [TITLE='Pie Chart'] 2,4,8; PEN=1...3
```

§11.4 统计分析

§11.4.1 回归分析

MODEL 指令定义响应变量和模型类型（线性、广义线性或非线性模型）。

FIT 指令配合模型。

RDISPLAY 指令显示拟合情况。

RKEEP 指令保存结果。

TERMS 指令指示一个最完全的模型，用于后续的模型分析。

ADD、ROPS 和WITCH 指令增加或减少模型中的项。

TRY 指令显示单一变量改变对模型的影响。

STEP 指令根据均方误差的比值进行模型变量的筛选。

PREDICT 指令用于预测。

下面的程序利用Draper 与Smith (1981) 的材料，对每月用水量与温度、产量、开工日数以及雇员数四种生产指标的关系，进行多元回归分析。

```
UNIT [NVALUES=17]
OPEN 'WATER.DAT'; CHANNEL=2
READ [CHANNEL=2] Temp,Product,Opdays,Employ,Water
MODEL Water
FIT Temp, Product, Opdays, Employ
TERMS Temp, Product, Opdays, Employ
ADD [PRINT=estimates] Product, Employ, Temp
DROP [PRINT=estimates] Temp
SWITCH [PRINT=estimates,accumulated] Temp, Employ
```

```
FIT [PRINT=*] Temp, Product, Employ
STEP [PRINT=estimates,changes; INRATIO=4; OUTRATIO=4] \
     Temp, Product, Opdays, Employ
```

继续使用第三节的例子：相应的Genstat 程序是：

```
Variate [nvalues=8] dose,y,n
read dose,y,n
1.691 6  59 1.724 13 60 1.755 18 62 1.784 28 56
1.811 52 63 1.837 53 59 1.861 61 62 1.884 60 60
model [distribution=binomial]y;nbinomial=n
terms dose
fit [print=m,s,e,f] dose
rkeep vcov=v
print v
```

结果是logit(p)=-60.74+34.29 log(dose) ，还可以计算出半数有效量ED_{50}。

§11.4.2 实验设计

Genstat 主要针对平衡设计，待拟合的模型用BLOCKSTRUCTURE、COVARIATE 和TREATMENTSTRUCTURE 指示，分析则采用ANOVA 指令，利用ADISPLAY 进一步显示结果，分析结果用AKEEP保存。使用GET 指令可以得到当前的模型，并且可以使用SET 指令来改变模型，DECIMALS 参数可以指示结果的小数位数，RESTRICT 指令指示仅对部分单元进行分析，使用RANDOMIZE 指令对处理进行随机分配。

```
1  "3x2 Factorial Design (Snedecor and Cochran 1980)"
2  UNITS [NVALUES=60]
3  FACTOR [LEVELS=!T(beef,cereal,pork]; VALUES=(1...3)20] Source
4  READ Gain
5  TREATMENTSTRUCTURE Source*Amount
6  ANOVA [PRINT=aovtable] Gain
7  ADISPLAY [PRINT=information, covariates, missingvalues]
8  ADISPLAY [PRINT=means]
9  ADISPLAY [PRINT=effects]
10 ADISPLAY [PRINT=%cv]
```

其中TREATMENTSTRUCTURE 指示ANOVA 语句拟合的处理因素，Genstat 采用了点(.)操作方法。此外，还有指令ADISPLAY 显示其它的结果。在几个误差项出现时，可以使用BLOCKSTRUCTURE 定义区组因素。使用COVARIATE 指令可以进行协方差分析，并继续使用ADISPLAY 和AKEEP 指令。

§11.4.3 多元分析

指一些同时分析多个变量的统计方法。关联的数据有两类，第一类是n 个样本p 个变量的数据，第二类可能是一种对称矩阵，包含了所有样本对或变量对的关联信息。象相关表示

的是变量间的关联，这样一类分析称做R-型的，另外一种是对单元间的关联分析，即是Q-型的分析。

基于平方和及乘积的方法有主成分分析(PCP) 和典型变量分析(CVA)，它们者可以使用FACROTATE 指令进行varimx 或quartimax 旋转。

生态学中的关联有序化(ordination) 或多维尺度变换(multidimensional scaling) 在Genstat 中进行的主成分分析、过程库中的对应分析，都属于这一类。Genstat 提供了一个更为一般的方法，即主坐标分析(principal coordinates analysis)，该方法经PCO 指令完成。也能用ADDPOINTS 增加新点。下面是一个主成分分析用例：

```
UNITS [NVALUES=12]
POINTER [VALUES=Height, Length, Width, Weight] Dmat
READ [PRINT=errors] Dmat[]
LRV [PRINT=Dmat; COLUMNS=2] Latent
PCP [PRINT=loadings] Dmat; LRV=Latent
FACROT [PRINT=rotation,communities] Latent[1]
```

§11.4.4 聚类分析

Genstat 提供了系统聚类和非系统聚类两种方法，列表如下：

功　　能	HCLUSTER 系统聚类	CLUSTER 非系统聚类
选项：		
PRINT	输出类型 (dendrogram,amalgamations)	输出类型 (criterion,optimum, units,typical,initial)
METHOD	聚类准则 (singlelink,nearestneighbour, completelink,furthestneighbour, averagelink,mediansort, groupaverage)	
DATA		分析数据矩阵或指针
CRITERION		分类准则 (transfer, swop)
INTERCHANGE		组间允许的移动
START		初始分类
参数：		
SIMILARITY	对称相似矩阵	
GTHRESHOLD	分组界值	
GROUPS	存贮形成的组	
PERMUTATION	聚类图中单元的次序	
AMALGAMATION	存贮聚类过程的链表。	
NGROUPS		要分到的目标类数

用例：

```
POINTER [NVALUES=4] Y
VARIATE [NVALUES=30] Y[]
READ [SERIAL=yes] Y[]
FACTOR [LEVELS=2; NVALUES=30] Optimum[2]
     & [LEVELS=5] Optimum[5]
CLUSTER [PRINT=criterion,optimum,typical,units; DATA=Y; \
        CRITERION=predictive] NGROUPS=5,2; GROUPS=Optimum[5,2]
CLUSTER [PRINT=criterion; DATA=Y; CRITERION=predictive] NGROUPS=6,5
```

§11.4.5 时序分析

Box-Jenkins 时序分析(TSM) 包括识别、估计和检查几个部分，在Genstat 中，提供了在时域和频域上的分析方法，计算一系表征时间序列的样本统计量如自相关、富里叶变换、ARIMA 模型、周期图、传递函数模型。

CORRELATION 构造变量间的相关、变量的自相关以及变量间的互相关。

FOURIER 计算实值或复值序列的富里变换。

ESTIMATE 估计Box-Jenkins 模型。

TDISPLAY 允许对ESTIMATE 进一步显示。

TKEEP 保存ESTIMATE 的结果。

FORECAST 预测时间序列未来的值。

TRANSFERFUNCTION 指示输入、输出序列和传递函数，以进行模型估计。

FILTER 使用时序模型对时序数据进行滤波。

TSUMMARIZE 时序模型特征显示。

Box 与Jenkins (1970) 的数据进行自相关计算。

```
VARIATE [NVALUES=132] Apt
OPEN 'airline data'; CHANNEL=2
READ [CHANNEL=2] Apt
CACULATE Dlapt=DIFFERENCE(LOG(Apt))
CORRELATE [MAXLAG=50; GRAPH=autocorrelations] Dlapt
```

利用此数据建立季节ARIMA 模型，程序是：

```
OPEN 'airline.dat'; CHANNEL=2
UNITS [NVALUES=132]
READ [CHANNEL=2] Apt
VARIATE [VALUES=0,1,0,1,1,12] Ord
      & [VALUES=0,0,0.00143,0.34,0.54] Par
```

```
TSM Airpass; ORDERS=Ord; PARAMETERS=Par
ESTIMATE [MAXCYCLE=0; PRINT=model] Apt; TSM=Airpass
FORECAST [MAXLEAD=12; FORECAST=Fcst12]
```

现假设有六个新的数据，放于文件airline2.dat，可用下面的程序括进来：

```
OPEN 'airline2.dat'; CHANNEL=3
READ [CHANNEL=3; SETNVALUES=yes] New6
FORECAST [PRINT=sfe; ORIGIN=6; MAXLEAD=0; FORECAST=New6]
```

其中的ORIGIN 指示了新数据的数目，设定MAXLEAD=0 避免了计算新的预测值。FORECAST 选项指定包含新数据的变量名，此处为New6。也可以包含新的数据并且产生预测。

```
FORECAST [ORIGIN=6; MAXLEAD=6; FORECAST=New6fcst6]
```

对于抗肺炎球菌血清的例子[5]，Genstat 程序如下：

```
VARIATE [NVALUES=5] DOSE, Y,X
READ DOSE,Y,N
0.0028  35  40
0.0056  21  40
0.0112  9   40
0.0225  6   40
0.0450  1   40
CALCULATE LOGDOSE=LOG(DOSE)
MODEL [DISTRIBUTION=BINOMIAL]Y;NBINOMIAL=N
TERMS LOGDOSE
FIT [PRINT=M,S,E,F] LOGDOSE
RKEEP VCOV=V
PRINT V
```

第k个滞后的样本相关是$r_k = (1 - k/n) \times C_k/C_0$ 其中

$$C_k = \sum_{t=1}^{n-k} \{(y_t - \overline{y}) \times (y_{t+k} - \overline{y})\} n_k$$

，n_k 是求和中所含的项的数目，$\overline{y}$是通常的样本均值。AUTOCORRELATION 允许用户存贮样本自相关。TEST 参数提供自相关为零的假设检验，其定义为：

$$S = n \times \sum_{k=1}^{m} r_k^2$$

当n很大而m相对于n很小时，S服从$\chi^2(m)$分布。

Genstat从自相关计算偏自相关，第k个滞后取值为：

$$corr(y_t, y_{t-k} | y_{t-1}, y_{t-2}, ...y_{t-k+1})$$

并记作$\phi_{k,k}$，可以想象成自回归预测方程中的最后一项：

$$y_t = c + \phi_{k,1} \times y_{t-1} + ... + \phi_{k,k} \times y_{t-k} + e_{k,t}$$

其计算方法：

$$\begin{aligned}\phi_{k,k} &= (r_k - \phi_{k-1,1} \times r_{k-1} - ... - \phi_{k-1,k-1} \times r_1)/\nu_{k-1} \\ \phi_{k,j} &= \phi_{k-1,j} - \phi_{k,k} \times \phi_{k-1,k-j}, j = 1...k-1 \\ \nu_k &= \nu_{k-1}/(1-\phi_{k,k}^2)\end{aligned}$$

由$\nu_0 = 1$开始，$\nu_k = variance(e_{k,t})/variance(y_t)$。

互相关函数公式为：$r_k = (1-k/n) \times C_k/(s_x \times s_y)$ 其中

$$C_k = \sum_{t=1}^{n-k}\{(x_t - \overline{x}) \times (y_{t+k} - \overline{y})\}$$

序列x与y可以不等长。Genstat用Crosscorrelation 计算互相关并提供类似自相关的检验。

指令FOURIER进行富氏变换，如：FOURIER R; TRANSFORM=F将自相关$r_0, ..., r_n$进行变换并将结果存于F，这些值相应于角频率$\pi \times j/m$ 即周期为$2m/j, j = 0, ..., m$

$$f_j = r_0 + \sum_{k=1}^{n}\{2r_k \times cos(\pi \times j \times k)/m\}$$

一般地实序列变换式为：

$$(a_j + ib_j) = \sum_{t=0}^{N-1}\{(x_t + iy_t) \times \exp(i2\pi \times j \times t/N)\}$$

ARIMA 模型为：$\phi(B)\{\nabla^d y_t^\lambda - c\} = \theta(B)a_t$，其中$B$为后移算子，$\nabla$为差分算子。$\phi(B) = 1 - \phi_1 \times B - ... - \phi_p \times B^p$, $\theta(B) = 1 - \theta_1 \times B - ... - \theta_q \times B^q$, c是∇y_t的均值，λ是Box-Cox指数转换参数。季节型ARIMA模型为:

$$\phi(B)\Phi(B^s)\{\nabla^d \nabla_s^D y_t^\lambda - c\} = \theta(B)\Theta(B^s)a_t$$

$\Phi(B^s)$与$\Theta(B^s)$季节自回归与移动平均，∇_s^D是差分阶数D。

传递函数模型为:$y_t = \nu(B)x_t + i(B)a_t$，其中$x_t$与$y_t$ 分别为输入序列和输出序列，其第二项又可以看成噪声序列。

第三部分

专用统计分析软件包

第十二章　MicroTSP

§12.1　MicroTSP 入门

§12.1.1　简介

MicroTSP 是基于大型机TSP软件包的微机产品。TSP的开发始于五十年代，主要用于时间序列分析，现已成为计量经济学和时序分析主要的软件包。TSP也有相应的微机版本。MicroTSP 于1981 年在Apple II 计算机上引入，1982 年推出第一版，Micro TSP 6.5 版1989 年开始投放软件市场，能力更强，使用更方便。

MicroTSP 提供了计量经济学和预测工作中实用的统计技术，其应用领域包括财务分析、经济研究、销售预测、费用分析、宏观经济预测、利率预测以及科学数据分析和评价。MicroTSP 方便易学、操作简便。

统计工具：描述统计、多元回归——普通最小二乘、加权最小二乘、多项式分布滞后(PDL)、ARIMA - Box-Jenkins (包括季节项)、ARMAX (带ARMA误差的结构估计)、相关图、非线性最小二乘、两阶段最小二乘、三阶段最小二乘——线性与非线性、似乎不相关回归(Seemingly Unrelated Regression) - 线性与非线性、Probit 和Logit 模型、向量自回归、VAR 冲激响应和方差分解、季节调整。

预测工具：预测、预测方程库、系统估计和VAR 模型求解和模拟(线性和非线性)、指数平滑、随机模拟方法。

数据管理：数据编辑、访问其它标准格式文件、时间转换等。

高分辨率图形：时序数据图示、散点图、条图、圆图、直方图、双标度图示、对数尺度图标、冲激响应函数。图形的软硬输出等。

其它功能：表格和统计输出可以打印或以文件形式存贮、批处理程序和宏调用、编辑、显示和管理TSP 模型和程序文件、支持数学协处理器。

MicroTSP 大部分功能由菜单驱动或MicroTSP行命令完成，如：

DATA —用于输入、扩充和修改时间序列数据。

GENR —在已有序列的基础上，利用各类型的计算公式，生成一个新的序列。

PLOT —在屏幕或打印机上生成高分辨率的趋势图。

SCAT —生成高分辨率的散点图（相关图）。

BAR —生成高分辨的直方图。

PIE —生成高分辨率的饼图。

PRINT —在屏幕上显示序列。

LS —最小二乘法（多元回归）。

AR —带有自回归误差校正项的最小二乘法。

TSLS —两阶段最小二乘。

NLS —非线性回归。

PROBIT，LOGIT —二分类数据建模。

SYS —估计线性或非线性方程组。

VAREST，VARSTAT —估计和分析向量自回归方程组。

COVA，IDENT，CROSS，HIST —描述统计量：相关系数，协方差，自相关系数，互相关系数以及在识别时间序列模型等方面非常有用的频率直方图。

AR，MA —在方程的估计中包含自回归和移动平均。

SAR，SMA —在季节变动分析中包含自回归和移动平均。

PDL —多项式分布滞后处理。

FORCST, FIT —利用已有的回归方程计算预测值。

SOLVE —求解联立模型并计算模拟值。

CREATE —建立内存工作文件。

SAVE —存储工作文件到磁盘。

LOAD —把磁盘上的工作文件加载到内存。

READ，WRITE —读写Lotus 格式(.WKS，WK1，.WR1.等) 文件，Lotus 打印格式(PRN) 文件，DIF 格式文件和其他格式的文件。

EDIT —建立和修改模型和其他文本文件。

F2 —命令回顾。

MicroTSP 软件提供了五个样本供学习使用。

§12.1.2 运行

MicroTSP 有四种工作方式：

- 菜单方式。利用系统提供的四个菜单，通过功能键F3—F6完成操作；
- 行命令方式。每输入一个命令，系统询问该命令其它信息，引导后续操作；
- 行命令带参数输入方式。用户把操作命令及参数一次输入，提交系统完成；
- 批处理方式。用户使用MicroTSP 本身的编辑程序或其他的字处理程序建立MicroTSP 的命令文件，供MicroTSP运行。

第一种是新用户常常使用。当对系统比较熟悉后，便可以使用第二、三种方式。批处理方式主要是便于用户把常规的处理过程变成一个批处理作业，反复使用。在批处理作业运行期间，用户可以做其他事情，而不必始终坐在显示器前。

使用命令TSP 启用系统后，出现系统提示：

>

用户既可以输入一个单独的命令，也可以输入包含命令及其选择参数的命令参数表，还可以通过功能键F3-F6 进入菜单操作。如采用菜单操作，可根据屏幕下方的提示：

F1－中断F2－命令重现F3－文件F4－数据F5－统计F6－TSP 控制

F3－F6 功能键用于选择TSP 软件包的四个主要菜单。按F3 选择文件菜单，

屏幕上的菜单格式为：

文件操作

(1) 工作文件（开始会话）

(2) 数据库文件操作

(3) 显示磁盘目录 DIR

(4) 改变子目录 CD

(5) 编辑文本文件 EDIT

(6) 文件改名 REN

(7) 删除文件内容 DEL

(8) 显示文件 TYPE

(9) 读外部文件 READ

(A) 写外部文件 WRITE

F1 中断(F3-F6 选菜单）

开始时选择菜单的第一行。选择可以有两种方式，其一是输入与有关命令对应的数字或字母；其二是利用键盘右面的光标移动键将光标移动到有关命令行上，然后按回车键。现输入1，表示开始会话过程。

此时将出现工作文件操作菜单：

(1) 在内存中建立工作文件 CREATE

(2) 加载工作文件到磁盘 LOAD

(3) 存储工作文件到磁盘 SAVE

(4) 扩充工作文件样本区间 EXPAND

(5) 按序列名排序 SORT

F1 中断(F3-F6 选菜单）

因为要建立工作文件，在第一行按回车，然后屏幕上会出现数据频度提示：

数据频度

(U) 非时间数据

(A) 年度数据

(Q) 季度数据

(M) 月度数据

F1 中断－中止程序运行

在这里输入A，表示在工作文件中输入年度数据。

开始时间？//1970

结束时间？//1980

// 后是回答的内容，这时屏幕上方的状况窗口的有关数据已被更新了，而且系统停止了提问并给出命令提示符“¿”表示用户可以输入其他的命令了。

现在可以利用数据编辑功能输入数据了，按F4 选择数据管理菜单，则数据管理的菜单内容会显示在屏幕上：

数据管理

(1) 调整样本区间	SMPL
(2) 利用已有序列生成新序列	GENR
(3) 数据编辑	DATA
(4) 季节调整	SEAS
(5) 序列编组操作	GROUP
(6) 内存中序列的改名	R
(7) 内存中序列的删除	D
(8) 图形操作	
(9) 显示数据报表	SHOW
(A) 打印数据报表	PRINT

F1 中断(F3—F6 选菜单）

输入3 选择数据编辑。TSP软件包的数据编辑程序开始运行。当软件包提示用户输入序列名时，即可输入您要输入的变量名。设为REV 告诉数据编辑程序为输入序列REV 做好准备。于是屏幕上显示初始日期为1970，并且在下边显示：

obs REV

光标定位在1970 右边一个高亮度区域内，提示用户在此输入样本观察值。每当用户按一次回车键，数据编辑程序都修改一个样本观察值并做好接受另一个观察值的准备。

假定用户在REV 序列中输入了如下观察值：

1970 697
1971 814
1972 963
1973 1122
1974 1224
1975 1369
1976 1539
1977 1780
1978 2161
1979 2605
1980 2915
1981 X

X 表示退出数据编辑系统，此时出现命令提示符。

若输入数据有误，在数据编辑中可以用命令进行修改，屏幕上方有关于这些命令的解释说明。例如，按回车键后，发现刚刚输入的观察值错误，可以输入命令B 使光标返回到上一个观察点，以修改刚刚输入的观察值。也可以用命令Ni 使光标定位到由 i 表示的观察点上，例如输入命令N75 后，光标定位到序列中1975 年的观察值上，如果要编辑1980 年的观察值，可以输入命令：N80。

数据编辑操作或其他操作之后，TSP 都会在屏幕上方显示最新的状态。现在序列目录中已有了REV。假如在REV 序列输入后发现序列中1976 年的观察值错误，可以调用数据编辑功能进行修改，这次不使用菜单方式，而直接输入DATA 命令，当软件包询问序列名时，输入：REV

按回车键，屏幕上将显示REV 序列的观察值，输入命令：N76 后，光标定位在1976 年的观察值上。键入正确的观察值后按回车键，序列中1976 年的观察值被新的观察值代替。输入X 字符后结束对REV 序列的编辑。

下面再输入一个变量GNP，这次将数据编辑命令与序列名一起输入，即输入：

DATA GNP

其区间与REV完全相同，顺序输入下列数据值：

992.7,1077.6,1185.9,1326.4,1434.2,1549.2,1718.0,1918.0,2156.1,2413.9, 2627.4

最后键入X 结束。TSP 软件包允许用户利用已有序列的方程式生成新的序列。按F4 调出数据管理菜单，然后选择2，即GENR 命令，则TSP提示：

方程式为？// RATIO=REV/GNP

这时在内存中生成了RATIO 序列，并且序列名出现在状态窗口中。

GENR 命令用于转换数据，用户可以说明十分复杂的代数公式，计算结果构成由方程中等号左边的命字命名的序列，如果由该序列命名的序列已经存在，新生成的观察值将代替序列中原有的观察值，否则将建立一个新的序列。

GENR 命令也可以用于计算标量和统计函数。

现在我们已经建立了一个新序列RATIO。为了显示它请按F4 功能键调出数据管理菜单，然后选择SHOW 命令。当系统询问序列名时回答RATIO，序列中的数据显示在屏幕上的表格区内：

obs	RATIO
1970	0.702125
1971	0.755382
1972	0.812041
1973	0.845899
1974	0.853437
1975	0.883682
1976	0.895809
1977	0.928050
1978	1.002273
1979	1.079167
1980	1.109462

按任意键后返回初始屏幕，处于系统提示下。

新序列生成后，可以进行各种处理如打印输出、屏幕显示，制图，回归分析、预测或存入磁盘，等等。

现在按F4 并选择SMPL 命令，用

1971　　1980

回答系统提问的样本区间，则SHOW 等操作就不使用1970 年的观察值了。

利用TSP 软件包可以非常方便地研究滞后序列。例如在命令提示符后直接输入GENR 命令和方程：

GENR PCHR=100*(RATIO-RATIO(-1))/RATIO(-1)

这里RATIO 是假日旅馆收入REV 与国民生产总值GNP 的比率，RATIO(- 1) 是RATIO 的滞后序列，滞后期为 1 。由上面方程计算的PCHR 为RATIO 的增长百分比。输入SHOW 命令可以显示新生成的序列PCHER。

按F6 选择TSP 控制菜单，则显示出下列命令菜单：

TSP 控制命令

(1) 结束会话/退回DOS EXIT

(2) 运行TSP批处理程序 RUN

(3) 运行DOS 命令 SYSTEM

(4) 打印机状态参数设置

(5) 装配TSP运行环境 CONFIG

(6) 参数选择及对应值 OPTION

(7) 内存使用报告 FREMEM

(8) 清屏并显示序列目录信息

F1 中断(F3－F6 选菜单)

选择1，则TSP 提问：

放弃当前内存工作文件吗？(Y/N) //Y

则退出TSP 软件包，返回到DOS 命令状态。也可用行命令EXIT 退出。使用SYSTEM命令可以暂时退回到DOS系统，DOS操作结束后用EXIT返回MicroTSP。

§12.2 MicroTSP 数据分析

§12.2.1 显示时间序列图形

如果用户的计算机配有图形设备，则可以用PLOT，SCAT，BAR 和PIE 命令生成高分辨率的图形。如果用户的硬件不能生成图形，则不能使用这些命令。

按F4 调用数据管理菜单。选择第8 项功能，则显示图形命令菜单：

(1) 曲线图（趋势图） PLOT

(2) 相关图 SCAT

(3) 直方图 BAR

(4) 圆图 PIE

(5) 频度直方图 HIST(G)

(6) 加载图形文件 LGRAPH

(7) 打印图形文件 PGRAPH

F1 中断(F3－F6 选菜单)

PLOT 命令显示时间序列图。横轴是时间方向，各个数据点之间以直线相连。选择PLOT后，出现比例尺参数设定方法菜单：

(A) 自动设定－单比例尺

(M) 手工设定－单比例尺

(N) 标准化图

(D) 双比例尺－图不交叉

(X) 双比例尺－图交叉

(R) 标准差标出的残差图

(S) 参数设定选择

F1 中断－中止程序运行

输入A 自动设定比例尺。TSP 询问序列名，这时输入：

RATIO

然后会在屏幕上看到高分辨率的曲线图。在观看完图后，输入X，返回命令接受状态。

SCAT 命令生成两个序列间的相关图。这次不使用菜单驱动，直接输入命令SCAT，TSP提问

序列名表？// REV GNP

系统会给出图形选择：

散点图选择

(S) 简单散点图

(C) 邻近两点连接式

(R) 给出回归线

(B) 邻近两点连线并给出回归线

F1 中断－中止程序运行

选择S，TSP 将以GNP 为X 轴方向，REV 为Y 轴方向给出两个序列的散点相关图。在看完图后，输入X，返回系统命令接受状态。

TSP 能够进行普通最小二乘法、两阶段最小二乘法、带有自回归误差校正的最小二乘估计。在所有估计中，用于估计的样本点都由当前的SMPL 确定。

§12.2.2 统计量的计算

COVA 命令计算并打印多个序列各自的均值、最大、最小值、标准偏差及它们的协方差矩阵和相关矩阵。例如可以输入命令：

COVA CONS GNP G R GNP(-1)

如果给出W选择，COVA将在计算描述统计量之前对每个序列进行加权处理。命令输入方式为：

COVA(W=POP) CONS GNP G R GNP(-1)

若使用了M 选择，COVA只显示均值，标准差和最大值、最小值。例如：

COVA(M) CONS GNP G R GNP(-1)

IDENT命令计算一个序列的自相关系数和偏相关系数。输入IDENT命令后，软件包要求给出序列名和要计算的相关系数的数目。另一种命令输入方式是在IDENT后面的括号中给出要计算的相关系数的数目，然后给出序列名。例如输入命令：

IDENT

IDENT(12) IBMPRICE

IDENT 计算出自回归系数的标准差。还计算Box-Pierce 的Q-检验值，即自相关系数的平方和。Q-检验值可用于进行自相关系数均为零的假设检验，即假定序列为白噪声。在零假设检验下，如果事先没有对序列进行ARIMA 分析，Q- 统计量服从χ^2-分布，自由度为自相关系数的数目。如果序列是检验ARIMA 估计的残差，则自由度为自相关系数的个数减去事先估计时的自回归和移动平均的项数。欲知详细的信息，请参阅Box 和Pierce (1970) [1]。

CROSS 命令计算两个序列的相关系数。输入CROSS 命令后，软件包提示给出要计算的相关系数的数目和序列名。也可以像IDENT 命令一样，在命令名后的括号中给出要计算的相关系数的数目，然后给出序列名。例如输入命令：

CROSS

CROSS(8) TB3 ASQ

HIST命令计算一个序列的频度图。在频度图中显示数据的频率分布。它将序列的数值范围(由最小值和最大值确定的数值区间) 等分成一些小区间，然后显示落入每个小区间的

观察值数目。作为隐含值，TSP 将数值区间等分成10 个左右小区间。但是通过在HIST 后面的括号中说明小区间的长度能改变系数的隐含区间分割方法。例如：

HIST RESID

HIST(2.5) AGE

标准的频度图输出是用文本字符格式，并同时显示序列的均值、标准差、最大最小值。用户可以输入P 打印输出结果。若有图形显示器，可以输入G 显示高分辨率的频度直方图。

包括G 选择参数的HIST 命令会立即在屏幕上以图形方式显示频数图。G 选择也可以和区间规格参数一起使用。若没有指出区间规格，TSP大致划分30 个左右的小区间来显示图形。输入例子如下：

HIST(G)

HIST(G,.05) IQ

§12.2.3 回归分析

按F5 选择统计运算菜单，则具体的内容显示如下：

统计运算和模型模拟

(1) 描述统计和统计检验

(2) 单方程估计方法

(3) 估计方程处理和预测分析

(4) 系统（文件）估计	SYS
(5) 向量自回归	VAR
(6) 求解模型（文件）	SOLVE
(7) 编辑系统文件或模型文件	EDIT
(8) 指数平滑	SMOOTH

F1 中断(F3-F6 选择菜单）

选择2，则显示出单方程估计方法菜单：

单方程估计方法

(1) 可带有ARMA 项的普通最小二乘法	LS
(2) 带有异方差性处理的普通最小二乘法	LS (H)
(3) 可带有ARMA 项的两阶段最小二乘法	TSLS
(4) 非线性回归	NLS
(5) 加权最小二乘法	LS (W)
(6) 加权两阶段最小二乘法	TSLS (W)
(7) 加权非线性最小二乘法	NLS (W)
(8) 非线性回归（NLS）的初值设定	PARAM
(9) 广义logistic模型	LOGIT
(A) 二元概率分布模型	PROBIT

F1 中断(F3-F6 选择菜单)

现在选择1，即进行最小二乘估计。系统将开始下面的会话：

请输入被解释变量名？//REV

系统提示：

解释变量命表中可以包含AR，SAR，MA，SMA 和PDL 参数项，

请输入解示变量名表？//C GNP

屏幕上显示统计结果。然后TSP 提示：

要显示协方差矩阵？(P，S，ENTER)

若回答S，则会在屏幕上显示协方差矩阵。之后提问

显示残差值，实际值和拟合值吗？(P, S, ENTER)

回答S，则显示出残差图(残差图为字符方式，不需要图形设备)，如果装配了图形设备，输入G，则会得到高分辨率的带有残差、实际值和拟合值的残差图。

最后，系统提问：

重新输出计算结果？(P, S, ENTER)

按回车键，将结束回归过程，回到命令提示。

㈠普通最小二乘法

计算普通最小二乘法，可使用不进行任何误差处理的LS命令。键入LS命令及一组序列名，第一个序列是被解释变量，其余的序列都是解释变量。可以利用普通最小二乘法建立一个简单的消费函数，即输入命令：

LS CONS C GNP

TSP 要求回归方程象处理其它解释变量一样处理常数项，这与许多回归程序是不一样的。在TSP 软件包中，有一个特殊的、预先生成的序列，名字为C。这个序列可以用于LS 及其他统计运算中，但不必象输入其他数据那样输入该序列，所以用户不要把自己建立的序列命名为C。

滞后序列可以出现在LS 或其他统计运算中。一个序列的滞后序列与原序列同名，只是要在序列名后的括号中给出滞后期。如：LS CONS C CONS(-1) GNP

设Y是被解释变量矩阵，X 是解释变量矩阵。LS 命令计算最小二乘法回归系数和各种有关的统计量。这些统计量包括系数的标准差和相应的 t －检验值、回归标准差、D-W 检验值。如果回归方程中包含常数项，则判定系数(R^2)、调整后的判定系数以及总体F－检验值也将显示出来。TSP提供了几组可选择的参数,使用户可以灵活地掌握那些回归结果在屏幕上显示或打印。基本回归结果显示之后，软件包将询问是否查看系数的协方差矩阵。可以输入S 在屏幕上显示这个矩阵,或输入P 将它在打印机上打印出来,或按回车键跳过这个选择。之后系统提问：

要显示残差值、实际值和拟合值吗？(P,S,G,ENTER)

回答P 或S 将打印或显示字符形式的残差图、被解释变量的实际值、拟合值和残差值。回答G 将以双比例尺方式生成残差、实际值和拟合值的高分辨率图形，也可按回车键跳过这个选择。

在LS 命令中可以使用两个选择来处理异方差性。H 选择要求TSP 使用相合的协方差矩阵代替前面给出的计算标准差和t-检验值时使用的协方差矩阵，请参阅[2]。如：LS(H) EDUC C AGE INCOME FATHER

W 选择加上一个等号一个权值序列名说明进行加权最小二乘法。若使用了这个选择，则在进行估计之前所有数据都要与权值序列相乘。如：

LS(W=SCALE) EDUC C AGE INCOME FATHER

㈡两阶段最小二乘法

TSLS命令计算两阶段最小二乘模型。两阶段最小二乘法有时也叫工具变量法。在TSLS命令中，被解释变量和解释变量按LS命令中变量的排列方式列出后键入一个@ 字符，然后列

出工具变量。例如,在消费方程中，可以把政府支出G、货币供给量的自然对数LM、趋势变量TIME 作为外生变量，因此可用做工具变量，输入命令如下：

TSLS CONS C GNP @ C G TIME LM

常量C 总是一个合适的工具变量,设Z是工具变量值构成的矩阵,Y和X分别是被解释变量和解释变量矩阵,TSLS根据以下方程计算估计方程的系数:

$$(X'Z(Z'Z)^{-1}Z'X)^{-1}X'X(Z'Z)^{-1}Z'Y$$

这些系数的协方差矩阵为: $XZ(X'Z(Z'Z)^{-1}Z'X)^{-1}$

与LS命令的计算过程一样,有关统计量也被计算。所有与残差有关的统计量的计算公式与LS命令的计算公式一样。这里的残差是“结构残差”Y-Xb, 而不是“第二阶段残差”。在TSLS 命令中可以使用W选择来进行加权两阶段最小二乘法。

注意：进行TSLS 估计时,即使方程中有常数项，R^2也可能是负值。

㈢一阶序列相关

当线性回归模型的扰动项序列相关时, 由普通最小二乘法估计出的系数尽管是无偏的,却是无效的。当扰动项出现一阶序列相关时，AR(1)命令提供了得到有效估计值的方法，为了使用序列相关校正方法估计消费函数，输入命令：

LS CONS C GNP AR(1)

用AR(1)方法进行误差序列校正的LS 计算过程使用了D.Cochrane 和G.H. Orcut [4] 提出的两阶段迭代方法。这种方法利用普通最小二乘法计算出的残差估计rho 的值，对被解释变量进行变换，使得变换后的估计方程的残差项基本上序列不相关，然后使用变换后的变量估计回归方程。变换方程是：

xi=xt-rho*xt-1

这种计算过程一直重复进行到rho收敛或迭代次数达到了预先给定的上限为止。一般来说，Cochrane-Orcutt 过程渐近于极大似然法, 但用小样本估计时它们是不同的。

包含AR(1) 命令的LS命令也给出通常的回归计算结果和根据rho 变换变量计算各种统计量及残差散点图。如果一阶AR 校正误差序列是恰当的, 输出的残差序列将是不相关的白噪声序列。

对于包含AR(1)命令的LS命令, 用SMPL 命令确定的样本区间一定要保证Cochrane-Orcutt过程使用的滞后观察值是存在的。例如，如果被解释变量和解释变量的第一个观察值对应的观测点均为1949 年, 则SMPL 命令说明的样本期最早只能从1950年开始。如果以工作文件的第一个样本点为起点，包含AR(1)命令的LS 命令会出错。

使用OPTION 命令,可以在执行LS 命令时控制迭代过程。通常，迭代20 次后,即使rho不收敛,也将停止执行LS 命令。选择MAXIT 能够改变对迭代的次数限制。例如，为将迭代次数限制增为100，只需输入命令：

OPTION MAXIT 100

LS 通过检查两次迭代得到的rho估计值之差来决定收敛程度。通常，当差值小于等于0.005时, LS 停止迭代并打印计算结果。为改变收敛判别标准，应选择CONVERGE。例如，为提高计算精确性，把判别标准改为0.001, 应输入命令：

OPTION CONVERGE 0.001

每一次迭代时, LS 试验给出步长。它首先检验采用这个步长能否降低残差平方和，如果残差平方和降低，就采用它并检验收敛程度，如果需要，就进行下一次迭代。如果所取的

步长不能降低残差平方和，取步长的一半执行迭代过程。假如还不能降低残差平方和，再把步长一分为二。如果此时仍无改进，LS 命令停止计算，并给出不能继续计算的信息。

MAXAZQ 选择控制LS 将步长一分为二的次数。通常只分两次，但命令

OPTION MAXSZQ 4

将其增加到4 次。一旦用OPTION 命令改变了一个参数,该参数在本次会话中一直有效。可以用一个OPTION 命令重新定义多个参数，如：

OPTION CONVERGE .01 MAXIT 50

㈣多项式分布滞后

估计方程中可以包含多项式分布滞后,一个方程中最多可以含有五个PDL项。每一项通知TSP 对一个序列计算多项式分布滞后系数,并根据多项式的形式确定这些系数。

PDL 项后面的括号中应给出一个序列名，如：LS SHPMNT C PDL(ORDERS) 表示SHPMNT 被ORDERS 序列的分布滞后拟合。输入以上命令后，TSP 要求给出分布滞后的补充信息。必然提供：

滞后长度－即滞后多少时间周期。

多项式的阶数。

是否强制系数在滞后分布的开始或末端趋近于零，或在开始和末端都趋近于零。用以下数字表示对估计系数的限制：

1－在分布近端趋近于零；

2－在分布远端趋近于零；

3－在分布近端和远端都趋近于零。

也可以在括号中序列名后面给出这些信息。如果不做任何限制，可以省略这些信息。例如输入命令：

LS SHPMNT C PDL(ORDERS,8,3)

LS SHPMNT C PDL(ORDERS(-1),12,4,2)

PDL 命令也可以用在TSLS 命令中。如果PDL 后面括号中给出的序列是外生的,应做为工具变量。为此，可以给出PDL(*)，这意味着所有的PDL变量做为工具变量。例如可以输入命令：

TSLS SHPMNT C TBILL MI PDL(ORDERS(-1),12,4,2)

@ FED FED(-1) PDL(*) 在PDL 的计算过程中，TSP 临时产生一些名为PDL1,PDL2,…的序列。这些序列用在LS 或TSLS 估计中。如果对PDL 估计很熟悉,可以将这些序列用作其它目的,例如作为TSLS 中的一个工具变量。

§12.2.4 高级统计技术

高级统计技术包括非线性最小二乘法和向量自回归模型。

㈠非线性最小二乘法

非线性模型有两种：一种是关于变量是非线性的；另一种是关于参数是线性的，当然也可以是关于二者均是非线性的。若一个方程关于变量是非线性的，但关于参数是非线性的，则能够用LS或TSLS方法来估计。例如，若有模型：

Y=D+A*LOG(L)+B*LOG(K) 关于变量L 和K 是非线性的,但关于参数是线性的,这样可以通过建立新的变量来使用普通最小二乘法:

GENR LOGL=LOG(L)

GENR LOGK=LOG(K)

LS Y C LOGL LOGK

但是这种方法只适用于方程的系数是线性的场合。当方程的系数是非线性的情况时，如方程:

```
Y=D*(L^A)*(K^B)+F*DUM
```

就无法建立新的序列然后用LS去估计。这时我们就可以使用非线性最小二乘法命令NLS。

NLS 使用特殊的方法来表示方程从而使其能够记着哪些参数是要估计的。这些参数依次叫做C(1),C(2),…估计上述方程的NLS 命令为：

```
NLS Y=C(1)*(L^C(2))*(K^C(3))+C(4)*DUM
```

在使用NLS命令前,必须为TSP提供待估计参数的初始值。PARAM 命令就是为此设置的。例如，下面的PARAM 命令就为上述NLS命令提供了一组可用的初值:

PARAM 1 213 2 .7 3 .3 4 0

这些命令设置C(1)=213,C(2)=0.7,…。当NLS 命令估计完成时, 估计出的系数还保留在C(1)、C(2)…中。如果这些数值可以用作下次估计的初始值，就不必使用新的PARAM 命令重设初值, 如果使用了新的PARAM 命令, 仅仅需要说明要修改的参数的初值。

一旦用PARAM 命令设置了参数初值,就可以用NLS 命令进行估计。NLS 估计过程如下：每一次迭代，通过对其中一个参数做少量的改变来计算有关参数的偏差，并查看方程变化了多少。然后将被解释变量与这些偏差做回归。在必要时，为了避免共线性问题，要使用曲线回归，但NLS 的最终结果不包含曲线回归。回归给出一个对参数计划进行修改的数值向量。NLS 评价这些计划修改的参数是否降低了残差平方和。如果残差平方和降低了，就对参数进行修改并开始新的一轮迭代。否则就对计划修改量进行分割。一般来说，在进行一定次数的分割之后对方程会有所改善。这种过程一直进行下去，直到修改的数量与参数本身相比非常小为止。

通过修改MAXIT、CONVERGE 和MAXSQZ 的参数值可以控制迭代过程。

有时NLS 甚至在第一次迭代就不能改善残差平方和,这样迭代过程就会停止。在大多数情况中，这种失败原因在于参数的初始值给得不好。例如，在上面的方程中，如果C(1) 的初值设为0, 则方程与C(2) 和C(3) 有关的偏差也将是0,这样迭代过程就无法开始了。

另一方面，对于许多行为很好的问题，不管定义了什么样的初值，收敛迅速。因此尝试精选好的初始值是没有必要的。一旦发现了一组能够导致收敛的初始值，就应予以保留。

NLS 在试图对参数修改的操作中一直给用户提示信息。当对计划修改量进行分割而改变了步长时,在屏幕上显示一个*号。在每一次成功地迭代之后，将显示参数值和残差平方和。

在取得了收敛之后，NLS 将给出与LS 相同的信息。虽然两种方法的理论只是近似的，但所有标准的统计结果和检验值的解释是相同的。

下面的例子是用NLS 估计CES 生产函数的参数：

PARAM 1 1 2 .1 3 .5 4 -1

NLS Q=C(1)*(C(3)*L ^ C(4)+(1-C(3))*K ^ C(4)) ^ C(2)

在NLS 中也可以使用W 或WEIGHT 选择来完成加权非线性回归。

㈡广义logistic和probit模型

广义logistic和probit模型研究反应为二分类的数据。概率反映一种事件发生的可能性，用0、1之间的一个数来表示。所以说明表达式的值要在这个范围之内。广义logistic和probit模型形式为：

$prob[y=0]=1/(1+exp(b_0+b_1x_1+b_2x_2+...))$ 及

$prob[y=0]=1-p(b_0+b_1x_1+b_2x_2+...)$

这里概率p(x)是具有正态分布，零均值, 方差为1 的随机变量的分布函数。

这两种表达式都使得概率值依赖于观察变量$x_1,x_2,...$ 这些变量分别乘以系数b_1,b_2,估计模型的目标是找出最好的系数值。当某个特定变量的系数为正值时，意味着变量的值越高，对应着y=0 的概率越小，换句话说，变量的值越高，对应于y=1 的概率越大。

广义logistic和probit通常是用最大似然法估计。达到最大值的过程需要迭代方法，但由于这两种函数的性质很好，所以在大多数情况下使用平滑方法。

使用LOGIT 命令进行广义logistic估计，LOGIT 要求给出反应变量名和回归变量，回归变量中要包含常数项。注意命令形式与LS 命令完全相同。但是被解释变量的值只能是0或1。若被解释变量取了其它的观察值，在估计时对应的样本点要被舍掉。命令输入例子：

LOGIT OWNVCR C SIZE INCOME URBAN CABLE

对于probit模型，使用PROBIT命令，其形式与LOGIT相同，例如：

PROBIT DEMOWIN C GROWTH WAR INFL

LOGIT 和PROBIT 命令每一次迭代时都提供了对数似然函数的值，估计结果包括回归系数、标准差、t-检验值和显著性水平。系数的解释必须要按LOGIT 和PROBIT 函数形式来进行。

此外，在结果中还包含被解释变量中0 和1 观察值的统计数。

㈢系统估计

系统估计是指估计一个完整的联立方程组。在TSP 中使用SYS 命令来完成。SYS 命令联立所有方程有两个主要的优点：首先，不同方程的残差经常是互相关联的，SYS 命令允许使用这种互相关性来改善估计的有效性；其次，用户可能要限定一个方程的系数与系统中其它一个或多个方程的系数是相同或相关的。这仅仅在联立估计所有方程时才是可能的。

系统估计也有潜在的缺点，如果错误地说明了一个系统中的一个方程，在单方程估计中，仅仅该方程不能正确地估计，而在系统估计中，所有的方程都不能被正确地估计。另外，系统估计方法很费时间，而且在TSP 中仅能估计70 或少于70 个系数的方程组。但在单方程估计中，每个方程都可以用到70 个系数。TSP 中, SYS 命令能够估计包含线性或非线性方程的方程组。使用的技术有：

(0) 普通最小二乘法

(W) 加权最小二乘法—在方程内取常数权值

(H) 迭代加权最小二乘法

(S) 似乎不相关回归

(I) 迭代似乎不相关回归

(2) 两阶段最小二乘法

(J) 加权两阶段最小二乘法

(3) 三阶段最小二乘法

(T) 迭代三阶段最小二乘法

在运行SYS 命令后,若指定了0 或2 选择(普通最小二乘法或两阶段最小二乘法), 则结果在线性方程情况下与LS 或TSLS 估计的结果是完全相同的。唯一的例外出现在对一些方程的系数进行了限制的情况下。非线性方程由修正的NLS 方法估计。这种修正的NLS 比NLS 更加复杂而且需要更长的时间才能达到收敛。

指定了W、H、J 或K 选择进行的加权估计除了对每个方程的标准差的倒数估计值加权外，与0或2选择是相同的。如果不同的方程有不同的方差，但不同方程的残差是不相关的，

则这种修正用于处理交叉方程的异方差性。如果没有交叉方程系数限制，这些技术将给出与0 与2 选择一样的估计结果。然而当存在交叉方程限制时，它们给出更好的结果。

似乎不相关回归(SUR) 和三阶段最小二乘法(3SLS) 是普通最小二乘法(OLS) 和两阶段最小二乘法(TSLS) 的对应系统方法。由SUR 或3SLS 进行的估计过程是:首先用单方程估计技术估计方乘, OLS 用于SUR, TSLS 用于3SLS。利用第一次迭代的方程系数计算出残差和残差协方差矩阵。在第二次迭代时使用方程残差之间的协方差对第一次估计的系数进行修改。

如果选择了S或3并且系统中的方程都是线性的, 则估计过程在两次迭代之后就完成了。如果选择了I 或T,或者系统中包含非线性方程, 则在第二次迭代之后重新计算残差和协方差矩阵。估计过程重复下去直到收敛为止。这种技术就是完全信息极大似然法(FIML)。

SYS 的输入是一个系统文件(文本格式),这个文件可以用TSP 的EDIT 命令或其它的文本编辑软件产生。系统文件中包含要估计的全部方程。除了方程，系统文件中还可能包含工具变量表和参数初值(仅在非线性模型中是需要的)。也可能包含一些定义方程。定义方程并不用在估计过程中但将转换到输出文件中以便于用进一步的模型模拟。

例如，可以建立一个模型文件MODEL1.SYS 使之包含下述内容:

```
INST LG RD2 AGYG XG XGXY TT4 DCPYP LDCYP LRSL DEP ⅡⅡ(-1) S(-1)
PARAM 1 .5 2 .5 3 1.5
GG=C(1)+C(2)+ⅡⅡ ^ C(3)+(1-C(2))*LG ^ C(3)
ⅡⅡ=(1-C(2))/C(3)+C(4)*GG+C(5)*TT4+C(6)*TT4(-1)+C(7)*LOG(DCPYP)+C(8)*ⅡⅡ(-1)
S=C(20)+C(10)*(GG+TT4)+C(11)*RDZ+C(12)*DEV+C(13)*S(-1)+[AR(1)=C(14)]
```

在文件的第一行中出现的INST 语句指出LG，RD2，…，DEP 是工具变量。如果没有对每个方程分别给出工具变量表，则对于2SLS，加权2SLS,3SLS 或迭代3LST，INST 语句是必须的。如果选择了OLS、加权LS、SUR 或迭代SUR, INST 语句是不需要的,此时若说明了INST 语句，则估计时将其忽略。在每个系统文件中只能包含一个INST 语句。

INST 语句为系统文件中的所有方程说明相同的工具变量表。但是，TSP 不要求所有的方程都使用相同的工具变量表。为了说明某个方程特定的工具变量表，只需在方程所给出一个“@”符号，然后指出工具变量表。例如：

```
CONS=C(1)+C(2)*GNP @ C GNP(-1) CONS(-1) M1
```

进行估计时，方程后面给出的特定工具变量优先于INST语句中给出的工具变量。

PARAM 语句为第一个方程说明参数初值: C(1)=0.5,C(2)=0.5,C(3)=1.5。由于其它的方程关于参数是线性的，所以没有必要说明参数初值。如果一个方程关于参数是线性的，则PARAM 语句是无效的。如果对于一个非线性方程没有说明PARAM 语句，则TSP 就使用系数向量的当前值作为参数初值，而不管这些当前值是什么。当在系统文件中包含另一个PARAM 语句时，则后面出现的语句优先用于SYS命令中。

SYS 使用与NLS 同样的方法表述方程。系数依次叫做C(1)、C(2) …没有必要连续地使用系数序号。唯一的限制是系数序号不能大于80, 例如C(81) 是非法的, 并且总的系数个数不能超过70 个。

方程既可以是系数非线性也可以是变量非线性或二者均非线性，方程系数约束可以通过系数运算来实现。例如，有三个系数出现在三个方程中，它们的和为1 ，则可以这样设定：C(1)在第一个方程中，C(2)在第二个方程中，然后说明(1-C(1) -C(2)) 作为第三个方程的系数。

一阶自回归误差校正项可以通过在方程后面的方括弧中说明AR(1)=C(常数) 来完成。注意在方括弧中不要含有空格。例如：

CONS=C(1)+C(2)*GNP+[AR(1)=C(3)]

通过在所有的方程中给出相同的自回归系数序号可以限定系统中的全部方程,也可以给每一个方程说明特定的自回归系数序号。SYS 命令不能估计移动平均误差项和高于一阶的自回归误差校正项。

系统文件中可以包含注释行。注释行以“@”字符开头。例如:

@ 这是注释行

与所有的TSP 命令一样,在用SYS 命令进行估计之前,必须用SMPL 命令说明所使用的样本区间。但是SYS并不需要对所有方程说明相同数目的样本点。如果在SMPL 说明的样本区间中的一部分有一个或多个方程中的序列出现缺值,但在SYS 命令中使用了MD 选择,则SYS 会正确地估计方程系统。

一旦用EDIT 建立了系统文件并用SMPL确定了估计用的样本区间,就为使用SYS 命令做好了准备。完整的SYS 命令格式为:

SYS(m,MD)系统文件名模型文件名

其中各个参数的意义如下:

m 是所使用的估计方法,其值可以为0、W、H、S、2、J、K、3 或T。

MD 选择告诉TSP,即使出现了缺值,估计过程仍继续下去。如果要求TSP 发现缺值样本点时即给出提示信息,请用一个逗号代替MD。

系统文件名:在上面描述过。是包含方程、工具变量表和参数初值的文件的名字。

模型文件名:当用SYS 命令估计一个系统文件时,若想把估计出来的模型输出到一个文件中以便用SOLVE 命令去求解,则要在SYS 命令中给出模型文件名。模型文件名是可选项,若省略了这个名字,则不会生成模型文件。

例如,为了用迭代三阶段最小二乘法估计MODEL1.SYS系统文件并生成MODEL1. DML模型文件,只需输入:

SYS(T) MODEL1.SYS MODEL1.MDL

另一种方法是由TSP提示如何输入。若在统计操作菜单中选择SYS 或直接输入SYS,则TSP 提示:

请输入系统文件名?

回答MODEL1.SYS。若想把估计结果输出到一个文件中,则可回答:

MODEL1.SYS MODEL1.MDL

这样MODEL1.MDL 中将包含估计出的模型。然后TSP 提问:

(0) 普通最小二乘法

(W) 加权最小二乘法—在方程内取常数权值

(H) 迭代加权最小二乘法

(S) 似乎不相关回归

(I) 迭代似乎不相关回归

(2) 两阶段最小二乘法

(J) 加权两阶段最小二乘法

(3) 三阶段最小二乘法

(T) 迭代三阶段最小二乘法

由于MODEL1 中包含关于参数非线性的方程,所以选择S 和I 是一样的,选择3 和T 也是一样的。为了由三阶段最小二乘法进行估计,请选择T。

在系统文件估计完成后,可以通过输入:

表10.1 1749—1924年太阳黑子年平均数(176个数据)

80.9	83.4	47.7	47.8	30.7	12.2	9.6	10.2	32.4	47.6
54.0	62.9	85.9	61.2	45.1	36.4	20.9	11.4	37.8	69.8
106.1	100.8	81.6	66.5	34.8	30.8	7.0	19.8	92.5	154.4
125.9	84.8	68.1	38.5	22.8	10.2	24.1	82.9	132.0	130.9
118.1	89.9	66.6	60.0	46.9	41.0	21.3	16.0	6.4	4.1
6.8	14.5	34.0	45.0	43.1	47.5	42.2	28.1	10.1	8.1
2.5	0.0	1.4	5.0	12.2	13.9	35.4	45.8	41.1	30.4
23.9	15.7	6.6	4.0	1.8	8.5	16.6	36.3	49.7	62.5
67.0	71.0	47.8	27.5	8.5	13.2	56.9	121.5	138.3	103.2
85.8	63.2	36.8	24.2	10.7	15.0	40.1	61.5	98.5	124.3
95.9	66.5	64.5	54.2	39.0	20.8	6.7	4.3	22.8	54.8
93.8	95.7	77.2	59.1	44.0	47.0	30.5	16.3	7.3	37.3
73.9	139.1	111.2	101.7	66.3	44.7	17.1	11.3	12.3	3.4
6.0	32.3	54.3	59.7	63.7	63.5	52.2	25.4	13.1	6.8
6.3	7.1	35.6	73.0	84.9	78.0	64.0	41.8	26.2	26.7
12.1	9.5	2.7	5.0	24.4	42.0	63.5	53.8	62.0	48.5
43.9	18.6	5.7	3.6	1.4	9.6	47.4	57.1	103.9	80.6
63.6	37.6	26.1	14.2	5.8	16.7				

SOLVE MODEL1.MDL

来运行动态模型(提供的模型已存储在MODEL1.MDL 中)。

§12.3 用例与样本程序

§12.3.1 用例

【例12.1】时间序列模型分析中，太阳黑子数据是比较经典的例子，现对它进行分析：

设上述数据存于文件S.TXT，读取数据的MicroTSP命令为：

>creat a 1749 1924

>read(s) s.txt sunspot

根据Pandit和吴贤铭的建模策略，首先拟合ARMA(2,1)和ARMA(4,3)模型：

>ls sunspot c ma(1) ar(1) ar(2)

结果为：

```
LS // Dependent Variable is SUNSPOT
Date: 12-14-1993 / Time: 14:33
SMPL range: 1751  -  1924
Number of observations: 174
Convergence achieved after 4 iterations
```

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
C	44.480818	3.9717037	11.199430	0.000
MA(1)	-0.1614766	0.1171300	-1.3786103	0.168
AR(1)	1.4284049	0.0884517	16.148989	0.000
AR(2)	-0.7245218	0.0793022	-9.1362137	0.000

R-squared	0.804150	Mean of dependent var	44.33736
Adjusted R-squared	0.800694	S.D. of dependent var	34.73446
S.E. of regression	15.50675	Sum of squared resid	40878.08
Durbin-Watson stat	1.979609	F-statistic	232.6706
Log likelihood	-721.8539		

¿ls sunspot c ma(1) ma(2) ma(3) ar(1) ar(2) ar(3) ar(4)
结果为：

LS // Dependent Variable is SUNSPOT
Date: 12-14-1993 / Time: 14:35
SMPL range: 1753 - 1924
Number of observations: 172
Convergence achieved after 10 iterations

VARIABLE	COEFFICIENT	STD. ERROR	T-STAT.	2-TAIL SIG.
C	44.671449	1.5564850	28.700212	0.000
MA(1)	0.9301516	0.8837230	1.0525375	0.293
MA(2)	0.3950168	0.5927001	0.6664700	0.505
MA(3)	-0.1018366	0.1219219	-0.8352611	0.404
AR(1)	0.3676083	0.8801520	0.4176646	0.676
AR(2)	0.2755596	0.7192365	0.3831280	0.702
AR(3)	-0.0348496	0.4872030	-0.0715299	0.943
AR(4)	-0.3663540	0.4745024	-0.7720804	0.440

R-squared	0.812607	Mean of dependent var	44.29767
Adjusted R-squared	0.804608	S.D. of dependent var	34.93502
S.E. of regression	15.44237	Sum of squared resid	39108.53
Durbin-Watson stat	1.999424	F-statistic	101.5951
Log likelihood	-710.7452		

由于ARMA(2,1)和ARMA(4,3)的残差平方和(Sum of squared resid) 分别为SSRI=40878.08和SSRII=39108.53，故F 统计量为：

F=[(40878.08-39108.53)/4]/[39108.5/(176-8)]< $F_{0.95}(4,\infty)$=2.37

表明无需拟合比ARMA(2,1)更高阶的模型。注意到在ARMA(2,1)中，MA(1) 的参数偏小。且其置信区间包含0，故F下面考虑AR(2)模型。

>ls sunspot c ar(1) ar(2)
结果为：

LS // Dependent Variable is SUNSPOT

```
Date: 12-14-1993 / Time: 14:37
SMPL range: 1751  -  1924
Number of observations: 174
Convergence achieved after 1 iterations
     VARIABLE   COEFFICIENT   STD. ERROR      T-STAT.   2-TAIL SIG.
        C        44.393596     3.7546702     11.823567     0.000
      AR(1)       1.3361112     0.0580720     23.007858     0.000
      AR(2)      -0.6501527     0.0581050    -11.189270     0.000
R-squared            0.801962    Mean of dependent var   44.33736
Adjusted R-squared   0.799646    S.D. of dependent var   34.73446
S.E. of regression   15.54746    Sum of squared resid    41334.70
Durbin-Watson stat   2.121557    F-statistic             346.2362
Log likelihood      -722.8203
```

比较AR(2)与ARMA(2,1)，其F统计量值为：
F=[(41334.78-40878.08)/1]/[40878.08/(176-4)]< $F_{0.95}(1,\infty)$=3.84
因此，从上面F值知AR(2)对太阳黑子活动数据是合适的。

【例12.2】非线性回归用例(王学仁等编《应用回归分析》，重庆人民出版社)：
设线性回归方程为：$Y=\alpha+(0.49-\alpha)\exp[-\beta(x-8)]$，观测数据(X,Y)为：
10 .48
20 .42
30 .40
40 .39
设数据存于文件NL.DAT，TSP 分析程序NL.PRM为：

```
create (u) 4
read(o) nl.dat x y
nls y=c(1)+(0.49-c(1))*exp(-c(2)*(x-8))
param 1 0 2 1
option output c:nl.out
pon
```

在TSP的"¿"提示符下打入RUN NL.PRM 命令，经九次迭代，结果：

```
 C(1)    0.403279    C(2)    0.509604     SSR     0.00252374
 C(1)    0.406802    C(2)    0.161704     SSR     0.00055278
 C(1)    0.397746    C(2)    0.102024     SSR     0.00025669
 C(1)    0.389350    C(2)    0.096182     SSR     7.7215E-05
 C(1)    0.384984    C(2)    0.086886     SSR     5.2099E-05
 C(1)    0.382316    C(2)    0.082154     SSR     4.6129E-05
 C(1)    0.381111    C(2)    0.080126     SSR     4.5305E-05
 C(1)    0.380774    C(2)    0.079565     SSR     4.5257E-05
 C(1)    0.380730    C(2)    0.079492     SSR     4.5257E-05
Convergence achieved after 9 iterations
```

```
============================================================================
                 COEFFICIENT   STD. ERROR      T-STAT.   2-TAIL SIG.
============================================================================
        C(1)      0.3807300     0.0085770     44.389472     0.001
        C(2)      0.0794917     0.0162171     4.9017227     0.039
============================================================================
R-squared               0.990717    Mean of dependent var   0.422500
Adjusted R-squared      0.986075    S.D. of dependent var   0.040311
S.E. of regression      0.004757    Sum of squared resid    4.53E-05
Durbin-Watson stat      1.958976    F-statistic             213.4379
Log likelihood          17.10316
============================================================================
```

即$\alpha = C(1) = 0.3807, \beta = C(2) = 0.07949$。

【例12.3】多方程系统估计。仍用第4章食品消费和价格的例子，系统方程为：

Y= C(10) + C(11) X1 + C(12) X2

Y= C(20) + C(21) X1 + C(23) X3 + C(24) X4

原始数据(Y,X1,X2,X3,X4)为：

```
98.485  100.323  87.4 98  1
......
106.232 113.49  127.1 93 20
```

设数据存于文件ECO.DAT中，TSP 程序ECO.PRM为：

```
create a 1 20
read(o) eco.dat y x1 x2 x3 x4
option output e:eco.out
pon
sys(o) ecomd.sys
```

模型ECOMD.SYS为：

```
inst x1 x2 x3 x4
y=c(10)+c(11)*x1+c(12)*x2
y=c(20)+c(21)*x1+c(23)*x3+c(24)*x4
```

在TSP" > "提示符下运行命令RUN ECO.PRM，完成参数估计。注意：

1. 对普通最小二乘不必指定工具变量。

2. 对两步最小二乘和三步最小二乘，程序中指定工具变量为：X1,X2,X3,X4。

3. 在上述程序中在采用普通最小二乘、两步最小二乘和三步最小二乘估计时，语句：SYS(*) ECOMD.SYS 中的星号(*)应分别换为0,2,3。

输出结果如下表：

```
============================================================================
     OLS         COEFFICIENT   STD. ERROR      T-STAT.   2-TAIL SIG.
```

```
=============================================================
     C(10)      99.839109     7.4941890     13.322203     0.000
     C(11)     -0.3160949     0.0903416    -3.4988839     0.003
     C(12)      0.3350539     0.0452963     7.3969450     0.000
=============================================================
     C(20)      58.276931     11.461841     5.0844304     0.000
     C(21)      0.1603437     0.0948678     1.6901806     0.110
     C(23)      0.2481402     0.0461884     5.3723480     0.000
     C(24)      0.2483209     0.0975180     2.5464119     0.022
=============================================================

=============================================================
     2LS        COEFFICIENT   STD. ERROR      T-STAT.   2-TAIL SIG.
=============================================================
     C(10)      99.839083     7.4941789     13.322218     0.000
     C(11)     -0.3160940     0.0903415    -3.4988790     0.003
     C(12)      0.3350533     0.0452963     7.3969295     0.000
=============================================================
     C(20)      58.276726     11.461841     5.0844123     0.000
     C(21)      0.1603444     0.0948677     1.6901901     0.110
     C(23)      0.2481415     0.0461884     5.3723729     0.000
     C(24)      0.2483215     0.0975180     2.5464184     0.022
=============================================================

=============================================================
     3LS        COEFFICIENT   STD. ERROR      T-STAT.   2-TAIL SIG.
=============================================================
     C(10)      99.215728     6.9048101     14.369074     0.000
     C(11)     -0.2703525     0.0812652    -3.3267929     0.004
     C(12)      0.2945319     0.0385021     7.6497632     0.000
=============================================================
     C(20)      62.364841     9.9066193     6.2952697     0.000
     C(21)      0.1459674     0.0844489     1.7284697     0.103
     C(23)      0.2117410     0.0356311     5.9425863     0.000
     C(24)      0.3308977     0.0606537     5.4555216     0.000
=============================================================
```

模型输出R-平方、调整R-平方、回归标准误、Durbin-Watson 统计量、因变量的均值与标准差，均方误差和、F-统计量，此处从略。

§12.3.2 样本程序

MicroTSP 6.5 系统同时附加了样本演示程序，对该软件的掌握很有帮助，现介绍如下。

★第一个样本程序，是想预测Holiday Inns hotel 的收入，演示数据处理、回归和预测。第一步是建立一个工作文件，它保存于计算机的RAM中，这可以由CREATE命令完成。使用该命令后，系统提问数据的频度，时间的区间，现在是1970 到1981的年度数据。也可以直接使用命令：

CREATE A 70 81

下一步使用数据编辑器来录入数据。序列名至多八个字符，使用DATA命令或在数据操作菜单下选择进入数据编辑。MicroTSP提示要编辑的序列名，然后录入数据：

```
697
814
963
1122
1224
1369
1569
```

假设应该录入的数据是1539，我们可以用命令.B返回去把它修正，继续录入：

```
1780
2161
2605
2915
```

最后一个观察是1980 的，因此，我们完成了数据录入，使用命令X退出。现在可以看出REV出现在状态窗口了。

也许我们应该把序列REV 存到磁盘上，当需要时FETCHed 调入。

> STORE REV

要在数据库文件中包括说明，可以使用LABEL 命令插入或调出。

> LABEL REV Total room revenue, in millions of dollars

当LABEL不带其它参数时，现有的信息就会出现在屏幕上，如：

> LABEL REV

为进行统计分析，还需要一些其它序列的数据，现使用FETCH命令。先用DIR命令浏览一下文件：

> DIR

所有数据文件都用扩展名.DB，因此可直接使用DIR *.DB 命令来浏览。结合上面的LABEL命令可以看出每个文件的描述。如对GNP.DB，使用命令：

> LABEL GNP

> FETCH GNP OCCUP

现在列出磁盘上可能对分析Holiday Inns 数据有用的序列的名字。

REV	Total room revenue, in millions of dollars
OCCUP	Occupancy rate - percent of rooms occupied on the average night
RRATE	Room rate - average revenue per occupied room per night
ROOMS	Number of rooms in the Holiday Inns system
GNP	U.S. Gross National Product in current dollars
PGNP	Implicit price deflator for GNP
UNEMP	U.S. Unemployment rate
CPR	Interest rate on prime commercial paper

得到其它的序列：

> FETCH RRATE ROOMS GNP PGNP UNEMP CPR

现在可以浏览我们的数据了，首先应告知数据的范围。使用SMPL 命令指定：

> SMPL 70 80 使用SHOW 命令得到数据的列表。

> SHOW REV OCCUP RRATE ROOMS GNP UNEMP

MicroTSP 也允许用户以图形方式浏览数据，在第二个演示程序中说明。

通常，数据在进行统计分析之前要进行转换，GENR 就是这样的命令。如产生序列RATIO，可以使用命令：

> GENR RATIO=REV/GNP

MicroTSP 处理数据滞后也很方便。首先，改变观察的范围：

> SMPL 71 80

产生一个序列，使它等于RATIO 变化的百分比，使用命令：

> GENR PCHR=100*(RATIO-RATIO(-1))/RATIO(-1)

GENR 命令也可用于动态模拟，以及对数据序列施加过滤。一个简单的用例是产生一个趋势变量，取值为0 in 1970, 1 in 1971, 2 in 1972,...., 11 in 1981。

```
> SMPL 70 70
> GENR TREND=0
> SMPL 71 81
> GENR TREND=TREND(-1)+1
```

事实上，MicroTSP 中有更为方便的方法，产生一个从1970 取值为零开始的趋势变量，使用命令：

> GENR TREND=@TREND(70)

> SHOW RATIO PCHR TREND

现在可以对上述数据进行一些统计分析了，MicroTSP 把最重要的结果先列出来，然后询问用户是否需要得更为详细，是在打印机(P)还是在屏幕(S)，或者跳过去(Enter)。

首先可以浏览数据的一些描述统计量：均值、标准差、最大值。最小值、协方差和相关。

> SMPL 71 80

> COVA REV OCCUP GNP UNEMP

在屏幕的最下一行，MicroTSP 会询问我们是否重复COVA 的输出。

数据的检查之后，我们可以产生关于REV 的预测方程，构造REV 关于常数(C) 、上年的国民收入GNP(-1)和收入REV(-1)。C 由系统设定，取值皆 1 。简单的最小二乘回归命令是LS，格式是：

> LS REV C GNP(-1) REV(-1)

MicroTSP 继续给出一些提示，与"S"和"P"相应，提示"G"表示高分辨率图形显示。MicroTSP 也把最小二乘的结果保留在内存中，这些结果可用于预测。使用命令SHOWEQ 命令可以浏览内存的模型。

> SHOWEQ

现要指明预测的范围，如仅仅是1981年的情况：

> SMPL 81 81

则可以使用命令FORCST，它带一个参数用于存贮预测的序列REVF，REVF在19 81年以前的值就是REV的值。

> FORCST REVF

最后可以统览一下：

> SMPL 70 81

> SHOW REV REVF

> SMPL 71 80

存贮工作文件，使用系统提供的SAVE命令，以后用LOAD就可以调入了。

> SAVE HOLIDAY

★MicroTSP 有许多命令用于产生高分辨图形，现在就用上面生成的工作文件HOLIDAY。

> LOAD HOLIDAY

由于REV和GNP是相关联的，这个关系可以通过图形来体现。MicroTSP 支持几种图形格式：线图、散点图、条图、圆图以及条图和线图的混合。图形可以在数据管理菜单(F4)下选择，PLOT 命令对应的是线图，MicroTSP 为线图提供了尺度选择，但我们可以用的它自动选择。

> PLOT(A) REV GNP

屏幕下方的一些选择，选择T (Type) 允许在屏幕上给出标号，用P (Print)则打印图形，S (Save) 则贮一个MicroTSP 图形文件，O(Options) 则允许用户改变图形，如图形的宽度、边界、字型、尺度等。

同样的数据可以用条图和散点图表达。

> BAR REV GNP

> SCAT REV GNP

★第三部分进一步演示MicroTSP 时序分析。假设用户对于Box-Jenkins 分析、两阶段最小二乘法、多项分布滞后是熟悉的。用MicroTSP构造一个汽车销售情况的简单模型，变量是：

TBQ 3-month Treasury bill rate

ASQ automobile sales in millions at annual rates

使用CREATE 产生一个季节型的工作文件。用特别的记号，如1958. 1 表示1958 年第一季度的数据。

> CREATE Q 58.1 78.4

> FETCH TBQ ASQ

> GENR TIME=@TREND(58.1)

首先产生一个ASQ 的单变量的Box-Jenkins 模型，把ASQ作为一个ARMA过程。为便于正确建模，首先看一下ASQ的相关图和偏相关图，由IDENT命令来完成。如浏览16个季节的自相关的偏自相关，打入命令：

> IDENT(16) ASQ

从相关图可以猜想, ASQ 是一个ARMA(2,0) 过程。LS 命令将估计这个ARMA 模型，由于ASQ具有正的均值，指定C作为常数项，但这里不是外生变量，而是AR() 和MA()的项，对1960.3 到1978.4 数据拟合模型，我们打入：

> SMPL 60.3 78.4

> LS ASQ C AR(1) AR(2)

现在可以看一下模型的新息(残差)的相关图，是否还存在自相关，MicroTSP 使用变量RESID 存贮新近估计的残差。

> IDENT(16) RESID

在滞后4 和8 处仍有一些相关，提示季节项的存在，现使用MA(4) 和MA(8) 项：

> LS ASQ C AR(1) AR(2) MA(4) MA(8)

虽然可以继续使用单变量的ARMA 模型，但我们想到利率会对销售额有一定的影响。可以对ASQ关于TBQ估计一个简单线性回归模型，然而模型将有序列相关，但是ARMAX技术使我们可以使用广义最小二乘法估计。

> LS ASQ C TIME TBQ TBQ(-1) TBQ(-4) AR(1) AR(2) MA(4)

看来TBQ 的确有益于预测汽车的销售情况。注意到TBQ有一个没有预料的正的系数，而滞后的TBQ则是负号。还可以使用SHOWEQ 命令显示模型的估计情况。

模型的系数也可用于GENR的计算，如：

GENR ASHAT = C(1)+C(2)*TBQ+C(3)*TBQ(-1)+C(4)*TBQ(-4)

给定利率解释销售的重要性，我们可以在ASQ方程中包括更多的TBQ滞后项，但可能会有共线性的问题。处理这一问题的一种方法是限制TBQ系数在多项式上，这用到了Polynomial Distributed Lags技术。使用这一技术时，MicroTSP 建立临时性序列PDL1,PDL2,... 用于估计。现对于ASQ，使TBQ 的八个滞后限制于三阶多项式，使用命令：

> LS ASQ C TIME PDL(TBQ,8,3) AR(1) AR(2) MA(4)

看来结果并没有改善多少，所以仍然回到原来的模型。TBQ 的系数仍然是正的，这可能由于ASQ与TBQ关联的结果。高的汽车需求可能对利率施加压力，现暂时忽略ARMA误差，试一下ASQ的两阶段最小二乘方程，这可以由位置符号@引导回归因子和工具变量来完成。

> TSLS ASQ C TIME TBQ TBQ(-1) TBQ(-4) @ C TIME TBQ(-1) TBQ(-2) TBQ(-3) TBQ(-4) ASQ(-1)

本例修正了simultaneity对TBQ系数的改变影响很小。

★第四个样本程序介绍一些更高级的技术，即非线性模型、probit 和二分类logit 模型。在计量经济学中有两种非线性，即变量或其系数的非线性，变量非线性而参数为线性则可由LS 或TSLS估计，如模型：

Y = a + b*LOG(L) + c*LOG(K)

关于L 和K 是非线性的，但其参数是线性的，可以首先创建变量，然后用普通的最小二乘。

> GENR LOGL = LOG(L)

> GENR LOGK = LOG(K)

> LS Y C LOGL LOGK

这种做法不适于参数是非线性的情况，如CES production function 模型：

$Q = a + b * LOG(c * L\hat{}d + (1 - c) * K\hat{}d)$

Q,L, 和K, 是对数log output, labor and capital。a,b,c 和d 是系数。

这时应使用非线性最小二乘命令NLS，它使用特殊的写法，其参数记为C(1), C(2), 等等。针对上述模型有：

> $NLSQ = C(1) + C(2) * LOG(C(3) * L\hat{}C(4) + (1 - C(3)) * K\hat{}C(4))$

使用NLS 之前，应提供参数的初值，这可以由PARAM完成，本例就是：

> PARAM 1 .5 2 -1 3 .5 4 -1

结果C(1) 为.5, C(2) 为-1, 等等。

现看此命令的用法，首先产生一个undated工作文件。

> CREATE U 30

把序列取到工作文件：

> FETCH Q L K

其次估计CES 函数。

> PARAM 1 .5 2 -1 3 .5 4 -1

> NLS Q=C(1)+ C(2)*LOG(C(3)*L^C(4)+(1-C(3))*K^C(4))

同线性估计一样，MicroTSP 也把方程存于内存，可以使用SHOWEQ 命令浏览。使用STOREQ 命令则可以把方程存于磁盘。使用FORCST 或FIT 可以进行预测。

现在转向probit 和logit 模型，一些人在工作训练中成功与否，受年龄、智力、文化的影响，logit 和probit 模型形式是：

Prob[PASS = 0] = 1/(1 + exp(a + b*AGE + c*IQ + d*EDUC))

Prob[PASS = 0] = 1 - P(a + b*AGE + c*IQ + d*EDUC)

仍然产生一工作文件，取序列并估计：

```
> CREATE U 38
> YFETCH PASS AGE EDUC IQ
> LOGIT PASS C AGE EDUC IQ
> PROBIT PASS C AGE EDUC IQ
> SHOWEQ
```

继续可使用FORCST, FIT, 或SOLVE。

★最后一个例子是考察portfolio of Wells Fargo Bank 并给出两阶段最小二乘和系统估计示例。数据是1963—1983年间Wells Fargo 的年度数据以及全国和加州地区在这个期间的经济情况。序列是：

SECU	Wells Fargo's holdings of securities, in millions of dollars
LOAN	Their holdings of loans
RSECU	Percentage return earned by Wells Fargo on securities
RLOAN	Percentage return on loans
SALES	Retail sales in California, in millions of dollars
RCP6M	Interest rate on 6-month commercial paper, percent
CONTR	Value of construction contracts in California,millions of dollars
TREND	Time trend, starting with a 1 in 1963

生成工作文件：PORT 是银行全部portfolio的大小，SPREAD是percentage spread between the interest earned on loans and the 6-month commercial paper rate。

由于RLOAN 序列始于1964, MicroTSP 不能计算SPREAD 在1963年的值，因此赋值NA(not available)。

LOARAT 是portfolio held in loans的比例。

LOASAL 是Wells Fargo's 贷款与销售的比。

CONSAL 是建设与零售的比。

```
> CREATE A 63 83
> FETCH SECU LOAN RSECU RLOAN SALES RCP6M CONTR TREND
> GENR PORT=LOAN+SECU
> GENR SPREAD=RLOAN-RCP6M
> GENR LOARAT=LOAN/PORT
> GENR LOASAL=LOAN/SALES
> GENR CONSAL=CONTR/SALES
```

分析的目的是了解银行从贷款获利的比例以及用贷款保持的portfolio比例。应该考虑：贷款的提供与贷款利率成正相关，为使顾客购买更旺，银行应使贷款利率相对于其它方面更加诱人，贷款需求与贷款利率成负相关。因此我们分析的目的是估计贷款的供求方程，以分出这两种影响。

对供给一方市场，我们对LOARAT 建模(fraction of Wells Fargo's portfolio held in loans)，假设Wells Fargo 在时返还贷款增加时增加贷款而在其它securities相对高时减少贷款。因而用LOARAT 对返还贷款RLOAN 、其它securities RSECU 及LOARAT滞后行回归。RSECU, LOARAT(-1), LOASAL(- 1) CONSAL, PORT, SALES, 和TREND 作为工具变量。

TSLS 命令是：

> SMPL 64 83

> TSLS LOARAT C RLOAN RSECU LOARAT(-1) @

C RSECU LOARAT(-1) LOASAL(-1) CONSAL PORT SALES TREND

估计的供给方程应予存贮。

> STOREQ SUPPLY

现看需求方程。使用贷款率、RLOAN 和其它一些主要的率的差很方便，该差是SPREAD。我们假设高的贷款需求源于高的零售，建设增加贷款利率和其它secur- ities间的差. 因此SPREAD的右端是LOASAL, 贷款对零售的比、LOASAL 的滞后、建设与零售的比CONSAL以及TREND。工具变量是LOASAL(-1) , CONSAL, and TREND plus RSECU, SALES, and PORT.

TSLS 命令是：

TSLS SPREAD TREND LOASAL LOASAL(1) CONSAL(-1) RSECU

LOARAT(-1) LOASAL(-1) CONSAL PORT SALES TREND

> STOREQ DEMAND

现在应转到完整的模型上来，第一个方程告知银行如何根据贷款和securities 利率作出决定，第二个方程告知应该如何让顾客有更多的购买力，完整就是把两者结合起来。MicroTSP 中，要求编辑一个模型文件，其包括模型的方程和其它有关信息。它有三种语句，即ASSIGN(告知MicroTSP 求解的序列存于不同的名) 、标识量(identity)、估计方程，本例为：

```
ASSIGN LOARAT PLORAT LOAN PLOAN SPREAD PSPRED
ASSIGN RLOAN PRLOAN LOASAL PLOASA
```

```
LOAN = LOARAT*PORT
RLOAN = SPREAD + RCP6M
LOASAL = LOAN/SALES
LOARAT = .2463962  +2.786724E-02*RLOAN -2.735787E-02*RSECU
         +.6360336*LOARAT(-1)
SPREAD = -4.277588 +.1052442*TREND -112.2502*LOASAL
         +103.5918*LOASAL(-1) +20.45162*CONSAL
```

可以使用MicroTSP的EDIT 编写模型：

> EDIT LOAN

EDIT 有许多点命令，使用.F 命令取得估计的方程。

.F SUPPLY

.F DEMAND

使用.L 命令浏览模型，使用.X 退出编辑。

使用TYPE LOAN命令可以证实模型确已存贮。

现用原始外生变量得到一个基础解。

```
> SMPL 76 83
> SOLVE LOAN
> GENR PSECU = PORT - PLOAN
> SHOW LOAN PLOAN SECU PSECU RLOAN PRLOAN
```

下一步是用模型模拟贷款市场，看建设是否在1976到1983要高10%。

> GENR CONSAL = 1.1*CONSAL

我们可以用EDIT 编辑一个与LOAN 类似的模型SIMLN，只要把前缀P改成S 即可。事实上软件已经事先准备好了，只消浏览一下。

```
> TYPE SIMLN
> SOLVE SIMLN
> GENR SSECU = PORT - SLOAN
> SHOW PLOAN SLOAN PSECU SSECU PRLOAN SRLOAN
```

仍然把CONSAL 恢复过来，并且看在贷款市场改变供给的影响。

```
> GENR CONSAL = CONSAL/1.1
> GENR PORT = PORT*1.1
> SOLVE SIMLN
> GENR SSECU = PORT - SLOAN
> SHOW PLOAN SLOAN PSECU SSECU PRLOAN SRLOAN
> GENR PORT = PORT/1.1
```

上述样本程序没有演示以下功能：

SYS	系统估计，包括三阶段最小二乘和似乎不相关回归
VAREST	向量自回归估计。
VARSTAT	从VAR 模型获得冲激响应函数及方差分解。
SAVE and LOAD	存贮工作文件。
EXPAND	改变工作文件的大小。
LGRAPH and PGRAPH	调入与打印磁盘图形。
READ and WRITE	与其它程序的数据交换(包括Lotus 123 工作表及 PRN 文件，WordStar 文件, DIF 文件) 及读取大型计算机而来的数据。
HIST	直方图。
CROSS	产生互相关图(cross correlograms).。
SMA and SAR	在ARMA和ARMAX模型引入季节(multiplicative) 移动平均及自回归误差项。
SMOOTH	指数平滑(single, double, Holt-Winters additive & multiplicative)
SEAS	季节调整。
STOREQ and FETEQ	存贮估计的方程。
SORT	按照一个或多个关键字对工作文件排序。
CONV	改变时间序列的频度，如四季资料到年份资料。
PON,POFF & TRACE	打印机控制。
RUN	批处理方式运行MicroTSP程序。
REN and DEL	磁盘工作文件换名和删除。
R and D	在工作文件中的序列更名和删除。
RND and NRND	产生随机数(GENR & models)。
OPTION	设定MicroTSP选项，包括输出再定向、屏幕和功能键控制等。

其它的特色，如：

- 统计输出和数据表格可转成磁盘文件或打印。
- 批处理操作，允许用户用哑元传递书写MicroTSP程序，从而允许宏操作。
- 这些程序可以包括FOR..NEXT循环。
- 在SHOW 和PRINT 中使用用户提供的格式输出数据表。
- 在PLOT 命令中定义输出格式。
- EDIT 可以编辑程序的模型文件。
- SOLVE 可以处理具有ARMA 误差的模型。

第十三章　GLIM

§13.1　GLIM 入门

§13.1.1　GLIM 简介

GLIM 第1版于70年代初问世，主要供专业统计人员使用。70年代中期3.22 版实现了商品化，1985年推出的3.77版最为流行，1993年又推出第4版。有别于其它通用统计软件包，使用GLIM要求用户对其建模过程有一个完整的概念。GLIM 提供了预分析的工具，如灵活的图形、优良的制表功能，回归、方差分析、列联表分析、生存分析，相应的误差分析等。其用于模型分析较之SAS、SPSS等的另一个优点是价格便宜。

GLIM 主要应用于三个方面。首先，它是一个强大的统计建模工具，供用户指定和拟合统计模型，评价拟合优度并给出估计量。建模过程不是预先设定的，这样用户就有了对建模过程最大程度的控制。其次，GLIM 可用于数据探索。最后，它是一个复杂的计算器，能够对单一的数或向量进行算术、逻辑、函数操作，宏操作。

§13.1.2　GLIM 系统组成

PC GLIM 3.77 系统组成：

EXAMPLES.GLM / EXAMPLES.LOG	样本程序及运行结果
EXMACLIB.GLM / EXMACLIB.LOG	宏调用样本程序及运行结果
GLIM.BAT / GLIM.LOG	系统运行批文件/运行记录文件
GLIMPROG.EXE / GLIMPROG.OVL	执行文件和覆盖文件
MACLIB.GLM	宏定义库
PROB.GLM	概率分布函数的宏
READ.ME	通道说明

§13.1.3　运行

设软件安装于$C:\backslash$ GLIM >，DOS 引导后，使用以下命令启动GLIM：

```
C:\> CD \GLIM <Enter>
C:\GLIM> GLIM <Enter>
```

或GLIM [参数1][参数2] $<$ *Enter* $>$

因GLIM.BAT 执行GLIMPROG.EXE，后者需要两个命令行参数，GLIM.BAT 则允许用户指定一个、两个或不指定，如：

```
C:\ GLIM>GLIMPROG %1 MACLIB.GLM
C:\ GLIM>GLIMPROG %1 %2
C:\ GLIM>GLIMPROG GLIM.LOG MACLIB.GLM
```

%1 是记录(transcript)文件，%2 表示宏定义文件。第三行即默认方式，代以GLIM.LOG，%2 代以MACLIB.GLM，因此只要对GLIM.BAT 进行适当的修改，可以使软件在任何工作盘上运行。

进入GLIM后，出现问号(?)提示，用户交互地以输入数据和指令。也可以先编好ASCII形式的命令文件，再读入执行。读取程序的语句是input/reinput，在该语句的后面要指定读入的通道号，通道号与DOS文件相联系。dinput 则用于读取data 语句中指示的量。return 或finish指令可以用于结束由input/reinput 或suspend引起的读取，skip 或exit 结束当前的栈也可以终止读取。

GLIM 4新增了manual命令，提供用户热线帮助。

【例13.1】例6.2问题的GLIM 程序如下：

```
$unit 8
$fact d 2 v 2 p 2
$data d v p count
$read
1 1 1 19  1 1 2 132
1 2 1  0  1 2 2   9
2 1 1 11  2 1 2  52
2 2 1  6  2 2 2  97
$yvar count
$erro poison
$fit d+v+p:+d.v:+v.p:+d.p $
$finish
```

GLIM 用$或 引导指令。很容易与其它软件类比，本例首先指定单元或记录个数、因素及水平，建立数据集。然后是模型部分，因本例是一个列联表分析，因变量是观察频数，误差为泊松分布。最后用fit命令把模型拟合出来，从主效应开始，依次增加交互项。

设程序存于文件loglin.glm，在GLIM中要运行它，使用以下命令：

$inp 7
File name? loglin.glm
$INP? $

运行指定通道号为7，第三行指示输入结束。

GLIM 作为计算器，如：

$calc 3.14159265/3 $

当运行结束时，使用命令$STop 返回至DOS系统下。

§13.1.4 GLIM 语言

GLIM 字符集：字母A—Z、数字0—9、符号+、-、*、/、**、()分别表示加、减、乘、除和括号，括号改变运算的优先级。特殊字符及其功能列表如下：

符号	名称	用途
$	美元符(dollar)	指令记号
:	冒号(colon)	重复符号
%	百分号(percent)	系统记号
#	井字号(hash)	替换符号
’	单引号(quote)	字串引号
&	与号(ampersand)	逻辑与
?	问号(query)	逻辑或
>	大于号(greater than)	大于号
<	小于号(less than)	小于号
_	下线(underline)	联接
[	左方括号(left-hand bracket)	左方括号
]	右方括号(right-hand bracket)	右方括号
;	分号(semi-colon)	定维记号
@	位置符号(at)	无效字符
|	竖线(modulus)	取模

变量的名字最多是四个字符，大小写字母意义相当。

令牌(tokens) 是输入字符的序列，其中包括指令名(directive names)、标识符(identifier)、值(values)、关键字(key words)、运算符号及以分隔符。由令牌可组成GLIM 的语句，一个语句由指令名及一系列项(item)组成，每个项都是令牌。GLIM 的一次运行(session) 是一套完整的语句，其定义是：[任务[$end 任务]] $stop，其中的任务是相关语句的组合，以end 结束。

标识符实际上可以指某种结构的数据或子文件(subfile)，有六种数据结构：

常量(scalar)	存单一的数
向量(vector)	存一列数
指针(pointer)	存向量名字
宏(macro)	存程序文本
函数(function)	影射实数
内部数据(internal)	系统变量值

常量分普通常量与系统常量，普通常量具有形式：%字母，故共有26 个，系统常量共有51 个。向量具有长度和水平数两种特征，用variate 和factor 指令产生的称为用户向量，与此对应的是系统向量，如：%fv 的长度由units 而定，存放的是fit 指令的拟合值。GLIM 的内部数据如%ssp，是平方和交叉乘积矩阵。

表达式：除了字符集中指示的以外，逻辑运算符：$<, <=, =, ==, /=, >=, >$ 以及&(AND)、?(OR)、/(NOT)，与一般高级语言相同。

GLIM 函数：由calculate 指令使用，或者由实参传至calculate 使用。

%ang(x)	方根的反正弦arcsin($\sqrt{x}$)
%exp(x)	自然指数exp(x)
%log(x)	自然对数ln(x)
%sin(x)	正弦函数sin(x)
%sqrt(x)	平方根函数$\sqrt{x}$
%np(x)	累积正态函数$\Phi(x)$
%nd	正态变量函数$N(x), N(\Phi(x)) = x; 0 < x < 1$
%tr	截尾函数
%gl(k,n)	产生一因素分组
%cu(x)	累积函数
%sr(0)/%sr(n)	标准伪随机函数，结果为$(0,1)$间的实数和$(0,n)$间的整数
%lr(0)/lr(n)	局部伪随机函数
%nd(sr(0))	随机正态偏差

函数%gl(最大值,重复数)根据指定的最大值进行若干次，很常用。

与广义线性模型相应，一个特定的模型可经误差的分布、线性部分构造和联系函数而确定。如对于正态误差分布，单位联系函数，是最简单的；列联表数据可看做来自poission分布，其联系函数为对数；量化的反应(quantal response) 结果r/n (n个对象中r个反应) 可视做具有二项分布，联系是probit 。通常可以使用GLIM默认的设置：误差为normal；连接函数为identity，尺度参量待估计；权为1，偏移量为零(偏移量出现于线性预报量中，没有参数，对每个观察有影响，如生存分析模型和稀释模型)，拟合为1 (常数项)。这些参数的关系可以组合成下表：

误差(error)	联系函数(link)	尺度参数
正态分布(Normal)	indentity	待估计
二项分布(Binomial)	logit	1
泊松分布(Poisson)	log	1
伽马分布(Gamma)	reciprocal	待估计

GLIM 用指令$link来指示联系函数，其后面的参数可以简写，如$link I 表示单位联系函数。除了上表给出的以外，还有S (平方根)、E(指数)、P(probit)、C(comp-log-log)。$是GLIM指令的分界符。

$error指令设定误差的分布形式，其参数也可以缩写，如：$error N 表示正态分布，除了上表给出的以外，还有G(gamma)即伽马分布。

尺度参数用指令$scale来设定。

现将GLIM 指令整理如下：

- $UNits n 表示标准向量的长度
- $DAta 变量名表表示待输入的变量名
- $Read 数据表读入数据

- $Yvariate 变量名指示因变量Y
- $Error 分布指示误差的类型
- $Fit 模型拟合模型

FIT 语句中效应的写法采用了Wilkinson与Rogers (1973) 的指定方法，这些记号如: $*, ., +, -, \backslash$的意义通常能从$DISP M 指令得到。

+ 表示效应的相加

. 表示简单交互

$*$ 表示交叉或层次交互

/ 表示嵌套

– 表示删除

因此，对于A,B,C三个效应的情况，可能有组合如：$A+B+A.B$、$A+B+A^2+B^2+A.B$。$A*B$与$A+B+A.B$等价，$A*B*C-A.B.C$与$A+B+C+A.B+B.C+A.C$等价，A/B与$A+A.B$等价，$(A+B)*C$与$A+B+C+A.C+B.C$等价，$(A+B)/C$与$A+B+A.B+A.B.C$等价。

GLIM 以美元符($) 做为语句分界符，如：

```
$units 18
$data freq
$read
15 11 14 17 5 11 10 4 8
10  7  9 11 3  6  1 1 4
$yvariate freq
$error p
$fit $
```

- $Display e r 显示拟合情况

这个语句使用重要，如使用M，L参数获得系统所分析的模型，E、A、U、V、C、S、T、R、W参数获得参数估计情况和残差，D则是离真度。

- $CAlculate 进行计算

$calculate mon=%gl(18,1) $fit mon
$display e r

- $INput 通道号文件读入程序文件
- $Argument 宏定义名参数表指示宏调用的信息
- $Use 宏定义名使用宏

- $Plot 纵横坐标画图

```
$calculate %a=50.84-%dv
$calculate %b=17-%df
$input 12 CHIT
$use CHIT %dv %df
$use CHIT %a %b
$plot %fv freq mon
$calculate f=%log(freq)
$calculate t=%log(%fv)
$plot t f mon
```

冒号(:) 用于命令的重复执行，如：calculate pw=1:x=2 $

- $STop 结束用于结束GLIM 的一次运行
- $FINish 指示程序文件结束
- $DINput 文件名通道号读数据文件
- $FOrmat Fortran 格式格式重定义 $data sex rev age
  $format (2x,f4.0,f1.0,5x,f2.0)
  $dinput 1
- $OUtput 文件名通道号结果写入外部文件
- $ECho 显示指示印出GLIM 所接受的所有信息
- $PRint 信息打印信息
- $LOok 常量/向量浏览常量/向量
- $Macro 宏定义名空格文本$End 宏的定义
- $ENVironment 代码获取系统信息

  它给出的信息包括C(通道分配)、D(数据)、E(外部PASS)、G(图形功能)、I(安装信息)、P(程序信息)、(随机数的种子)、S(系统)、U(可使用的空间)。
- $DElete 宏名表/变量取消宏定义/变量
- $EXTract 结构提供系统结构
- $Tabulate 变量;变量造表
- $TPrint 变量; 变量打印表的内容
- $End 结束一个分析
- $Assign 名=名,名=名变量赋值

- $Offset 向量引入一个偏移量
- $FActor 因素水平指示名义或因素变量
- $Variate 测量变量指示标识的长

不同的分析可以子文件的形式共存于文件中，即$subfile 文件名1,. . . , $subfile 文件名n $finish。GLIM 最多可以嵌套16 层，当前的层数存于系统量%CL 中，GLIM 每开始一项任务，其栈都重新初始化。GLIM 使用!引导注释。

控制指令有alias、cycle、recycle 指示控制拟合的计算。一次拟合的结果通过系统变量%fv、%lp、%wt、%wv、%dv、%
以及%dr、%va、%di 来观察。

指令tabulation, sort, look, tprint, print, plot, hist 进行向量制表、排序、按列或按表浏览、散点图和直方图，而最中心的内容是指定和拟合广义线性模型，对同一批数据指定不同的模型，增加或者减少包含的项。

输入/输出通道和宏调用：通道与DOS 的设备或文件联系，用$enviroment c 能观察到这些定义。

下例从6号通道读取外部文件test.dat数据矩阵：

```
$unit 9 $
$data x y $
$dinput 6 $ !若数据文件的宽度超于80列，使用$dinput 6 132 $
File name ? test.dat
$look x y $
```

若x与y的数据是先后次序排列，用以下的语句读取：

```
$unit 9 $
$data x $
$dinput 6
File name ? test.dat
$data y $
$dinput 6 $
$look x y $
```

第二次读取不需要继续给定文件名。

据READ.ME文件的提示，通道3指定为GLIM.LOG，通道5指定为MACLIB.GLM，如启用子文件TEST，使用命令：

```
? $INPUT 5 test $
```

屏幕显示：

```
****** Successful Macro Library Access *******
```

MACLIB.GLM 是以文本文件的形式提供给用户的宏义，其通道号存于系统变量

标识：GLIM 3.77 macro library, release 1.0, January 1985

子文件名	宏名	描述
数据描述与显示		
SUMM	SUMM	变量(variate)的综合统计量
STEM	STEM	茎叶图
SMOO	SMOO	变量(variate)的Tukey 平滑
统计工具		
CHIP	CHIP	χ^2概率
正态模型		
QPLOT	QPLOT	正态概率图
QPLOT	STAN	标准化残差
QPLOT	JACK	大折刀(Jackknife) 残差
TNOR	TNOR	使用适合度χ^2的正态性检验
TNOR	WDASH	Shapiro Francia W' 正态性检验
NORMAC	RSQ	R 平方统计量
NORMAC	TVAL	参数估计的t 值
LEV	LEV	杠杆值
BOXCOX	BOXCOX	关于y 变量的Box-Cox 转换族
BOXCOX	BOXFIT	固定λ的Box-Cox 转换
PRESS	PRESS	预测误差平方和
泊松模型和列联表		
二项分布模型		
伽马模型		
生存分析		
WEIB	WEIB	对于截尾数据拟合指数和威布尔分布
WEIB	RESP	使用WEIB宏后的残差图
其它		
TUNI	TUNI	变量是否为均匀分布的χ^2拟合度检验

MACLIB.GLM 说明了各个宏的入口参数和产出结果。

宏用于重复，专用的过程，循环和一些复杂的例行程序，如：

```
$macro n %nd($calc y= #n:...:z= #n $ !用于重复
$macro m :+a*b*c*d-a.b.c.d $endmac $
$fit #m $
```

可以借助于GLIM提供的宏功能进行专门的功能，如下面是一个结合系统变量进行残差图示的例子[5]。

```
$macro rplot$
$calc resid=%yv-%fv $
$sort resid $
$calc n=%cu(1) $
```

```
$calc norm=%nd((n-0.375)/(%(%nu+0.25)) $
$plot resid norm '*' $
$endmac $
$use rplot $
```

宏调用可以不限于一次，结合while指令可以进行多次调用，如：

$calc %a=1 $
$while %a update $
update 的结构是：
$macro update $
$calc %z1=%z1+1 $
... $calc %a=%if(%z1¡10,1,0) $
$endmac $

超于10次时停止，也可以根据条件进行切换和执行，如：

$calc %a=2 $
...
$switch %a one two thr$
$endmac $
$switch %a update $

根据%a的不同取值执行宏调用。
其它指令如：
$accuracy 4 $ 指定系统保留四位小数。
$calc x(8)=10.1 $ 改变x的第8个值。
$edit 2 3 x 0.2 0.1 $ 结果如同：$%calc x(2)=0.2:x(3)=0.1，其中的冒号表示重复最近一个指令。

对向量x进行排序只消使用指令：$sort x $!一个参数，按x的取值进行排序。
$sort y x $!两个参数，x次序不变，排序结果存于y。
$sort z y x $!三个参数，x不变，使用排序结果对y进行排序，结果存于z。
$units 10 $
$calc r=%sr(0)$
$sort s 1 r $

其好处是能够产生不重复的10个随机数。

$sort s 1 s $!记取排序的次序号。

产生滞后：

$assign a=3,9,4,6,5,1,8,2,10,7 $
$sort b a -2 $

结果是a的数据提前一行，因为最末一个数是无效的，故用以下语句：

$calc diff=b-a:wt=(%cu(1)/=10)$
$weight wt$
$look diff wt $
$sort b a 2 $!数组b含a的滞后值。
$look b a $

transrpit指令管理记录文件，很有用：

$units 9 $
$data x y $
$trans $
$dinput 8 $
File name: test.dat
$plot y x $
$trans i w f h o $
$plot y x $

先关闭记录，等结果满意后再存取。i，w，f，h，o 对应input(输入) ，warnin g messages(警告信息)，fault messages(错误信息)，help messages(帮助信息)，ordinary output(正常输出)。

当运行中途停止时，可以使用dump和restore指令保存和恢复现场。

【例13.2】以下程序说明了宏的用法[1]，CHIT 对给定的卡方检验算出概率水平，有χ^2值及自由度两个参量，结果是用GLIM.LOG给出的。

```
$macro CHIT
$calc %p=(%2==1)*(2-2*%np(%sqr(%1)))+(%2==2)*(%exp(-%1/2))
      +(%2>2)*(1-%np(((%1/%2)**(1/3)-1
      +2/(9*%2))/%sqr(2/(9*%2))))
$print 'CHI2 P='%p' for CHI2='%1' WITH '*-4%2 'd.f.';
$$endma
$return
$macro UCHI
   $use CHIT $calc %dv=%d-%dv:%df=%e-%df $use CHIT
$endma
$units 4 $data FREQ $read
72 714 655 41
$yvar FREQ $error p
$assign clas=1,-1,-1,1
$calc c1=2*(%gl(4,1)-2.5) : c2=(c1/2)**2-1.25
$fit $use CHIT %dv %df
$calc %d=%dv :%e=%df $disp er
$fit clas  $use UCHI $disp er
$fit c1+c2 $use UCHI $disp er
$finish
```

运行结果：

```
[o] scaled deviance = 1266.8 at cycle  4
[o]            d.f. =    3
[i] $calc %d=%dv :%e=%df $disp er
[o] CHI2 P=  0.     for CHI2=   1267. WITH    3.d.f.
[o]          estimate        s.e.     parameter
```

```
[o]     1       5.915     0.02594     1
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72    370.50    -15.508
[o]      2       714    370.50     17.846
[o]      3       655    370.50     14.780
[o]      4        41    370.50    -17.118
[o] scaled deviance = 11.158 at cycle  3
[o]            d.f. =  2
[o] CHI2 P=  0.0038 for CHI2=   11.16 WITH    2.d.f.
[o] CHI2 P=  0.     for CHI2=   1256. WITH    1.d.f.
[o]          estimate        s.e.     parameter
[o]     1       5.281     0.04892     1
[o]     2      -1.247     0.04892     CLAS
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72     56.50      2.062
[o]      2       714    684.50      1.128
[o]      3       655    684.50     -1.128
[o]      4        41     56.50     -2.062
[i] $fit c1+c2 $use UCHI
    $disp er [o] scaled deviance = 1.4458 at cycle  3
[o]            d.f. = 1
[o] CHI2 P=  0.2292 for CHI2=   1.446 WITH    1.d.f.
[o] CHI2 P=  0.     for CHI2=   1265. WITH    2.d.f.
[o]          estimate        s.e.     parameter
[o]     1       5.271     0.04937     1
[o]     2    -0.06409     0.02066     C1
[o]     3      -1.255     0.04922     C2
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72     67.23      0.582
[o]      2       714    728.31     -0.530
[o]      3       655    640.69      0.565
[o]      4        41     45.77     -0.705
```

【例13.3 】二分类数据分析常用probit、logit 和极值分布模型，这三者具有相同的模型形式：$\pi(x) = F(\alpha + \beta x)$。毒理实验中，许多毒物剂量对数的容许值(tolerance)分布通常近似正态分布，则$\pi(x) = \Phi[(x-\mu)/\sigma], \Phi(.)$ 是标准正态分布的累积分函数，故$F = \Phi, \alpha = -\mu/\sigma, \beta = 1/\sigma, \Phi^{-1}(\pi(x)) = \alpha + \beta x$, 即probit 模型；而对$\pi(x) = exp[-exp(\alpha + \beta x)]$，有$log[-log(\pi(x))] = \alpha + \beta x$，它与极值分布对应。$G(x) = exp(-exp[-(x-a)/b]), b > 0, -\infty < a < \infty$，均值为a+0.577b，标准差为$\pi_b/\sqrt{6}$。

下面是一个生物检测(bioassay) 的例子[2]，是一个甲壳虫接触气性二硫化炭 5 小时后的死亡情况，死亡与否是一个二分类数。程序依次对logit、probit、complementary log-log 计算估计值。

```
$c Fitting logit/probit/extreme-value models
$unit 8
$data dose kill number
$read
$1.691 6  59 1.724 13 60 1.755 18 62 1.784 28 56
$1.811 52 63 1.837 53 59 1.861 61 62 1.884 60 60
$yvar kill
$error bin number
$fit dose
$disp e r v$
$link p
$fit dose $
$disp e r v$
$link c
$fit dose $
$calc survive=number-kill $
$c the yvar is replaced with kill
$yvar kill
$fit dose $
$disp e r v$
$dele dose kill number survive$
$finish
```

link 语句中C 表示双对数联系，E 表示指数联系，G 表示logit 联系，I 表示单位联系，L 表示对数联系，P 表示probit 联系，R表示倒数联系，S表示方根联系。以上程序运行结果如下：

scaled deviance = 11.116 at cycle 4 d.f. = 6

scaled deviance = 9.987 at cycle 4 d.f. = 6

scaled deviance = 3.5143 at cycle 4 d.f. = 6

因为尺度参数均为1，则规格化deviance 与deviance 相同，同时还可以看出，三个模型以最后的complentary log-log 较佳。

估计值	标准误	参数的方差-协方差	
1 -60.74	5.181	26.84	
2 34.29	2.913	-15.09	8.484
1 -34.96	2.648	7.012	
2 19.74	1.487	-3.937	2.213
1 -39.52	3.234	10.46	
2 22.01	1.796	-5.806	3.226

拟合值：

unit	死亡数	总数	logit	probit	comp log-log
1	6	59	3.503	3.407	5.653
2	13	60	9.820	10.686	11.282
3	18	62	22.421	23.438	20.942
4	28	56	33.875	33.784	30.339
5	52	63	50.048	49.559	47.681
6	53	59	53.339	53.370	54.188
7	61	62	59.239	59.682	61.117
8	60	60	58.755	59.239	59.948

§13.2 广义线性模型简介

§13.2.1 一般理论

一个随机变量的统计模型隐含着这样一个思想：即被研究的变量有一个确定的结构，能够解释现有资料和进行预测。这也是广义线性模型的思想。它的三个组成部分是，随机部分、系统部分和联系部分。随机部分指示了因变量的概率分布，系统部分是自变量的线性函数，联系部分指明了系统部分与随机部分期望值间的关系。

广义线性模型的随机部分由指数族分布的独立观测组成，它们的分布形式为：

$$f(y,\theta,\phi)=\exp\{[y\theta-b(\theta)]/a(\phi)+c(y,\phi)\}$$

其中ϕ是尺度化参数或离散参数，在列联表分析中常常取值为1，当$\phi>1$ 时模型方差过大，称为过度离散。

如正态分布$\theta=\mu, b(\theta)=0.5\theta^2, a(\phi)=\sigma^2, c(y,\phi)=-0.5[log(2\pi\phi)+y^2/\phi]$。$\phi=1$ 时，上式可以化成如下的形式(Agresti, A. 1990)：

$f(y;\theta)=a(\theta)b(y)exp[yQ(\theta)], Q(\theta)$ 称为自然参数。

当$\eta=Q$时称作典型联系函数，很常用。

如对于logit 模型，有：

$f(y;\pi)=\pi^y(1-\pi)^{1-y}=(1-\pi)exp[ylog(\pi/(1-\pi)], Q(\pi)=log[\pi/(1-\pi)]$

表 13.1 几种分布所对应模型的各个部分

随机部分	联系函数	系统部分	模 型
正态分布	单位联系	连续的	回归
正态分布	单位联系	分类的	方差分析
正态分布	单位联系	混合的	协方差分析
伯努利分布	logit	混合的	logistic 回归
泊松分布	对数联系	混合的	对数线性模型
多项分布	广义的logit	混合的	多项反应模型

对于Poisson 模型，有：

$$f(n;m)=\exp(-m)(m^n)/n!=exp(-m)(1/n)exp(nlog(m)]$$

故其联系函数是log。与广义线性模型的定义$\mu=X\beta,\eta=g(\mu)$ 相比，有：$\eta=log(m)=X\beta$，是一个对数线性模型。列联表中的计数可以看成来自泊松分布。

一个观察y的标准线性模型是$y=X\beta+\varepsilon,\mu=X'\beta$构成系统部分，随机部分与系统部分的关联是$\eta=g(\mu),g$是$\mu$的任何单调可微函数。最常用的链接或联系函数为：

单位(identity) 联系 $\eta=\mu$

对数比数(logit) 联系 $\eta=ln[\mu/(1-\mu)],0<\mu<1$

概率单位(probit)联系 $eta=\Phi^{-1}(\mu),0<\mu<1,\Phi$ 为$N(0,1)$ 的分布函数

重对数(log-log) 联系 $\eta=ln[-ln(1-\mu)],0<\mu<1$

指数(power) 联系 $\eta=\begin{cases}\mu^r & r\neq 0\\ \log(\mu) & \text{r=0}\end{cases}$

指数分布族包括正态、伽马、泊松、二项、贝塔、负二项、卡方和逆正态分布等，它们关联的的分析模型列于下表。

广义线性模型的参数估计常用迭代加权最小二乘法。

表示模型拟合效果的量有偏差或离真度(deviance) 和广义Pearson χ^2统计量。设对给定模型的似然为$L(\hat{\mu};y)$，由数据而来的最大似然为$L(y;y)$，则规格化偏差$=2[L(y;y)]-L(\hat{\mu};y),\hat{\mu}=y$是估计值。

模型的检验常结合残差分析、数据诊断与图形分析等手段，类似通常的回归诊断问题。

许多统计软件可以进行广义线性模型分析，如GLIM、Genstat、S 以及SAS 6. 08 中的PROC GENMOD 等。第二章介绍了线性回归分析、logistic 回归和Cox回归，这里则对列联表的对数线性模型进行更详细的说明，模型中危险因子、混杂因子和反应均为离散性的数据，可整理成列联表格式。模型在SAS 、SPSS/PC+、BMDP、SYSTAT等软件包均可实现。

§13.2.2 列联表分析用例

我们知道$R\times C$表的独立性检验时，每个观察格子的的期望值边缘概率的积乘以总的格子数，将这一等式两取对数，就成为一个对数线性模型，即对数的线性和，联系函数也就是对数。根据列联表变量是有序或名义的类型，对应不同的对数线性模型。

最简单的是2×2 表，其比数比(odds ratio) 是$\theta = n_{11}n_{22}/n_{n_12}n_{21}$，$log(\theta)$ 的渐近标准误是$\sigma(log[\theta]) = \sqrt{\Sigma_i\Sigma_j 1/n_{ij}}, i,j = 1,2$。有时也在此式的每一个 n 值上加上0.5，标准误的公式也类似。两维列联表中对数线性模型参数λ_{12}可经$0.25ln(\theta)$而估计。

格子数为正值时的对数线性模型：$\log(m) = X\beta$。如在2×2表下，$\log(m_{ij}) = \mu + \lambda_i^x + \lambda_j^y$，约束为$\Sigma_i\lambda_i^x = \Sigma_j\lambda_j^y = 0$，此即：

$$\log\begin{pmatrix} m_{11}\\ m_{12}\\ m_{21}\\ m_{22}\end{pmatrix} = \begin{pmatrix} 1 & 1 & 1\\ 1 & 1 & -1\\ 1 & -1 & 1\\ 1 & -1 & -1\end{pmatrix}\begin{pmatrix}\mu\\ \lambda_1^x\\ \lambda_1^y\end{pmatrix}$$

如：$\log(m_{12}) = \mu + \lambda_1^x + \lambda_2^y = \mu + \lambda_1^x - \lambda_1^y$

对于独立Poisson模型

$$L(m) = \Sigma_i n_i\log(m_i) - \Sigma_i m_i = \Sigma_i n_i(\Sigma_j x_{ij}\beta_j) - \Sigma_i\exp(\Sigma_j x_{ij}\beta_j)$$

$$\frac{\partial L(m)}{\partial\beta_j} = \Sigma_i n_i x_{ij} - \Sigma_i m_i x_{ij}, X'n = X'\hat{m}$$

$$\frac{\partial^2 L(m)}{\partial\beta_j\beta_k} = -\Sigma x_{ij}\frac{\partial m_i}{\partial\beta_k} = -\Sigma_i x_{ij}x_{ik}m_i$$

设有k个2×2 表，设定各表的边缘合计为$\{n_{+1k}, n_{+2k}, n_{1+k}, n_{2+k}\}$ 时服从超几何分布，而且只用n_{11k}应能够确定$\{n_{+1k}, n_{+2k}, n_{1+k}, n_{2+k}\}$，

$$m_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$$

$$V(n_{11k}) = n_{+1k}n_{+2k}n_{+1k}, n_{+2k}/n_{++k}^2(n_{++k} - 1)$$

因有条件独立，它们可以相加，故有Mantel & Haenszel 统计量：

$$M^2 = (|\Sigma n_{11k} - \Sigma m_{11k}| - 0.5)^2/\Sigma V(n_{11k}) \sim \chi_{(1)}^2$$

假设$\{n_i, i = 1, ..., n\}$ 是一个列联表中的观察，n_i非负，其最简单的情形是泊松分布，方差与均值为m_i。它具有性质$n = \Sigma n_i$ 仍为泊松分布，参数为Σm_i。泊松分布用于时空上随机发生的事件数，如某地某年某月n_1 (自然流产数)、n_2 (引产数)、n_3 (活产数) 具有泊松分布。由于这样的泊松抽样是随机样本，若$n = \Sigma n_i$ 而每个n_i 以n 为条件，n_i 不再独立，它们服从多项分布。在流行病学前瞻性研究中，对应于研究因素各水平的边缘合计视为固定，而把各水平上反应变量视为独立的多项分布样本；在回顾性研究中，反应变量的各水平的边缘合计值视为固定而把研究因素的各个水平视为多项分布的样本；在横断面研究中，则可以把总的计数视为固定。

与析因分析类似，层次对数线性模型分析可以用于研究因素的交互，高层的交互影响隐含了低一级的效应，三维的饱和模型是：

$$log(m_{ijk}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz} \qquad (XYZ)$$

对独立模型，上式即是：

$$log(m_{ijk}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z \qquad (X, Y, Z)$$

相当于：

$$l(m) = n\mu + \Sigma_i n_{i++}\lambda_i^x + \Sigma_j n_{+j+}\lambda_j^y + \Sigma_k n_{++k}\lambda_k^z$$

其它形式的模型有:

$$\begin{aligned}
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} && (XY, Z)\\
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{jk}^{yz} && (XY, YZ)\\
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{jk}^{yz} + \lambda_{ik}^{zx} && (XY, YZ, XZ)
\end{aligned}$$

自由度的分解：$\Sigma\Sigma\Sigma\pi_{ijk} = 1$，即共有$IJK - 1$ 个线性无关参数。对独立模型，直接估计使用$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$，使用$\pi_{i++}\pi_{+j+}\pi_{++k}$共有$I - 1 + J - 1 + K - 1$ 个参数，故自由度为$(IJK - 1) - (I + J + K - 3) = IJK - I - J - K + 2$。一般的情况如下，第二列是自由度。

(X,Y,Z)	IJK-I-J-K+2
(XY,Z)	(k-1)(IJ-1)
(XZ,Y)	(J-1)(IK-1)
(YZ,Z)	(I-1)(JK-1)
(XY,YZ)	J(I-1)(K-1)
(XZ,YZ)	K(I-1)(J-1)
(XY,XZ)	I(J-1)(K-1)
(XY,XZ,YZ)	(I-1)(J-1)(K-1)
(XYZ)	0

模型的估计采用极大似然法(ML)和迭代比例拟合(IPF) 方法。SAS/IML 还有专门的IPF函数。相比于ML，IPF 总有解并收敛至极大似然解。模型的拟合优度检验可以用Pearson χ^2 和似然比统计量，它们的公式是：

$$\chi^2 = \Sigma_i\Sigma_j\Sigma_k(n_{ijk} - \hat{m}_{ijk})^2/\hat{m}_{ijk}$$

对(XYZ) 自由度=$IJK - I - J - K + 2$。对于2×2表，当n_{ij}较小时可用修正的公式。

$$G^2 = 2\Sigma_i\Sigma_j\Sigma_k n_{ijk} log(n_{ijk}/\hat{m}_{ijk})$$

利用方差稳定变换，使每个格子的分布基本接近标准正态分布，如Freeman- Turkey变换：$\sqrt{n} + \sqrt{n+1} - \sqrt{4m+1}$，它的平方和符合$\chi^2$分布，自由度与理论数$m$ 有关。

列联表分析，常要进行χ^2的分解，这时应当保证：①、分表自由度应与原表相同；②、原表中的每一个格子当且仅当在一个分表中；③、原表的边缘合计须为一个小表的边缘合计。在经验上,可以看一下分表的G^2 是否与总表相同相同。

以第 6 章表6.2死刑的资料分析为例。不计受害者的种族，执行死刑的白人为12%，黑人为10%，黑人较白人要低；但控制了受害者的民族，黑人要高。可见对其边缘合计表分析，会出现结论不一致的现象，称作Simpson's paradox。又如在不平衡的关于疗效的研究中，一种疗法可能对男性病人和女性病人都是好的，但对于所有病人就不一定好。

表 13.2 表6.2的边缘合计表

	是	否	小计
白人	19	141	160
黑人	17	149	166
总计	36	290	326

表 13.3 死刑例子的估计结果

模型	G^2	差值	自由度
(D,V,P)	137.93		4
		129.80	
(DV,P)	8.13		3
		6.25	
(DV,VP)	1.88		2
		1.18	
(DV,VP,DP)	0.70		1
(DVP)	0.00		0

利用对数线性模型分析得下表：

可见，(DV,VP)模型是可以接受的。

与尺度模型相比，有序资料分析具有优点：①、模型参数易于解释；②、名义模型会出现饱和模型而此时不会出现；③、检验利于寻找关联交互的类型。如两维列联表两个变量均为有序的线性关系(linear by linear association) 模型是：

$$\log(m_{ij}) = \mu + \lambda_i^y + \lambda_j^y + \beta u_i \upsilon_j$$

u_i, υ_j 是行列的分数，$\beta = 0$则为独立模型。

对于uniform association 仅有一个参数β，模型是：

$$\log(m_{hj}m_{ik}/m_{hk}m_{ij}) = \beta(u_i - u_h)(\upsilon_k - \upsilon_j), h < i, j < k$$

使用局部比数比$\theta_{ij} = m_{ij}m_{i+1,j+1}/m_{i,j+1}m_{i+1,j}$是有用的，对上述线性x线性模型有$\log(\theta_{ij}) = \beta(u_{i+1} - u_i)(\upsilon_{j+1} - \upsilon_j)$，在等距分数时，所有比数比是一样的，故称均匀关联。有序模型的一种构造方法是使用相邻两个概率的比。在线性x 线性关联下为：

$$log(\pi_{j+1|i}/\pi_{j|i}) = log(m_{i,j+1}/m_{ij}) = (\lambda_{j+1}^y - \lambda_j^y) + \beta(\upsilon_{j+1} - \upsilon_j)u_i$$

对于单位间隔$\{\upsilon_j\}$，可以简写为$\alpha_j + \beta u_i$。

表 13.4 工作满意度分析结果

模型	G^2	df	P
独立(I)	12.03	9	0.211
均匀关联($L \times L$)	2.39	8	0.967
条件独立($I\|L \times L$)	9.64	1	0.035

记$(\pi_1(x), ..., \pi_J(x))$ 是反应概率，相邻两组的logits 是$L_j = log[\pi_j(x)/\pi_{j+1}(x)], j = 1, ..., J-1$。若要拟合模型$L_j = \alpha_j + \beta' x$，可用关系式

$$L_j^\star = log[\pi_j(x)/\pi_J(x)] = \Sigma_{k=j}^{J-1}\alpha_k + \beta'(J-j)x = \alpha_j^\star + \beta' u_j, j = 1, ..., J-1$$

仍然用$\upsilon_1 \le \upsilon_2 \le ... \le \upsilon_j$ 做分数，模型$M(x) = \Sigma_j \upsilon_j \pi_j(x) = \alpha + \beta x$ 称作平均响应，表示了条件均值与自变量之间的线性关系。

使用有序反应资料的另一种方式是利用$F_j(x) = \pi_1(x) + ... + \pi_j(x), j = 1, ..., J$ 累积logit是$L_j = logit[F_j(x)] = log[F_j(x)/(1-F_j(x))]$，最简单的是$L_j(x) = \alpha_j, j = 1, , J-1, \alpha_j$ 称做断点，因为L_j是$F_j(x)$的增函数，所以α_j 是不减的，进一步，括入解释变量时，使用模型$L_j(x) = \alpha_j + \beta' x, j = 1, ..., J-1$。因为$L_j(x_1) - L_j(x_2) = \beta(x_1 - x_2)$ 从而称做比例比数比模型(proportional odds model)。对于单个自变量的情形，可以将函数用图表示出来，固定j时其图象类似于logistic 回归线。为了使$\beta_j > 0$有习惯上的解释，常常把模型写作$L_j(x) = \alpha_j - \beta' X, j = 1, ..., J-1$。相对于累积比数比模型和比例比数比模型，有累积链接模型(cumulative link model)：

$$\begin{aligned} G^{-1}[F_j(x)] &= \alpha_j - \beta' X \\ G^{-1}(u) &= \log(u/(1-u)) \\ G^{-1}(u) &= \log[-log(1-u)] \\ G^{-1}(u) &= \Phi^{-1}(n) \end{aligned}$$

等等。另外，常把多分类反应中的一个水平作为基线，如所有分类与最后一个水平比较。SAS CATMOD 提供了ALOGTS, CLOGITS, MEAN, LOGIT 几种选项来处理上述几种情况。

现对第 6 章工作满意度的资料进行分析，结果如下：

独立模型结果$G^2 = 12.03$, 自由度=9，关联是很弱的，但却忽略了“对工作满意的程度随工资增加”，使用均匀关联后，拟合得到改善$G^2 = 2.39$, 自由度=8，若用单位赋分，关联参数的估计为0.112(标准误0.036)，正值表示满意度随着工资的增加而增加，局部比数比是exp(0.112)=1.12，可信区间为exp(0.112±1.96*0.036)，即(1.04,1.20)，在端点的比数比则为exp(0.112(4-1)(4-1))=2.74 即高收入是低入的2.74倍。

对数线性模型与logit 的区别：在参数的解释方面，它不区分反应变量和原因变量，这会影响我们对于模型的选择；对于层次模型来说，高阶效应的存在就意味着组成它的低阶效应的存在。在存在反应变量时，对数线性模型相应于该反应的logit 模型，在反应量具有两种以上的分类时，与广义logit 模型相应。

工作满意度分析的程序如下：

```
$unit 16
$factor income 4 satisf 4
$data income satisf count
$read
 1 1 20  1 2 24 ... 1 4 82
 ...     ...    ... ...
 4 1  7  4 2 18 ... 4 4 92
$calculate uv=income+satisf$
$yvar count
$error pois
$fit income=satisf+uv$
$calculate v=satisf$
$fit income+satisf+income v$
$finish
```

现以精神损害(well，mild，moderate，impaired)与生活事件(X_1)及社会经济状况(X_2，1=high，0=low)的关系研究为例说明比例比数比模型(Agresti, A 1990)。

```
well 1 1  mild 1 5  moderate 0 0 impaired 1 8
well 1 9  mild 0 6  moderate 1 4 impaired 1 2
well 1 4  mild 1 3  moderate 0 3 impaired 1 7
well 1 3  mild 0 1  moderate 0 9 impaired 0 5
well 0 2  mild 1 8  moderate 1 6 impaired 0 4
well 1 0  mild 1 2  moderate 0 4 impaired 0 4
well 0 1  mild 0 5  moderate 0 3 impaired 1 8
well 1 3  mild 1 5               impaired 0 8
well 1 3  mild 1 9               impaired 0 9
well 1 7  mild 0 3
well 0 1  mild 1 3
well 0 2  mild 1 1
```

使用stata的ologit命令拟合模型：$L_j(a) = \alpha_j - \beta_1 X_1 - \beta_2 X_2, L_j$是累积分布取logit。

```
. label define aa 0 well 1 mild 2 moderate 3 impaired
. infile a:aa x1 x2 using table98.raw
. ologit a x1 x2,table
```

对数似然值为-49.55。

```
Ordered Logit Estimates                          Number of obs =     40
                                                 chi2(2)       =   9.94
                                                 Prob > chi2   = 0.0069
```

```
Log Likelihood = -49.548948                              Pseudo R2     = 0.0912

------------------------------------------------------------------------------
       a |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |  -1.111234   .6108775     -1.819   0.069      -2.308532     .086064
      x2 |   .3188611   .1209918      2.635   0.008       .0817216    .5560006
---------+--------------------------------------------------------------------
   _cut1 |  -.2819054   .6422652             (Ancillary parameters)
   _cut2 |   1.212789   .6606523
   _cut3 |   2.209368   .7209676
------------------------------------------------------------------------------
       a |      Probability            Observed
---------|--------------------------------------
    well |   Pr(      xb+u<_cut1)        0.3000
    mild |   Pr(_cut1<xb+u<_cut2)        0.3000
moderate |   Pr(_cut2<xb+u<_cut3)        0.1750
impaired |   Pr(_cut3<xb+u)              0.2250
    . ologitp p0-p3
```

使用label语句对对因变量反序编码，可以直接对回归的系数进行解释。当生活事件分数增加时精神损害程序加大，在高的社会经济水平下减低。给定生活事件分数下，低于任何水平的精神损害的比数比，在高的或低的社会经济状况下均为exp(1.111)=3.04倍。最后的ologitp给出了该模型下的概率预测值。

为了适应数据的特定结构，对数线性模型有很多特殊的处理方法。如配对资料的对称模型、Bradley-Terry 模型、准独立(Quasi-independent) 模型等，它们的含义和实现方法可参考Agresti, A.(1990) 和Lindsay, J.K.(1989)，BMDP 手册。

GLIM4 提供了许多新的功能，如包括了逆高斯(IG)分布以及指数和负二次连接。模型定义语句包括了正交多项式，用户自定义矩阵，以及连续变量的交叉乘积。这些功能由一系新的指令和函数来完成，如ELIMINATE 指令可以使分析和计算大大简化，用于配对病例对照研究、多项反应模型、Cox 比例风险模型。函数包括了卡方、t、F、贝塔、二项、泊松分布的概率和分位点，不完全伽马函数、对数伽马、双伽马、三伽马函数。与早期版本的Gauss-Jordan算法相比，新版本增加了Givens算法，后者更稳定和更精确。数据结构中增加了表类型，GRAPH 命令用于高分辨图形，GLIM4提供了宏编辑器。

GLIM4 的MANUAL命令提供了在线帮助功能，包括用例说明。

广义线性模型的理论意义决定了GLIM的应用价值。有关GLIM在随机过程中的用法可见[11]，其中介绍了Markov链、点过程和更新过程、生存曲线包括Cox 模型、生长曲线、时间序列、重复测量等方面的应用，并附有相应的GLIM程序。除了GLIM 以外，SAS 6.08 PROC GENMOD 和Genstat 5 也能进行广义线性模型分析，这些软件在应用时各有特色，只有对广义线性模型有一定的了解才能应用自如。

广义模型有许多推广，如Hastie和Tibshirani 提出的广义和模型(GAM) ，Liang 和Zeger 提出的广义估计模型(GEE)。S-Plus 的glm和gam 分别用于拟合广义线性模型和广义和模型。

广义线性模型的一般介绍及其推广可以参考有关文献，SAS 6.12 GENMOD可以进行GEE模型分析。

第十四章　LISREL

§14.1　LISREL 操作使用

LISREL 是线性结构方程模型的专用软件包，这里的介绍基于LISREL VI，它对结构模型的表达形式与LISREL IV、V 有所不同，它能自动计算初值，给出修正指数、增加模型的参数、结果绘图，可以自由格式读取数据。用五种方法(IV 、2SLS、OLS、GLS和ML法) 进行模型估计。结构方程模型是典型的确证型分析。

§14.1.1　PC LISREL 运行环境

LISREL 运行不需数学协处理器，但可以自动使用它，应用协处理器可以大大加快软件的运行速度。在没有协处理器的AT 机上运行时建议设置环境变量NO87 =FALSE。建议在配置文件CONFIG.SYS 中使用DEVICE=[d:][PATH]ANSI.SYS 和BREAK=ON 设置。PC LISREL VI 系统包括以下文件：

README.DOC	有关LISREL的信息
CONFIG.SYS	LISREL 系统配置文件
LISREL.BAT	执行LISREL 的批文件
PCLIS1.EXE	第一阶段LISREL 执行文件
PCLIS2.EXE	第二阶段LISREL 执行文件
INPUT	检验程序
DATA	检验数据
OUTPUT1	第一阶段检测输出
OUTPUT2	第二阶段检测输出
STACKED.BAT	运行批处理的样本程序
KLEIN-1.INP	Klein 美国经济模型，ULS 估计
KLEIN-2.INP	Klein 美国经济模型，ML 估计
KLEIN.DAT	Klein 模型数据
MTMM-CSA.INP	MTMM (Multi-Trait, Multi-Method) 心理学分析示例

辅助程序DISP.EXE 用于在屏幕上逐屏显示运行结果，FPRINT.EXE 用于在打印机上输出运行结果。DISP.EXE 与DOS中MORE.COM执行不同之处在于它滤掉了回车控制符。

STACKED.BAT 内容为：

```
ECHO OFF
PCLIS1 KLEIN-1.INP KLEIN-1.UT1 KLEIN-1.UT2
PCLIS2
PCLIS1 KLEIN-2.INP KLEIN-2.UT1 KLEIN-2.UT2
PCLIS2
PCLIS1 MTMM-CSA.INP MTMM-CSA.UT1 MTMM-CSA.UT2
PCLIS2
ECHO ON
```

可见PCLIS1.EXE 可以带有三个命令行参数，第一个为程序，第二个和第三个是生成模型文件的结果文件。ECHO OFF/ON 是DOS内部命令，用于批处理执行时屏幕显示的关闭与开启。

§14.1.2 进入系统

运行LISREL 批处理文件LISREL.BAT：

C:\LISREL>LISREL <Enter>

这时提示程序名、模型信息文件名和结果文件名。运行结束后，用浏览或打印程序显示和打印结果文件。

也可以直接在命令行上同时给出三个文件名，如：

C:\LISREL>LISREL PEER.INP FILE1 FILE2

表示利用程序PEER.INP 把运算结果放在FILE1 和FILE2 文件中。在软件提示运行程序名时直接打<Enter> 键则显示当前目录中的所有文件名。

使用命令C:\LISREL>LISREL INPUT OUT1 OUT2 <Enter> 产生结果文件OUT1 和OUT2 与OUTPUT1 和OUTPUT2 进行比较。

在IBM 机上的版本使用以下的运行格式：

//EXEC LISREL, SIZE=nnn 其中nnn 为占用空间的千字节数目。

§14.1.3 LISREL 建模过程

使用LISREL 一般有以下步骤：

. 构思通相互关联的结构或通径图；

. 根据结构，写出结构方程组；

. 收集资料；

. 利用收集的数据进行实算；

. 结果判断；

. 模型的选择；

. 实际意义的解释。

LISREL 的运行分为两个阶段，第一阶段读入并检查程序和数据，计算初值和待分析的矩阵。第一阶段的结果存入第一个结果文件，默认为OUTPUT1，利用它可以检查设定是否与期望相符。第二阶段接受第一阶段生成的二进制的中间文件并且进行其余的计算，若有输出，则隐含存入文件OUTPUT2中。微机系统LISREL大型计算机上的版本语句略有不同。

§14.1.4 LISREL 控制卡和用例

使用LISREL软件，首先要熟悉由希腊字母表示的矩阵记号，完整的程序应当指示分析的标题、数据、模型、输出，每种指示用两个字母做为关键字，称为控制卡，一个卡一行写不下时，可在参数的位置放以"C"做标记。对于每个分析，可以使用一个标题，使用的方法是在第80列上放非空的字符，则标题出现在LISREL每页的输出上，对应标题、数据、模型和输出。LISREL使用四种最基本的控制卡，以下加以介绍：

1. TItle 标题卡. 是LISREL 程序的第一卡。

2. DAta 数据卡. 指定LISREL 分析的数据。

DA 卡有参数NGroup=组数，默认为1；NInputvar=输入变量数，默认为0；NObs=样本大小，默认为100。在N 未知时，设定NO=0，程序自动计算N值；MA=矩阵类型：MM 矩统计量(moment) 矩阵、CM 协方差阵、KM 相关阵、AM 增广矩统计量矩阵(即交叉乘积矩阵)，默认为CM。

在DA 卡的后面，还可进一步使用四种信息，即卡的标号、数据格式、矩阵的形式以及变量与数据的对应关系。用LAbels 指定标号。数据格式与Fortran数据格式说明类似，如：(4F6.2)，* 表示自由格式。矩阵形式有：RA 原始数据阵、MM 矩的矩阵、CM 协方差阵、KM 相关阵、FU 满元矩阵、SY 对称阵、ME 均值、SD 标准差。每个卡均可以带上UNIT=n,FOmat=ON REwind的三个参数。给隐式变量的标号是LK 与LE 卡，相应于ξ与η量。数据顺序由SE 卡指示，如：SE 1 3 4/ 表示数据顺序是1、3、4，其中的反斜杠表示舍弃后面的量。

3. MOdel 模型卡. 指定LISREL 分析的模型，内容包括：模型中观察变量(X 与Y) 的数目，模型中的隐含变量数目(ξ与η)，每个待分析矩阵的形式(完整矩阵、对角阵、三角阵、单位阵、零矩阵)、每个矩阵估计的状态(如各元素取值，是固定还是不固定)。仅不分析LISREL模型时使MO 卡缺失，此时仅给出待分析的矩阵。

MO 卡可带有NY,NX,NE,NK量指示y, x, ξ及η的数目p, q, m, n。

4. OUput 输出卡. 指定LISREL 的输出。

OU 卡是针对五种方法的，其组合为：
IV 仅计算辅助变量(IV) 估计；
TS 两步最小二乘(TSLS) 估计，类似于IO(Initital only 仅仅初值)；
UL 给出IV 及无约束最小二乘(ULS) 估计；
GL 给出TSLS 与广义最小二乘(GLS) 估计；
ML 给出TSLS 与ML 估计。

未加指示时用ML 方法，指示方法很多时使用最未一个。与之有关的是NS 卡，指示用给定的初值开始最速下降法的计算。其它的参数有：

PT 打印技术方面的输出(Technical Output)；
SE 打印标准误；
TV 打印T-值；
PC 打印估计量的相关；
RS 打印与协方差阵有关的信息、正态化残差及$Q--Q$图；
EF 打印总的效应；
VA 打印方差与协方差；
MR 作用等价于RS、EF 与VA；
MI 打印修正指数；
FS 打印因子分数的回归；
FD 打印一阶导数；
SS 打印标准解；
AL 给出所有输出；

NS 不用自动初值
TO 改动默认的132 字符行为80 字符；
ND 输出的小数位数(0-8)，默认为3。
TM 解此问题的最大CPU 秒数，默认为60；

OU 卡应为求解问题时的最后一个卡，其打印输出采用了Fortran 语言对打印机的控制方法。

软件运行时的错误信息可由多种原因引起[3]：使用的关键字与LISREL 的习惯不符，省略了必需的斜杠、等号、逗号或空格，括号和引号不匹配，TI、DA 、MO 和OU 卡的顺序写错，输入相关阵漏掉了1，使用了小写，初值与实际值相差太远，在使用自己的初值时在OU中遗漏了NS。

LISREL 用几种手段进行适合度的评价，注意的量是：参数估计值、标准误(ML)、复相关的平方、决定系数、参数估计的相关(ML)。其中出现不合理的值则提示模型根本上是错误的，如方差为负、相关大于1、方差或协方差阵非正定、复相关为负、估计量的标准误太大，参数值高度相关等。当所有观察为多元正态分布、分析使用样本协方差阵和样本量足够大时，可使用程序提供的χ^2检验，它是一个似然比统计量，此值对于样本含量的大小是敏感的，模型假设(线性假设、共线性和可加性) 不合理也是常见的原因。另外可以使用拟合度指数(GFI)、调整拟合度指数(AGFI) 和剩余平方和(RMR)。GFI表示模型所解释的方差协方差的相对大小，AGFI 与GFI 的区别仅仅在于后者调整了模型的自由度。RMR 是样本协方差阵与假设协方差阵元素间平均的差异，因此对于拟合较好的模型来说，此值应该很小，如小于0.05。利用这些指标可以比较针对于同一批资料两个模型的好坏。

LISREL 给出了一些指标来评价参数的取值。第一项是t 值，一般说来，$t > 2$ 则认为显著，若t 值不显著，则参数可在以后的拟合中设为固定。第二项指标是标化残差，它是样本协方差阵和假设协方差阵的差，大于 2 的值指示模型指示的问题。为了进一步研究模型的拟合，LISREL 提供了$Q--Q$ 图，它是标化残差的图，在残差几乎重合于45° 的线时表示拟合效果较好。最后，针对每个固定参数，软件给出了修正指数(MI)，它表示一个固定参数在松驰时期望使模型χ^2值下降多少，是拟合函数关于固定和约束参数的导数。对于自由参数，修正指数为零。修正指数可与自由度为1 的卡方比较，若此值比较大，则表明松驰此参数可以使拟合最大限度地得到改善，修正的方法通常的逐个参数地进行而且参数的松驰应该有意义。使用修正指数来选择模型并不是一个值得推荐的方法。

现把模型指示综合列于下表：
模型参数可能的取值列于下表：
其中：
ZE= 0 是零矩阵
ID= I 是单位阵
IZ=[I 0] 或[I 0]’ 为分块单位阵与零
ZI=[0 I] 或[0 I]’ 为分块零与单位阵
DI=对角阵
SD=子对角阵, 即对角元为0 的下三角阵
SY=非对角的对称阵
ST=对称阵,对角元为1的相关阵
FU=方阵或不对称方阵

表 14.1 LISREL 常见的几种模型及指示方法

模型	指示	默认	参数
$x = \Lambda_x \xi + \delta$	NX,NK	NY,NE	$\Lambda_x, \Phi, \Theta_\delta$
$y = By + \Gamma x + \zeta$	NY,NX	NE,NK	B, Γ, Ψ
$y = \Gamma x + \zeta$	NY,NX	NE,NK	$B = 0$
$y = By + \zeta$	NY	NX,NE,NK	B, Ψ
$y = \Lambda_y (1 - B)^{-1} (\Gamma \xi + \zeta) + \varepsilon$	NY,NE,NK	NX	$\Lambda_y, B, \Gamma, \Phi, \Psi, \Theta_\varepsilon$
$y = \Lambda_y (\Gamma \xi + \zeta) + \varepsilon$	NX,NE,NK	NX	$B = 0$
$y = \Lambda_y (1 - B)^{-1} \zeta + \varepsilon$	NY,NE	NX,NK	$\Lambda_y, B, \Psi, \Theta_\varepsilon$
$y = \Lambda_y \zeta + \varepsilon$	NY,NE	NX,NK	$B = 0$

表 14.2 LISREL 各种模型参数可能的取值

名称	LISREL 记号	LISREL 名称	阶数	可能的形式	默认值	默认为固定或自由
Lambda-y	Λ_y	LY	NY,NE	ID,IZ,ZI,DI,FU	FU	FI
Lambda-x	Λ_x	LX	NX,NK	ID,IZ,ZI,DI,FU	FU	FI
BEta	B	BE	NE,NE	ZE,SD,FU	ZE	FI
GAmma	Γ	GA	NE,NK	ID,IZ,ZI,DI,FU	FU	FR
PHi	Φ	PH	NK,NK	ID,DI,SY,ST	SY	FR
PSi	Ψ	PS	NE,NE	ZE,DI,SY	SY	FR
Theta-epsilon	Θ_ε	TE	NY,NY	ZE,DI,SY	DI	FR
Theta-Delta	Θ_δ	TD	NX,NX	ZE,DI,SY	DI	FR

施加约束的方法：如：FR PS(1)、PS(2) PS(6-8) 可指示第一个、第二个及第六至第八个为自由参数。还可以结合相等条件的约束，见EQ 卡。

对于TE 与TD 指示，默认是对角和自由的。TD=SY 示对称阵、对角线自由，否则固定，TE=SY,FR,TD=SY,FR 示整个阵是自由的。

EQ 卡指示相等的约束

VA 卡与ST卡用于给出初值，如：ST 0.5 BE(1,1)-BE(3,3)、ST 0.5 ALL。

MA 卡即矩阵卡，用自由格式读入时可用斜杠中止数据读取。

【例14.1】Blau and Duncan(1967) 资料，有关的说明见第二节。设数据存放在文件BLAU.DAT中，其排列为下三角阵：

```
1.000
0.516  1.000
0.453  0.438  1.000
0.332  0.417  0.538  1.000
0.322  0.405  0.596  0.541  1.000
```

现用以下程序调用这个数据进行分析：

```
Model for Blau and Duncan Stratification Data
DA NI=5 NO=20700
LA
*
'X1' 'X2' 'Y3' 'Y4' 'Y5'
KM SY FILE=Blau.DAT
SE
3 4 5 1 2
MO NX=2 NY=3 BE=SD PS=DI
FI GA 2 1 GA 3 1
OU TV RS MI FS EF
```

程序指定标题为Model for Blau and Duncan Stratification Data，数据卡指明共有五个输入变量，观察例数为20700。变量标号格式为自由格式，记为通常的x_1, x_2, y_1, y_2，读取相关阵的数据，文件名为blau.dat。变量顺序是3 , 4, 5, 1, 2，模型指定 B 矩阵是下三角阵Ψ阵为对角阵。Γ阵的两个元素应指明为固定的，分析输出内生变量及外变量影响的总效应等。

极大似然法计算结果表明，各量的符号正如期望，t-值均有显著性，从修正指数来看，Y_4与X_1间的修正指数最大，但同时它的正态化残差也最大。

【例14.2】Duncan, Haller, and Portes(1968) 抱负的影响数据，相关阵为：

```
Respondent
x2 1.
x1 .1839 1.
x3 .2220  .0489 1.
y1 .4105  .2137  .3240 1.
```

```
y2 .4043  .2742  .4047  .6247 1.
Best friend
x5 .3355  .0782  .2302  .2995  .2863 1.
y6 .1021  .1147  .0931  .0760  .0702  .2087 1.
x4 .1861  .0186  .2707  .2930  .2407  .2950 -.0438 1.
y4 .2598  .0839  .2786  .4216  .3275  .5007  .1988  .3607 1.
y3 .2903  .1124  .3054  .3269  .3669  .5191  .2784  .4105  .6404 1.
```

设放于文件PEER.DAT 中，分析程序为：

```
PEER INFLUENCES ON AMBITION(MODEL 2)
DA NI=10 NO=329
LA
*
'x2' 'x1' 'x3' 'y1' 'y2' 'x5' 'x6' 'x4' 'y4' 'y3'
KM SY FILE=PEER.DAT
SE
 4 5 10 9 2 1 3 8 6 7
MO NY=4 NE=2 NX=6 FIXEDX BE=FU
FR LY 2 1 BE 1 2
FI GA 1 5 GA 1 6 GA 2 1 GA 2 2
ST 1.0 LY 1 1 LY 4 2
EQ BE 1 2 BE 2 1
EQ LY 2 1 LY 3 2
EQ GA 1 1 GA 2 6
EQ GA 1 2 GA 2 5
EQ GA 1 3 GA 2 4
EQ GA 1 4 GA 2 3
EQ PS 1 1 PS 2 2
EQ TE 1 1 TE 4 4
EQ TE 2 2 TE 3 3
OU SE TV EF SS
```

输入变量数为10, 4 个为Y 变量，6 个为X 变量，观察数为329，结果要求输出标准误、T 值和总效应、标化解。

```
    Y1         Y2         Y3         Y4         X1         X2
    --------   --------   --------   --------   --------   --------
Y1     1.000
Y2      .625      1.000
Y3      .327       .367      1.000
Y4      .422       .328       .640      1.000
X1      .214       .274       .112       .084      1.000
```

```
X2      .411       .404       .290       .260       .184      1.000
X3      .324       .405       .305       .279       .049       .222
X4      .293       .241       .411       .361       .019       .186
X5      .300       .286       .519       .501       .078       .336
X6      .076       .070       .278       .199       .115       .102

      X3         X4         X5         X6
    --------   --------   --------   --------
X3     1.000
X4      .271      1.000
X5      .230       .295      1.000
X6      .093      -.044       .209      1.000
    DETERMINANT =  .700505D-01
```

因为原始相关矩阵中变量顺序的需要变化，使用卡片SE 把它们调整过来。软件给出相关阵的行列式值为‘700505D-01。在LISREL 的第一个输出文件中包括了待求解的参数，在$\Lambda_y, B, A, \Psi, \Theta$阵相应的位置指示出来，本例$\Phi$是固定的。

本例中使用两阶段最小二乘法(TSLS) 进行初值估计。

最后的极大似然估计下Y 的总决定系数为.937，结构方程复相关系数的平方为：.564和.571。结构方程总的决定系数为.729。模型χ^2值为30. 63, 自由度为24,P=0.165，拟合度指数为0.982，调整拟合度指数为0.900，均方误差为0.023。在设定Ψ为对角矩阵、$i_{21}=0, \beta_1=\beta_2$ 的条件下，$\chi^2=26.90$，自由度为17，$p=0.06$。

【例14.3】Klein 宏观经济学模型的ULS程序：

```
KLEIN'S MODEL I ESTIMATED BY IV AND ULS
DA NI=15 NO=21
LA file=klein.dat
RA file=klein.dat
SE
1 4 3 10 14 11 13 12 7 8 15 9 2 5 6
MO NY=8 NX=7 BE=FU GA=FI PS=FI
FR BE(1,4) BE(1,7) BE(2,4) BE(3,8)
FR GA(1,5) GA(2,5) GA(2,6) GA(3,4) GA(3,7)
FR PS(1,1)-PS(3,3)
VA 1 BE(4,5) BE(5,1) BE(5,2) BE(6,2) BE(7,3) BE(8,5) C
GA(5,3) GA(6,6) GA(7,1) GA(8,2)
VA -1 BE(4,7) GA(5,2) GA(8,1)
OU UL
```

Σ矩阵在使用使用所有十五个参数时是奇异的，因此ML法不可行。由于五个y变量是冗余的，去掉后估计可行。

```
KLEIN'S MODEL I ESTIMATED BY ML
DA NI=15 NO=21
```

表 14.3 Klein 氏I 类模型估计结果

参数	TSLS	ULS	ML
a_1	0.02	0.04	-0.23
a_2	0.22	0.21	0.38
a_3	0.81	0.81	0.80
b_1	0.15	0.05	-0.80
b_2	0.62	0.68	1.05
b_3	-0.16	-0.17	-0.15
c_1	0.44	0.39	0.24
c_2	0.15	0.20	0.28
c_3	0.13	0.15	0.23

```
LA file=klein.dat
RA file=klein.dat
SE
1 4 3 7 8 15 9 2 5 6 /
MO NY=3 NE=8 NX=7 FI LY=IZ BE=FU GA=FI PS=FI TE=ZE
FR BE(1,4) BE(1,7) BE(2,4) BE(3,8)
FR GA(1,5) GA(2,5) GA(2,6) GA(3,4) GA(3,7)
FR PS(1,1)-PS(3,3)
VA 1 BE(4,5) BE(5,1) BE(5,2) BE(6,2) BE(7,3) BE(8,5) C
GA(5,3) GA(6,6) GA(7,1) GA(8,2)
VA -1 BE(4,7) GA(5,2) GA(8,1)
ST 5 PS(1,1) PS(2,2) PS(3,3)
OU NS
```

程序由文件KLEIN.DAT文件读取标号和数据(Fortran格式)，用到了上述各种卡控制。运行结果见下表：

§14.2 线性结构方程模型简介

单方程的线性模型和和多元线性模型考虑了多个自变量间的相关，但是它们将各个自变量平均地加以处理。线性结构方程模型(LInear Structural RELationship Model) 则是处理多个变量相互关联结构的一种方法，它能对观察相关及就量间的相互影响给予解释，模型已广泛用于计量经济学、心理学和社会学等学科。该方法常结合通径分析(path analysis) 和确证型因子分析(confirmatory factor analysis) 见于标准的统计专著中。

LISREL 模型由Joreskog 于1973年引入，模型归结为以下三个方程：结构子模型： $\eta = B\eta + \Gamma\xi + \zeta$

y测量子模型： $y = \Lambda_y \eta + \varepsilon$

x测量子模型： $x = \Lambda_x \xi + \delta$

结构模型：

$B_{(n\times n)}$ 是关联内生变量的回归矩阵

$\Gamma_{(m\times n)}$ 是关联m 个外生变量与n 个内生变量的回归矩阵

$\Phi_{(m\times m)}$ 是m 个外生变量ξ的对称方差协方差阵

$\Psi_{(n\times n)}$ 是n 个内生变量剩余ζ的对称方差协方差阵

$\eta_{(n\times 1)}$ 是内生变量向量

$\xi_{(m\times 1)}$ 是外生变量向量，$\xi \sim N(0,\Phi)$

$\zeta_{(n\times 1)}$ 是剩余向量，$\zeta \sim N(0,\Psi)$

测量模型：

$\Lambda_{x(p\times m)}$ 是外生变量与p 个测量变量间的回归矩阵

$\Lambda_{y(q\times n)}$ 是内生变量与q 个测量变量间的回归矩阵

$\Theta_{\delta(p\times p)}$ 是外生变量测量误差间的对称方差协方差阵

$\Theta_{\varepsilon(q\times q)}$ 是内生变量测量误差间的对称方差协方差阵

假设是：ζ与ξ不相关

ε与η不相关

δ与ξ不相关

ζ,ε与δ互不相关

B的对角元是零，但$I-B$非奇异

结构方程指示隐含变量间的关系，测量模型把隐含量同其示性量关联起来，结构方程模型的假设通常有两大类：一种是因果联系的假设，一种是关于误差的分布。模型分析所用变量一般是标准化的，但在说明时常常不专门区分。

模型有三类变量，内生的(endogenous)、外生的(exogenous) 量以及扰动量(disturbance)。上式中 y 为内生变量，是由模型内确定。 x 为外生变量，是事先给定的。内生变量可受其它内生变量、外生变量和扰动量的影响，内生量在模型中可作为原因出现，但不作为结果出现。除了这种区分外，模型变量还可区分为可测的量与不可测的量。表示模型的结构关系，还可以用通径分析那样的记号，表达变量间的关系。如□表示可测量、→表示变量间的作用、〇表示不可测量，等等。不同的误差或扰动(disturbances)不相关即它们之间没有双箭头，因果结构是单向的，则模型是可逆归的(recursive)。用矩阵的记号，没有双箭头意味着误差的协方差阵是对角的，作用是单向的表明 B 矩阵是下三角阵。把这个条件稍微放松一些，若模型不符合逆归模型的条件，但是对于内生变量和误差的子集之间存在，这时称做块逆归的(block-recursive)。

模型一旦被指定，就涉及参数的估计问题，也称做识别问题(identification problem)。在单一方程线性模型中，若设计矩阵是满秩的，模型的一般假设就能保证其参数被估计；结构方程模型中一些分布的假设一般来说不能保证模型能够识别，这时要施加一些额外的限制。这种限制采取的形式通常有两种：一是指定一些结构上为零，这与对数线性模型有些类似；另外就是对误差协方差阵的限制。一个参数若能估计就称为识别(identified)，否则就是不能识别(underidentified) 或过识别(overidentified)，过识别是指一个参数能有几个解，只有当一个参数恰好有一个解时，才是恰好识别(exactly identified)。模型识别的条件是待估计参数的数目不能超于已知量的数目。

有许多方法能用于结构方程模型的识别，LISREL 提供了工具变量法、两阶段最二乘法等几种估计方法，从模型正态误差出发进行模型参数估计在计量经济学中常称为充分信息极大似然估计(FIML)，LISREL 模型指定时已限定某些参数为0 或1，所以是一个约束最优化

方法，往往要采用数值方法。

记F是目标函数，k 是观测x,y变量数，d是模型自由度，t是被估计的独立参数数目，有关公式如下：

$$\begin{aligned}
\text{F(ULS)}= &\quad 0.5tr[(S-\Sigma)^2]\\
\text{GFI}= &\quad 1-tr(S-\Sigma)^2/tr(S^2)\\
\text{F(ML)}= &\quad \log|\Sigma|+tr(S\Sigma^{-1})-log|S|-(p+q)\\
\text{GFI}= &\quad 1-tr(\hat{\Sigma}^{-1}S-I)^2/(\hat{\Sigma}^{-1}S)^2\\
\text{AGFI}= &\quad 1-[k(k+1)/2d](1-GFI)\\
\text{RMR}= &\quad \sqrt{2\Sigma_{i=1}^{k}\Sigma_{j=1}^{i}(s_{ij}-\hat{\sigma}_{ij})^2/k(k+1)}\\
\text{AIC}= &\quad \chi^2+2t\\
\text{CAIC}= &\quad \chi^2+(1+\log N)t\\
\text{ECVI}= &\quad \chi^2/(N-1)+2t/(N-1)
\end{aligned}$$

RMR是均方误差。

模型χ^2值是最小目标函数的拟合值$\times(N-1)$，自由度取作$0.5(p+q)(p+q-1)-t$或$0.5k(k+1)-t$。

据R.A. Johnson(1988)，模型模型的拟合战略有：1.如果可能，用几种准则估计模型中的参数，然后注意：各量的符号与大小是一致的吗？所有的方差估计为正吗？剩余矩阵是否对称？2.利用协方差阵和相关阵分别进行计算，标化观察变量后，结果有何影响；3.把一个大的数据集分成两部分，分别重复 1.、2.做法，然后观察两部分结果之间以及与全部数据之间的稳定性。

LISREL 方法的优点在于它能给出变量作用的间接效应、把变量对间统计关系区分成因果和非因果的组分，这一点与通径分析有相同之处，除了与通径分析的关联外，与因子分析也有密切的关系，LISREL 进行的因子分析模型通常称作确证型因子分析与探索性因子分析(Exploratory Factor Analysis) 相区分。因为后者并不是先验地假定变量间有给定的关系存在。

另外要注意的是，结构方程模型虽用于描述多个变量关联及相互作用的大小，但有时称作Causal Model 并不恰当，因为许多研究旨在发现因果关联，很少是把简单的相关或预测作为目的。同时，进行因果推断时，同样存在着固有的缺陷[2]。

Blau and Duncan stratification 模型[2,3] (观察例数为20,700)

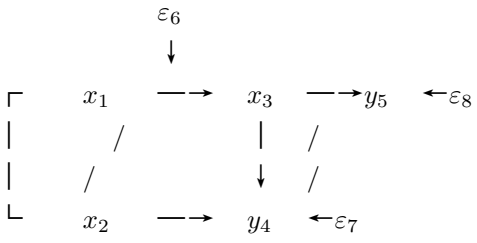


图12.1 递归结构方程模型

变量含义：x_1-父亲的教育，x_2-父亲的职业，y_3-子女的教育，y_4-子女的第一个工作，y_5-子女在1962 年的职业。

$$
\begin{aligned}
y_3 = &\quad \gamma_{31}x_1 + \gamma_{32}x_2 + \varepsilon_6 \\
y_4 = &\quad \beta_{43}y_3 + \gamma_{42}x_2 + \varepsilon_7 \\
y_5 = &\quad \beta_{53}y_3 + \beta_{54}y_4 + \gamma_{52}x_2 + \varepsilon_8
\end{aligned}
$$

比较一般模型$y = By + \Gamma x + E$，有：

$$
B = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{43} & 0 & 0 \\ \beta_{53} & \beta_{54} & 0 \end{pmatrix}
\Gamma = \begin{pmatrix} \gamma_{31} & \gamma_{32} \\ 0 & \gamma_{42} \\ 0 & \gamma_{52} \end{pmatrix}
E = Diag(\varepsilon_6 \quad \varepsilon_7 \quad \varepsilon_8)
$$

现在，观察相关矩阵为：

```
1.000
0.516  1.000
0.453  0.438  1.000
0.332  0.417  0.538  1.000
0.322  0.405  0.596  0.541  1.000
```

极大似然法结果：

$$
\begin{aligned}
y_3 = &\quad 0.309x_1 + 0.278x_2 + 0.738\varepsilon_6 \\
y_4 = &\quad 0.224x_2 + 0.440y_3 + 0.670\varepsilon_7 \\
y_5 = &\quad 0.115x_2 + 0.395y_4 + 0.281y_4 + 0.566\varepsilon_8
\end{aligned}
$$

Duncan、Haller、Portes 的peer influence 模型[2,3]. 内生变量具有多个示性量并假设外生变量度量无误差，相互作用关系图示如下：

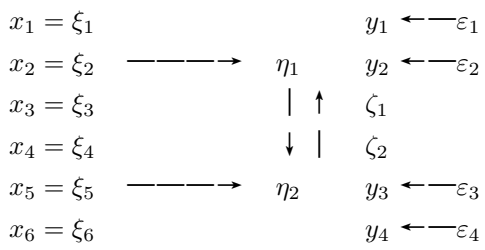


图12.2 含有隐含变量(或伪变量、哑变量)的模型

变量含义：

x_1-应答者父母的抱负，x_2-应答者的智力

x_3-应答者的社会经济状况

x_4-最好朋友的社会经济状况，x_5-最好朋友的智力

x_6-最好朋友父母的就业愿望

y_1-应答者的就业愿望，y_2-应答者的教育愿望

y_3-最好朋友的智力教育愿望，y_4-最好朋友的就业愿望
η_1-应答者的雄心，η_2-最好朋友的雄心

外生量的关系为：$x=\xi$ (即$\Lambda_x=I_6,\delta=0),\Phi=\Sigma_{xx},\theta=0$

y测量子模型是：

$$\begin{pmatrix}y_1\\y_2\\y_3\\y_4\end{pmatrix}=\begin{pmatrix}1&0\\\lambda_{21}&0\\0&1\\0&\lambda_{42}\end{pmatrix}\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\end{pmatrix}$$

$\Theta_{\varepsilon 4\times 4}=diag[\theta_{11}\quad\theta_{22}\quad\theta_{33}\quad\theta_{44}]$ 即内生变量的度量误差不相关。

结构子模型为：

$$\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}=\begin{pmatrix}0&\beta_{12}\\\beta_{21}&0\end{pmatrix}\begin{pmatrix}\eta_1\\\eta_2\end{pmatrix}+\begin{pmatrix}\gamma_{11}&\gamma_{12}&\gamma_{13}&\gamma_{14}&0&0\\0&0&\gamma_{23}&\gamma_{24}&\gamma_{25}&\gamma_{26}\end{pmatrix}\begin{pmatrix}\xi_1\\\xi_2\\\xi_3\\\xi_4\end{pmatrix}+\begin{pmatrix}\zeta_1\\\zeta_2\end{pmatrix}$$

现在，$[yx]'\sim N(0,\Sigma)$，模型求解的结果是：

$$\hat{\Lambda}_y^\star=\begin{pmatrix}.7667&0\\.8148&0\\0&.8299\\0&.7716\end{pmatrix}\quad\hat{\Theta}^\star_{\varepsilon(4\times4)}=diag[\begin{matrix}.4121&.3361&.3112&.4046\\(.0512)&(.0521)&(.0459)&(.0462)\end{matrix}]$$

$$\hat{B}_y^\star=\begin{pmatrix}0&-0.1994\\&(.1027)\\-0.2176&0\\(.1103)&\end{pmatrix}\quad\hat{\Gamma}^\star=\begin{pmatrix}.2103&.3256&.2848&.0937&0&0\\(.0506)&(.0574)&(.0576)&(.0648)&&\\0&0&.0746&.2758&.4205&.1922\\&&(.0624)&(.0532)&(.0546)&(.0468)\end{pmatrix}$$

可见η_1与η_2的许多解是可比的。

Klein 关于美国经济的 I 类模型是一个经典计量经济学模型，LISREL 讨论了它的实现方法，其它的如Greene, W.H.(1990) 也做了介绍，它是根据两次世界大战间美国每年的经济数据而构造的八个方程的方程组。数据是一个时间序列，所以模型是动态的。模型有三个行为方程：

$C_t=$	$a_1P_t+a_2P_{t-1}+a_3W_t$	集团消费
$I_t=$	$b_1P_t+b_2P_{t-1}+b_3K_{t-1}$	净投资
$W_{t*}=$	$c_1E_t+c_2E_{t-1}+c_3A_t$	个人工资
五个量	$P_t=Y_t-W_t$	总利润
	$Y_t=C_t+I_t+G_t-T_t$	总收入
	$K_t=K_{t-1}+I_t$	年末资本存量
	$W_t=W_{t*}+W_{t**}$	总工资
	$E_t=Y_t+T_t-W_{t**}$	私有工业总产值

左边是内生变量，含义在式子右端给出。$C_t,I_t,W_t,P_t,Y_t,K_t,W_t,E_t$分别记做$y_1-y_8$。预先定义的量是外生变量。$W_{t**}$=政府工资，$T_t$=税收，$G_t$=政府非工资支出，$A_t=$1931年以来所过年数。衍生内生变量是$P_{t-1},K_{t-1},E_{t-1}$，分别记作$x_5,x_6,x_7$。Klein 使用了1921-1941年间的数据进行分析，该模型常用于检验计量经济学估计量。

B,Γ及Ψ矩阵分别是：

$$
\begin{array}{c}
\\ C_t \\ I_t \\ W_{t*} \\ P_{t*} \\ Y_t \\ K_t \\ W_t \\ E_t
\end{array}
\begin{array}{c}
\begin{array}{cccccccc} C_t & I_t & W_{t*} & P_t & Y_t & K_t & W_t & E_t \end{array} \\
\begin{pmatrix}
0 & 0 & 0 & a_1 & 0 & 0 & a_3 & 0 \\
0 & 0 & 0 & b_1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & c_1 \\
0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0
\end{pmatrix}
\end{array}
\begin{array}{c}
\begin{array}{ccccccc} W_{t**} & T_t & G_t & A_t & P_{t-1} & K_{t-1} & E_{t-1} \end{array} \\
\begin{pmatrix}
0 & 0 & 0 & 0 & a_2 & 0 & 0 \\
0 & 0 & 0 & 0 & b_2 & b_3 & 0 \\
0 & 0 & 0 & c_3 & 0 & 0 & c_2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
\end{array}
$$

及

$$
\begin{pmatrix}
i_{11} & & & & & & & \\
i_{21} & i_{22} & & & & & & \\
i_{31} & i_{32} & i_{33} & & & & & \\
0 & 0 & 0 & 0 & & & & \\
0 & 0 & 0 & 0 & 0 & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

SAS/STAT 的PROC CALIS；SPSS LISREL；BMDP EQS；1989 年版的SYSTAT EzPATH均可用于结构方程建模。SAS 技术报告[7]详细解释了LISREL、EQS 及COSAN 等的联系，并配有实例。第 4 章介绍了CALIS 用于计量经济学分析的例子。SAS/ETS PROC SYSLIN 等也常用于计量经济学模型分析。

附Klein.dat 文件内容：

```
(15A4)          INPUT LABELS FOR KLEIN'S MODEL
  C P-1  W*   I K-1 E-1 W**   T   A   P   K   E   W   Y   G
(15F5.1)          INPUT DATA FOR KLEIN'S MODEL
 41.9 12.7 25.5 -0.2182.8 44.9  2.7  7.7-10.0 12.4182.6 45.6 28.2 40.6  6.6
 45.0 12.4 29.3  1.9182.6 45.6  2.9  3.9 -9.0 16.9184.5 50.1 32.2 49.1  6.1
 49.2 16.9 34.1  5.2184.5 50.1  2.9  4.7 -8.0 18.4189.7 57.2 37.0 55.4  5.7
 50.6 18.4 33.9  3.0189.7 57.2  3.1  3.8 -7.0 19.4192.7 57.1 37.0 56.4  6.6
 52.6 19.4 35.4  5.1192.7 57.1  3.2  5.5 -6.0 20.1197.8 61.0 38.6 58.7  6.5
 55.1 20.1 37.4  5.6197.8 61.0  3.3  7.0 -5.0 19.6203.4 64.0 40.7 60.3  6.6
 56.2 19.6 37.9  4.2203.4 64.0  3.6  6.7 -4.0 19.8207.6 64.4 41.5 61.3  7.6
 57.3 19.8 39.2  3.0207.6 64.4  3.7  4.2 -3.0 21.1210.6 64.5 42.9 64.0  7.9
 57.8 21.1 41.3  5.1210.6 64.5  4.0  4.0 -2.0 21.7215.7 67.0 45.3 67.0  8.1
 55.0 21.7 37.9  1.0215.7 67.0  4.2  7.7 -1.0 15.6216.7 61.2 42.1 57.7  9.4
 50.9 15.6 34.5 -3.4216.7 61.2  4.8  7.5  0.0 11.4213.3 53.4 39.3 50.7 10.7
 45.6 11.4 29.0 -6.2213.3 53.4  5.3  8.3  1.0  7.0207.1 44.3 34.3 41.3 10.2
 46.5  7.0 28.5 -5.1207.1 44.3  5.6  5.4  2.0 11.2202.0 45.1 34.1 45.3  9.3
```

```
48.7 11.2 30.6 -3.0202.0 45.1  6.0  6.8  3.0 12.3199.0 49.7 36.6 48.9 10.0
51.3 12.3 33.2 -1.3199.0 49.7  6.1  7.2  4.0 14.0197.7 54.4 39.3 53.3 10.5
57.7 14.0 36.8  2.1197.7 54.4  7.4  8.3  5.0 17.6199.8 62.7 44.2 61.8 10.3
58.7 17.6 41.0  2.0199.8 62.7  6.7  6.7  6.0 17.3201.8 65.0 47.7 65.0 11.0
57.5 17.3 38.2 -1.9201.8 65.0  7.7  7.4  7.0 15.3199.9 60.9 45.9 61.2 13.0
61.6 15.3 41.6  1.3199.9 60.9  7.8  8.9  8.0 19.0201.2 69.5 49.4 68.4 14.4
65.0 19.0 45.0  3.3201.2 69.5  8.0  9.6  9.0 21.1204.5 75.7 53.0 74.1 15.4
69.7 21.1 53.3  4.9204.5 75.7  8.5 11.6 10.0 23.5209.4 88.4 61.8 85.3 22.3
```

第十五章　Epi Info

§15.1　简介

Epi Info 是1988 年美国疾病控制中心(CDC) 流行病学室编制的用于流行病学调查分析的软件，运行于IBM PC 及其兼容机。1990 年该室与WHO 全球爱滋病控制小组合作，推出了功能更齐全的第5 版。它同时具有文字处理、数据库管理与统计分析这三种功能。在此基础上的汉化版Epi Info 在菜单提示帮助方面全为中文，且可使用中文变量名(字段)和文件名，便于不熟悉英文者使用和在国内推广。除了在建立调查表时要输入中文外，以后在使用中文时必须输入中文的时候不多而可由文件菜单及变量名菜单中直接调用汉字文件名及字段名，因而使用汉化版也十分方便。EPI INFO 新版本可在ftp.cdc.gov得到，在功能上有更多的改进。EPI INFO 是流行病学研究设计、数据录入和分析的最佳选择之一。以dBASE Ⅲ为例，编写修改屏幕格式文件是一件费力的工作，而且由于它至多只能处理128 个变量，极易在变量数较多时造成数据库间关系的混乱，Epi Info 则很好解决了这一问题。

Epi Info 主菜单中有主控程序EPI.EXE 连结主菜单中的主要程序。这些主要程序的名称及功能如下：

1. EPED 文本编辑程序。可用于建立调查表或编写各种程序。其中的EPIAID 程序可帮助准备文稿及设计流行病学调查。EPED 类似Wordstar 的编辑功能。计算机所编辑的表格样式，可以直接作为dBASE 那样的屏幕格式文件进行数据录入。每数据项后加上相应类型和长度的格式描述符后进入录模块，即可进行录入。该模块的EPAID可用于流行病调查辅助设计和报告撰写。

2. ENTER 数据录入程序。在由EPED 或由其它文字处理器建立调查表的基础上, 形成Epi Info 数据文件。可用于输入、查找以及修改数据文件格式。开始时，一个以.QES为扩展名的调查表文件是必要的，.REC 一旦建立，可被ENTER 反复直接调用。

   ENTER 可以进行光标的任意移动，并提供有特定记录的定位方法。每个记录录入结束后，软件提问是否将录入的数据记入磁盘，也可以取消提示。

3. ANALYSIS 数据分析程序。可由Epi Info 或dBASE 数据文件中显示记录及频数; 处理交叉表。进行流行病学中常用的统计计算: 如比数比、相对危险度、可信限，Fisher 精确检验及卡方检验，Mantel Haensze1 分层分析，单向方差分析，线性回归及配比病例对照研究的分析。此外，可对记录作选择、排序；有”IF” 语句及数学和逻辑运算、绘图、编制复合报表及产生新数据集的功能。还有一个程序语言可供编程计算。

4. CHECK 数据核对。可用于设置变量值范围、合法数值、自动编码及数据输入(ENTER)的跳越模式。可进行字段间数学和逻辑运算、复合跳跃、同一过程中访问几个文件。并可支持用户提供的自编程序。

5. STATCALC 统计计算。可对由键盘输入表格的数值进行统计。包括$2 \times n$表单一的或分层的2×2表、样本大小估计及单一或分层的趋势分析。

6. CONVERT 数据文件输出。把Epi Info 的数据文件转换到12 种数据库或统计软件的格式，大多数计算机程序可以接受至少一种格式。所支持的文件格式有：

1.EAS、2.SAS3、3.定界的、4.BASIC、5.SPSS-X、6.EpiStat、7.dBase II 、8.dBase III、9.Lotus 1-2-3、10.SPSS/PC+、11.Statpac 、12.固定度度的格式数据。

7. IMPORT 数据文件输入。从其它系统输入数据供Epi Info 使用。IMPORT 可将其它系统产生的文件调入Epi Info 作处理，所支持的文件格式有：1. 固定长度的格式记录；2. 逗号定界的；3. Lotus 1-2- 3 ；4. dBASE II/III。对其中的第1 和第2 项，首先必须产生一个描述数据的调查表并用ENTER 产生一个空的.REC 文件，对第3 和第4 则不需要。

8. MERGE 数据文件连接。将相同或不同格式的调查表产生的文件合并。可用以合并各计算机输入的数据文件并可用新输入的数据对过去的记录进行更新。

9. VALIDATE 比较由两个操作员录入的Epi Info 文件并报告其差别。如果能指定一个包含特定的标识的关键字段则两个文件中的记录不要求顺序相同。

除以上主要程序外，Epi Info 还有一些其它文件。

1. Help 文件。该文件包括了指导手册中大部分内容。在Epi Info 程序中可用¡F1¿键查到。

2. Sample 程序。提供了两个完整的样本监测系统; 两个流行病学调查文件及一个营养调查的人体测量学系统以供使用。汉化版提供了一个上海市肿瘤报告的例子。

3. Tutorials。一个完整的介绍EPED 和ANALYSIS 特性的人机对话课程。

在文字处理方面，除一般文字处理外，能以全屏幕编辑方式建立调查表，同时也建立了数据结构。根据调查表输入数据即可建立相应的数据库。

软件的统计功能限于一些在流行病学方面常用的统计分析。但由于它能同包括SAS，SPSS，dBASE及BASIC 等12 种数据类型进行互相转换，因而所建立的数据很容易由其它统计软件进行复杂的统计处理。

§15.2 汉化版的运行环境和安装

§15.2.1 硬件配置要求

Epi Info 对硬件的要求不高，如：
IBM\PC机或其兼容机CPU 80286 以上；
内存≥1 MB 至少有一个软盘驱器；
硬盘驱动器至少10 MB 根据数据量而增加；
显示器彩色VGA 或更高分辨率；
打印机LQ1500等、HP激光打印机、绘图仪等。

§15.2.2 软件配置要求

DOS 3.3 及以上版本。
UCDOS 2.01及以上版本或西山DOS 5.0及以上版本。

§15.2.3 安装或复制

Epi Info 第5 版是用Turbo Pascal 语言及汇编语言编写的。源程序约占7 .5兆，经编译后共105 个文件，容量约2 兆(1984613字节)。经压缩后存放于4 张360K 5 1/4 英寸软盘中。不经恢复不能使用。因而需要用Install 程序进行安装。汉化版使用方法与原版基本相同。

安装或复制时需要将安装盘置于相应驱动器中，(根据安装盘的种类) 键入Install 并回车即出现提示。按提示操作即可以人机对话方式装入全部程序。或者复制另一安装盘以推广本软件。

安装之后，在所安装的磁盘中即已建立了子目录\EPI5、\RSURV 以及\SURV。

如果在相应磁盘中已经安装过Epi Info 软件，则在安装时会逐一询问是否要复盖原来的文件，这样比较化费时间，不如先把原有的所有\EPI5 中的文件删去，然后删去子目录后再进行安装。如果你在\EPI5 子目录中已建立了你自己的一些调查表或记录文件，而你希望更新Epi Info 且要保存所有的调查表和记录文件,则可以把第一张盘插入A: 驱动器，在\EPI5 子目录中给出命令A:WIPE_E5 即可。Epi Info 安装到最后时，程序要求，如果此软件不是来自CDC 或USD. INC，用户可向他登记,登记后可以在以后得到他们无偿提供的有关Epi Info 的信息。

§15.3 使用

由于Epi Info 具有中文的教学软件、Help 文件、实例及大量菜单及提示，因而学习和使用十分方便。

首先进入中文DOS，然后用C:>CD\EPI5 进入子目录，即出现提示C:\EPI5¿。可用以下三种方法中任一种运行程序:

1. 直接启动所需程序，这样可以节省内存。如键入C:\EPI5> STATCALC 即可启动统计计算程序。各程序格式如下：

命令	命令行参数
ENTER	< .REC 文件> {< .QES 文件>}
ANALYSIS	<程序文件名> | <数据文件名> /BW /8
CHECK	< .REC 数据文件名>
STATCALC	{打印机或¡文件名>}
CONVERT	< .REC 文件><新文件>< #1-12>
IMPORT	< .REC 文件><外来文件>< #1-4>
MERGE	<f1><f2><f3>< #1-4> {<对应字段><Y/N>}
VALIDATE	< .REC 文件1>< .REC 文件2><关键词>

命令行参数是可选的，若键入命令时没有指定，系统则给予提示。设有一个全国性调查，各省上报数据文件为“省名.REC”，用DOS批文件直接把它们转到SAS中处理，批文件的内容为：

```
convert ANHUI    ANHUI    2 y
convert BEIJING  BEIJING  2 y
convert GUANGXI  GUANGXI  2 y
convert HUNAN    HUNAN    2 y
```

...

可见其简便。

2. 键入EPI，启动主控程序即出现主菜单。主菜单中每一主要程序名之中有一另一种颜色的字符(如黄色字符串STATCALC 中的S 为蓝色字符)。键入该字符即可运行该主要程序。

3. 在主菜单中有一光标条(cursor bar) 位于第一个主程序名上，用↑和↓键可以移动此光标到其它主程序名。按回车键即可启动此程序。此时如按空格键，则可在菜单下方提示框中直接输入命令。如光标在EPED 主程序上，按空格键后，即可在提示框中EPED 后键入文件名，按回车键后即可对此文件进行编辑。

   Epi Info 各主程序在屏幕上方列出各功能键的作用、可用功能键进行选择。如在EPED按<F2>键，可出现文件功能子菜单。Epi Info 各主程序都有逐级子菜单可列于屏幕上。各子菜单也可用上述第2、3种方法进行选择。在文字处理等程序也可用控制符进行操作，如EPED 中常可用与Wordstar 类似的控制符操作。

   用<ESC> 键可退到上一级菜单，用<F10>键退到主菜单或退出Epi Info。使用中并可随时进入或退出DOS，使用DOS 命令。

   汉字输入比较麻烦，在Epi Info 中常可以利用文件名清单或变量名清单用光标条进行选择。如在EPED 中用F2 功能键中"打开本窗口文件"功能，在要求输入文件名时输入*.*，或在路径后输入*.*, 回车后即显示相应子目录中文件名清单, 用光标条选择即输入了文件名。在ANALYSIS 中，调用某一记录文件后用<F3> 功能键即可列出字段名清单。同样，用光标条选择并按回车后即可输入此变量名，方便了汉字文件名或变量名的输入。

【例15.1】肿瘤报告卡的处理。CANCER.QES中汉化Epi Info中处理肿瘤报告卡的例子，可直接用于屏幕数据输入，这里主要是介绍它所用的数据格式。其中## #、###.##、<mm/dd/yy>及----等限定了数据录入为整型、实型、日期型和字符型格式。

【例15.2】汉化Epi Info 中假想了一个居民饮水与疾病调查，调查地区为北京市、上海市、江苏省和浙江省，调查的基本单位是户，每个被调查的人可能有许多次随访，所以数据建立时应设家庭编号和个人编号关键字。利用CHECK 可以建立几个数据集的关系，在每个问卷文件中已进行相关的说明，文件名在方括号(【】) 内。这几个文件也可以经其它文字处理软件生成。

【HOUSE.QES】

家庭记录

记录连接说明. 每个家庭记录以识别符"家庭编号"与一个或多个个人记录连接，在数据输入时，"家庭"文件为主文件. 在分析时，"随访"文件是主文件，而个人以"个人编号"与之相联系，然后家庭以"家庭编号"字段与合并了的"随访-个人"记录相联系.

家庭编号<idnum>

地址_____

《肿瘤调查表》

□ 总号_____

上海市肿瘤病例报告卡 <ICD编码> ###.#

门诊号_____ 更正诊断报告栏

住院号_____ (原报告诊断有误时填写)

患者姓名____ 性别__ 实足年龄### 原诊断.....

出生年月<mm/dd/yy> 民族..... 原诊断日期.....

职业(具体职务).....

工作单位_____ 诊断依据#

常住户口在....1:本市市区2:本市郊区 1:临床6:病理(继发)

地址_____ 2:X,超声波7:病理(原发)

诊断..... 内窥镜,CT

病理学类型_____ 3:手术,尸检8:尸检

诊断日期<mm/dd/yy> (无病理) (有病理)

报告单位### 报告日期<mm/dd/yy> 4:生化,免疫9:不详

D <mm/dd/yy> 死亡原因## 报告医师.... 5:细胞学,血片0:死亡补发病

沪卫生局批准

县市_____ 省市自治区_____

卧室数: ##

自来水: <Y>

【PERSON.QES】

个人记录

记录连接说明. 个人记录以"家庭编号"识别符和家庭记录联系,

并以"个人编号"识别符与随访记录相联系.

个人编号<idnum> 家庭编号#####

姓名_____

年龄### 性别__ 患病<Y>

【VISIT.QES】

随访

说明在Epi Info中三个文件的连接. 这一随访记录由"个人"文件中的个人建立. 每个人可有几次随访. 这两个文件中的记录

以"个人编号"字段互相连接.

个人编号#####

随访日期<mm/dd/yy>

随访记录_____

```
【HOUSE.CHK】
家庭编号
  KEY UNIQUE
END
 省市自治区

  Legal
    北京
    上海
    江苏
    浙江
  END
END
 自来水

  ENTER
  ENTER PERSON.REC 家庭编号
END
【PERSON.CHK】

个人编号
  KEY
END
 患病

  ENTER visit 个人编号
END
【VISIT.CHK】

个人编号
  KEY
END
```

以上各.CHK文件中汉字开始END结束的内容是对对该字段的说明。实际调查的项目可能更多，逻辑关系也会更复杂，这个例子仅仅是一个示范。准备好上述几个文件后就可以用ENTER录入了，录入时ENTER自动在几个数据文件中跳转。

【例15.3】下是1991-1994年医院满意度调查的例子，在“数据处理与综合应用”一章对之有更多的介绍。

卫生部“纠风调查”病人问卷

同志，您好！欢迎您参加卫生部组织的民意测验，我们希望通过您的真诚回答，反映出群众对卫生系统行业作风的意见和看法，本问卷不记单位和姓名，我们保证对您的回答保密。谢谢您的合作！

医院名称NAME：_____

医院编号{ID1}：######

病人编号{ID2}：######

病人类别{SS}：#

一您这次在哪级医院就医？	{N1}	#
二您这次在哪科就医？	{N2}	##
三您的主要职业：	{N3}	##
四您的医疗费用支付方式为	{N4}	#
五您对医生服务态度是否满意？	{N5}	#
六您认为医生技术	{N6}	#
七您对护士的服务态度是否满意？	{N7}	#
八您认为护士技术	{N8}	#
九您对这所医院的医疗质量是否信任？	{N9}	#
十您对挂号处工作人员的服务态度是否满意？	{NA}	#
十一您对药房工作人员的服务态度是否满意？	{NB}	#
十二您对医院膳食是否满意？	{NC}	#
十三您对医院的环境卫生条件是否满意？	{ND}	#
十四您这次就医期间是否托关系？	{NE0}	#
㈠中间人是否本院职工	{NE1}	#
㈡中间收礼？	{NE2}	#
㈢托人价值	{NE3}	#
十五您这次就医期间是否送钱物？	{NF0}	#
㈠送礼价值为	{NF1}	#
㈡送礼原因是	{NF2}	#
㈢送礼态度是	{NF3}	#
十六您在这次就医过程中有无宴请	{NG0}	#
㈠宴请价值为	{NG1}	#
㈡宴请原因是	{NG2}	#
㈢宴请态度是	{NG3}	#

十七您对医院最不满意的是什么？有什么{建议}(文字叙述)

_____

_____

由于问卷项目必须符合一定的逻辑关系或条件，应在CHK 型文件中加以说明，如合法范围、汉字标号以及跳转与重复等。使用汉之标号便于录入过程中进行特定项目的核实，使用跳转则避免了可能存在的数据的不合理。设文件名为SURV.CHK，其内容如下：

```
NAME
  Repeated
END

SS
  Comment Legal
    1 门诊
    2 住院
    3 出院
  END
END

N1
  Comment Legal
    1 省级
    2 地、市级
    3 县级
  END
END

N2
  Comment Legal
    1 内科
    2 外科
    3 妇产科
    4 儿科
    5 中医科
    6 眼科
    7 耳鼻喉科
    8 口腔科
    9 皮肤科
    10 其它
  END
END

ND
  Comment Legal
    1 满意
    2 较满意
    3 一般
    4 不满意
    5 很不满意
  END
END

NE0
  Comment Legal
    1 是
    2 否
  END
  Jumps
    2 NF0
  END
END

NE1
  Comment Legal
    1 是
    2 否
  END
END

NE2
  Comment Legal
    1 是
    2 否
  END
END
```

```
N3
  Comment Legal
    1 工人
    2 农民
    3 军人
    4 教师、医务人员或科验
    5 企事业机关干部
    6 个体从业者
    7 商业从业者
    8 无职业者
    9 学生
    10 其它
  END
END

N4
  Comment Legal
    1 公费
    2 劳保
    3 半自费
    4 自费
    5 商业性医疗保险
  END
END

N5
  Comment Legal
    1 满意
    2 较满意
    3 一般
    4 不满意
    5 很不满意
  END
END

N6
  Comment Legal
    1 好
    2 较好
    3 一般
    4 不好
    5 很不好
  END
END

NE3
  Comment Legal
    1 50元以下
    2 50-
    3 100-
    4 200-
    5 500-
  END
END

NF0
  Comment Legal
    1 是
    2 否
  END
  Jumps
    2 NG0
  END
END

NF1
  Comment Legal
    1 50元以下
    2 50-
    3 100-
    4 200-
    5 500-
  END
END

NF2
  Comment Legal
    1 出于感激
    2 想得到方便和照顾
    3 担心医务人员工作不认真
    4 受它人影响
    5 医务人员暗示的
    6 医务人员直接索要的
  END
END

NF3
  Comment Legal
    1 拒绝收
    2 事后全部退还
```

```
N7
  Comment Legal
    1 满意
    2 较满意
    3 一般
    4 不满意
    5 很不满意
    6 没接触
  END
END

N8
  Comment Legal
    1 好
    2 较好
    3 一般
    4 不好
    5 很不好
    6 没接触
  END
END

N9
  Comment Legal
    1 信任
    2 较信任
    3 不信任
    4 很不信任
    5 说不好
  END
END

NA
  Comment Legal
    1 满意
    2 较满意
    3 一般
    4 不满意
    5 很不满意
    6 没接触
  END
END
    3 收礼后照价付了钱
    4 受礼后付了部分钱
    5 推辞过，但还是收下了
    5 没有拒绝就收下了
  END
END

NG0
  Comment Legal
    1 是
    2 否
  END
  Jumps
    2 建议1
  END
END

NG1
  Comment Legal
    1 50元以下
    2 50-
    3 100-
    4 200-
  END
END

NG2
  Comment Legal
    1 出于感激
    2 想得到方便和照顾
    3 担心医务人员工作不认真
    4 受它人影响
    5 医务人员暗示的
    6 医务人员直接索要的
  END
END

NG3
  Comment Legal
    1 拒绝收
    2 事后全部退还
    3 收礼后照价付了钱
    4 受礼后付了部分钱
```

```
        NB
          Comment Legal
            1 满意
            2 较满意
            3 一般
            4 不满意
            5 很不满意
            6 没接触
          END
        END

        NC
          Comment Legal
            1 满意
            2 较满意
            3 一般
            4 不满意
            5 很不满意
            6 不清楚
          END
        END
```

```
            5 推辞过，但还是收下了
            6 没有拒绝就收下了
          END
        END
```

为了节省篇幅，后面一部分与前面部分同时列出。

第一项设为重复，因此对特定医院的病人，汉字名只需输入一次就够了。其它项目有相应的合法值和说明，只要在ENTER 下打入F9，就可以行到相应的标号提示，最后一些项目当出现“否”时，其下面的内容略去不输，比dBASE 自动填零进了一步。

第四部分

数据管理与图形文字处理

第十六章　数据管理和综合应用

§16.1　数据管理及其计算机软件

为了更有效使用统计软件包，有一个计算机数据管理的映象是至关重要的。如方差分析手工的做法是先把数据分组，但在软件包处理时，是放在线性模型的框架之中的，需要用一个分组变量。反之，若在SAS 的数据步中使用简单的命令则可给分组数据加上分组标志。

数据管理即对数据的操作，内容包括数据库创建、数据录入、数据编辑、查找、索引、合并、追加和汇总、存档等；也即从表格录入、逻辑检错，到汇总分析和产出的全过程。

数据库是公用的为特定目标服务的数据集合，用于满足多种类型终端用户的需要，数据库管理系统(DBMS) 是数据库与用户间的接口。数据库模型有三种，即层次型、网状型和关系型，以关系型数据库最为重要。它用表格来表示实体与实体间的关系，用行或横栏表示记录或不同的观察对象，用列表示具有相同特征的数据。如人口普查中的数据项目有年龄、性别、职业、文化程度，住址等，横栏是不同的人，各项目表示了不同人的特征。关系型数据库的一个重要特征就是一个表的记录可以与另一个表中的记录关联。存贮信息的不同表形成一个数据字典，它记录了数据环境的逻辑和物理方面的安排。它可以是活动的或被动的，后者对于数据定义进行维护，但对数据库的访问无控制。一个好的数据字典应能够：ａ、对数据格式和类型提供标准定义；ｂ、维护对应用程序提供的交叉参照数据列表；ｃ、对与系统有关的项目，包括用户数、缓冲区数目和大小、那一个终端与系统连接、那一个程序与用户受系统特定变动的影响，等等。与数据库不同，数据字典记录的是数据库自身的结构，数据字典的内容可视做关于数据与系统的综合资料，有时称做公用数据字典(common data dictionary, CDD)。SPSS/PC+的数据录入工具DE是一个范例。SAS使用Windows下的动态数据交换可直接读写EXCEL的数据。

数据库是计算机领域中最重要的技术之一，已成为计算机软件独立的分支。数据库技术产生于六十年代初期。现已广泛应用于工农业生产、交通运输、商业、行政管理、科学研究、医疗卫生事业和国防建设等。

各种统计软件包有其自身的数据格式，称为系统文件，其中包括数据的创建时间、变量的格式、标号等。其优点在于操作方便，处理速度快，但其内容多在终端上不能正确显示，也难以用其它软件进行编辑，必须经中介的文件格式进行软件包的数据交换。软件包在数据管理上有其自身的特点，如StatGraphics 软件进行数据分析时，能够进行数据文件之间的跨越调用。多数软件包能够处理缺失值，在运算时，或者当缺失发生在分析的变量时删除该变量，或者当一个记录出现缺失时，删除该记录，进行统计分析时应注意适当选用。处理的原始数据，既可以是坐标类型原始数据，也可以是前期运行后的对称矩阵。另一种软件间常用的数据交换格式是ASCII 类型文件，PC 机上比较标准的格式如dBASE 数据库文件、电子报表如LOTUS 1-2-3 格式等。

dBASE 是一个优秀的微机关系型数据库管理软件，dBASE Ⅱ、dBASE Ⅲ、dBASE ⅢPLUS及dBASE Ⅳ均是Ashton-Tate 公司推出的dBASE 产品。利用dBASE 进行数据管理，首先要根据调查项目建立数据库结构，包括各个字段的名称、类型、长度、小数位数等。以后便可进行数据的录入、编辑、插入等操作，还可以增加新的项目或字段。这些操作即是数据库的维护，各操作对应不同的数据库管理功能。这些命令的组合，形成了数据库管理程序。在VAX/VMS系统和UNIX 系统下也有dBASEⅣ相应的软件。FoxBASE+ 是美国FOX 软件公

司于1987 年2 月推出的关系数据库系统，它与Ashton-Tate 公司的dBASE lllPLUS 完全兼容，而且有多方面扩展。其速度更快，适用机种广泛，支持的操作系统多，最近推出的FoxPro 2.0 则保持了同FoxBASE+ 的兼容并对FoxBASE+ 的功能进行了拓展。

其它流行的数据库软件如ORACLE 最早用于IBM 大型机,开发于1979 年，使用结构查询语言SQL，现已用于MS-DOS、Unix、VM/SP、MVS/SP、MVX/XA 及VMS 系统，本章结合SAS用例加以介绍。SQL 与宿主语言结合的另一种方式是嵌入方式，目前SQL 语言标准允许嵌入的宿主语言有Fortran、COBOL、Pascal、PL/I 四种。SAS和SPSS均对它有相应的支持。

综合应用是即充分利用现有计算机各种系统资源，进行计算机与软件包之间数据与程序的交换，以扬长避短。本章介绍几个PC机与VAX/VMS系统操作的例子。

§16.2 原始数据的录入和管理

计算机数据处理和统计分析，首先要进行原始数据的录入。如在流行病学研究中，常常要进行一般项目的调查，如姓名、年龄、性别等，调查表可能是以下的形式：

识别码(ID) □□□

一、姓名(NAME) _____ ----------

二、年龄(AGE) _____ 岁 □□□

三、性别(SEX) 1.男(M) 2.女(F) □

四、身高(HEIGHT) _____ 厘米 □□□.□□

五、体重(WEIGHT) _____ 公斤 □□□.□□

. . .

第二列专为计算机录入使用，使用中英文数据库管理软件，数据库变量名可为中文或英文，(英文名放在括号内)，各变量的信息为：

名称	(英文名)	类型	宽度	小数位数	起止位置
识别码	(ID)	数字型	3	0	1-3
姓名	(NAME)	字符型	10		4-13
年龄	(AGE)	数字型	3	0	14-16
性别	(SEX)	字符型	1		17
身高	(HEIGHT)	数字型	5	2	18-23
体重	(WEIGHT)	数字型	5	2	24-29

对调查、实验获得的原始表，首先要进行标识项目的检查，从表格上应该看出它是来自那个单位，其隶属关系如何。同时，标明调查日期。为了方便计算机分类处理，不同的表格可以编上一些标识码；其次是数据项之间合理性的判断。对于大量的数据，只要有可能，对数据进行编码；使用最详细的记录，如记录年龄时记录生日；经常对数据进行备份；对每个量用一个代码表示缺失值。数据处理和分析的流程是：

建库→原始数据录入→逻辑检查→数据分析→报告。

本例使用前面介绍的Epi Info 软件可直接录入并处理。现用流行的dBASE 软件建库，使用CREATE、APPEND 命令即将数据录入，生成.DBF 文件，使用带有SDF 选项的COPY 命令，可将数据拷成标准格式的ASCII 码文件，由所用的软件读取。

dBASE数据库命令的语法可在进入数据库系统后，借助软件的帮助命令(HELP) 获得。dBASE用前四个英文字母作为关键字。其命令子句也是其可选项，如：

COPY <范围> TO <文件名> FIELDS <字段名表> [FOR/WHILE< 条件表达式>] [SDF/DELIMITED]。

其中大写字母表示关键字，方括号表示可选项，当关键字被指定时，其后的尖括号为必选项。<范围>有三种：ALL 表示所有记录；RECORD <表达式> 是表达式所指定的单一记录；NEXT <表达式> 表示从当前记录开始后指定数目的记录，表达式指定了记录的数目。字段名表是一系列用逗号分开的字段名，只有被指定的相关的字段才被操作。在dBASE 的语句中，使用FOR/WHILE 限定作为数据库操作的逻辑表达式，最后的SDF/DELIMITED 子句指定系统生成标准格式或带有分界符的ASCII 文件。考虑到一些微机统计软件尚不能直接使用dBASE 的数据库，SDF /DELIMITED 子句很有用。

下面列出的是其最常用的指令，各指令的用法可参有关资料。

USE	打开和关闭数据库。
APPEND	向数据库追加记录。
EDIT	编辑记录。
DELETE	删除记录。
INSERT	插入记录。
JOIN	数据库文件间的连接。
CHAGNE	改变字段和记录内容。
CREATE	创建数据库文件、报表格式文件和标签文件。
UPDATE	改变记录内容。
DISPLAY	分屏显示数据库结构和记录内容。
LIST	连续显示数据库结构和记录内容。
MODIFY	修改数据库结构、命令文件、报表格式文件和标签文件。
SORT	按关键字段对数据库进行排序。
TOTAL	按关键字段对数据库的数据求和。
INDEX	建立数据库索引文件。
REPORT	生成报表格式文件。
LABEL	生成标签文件。

dBASE 常用文件类型，可以通过文件的扩展名来区分，这样，可以对有关文件进行适当的维护，如：

.DBF	数据文件，用于保存数据和暂存文件。
.PRG	命令文件，用于完成数据管理的某种功能。
.FOX	编译程序，功能与.PRG 文件相同。
.FMT	格式文件，进行数据录入时的屏幕格式和产出时的格式定义。
.BAK	备份文件，在数据库文件修改时可以生成。
.NDX	索引文件，用于查找、排序等功能。
.FRM	报表文件，用于产生报表。
.LBL	标签文件，用于打印数据库项目的标签。
.MEM	内存变量文件，可存放dBASE 运行时的内存变量。

需要管理多个数据库时，可启用不同的区。数据库管理程序是运行采用DO < .PRG 命令文件> 执行。

与软件包的系统文件相比，dBASE 文件没有专门的缺失值定义，使用SAS 和SPSS/PC+ 等软件在dBASE 数据转贮时进行了特殊的处理。另外，dBASE 不能使用变量标号，因而数据文件不很直观；对于每个记录，则可使用注释字段的方式来解决。单个dBASE 文件最多处理128 个变量。

§16.3 软件包数据管理

§16.3.1 SAS

具有强大的数据管理功能。大部分操作是通过SAS/BASE 来完成，其强大的语言特色也主要表现在其数据步上。SAS也用一系列专用过程如CATALOG、DATASETS、COMPARE、APPEND、COPY、TRANS用于目录和数据集的管理，CONTENTS 用于浏览不同存贮类型文件的内容。SAS/FSP 提供了全屏幕操作，包括文件的创建、编辑等，有专门的屏幕控制语言(SCL) 。另外，SAS/ACCESSS、SAS/CONNECT 、SAS/SHARE 等工具用于各种计算机系统和软件包间的数据交换。在PC SAS上，使用PROC DBF 和DIF 进行dBASE 等文件的转换。使用DBF 生成dBASE 文件时，由于后者没有缺失值和标号，用填满9 的16 位宽的字段表示缺失值，转出时，为了保证数据格式的正确，应使用FORMAT 语句对变量格式行说明。假设现有ASCII 文件，以外部文件的格式读入，则应在DATA 步中使用INFILE 语句，若是dBASE 的COPY 命令生成的标准数据集(SDF)，可以采用带有格式的INPUT 语句在DATA 步读入，如上节的例子，INPUT ID 1-3 NAME $ 4-13 AGE 14-16 SEX 17 HEIGHT 18-22 WEIGHT 23-27；使用dBASE 的COPY 命令生成的有分界符的文本文件，则可使用DELI WITH BLAN 子句，一个记录的数据项之间用特殊分界符分开，可在INFILE 语句中指示dlm='分界符'。当文本文件宽度超常时，应指定LRECL= <记录宽>，其它的选项可参考有关说明书。

【例16.1】以下程序读入一个对称阵，当一行读不满时，使用MISSOVER 避免了到下一行继续读入数据。

* 杨维权等:《多元统计分析》;

```
data p341(type=corr);
infile cards missover;
input _name_$ x1-x3 _type_$;
cards;
x1  corr 1.0000000
x2  corr -.3333333  1.0000000
x3  corr 0.6666667  0.0000000  1.0000000
.   n    5          5          5
;
```

数据作为统计过程的选项，使用DATA= 引用被分析数据，OUTSTAT= 生成包含统计量的数据集，OUT= 指示生成的带有原始数据的文件。象PROC REG 一类过程可以使用专门的OUTPUT OUT= 语句保留计算结果。

SAS 在系统文件管理上，采用两水平的文件名，首先把计算机物理路径赋给一个逻辑的库名，在以后的操作中使用“库名.数据集名”的格式调用。SAS 把这样生成的数据集称为永久性数据集。

【例16.2】下面的程序利用DATASETS 过程进行永久性数据集间的操作。

```
LIBNAME MY 'C:\MYDIR1';                /* 定义库名*/
LIBNAME YOUR 'C:\YOUR\MYDIR2';
PROC DATASETS LIBRARY=MY COPY OUT=YOUR; /* 调用DATASETS 过程*/
SELECT CLASS1 EXAM1;
```

```
CHANGE EXAM1=EX1 EXAM2=EX2;
DELTE CLASS2;
MODIFY CLASS1; /* 修改数据集CLASS1 */
RENAME NAME1=NAME GOUK1=TOTAL;
LABEL EIGO='TEST-1' SUGAKU='TEST-2';
RUN;
```

【例16.3】下面程序用CONTENTS 把D:\SAS\SASINST 目录中所有文件列出来。

```
LIBNAME INST 'D:\SAS\SASINST';
PROC CONTENTS DATA=INST._ALL_; RUN;
PROC PRINT DATA=INST.CLASS;    RUN;
```

CONTENTS 关于数据集的产出包括数据集名、记录数、变量数、数据库及变量标号、记录长度及类型和格式。如第四章用例数据集CLASS.SSD 的内容如下：

```
Data Set Name:  SS.CLASS                Type:
Observations:   19                      Record Len: 37
Variables:      5
Label:          Student information
          -----Alphabetic List of Variables and Attributes-----
#  Variable  Type  Len  Pos  Label
3  AGE       Num     8   13  Age in years
4  HEIGHT    Num     8   21  Height in inches
1  NAME      Char    8    4  First name
2  SEX       Char    1   12  Gender
5  WEIGHT    Num     8   29  Weight in pounds
```

数据列表：姓名(NAME) 与性别(SEX) 是字符型数据，OBS 栏指示记录号。

OBS	AGE	NAME	SEX	HEIGHT	WEIGHT
1	15	JANET	F	62.5	112.5
2	11	JOYCE	F	51.3	50.5
3	14	JUDY	F	64.3	90.0
4	14	CAROL	F	62.8	102.5
5	12	JANE	F	59.8	84.5
6	12	LOUISE	F	56.3	77.0
7	13	BARBARA	F	65.3	98.0
8	15	MARY	F	66.5	112.0
9	13	ALICE	F	56.5	84.0
10	12	JOHN	M	59.0	99.5
11	12	JAMES	M	57.3	83.0
12	14	ALFRED	M	69.0	112.5
13	15	WILLIAM	M	66.5	112.0

```
14     13    JEFFREY     M      62.5       84.0
15     15    RONALD      M      67.0      133.0
16     11    THOMAS      M      57.5       85.0
17     16    PHILIP      M      72.0      150.0
18     12    ROBERT      M      64.8      128.0
19     14    HENRY       M      63.5      102.5
```

【例16.4】下面是一个论文报告的例子，采自SAS 6.07 PROC SQL 示范程序。是数据查询管理的好范例，程序paper.sas 创建一个表，用于以后的查询。变量为作者名、专题类别、标题、开始时间和持续时间。

```
data paper;
 input author$1-8 section$9-16 title$17-43 @45 time time5.
       duration;
 format time time5.; label title='Paper Title';
 cards;
 Tom     Testing Automated Product Testing    9:00 35
 Jerry   Testing Involving Users              9:50 30
 Nick    Testing Plan to test, test to plan  10:30 20
 Peter   Info SysArtificial Intelligence      9:30 45
 Paul    Info SysQuery Languages             10:30 40
 Lewis   Info SysQuery Optimisers            15:30 25
 Jonas   Users   Starting a Local User Group 14:30 35
 Jim     Users   Keeping power users happy   15:15 20
 Janet   Users   Keeping everyone informed   15:45 30
 Marti   GraphicsMulti-dimensional graphics  16:30 35
 Marge   GraphicsMake your own point!        15:10 35
 Mike    GraphicsMaking do without color     15:50 15
 Jane    GraphicsPrimary colors, use em!     16:15 25
 ;
```

下面程序便用PROC SQL 过程对上述数据进行查询操作：

1、选择

```
%include paper;
proc sql;
  * 以下语句列出表的所有信息;
  select * from paper;
  * 多长时间结束?;
  select author, title, time, duration,
              time + duration*60 as endtime
     from paper;
  * 现在加上一些标号和格式以更明了;
```

```
  select author, title, time,
                duration label='How Long it Takes',
                time + duration*60 as endtime format=time5.
     from paper;
  * 哪些论文是上午报告? 使用where 子句控制;
  select author, title, time,
                duration label='How Long it Takes',
                time + duration*60 as endtime format=time5.
     from paper where time < '12:00't;
```

2、创建

```
%include paper;
proc sql;
  /* SQL 用三种方式创建SAS 数据集*/
  /* 1) 作为其它表的空拷贝, 2) 作为任何有效SQL select 表达式的结果，3) 从传统的SQL 数据操纵语言(DML) 生成。下面的表P2 是PAPER的拷贝，P3 包括了所有12:00 以后交的论文*/
  create table p2 like paper;
  create table p3 as select * from paper where time > '12:00't;
  * 下面是一个新表;
  create table counts(section char(20),papers    num);
```

3、删除

```
%include paper;
proc sql;
  /* 创建一个表，用于演示数据的增删*/
  create table counts( section char(20), papers num );
  insert into counts values('Graphics', 4)
                          values('Info Sys', 3)
                          values('Testing',   2)
                          values('Users',      3)
                          values(",            1);
  * 删除以前的情况;
  select * from counts;
  * 用where 删除那些对于section 变量无意义的论文;
  delete from counts where section is null;
  select * from counts;
  * 宣读的论文仍然很多，含论文数目最小的专题将被取消;
  delete from counts
     where papers = ( select min(papers) from counts );
  select * from counts;
```

4、插入

```
%include paper;
proc sql;
```

```
/* 有两种方法把数据插入SAS 数据集*/
* 1) 插入常数值, 2) 使用SQL select 选择的数据;
* Jost 提交一份外语问题的新论文;
insert into paper(author, title, time)
      values('Jost', 'Foreign Language Issues', '11:15't);
* 插入以后的结果;
select * from paper;
* 创建一个新表Counts, 包括专题号及其论文数目;
create table counts( section char(20), papers num );
insert into counts
select section, count(*) from paper group by section;
select * from counts;
```

5、连接

先构造一个专题和园桌讨论会的数据，内容为专题、所在房间号、专题召集人、园桌会议主持人和讨论题。

```
data section;
 input section$1-8 room$ convenor$;
 cards;
Graphics  Sable  Denise
Info Sys  Kudu   Peter
Testing   Sable  Jenny
Users     Kudu   Sally
data roundt;
 input leader$1-8 subject$9-30;
 label subject='Roundtable Subject';
 cards;
Mary    External DBMS's
Nick    Testing Networks
Jerry   User Specifications
Peter   Selling Solutions
Jim     Distasteful Jokes
Marge   Designing Fonts
```

```
proc sql;
  * 哪些作者是园桌会议的主持人?;
  select author, title, subject
     from paper, roundt where author = leader;
  * 将讨论哪些论文和园桌论题?;
  select author, title, subject
     from paper full join roundt on author = leader;
  * 讨论的内容是什么，谁负责？;
  select coalesce(author, leader) as person, title, subject
```

```
    from paper full join roundt on author = leader;
  * 下一位报告人是谁?;
  select p.author label='Just Heard',
         n.author label='Then Try', n.title, n.section, n.time
    from paper p, paper n
  where n.time between p.time + p.duration*60
                   and p.time + (30+p.duration)*60
  order by p.author;
  * 谁非常忙，交论文、主持园桌会议和专题?;
  select distinct author from paper, section, roundt
  where author=convenor and author=leader;
  * 有必要在门厅处张贴一个简报，写明论文、园桌论题内容及负责人;
  select coalesce(author,convenor,leader) as person,
         title label='Gives Paper:',
         section.section label='Convenes Section:',
         subject label='Leads Roundtable on:'
    from paper full join section on author=convenor
                full join roundt   on   coalesce(author, convenor)
                =leader
  order by 1;
6、更新
%include paper;
  /* update 语句就地更新SAS 数据集*/
  * 使用的值有：1) 常量, 2) SQL 查询的结果, 3) 关于原始值的表达式;
  * 为了更新，先插入新到的Jost 的情况;
proc sql;
  insert into paper(author, title, time)
    values('Jost', 'Foreign Language Issues', '11:15't);
  * Jost 无专题，给它命名为"Users";
  update paper set section='Users' where author='Jost';
  select * from paper;
  /* Jost 认为他的论文会和别人的一样长，因此使用SQL update 语句设其持续时间
为所有论文的平均值。UNDO_POLICY 的默认选项可使PROC SQL 唯一地访问被插
入的数据集(sql.newprice)，对同一个表第二次引用时会失败，使用UNDO_POLICY=OPTIONAL
使查询得以继续*/
  reset undo_policy=optional;
  update paper
      set duration = ( select avg(duration) from paper )
    where author = 'Jost';
  reset undo_policy=required;
  select * from paper where author='Jost';
  * 因为大家都想多睡一会儿，因此每篇论文宣读也较原来的推迟30 分钟;
```

```
  update paper set time = time + '0:30't;
  select * from paper;
```

7、视图

```
%include paper;
```

创建视图，给原始数据增加一列，根据是starttime 和duration 得到的结束时间。

```
proc sql;
        create view pt as
        select author, title, time, duration label='Duration',
                time + duration*60 as endtime format=time5.
          from paper;
  * PROC SQL 中视图若同真正数据集一样;
  select * from pt where endtime > '14:00't;
  * 做一个论文宣读的清单和时间表;
  * SAS 过程可以访问SQL 视图，就象真数据库一样;
proc print data=pt;run;
proc timeplot data=pt;
  plot time='<' endtime='>' / overlay ref='12:00't hiloc;
  class author;
  run;
  * 论文宣读的顺序是什么？;
proc sql;
  create view pt_time as select * from pt order by time;
proc print data=pt_time; run;
  * 这个新的次序影响时间表吗？;
proc timeplot data=pt_time;
  plot time='<' endtime='>' / overlay ref='12:00't hiloc;
  class author;
  run;
```

【例16.5】使用SAS for Windows 与MS Excel进行数据交换。假设在MSExcel 有一个活动的工作表文件名为sheet1，现用SAS 产生三个随机数写入该活动文件，然后再把它读取打印，程序如下：

```
filename random dde 'excel |sheet1 ! r1c1:r100c3';
data random;
  file random;
  do i=1 to 100;
       x=ranuni(i);y=10+x;z=x-10;put x y z;
  end;
run;
filename monthly dde 'excel | sheet1 ! r1c1:r10c3';
```

```
data monthly;
   infile monthly;
   input var1 var2 var3;
run;
proc print;run;
```

§16.3.2 SPSS/PC+

命令GET 和SAVE 用于读入(GET) 和存贮(SAVE) SPSS/PC+ 系统文件，其格式为：

GET /FILE='文件名'.

SAVE /OUTFILE='文件名'.

SPSS/PC+ 的WRITE 命令可用于转贮ASCII 文件，FLIP 用于进行数据的转置。

在SPSS/PC+ 的显示管理方式下，运行命令DE. 即启用它的全屏幕数据管理工具DATA ENTRY。DATA ENTRY 工具除进行系统文件的录入、读写外，也可以读入dBASE/Lotus 等格式的文件。

DM 主菜单(Main Menu) 内容如下，用↑F1-↑F10(Shift 与功能键的组合)来调用。DE 的功能分成七类，在各部分中，功能键有不同的定义，使用F1 进入菜单，然后打需要帮助的功能键，使用空格键退出帮助窗口，使用ESC 键退出帮助状态。

↑F1 Help	帮助信息
↑F2 Files	文件操作
↑F3 Forms	屏幕表式
↑F4 Dictionary	数据字典
↑F5 Data	数据管理
↑F6 Cleaning	数据清理
↑F7 Skip&Fill	跳转定义
↑F8 Options	运行选项
↑F9	(未用)
↑F0 Exit	退出

在文件操作中，可以定义一个新文件，读取或存贮SPSS/PC+ 系统文件，或者电子报表、数据库文件、交换文件，或ASCII 码文件；把现有的数据字典拷贝到新文件，列目录；编辑、读取或存贮ASCII 数据模板。

使用逻辑定义可以定义和修改数据的范围、规则(RULES, 变量间的逻辑表达式) 。范围和规则定义之后，可以对数据进行扫描并且打印符合条件的记录，可利用数据部分来检查或改正这些记录。

数据字典可以为新文件定义变量，或者修改现有文件中变量的定义。新文件只有定义了变量以后才能录入数据，从现有文件增加或删除变量，必须存贮这个文件，然后读入开始录入。

数据管理用于数据录入，或者修改数据。这可以由电子报表或表样方式录入。前者以记录号为行，以变量为列，每屏20行，后者按表部分的定义显示。

跳转定义允许定义和修改跳转逻辑规则，每个规则与一个变量关联。若规则在数据部分作了修改，跳转规则可用于填充其它变量的取值以及决定下一个编辑的变量。

运行选项为DE 程序运行设定或修改环境。

```
————————Create/Edit Dictionary————————
                    Help F1    aF1 Help
          Define Variable F2   aF2
            Edit Variable F3   aF3
            Copy Variable F4   aF4
       Edit Value Labels F5    aF5 Edit Field Help
       Copy Value Labels F6    aF6 Copy Field Help
                          F7   aF7
Set Display Mode (All) F8      aF8
          Delete Variable F9   aF9
                          F0   aF0
————————Press Function Key to Select————————
```

退出DE 程序将返回DOS 或SPSS/PC+。

在DE运行的任何时刻打F1，系统给出相应的提示。控制菜单(Ctrl Menu) 使用Ctrl 键与功能键的组合，如^F5显示缓冲区，^F9提供变量信息，^F10表示完成。

用例，现拟建一个名为NEW.SYS 的系统文件。选择↑F2 读入文件部分，使用F4 定义新文件，退出。然后使用↑F4定义数据字典，其屏幕格式如下：

打F2定义变量。现设变量为年龄(age)，标号为"age in years'，宽度为3，缺失值为-999，有关信息如下：

```
Variable Name     age
Variable Label    age in years
Type of Variable  Numeric
Variable Length   3
Decimal Places    0
Display Mode      Edit
Missing           999
```

以^F10 完成。

使用↑F6，选择F2(Define Range/Rule)定义年龄的取值范围，现规定为1-150，即1 thru 150，以^F10退出。

使用↑F7，F2 (Define Skip Rule)定义变量间的跳转定义，^ F10退出，有关的跳转表达式操作符如下。

————————————Available Operators————————————

-	mod	and	&	lt	<	″	not	if	exit
+	in	or	*mid*	le	<=	**	implies	else	->
⋆	thru	eq	=	ge	>=	,	(...)	display	;
/	by	ne	″=	gt	>	<>	{...}	nextcase	

本例只定义年龄一个变量故不使用该选项。

为了今后数据录入方便，定义数据录入的屏幕格式，使用↑F3，F2(Generate default form)，使用F10(draw box)可以画方框，结束时用^F10退出。

—File Type to Save—

SPSS/PC+
SPSS Portable
123 Rel 1A
123 Rel 2
Symphony Rel 1.0
Symphony Rel 1.1-2.0
dBase II
dBase III
dBase IV
Multiplan Symbolic

——————————————Variables ——————————————————
ALL TO $CASENUM $DATE $WEIGHT AGE
—————————————— AGE ——————————————————

Age in years
Type: Numeric Missing value: 999 Width: 3 Decimals: 0
No value labels *
————Type Esc or punctuation character to remove menu——

录入样本数据：退回主菜单后使用↑F5，或者按照定义的屏幕格式录入，或者按照电子表格那样的形式录入，该项功能用F10切换。数据录入过程中，如出现非法数据或条件，屏幕自动跳出给予相应的警告信息。

全屏幕数据录入工具运行环境可用↑F8来改动，F2为读取现有设置。

录入结束后，仍以↑F2，选择文件存贮，系统提示：

择第一项，存贮为SPSS/PC+格式，压缩存贮。

现用↑F10退出DE 至SPSS/PC+ 系统看建立的效果。

```
SPSS/PC:get /file 'new.sys'.
 The SPSS/PC+ system file is read from
     file new.sys
 The file was created on 12/02/93 at 15:47:59
 and is titled SPSS/PC+ System File Written by Data Entry II
 The SPSS/PC+ system file contains
       7 cases, each consisting of
       4 variables (including system variables).
       4 variables will be used in this session.
```

打F1 键，选择变量列表(Var List)，其各变量内容如下：

若活动文件已经存在，将给予<active file>的提示，下面是BASETEST. INC文件生成的数据提示。

————————————<active file>————————————

Title: SPSS SYSTEM FILE. IBM PC DOS, SPSS/PC+ V3.0
Date of Creation : 12/2/93
Time of Creation : 15:46:07
Number of Variables: 18 (23)
Number of cases : 100
System File Version 2, Compressed

————————Press space to continue-————————

SPSS/PC+ 也可读入矩阵类型的数据，下面示例是上面SAS 因子分析用例在SPSS/PC+ 中实现的程序。

```
DATA LIST MATRIX FREE/ X1 TO X3.
N 88.
BEGIN DATA.
1.0000000
-.3333333  1.0000000
0.6666667  0.0000000  1.0000000
END DATA.
FACTOR READ=COR TRIANGLE /VARIABLES=X1 to X3/CRITERIA FACTORS (3)
      /EXTRACTION ML /PRINT=CORRELATION EXTRACTION ROTATION FSCORE.
```

JOIN 命令进行文件的合并(MATCH) 或追加(ADD)，其格式为：

JOIN MATCH FILE='文件名' /KEEP=变量名/DROP=变量名/RENAME (旧名=新名) /MAP (/BY 变量名)

JOIN ADD 的用法与之类似。

生成总计(AGGREGATE) 文件，其格式为：

AGGREGATE OUTFILE='文件名' /PRESORTED /BREAK=变量表(A|D) /MISSING=columnwise /AGGVAR '标号'... =函数(变量表,参数)

转入其它格式的数据文件:

TRANSLATE FROM FILE='文件名' /TYPE=WKS ... /DROP=变量表/KEEP=变量表/FIELDNAMES RANGE 范围/MAP.

转贮其它类型的文件使用TRANSLATE TO 命令，其格式与TRANSLATE FROM 相类似。可进行交换的数据格式有：LOTUS(WKS、WK1、WK3) 和SYMPHONEY 数据(WRK、WR1、SLK)，以及DBASE 数据(DB2、DB3、DB4)。

SPSS-X 格式文件转入：

IMPORT FILE='文件名' /KEEP=变量表/DROP=变量表/RENAME=旧名=新名

SPSS-X 格式文件转贮：

EXPORT OUTFILE='文件名' /KEEP=变量表/DROP=变量表/RENAME=旧名=新名/MAP /DIGITS=小数位数

§16.3.3 BMDP

这里介绍DM模块。DM模块各段落的用法可经HELP段来查看，如：HELP READ. / 给出READ段的用法。除END和FINISH外，几乎DM的所有段落中均具有FILE=c，以下除非专门指明，它都表示内部工作文件。

1. READ 段：读取系统文件。SFILE = c. 输入文件名。
   REWIND. 读取数据前重绕(rewind) 输入文件。
   FORMAT= 'c'. 输入记录格式。
   VNAME = list. 变量名。
   VARIAB= #. 变量数。
   RLABEL= v1,v2. 记录标号。
   BLANK = ZERO|MISS. 使用Fortran 类型格式时，空格的处理方法。
   MCHAR = c. 数据中的缺失值。
   CODE = c. BMDP 文件码(文件中的记录类型)。
   CONTENT= c. BMDP 文件内容。
   LABEL = 'c'. BMDP 文件标号。
   KEEP = list. 保持变量(文件中的一种记录类型)。
   DELETE= list. 删除变量(文件中的一种记录类型)。
   NEWNAME= list. 从BMDP 文件读取变量的新名.
   RECT = list. 字母数字记录类型标识。
   RECN = list. 数值类型标识。
   RECID = #1,#2. 记录标识为输入记录#1 到#2 间的字符。
   LEVEL = list. 记录类型的层次水平。
   注：使用FORMAT(t)=, VNAME(t)=, 读取多种类型的记录。

2. SORT 段：对数据文件的记录排序。
   KEY = list. 排序关键字。
   ORDER = list. 用A, D, 或C, 指示记录按排序关键字进行升序、降序、或字符类型排序。
   KEEP = list. 生成排序文件中保持的变量名。
   DELETE = list. 删除变量。
   NEWNAME= list. 排序文件变量名。
   NEWFILE= c. 输出工作文件，未指定时使用输入文件名。

3. EXTRACT 段：抽取记录和变量。
   RECT = list. 抽取的记录类型。
   KEEP(t)= list. 从类型t 中抽取的变量名。
   DEL(t) = list. 从类型t 中删除的变量名。
   NEWN(t)= list. 类型为t 的保持变量的新名。
   NEWFILE= c. 输出工作文件名。

4. CHECK 段：标记缺失值和超出指定围的值，也用于经'hot-deck' 过程填充缺失值。
   MISS = list. 每个变量的缺失值。
   MIN = list. 每个变量的最小值。
   MAX = list. 每个变量的最大值。
   HOTD = list. hot-deck 过程工作的变量。
   注：使用MISS(t)=, MIN(t)=, 等处理多种记录类型。

5. GROUP 段：指示变量分组，每种记录类型分别使用。
   RECT = list. 记录类型。
   CODE(v)= list. 变量代码。
   CUTP(v)= list. 变量分隔点。
   NAME(v)= list. 变量的分类名。
   FROMF = c. 指定含有分组信息的文件名。
   FROMV = list. 含有分组信息的FROMF 文件变量。
   VARIAB = list. 接受分组信息的变量名。

6. TRANSFORM 段：每次对一个记录有效。
   RECT = c. 转换记录类型。
   KEEP = list. 保持变量。
   DELETE = list. 删除变量。
   NEWNAME= list. 保持变量新名。
   NEWFILE= c. 输出工作文件名。
   RETAIN. 记录之间保持新变量的值，
   影响转换的语句具有形式：
   v = 表达式。
   IF (逻辑条件) THEN (操作语句.) ELSE (语句.) UNDEFINED (语句.).
   FOR v = list DO (语句.).
   WHILE (l逻辑变量) DO (语句.).
   SHOW (变量值、变量、或表达式).
   TEXT (显示文本).

7. MERGE 段：两个或多个文件的联接。
   FILES = list. 合并的工作文件名。
   KEY(f) = list. 文件f 中引导合并的变量。
   ORDER = list. A, D, 或C, 的列表，指示关键字是升序、降序或字符顺序。
   KEEP(f)= list. 文件f 保持的变量。
   DEL(f) = list. 文件f 删除的变量。
   NEWNAME= list. 合并生成文件的变量名。
   NEWFILE= c. 输出工作文件名，不指明时为第一个输入文件。

STOP = list.	若相同的关键变量值在多于一个文件内找到，用这些指示
FIRST = list.	指明保持重复的第一个、最后一个、所有的或不保持。
LAST = list.	
ALL = list.	List: 是字符串，每个的长度= 文件数。
NONE = list.	如：FIRST='.B.D','..CB'. 表明若记录在第二个和第四个
UPDATE= list.	找到，则保持第二个，若在第三个和第四个找到，则保持
PRINT = list.	第三个。

8. JOIN 段：两个或多个文件记录的合并。
   FILES = list. 工作文件名。
   KEY(f)= list. 引导合并的变量名。
   ORDER = list. A, D, 或C 的列表，指示合并顺序。
   KEEP(f)= list. 保持变量名。
   DEL(f) = list. 删除变量名。
   NEWNAME= list. 合并文件变量名。
   NEWFILE= c. 输出工作文件名，不指明为第一个工作文件。

PAD = list.	若指定在文件中不出现，HOTREC 和HOTKEY 使用前一个记
DROP = list.	录和符合条件的记录来代替缺失值。
HOTREC= list.	
HOTKEY= list.	List: 字符串；每个长度与文件数相同。
STOP = list.	如: STOP='.BC'. 表明第一个文件出现缺失时停止。
PRINT = list.	

9. AGGREGATE 段：把记录归并到新文件。
   WITHIN = list. 按WITHIN 变量分组归并(输入文件应排序)，或：
   RECT = c. 类型为c 的记录被归并。
   MAXCOPY= #. 每个归并集合的最多记录数(默认20)。
   KEEP = list. 保持变量名。
   DELETE = list. 删除变量名。
   NEWNAME= list. 保持变量的新名。
   NEWFILE= c. 输出工作文件名。
   RETAIN. 保持新变量值。
   APPEND. 追加结果到每个集合。
   影响归并的语句为：
   v = 表达式。
   IF (逻辑条件) THEN (执行语句.) ELSE (语句.) UNDEFINED (语句.).
   FOR v = list DO (语句.).
   WHILE (逻辑变量) DO (语句.).
   SHOW (值、变量或表达式).
   TEXT (要显示的文本).

10. PACK 段：从记录的集合中选出一个变量。
    WITHIN = list. 按WITHIN 的分组压缩(输入文件应排序)，或：

RECT = c. 类型为c的记录被压缩。
VARIAB = list. 待压缩的变量。
KEYS = list. 引导压缩的变量。
LEVEL = list. 关键变量的水平数。
CODE(v)= list. 关键变量的编码。
CUTP(v)= list. 关键变量的间隔。
PICK = list. 对CODES 使用EXACT, CLOSEST, BELOW, 或ABOVE。
对CUTPOINTS 使用FIRST, LAST, SMALL, 或LARGE。
MAXPAD = #. 压缩的最大记录数。
KEEP = list. 除了压缩变量外保持的变量。
NEWNAME= list. 压缩文件变量新名。
NEWFILE= c. 输出工作文件名，默认为第一个文件名。

11. UNPACK 段：把记录分解到几个记录。
VARIAB = list. 操作变量名。
SEQU = list. 产生的case sequencing 变量名。
LEVEL = list. case sequencing 水平数，或
CODE(v)= list. case sequencing 变量v 的值。
KEEP = list. 除unpacked和sequencing变量外拷贝到每个未压缩记录的变量。
NEWNAME=list. 在未压缩文件中的变量名。
NEWFILE= c. 输出工作文件名。未声明时为输入文件名。

12. PRINT 段：打印数据、文件名和内容、软件信息，亦用于控制行、页的大小和输出多少。
NAMES. 显示工作文件名、记录数、记录类型、变量名。
NEWS. 打印程序的限制及错误。
PAGE = #. 每页行数。
LINEs = #. 每行字符数。
LEVEL = c. 详细程度: MINIMAL, BRIEF, NORMAL, or VERBOSE.
POINTERS. 显示文件名、记录类型、变量名、纠错中的文件指针。
以下8 个参数用于打印数据：
FILE = c. 工作文件名。
VAR = list. 打印的变量名或指标，可以使用VAR=ALL。
FIELD = list. 打印变量的域宽。
FORMAT= 'c'. 变量的打印格式。
NUMBER= #. 每页打印的记录数。
CASES = #. 打印记录数。
HEAD = 'c'. 每页首打印的标题。
RECT = c. 记录类型名(非方形数据文件)

13. MAP 文件：打印映象，即文件中记录的结构。

WITHIN= list. 同一组记录在一条线上画出(输入文件应排序)。
VARIAB= v. 使用1,2,...,9,A,B,...,Z 编码的变量。

TIME = v. 水平轴画的变量。
DELTA = #. 水平轴的字符增量。
RANGE =#1,#2. 水平轴范围。
CUTP = list. 定义编　码的间隔值。
CODE = list. 定义编码。
SYMB = list. 代换1,2,...,9,A,B,...,Z 的记号。
RECT = c. 显示的记录类型。

14. STATISTICS 段：报告变量的统计量，可以针对部分记录。
WITHIN= list. WITHIN 变量分组(输入文件应选排序)。
VARIAB= list. 需要显示统计量的变量。
CELLWISE. 记录的所有统计量同时报告，NO CELLW 则是每变量一组。
RECT = c. 显示统计量的记录类型，参数仅用于含有多种记录类型的文件，对这种类型的每个连续记录报告统计量。

15. HISTOGRAM 段：打印变量直方图，可以针对部分记录。
WITHIN= list. WITHIN 变量分组直方图(输入文件应先排序)，或：
RECT = c. 类型c 的记录集合组成一个直方图。
VARIAB= list. 产生直方图的变量。
MIN = list. 每个变量的最小尺度。
MAX = list. 每个变量的最大尺度。

16. SAVE 段：存贮方形文件到BMDP 文件、FORMATTED 或二进制文件。

FILE = c. 存贮文件名。
SFILE = c. 操作系统认可的文件名，默认值为FILE 的文件名。
NEW. 从SFILE 中删除BMDP 文件。
CODE = c. BMDP 文件码，未指定时为FILE 的名字。
LABEL = 'c'. BMDP 文件标号，至多40 字符。
KEEP = list. 存贮的变量。
DELETE= list. 删除的变量。
FORMAT= 'c'. 输出格式或BINARY 字，默认为BMDP 文件。
RECT = c. 记录存贮类型。

17. DELETE 段：删除一个或多个工作文件。
FILE = list. 删除文件名，该选项不删除操作系统下的外部文件。

18. END 段：删除所有工作文件，但不终止程序。

19. FINISH 段：终止程序并且把控制返回到系统。

20. CONTROL 段：控制程序执行环境并且为程序错误提供诊断信息。
INTERACT. 设定执行状态为交互式，NO INTERA 则相反。
FILE = c. 程序指令从名为c 的文件读取。文件结束后返回。
MACRO = c. 宏文件名。
ERROR = c. 程序停止的水准NONE, INTERACT, NORMAL, 或STRINGENT。
DUMP. 批处理方式下，打印完整的BMDP 存贮显示，在交互方式下，显示选择的数组。
LENGTH= #. 所使用存贮区的长度，减少存贮区的大小可节省DEBUG=TEST或INFO下CPU 的时间。
DEBUG = NORM. 不做特殊纠错。
TRACE. 报告子程序进入/退出信息。
TEST. 用期性检查内存。
INFO. 检查内存，空间分配、GETME 调用以及子程序跟踪情况。

把第 6 章例6.4的程序增加SAVE段，指定存于文件2L，编码为2L，程序为：

```
save code='2L'. content=data. new. file is '2L'. code='2L'./
```

现转入DM模块，用READ段读取，指定工作文件为UU：

```
read sfile='2L.'. file='uu'. code='2L'./
```

使用PRINT段浏览其内容：

```
print head='The Data of Example 5.3'. names. pages=60. lines=72. var=all. /
```

利用SORT段对数据排序，并使用MAP段：

```
sort key=group, survival. /
map variables=survival. within=group. /
```

利用STATISTICS段求取描述统计量：

```
statistics variables=survival. within=group. /
```

利用HISTOGRAM段绘直方图：

```
histogram variables=survival. within=group. /
```

这里也举一个读取矩阵数据的例子[3]，程序如下：

```
problem   title is 'BMDP4M'./
input     type is correlation.
          shape is square.
          variables are 12.
          format is '(12F5.0)'./
variables names are le,aso,in,..../
plot      initital is 0.
          final is 0.
          fscore is 0.
```

```
factor    method is pf.
          constant is 0.
          iterate is 25.
          commun is smcs.
rotate    method is vmax.
          normal./
end/
```

...数据矩阵...

程序调用4M模块进行因子分析，同一般程序一样，分为两部分。第一部分是数据的录入，内容有变量数、变量名、数据的形式、数据的格式。第二部分是进行因子分析，包括因子的抽取方法、因子数目、禄始公因子方差的选择、旋转方法。

数据的类型有DATA,CORR,COVB,LOAD,FSCF几种， 形状有方SQUARE,LOWER两种，格式可以是固定的或自由(FREE)的。因子分析方法有PC,PF,ML,LJ几种，公因子方差除了SMCS以外有UN,SM,MAX几种，旋转方法有VMAX,NONE,QRMAX, EQM, DQ, DOBL ,ORTHOG,ORHTOB几种。

§16.3.4 SYSTAT

源于SYSTAT 模块化的特点，SYSTAT 专用DATA 模块进行数据管理并为后续的分析作准备，SYSTAT 的语言特色也在该模块体现得最好，如一系列统计函数和类似BASIC 语言的语句。运行后在系统指示下打入EDIT，即进入全屏幕编辑方式。SYSTAT 4.0 提供了专门的EDIT 模块进行全屏幕数据管理，在DATA/EDIT 模块经SWITCH TO 命令转入其它分析模块。实际上，不同模块间可经此命令相互切换。下面是一个全屏幕编辑运行的示意图，在DATA 块中打入命令USE IRIS 和EDIT，系统进入编辑状态。第一列是例号，后面几列是对应各变量的观察值。

SYSTAT Editor IRIS.SYS

Case	SPECIES	SEPALLEN	SEPALWID	PETALLEN	PETALWID
142	3.000	6.900	3.100	5.100	2.300
143	3.000	5.800	2.700	5.100	1.900
144	3.000	6.800	3.200	5.900	2.300
145	3.000	6.700	3.300	5.700	2.500
146	3.000	6.700	3.000	5.200	2.300
147	3.000	6.300	2.500	5.000	1.900
148	3.000	6.500	3.000	5.200	2.000
149	3.000	6.200	3.400	5.400	2.300
150	3.000	5.900	3.000	5.100	1.800
151					

编辑时可用ESC 键进行命令行与数据录入间的切换。输入的变量为字串类型时，变量名前应导以单引号(')。在命令行打入HELP有下面的菜单提示。

EDIT 产生和编辑SYSTAT 文件。

光标命令以及等效功能键(alternative keys) 有:

←(Cntl-S)	→(Cntl-D)	↓(Cntl-X)
Ins (page left, Cntl-A)	Del (page right, Cntl-F)	PgDn (Cntl-C)
PgUp (Cntl-R)	Home (Cntl-W)	End (Cntl-Z)

编辑(EDIT) 命令有:

Esc (Cntl-Q)	进行数据窗与命令行间的切换。
USE <file name>	把数据区用SYSTAT 文件填充。
SAVE <file name>	把工作区存入一个SYSTAT 文件。
FEDIT <filename> |*| > |#|	启用SYSTAT 的文本编辑器。
FIND <expression>	把光标移至被选定的观察号。
FORMAT <#>	设定显示的小数位数。
FPATH <path>/GET|OUTPUT|SAVE| SUBMIT|USE|FEDIT|TRANSFER	给指定的文件设定路径前缀。
LET <var>=<expression>	转换或生成变量。
IF <expression> THEN LET <var>=<expression>	条件转换。
REPEAT <#>	把工作区填充至指定数目<#> 的观察。
TYPE <type of matrix>	指示CORR, COVARIANCE 等类型。
HELP <command>	提示信息。
NEW	清除工作区以供新数据集使用。
DOS 'DOS command'	执行一个MS-DOS 或PC-DOS 命令。
SWITCHTO 'module'[<file>/ECHO]	切换至另一个SYSTAT 模块。
END or QUIT	返回DOS

系统文件的左右合并是通过USE 命令来完成的，要进行这种合并，只需同时指示几个文件名及其相应的变量。SYSTAT 以PUT/GET 命令存/取一个ASCII 文件。GET 在读入一个文件时，要求首先要运行SAVE 命令指示要存贮的系统文件名。当读入ASCII 数据宽度超常时，应用LRECL= 命令指示记录的宽度。SYSTAT 活动数据的转置也用TRANSPOSE 命令。

SYSTAT 使用IMPORT 命令把一个外部文件转换成SYSTAT 文件，其句法是：

```
IMPORT <file> [(<var1>,<...>)] /,
 TYPE= LOTUS | LOTUS2 | SYMPHONY | SYMPHONY11 | DBASE2 | DBASE3 | DIF |
 MAP | PORTABLE [RANGE=<range>] [ROWS=<#>-<#>]
```

如：IMPORT 'MYFILE.WK1' / TYPE=LOTUS2 ROWS=1-50　意为把Lotus 1-2-3 第二版的文件MYFILE.WK1 转换成SYSTAT 文件，仅仅使用1-50 行的内容。

EXPORT 命令把一个SYSTAT文件转成其它格式的文件，其句法是：

```
EXPORT <file> [(<var1>,<...>)] /,
 TYPE= LOTUS | LOTUS2 | SYMPHONY | SYMPHONY11 | DBASE2 | DBASE3 | DIF |
 MAP | PORTABLE [ROWS=<#>-<#>]
```

如：EXPORT LOTUSFIL / TYPE=LOTUS2 ROWS= 1- 50　由内存的文件生成一个Lotus1-2-3 文件。

微机SYSTAT 的一个特点是提供了PC 与Macintosh 机间SYSTAT 文件的转换功能。

§16.3.5 Stata

Stata 的系统文件以.DTA 作为扩展名，这类文件存贮了数据的格式、标签等。

命令use 从磁盘调一个Stata 格式的数据到内存，其格式为：

use 文件名[, clear nolabel]

clear 允许所有情况下调入内存，不论内存改动的数据是否已存盘。

nolabel 不允许存贮的数据中的标号被调入。

使用describe using 文件名可以浏览文件的内容。

存贮Stata 格式数据的命令是：

save filename [, replace nolabel]

replace 允许覆盖已存在的数据集。

nolabel 省略数据集中的村标号。

文件扩展名不指定时，用.dta 或者.xp，此时文件含有一个交叉乘积矩阵。

如：save myfile

File myfile.dta already exists

r(602) ;

系统报错，增加选项replace，命令为：save myfile, replace

存贮的数据将以压缩二进制格式存放。

Stata 的.DCT 文件包含一个数据字典，它描述了文件所含的变量及其格式、标号，以及数据存放方式等，数据可以在这些描述的下面或者放在其它文件，但它仍然是一个ASCII 文件。Stata 读取该格式的数据时需要指示dictionary 选项。其它软件可以以固定格式或自由格式读取。Stata 专门用.RAW 扩展名指示ASCII 格式的数据文件，infile/outfile 命令读取/存贮ASCII格式的数据文件。infile 的格式是：

infile [变量表[_skip[(#)] [变量表[_skip[(#)] ..]]]] using 文件名[in 范围] [if 表达式] [, automatic byvariable(#)]

文件默认具有.RAW 扩展名。指定变量列表时，文件中所含数据是自由格或用逗号分开的。若不指示变量列表，则文件中含有一个数据字典。若文件扩展名未指示，则隐含使用.DCT。

执行infile 命令时，内存中不应该有数据，这可以预先执行drop _all 命令。也可参照help maxvar 给出的说明。

现有数据文件为myfile.raw，内容为

1 2 3 1, 2 3

4 5 6 或4,5 或1 2 3 4,5 6

6

三个变量读入时赋为A、B、C，则可用命令: infile a b c using myfile

变量列表中使用_skip可以跳过一些量，如：

infile a _skip c using myfile

infile _skip(2) c or infile _skip _skip c

第一句只读变量a 和c，而第二句则只读变量C。

infile str20 name age sex using myfile

infile str20(name age) sex using myfile

infile str20 name age int sex using myfile

第一句中读入量name 为长度为20 的字串，age 和sex 为浮点数。第二句中name 和sex 为字串而sex 为浮点数，第三句中name 为字串，age 为点数，sex 为整数。

infile 也可以读入非数值类型的变量并产生数值类型标号，可用automatic 来做到，如对数据：

”James Smith” 38 ”male”

Branton 32 female

”Bill Ross” 27,’male’

用命令：infile str20 name age int sex:sexlbl using myfile, automatic

整型量sex 对于male 将取值为0，对于female 则取值为1。

与infile 相仿，outfile 则是把数据以ASCII 码形式写入磁盘，语法为：

outfile [变量表] using 文件名[if 表达式] [in 范围] [, comma dictionary nolabel replace]

最后的选项指定Stata 生成用逗号分隔或字典格式文件，nolabel以数值记录标号变量的值。

Stata 管理数据有以下约定：

. 字串总是用双引号括起。

. 除comma 以外的所有格式，数据均以表的形式存贮。

. outfile 的行不超于80 个字符，因而一个观察可以点数据文件的几行上。

. 在comma 格式中，数值缺失值记作”,,”，否则以圆点”.”存贮。

. 所有格式的缺失字串均以双引号(””) 记录。

在Stata内部录入数据之前，消除已调入的数据，使用命令drop _all 和label drop _all 消除内存数据和标号。录入数据只消使用命令input 变量列表，如：

```
. input id mpg weight price
           id        mpg     weight      price
  1. 1 22 2930 4099
  2. end
```

每次录入以end 结束，一旦内存数据生成，继续录入使用input 即可。

```
. input
          id        mpg     weight      price

2. 2 17 3350 4749
3. 3 22 2640 3799
4. 4 20 3250 4816
5. 5 15 4080 7827
6. end
```

录入字符量时应施以前缀str#，#的取值为2 至80。因为未声明时，默认为float，如：

```
. drop _all
. input str14 make mpg weight price
```

大批量数据这样的录入仍然很繁琐，使用infile 读入ASCII 数据较方便。

对DOS 用户来说，Stata 可以读入Lotus, Symphony, dBase, Gauss, SPSS，或SYSTAT 格式的数据，这个工具称为Stat/Transfer，它是一个菜单驱动的模块。

§16.3.6 DBMS/COPY

可以转换的数据类型有：Lotus, Quatro, Clipper, Database, Smart, ASCII, ACT!, Datalex, ABstat, Bass, BMDP, CSS, 4CasT/2, Forecastpro,Microstat -II, NCSS, Probe, RATF, SigmaPlot, StatGraphics, SyGraph, Excel, Autobox, Gauss, GLIM, Minitab, SAS, SCA, Soritex, SPSS, Stata, Statpac, SYSTAT，S-Plus等也就是说，本书涉及的多数软件可经它转换。该软件使用方便，在ASCII 转至其它有格式文件时也生成一个数据字典。

以SPSS/PC+为例，它可以调用DBMS/COPY进行与其它软件包间的数据交换。

```
DBMSCOPY FROM ' ' TO ' '.
PLOT /plot y with x.
NPPlot/ variables x.
QED
SET BOX='-|-++++....'.
```

公司的地址：P.O.Box 56627, Houston, TX 77256, 电话(800) Stat-wow即(800)7827-969。

§16.4 数据交换用例

不同计算机系统和软件间数据与程序的交换交换受许多因素的制约，考虑经过网络传输时，要对计算机网络有所了解，如DOS与VMS系统间的交换，常用的途径有：

```
. Pathworks (PCSA), DEC
. PC-NFS, SUN Microsystem
. DECNET, DEC
. Kermit, Columbia University (treeware)
```

对于DOS和UNIX之间的交换，常用PC-NFS、TCP/IP和Kermit。

§16.4.1 程序交换用例

【例16.6】VAX/VMS SAS 样本程序的使用

VAX/VMS SAS 6.07 提供了许多.SAS 样本文件，通过网络传输至PC 机时，由于其仅仅使用换行符，没有硬回车，致使大多数PC 机的编辑软件不能调用，也不能在PC SAS 下调入，这时可以使用DOS 5.0 中的编辑EDIT，也可以采用下面的BASIC 程序进行转换：

```
INPUT "请输入文件名";INP$
INPUT "请输出文件名";OUT$
OPEN INP$ FOR INPUT AS #1
OPEN INP$ FOR OUTPUT AS #2
WHILE NOT EOF(1)
   A$=INPUT$(1,1)
   IF A$=CHR$(10) THEN PRINT #2, ELSE PRINT #2,A$;
WEND
CLOSE #1,#2
END
```

运行时指定VAX/VMS SAS 样本程序为源文件，指定PC 机文件名为目标文件。程序的处理办法是把换行符换成DOS 下的回车换行符，这样一处理，就可以在微机上正常编辑使用了。第三章介绍的DOS2UNIX/UNIX2DOS功能与之类似。

【例16.7】VAX/VMS 下SAS 对BMDP 的调用

由于BMDP 模块化的特性，使得在SAS 内启用BMDP 很方便。下面是SAS 用户手册上的用例，建立数据集，启用BMDP 进行分析，产成结果用CONVERT 过程转入SAS。

```
DATA TEMP;
   INPUT A B C@@; CARDS;
   1 2 3 4 5 6 7 8 9
PROC CONTENTS;
   TITLE 'CONTENTS OF SAS DATA SET TO BE RUN THROUGH BMDP1D';
PROC BMDP PROG=BMDP1D DATA=TEMP;
   PARMCARDS; /* 指示 BMDP 语句引用开始 */
   /PROB TITLE='SHOW SAS/BMDP INTERFACE'.
   /INPUT UNIT=3. CODE='TEMP'.
   /SAVE CODE='BOUT'. NEW. UNIT=4.
   /END
   /FINISH
;
PROC CONVERT BMDP=FT04F001 OUT=FROMBMDP;
PROC CONTEMTS;
   TITLE 'SAS DATA SET CONVERTED FROM BMDP SAVE FILE';
PROC PRINT;
```

§16.4.2 数据交换用例

【例16.8】用SAS/RTERM 进行PC 与VAX/VMS SAS 数据交换

卫生部进行1991 年全国医院行业“纠风”的调查时，对门诊病人、住院病人或出院病人进行询问，以了解不同级别医院、不同科别的病人就医时对医生、护士的满意情况，同时看病人在医疗福利类型的影响情况，共调查 2 万余例。在dBASE lll数据文件约7 兆，生成SAS数据文件约10 兆，在微机上制表太费时间，此时拟转至VAX 机上完成。

利用PROC DOWNLOAD 过程. 在AUTOEXEC.SAS 中已用语句filename rlink 'd:\sas\saslink\logvms.scr'; 进入SAS 系统，在PGM 窗口命令行上打入: SIGNON，据提示输入帐户名和口令进行登录。然后远程提交(rsubmit) 程序。

```
filename pc 'd:\dBASE3\bank.dbf';
proc dbf db3=pc out=bank;
run;
DM 'rsubmit';
libname user '[]';
proc upload data=bank out=user.file;
run;
```

则把数据库文件传至VAX 机，数据库的格式亦完整地由PC 过录到VAX 机。由于计算机网络的发达，使用SAS的传输格式更为方便：

```
libname us xport 'us.tds';
libname counties xport 'counties.tds';
proc copy in=maps out=us mtype=data;
  select us;
run;
proc copy in=maps out=counties mtype=data;
  select counties;
run;
```

将SAS/GRAPH中maps的图形数据集转为传输格式，在主机上使用类似的语句转为SAS数据集。

【例16.9】SPSS/PC+ 到VAX/VMS SAS 数据的转用

北京阜外医院心外科拟进行心脏瓣膜移植的研究，该医院拥有SPSS/PC+ 软件。由于被分析的变量和生成的变量数目很大，超过了128 个，故不用dBASE Ⅲ的格式存放而改用SPSS/PC+ 格式存放，为了能在VAX/VMS SAS 上使用，首先把SPSS/PC+ 格式换成小型机上SPSS-X格式，使用EXPORT 命令，文件经DECnet 直接拷贝到VAX 机。然后于VAX 机重新登录，运行SAS 软件和启用转换的数据集。

由上例的做法得到启示，1992 年医院满意度调查分析，直接经SPSS/PC+ 进行转换，速度可以改善。程序如下：

```
SET /MORE OFF /LISTING='D:\FOX\BANK.LOG'.
trans from '\fox\bank.dbf' /TYPE DB3.
EXPORT /OUTFILE '\fox\bank.sys' /MAP.
EXIT.
```

把程序运行情况存于BANK.LOG，第二句把BANK.DBF 转成SPSS/PC+ 文件，第三句进行转换，结果生成bank.sys，转换时用MAP 选项列出转换信息。仍经DECnet 网把bank.sys 传至VAX。VAX 上的文件转换程序为：

```
filename user 'bank.sys';
proc convert SPSS=user out=data1;
run;
libname user '[]';
data user.file;
set work.data1;
run;
```

启用专用过程PROC CONVERT，程序运行结果是在VAX 机当前目录下生成一个名为FILE的SAS 系统文件。

§16.4.3 综合用例

【例16.10】1993 年卫生部行业纠风调查的数据处理是比较典型的，现加以介绍。调查分

为几个步骤，正如第二章介绍的那样，首先确定调查时间、对象、内容，调查一览表与病人问卷表等。

医院抽法：拟按部级、省级和地市级(计划单列市)几个水平。因为分层后可能不足，最终进行个别调整。

抽样框架的确定：利用卫生单位代码库作为原始抽样框(FRAME) ，它是一个dBASE格式的数据库文件，含有各单位的行政区划代码、相应的医院情况信息等，这样就能利用dBASE或FoxBASE+中的SET FILT TO 命令计算出各种条件下医院的总数，其信息可用SET ALTE TO 命令和? 命令到文本文件中，设计数行用*** 作标记，则用BASIC 程序读取之，结合随机函数将随机号与流水号以及医院名称等有关信息同量连续列出，设列表文件为SAMPLED.TXT，BASIC程序如下：

```
OPEN "I", #1, "sampled.txt"
OPEN "O", #2, "result"
RANDOMIZE TIMER
DO WHILE NOT EOF(1)
    LINE INPUT #1, line$
    IF INSTR(line$, "***") <> 0 THEN
       num = VAL(LEFT$(line$, 10))
       print #2, line$
    ENDIF
    i = 1
    WHILE i <= num
       LINE INPUT #1, line$
       index$ = STR$(i)
       sel$ = STR$(INT(RND * num) + 1)
       line$ = index$ + SPACE$(5 - LEN(index$)) + sel$ +
                SPACE$(5 - LEN(sel$)) + line$
       PRINT #2, line$
       i = i + 1
    WEND
    LOOP
END
```

随机数的种子是当前时间(TIMER)，随机数范围随医院的数目而定。

抽样框的形式为：

⋆⋆⋆⋆ 医院数目

随机号	流水号	行政区划码	医院名称	床位数	.....
xxxxxx	xxxxxx	xxxxxxxxxx	xxxxxxxx	xxxxxx	xxxxxx

从任意一个随机号开始，读取几个随机号，其内容到对应的流水号中查取即得到抽中的医院号，这样做法也避免了每年编程抽取的麻烦。

编程与发盘。91 年、92 年采用dBASEIII/FoxBASE+ 程序，当时EPI INFO 还没有汉化，而且部分省市没有286 以上计算机，其CGA或MDA显示器只能显示10 行汉字。进行此选择是比较合适的。93 年度由于已在几个调查中使用，所有省市卫生厅已配备286计算机或具备使

用EGA/VGA 显示器、25 行汉字的能力，采用汉化EPI INFO 是必要的，也考虑今后对该软件的进一步推广应用。利用原始调查表，制做.QES 文件，为后继SAS 软件的处理，变量名大多使用英文，放在大括号内，设其文件名为SURV.QES，其内容见【例15.3】。

以上程序下发的同时，还提供了往年数据操作处理的样本程序，以及相应的EPI INFO 分析模块，这样在地方水平上也能够方便产出。

使用ENTER 录入，生成数据文件SURV. REC，则可用于生成dBASE 或FoxBASE+ 文件，或SAS、SPSS 文件了。由于全国单位较多，可将转换过程编入DOS 批处理文件CONV.BAT，其内容如下：

```
REM 本程序用于将EPI INFO 文件转成DBASEIII。
REM DOS 文件名至多有八个字符，故有一些省市名不完全。

convert ANHUI    ANHUI    8 Y
convert BEIJING  BEIJING  8 Y
convert FUJIAN   FUJIAN   8 Y
convert GANSU    GANSU    8 Y
convert GUANGDON GUANGDON 8 Y
convert GUANGXI  GUANGXI  8 Y
convert GUIZHOU  GUIZHOU  8 Y
convert HAINAN   HAINAN   8 Y
convert HEBEI    HEBEI    8 Y
convert HEILONGJ HEILONGJ 8 Y
convert HENAN    HENAN    8 Y
convert HUBEI    HUBEI    8 Y
convert HUNAN    HUNAN    8 Y
convert JIANGSU  JIANGSU  8 Y
convert JIANGXI  JIANGXI  8 Y
convert JILIN    JILIN    8 Y
convert NEIMENG  NEIMENG  8 Y
convert NINGXIA  NINGXIA  8 Y
convert QINGHAI  QINGHAI  8 Y
convert SHANDONG SHANDONG 8 Y
convert SHANGHAI SHANGHAI 8 Y
convert SHAANXI  SHAANXI  8 Y
convert SHANXI   SHANXI   8 Y
convert SICHUAN  SICHUAN  8 Y
convert TIANJIN  TIANJIN  8 Y
convert XINJIANG XINJIANG 8 Y
convert YUNNAN   YUNNAN   8 Y
convert ZHEJIANG ZHEJIANG 8 Y
convert LIAONING LIAONING 8 Y

REM 完成!!!
```

由于各省的数据库中没有省名信息，拟最终合并时加入，考虑到EPI INFO 虽然录入方

便，但它对数据细处的操作不如dBASEIII或FoxBASE+，因此仍用后者编程解决。为了处理的方便，将各省报盘相同的文件名SURV.REC (或其它改动后的文件名) 拷入子目录时即改名为相应的省名，如：

COPY A:SURV.REC BEIJING.REC

这样前面的格式转换程序和下面的数据合并程序都可以自动完成。

合并程序由两部分完成，即对各省调用替换、追加和具体追加和替换，对应的文件是ASSE.PRG和REPL.PRG，其内容为：

⋆ ASSE.PRG

⋆⋆⋆ 把现有数据合并起来

```
set safe off
set echo off
set talk off
sele 1
use
sele 2
use
use bank
zap
set safe on
do repl with 'ANHUI'
do repl with 'BEIJING'
do repl with 'FUJIAN'
do repl with 'GANSU'
do repl with 'GUANGDON'
do repl with 'GUANGXI'
do repl with 'GUIZHOU'
do repl with 'HAINAN'
do repl with 'HEBEI'
do repl with 'HEILONGJ'
do repl with 'HENAN'
do repl with 'HUBEI'
do repl with 'HUNAN'
do repl with 'JIANGSU'
do repl with 'JIANGXI'
do repl with 'JILIN'
do repl with 'NEIMENG'
do repl with 'NINGXIA'
do repl with 'QINGHAI'
do repl with 'SHANDONG'
do repl with 'SHANGHAI'
do repl with 'SHAANXI'
```

```
do repl with 'SHANXI'
do repl with 'SICHUAN'
do repl with 'TIANJIN'
do repl with 'XINJIANG'
do repl with 'YUNNAN'
do repl with 'ZHEJIANG'
do repl with 'LIAONING'
set talk on
retu
```

程序首先把数据库BANK.DBF的内容清空，用SET SAFE OFF 指定操作为自动。注意数据库BANK 可由空的SURV.REC 文件经CONVERT 而来，但必须事先在dBASE 或FoxBASE+下运行了MODI STRU 命令以增加字段PROV 用于存贮省名。

```
* REPL.PRG
** 用于进行全国数据汇总
parameter name
if .not.file("&name..dbf")
   retu
endi
sele 1
use &name
num=recc()
use
sele 2
use bank
appe from &name
skip -num+1
repl next num prov with '&name'
retu
```

REPL.PRG 的操作有两部分，即在第一个区内对各省数据库&NAME 记录计数，第二部分在第二区对已追加到BANK.DBF 参加计数的记录进行PROV变量赋值。

设PC SAS 上的AUTOEXEC.SAS 文件内容为：

```
filename rlink 'd:\sas\saslink\decnet.scr';
options remote=cterma;
run;
```

SAS/BASE 中的SASZRLNK.EXE 应进行替换。

在DOS 下设定环境变量CTERMA，如设VAX 主机节点名为VAX1，则设法很简单，使用DOS 命令：SET CTERMA=VAX1 即可，设定后可单独打SET 命令确实一下，此步可参照CTERM 中的READ.ME 进行。

运行STARTNET.BAT 上DECnet 网，进入SAS，并在命令行打入命令：

SIGNON

系统提示账户和口令，系统自动进行远程调用进入VAX/VMS SAS，可以使用远程提交命令RSUBMIT 了，为了简便，上述过程可放在SAS 中的显示管理命令DM 中，转贮数据集的程序为：

```
libname local '.';
filename bank 'd:\surv\bank.dbf';
proc dbf db3=bank out=local.bank;
run;
dm 'rsubmit';
libname remote '[]';
proc upload data=local.file out=remote.file;
run;
```

BANK.DBF 是全国数据库，用PC SAS 转成SAS数据集放在LOCAL库名即SAS 子目录下，经VAX主机SAS PROC UPLOAD 完成数据转贮。

制表分析程序为：

```
/*********************************************
标题：医院满意度调查分地区、分省份描述分析
作者：卫生部卫生统计信息中心
日期：1994 年1 月

产品：SAS/BASE
过程：DBF、FORMAT、DATASETS、TABULATE
*********************************************/


title1 '1993 年度纠正医院不正之风调查汇总表';
proc printto print='d:result' new;
run;
dm 'rsubmit';
libname remote '[]';
options ls=130 ps=300 nodate formchar=' ----------';
options missing=' ' nocenter mprint;
proc format;
     value yesno 1='是'      2='否';
     value ssfmt 1='门诊'    2='住院'     3=' 出院';
     value  i    1='省级'    2='地市级'   3=' 区县级';
     value ii    1='内科'    2='外科'     3=' 妇产科'
                 4='儿科'    5='中医科'   6=' 眼科'
                 7='耳鼻喉科'8='皮肤科'   9=' 口腔科' 10='其它';
     value iii   1='工人'    2='农民'     3=' 军人'
                 4='教师等'  5='干部'
```

```
                6='个体'    7='商业服务'
                8='离退休' 9='无职业者' 10='学生'   11='其他';
value iv        1='公费'    2='劳保'      3='半自费'
                4='自费'    5='商业性医疗保险';
value sat       1,2='满意较满意'    3='一般'
                4,5='不满意很不满意'6='没接触';
value vi        1,2='好较好'        3='一般'
                4,5='不好很不好'    6='说不好';
value viii      1,2='好较好'        3='一般'
                4,5='不好很不好'    6='没接触';
value ix        1,2='信任较信任'    3,4=' 不信任很不信任'
                5='说不好';
value value 1='50 元以下'       2='50-'3='100-'
                4='200-'            5='500 以上';
value for       1='出于感激'        2=' 想得到方便和照顾'
                3='担心不认真看病' 4='受人影响' 5='暗示的'
                6='直接索要';
value act       1,2='拒绝收和事后退还' 3='照价付了钱'
                4='付部分钱'          5,6=' 推辞过和没有拒绝';
value $prov '安徽省'
                'Abeijing'  ,'BEIJING'='北京市'
                'Mfujian'   ,'FUJIAN'='福建省'
                'Zgansu'    ,'GANSU'='甘肃省'
                'Sguangdong','GUANGDON'='广东省'
                'Tguangxi'  ,'GUANGXI'='广西'
                'Wguizhou'  ,'GUIZHOU'='贵州省'
                'Uhainan'   ,'HAINAN'='海南省'
                'Chebei'    ,'HEBEI'='河北省'
                'Hheilongj' ,'HEILONGJ'='黑龙江省'
                'Phenan'    ,'HENAN'='河南省'
                'Qhubei'    ,'HUBEI'='湖北省'
                'Rhunan'    ,'HUNAN'='湖南省'
                'Jjiangsu'  ,'JIANGSU'='江苏省'
                'Njiangxi'  ,'JIANGXI'='江西省'
                'Gjilin'    ,'JILIN'='吉林省'
                'Eneimeng'  ,'NEIMENG'='内蒙古'
                'bningxia'  ,'NINGXIA'='宁夏'
                'aqinghai'  ,'QINGHAI'='青海省'
                'Oshandong' ,'SHANDONG'='山东省'
                'Ishanghai' ,'SHANGHAI'='上海市'
                'Yshaanxi'  ,'SHAANXI'='陕西省'
                'Dshanxi'   ,'SHANXI'='山西省'
```

```
                'Vsichuan'  ,'SICHUAN'='四川省'
                'Btianjin'  ,'TIANJIN'='天津市'
                'cxinjiang' ,'XINJIANG'='新疆'
                'Xyunnan'   ,'YUNNAN'='云南省'
                'Lzhejiang' ,'ZHEJIANG'='浙江省'
                'Fliaoning' ,'LIAONING'='辽宁省';
run;
%macro format;
     class prov rprov ss _numeric_;
     keylabel n='计数' all='合计';
     format ss  ssfmt.   n1  i.     n2  ii.    n3  iii.
            n4  iv.      n5  sat.   n6  vi.    n7
            n8  viii.    n9  ix.    NA NB NC ND sat.
            NE0 yesno. NE1 yesno. NE2 yesno. NE3 value.
            NF0 yesno. NF1 value. NF2 for.   NF3 act.
            NG0 yesno. NG1 value. NG2 for.   NG3 act.
            prov rprov $prov.
%mend;
%macro table(a,b,box);
       table &a all,all &b*(n pctn<&a all>='列 %'*f=5.2
             pctn<&b all>='行 %'*f=5.2)/rts=16 box=&box;
%mend;
data remote.tran;
   set remote.file;
   length rprov $20.;
   if prov='ANHUI'     then rprov='Kanhui';
   if prov='BEIJING'   then rprov='Abeijing';
   if prov='FUJIAN'    then rprov='Mfujian';
   if prov='GANSU'     then rprov='Zgansu';
   if prov='GUANGDON'  then rprov='Sguangdong';
   if prov='GUANGXI'   then rprov='Tguangxi';
   if prov='GUIZHOU'   then rprov='Wguizhou';
   if prov='HAINAN'    then rprov='Uhainan';
   if prov='HEBEI'     then rprov='Chebei';
   if prov='HEILONGJ'  then rprov='Hheilongj';
   if prov='HENAN'     then rprov='Phenan';
   if prov='HUBEI'     then rprov='Qhubei';
   if prov='HUNAN'     then rprov='Rhunan';
   if prov='JIANGSU'   then rprov='Jjiangsu';
   if prov='JIANGXI'   then rprov='Njiangxi';
   if prov='JILIN'     then rprov='Gjilin';
   if prov='NEIMENG'   then rprov='Eneimeng';
```

```
   if prov='NINGXIA'   then rprov='bningxia';
   if prov='QINGHAI'   then rprov='aqinghai';
   if prov='SHANDONG'  then rprov='Oshandong';
   if prov='SHANGHAI'  then rprov='Ishanghai';
   if prov='SHAANXI'   then rprov='Yshaanxi';
   if prov='SHANXI'    then rprov='Dshanxi';
   if prov='SICHUAN'   then rprov='Vsichuan';
   if prov='TIANJIN'   then rprov='Btianjin';
   if prov='XINJIANG'  then rprov='cxinjiang';
   if prov='YUNNAN'    then rprov='Xyunnan';
   if prov='ZHEJIANG'  then rprov='Lzhejiang';
   if prov='LIAONING'  then rprov='Fliaoning';
run;
proc datasets library=remote;
     modify tran;
     label prov ='省市名'     rprov='省市名'
      name ='医院名称'        id1='医院编号'
      id2  ='病人编号'        ss ='病人类别'
      n1 ='在哪级医院'        NE0='是否托关系'
      n2 ='在哪科就医'        NE1='本院职工'
      n3 ='主要职业'          NE2='中间收礼'
      n4 ='费用支付方式'      NE3='托人价值'
      n5 ='医生服务态度'      NF0='送钱物'
      n6 ='医生的技术'        NF1='送礼价值'
      n7 ='护士服务态度'      NF2='送礼原因'
      n8 ='护士的技术'        NF3='送礼态度'
      n9 ='医疗质量是否信任'  NG0='宴请'
      na ='挂号处态度'        NG1='宴请价值'
      nb ='药房态度'          NG2='宴请原因'
      nc ='医院膳食'          NG3='宴请态度'
      nd ='环境卫生'
;
     format ss _numeric_ 20.;
run;
proc tabulate data=remote.tran f=6. noseps;
     %format;
     %table(rprov,n1,' ');
     %table(rprov,n2,' ');
     %table(rprov,n3,' ');
     %table(rprov,n4,' ');
     %table(rprov,n5,' ');
     %table(rprov,n6,' ');
```

```
     %table(rprov,n7,' ');
     %table(rprov,n8,' ');
     %table(rprov,n9,' ');
     %table(rprov,na,' ');
     %table(rprov,nb,' ');
     %table(rprov,nc,' ');
     %table(rprov,nd,' ');
     %table(rprov,ne0,' ');
     %table(rprov,nf0,' ');
     %table(rprov,ng0,' ');
run;
proc tabulate data=remote.tran f=6. noseps;
     where ne0=1;
     %format;
     %table(rprov,ne1,'托关系者分类');
     %table(rprov,ne2,'托关系者分类');
     %table(rprov,ne3,'托关系者分类');
proc tabulate data=remote.tran f=6. noseps;
     where nf0=1;
     %format;
     %table(rprov,nf1,'送钱物者分类');
     %table(rprov,nf2,'送钱物者分类');
     %table(rprov,nf3,'送钱物者分类');
proc tabulate data=remote.tran f=6. noseps;
     where ng0=1;
     %format;
     %table(rprov,ng1,'宴请者分类');
     %table(rprov,ng2,'宴请者分类');
     %table(rprov,ng3,'宴请者分类');
run;
```

上述程序中，LIBNAME REMOTE '[]'; 语句在VAX 机生成一个逻辑库名。程序首先进行格式和宏定义，并使用汉字，调用时很简便，为了看其运行的具体程序，则在OPTIONS 语句中指定MPRINT 打印。注意省名在具体使用时，在各省的汉语拼音前加了一个A-Z 等的序号，这样产出时是按照全国大区分的，但格式化过程的说明可以同时指定，其它如“好”、“较好”一类的合并也是如此。TABULATE 的选项中特别指定了参数FORMCHAR='————-' 这样产出的表是一般统计学书上所习惯使用的格式，过程也用WHERE 语句限定处理符合条件的数据子集。

最后得到完整的汇总产出表。

运行RSUBMIT后，仍用SIGNOFF对系统复位。

其它方案也是可行的，如合并成大的dBASE数据库文件后，可利用EPI INFO的CONVERT功能直接转到SAS 的数据步程序，仍然用VAX/VMS SAS。但可能由于系统间的差异，转成的

程序在运行时会报一些错误；也可以经SPSSX 的格式传输，进而利用SAS PROC CONVERT。

分析报告的撰写，实际是以上工作的小结。

第十七章　高分辨统计图形

§17.1　统计图形与图形格式

图形是重要的描述和展示手段，这里略提一下图形的格式，随后介绍几种有代表性的软件包如SAS、SPSS、Stata 以及AutoCAD、Harvard Graphics的处理方法。它们均可生成HP的描述格式(HPGL)，WordPerfect 5.1采用实用程序GRAPHCNV. EXE 将之转为.WPG格式供嵌于文本。汉化AutoCAD 与SAS/GRAPH还能够加注汉字，常用的格式有：

CGM	ANSI Computer Graphics Metafile
CGM	CGM Harvard Graphics
CGM	CGM Lotus Freelance Plus
PCX	PC Paintbrush Bitmap
TIF	TIFF Bitmap
BMP	Windows Bitmap
WMF	Windows Metafile
(E)PS	(Encapsulated) Postscript
DXF	Autocad DXF file
PLT	Hewlett Packard Graphics Language

BMP是WINDOWS PAINTBRUSH采用的格式，TIFF是FAX传输中的标准，由多数扫描器支持。MS Word支持的格式是.IMG。PostScript是由Adobe 开发的一种页描述语言，以文本形式由标准输出设备解释。EPS 也是用文本文件存贮重构的图形。如WordPerfect 支持的格式有：

```
Compuserve GIF              *.gif
JPEG                        *.jpg
Bitmap                      *.bmp
Computer Graphics Metafile  *.cgm
Encapsulated Postscript     *.eps
HP Graphics Language        *.hpg
PC Paintbrush               *.pcx
Sun Rasterfile              *.ras
Tagged Image Format         *.tif
WordPerfect Graphic 1.0     *.wpg
WordPerfect Graphic 2.0     *.wpg
X Bitmap                    *.xbm
X Window Dump               *.xwd
```

§17.2　统计绘图的实现

§17.2.1　SAS/GRAPH

SAS 的图形模块而是一个功能强大的图形工具箱，以下简介其功能，详细内容可参考SAS/GRAPH手册。SAS 也产生低分辨图形，CHART产生垂直或水平条图(直方图)、立体

图、圆图和星形图。PLOT 绘制每观察变量间的图形。

SAS 绘图语句大致分为三类，它们都有丰富的选项：

文本控制语句：TITLE(标题)、FOOTNOTE(脚注)、NOTE(附注)。

设计控制语句：AXIS(图轴)、LEGEND(图例)、PATTERN(阴影)、SYMBOL(符号)。

系统控制语句：GOPTIONS(绘图选项)。

其它如ANNOTATE=图形中使用的注释数据集，BY 指示分类。

FONTS 在SASHELP 库中，是一系列字型的名称、类型、描述、更新日期等信息。ANNOTATE数据集包含了一些命令或函数，指示SAS/GRAPH 增强图形的效果，如在图形中增加标号，在地图中放置符号和城市名，用线连接两个点，形成复合图形。

绘图过程主要有：

GANNO —输出附注数据集图形。

GCHART —产生垂直与水平条图(直方图)、立体图、圆图和星形图。计数可以是频数、累积频数和百分数，和与均值。

GCONTOUR —轮廓图，用二维的图示表达三维数量关系。

GDEVICE —检查和改变图形设备的设备参数。

GFONT —产生或显示字形。

GIMPORT —输入其它软件或机型的图形到SAS/GRAPH，输入格式是计算机图形交换文件(compter graphics metafile, CGM)。SAS 本身也可以产生CGM 格式的文件。

GKEYMAP —产生设备映象和键位图。

GMAP —产生显示某变量随地域变化的两维(choropleth) 或三维(block, prism, 和surface)彩色地图。

GOPTIONS —显示图形选项列表。

GPLOT, GPRINT —与PLOT和PRINT对应的过程。

GPROJECT —将圆形坐标的数据转为直角坐标系下的数据供GMAP使用。

GREDUCE —减少用于绘图的点数。

GREMOVE —合并地图数据集所定义面积单元，地区的边界被擦去。

GREPLAY —重显和管理图形目录项，同时产生重显彩色地图的模板和。

GSLIDE —显示由TITLE、FOOTNOTE 和NOTE语句产生的文本、直线图，同时可以显示由附注数据集产生的图形。

GTESTIT —产生三个图形，以提供设备设定的有关信息。

G3D —产生三维图形。

G3GRID —对分布不规则的数据点进行插值，产生G3D或GCONTOUR 的数据集以绘制三维表面或轮廓图。

SAS 的图形窗口用于显示图形，如同其它窗口一样，可以缩放。使用时首先设定GWINDOW选项，在命令行上使用时需要打入命令GRAPH1..4 < libref> catalog-name<.entry-name><.GRSEG>。

在图形显示设备上显示图形必须指示图形设备名，这可以通过GOPTIONS 的选项DEVICE=完成。在PC 机上显示主机的图形可经SAS/RTERM、SAS/CONNECT 或图形文件(graphics stream file, GSF)，在VMS 下要调入PC，则应删除回车换行控制(carriage control)。下面是一个用例：

```
$ANALYZE/RMS/FDL myfile.ext
$EDIT/FDL filename
```

```
  MODIFY RECORD CARRIAGE_CONTROL NONE
  MODIFY RECORD FORMAT           UNDEFINED
  EXIT
$CONVERT/FDL filename myfile.ext newfile.ext
```

在newfile.ext 调至PC 时，使用二进制拷贝(COPY/B 文件名) 送至设备。

GSF 格式的形成需要以下命令：

FILENAME GSASFILE 'MYFILE.EXT';

GOPTIONS DEVICE=设备名GAACESS=GSASFILE NOPROMPT HANDSHAKE=NONE;

若安装了SAS/CONNECT，主机的图形可在PC 机的终端上显示，首先应signon，然后远程提交(RSUBMIT) 语句GOPTIONS DEVICE=GRLINK; 其次提交(SUBMIT) 语句GOPTIONS DEVICE=设备名。

另外，产生CGM 格式文件的命令是：

FILENAME GSASFILE 'myfile.ext';

GOPTIONS DEVICE=设备名GACCESS=GSASFILE;

DEVICE=HP7470 是针对两种颜色的，HP7475 有六种颜色，HP7220 有八种颜色，HP7550 用于hardware polygon fills。

产生EPS 格式文件的命令是：

FILENAME GSASFILE 'myfile.ext';

GOPTIONS DEV=PSEPSF GACCESS=GSASFILE GSFLEN=80;

设备名也可以是DEVICE=PSLEPSF。

产生TIFF 格式文件的命令是：

GOPTIONS DEVICE=设备名GACCESS=GSASFILE GPROTOCOL=SASGPASC GSFLEN=80;

另外，使用SAS/QC 产生质量控制图。

CAPABILITY 产生描述统计量，研究受控过程与设计的异同，用于过程控制。

CUSUM 产生累积和式控制图。

ISHIKAWA 产生鱼刺图。

MACONTROL 产生移动平均控制图。

PARETO 产生Pareto 分布图。

SHEWHART Shewchart 控制图。

【例17.1】山东省生育资料趋势面分析

原理简介[1] 趋势面分析可作为回归分析在“数学地质”中的应用。在地质勘探中，测得某个地区大面积上许多测点的地质数据，需要通过这些地质数据，了解这块地区某种地质特征的变化趋势并找出它的异常部位，以帮助了解这一地域上的矿床分布趋势、找矿方向和直接寻找矿体。

趋势面分析的具体方法是：求一趋势面来逼近原始数据，即对于给定的原始数据求回归曲面，再利用回归曲面的变化趋势以及该曲面与原始数据的差异程度，来分析地质特征的正常趋势和异常部位。

常用多项式函数和三角函数表示趋势面，但以前者为常用。多项式的次数要据数据的特点而定，一次多项式为一平面，二次多项式则为抛物面、双曲面或椭球面。

以二次为例，某地域的n个点(x_i, y_i) 上的数据为z_i，二次趋势面是在多项式

$$\tilde{z} = c_0 + c_1 x + c_2 y + c_3 x^2 + c_4 xy + c_5 y^2$$

中求得使$\sum(\tilde{z}-z)^2$最小的$\tilde{z}=\hat{c}_0+\hat{c}_1x+\hat{c}_2y+\hat{c}_3x^2+\hat{c}_4xy+\hat{c}_5y^2$，$\hat{z}$称做趋势值。

进行趋势分析，既可以了解某地域某地质特征的大体变化规律，也可以突出局部异常。按照一定的趋势间隔画出趋势面等值线图来反映趋势变化，以及由剩余值的大小的等值线图从而由其大小和变化状态判断局部异常。

现对山东省生育率资料趋势分析，目的是看一下生育率呈什么趋势，用原始数据绘图，毫无规律，使用聚类分析可看出按生育率的高低，各地区可分到那一类中，采用响应曲面分析和绘制轮廓图，能进一步明确其趋势。

```
    名称:  ART
    标题:  山东省生育率资料分析
SAS产品:  GRAPH
    系统:  VAX
  关键字:  GRAPHICS GCONTOUR G3GRID AXIS LEGEND
    过程:  GCONTOUR G3GRID
    数据:  山东省生育率资料分析

    写作:  卫生部统计信息中心          修改日期: 17DEC92
    参考:
    其他:
```

```
    /* 图形环境设置*/
options ps=500 nocenter nonumber nodate;
/* 使用汉字字库*/
libname gfont0 'hanzifont:';
libname sas '[]';
/* 读取数据*/
data tan;
infile cards dlm=',';
input fert name $ y x; /* x y 系经纬度值*/
cards;
1.42,  章丘县, 36.79,     117.38
1.82,  长清县, 36.48,     116.80
1.88,  平阴县, 36.20,     116.41
2.82,  济阳县, 37.02,     117.20
2.47,  商河县, 37.32,     117.20
1.67,  胶南县, 35.86,     119.84
  ...   ...   ...   ...   ...   ...
/* 响应曲面分析*/
proc rsreg data=tan out=rstan;
        id name;
        model fert=x y /predict residual;
run;
proc print data=rstan;run;
```

```
/*快速聚类分析，结果可用于绘制轮廓图*/
proc fastclus data=rstan maxclusters=6 out=fstan;
        where _type_ eq "PREDICT";
        id name;
        var fert;
run;
/* 按地区打印*/
proc print data=fstan;
run;
/* 按类别打印*/
proc sort data=fstan;
        by cluster;
run;
proc print data=fstan;
run;
/* 准备注释数据*/
data county1;
    length text $40 function color style $ 8;
    retain function 'label' xsys ysys '2' hsys '3' when 'a';
    set tan; /* 原始值*/
    color='blue'; size=3;
    text='J'; position='5';
    style='Special';
    output;
    text=name; position='0'; color='green';
    style='k'; size=1.5; /* 楷体汉字*/
    output;
    text=compress(fert); position='2'; color='green';
    style='swissb'; size=1.5;
    output;
run;
data county2;
    length text $40 function color style $ 8;
    retain function 'label' xsys ysys '2' hsys '3' when 'a';
    set rstan; /* 响应值*/
    fert=int(fert*100+0.5)/100;
    where _TYPE_ eq "PREDICT";
    color='blue'; size=3;
    text='J'; position='5';
    style='Special';
    output;
    text=name; position='0'; color='green';
```

```
    style='k'; size=1.5; /* 楷体汉字*/
    output;
    text=compress(fert); position='2'; color='green';
    style='swissb'; size=1.5;
    output;
run;
/* 绘图文件*/
goptions device=hp7470 gsfmode=replace border;
filename or 'art1.plt';
filename rs 'art2.plt';
axis1 label=('Longitude');
axis2 label=('Latitude');
legend1 label=(j=c 'Fertility');
legend2 label=(j=c 'Response Surface of Fertility');
title f=h1 '山东省生育率资料分析';
footnote j=r 'CHSI/MOPH';
** 原始数据;
proc g3grid data=tan out=tangrid;
      grid y*x=fert / naxis1=40 naxis2=40;
run;
goptions gsfname=or;
proc gcontour data=tangrid;
      plot y*x=fert /levels=1.03 to 3.76 by 0.5 join
                     vref=34.65 to 37.98 by 0.5
                     href=115.12 to 122.42 by 0.5
                     haxis=axis1 vaxis=axis2 legend=legend1
                     annotate=county1;
run;
** 分析数据;
proc g3grid data=rstan out=tangrid;
      grid y*x=fert /naxis1=40 naxis2=40;
run;
goptions gsfname=rs;
proc gcontour data=tangrid;
      plot y*x=fert /levels=1.33 to 3.58 by 0.5 join
                     vref=34.65 to 37.98 by 0.5
                     href=115.12 to 122.42 by 0.5
                     haxis=axis1 vaxis=axis2 legend=legend2
                     annotate=county2;
run;
```

为了防止汉字读错，数据中使用逗号分隔，程序使用语句INFILE CARDS 来读取。汉字字库放于gfont0 中，绘图的标题采用了仿宋体，注释数据集中包含了要在轮廓图中追加信息的大小、颜色、字体等，这里把结果输出到图形文件。程序后面一部分是用来绘制地图的，此处从略。注意程序并没有使用GPROJECT 过程对经纬度值进行转换。

§17.2.2 SPSS/PC+

对图形部分的处理比较简化，它直接与Harvard Graphics 等软件包连接从而引入相应的功能。SPSS/PC+ 也拥有自己的GRAPH-IN-THE-BOX 图形捕获程序，还能与Ashton-Tate的Map-Master结合。图形环境设置方法与SAS相仿，即在SET 前加一图形标志(字母G)，但比较简单，如：

GSET PACKAGE HARVARD /HIGHRES 'C:\SPSS\GRAPHICS\gfile' /LOWRES=OFF.

利用GSHOW.命令可以得到当前的图形设置。

【例17.2】SPSS/PC+默认使用Harvard Graphics软件包绘图，HG 软件可以放在SPSS子目录，但在DOS的路径中亦可。HARVARD.INC 是一个典型的用例，程序运行结束后将生成Harvard Graphics 图形文件。程序的前半部分与BASETEST.INC相同，做图部分的语句是：

GRAPH BAR=MEAN(SALARY82) BY GRPAGE.

如第5章中的介绍，程序的前半部分首先读入列表数据，按变量定义缺失值，利用RECODE对年龄分组，再利用VALUE LABELS 语句给不同的分组值进行说明。最后，使用GRAPH语句按照年龄分组绘出SALARY82 均值直条图。

现利用上述活动文件继续运行如下指令：

```
SORT CASES BY grpage.
GRAPH /PACKAGE HARVARD /PIE grpage.
GRAPH /BAR COUNT BY SEX.
GRAPH /LINE COUNT BY GRPAGE.
GRAPH /BAR MEAN (SALARY79 SALARY80 SALARY81 SALARY82) BY GRPAGE.
GRAPH /LINE COUNT BY GRPAGE.
```

第一句仍然是按年龄组排序，第二句绘性别的条图，第三句绘折线图，第四句是关于年龄段绘几年的工资均值图(复合条图) ，第五句是各年龄段频数条图。SPSS/PC+每绘一个图时，系统自动进入HG软件，在屏幕上显示绘出的图形，继续打一键后显示参加绘图各序列的值已进入HG，此时在HG的控制下输入标题和修饰，用F2预演图形，^S存贮为.CHT格式的图形文件。

§17.2.3 Stata

Stata的图形功能在第 8 章第三节已做了较多的介绍，Stata 的一维做图很有特点，用于比较数据的分布情况，画一个轴，在有观察的地方打一短线，类似于星星的谱线。一维条形图图仍用graph 命令完成，常结合Box 给出，观察点很多时出现黑区，这时可用jitter(#) 选项指示重复的点数，如jitter(2) 表示用两个点表示一个观察。例：

graph temjuly, by(region) oneway

graph temjuly, by(region) oneway box

若在graph 命令中指定saving(文件名[,replace])，则可以把在显示器上的图形以文件形式存贮起来，命令中的REPLACE选项控制对同名文件进行覆盖。图形文件一般用.GPH作扩展，

它可以转成LOTUS1-2-3 格式的图形文件(.PIC)，再经WordPerfect 5.1 的Graphcnv.EXE 工具转成.WPG 文件。也可将图形以HPGL格式存贮。如stata教学盘上有一个图形文件(PIE.GPH)，使用以下命令把它输入HPGL 文件。

C:\STATA>gphpen pie /opie /dhp7470ls

生成后的文件又能经WordPerfect的graphvcnv工具转成.WPG格式。

又设在UNIX下，使用以下stata和UNIX系统命令：

```
. graph x, histogram saving(myfile)
. exit
unix% gphpen -dps -omyfile myfile.gph
unix% lpr myfile.ps
```

则将变量x的直方图以PostScript格式输出到打印机。

§17.2.4 Harvard Graphics

简称HG，常用于绘制统计图，如圆图、线图、直方图，可以是二维或三维并对这些图能追加各种修饰。该软件的特点使用简单，不论是专业非专业统计或计算机人员都适用。该软件可把多个图形做成幻灯文件，每间隔一定时间在屏幕上重显。HG 3.0 已有功能齐全的手动绘图工具，也支持鼠标。另外，SPSS/PC+ 的数据绘图功能可与Harvard Graphics 联合使用。

HG本身具有数据处理的功能，对于时间序列数据的处理尤其方便，直接指定时间区间，系统生成各时间间隔上的数据。HG利用各序列的数据可以构造新的序列，这一点与电子报表很相似。HG除了利用系统提供的数据录入功能外，HG利用的外部数据有Lotus、ASCII、dBASE、MS-Excel。HG也能输入Lotus、MS-Excel和.CGM格式的文件。

大部分的图形输出设备HG都支持，如IBM系统图形打印机、EPSON点阵打印机、HP等系统激光打印机以及绘图仪等。

为了方便做图，HG提供了大量的样本程序和符号库、图形框架(gallery) 等，做图时只要直接调入这些框架，略做修改即成新图。HG还提供了宏命令操作的功能，类似于第18章介绍的WP5.1，首先运行MACRO.COM，用Alt-0 菜单指定宏操作记录文件名，再进入HG和进行通常的操作，操作的内容自动记入文件中，仍用Alt-0 调出控制菜单并选择Unload退出。下一次HG运行时选择宏演示即可。HG提供了较完整的演示程序，对于HG更详细的介绍以及HG for WINDOWS可参有关文献。

以HG2.30为例，系统安装结束到HG目录后，在DOS提示下打入：

C:\ >CD HG

C:\HG> HG

则进入系统，调入图形文件SPSS_HVD.CHT，选择菜单的大写字母或数字进入相应的功能，各功能选择也可以经鼠标完成。其功能为产生一个新图、进入/ 编辑图形、屏幕做图/图注、取/存/删除图形、输入/输出图形、设备图形输出、幻灯菜单、图册菜单、设置、退出。

第一次运行或参数需要改变时，选择Setup功能。

系统在线帮助使用F1，选择F2将显示现有图形，选择F3可运行应用程序，选择F4进行图形中的文字拼写检查，选择F8进行图形特性选定。

上图对一些其它的功能键未做说明，如^G表示读图形文件、^S存贮文件等，利用HG的在线帮助可以获得这些信息。

HG的图形文件扩展名一般为.CHT，SPSS/PC+ 设图形名为SPSS_HVD.CHT。利用它的Export功能可以把图形输出到外部文件，HG2. 3 支持的外部图形文件格式为Professional Writer(PW)、Encapsulated Postscript和HPGL，对于第一种格式，可以在PW软件中利用*G命令把图形引入文本中实行图文并排，第二种、第三种格式输出图形描述文件，第三种即是熟悉的.PLT格式，能经WP 5.1 GRAPHCNV.EXE 工具转成.WPG格式加以使用。

§17.2.5 AutoCAD

AutoCAD 应用于DOS、Extended DOS、OS/2、Macintosh、Xenix、Unix、Aegis 及VMS系统。其第十一版运行于80386和DOS3.30 以上时可追加AutoCAD 开以系统和高级模型扩充(AME)。AutoCAD 绘图软件包是一个应用广泛的绘图软件。它功能强大，各版本间良好兼容，该软件把图形作为实体来处理，有关的概念和做法，已被其他新版本的绘图软件如Harvard Graphics 3.0所吸收，软件专门提供了自己的编程语言AutoLISP。软件提供给用户制图的实体(object) 或图形元素有：直线(lines)、圆(Circles)、孤(Arcs)、文本(text)等，用于二维、三维图形的制做。也可进行更复杂的操作，如图层(layer)进行多个图形的叠加，制做幻灯片文件等。也有象文字编辑软件那样的块操作(blocks)。图形可经键盘或鼠标器(mouse) 或图形输入版(tablet)输入。图形可以象文字那样擦去(erase)、移动(move) 、镜象(mirror)、阵列(array)、拷贝(copy)、缩放(zoom)、摇动(pan)和插入(insert)等编辑。此处括号内也是该软件所用的命令。软件具有令人满意的输出效果，可在点阵和激光打印机、绘图仪上输出。也可打印生成.PLT 文件。其.DXF和.PLT文件均可转成WordPerfect 5.1 图形文件。

现以AutoCAD 2.6 为例，设软件被安装于目录ACAD 中，可按如下操作进入：

C:\ >CD\ACAD

C:\ACAD>ACAD

这时，出现系统主菜单。

```
          A U T O C A D
Copyright (C) 1982,83,84,85,86,87 Autodesk, Inc.
Version 2.6 (4/3/87) IBM PC
Advanced Drafting Extensions 3
Serial Number:  97-835365
NOT FOR RESALE

Main Menu

  0.  Exit AutoCAD
  1.  Begin a NEW drawing
  2.  Edit an EXISTING drawing
  3.  Plot a drawing
  4.  Printer Plot a drawing

  5.  Configure AutoCAD
  6.  File Utilities
  7.  Compile shape/font description file
  8.  Convert old drawing file
```

```
Enter selection:
```

选择6是文件应用菜单，有以下的选择。

```
File Utility Menu

  0.  Exit File Utility Menu
  1.  List Drawing files
  2.  List user specified files
  3.  Delete files
  4.  Rename files
  5.  Copy file

Enter selection (0 to 5) <0>:
```

选择0则退出系统，选择1或2则进行图形的处理，此时若要保存生成的图形，则打入END命令，否则用QUIT放弃，这时回到系统主菜单。使用SAVE 命令则可以边绘图边存贮。AutoCAD的图形文件以.DWG 做为后缀。上述文件操作的菜单也可以由命令方式下的FILES 功能来实现。

AutoCAD 可以接受由键盘、鼠标或数字化仪等设备的输入。在屏幕底部的Command: 提示下，打入各功能相应的命令。可用Ins 把光标置于菜单之上，用光标上下移动，以空格或回车键选择。执行一次命令后要重复执行该命令，则仅仅打入空格或回车键即可。要废弃打入的命令，仍可用Ctrl-C。使用HELP 或? 命令，则可列出所有命令的语法和解释。

图形针对目标的操作，有自己的约定，AutoCAD 询问：

Select objects or Windows or Last:

这时可用三种方法定位：

. (point) 称做指出目标。系统扫描图形，定出标记有十字的实体。

M (multiple) 允许一次进行多个实体的选择，按回车键后始定位、操作。

W (window) 允许指定包含窗口内所有的实体。

AutoCAD 定位方法有两类，第一类是绝对坐标(World Coordinate System. WCS)，第二类是用户的相对坐标(User Coordinate System, UCS)，前缀以@符号。

绝对坐标是坐标的实际值，12.5, 3 表示x=12.5, y=3。

相对坐标前导以@符号，由(10,4.5)起相对坐标为@2,-3.5的点是(12,1)。

极坐标前仍导以@符号，由(10,4.5)起极坐标为@5¡30表示新点距(10,4.5) 为5 个单位，角度为30度。对于三维图形中的点，增加一个的Z 位置，如：

绝对：2,13,6

相对：@2,3,1

AutoCAD 能使用一个圆点指示滤掉那个坐标，如：.X 表示在以后的命令中不要求指定X坐标的值。其它的定点方法如：

CENter of circle	(CEN,圆心)	END point of line	(END,线的端点)
INSertion point	(INS,插入点)	INTersection point	(INT,交点)
MIDpoint of line	(MID,中点)	NEArest point	(NEA,最近点)
a NODe	(NOD,节点)	the PERpendicular	(PER,垂线)
The QUAdrant	(QUA,孤度)	TANgent to a circle	(TAN,切线)

对象的指示方法，点表示一个对象，M 表示多个对象，L 表示最近的对象，W 表示窗口内的对象，C 表示交叉窗口中的对象。A/R/U 表示对象增/减态/取消改动。进行删除操作后，屏幕可以重画(REDRAW) 以得到满意的显示效果。

屏幕做图用光标键或鼠标直接在屏幕上移动。使用键盘操作时，PgUp与PgDn两个键用于调整光标移动的速度。选择点后，空格或回车键确认，使用end键放弃。

制做幻灯文件使用命令MSLIDE，幻灯文件用扩展名.SLD，观察幻灯文件使用命令VSLIDE。现有幻灯文件BAR.SLD、LINE.SCR 以及PIE.SLD，可用以下的方法编于批处理命令文件THREE.SCR：

Command: EDIT

File to Edit: THREE.SCR

实际上是调用DOS EDLIN 进行编辑。因而也可以在DOS 编辑THREE.SCR文件。

```
 1: VSLIDE BAR.SLD
 2: VSLIDE *LINE
 3: DELAY 2000
 4: VSLIDE
 5: VSLIDE *LINE
 6: DELAY 2000
 7: VSLIDE
 8: VSLIDE *PIE
 9: VSLIDE
10: DELAY 3000
11: RSCRIPT
```

其中幻灯文件前面的星号表示把幻灯文件调入内存，演示时节省时间，DELAY 表示延时。

在AutoCAD 中调用方法如下：

Command: Script

File name: THREE

即可进行重复显示，用^C则中断显示。

AutoCAD 能进行一些最基本的DOS 操作，如：DIR，DEL，TYPE，用UTILITY 的CATALOG 进行宽的目录显示。也可用SHELL 命令返回DOS 系统。

调用AutoLISP 编写的程序类似LISP 语言的使用，即使用命令：

(load 文件名)

图形交换文件的生成使用EXFOUT命令。

Command: DXFOUT

File name: 文件名

输出的文件也可用DXFIN 读入AutoCAD。AutoCAD 系统提供了一个使用图形交换文件的样本BASIC 程序。这里只关心其对　WordPerfect 的转换。使用绘图命令PLOT 可将图形绘出或以文件保存起来，生成的文件为HPGL 格式，扩展名为.PLT。

```
LaserPlotter 1.3 for HP LaserJet    (C) Copyright 1986 Insight Development Corp.
Settings       Files       Go       Configuration       Exit
Get/Save/Change values in the SETTINGS SHEET

    Use arrow keys to highlight an option, then press Return to select it
                                  - OR -
                Just enter the first letter of the option

                    Press Esc to revert to previous menu

----------------------+------  SETTINGS SHEET ------+-----------------------
  PLOTTER EMULATION   |                            |   SETTINGS SHEET FILE
      HP 7470A        |                            |
----------------------+                            +-----------------------
    CONFIGURATION          Pen     Size     Color       Work Disk     D:
  Active printer            1        1      Black       Copy count    1
    LaserJet Plus           2        1      Black
  Output destination                                    Origin  X  0.00 in
    LPT1                                                        Y  0.00 in
  Paper source
    Paper tray                                          Resolution   75 dpi
  Size paper                                            Inverse      No
    Letter 8 1/2 x 11                                   Adjust-Size  100%
```

图 17.1 LP 主控菜单

§17.2.6 LaserPlotter

LaserPlotter 是激光打印机仿真绘图仪的软件，它可以将HPGL图形文件仿真HP 绘图仪在激光打印机上印出。在DOS系统提示下，执行实用程序LP.EXE。

C:\LP>LP

出现以下的菜单：

Settings子菜单能存贮设置内容。如提高分辨率则可设最高的300 DPI，有时图形的大小需要调整，等等，每次设置较麻烦，可一次设好存贮。设记录设置内容的文件名为MYCONFIG，则以后调用时使用命令：

C:\ >LP MYCONFIG

Files 子菜单下选择需打印的文件，一次可打印多个。

Configuration 子菜单下设置输出端口或磁盘文件等等。

在统计软件包或Harvard Graphics、AutoCAD 等软件包图形文件到HPGL输出时选择被仿真的绘图仪名称，如上图中显示仿真的绘图仪为HP 7470A plotter。

LP 的使用比较简单。

第十八章　文字处理与报告撰写

§18.1　概述

有了合格的资料和恰当的分析，下一步是用简练的语言、准确的描述、规范的体例把所得到的结论表达出来，再加上精美的输出则是珠联壁合。本章介绍一些常用的辅助工具，主要是字处理软件。

WordPerfect (以下简称WP)，是西文字处理软件中的姣姣者，许多文字处理软件受到其影响，如根据WP 5.1 写成的WP到LaTex转换程序，使后者的应用得到简化。WP 5.2 已经有相应的汉化版本。目前流行的WP6.1版，与5.1 多有类似之处。Windows所支持的多任务和动态数据交换等使它的功能得到加强。

许多软件包只在英文状态运行。程序中的注释、标号和提示信息多可以用汉字写成，如SAS 中的TITLE、FOOTNOTE 和LABEL 语句，有时则需专门的处理，如SAS的图形汉字。中文字处理软件始于西文软件汉化，如中西文WordStar，许多软件如EPI INFO的功能键定义与之兼容。比较成熟的中文字处理软件如CCED 和WPS 是若干年努力的结果。MiscroSoft等公司已对中文进行直接支持。

编程软件如PE II、SideKick 特点是占磁盘空间和内存小。PE II可以功能键重定义或执行宏定义。SK 则是一个优秀的常驻内存程序，在保存运行效果时很有用。IBM PC DOS 7.0 中的E 编辑与PEII功能极为相似，SK的增强版SK Plus 也有Windows 95相应的新版本，Windows运行软件往往可以用剪贴版实现SK 类似拷屏功能。

本章把统计报告撰写作为条目提一下。如调查报告的书写，内容有[1]：

1. 调查研究的主题以及进行该项调查的意义；
2. 总结和阐述已有的有关研究成果；
3. 介绍本次调查的范围、对象、时间、地点、方法、程序，如果是抽样方法的话，必须给出抽样方法、样本容量，样本的构成成分，必须对样本进行评价；
4. 对调查报告资料进行分析和研究讨论，逐步提出研究结论；
5. 总结；
6. 提出解决问题的办法或建议。

§18.2　几种字处理软件使用简介

§18.2.1　WordPerfect 5.1

WP 具备文字、数字与图形处理强大的功能，支持网络运行和鼠标，拥有菜单提示和在线帮助功能，可对英文章行拼写检查和同义词查找，根据不同的显示器和打印机显示和打印丰富的字符、数学公式，等等，是微机上一个成功的字处理软件。

㈠　基本操作

WP 对各种功能进行了定义，使用时应注意操作与WP命名相结合。

进入与退出：把软件建立在DOS 路径以内，打入WP<Enter> 即可。使用菜单或命令exit (F7) 退出。

目录与文件操作：使用Dir (F5) 命令，可进行文件的显示、打印、读入、删除等操作。也可使用F10存贮文件，Shift-F10读入文件。系统提供了文件PRINTER. TST 检测用户所用打印机具有的基本功能。

光标移动：利用小键盘的组合完成，列表如下：

移动	键序列	移动	键序列
单个字符	→或←	移至前一屏	Home ↑或重复- 号
单字	Ctl-→或Ctrl-←	移至后一屏	Home ↓或重复+ 号
行首	Home ←	页顶行	Ctrl-Home ↑
行尾	Home →	页底行	Ctrl-Home ↓
句尾	Ctrl-Home	前页顶行	PgUp
某字符	Ctrl-Home <字符>	前页底行	PgDn
上移一段	Ctrl- ↑	指定页	Ctrl-Home <页号>
下移一段	Ctrl- ↓	上一光标位置	Ctrl-Home Ctrl-Home
屏顶	Home ↑或- 号	上一块的首行	Ctrl-Home Alt-F4
屏底	Home ↓或+ 号	文首	Home Home ↑
屏幕右边界	Home →	文首控制码之前	Home Home Home ↑
屏幕左边界	Home ←	文件尾	Home Home ↓
行尾	Home Home ←		

编辑命令：如插入、删除、查找与替换、行列及段落字块(F4)操作，下面是一个删除功能键列表。

删除	按键
字符	BackSpace 或Del
单字	Ctrl-BackSpace
多个单字	Esc n Ctrl-BackSpace
光标左边	Ctrl-←Ctrl-BackSpace
光标右边	Ctrl-→Ctrl-BackSpace
至字首	Home BackSpace
至字尾	Home Del
至行尾	Ctrl-Del
至句尾	Alt-F4.Del
至段尾	Ctrl-PgDn
至块尾	Alt-F4 <字块> BackSpace 或Del

使用F2进行匹配字串的搜索，向前搜索则继续使用F2，向后搜索则使用Shift -F2 或在提示下使用 ↑ ；为了对控制码进行搜索，应当使用扩展搜索，即先打Home 键，再进行搜索，F2(搜索)、Shift-F2(反向搜索) 或Alt-F2(替换)。通配的方法是在字串中使用^X(Ctrl V, Ctrl X)。不同文件中特定模式的查找应使用其文件管理功能部分。步骤：打F5，到达指定目录后，再打9(Find)，然后指示被查找的文件名、文件概要，指定的某页或全文，以及字符模式：其通配模式如下：

类型	符合条件的字词
teach	teach 或teacher
p?n	panic, pen, pin 或pun
chi*	china或chinese
London: Paris	London to Paris
London, Paris	London or Paris

控制码的显示可借助于Reveal Codes (Alt-F3) 切换，使用这一功能，可以灵活地进行控制码的删除、移动、复制等操作。系统还提供了宏命令文件CODES .WPM，使用它可以打印出文章中的控制码和文本，也可以把两者写入一个文件。

字体控制与修饰：与字体修饰有关的操作在F8。对已录入的文字，首先进行有关的字块标记，再以^F8变换字体的形状和大小。也可在打字之前用Ctrl -F8键设定。基本字型也可以在setup 内完成。在Setup 下使用键盘定义文件SHORTCUT.WPK 则可用热键直接启用各种字体修饰。

标题与脚注：使用Ctrl-F7 进行产生和编辑。

纠错与同义词：使用spell (Ctrl-F2) 和Thesaurus (Alt-F1)。

图象处理：使用主菜单的Graphics 选项或Alt-F9 产生和编辑。在此菜单下，可进行调入和编辑.WPG 图形、调入或产生数学公式、用户方框的制做。CHARMAP .TST 存放特殊字符，可与汉字的区位码表相类比，操作时用^V指定其位置，则字符即被编入文件。嵌入文章中的数学符号和公式可经调用INLINE.WPM来实现。即于WP 下打入Alt-F10键并指示宏名INLINE.WPM。编辑结束时以exit (F7) 退出。

合并打印：需要定义初级文件和次级文件，其定义与F9 有关。

文件转换：可以转成DOS 文本文件以及4.2 和5.0 版文件。

图象转换：使用graphcnv.exe 工具完成。

WP 5.1 已对相应功能进行了定义，如帮助就是HELP。这样可以随时用帮助和索引。连续打两次help (F3，有时为F1)，可显示键盘模式。

打help 后，打其它的键则有相应的功能索引，如下面是先打帮助键再打字母D 后的结果，其中WP 键是软件对特定击键方法所给予的命名。

功能[D]	WordPerfect 键	击键方法
Date Format	Date/Outline	Shft-F5,3
Date Format (Default)	Setup	Shft-F1,4,2
Date of File Creation	Format	Shft-F8,3,5,1
Date/Time	Date/Outline	Shft-F5
Decimal/Align Character	Format	Shft-F8,4,3
Decimal Tab Setting	Format	Shft-F8,1,8,d
Default Codes	Setup	Shft-F1,4,5
Default Directory	List	F5
Default Settings	Setup	Shft-F1
Define Macros	Macro Define	Ctrl-F10
Define Paragraph/Outline Numbering	Date/Outline	Shft-F5,6
Define Printer	Print	Shft-F7,s
Define Text (Highlight)	Block	Alt-F4
Define ToC, Lists, ToA, Index	Mark Text	Alt-F5,5
Delete	Delete	Backspace or Delete
Delete Block (Block On)	Block Delete	Backspace or Delete
Delete File	List	F5,Enter,2
Delete Text (Block On or Off)	Move	Ctrl-F4,1-3,3
Delete to End of Page	Delete End of Page	Ctrl-PgDn
Delete to End of Line	Delete End of Line	Ctrl-End

More... 打d 继续

在setup (Shift-F1) 中选择键盘宏定义为ALTERNATE.WPK 时，帮助键为F1，F3 是Esc，而Esc 成为取消键，用时注意哪个键是help，其他情形下的使用与此相仿。WordPerfect 5.1 的在线帮助是在任何功能选单出现时，打帮助键即获得相应的说明。

其主要功能如下，取自其演示盘的显示结果。

合并(Merge)
纲要(Outline)
拼写检查(Speller)
同义词(Thesaurus)
行号(Line Numbering)
题头/脚注(Headers/Footers)
文件管理(File Management)
产生索引(Index Generation)
屏幕列定义(On-Screen Columns)
目录(Table of Contents)
脚注/尾标(Footnotes/Endnotes)
作者索引表(Table of Authorities)
编辑两个文件(Dual-Document Editing)
标号(Labels)
鼠标支持(Mouse Support)
长文件名(Long Filenames)
下拉菜单(Pull-Down Menus)
公式编辑(Equation Editing)
自动备份(Automatic Timed Backup)
教学程序(Computer-Based Tutorial)
风格(Styles)
Kerning
Soft Keyboards
助记菜单(Mnemonic Menus)
Master Documents
文件比较
打印预演(Preview)
密码保护(Password Protection)
交叉索引(Cross-Referencing)
改变字体
宏定义
定义合并
文本与图形混排
制表(Tables
电子报表联接
在线帮助
相对与绝对跳格
移行(Hyphenation)

㈡ 安装、使用与功能概要

【安装】WP 5.1 的安装很简单，执行INSTALL.EXE，则有以下的提示：

1 - Basic	实行一个标准安装
2 - Custom	为用户特定要求的安装。
3 - Network	网络安装。
4 - Printer	安装更新的打印驱动文件(.ALL)。
5 - Update	安装更新部分的WordPerfect 5.1产品。
6 - Copy Disks	把安装盘上的每个文件拷贝到指定位置 (对于安装所有的.ALL文件很有用)

如可方便地把所有被安装文件预先放在 C:\WP51，且安装于D:\WP51，结果如下。光标停止处有相应的提示。

1 - 安装文件来自 C:\WP51
2 - 安装文件到 D:\WP51
3 - 安装到磁盘
4 - 检查CONFIG.SYS 和AUTOEXEC.BAT
5 - 检查WP{WP}.ENV
6 - 选择和安装打印机和退出
→ 7 - 退出

退至DOS系统下

WP 提供了实用WPINFO.COM 工具，可以显示所用计算机的软硬件以及系统配置文件AUTOEXEC.BAT 和CONFIG.SYS 的内容。

【配置】用Shift-F1行配置(setup)，包括屏幕、文件位置等，设置的退出可用EXIT(即F7)。

不使用屏幕提示时，可用Alt-= 键调出系统菜单。

【文件管理】用F5可进行丰富的文件管理功能，可以列出文件目录，并进行有关的操作：如：

WordPerfect 的文件管理
读入文件
删除文件
移动或文件改名
打印文件
长/短目录
在读入之前观看文件的内容
改变到其它驱动器
拷贝文件
使用关键字或短语寻找串所在文件
文件名字搜寻定位

1 Retrieve; 2 Delete; 3 Move/Rename; 4 Print; 5 Short/Long Display;

6 Look; 7 Other Directory; 8 Copy; 9 Find; N Name Search 6

文件存贮可以保持与以前版本间的兼容，如WP 4.2 和5.0。也可以读入和存贮ASCII 文件。利用其Retrieve 功能可读入Lotus 1-2-3 电子表格数据，WP 的文件可以使用SUMMARY 功能进行文件的有关说明(详见HELP，S)。

【文件打印】可经主菜单File,Print或Shift-F7后可进行文件的打印。

1 - 打印全文
2 - 打印本页
3 - 打印磁盘文件
4 - 控制打印机
5 - 打印指定的页
6 - 打印预演
7 - 初始化打印机
选择

S - 选择打印机	HP LaserJet Series II
B - Binding 偏移	0”
N - 打印份数	1
U - 产生多页的设备	WordPerfect
G - 图形质量	Medium
T - 文本质量	High

此时按帮助键即可获得有关说明，打相应字母进行有关的选择。打印预演很有用，可根据所选打印和具体的显示器，在屏幕上显示将打印的内容。选择S时，进行印机的选择，结果可以是：

* HP LaserJet Series II
Canon LBP-8III
Panasonic KX-P1124

已定义了三种打印机，无自己的打印机，则可以从480 多种打印机表中选择一种做为第四种打印机。则择A，此时有下以的结果：

选择打印机:其它打印机

Apple LaserWriter

Apple LaserWriter IINT

.....

用光标移动来选择，选后出现对此打印机的说明及另一个菜单，供有关的定义使用，如定义打印输出口等，也可根据打印驱动文件列表获得。打印机的详细控制可经PTR.EXE和PTR.HLP 文件来完成。可根据需要指定不同的打印文件名(如.ALL 文件和.PRS 文件) 进行相应改动。

如果要接通网络打印机，则可的选择了打印机名后选择Edit进行修饰。

【排版功能】有报栏格式(Newspaper) 与平行格式(Parallel)两种类型的列。前者允许列与列间、页与页间的保护；后者则保证列编辑时，相关的列是平行的。其字体控制也很随意。特殊字符的录入使用Ctrl-V，这时有两种输入方法，数字录入法和助记符录入法。前者要求键入两个数字，与汉字的区位码有些类似，如8, 1 —8,42 是大小写希腊字母，码表存放在文件CHARACTR.DOC 和CHARMAP.TST。助记符是输入一些特殊的字符，如在Ctrl-V的Key=提示下键入/2或/4就可在屏幕上出现1/2和1/4，这一功能对使用特殊字符较多的录入是有用的，如法文的录入。不仅每页可以有不同位置的页码，而且每行也可以使用行码。

【块标记】由F12或Alt-F4结合Ctrl-F4等完成，进行块标记后，可进行一些特殊的操作，内容有：

块移动	按某项排序
块删除	搜索和替换
字体修饰	标记作为表、目录、索引的内容
风格	存贮
居中	追加至另一文件尾部
大小写转换	拼写检查
打印	

【语词功能】如拼写检查和同义词。编辑状态下可以进行英文单字拼写检查(Ctrl-F2)，系统将提示本页或全文拼写检查，对系统认为有错误的单词提示用户跳过(SKIP)，增加(ADD)，等等。

同义字库(Thesaurus) 有10,000个关键字(headwords) 会提示同义词、反义词以及交叉查找和指示关键字条。

在Shift-F8的其它功能项下涉及不同语言的WordPerfect，反映在WordPerfect 的系统文件上，如WPWPUS.HYC，含有US 表示是美国字库。该文件存有移行规则THS 是同义词库。WP.LRS 是Language Resource File，它是Secondary Merge 格式的文件，含有许多国家的语言表达法，其它的如WPWP.SPW 表示Spell by Word 文件，等等。

【公式编辑】使用Alt-F9产生和编辑。进入公式编辑后，根据需要指定公式的大小、编号，以及左齐、右齐或居中。公式的编号可以自动，因此当进行增删和块操作后，系统重新编码。公式明确键的含义很方便，如Setup表示SHIFT-F1，List 表示F5，显示表示F9。WP提供了inline.wpm宏，可以将数学符号插入文字中。另外，可以读入和存贮外部文件，这样可以使用公式模板。

【目录、索引和作者索引】用WP可以进行复杂功能。下面是一个生成的目录：

```
                    Table of Contents

Preface  . . . . . .  i
Table of Contents  . . . . . .  ii
Tables . . . . . . iii
Graphs . . . . . .  iv
Permits  . . . . . .  1
     Herd Count  . . . . . .  1
```

其做法比较简单，首先给出一个目录名称，一般使标题居中(录入前用Shift -F6或对已录入的内容进行块标记再打Shift-F6)。然后使用文本标记(MARK TEXT或Shift-F5, D, C 定义目录的层次。接下去是定义目录项，方法仍然是标记作为目录的文本，使用Alt-F5命令并选择C和回答目录项属于哪一个层次。逐个目录项定义好以后，使用Alt-F5, G, G 产生目录。

下面是一个索引用例：

```
                        Index

Animals
     Antelope  . . . . . . 1, 3
     Big Horn Sheep  . . . . . . 1, 3
     Buffalo . . . . . . 1, 3
Big Game . . . . . .  3
License  . . . . . . 3, 4
```

其作法类似于目录。

作者索引用例如：

```
                    Table of Authorities

                                                        Page
 Cases:
 Cornella v. Schweiker, 741 F.2d 170
   (8th Cir. 1984)  . . . . . .  3, 11, 15, 16
 Hensley v. Eckerhart, 461 U.S. 424 (1983)  . . . . . .  1, 10, 11
```

生成的步骤是：先定义页号(Shift-F8,2,6,1)，再用Alt-F5 进行定义(Alt -F5,D,A)，标记文本区中第一个作者名用并使用命令Alt-F5选择A，在提示下给出Section号、完整或简略形式的名称，此后便可以利用Alt-F5,D,E 进行作者目录标题等的修饰。然后在文本区每个名称处使用Alt-F5, A, 最后用Alt-F5,G,G 产生一个完整的作者索引表。

索引的产生还可以使用语词索引(concordance)文件，其步骤是：清屏，然后键入索引关键字，打<Enter>使每个关键字为单独一行，存贮该文件(如果每项目标题有所不同，则还需以上步骤：Alt-F4标记未在语词索引文件的词，键入Alt -F5,3并指示不同的标题)。在主文件准备如光标，键入Alt-F5, 5, 3 指定语词索引文件名，根据提示定义索引风格，然后打Alt-F5, 6, 5, Y。

现要在输出报告的每页下方打出“共几页，现在为第几页的信息”，可在文首利用页格式定义Page(Shift-F8,P,F)，进一步选择A, P, 此时进入每页标题的编辑态。打入Page ^B of ，其

中^B 表示插入当前的页号，文本区总页码的插入是使用Alt-F5, R, R, P 则提示成为Page ^B of ? ，然后打Exit(F7)退出，系统提示一个被参照页码的代号，键入之，如LP。现将光标移至文末(Home Home, ↓)，启用命令Alt-F5, R, T，指定参照的代码，默认为LP，就完成了定义，最后使用Alt-F5, G, G, Y 操作完成，此时用Reveal Codes (Alt- F3 或F11) 来观察，开始定义处总页码已被自动加上。

【制表】下面是一个表的范例：

WordPerfect 公司的产品		
Product	产品描述	软件平台
WordPerfect	字处理	PCs, PC Networks, IBM 370, VAX, Mac, DG, Unix, Apple II, Amiga, Atari
PlanPerfect	电子报表、与WordPerfect 兼容，可行接口的选择	PCs, PC Networks, VAX, DG

WP 5.1 可以包括文字或者数字。表格可以多达32列和32,765行。有时表太大，可以缩小基本字型，改变纸张大小，纸张大小与所选择的打印机有关。表格可以任意合并，改变线框。在表的栏目中使用公式，则可显示相应的计算结果。

包括四种数学操作符(+ - * /)，数字位数可选，小数位数可达15位。也可与电子报表(spreadsheet)产生连接。其步骤是：设定光标，键入Ctrl-F5, 5 据需要指示1(输入)或2(产生连接)和文件名并打<Enter>，若只输入电子报表中的某个范围，则指示Range。输入的格式可以是表格或文件。

【合并打印】由定义主文件、辅助文件和文件生成三个步骤完成。主要由F9相关的功能键完成。合并时(Ctrl-F9)需要指定主文件名字，其中包含了信件的正式文本和将被替换的字段如日期、地址位置；辅助文件，含有与主文件对应的记录。每个记录的字段以{END FIELDS}(F9)标记，记录以END RECORD(SHIFT-F9, E)结束。除了地址以外，还可以加上其他信息，如指定合并时的录入等，以下是一个范例，右上的方框中有关的选择可以用Shift-F9,M调出。

【使用框架】WP 5.1 的框架允许存贮和使用格式化指令来对文章段落进行编号。在进行段落的移动后，号码自动更换，使用时注意框架的开启。下面是一个用例：

```
I.  Installation
      A. Package Components
      B. Installation Guideline
II. Getting Started
      A. User Interface
           1. Function Key
```

【宏的使用】宏实际上是一个例行的程序，在WP 5.1 中使用Ctrl-F10 进行宏定义，首先输入宏的描述，即指明宏的用途，然后系统自动把键入的内容记录下来。在编辑状态，可使用Ctrl-PgUp 调出宏命令菜单供选择。如要使用光标键进行宏定义，则应先打入Ctrl-V。WP 5.1 也可对现存的宏进行修改。WP 5.1 提供了几个宏供使用，如INLINE.WPM 用于在每行中

嵌入公式和字符，其内容如下图所示，大括号表示WP 5.1 的某项功能定义，数字相应于特定的选择。

```
{;}Macro for creating In-line Equations.
   These equations are placed in the User box list.
   In your Initial Codes you may want to set up the User box options
   so that inside and outside border spacing=0.
   If you want in-line equations in a different list change (1),
    maybe (2) and your initial codes accordingly.
   Set this up as a handy Alt-key macro~
{IF}{STATE}&4~
{MENU OFF}{DISPLAY OFF}{Graphics}
4              {;}user box list (1)~
1              {;}create~
24             {;}contents=equation (2)~
43             {;}type=char type~
54             {;}vert alignment=baseline~
74             {;}size=auto both~
{MENU ON}{DISPLAY ON}9  {;}go into equation editor
{END IF}
```

WP 5.1 的键盘定义文件有ALTERNATE.WPK、ENHANCED.WPK、EQUATION.WPK、MACROS.WPK 和SHORTCUT.WPK，可在Setup 下编辑(Edit)、调入(Action)。如MACROS.WPK中的Ctrl-C 是一个算术计算器。可以获得编辑宏定义的模板。

㈢工具程序

【SPELL.COM】拼写工具。在系统下打入SPELL ¡RETURN¿，出现以下的提示：

Spell – WordPerfect Speller Utility

0 - 退出；
1 - 改变/产生字典
2 - 追加新字
3 - 删字
4 - 字典优化
5 - 显示常用字
6 - 检查字所在的库
7 - 查字
8 - 按音序查字
9 - 把4.2 版字典转为5.1 版
A - 组合其他5.0 或5.1 版字典
B - 压缩/扩充补充字典
C - 抽出追加的新字

【CURSOR.COM】可以改变光标大小。在WP 5.1 运行前使用命令CURSOR/GB 可使用光标显示更为直观，这对使用LCD 显示器很有用。

【CONVERT.EXE】WP51 提供了丰富的文件转换功能，CONVERT.EXE 是一个外部命令，执行后有以下的提示：

输入文件名? printer.tst
输出文件名? tst
0 EXIT
1 WordPerfect to another format
2 Revisable-Form-Text (IBM DCA Format) to WordPerfect
3 Final-Form-Text (IBM DCA Format) to WordPerfect
4 Navy DIF Standard to WordPerfect
5 WordStar 3.3 to WordPerfect
6 MultiMate Advantage II to WordPerfect
7 Seven-Bit Transfer Format to WordPerfect
8 WordPerfect 4.2 to WordPerfect 5.1
9 Mail Merge to WordPerfect Secondary Merge
A Spreadsheet DIF to WordPerfect Secondary Merge
B Word 4.0 to WordPerfect
C DisplayWrite to WordPerfect

其中的RFT 格式(第3项)，可以与Microsoft Word 软件转换文件。

在选择了号码1后，还会有以下的菜单：

输入文件名? printer.tst
输出文件名? tst

0 EXIT
1 Revisable-Form-Text (IBM DCA Format)
2 Final-Form-Text (IBM DCA Format)
3 Navy DIF Standard
4 WordStar 3.3
5 MultiMate Advantage II
6 Seven-Bit Transfer Format
7 ASCII Text File
8 WordPerfect Secondary Merge to Spreadsheet DIF

源文件和目标函数及其类型亦可由CONVERT加命令行参数的形式指定。

【GRAPHCNV.EXE】是一个图形转换工具，在DOS提示下，打入"graphcnv /h"，即显示程序使用方法。运行后提示待转换的文件名，这些文件的格式可以是DXF 、CGG、EPS、IMG、HPGL、PIC、PCX、TIFF格式和Dr. Halo及Macintosh/paint，生成WP 的图形文件(.WPG)，由WP 5.0 或以后版本调用。Graphcnv 命令格式：

graphcnv 被转换文件名　生成的.WPG 文件名

文件名中可以包括路径名，可以使用通配符"*" 和"?"，转换指定目录中符合条件的所有文件。生成文件均以.WPG 作为扩展名。命令开关：

/l 登录转换标准打印设备的状态信息。可使用/l=文件记录转换过程。

/o 覆盖文件替换提示信息(”replace files ?”)。

/W HPGL 文件中的行设为最小。避免重复画笔的厚度。

/b=# 设定WPG 文件的背景颜色。号码#可以为1-8。默认为是增强白色。

/c=w,/c=b 把输入文件中所有的颜色转成黑白(不用于bitmap 图象)

/c=2 把输入文件中所有颜色值转成黑白单色。

/c=16 把颜色值转成WP标准16色。

/c=256 把颜色值转成WP标准256色。

/g=16 把颜色值转成WP标准16色(grey palette)。

/g=256 把颜色值转成WP标准256色灰色。

颜色转换不成功时则转换为WP 标准的256 色。颜色号码的值与DrawPerfect 中的菜单选项对应。1 =黑，2 =蓝，3 =绿，4 =深兰，5 =红，6 =深红，7 =棕，8 =灰，9 =深灰，...。DrawPerfect 支持.CGM、.EPS、.HPGL 、. PCX、.TIFF 和.WPG文件格式。

/f=# 把填充的颜色转成指定的颜色。

/n=# 把所有颜色转成指定的颜色。

如：graphcnv /c=16 /f=1 把填充颜色变为黑色。

【PTR.EXE】是一个打印驱动程序维护工具，PTR. HLP 是其相应的帮助程序，用户可定义、修改自己的打印驱动程序，/CONVERT 命令行参数把WP 5.0 的打印驱动程序转到5.1。

【MACROCNV.EXE】即宏定义转换工具，用于把WP 5.0 的宏定义文件转到5.1。

【GRAB.COM】是屏幕图形捕获工具，通过执行grab.com 完成，打GRAB/H 给出其相应的使用信息。用<ALT><SHIFT><F9> 键激活屏幕图形拷贝。此时在屏幕上出现一个小方框，使用<SHIFT> 加上箭头键改变方框的大小，用<INS> 键切换移动方框和方框大小的步距。框住图形时，打<Enter> 键，文件grab.wpg 自动生成。

在WP 中调用则可以用”GRAPHICS图形”(<ALT><F9>) 菜单调入实现图文并排。一次拷贝多个图形时，图形文件依次命名为grab1.wpg、grab2.wpg等等。可以使用<F1> 取消捕获。GRAB 工具有以下开关：

/D=路径前缀指示路径前缀，长度为1 至80 字符。

/F=根文件名指示图形文件前缀，长度为1 至7 个字符。

/R 终止和取消工具驻留。

/8 允许使用8514 仿真。

/I 忽略DOS-忙标志。

/M 强迫以单显bitmaps 格式。

/S 当光标键按下时，可用空格键进行方框移动和改变大小的切换。

/K 保持上次的剪裁方框。

/P=2 在Hercules 图形适配器下捕获第二页。

/C 或/C0-5 键盘重捕获次数，/C 为一次(默认), /Cn n= 0-5允许n 次重试。

GRAB 可用于向WP 外壳剪贴板拷贝图形。调入GRAB，运行SHELL程序，用<Alt><Shift><− > 捕获。

WP 提供了许多图形的范例(.WPG)，在系统安装时可以指示装入。然后在图形生成时(Alt-F9, Figure, Create) 指定文件名(Filename,List/F5, Retrieve) 调入。

§18.2.2 PE II 软件

PE II 是一个优秀的全屏幕编辑软件，最适合进行程序编辑和简单的中西文文章编辑，尤其是它的列块操作给人的印象深刻。它不仅占内存少，甚至在编辑态返回DOS 多次重新调用，而每次返回时，原来的状态都能相应恢复。它最多可用四个窗口编辑相同或不同的文件，同时打开和编辑20 个文件，并且可在这些文件之间进行复制、移动等编辑操作。这一功能也可用于一个文件前后对比编辑。PE II 也提供了用户键盘自定义的能力，流行的“联想”汉卡加强了它的制表和中文处理功能，在UCDOS 等25 行汉字显示下使用命令行参数/B亦可显示和编辑汉字。

PE II 有PE2.PRO (定义文件)、PE2.HLP (帮助文件)、PE2.EXE (执行文件)三个基本文件，PE2.PRO 存放它的功能键定义，执行时由PE2. EXE 时自动完成。PEII通过编辑文件PE2.HLP 完成的，建议把PE2.HLP 属性改成只读。随PE2 软件盘还有一个安装程序，一套演示程序和文件，但仅用PE2.EXE 也可以进行编辑。

命令格式：C:\PE2> PE2 [参数] [文件名] <Enter>

即在DOS 系统提示符下，打入PE2 ¡Enter¿ 进入，命令行参数是：

/B，强迫以系统调用方式进入PE II，这会影响显示速度，但可直接输入汉字。

/Pprofilename，指示PE II 初始化时处理的命令文件，其默认值为/PPE2 .PRO，该文件不存在时，进入编辑后显示“profile not found”，系统使用内部的命令定义。

/Q，进入PE II，但省略(quiet) 前面的说明。

/Rnnn，为命令翻译器的拷贝副本保留nnn 千字节RAN，默认值为/R2，这个参数很少用。

/Sspillfilename，指示溢出文件名，当系统提供给PE II 需要的RAM 不足时，PE II 把部分内容写到一个溢出文件中，隐含为PE2TMP，溢出文件的扩展名由PE II 自己来定。

/W，进入PE II，但略停顿等待打一键后方进入。

PE II 的键面定义列表如下，其中的简写含义为，s = Shift 键；a = Alt 键；c = Ctrl 键。

功能键	功能
F1	帮助屏幕
F2	保存当前文件并继续编辑
F3	即FILE，保存文件并退出
F4	即QUIT，退出当前文件
F5	删除一行
F6	从当前光标删至本行末
F7	打印当前文件
F8	切换所编文件
F9	插入一行
F10	插入一行并与上一行对齐

如果按F4 (QUIT)，则出现：“退出吗(键入y/n)”，若按N，则QUIT 命令作废；若按Y，则退出不保存。

光标移动

↑	上移一行a–/=	左/右移一个字
↓	下移一行C-Home	文件的顶部
←	左移一列C-End	文件的底部
→	右移一列C-PgUp	屏幕顶部
Home	行首C-PgDn	屏幕底部
End	行尾C- ←	左移40 列
PgUp	前一页C- →	右移40 列
PgDn	后一页C- L	本行屏幕居中

文本标记

a-B – 块标记,用于列块, 垂直或水平行

a-C – 字符标记,用于句子, 词和字符。

a-L – 行标记,用于一行或一段。

a-U – 取消所有标记

a-W —标记下一个字，也可用于连续向下一个字的移动

文本标记的用法

a-Z – 复制文本标记区,保留原有文本。

a-M – 移动标记区,删除原有文本。

a-O – 复盖文本标记区,保留原有文本。

a-D – 删除文本标记区

a-F – 块填充

清除文本

F5 – 清除一行

F6 – 清除光标右边的文本

s-F6 – 清除光标左边的文本, 并把剩余部分移到行首

恢复文本

s-F4 – 恢复当前行

c-U – 显示最近若干次改动前的未命名文件

文本格式化

a-S – 在光标处拆行

a-J – 在光标处合并行并保留空格[与c-j不同]

a-P – 在光标处排段

c-C – 把文本在当前边界设置下居中

分割屏幕(多窗口)

c-S – 将屏幕分为四个窗口或上下或左右两个窗口

c-V – 将光标移到同一文件的下一个窗口

c-W – 移动光标到下个窗口

c-Z – 把当前窗口放大充满整个屏幕

文本的定位和修改

l/text/ – 定位起初文本，其中的分界符'/' 可以任意设定。

c/old/new/ – 定位起初文本并询问确认修改，具有-*m 几种选择

s-F5 – 确认修改

c-Enter – 从文本区执行命令

其它命令

c-p	打印标记区
c-k	功能键定义显示
c-r	复制上一行的内容
c-d	返回到DOS [以EXIT 返回PE II]
c-f	将命令行的内容复制到文本区
c-t	将文本区内容复制到命令行，这对于一些特殊的替换操作很有用
s-f7	左移标记区
s-f8	右移标记区
s-f9	列目录[在命令行打DIR，效果与之相同]
s-f10	编辑.DIR 列出的文件
a-f10	显示光标处字符的ASCII 码值

不同键的定义可以在命令行打? KEY keyname 进行询问，如? key s-F1。用Ctrl-T/Ctrl-F完成光标所在行到命令行的复制或命令行到文本行的复制，用于录入一些难于录入和清除的字符时特别有用。

【附】PE II 集合与宏定义

PE II 的上述功能，可经过PE II 提供的62 个基本命令的组合来完成，这些命令显示或执行时，以完整或缩写的形式出现。系统默认的定义存于内部文件.keydefs，可以使用EDIT命令来观察。需要永久改动功能键的定义时，可把改动结果存于PE II.PRO 中在系统启动时定义或存于其他文件，使用macro 命令完成定义。宏定义文件中可以使用* 以及/* */ 做注释。编辑态动态地改动某个命令，可以使用define 命令，如define enter=[il] 把¡Enter¿ 键定义为行的插入。

◆ 光标定位

[left XX] [right XX] [up XX] [down XX] 将光标向各个方向移动XX 位置，默认为1。

[line XX] [column XX] 将光标定位在指定的行列上。

[wb] [we] 光标定位到当前词首、词尾。

[tw] [bw] 光标定位到上一个字、下一个字。

[tb] [bt] 光标定位到上一个制表位置、下一个制表位置。

[te] [be] [le] [re] 光标定位到窗口的上下左右四个边缘位置。

[lg] [rg] [pg] [in] 光标定位到版式设置的左边界、右边界或缩进位置。

[bl] [el] 光标定位到行首或行尾。

[to] [bo] 光标定位到文首或文末。

[bm] [em] 光标定位标记区首或尾。

[pu] [pf] 光标定位上一页或下一页。

[fh] [fb] 光标定位到第一个非空字符或下一个空行上。

◆ 方式转换

[lm] [rm] 插入状态或替换状态。

[it] 切换插入/替换状态。

[cc] [ca] 光标到命令行或数据区。

[cg] 光标命令行/文本区世切换开关。

◆ 插入删除

[il] [dl] 插入和删除一行。
[dc] [ro] 删除光标处或光标前字符。
[ed] [ee] 删除至行首/行尾。
◆ 标记区操作
[mb] [ml] [mc] 标记方式。
[um] 取消标记。
[mm] [cm] [ob] 移动、复制、覆盖标记区内容。
[dm] [fm] [pm] 删除、填充和打印标记区。
[uc] [lc] 标记区大小写切换。
[ps] [po] [ck] 标记区入栈、出栈和清栈。
◆ 窗口操作
[zw] 窗口放大。
[ss] 屏幕分割。
[nw] [nv] 下一个窗口或下一视图成为当前窗口。
◆ 文件操作
edit [文件名[notabs]] 编辑新文件，edit 可略为e。
save/file [文件] NOTABS 存文件并去掉跳格键
name 当前文件改名。
quit 退出妆前文件编辑。
print 或p 打印整个文件。
◆ 屏幕调整
[cl] 行居中。
[scrollup] [scrolldown] [scrolleft] [scrollright] 屏幕向上下左右滚动。
◆ 行调整
[sl] [sr] 标记区内容左右移动。
[sp] [jo] 分裂当前行或连接两行。
[ci] 当前行内容调整到边界中央。
[rf] 排版。
◆ DOS 操作
[dos] 退至DOS 外壳。
dir 显示文件目录。
erase 文件名删除文件。
rename 旧文件名新文件名文件改名。
chdir 改变当前磁盘目录。
◆ 查找和替换
l\xx\[-] [m] 查找文本/标记区字串。
c\xx\yy\[-] [m] [*] [co] 替换文本/标记区字串、确认替换。
◆ 演示功能
[df] [ds] 快速或慢速演示。
[de] 演示结束。
◆ 其它
[ex] 执行命令行上命令。

[ud] 取消当前行上任何修改。
[rd] 屏幕重画。
[ce] 消除显示区内容。
[bp] 振铃。
[es] 输入特殊字符。
[nu] 用于define 语句中的空定义。
[xxx] 三位十进制数，指示相应的ASCII 码值。

◆ 设置与查询命令

set backup xx	[?k] 设置和查询删除内容保留次数。
set margins xx xx xx	[?m] 设置和查询边界值，包括左边界、右边界、版首行缩进。
set searchcase exact/any	[?s] 设置和查询匹配条件。
set abbrev on/off	[?a] 设置和查询命令缩写状态控制。
set hscroll on/off	[?h] 设置和查询屏幕滚动方式。
set display xx	设置显示器类型。
set BLANKCOMPRESS on/off	[?b] 空格压缩。
set TABEXPAND on/off	[?e] 查询标尺展开。
set TABS xx xx	[?t] 设置查询标尺。
	[?y] 查询可用内存空间。
	? diskspace 查询磁盘空间。
	[?c]/? CHAR 查询字符的ASCII 码值。
	[? dir] 查询当前磁盘目录。

如：

```
s margins 1 79
d enter=[il][in]
d f1=[e ce2.hlp]
d f2=[cc][bl][ee]"save"
d f3=[cc][bl][ee]"file"
d f4=[q]
d f5=[bl][ee]
d f7=[cc][bl][ee]"print"
d f8=[e]
d f10=[il][ps][ml][bw][bl][wb][bm][po]
d s-f10=[ps][ca][ml][cm][column 9]"."[column 14][ee][c/ //-*m][ bl] [sr][sr]"e "[ct][down][dm][ex][po]
d a-n=[il][012]
d a-f2=[? diskspace]
d a-f3=[? dir]
```

§18.2.3 中西文WordStar 软件

使用很普遍，而且许多软件如SK、Turbo Pascal编辑器、QUICKBASIC 编辑等的功能键与之相兼容，故有必要掌握。WordStar 3.X 有三个文件较为关键，即WS .COM、WSMSGS.OVL

以及WSOVLY1.OVR。

在系统提示下，打入WS 就进入主菜单。如：C>WS <Enter>

《起始命令》

D 进入编辑	E 更换文件名
P 打印文件/中断	O 拷贝文件
R 运行命令	Y 删除文件
N 编辑非文书文件	X 退出

键入相应的英文字母即可实现各个功能。

西文WordStar 尚具以下功能。

F 显示磁盘目录

H 设帮助水平

L 改变当前磁盘

M 合并打印

S 运行SPELSTAR

选择P 进行打印时，打印可以输出至一个文件、指定开始与结束页码、页间暂停(以P继续)、中止打印(用P出现提示后，再打Y)，打Esc键则中止询问，用默认的设置。

选择D 或N 进入后，可以使用以下功能键定义：

光标移动命令：

Ctrl-E 或↑	光标上移一行
Ctrl-X 或↓	光标下移一行
Ctrl-D 或→	光标右移一个字符
Ctrl-S 或←	光标左移一个字符
Ctrl-F	光标左移一字
Ctrl-A	光标右移一字
Home 或Ctrl-Q-E	光标至屏首
End 或Ctrl-Q-X	光标至屏尾
PgUp	向前翻页
PgDn	向后翻页
Ctrl-Q-PgUp 或Ctrl-Q-R 或F9	文首
Ctrl-Q-PgDn 或Ctrl-Q-C 或F10	文末
Ctrl-W	屏幕向上滚动
Ctrl-Z	屏幕向下滚动
Ctrl-G	删除光标所在字符
Ctrl-H 或Del	删光标前面字符
Ctrl-T	删除光标处的一个字
Ctrl-Y	删光标所在行
Ctrl-Q-Y	删至行末
Ctrl-Q-Del	删至行首

字块操作命令：

Ctrl-K-B 或F7	定义块首
Ctrl-K-K 或F8	定义块尾
Ctrl-Q-B	光标至块首
Ctrl-Q-K	光标至块尾
Ctrl-K-C	拷贝字块
Ctrl-K-V	移动字块
Ctrl-K-H	隐藏字块
Ctrl-K-N	行块与列块切换
Ctrl-K-R	把文件做为字块读入
Ctrl-K-W	把字块做为文件记盘
Ctrl-K-Y	块删除
Ctrl-K-P	打印字块
Ctrl-Q-B	光标至块首
Ctrl-Q-K	光标至块尾
Ctrl-K-0 .. Ctrl-K-3	定义标记0..9
Ctrl-Q-0 .. Ctrl-Q-3	到达标记0..9

查找与替换：

Ctrl-Q-F	查找
Ctrl-Q-A	查找并替换
Ctrl-L	寻找下一个

方式转换与插入：

Ctrl-V 或Ins	插入状态切换
Ctrl-O-I	自动缩边切换
Ctrl-O-T	标尺显示切换
Ctrl-N	插入新行
Ctrl-M	插入回车符
Ctrl-P	打入控制符
Ctrl-I 或Tab	插入跳格

屏幕编辑命令：

Ctrl-O-L 或F3	设左边界
Ctrl-O-R 或F4	设右边界
Ctrl-O-S	设行距
Ctrl-O-C	行居中
Ctrl-B	段落排版

文件操作与退出命令：

Ctrl-J	帮助命令
Ctrl-U	终止命令
Ctrl-J-H	设帮助水平
Ctrl-K-S	存盘并继续编辑
Ctrl-K-D	存盘并返回主菜单
Ctrl-K-Q	放弃存盘并返回主菜单
Ctrl-K-X	存盘并且退至DOS
Ctrl-K-F	显示文件目录
Ctrl-K-O	拷贝文件
Ctrl-K-E	文件改名

WordStar的列块移动功能与PE II有所不同，它可以跨越标记区，似乎较好一些，下面介绍的WPS与WordStar类似。

除上面列出的以外，尚有一些点命令，如.PL 页长、. PO 打印开始列数等。Ctrl-Q-Q 为快速执行命令。合并打印也有一套专用指令。WordStar4.0 版本以后，其功能大大增强，而功能键定义与上述略有不同。

§18.2.4 Sidekick

Sidekick 是一个常驻内存(Terminate and Stay Resident, TSR) 的软件，文件字节不多，但功能很强，如记事编辑、日历、计算器、ASCII 码表、日程表等，各功能用不同的窗口(见下图)。此图的获得亦是通过它的NotePad 中的输入(Import, F4)，把屏幕内容扫描到文件中的，其编辑功能键与Turbo Pascal 、WordStar 等是兼容的，后者在下面将予介绍。其最大的特点是可以在许多软件运行的条件下，以ctrl-alt或同时按下左右Shift 键自由进退。这样在使用GLIM、Stata一类软件时可以不必退出系统进行编辑，利用SK 在程序编辑和运行之间进行热键切换。与

其它软件相比，它能够分辨跳格符(^I，TAB)，PE II 以及MS-DOS 中的EDLIN 默认采用压缩格式存贮，在打开文件时自动放开。但编辑生成的ASCII 数据集在Fortran 或LISREL 软件使用时往往容易出错。因此，对于部分用户来说，SK 软件是很值推荐的。

SideKick Plus 的功能有所加强，包括ASCII码表, 记事本, 万年历, 日志簿, 电话拨号(Phone dialer), 数据库函数, 文件管理(file manager), outline processor, 在线通讯及kill-the-space-things 游戏。程序最大的部分是弹出式服务(pop-up access)，在1MB 以上的内存下使用比较好。

SideKick 1.56A 有两个基本文件，即：

SK.HLP　帮助文件

SK.COM　执行文件

只有SK.COM 亦可，只是不能调用在线帮助。为了使用其在线帮助，SK. HLP 文件应于当前目录。用SKINST决定SK的功能键定义与编辑文件大小，这点与Turbo Pascal 类似。

进入：在DOS系统提示符下,打入：SK ¡Enter¿　即可进入。

在联想汉卡下可以编辑汉字文件。

编辑文件时，键入F4 ，即进入屏幕“抄写”，使用Ctrl-K-B (或F7) 定义块首，然后用→ ← ↑ ↓或Ctrl-→Ctrl-←移动光标，用Ctrl-K-K (或F8) 定义块尾，以Ctrl-K-C 就把屏幕内容拷贝过来。这里指出的是，本书中的许多文字材料都是利用它的屏幕抄写功能，在软件运行时将其运行信息，不加改动地保留下来的。注意它不抄写图形，图形抄写可用如：PZP 软件、SPSS/PC+ graph-in-the- b ox、WordPerfect 5.1 GRAB、WINDOWS Clipper 等。

Sidekick的帮助系统是在线的，你可以在特定窗口打F1或Alt-H得到使用信息。

Notepad 是一个全屏幕、与WordStar/TURBO Pascal 兼容的文本编辑器并有自己的特点。括剪贴和自动给出日期/时间。Calculator 是一个计算器，进行十进制、十六进制和二进制计算。Calendar 可安排每日日程。Dialer 从屏幕上或自己的电话号码簿中取得电话号码，Sidekick 可以使用modem进行呼叫。ASCII 表显示256-字符的ASCII 字母表，同时显示十进制、十六进制码值和字符、助记符。Setup 用于改变Sidekick所使用的标准文件名和目录，永久记录这些值及窗口位置、大小数据。从主菜单中用以下方法选择窗口：

1 按住Alt-键并打窗口名中高亮度显示的大写字母。

Alt-H 调Help　　Alt-D 调Dialer

Alt-N 调Notepad　　Alt-A 调ASCIItable

Alt-C 调Calculator　　Alt-S 调Setup

Alt-L 调caLendar

若按Alt键超过1.5秒，系统认为需要帮助，此时显示主菜单。

2 直接打窗口名中高亮度显示的大写字母。

3 使用 ↑ 或 ↓ 键到指定的窗口，再打Enter。

4 打与每个窗口关联的功能键。

推荐第一种方法，因为这时不需要主菜单被激活。

屏幕底行应该特别注意，这里给出了可供使用的功能键。进入不同的窗口时功能键内容会改变。而此时仔细的用法可由F1或Alt-H得到。

经Alt-键使用多窗口，直到所有窗口出现，而只有最顶层的被激活。使用Esc 健关闭当前窗口。使用特定窗口对应的Alt健可进入该窗口，而其它窗口仍显示着。

多窗口情况下屏幕的一部分由窗口遮住，这时可以用Sidekick的窗口移动功能把窗口移

开。这一操作需要ScrollLock键激活(看右下角的标志)，因用光标键移动，所以NumLock应该关闭。移动的窗口的位置可以经Setup 窗口中的Save Window Setup功能来永久记录。

Sidekick在用户操作的任何时刻提供帮助。Sidekick的功能键列如下：

	WORDSTAR	PC
光标移动		
字符左移	Ctrl-S	←
字符右移	Ctrl-D	→
字左移	Ctrl-A	Ctrl-←
字右移	Ctrl-F	Ctrl-→
行上移	Ctrl-E	↑
行下移	Ctrl-X	↓
上卷	Ctrl-W	
下卷	Ctrl-Z	
上一页	Ctrl-R	PgUp
下一页	Ctrl-C	PgDn
行左端	Ctrl-Q-S	Home
行右端	Ctrl-Q-D	End
页首	Ctrl-Q-E	Ctrl-Home
页底	Ctrl-Q-X	Ctrl-End
文首	Ctrl-Q-R	Ctrl-PgUp
文末	Ctrl-Q-C	Ctrl-PgDn
文末加日期/时间	Ctrl-Q-O	
标记区首	Ctrl-Q-B	
标记区尾	Ctrl-Q-K	
上一光标位置	Ctrl-Q-P	
插入和删除		
插入态开/关	Ctrl-V	Ins
插一行	Ctrl-N	
删一行	Ctrl-Y	
删至行末	Ctrl-Q-Y	
删右边一字	Ctrl-T	
删光标处字符	Ctrl-G	Del
删左边字符	Ctrl-H	Backspace
块命令		
定义块首	Ctrl-K-B	F7
定义块尾	Ctrl-K-K	F8
标记单字	Ctrl-K-T	
隐藏/显示块	Ctrl-K-H	
拷贝块	Ctrl-K-C	
移动块	Ctrl-K-V	
删除块	Ctrl-K-Y	
从磁盘读块	Ctrl-K-R	
写块至磁盘	Ctrl-K-W	
块的排序	Ctrl-K-S	
打印块	Ctrl-K-P	
其它命令		
存贮文件	Ctrl-K-D	F2
粘贴块	Ctrl-K-E	
剪贴		F4

块粘贴可以使块内文本复制到任何其它程序。剪贴允许用户把屏幕内容复制到Notepad。

跳格	Ctrl-I
重复最末一次查找	Ctrl-L
控制字符前缀	Ctrl-P
行重新格式化	Ctrl-B
设右边界	Ctrl-O-R
查找	Ctrl-Q-F
查找与替换	Ctrl-Q-A
自动跳格开/关	Ctrl-Q-I
恢复删除的行	Ctrl-Q-L
读取日期和时间	Ctrl-Q-T
图形开/关	Ctrl-Q-G

图形开关Ctrl-Q-G决定图形字符的显示，开启时可以看到256个ASCII字符。

右边界与排版。右边界默认值是65，使用Ctrl-O-R命令来改变它。此后便可以用Ctrl-B来格式化段落了，一个段落以空行结束。右边界设成250来取消排版。

功能键。F1 Help. 显示详细的实时使用信息。

F2 Save. 将文件内容存贮到磁盘。等价于WordStar命令Ctrl-K-D。

F3 New note file. 定义新文件。默认为NOTES。

F4 Cut and paste. 拷贝进入Notepad之前屏幕所显示的内容。

F9 Expand window. 扩大窗口，第一次按键后结合光标键扩大窗口，第二次回至常。

F10 Contract window. 窗口缩小，与F9用法类似，但作用相反。

填加日期和时间。第一行第一列为.LOG的文件为记录文件。每次调入后，系统自动将光标移到文尾并填加当前日期和时间。Ctrl-Q-O命令作用与之相同。命令Ctrl-Q-T把日期和时间加到光标位置。Ctrl-Q 可引导两个字符的控制命令，进行光标移动、删除及其它有关操作。

Ctrl-Q-S、Ctrl-Q-D、Ctrl-Q-E、Ctrl-Q-X、Ctrl-Q-R、Ctrl-Q-C、Ctrl-Q-O、Ctrl-Q-B、Ctrl-Q-K、Ctrl-Q-P、Ctrl-Q-Y、Ctrl-Q-F、Ctrl-Q-A、Ctrl-Q- I 、Ctrl-Q-L、Ctrl-Q-T、Ctrl-Q-G，对应的功能上面已做说明。

字串查找。字串可以带有控制字符，后者用Ctrl-P做前缀。查找与替换的字串可以用以下键来编辑：

字符左移	Ctrl-S 或←
字符右移	Ctrl-D 或→
字左移	Ctrl-A
字右移	Ctrl-F

字右移(Ctrl-F)显示前一次查找的内容供编辑之用。Ctrl-A可用于查找字串时的通配符。字符最长是30。字符替换时的操作与此相仿，有关的选项如下：

B 从光标位置向后查找和替换直至文首。

G 全程查找和替换，作用范围与光标位置无关。

n n = 任意数。查找和替换n次。

N 不显示替换时的确认提示Replace (Y/N)。

U 忽略大小写。

W 仅查找和替换整字，不计嵌到其它字中的情形。

用例：

N10 查找并替换10次，替换时不提示。

GWU 全文字查找和替换，忽略大小写。

查找与替换中途停止时用Ctrl-U命令取消，操作由Ctrl-L重复。

Ctrl-K 也能引导一系列两字符的控制命令，即：

Ctrl-K-B、Ctrl-K-K、Ctrl-K-T、Ctrl-K-H、Ctrl-K-C、Ctrl-K-V、Ctrl-K-Y、Ctrl-K-S、Ctrl-K-E、Ctrl-K-R、Ctrl-K-W、Ctrl-K-P、Ctrl-K-D。

其功能前面明已经说明。

若给出的文件名含有通配符(? 或*)。Sidekick会显示系列符合条件的文件名。*与?的含义与DOS下相同。*.* 指所有文件，*.TXT 则指所有扩展名为TXT的文件，??.? 指文件名长度为2扩展名长为1的文件。存贮的文件名不能以'BAK'作为扩展名，因为这类文件由Sidekick用作备份。

块排序需指定关键字所在的列号。

在Sidekick中可以定义粘贴键，通常用Alt键的组合如Alt-A、Alt-9。键的使用要在块显示时，定义可由Ctrl-K-E Del命令取消。功能键及其与Shift, Ctrl, 或Alt键的组合可以用作粘贴键。

§18.2.5 中文字处理软件WPS

由北京大学和香港金山公司开发的超级汉字操作系统是一个使用方便、功能齐全的桌面印刷系统。以其第五版为例，系统由以下各部分组成。

1 、Super-CCDOS V5.0 中文操作系统

2 、Super-WPS V2.0 中文处理系统

3 、Super-Star V1.0 明星图文编排及印刷系统

4 、辅助程序：如批处理、文本说明等。

系统的进入可以通过安装 S u p e r 汉卡或软件安装而完成。

启动原理框图：

主机自检 → 启动PC DOS → 进入WPS子目录 → 启动CCDOS

WPS5.1 往往使用XMS，加一相应的驱动程序则速度会大大改善。如在CONFIG. SYS中加入：DEVICE=HIMEM.SYS、DEVICE=SMARTDRV.SYS 1024 512，DOS 5.0 也提供这两个驱动程序。结合DOS 的PATH 和APPEND 命令可在任意子目录下调用系统文件。有时计算机内存受限，则可以调用其一级字库。

SPLIB←┘ 调入字库，其中←┘ 表示<Enter>，下同。SPLIB 可以使用参数/2 调用二级字库，使用/[nnnn] 指定程序运行缓冲区大小，这对改善编辑时的显示速度也是很有益的。

SPDOS←┘ ┌装入字库管理模块
        ┤ 装入键盘及显示管理模块
        └装入拼音汉字输入法模块

可 ┌WBX←┘ 装入五笔字型输入法模块

选 └BXM←┘ 装入表形码输入法模块

启动WPS 和SPT: WPS←┘ 或WPS 文件名←┘ ，可使用/N 指示为文本编辑，/V 指示不保留进入WPS 编辑以前的屏幕。

运行SPT 命令进入明星排版系统：SPT←┘

当WPS启动后，屏幕首先显示一个主功能选单，并请用户选择，故不用记忆其功能键。

菜单←文件:DEMO.WPS 行=00057 列=060 控制ON 插入行→

文件操作	块操作	删除字符	光标移动	寻找/替换	打印控制	版面控制	编辑控制	窗口	其它功能
保存文件	定义块首	删除一字	到文件头	寻找	选择字体	字符升高	置左边界	水平分割	模拟显示
存盘返回	定义块尾	删前一字	到文件尾	寻找且替换	选字型号	字符后退	置右边界	垂直分割	文件打印
放弃存盘	块取消	删除一句	到块首	寻找某行	选西文体	字间距	标尺显示	下一窗口	计算器
存盘退出	块复制	删除一行	到块尾	寻找下一个	选择修饰	行间距	标注显示	窗口调整	当前日期
读文件	块移动	删到行首	上一位置		选择划线	左空点数	水平制表		当前时间
块写文件	块删除	删到行尾			选择背景	设定分栏	TAB 宽度		当前星期
DOS命令	拷DOS 块	恢复删除			选择前景	设定栏空	自动制表		计算结果
设置密码					选择阴影		制表连线		重复执行
							取消连线		
							段落重排		

^KS 正在编辑的文本存盘, 继续编辑

半角双拼双音:

WPS 系统的绝大多数的功能都可以从编辑状态发出命令来完成。如上图所示，用ESC键进行功能的选择。用→启用下有相应的下拉式菜单。大亮块可以用光标键来移动，如果按了Enter 键,则大亮块所示的命令就被执行。键盘命令表达式的语法定义:(字符“^”是Ctrl 键的缩写)。如：^KJ 同时按Ctrl 和K,然后按J,进入帮助系统。^J 则是同时按Ctrl 和J, 进入键盘命令解释帮助系统。

光标移动键如小键盘上的→ ← ↑ ↓Home End PgUp PgDn。Ins 是插入和改写状态的转换开关键。Enter 在插入态是光标处插入一硬回车，可插入新行或分断一行，光标移到下一行左端；在改写态则光标移到下一行左端。Esc 系从编辑状态进入选单操作方式，或退出选单、取消命令。Backspace 或←键删除光标左边的一个字符或汉字。Del删除光标指示的一个字或汉字。^ T 、^ Y 等功能与WordStar相仿。使用多窗口(F6，^])进行多个文件间的块操作补非常方便。

当进入选单操作方式时，大亮块指示的功能在屏幕的底行有与其相应的键盘命令表达式和中文解释，便于用户理解，加快学习使用的过程。

文稿输入：输入的文字显示在光标所指定的位置。

1. 英文字符输入法：按Alt+F10，然后从键盘直接键入。全角 / 半角字宽的转换由Ctrl+F9进行。

2. 汉字输入法：按Alt＋F1到F8然后根据所选汉字输入法的规则输入汉字。

3. 1 － 9 区图形字符输入法：按Alt＋F9 重复按几次就显示第几区的符号，再用－或＝左右寻找，然后键入相应数字选择。

4. 手动制表：有粗线和细线二种。粗线，按Alt ＋← ↑ ↓ →。细线，按Ctrl ＋← ↑ ↓ →

注意：画线区域限制在文末符“■”前，如要超过此区，必须先用回车键插入新行以扩充线区。

打印控制：按Esc后选“打印控制”项的功能来定义用户需要的控制符号;并随时可用模拟显示 ^ KI来从屏幕上观察控制效果(从光标处开始显示)，如欲停止显示，按一次Ctrl＋Break即可回编辑状态。实现了“所见即所得”。

当用户对编辑结果和模拟显示效果满意后，就需要用打印机在纸上打印出来，这时请按^KP即可。从打印机上得到的结果与模拟显示结果是一样的。

当用户有了一定的使用经验后，即可处理更复杂的文稿和版面：

1 、块操作，磁盘读写

2 、多窗口操作

3 、左右边距

4 、计算器

5 、指定字段的查找定位和替换新字段

6 、其他功能组合

在系统软盘上有一个名为SAMPLE.WPS的样本文件，它既可向用户演示本系统的打印效果，又可向用户示范怎样去控制实现这样的功能（用 D 方式读入SAMPLE.WPS，用模拟显示功能^KI观察控制效果）。

SPT 图文编辑系统使用TIF 格式的文件。在系统下用^F10 可以调出系统控制的选单，仍以←、→、↑ 、↓行选择。第一项控制系统的录入方式，如区位、五笔字型、英文数字等。以回车键行录入方式的选择，以ALT-ESC 撤出该录入方式；其辅助功能提供ASCII码表、日程表、计算器、汉字区位码，利用其辅助功能还可以退出系统。若使用WPS 5.1，此时仍可以进入WPS 的编辑，只是退出编辑后返回到西文状态。在WPS 4.03 版中也有类似的功能。

Super-CCDOS 5.1 若使用字库XSDOS.LPH，由于文件较大，在支持大硬盘分区的DOS 5.0 下读字库有问题，应行专门的处理。如将计算机C: 盘分区做成32 MB，存放字库XSDOS.LPH，把SAS 等占空间较大的软件放在大分区的D: 盘中，然后改动SPLIB.EXE 文件，如使用PCTOOLS查找008B处的指令代码“80,7F,16,F8”并将其中的16改成17，查找00EB处指令代码“8B, 47, 10”将其中的10改成11。改动对MS-DOS 6.0 和DR DOS 6.0亦有效。

WPS 系统兼容四通2401 格式，其早期版本可行两者文件的互换，在5.0 以后则只支持四通文件到WPS文件的单向转换。据实际使用经验，四通打字机的软盘出现“磁盘格式不对”的错误，经常造成信息的损失，幸运的是大多数情况下可以找到解决的方法。四通盘容量是720KB，可经计算机的小盘驱动器驱动，可用计算机的工具软件将原四通盘备份一次。针对备份盘进行CHKDSK、NORTON软件处理或其他操作，找回文件碎片，而西山WPS系统可直接使用这些找回的信息。

下面介绍两个与WPS有关的实用程序，它们均可经C 编译生成相应的执行文件。第一个是计算机与四通格式互换的程序，其使用标准的ASCII 格式。四通的有格式文件可经西山生成ASCII文件，经计算机处理后再使用该程序转回四通格式。/*《中国计算机用户杂志》1992.3 */

```
#include <stdio.h>
#define MAXLEN (48*1024)
char r=0;
```

```
main(argc,argv)
int argc;char *argv[];
{
    char c=1;
    if (argc==1){

      puts("四通文件和微机文件转换程序\n");
      puts("Usage: MSPC -[rmp] < [源文件名] > [目标文件名]");
      puts("   -m 四通文件==> 微机文件");
      puts("   -p 微机文件==> 四通文件");
      puts("   -r 转换控制字符(default)");
      exit(0);

    }
    if (**++argv=='-'){
        if((strchr(*argv,'r')!=NULL)||(strchr(*argv,'R')!=NULL)) r=1;
        if((strchr(*argv,'m')!=NULL)||(strchr(*argv,'M')!=NULL)) c=1;
        if((strchr(*argv,'p')!=NULL)||(strchr(*argv,'P')!=NULL)) c=0;
    }
    if (c) mtop();
    else ptom();
}
mtop()
{   register unsigned char *buft,c=0,e=0,*buf;
        short rnum,wnum;
        buf=(unsigned char *)malloc(MAXLEN);
        while((rnum=fread(buf,1,MAXLEN,stdin))>0){
           wnum=rnum;
           buft=buf;
           for(;rnum>0;){
             rnum--;
             if(*buft>=0xa1) c=(c==1)?0:1;
             else if (c) { *buft+=128;c=0;}
             else if (e) { *buft=32;e=0;}
             else {
                if (r) {
                   if (*buft=0x8d) *buft=0xd;
                   else if (*buft==0x8a) *buft=0xa;
                   else if ((*buft&0x80)&&(*buft-128<32)){
                           *buft=32;e=1;}
                   }
              } buft++;
           }fwrite(buf,1,wnum,stdout);
```

```
        }free(buf);}
ptom()
{
    register unsigned char *buft,c=0,*buf;
    short rnum,wnum;
    buf=(unsigned char *)malloc(MAXLEN);
    while((rnum=fread(buf,1,MAXLEN,stdin))>0){
        wnum=rnum;buft=buf;
        for (;rnum>0;) {rnum--;
          if (*buft>=0xa1) {
             if (c) *buft-=128;
                c=(c==0)?1:0;}
             else {c=0;
             if(r && *buft<32)*buft=0x80;}
        buft++;}
        fwrite(buf,1,wnum,stdout);}
        free(buf);
}
```

设现已用TURBO C 2.0 将上述文件编译、连接并生成可执行文件MSPC.EXE。要把西山文件转成四通文件，则先用西山的文件经主编辑主菜单下的文件服务功能转为标准文本文件TRANS.ASC，然后进行操作：

C>MSPC -p <TRANS.ASC >TRANS.240 <Enter>

则将ASCII 类型文件TRANS.ASC 转为四通文件TRANS.240 了。

第二个实用程序是处理WPS密码的，因为有时使用WPS时会不慎加入密码或忘记密码，这个小程序可能有所帮助。经分析，WPS文件密码存放在文件的732-740字节。其八位二进制字节高四位与低四位交换后取反，小写字母换为大写。功能键ASCII 码按零算。

/*《计算机世界》1994.2.2 */

```
#include<stdio.h>
#include<process.h>
main(int argc,char *argv[])
{
  FILE *fp;
  int i,j,a[8],b[8];
  printf("\nWPS Password Retriever V1.0\n");
  printf("(C)Copyright China Computer World 02/02/94 \n\n");
  if(argc!=2){printf("Usage: PASS filename\n");
     exit(-1);}
  if((fp=fopen(argv[1],"r"))==NULL){
      printf("Can't open %s\n",argv[1]);
      exit(-1);}
   fseek(fp,733L,SEEK_SET);
```

```
    if ((a[0]=fgetc(fp))==0){
         printf("No Password\n",a[0]);
         exit(-1);}
    else printf("Internal Codes: %2x",a[0]);
    i=1;
    while(i<8){
         if ((a[i]=fgetc(fp))==0) break;
         printf("%02x",a[i]);
         i++;}
    putchar('\n');
    fclose(fp);
    printf("Password:");
    for(j=0;j<i;j++){
       b[j]=~(a[j]<<4|a[j]>>4)&255;
       if(b[j]==0) printf("Function keys");
       else if (b[j]==8) printf("Backspace");
       else if (b[j]=='\t') printf("Tab");
       else if (b[j]==27) printf("Esc");
       else if (b[j]==' ') printf("Space");
       else if (b[j]>='!' && b[j]<='~') printf("%c",b[j]);}
    putchar('\n');
    exit(-1);
}
```

下面是WPS 帮助信息。

一、输入及编辑

^D 或→键	光标右移一个字符
^S 或←键	光标左移一个字符
^E 或↑键	光标上移一行
^X 或↓键	光标下移一行
^QE	光标移到当前窗口左上角
^QX	光标移到当前窗口最后一行末尾
^R 或PgUp键	向上翻一页面
^C 或PgDn键	向下翻一页面
^A	光标左移到当前句句首
	以下字符为一句的有效结束符号
	Tab,空格,"!$&()+-*/,.:;=?\\|[]
	{} <>",回车,分页,文末符
^F	光标移至下一句的句首
^QS 或Home键	光标左移到当前行行首
^QD 或End键	光标右移到当前行行尾
^W	窗口向上滚动一行, 屏幕向下滚动一行
^Z	窗口向下滚动一行, 屏幕向上滚动一行
^QR 或^Home	光标移到当前文件开头
^QC 或^End	光标移到当前文件末尾
^V 或Ins键	插入 / 改写状态转换
^I 或Tab键	使光标向右跳到下一个制表站
	在N打开方式下: 光标移动一个Tab位置
^G 或Del键	删除当前光标的一个字符或汉字
^H 或Backspace	删除当前光标的前一个字符或汉字
^T	删除光标指向的字符及其后一句内的所有字符
^Y	删除光标所在的一行
^QY 或^	删除光标处到行尾的所有字符
^QH 或^←	删除光标处到行首的所有字符
^U	恢复最近一次删除的内容到光标位置
	注: 不能恢复块删除的内容
^M 或Enter键	在当前光标处插入一个硬回车
	在改写状态下: 将光标移到下一行的行首
^N	在当前光标处插入一个硬回车,光标不移
^PP	在当前光标处插入一个分页符

二、文件操作

^KW	将定义的块写入磁盘文件
^KR	将磁盘文件读入到当前光标位置
^KS	正在编辑的文本存盘，继续编辑
^KD 或F2	正在编辑的文本存盘，停止编辑，返回主选单
^KX	正在编辑的文本存盘，停止编辑，退出WPS系统
^KQ 或F3	停止编辑，文本不存盘，返回主选单
^OP	在D打开方式下设置文件密码
	设置密码后必须存盘才能使密码有效

三、块操作

^KB 或F4	将光标所在位置设成块首
^KK 或F5	将光标所在位置设成块尾
^QB	将光标移到块首位置
^QK	将光标移到块尾位置
^KH	取消已定义的块
^KV	将块移到当前光标所在位置
^KC	将块复制到当前光标所在位置
^KY	将块删除
^KN	块的行 / 列方式转换
^KL	拷贝CCDOS屏幕内容到当前光标处

四、查找和替换

^QF 或F7	在文本中查找指定的字句
^QA	在文本中查找指定的字句，并用一些字句来替换
	查找或替换时的控制符:
	^S - 通配任何字符
	^A - 通配任何ASCII字符
	^C - 通配任何汉字或打印控制符
	^P^M - 表示回车符
	^P^L - 表示分页符
	^P^J - 表示软回车
^L	重复前一次查找或替换命令
^QV	光标返回到上一次工作点位置
	上一次工作点指上一次查找或替换时的位置
^QL	光标移动到指定行行首

五、格式编排及制表

^OL	设置文本左边距
^OR	设置文本右边距
^B	根据新的左右边距对文本进行段落重排
^OF	标尺显示开 / 关
^OI	制表站的设定
^OC	控制符号显示的开 / 关
^OK	设置Tab宽度
^OA	自动制表功能
^OS	制表连线，将块首与块尾连接一制表线
^OY	删除块首与块尾之间的制表线
^OE	设置左边界字的点数
	为了使文本中自动产生的左边界字符的字号不受文本中字号大小的控制, 用左边界字的点数来表示文本中左边界字的字号
^OZ	设置分栏打印栏空格数
	栏空格数×8为栏与栏之间的点数

六、打印控制

^PA	选择字体: 宋体、仿宋体、楷体、黑体
^PB	选择字型: 标准、长型、扁型、自定义、特大、统一
^PC	选择上下划线
^PD	选择汉字修饰
^PE	定义字符背景内容
^PF	选择英文字体
^PG	字符后退n个半角字(0 127)
^PH	字符升高n个点(-63 64点)
^PK	定义字符间距(-63 64点)
^PL	定义行间距(0 127点)
^PM	定义字符阴影
^PN	定义字符前景
^PS	设定分栏打印

七、窗口操作及其它

^KZ 或F6	定义一个窗口
^QN 或^]	将光标转到下一窗口
^KO	窗口尺寸的调整
^KI 或F8	从当前光标位置模拟显示
^KP 或F9	从当前光标位置打印
^KA 或^Ins	请求计算器服务功能
^QQ	重复执行命令功能
^KJ 或F1	进入帮助系统
^KF 或F10	执行DOS或CCDOS命令
^OD	取当前日期到当前光标位置
^OT	取当前时间到当前光标位置
^OW	取当前星期到当前光标位置
^OM	取计算器结果到当前光标位置

用PgUp、PgDn键翻页，ESC键返回

附录一　参考文献

第 1 章

[1] 陈希孺:《统计学概貌》，科学技术文献出版社，1989.

[2] Moore, D.S. (1990), Uncertainty. in Steen, L.A. (eds), On the Shoulders of Giants, New Approach to Numeracy. National Academy of Sciences.

[3] Dowdy, D. and Wearden S.(1991), Statistics for Research, second edition, John Wiley & Sons Inc.

[4] Turkey, J.W. Exploratory Data Analysis, Reading, MA., Addison-Wesley.

[5] Cox, D.R. and E. J. Snell(1981), Applied Statistics: Principles and Examples, Chapman and Hall.

[6] Chatfield C.(1988), Problem Solving—A Statistician's Guide, Chapman and Hall.

[7] Mcpherson G.(1990),Statistics in Scientific Investigation: Its Bases, Application, and Interpretation. Springer Verlag.

[8] Afifi, A.A. and Clark V.(1990), Computer-aided Multivariate Analysis, 2nd Edtion, Van Nostrand Reinhold Company.

[9] Norman Cliff(1987), Analyzing Multivariate Data, Harcourt Brace Jovaovich，Inc.

[10] James Stevens(1992), Applied Multivariate Statistics for the Social Sciences, Second Edition. Lawrence Erlbaum Associates, Inc.

[11] Jobson, J.D.(1991), Applied Multivariate Data Analysis, Vol I. Regression and Experimental Design, Springer-Verlag.

[12] Ross, P.W. eds, The Handbook of Software for Engineers and Scientisits, CRC and IEEE Press, 1996.

第 2 章

[1] 上海医科大学:《中国医学百科全书》第一卷，预防医学。上海科学技术出版社，1991.12.

[2] Box, G.E.P. and N. R. Draper (1987), Empirical Model-Building and Response Surfaces, John Wiley & Sons, Inc.

[3] Plackett,R.L. and J.P.Burman(1946), The Design of Optimum Multifactorial Experiments, Biometrika, 33, 305-325.

[4] Box, G.E.P and D. W. Behnken (1980), Some New Three Level Designs for the Study of Quantitative Variables, Technometrics, 2, 455-475.

[5] Box, G. E. P and K. B. Wilson (1951), On the Experimental Attainment of Optimum Conditions, J. Roy. Statist. Soc., B13, 1-38.

[6] Andre I Khuri, and J. A.Cornell(1987), Response Surfaces: Designs and Analysis, Marcel Dekker, New York.

[7] SAS/QC Software: Reference Version 6, First Edition. SAS Institute, Inc.

[8] Seber, G.A.F.(1977), Linear Regression Analysis. John Wiley & Sons, Inc.

[9] Stevens, S.S. (1951), Mathematics, Measurement, and Psychophysics, in stevens, S. S. ed. Handbook of Experimental Psychology, New York, Wiley.

[10] Afifi, A.A. and Clark V.(1990), Computer-aided Multivariate Analysis, 2nd Edtion, Van Nostrand Reinhold Company.

[11] Cliff N. (1987), Analyzing Multivariate Data, Harcourt Brace Jovanovich，Inc.

[12] Jobson, J.D.(1991), Applied multivariate Data Analysis, Vol I. Regression and Experimental Design, Springer-Verlag.

[13] Hoaglin, D.C., Mosteller, F. and Turkey, J.W.(1983), Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, Inc.

[14] Dixon, W.J.(1951), Ratios involving extreme values, Ann. Math., 22, 68-75.

[15] Dowdy, D. and Wearden S.(1991), Statistics for Research, second edition, John Wiley & Sons Inc.

[16] Mendenhall, W. and T. Sinsich (1992), Statistics for Engineering and the Sciences. Dellen Publishing Company, 2rd Edition.

[17] Madansky, A. (1988), Prescriptions for Working Statisticians. Springer-Verlag.

[18] Johnson, R.A. (1988), Applied Multivariate Statistical Analysis, Prentice-Hall, Inc.

[19] Hartley, A.J.(1983), Appropriate uses of multivariate analysis, Ann Rev. Public Health, V4 P155-180.

[20] Rao, C.R.(1973), Linear Statistical Inference and Its Applications Wiley, New York.

[21] M.S. Srivastava and E.M. Carter (1983), An Introduction to Applied Multivariate Statistics, Elsevier.

[22] Crowder, M.J. and D. J. Hand (1990), Analysis of Repeated Measures，Chapman and Hall.

[23] Cox, D.R. and Hinkley, D.V.(1974), Theoretical Statistics. Chapman and Hall, London.

[24] Snedecor, G.W. and Cochran, W.G. (1980), Statistical Methods (7th Edition), Iowa State University Press.

[25] 周概容:《概率论与数理统计》，高等教育出版社，1984.

[26] Collett, D. (1994), Modelling Survival Data in Medical Research, Chapman and Hall.

第 3 章

[1] Yu- cheng Liu，Gibson, A.G.(1984), Microcomputer Systems: The 8086/8088 Family —Architecture, Programming & Design. Prentice -Hall, Inc.

[2] 朱毕:《计算机病毒的预防与消除》，中国科学院希望电脑公司，1991.1.

[3] Gregory, W. and W. Wojtkowski (1990), Applications Software Pro gramming with Fourth-generation Languages, Boyd & Fraser Publishing Co.

[4] Schofield, C.F.(1989), Optimizing Fortran programs, Ellis Horwood Ltd.

[5] Sanchez, J. (1990), Assebly Langauge Tools and Techniques for the IBM Microcomputers. Prentice Hall, Inc.

[6] Forsythe,G.E.,M.A. Malcolm and C.B. Moler(1977), Computer Methods for Mathematical Computations, Printice-Hall, Englewood Cliffs, N.J.

[7] Maindonald, J.H.(1984),Statistical Computations, John Wiley & Sons Inc.

[8] Press, W.H., B.B., Flannery，S.A., Teukolsky and W.T., Vetterling (1986), Numerical Recipes - The Art of Scientific Computing, Cambridge University Press.

[9] Wilkinson,J.H. and C. Reinsch(1977), Linear Algebra, Vol. 2 of Handbook for Automatic Computations, Springer-Verlag, N.Y.

[10] Wilkinson, J.H. and Reinsch, J.(1971), Handbook for automatic computation. Vol III. Linear Algebra, Springer-verlag, Berlin.

[11] Sewell, G.(1990), Computational Methods of Linear Algebra, Ellis Horwood Ltd.

[12] Kennedy, W.J. and J.E.Gentle (1981), Statistical Computing. Marcel Dekker, Inc.

[13] Enslein,K.,A. Ralston and H.S. Wilf, eds(1977),Statistical Methods for Digital Computers (Vol.3 of Mathematical Methods for Digital Computers) Wiley-InterScience, N.Y.

[14] 郭祖超主编:《医用数理统计方法(第三版)》，人民卫生出版社，1983.

[15] 冯康:《数值计算法》，国防工业出版社.

[16] 刘德贵、费景高等:《Fortran 语言算法汇编》，国防工业出版社，1983.

[17] 上海计算技术研究所:《电子计算机算法手册》，上海教育出版社，1982.

[18] 中国科学院计算中心概率统计组:《概率统计计算》，科学出版社，1979.

[19] 中国科学院沈阳计算技术研究所等:《电子计算机常用算法(增订版)》，科学出版社，1983.

第 4 章

[1] Sincich, T.(1982), Statistics by Example, 3rd edtion Dellen Publishing Company, a division of Macmillan, Inc.

[2] Bollen K. A.(1990), Regression Diagnostics. in J. Fox and J. Scottloag eds.Modern Methods of Data Analysis, Sage publications inc.

[3] 吴国富, 安万福, 刘景海:《实用数据分析方法》，中国统计出版社，1992.

[4] Rawlings, J.O.(1988), Applied rgression analysis: A research tool. Wadsworth, Inc.

[5] Chatterjee S. and Price, B. (1991), Regression Analysis by Example, Second Edition, John Wiley & Sons, Inc.

[6] Draper, N.R. and Smith, H.(1981), Applied Regression Analysis (2nd Edition), Wiley, New York.

[7] Wetheril, G.B. (1986), Regression Analysis with Applications, Chapman and Hall.

[8] 陈希孺、王松桂:《近代回归分析》，安徽教育出版社，1987.

[9] 谭启华、何大卫：医用线性回归变量变换及诊断,《中国卫生统计》第十卷第一期，1993.

[10] Cody, R.P. and Smith J.K. (1991), Applied Statistics and the SAS Programming Language, North-Holland.

[11] Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978), Statistics for Experimenters, New York: John Wiley & Sons, Inc.

[12] Searle, S.R. (1987), Linear models for unbalanced data, John Wiley & Sons, Inc.

[13] Searle, S.R., Casella, G. and C.E.McCulloch (1992), Variance Components, John Wiley & Sons, Inc.

[14] Henderson, C.R. ANOVA, MIVQUE and ML Algorithms for estimation of variances and covariances, in David, HA and David, HT eds. Statistics,An Appraisal. The Iowa State University Press, Inc.

[15] Karlin, S, et.al (1983), Path analysis in epidemiology —A Critique. American Journal of Human Genetics, 35:695-723.

[16] Maindonald, J.H.(1984),Statistical Computations, John Wiley & Sons Inc.

[17] M.S. Srivastava and E.M. Carter (1983), An Introduction to Applied Multivariate Statistics, Elsevier.

[18] Mike V. and K.E.Stanley (1982), Statistics in Medical Research, John Wiley & Sons, Inc.

[19] Agresti, A.(1990), Categorical data analysis, Johns Wiley & Sons, Inc.

[20] Cox, D.R.(1970), Analysis of Binary Data, Chapman and Hall, London.

[21] Hosmer, D.W. and S. Lemeshow (1989), Applied logistic regression. John Wiley & Sons, Inc.

[22] Lee T.E. (1992), Statistical Methods for Survival Data Analysis, Second Edition, John Wiley & Sons, Inc.

[23] SAS Technical Report P-229, SAS/STAT Software Changes and Enhancements, Release 6.07, p251-286, SAS Institute Inc., Cary, NC, USA,1992.

[24] 杨维权，刘兰亭，林鸿洲编:《多元统计分析》高等教育出版社，1989.

[25] SAS/QC Software: Reference Version 6, First Edition. SAS Institute, Inc.

[26] Box, G.E.P. and N. R. Draper (1987), Empirical Model-Building and Response Surfaces, John Wiley & Sons, Inc.

第 5 章

[1] Schulman, R.S. (1992), Statistics in Plain English with Computer Application, Van Nostrand Reinhold.

[2] 杨维权，刘兰亭，林鸿洲编:《多元统计分析》高等教育出版社，1989.

第 6 章

[1] 黄因敏: 几种大型统计分析软件包的介绍及性能比较,《中国医药信息大会论文集(三)》，781-791，1987年4月.

[2] Dixon, W.J. (1981), BMDP Statistical Software, 1981. Berkeley, CA.

[3] Agresti, A. (1990), Categorical Data Analysis. John Wiley & Sons, Inc.

[4] Gehan, E.A.(1965), A Generalized Wilcoxon test for comparing arbi trarily singly-censored samples, Biometrika 52:203-223.

[5] Collett, D. (1991), Modeling Binary Data. Chapman and Hall.

[6] Smith, W. (1932), Filtration of antipneumococus serum. Journal of Pathology 35, 509-526.

[7] BMDP USER'S DIGEST: A Condensed Guide to the BMDP Computer Programs BMDP Ststiatical Software, Inc., 1984.

[8] Afifi, A.A. and Clark V. (1990) . Computer-aided Multivariate Analysis, 2nd Edtion, Van Nostrand Reinhold Company.

[9] BMDP, Statistical Software Inc. (1988), BMDP Statistical Software, Los Angeles.

[10] Kalbfleisch, J.D. and Prentice, R.L.(1980), The Analysis of Failure Time Data, New York: John Wiley & Sons, Inc.

第 7 章

[1] Wilkinson, L.(1987), SYSTAT: The System for Statistics. Systat Inc.，Evanston, Illinois

第 8 章

[1] Stata 2.0, Computing Resource Center, 1081 National Boulevard, Los Angeles, California 90064, 1988.

第9章

[1] Ripley, B.D. (1994). Introductory Guide to S-Plus, statlib.

[2] Chambers, J.M. and Hastie, T.J. (1992). Statistical Models in S, Wadsworth & Books.

[3] Ross, P.W. eds, The Handbook of Software for Engineers and Scientisits, CRC and IEEE Press, 1996.

[4] S-PLUS User's Manual: Version 3.2 MathSoft, Inc. StatSci Division, Seattle WA.

第10章

[1] Ross, P.W. eds, The Handbook of Software for Engineers and Scientisits, CRC and IEEE Press, 1996.

第11章

[1] Digby, P.G.N., Galwey, N.W., and Lane, P.W.(1988), Genstat 5: a second course. Oxford University Press.

[2] Lane, P.W., Galwey, N.W., and Alvey, N.G.(1988), Genstat 5 : an introduction. Oxford University Press.。

[3] Genstat 5 Reference Manual. Oxford Science Publications 1988.

[4] Wilkinson, G.N. and Rogers, C.E.(1973), Symbolic description of factorial models for analysis of variance. Applied Statistics 22, 392-9.

[5] Collett, D. (1991), Modeling Binary Data. Chapman and Hall.

第12章

[1] Box and Pierce(1970), Distribution of residual autocorrelations in autoregressive integrated moving average time series models, JASA, Vol 65, PP 1509-1526.

[2] White, H. (1980), A heteroskedasticity-consustent covariance matrix and a driecct test for heteroskedasticity, Econovetrca 48,

[3] Pao, P. and Z. Griliches (1969), Small sample properties of several two-stage regression methods in the context of autocorrelated errors, JASA, Vol 64, pp 253-272.

[4] Cochrane, D. and G.H. Orcut (1949), Application of least squares regression to relationships contaioning autocorrelated error terms" JASA, Vol 44, PP 32-61。

[5] Box, G.E.P. and Jenkins, G.M.(1970), Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco.

第13章

[1] Lindsay, J. K.(1989), The Analysis of Categorical Data Using GLIM, Springer-Verlag.

[2] Agresti, A. (1990), Categorical Data Analysis, John Wiley & Sons,Inc.

[3] Baker, R.J. and J.A. Nelder GLIM Manual Version 3.77.

[4] Healy, M.J.R.(1988), GLIM: An Introduction. Clarendon Press, Oxford.

[5] Crawley, M.J.(1993), GLIM for Ecologists. Blackwell Scientific Publications.

[6] Collett, D. (1991), Modeling Binary Data. Chapman and Hall.

[7] McCullagh, P. and Nelder, J.A.(1983), Generalized linear models. Chapman and Hall, New York.

[8] Agresti, A. (1990), Categorical Data Analysis. John Wiley & Sons,Inc.

[9] 陈希孺、王松桂:《近代回归分析》，安徽教育出版社，1987.

[10]王学仁、陈希镇，广义线性模型——回归模型的统一理论,《应用概率统计》，第六卷，第二期，1990年8月.

[11]Lindsay, J. K. (1992), The Analysis of Stochastic Processes using GLIM, Springer-Verlag.

[12]Wilkinson, G.N. and Rogers, C.E.(1973) Symbolic description of fac torial models for analysis of vairance. Applied Statistics 22,392-9.

[13]Hastie, T.J. and Tibishirani (1990) Generalized Additive Models. Chapman and Hall.

[14]Liang, K.Y. and Zeger, S.L.(1986) Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.

[15]Francis, B., Green, M. and Payne, C.(1993) The GLIM System Release 4 Manual, Clarendon Press, Oxford.

第14章

[1] Karl G. Joreskog and Sorbom, D.(1985), LISREL VI; Analysis of Linear Structural Relationships by the Method of Maximum Likelihood, Instrumental Variables, and Least Squares Methods ，Uppsala: University of Uppsala.

[2] Fox J.(1984), Linear Statistical models and related methods. John Wiley & Sons.

[3] Blau, P and Duncan, O.D.(1967),The American occpational structure. New York, Wiley.

[4] Byrne M. B.(1989), A primer of LISREL —Basic Applications and Programming for Confirmatory Factor Analytic Models. Spinger-Verlag.

[5] Duncan, O.D.(1975),Introduction to Structural Equation Models, New York, Academic Press.

[6] Johnson, R.A. (1988), Applied Multivariate Statistical Analysis, Prentice-Hall.

[7] SAS Technical Report P-229, SAS/STAT Software Changes and Enhancements, Release 6.07, SAS Institute Inc., Cary, NC, USA,1992.

[8] John, C. Loehlin (1992), Latent Variable Models: An Introduction to Factor Path and Structural Analysis, Lawrence Erlbaum Associates.

第15章

[1]. Andrew G. Dean: Epi Info Version 5 User Guide. Center for Disease Control, Atlanta, Georgia, 1990

[2]. 俞顺章、俞国培等译：疾病数据的管理和分析Epi Info 软件使用手册. 上海医科大学出版社1992.12.

第16章

[1] SAS User's Guide: Basics, Version 5 Edition. SAS Institue Inc 1985.

[2] Cody, R.P. and J.K. Smith (1991), Applied Statistics and the SAS Programming Language. Third Edition, North Holland.

[3] Comery, A.L. and H.B., Lee(1992), A First Course in Factor Analysis，Second Edition, Laurence Eribaum Associates, Inc.

[4] Pendhanker, S.S. and R.A., Biegel(1992), dBASEⅣfor VMS and UNIX—A Techinical Support Approach, Van Nostrand Reinhold.

第17章

[1] 胡国定、张润楚：《多元数据分析方法—纯代数处理》南开大学出版社，1990.

[2] Grabowski R. (1991), AutoCAD technical reference, Delmar Publishing Co.

第18章

[1] 国家技术监督局等编:《当代青年必备知识选编》，1993年6月.

[2] Dashefsky, H.S. and Sax, N. (1992) , Desktop Publishing with WordPerfect for Windows, Wincrest, McGraw-Hill.

附录二　统计软件包信息

§附录B.1　名称缩写与英文对照

SAS	Statistical Analysis System
SPSS	Statistical Package for Social Sciences
BMDP	BioMeDical computer Package
SYSTAT	System for Statistics
Genstat	GENeral STATistical modelling
GLIM	Generalized Linear Interactive Modelling
LISREL	LInear Structural RELation models
EGRET	Epidemiological, Graphics, Estimation and Testing Program
IMSL	International Mathematics and Statistics Library
WP	WordPerfect
PE II	Personal Editor II
SK	SideKick
WS	WordStar

§附录B.2　软件公司地址

1. SAS Institute Inc.
   (SAS PC)
   SAS Campus Dr.
   Cary, NC 27513
   USA.
   Tel (919)677-8000
   Fax (919)677-8123

2. SPSS Inc.
   444 North. Michigan Avenue
   Chicago, IL 60611.
   USA
   Tel (312)329-2400
   Fax (312)329-3668

3. Genstat 5 & GLIM

NAG,Ltd	或	Numerical Algorithms Group Inc.
Wilkinson House		1400 Opus Place, Suite 200
Jordan Hill Road		Downer's Grove
Oxford OX2 8DR		IL 60515-5702
UK		USA
Tel 865 511245		Tel 708 971 2337
Fax 865 310139		Fax 708 971 2706

4. BMDP Statistical Software
   1440 Sepulveda Blvd, Suite 316,
   Los Angeles, CA 90025
   USA
   Tel 213-4-7799

5. Marketing Department, SYSTAT Inc.
   1800 Sherman Ave.
   Evanston, IL 60201
   USA
   Tel (708)854-5670
   Fax (708)492-3567

6. Stata
   Stata Corporation
   702 University Drive East
   College Station, TX 77840
   USA
   Tel (409)696-4601, (800)782-8272
   Fax (409)696-4601
   E-mail: stata@stata.com

7. Minitab, Inc.
   3081 Enterprise Drive
   State College, PA 16801
   USA
   Tel (800)448-3555
   Fax (814)238-4383

8. AT&T Software Sales
   P.O.Box 25000
   Greenshoro, North Carolina 27420

(800)828-UNIX

9. EGRET
   Statisticas & Epidemiology Research Corporation
   909 Northeast 43rd Street, Suite 310
   Seattle, Washington 98105
   USA

10. 汉化Epi Info
    上海200032 医学院路138 号上海医科大学卫生统计学教研室金丕焕
    英文版URL: ftp//ftp.cdc.gov (EPI INFO 和EPI MAP)

11. IMSL Inc.
    2500 Park West Tower One
    2500 City West Boulevard
    Houston, TX 77042-3020
    USA

12. MATHEMATICA
    Wolfram Research, Inc.
    100 Trade Center Drive
    Champaign, Illinois 61820-7237
    USA
    Tel (217)398-0700
    Fax (217)398-0747

§附录B.3 标准软件参考资料

1. SAS
   SAS/FSP Guide Version 6, 1987
   SAS Introductory Guide for Personal Computers Release 6.03,1988
   SAS Procedure Guide Release 6.03, 1988
   SAS Language Guide for Personal Computers Release 6.03,1988
   SAS/STAT Guide for Personal COmputers Version 6, 1987
   SUGI Supplemental Library User's Guide Version 5, 1986

2. SPSS
   SPSS Data Entry II, 1988
   SPSS-X User's Guide, 3rd Edtion, 1988
   SPSS/PC+V2.0, Base Manual, 1988

SPSS/PC+ Advanced Statistics V2.0, 1988
SPSS/PC+V3.0 Update Manual
SPSS 6.1 for Windows Update, SPSS Inc. 1994
SPSS Professional Statistics 6.1
SPSS Advanced Statistics 6.1
SPSS for Windows Base System
SPSS for Windows User's Guide Release 6.0
SPSS Tables, Trends, Categories, CHAID and LISREL 7

3. BMDP

BMDP Statistical Software Manual, Volume 1,1988
BMDP Statistical Software Manual, Volume 2,1988

4. SYSTAT

SYSTAT: The System for Statistics, Evanston, Illinois, 1988.

5. Stata

Stata 2.0, Computing Resource Center, 1081 National Boulevard, Los Angeles, California 90064, 1988.

Stata Reference Manual: Release 3, 5th edition, Computing Resource
Center, Santa Monica, CA.
Stata Reference Manual Release 5, Stata Press, College Station, Texas, 1996.

6. Lisrel

Karl G. Joreskog and Sorbom, D.(1985), LISREL VI; Analysis of Linear Structural Relationships by the Method of Maximum Likelihood, Instrumental Variables, and Least Squares Methods，Uppsala: University of Uppsala.

Joreskog, K.G., Sorbom, D.(1989). LISREL 7: User's Reference Guide, Scientific Software, Mooresville, 1st, ed. 1989.

Joreskog, KG, Sorbom, D(1990) LISREL 7: A Guide to the Program and Applications, Chicago: SPSS Inc.

7. GLIM

Baker, R.J. and Nelder, J.A. (1978) . GLIM Manual, Release 3. Numerical Algorithms Group and Royal Statistical Society, c/o Numerical Algorithms Group, Oxford, U.K.

Baker, R.J. and J.A. Nelder GLIM Manual Version 3.77

Baker, R.J.(1981), GLIM-4 et al. GLIM newsletter, issue 5(December)

Francis, B., Green, M. and Payne C.(1993). The GLIM SYSTEM Release 4 Manual, Clarendon Press, London.

8. Genstat

Genstat 5. Reference Manual. Oxford Science Publications 1988.

Release 1.3 of the GENSTAT 5 Statistics System, Oxford, Clarendon Press, 1988.

9. S-PLUS

Becker, R.A. and J.M., Chambers (1980), S: A Language and System for Data Analysis. Bell Laboratories, Murray Hill, N.J.

Becker, R.A. and J.M., Chambers (1984), Design of the S system for Data Analysis, Communications of the ACM, May 1984, No.5.

Becker, R.A., Chambers, J.M., and A.R., Wilks (1988) . The New S Language, Belmont, California:Wordsworth.

10. Minitab

Ryan, T.A., Joiner, B.L. and B.F., Ryan (1981), Minitab Reference Manual. Minitab Project, Philadelphia.

Ryan, B.F., Joiner, B.L. and Ryan, T.A., Jr.(1985), Minitab Handbook, 2nd edition, Duxbury, Boston.

Schulman, R.S. (1992), Statistics in Plain English with Compter Applications, Van Nostrand Reinhold, N.Y.

Minitab Graphics Manual. Minitab Inc. State College, PA.

Minitab QC Manual. Minitab Inc. State College, PA.

Minitab Reference Manual. Minitab Inc. State College, PA.

Ryan, B.F., Joiner, B.L. and T.A., Rayn, Jr., Minitab Handbook. PWS -Kent Publishing Company, Boston, MA, 1992.

11. IMSL

IMSL (1982), IMSL Library Reference Manual. IMSL. Inc., Houston, Tex.

IMSL Library: Fortran Subroutines for Mathematics and Statistics，Edition 9.2, Houston, 1984.

IMSL (1987), IMSL User's Manual, math/library version 1.0, IMSL, Houston.

IMSL (1991), IMSL stat/library user's manual version 2. 0 IMSL, Houston.

12. MATHEMATICA

Martha, L. Abell and James, P. Braselton (1992) . The Mathematica Handbook, Academic Press.

13. NAG

NAG (1993). The NAG FORTRAN Library Manual, Mark 16. Oxford: NAG.

14. GAUSS

GAUSS (1991). The GAUSS System, version 2.2, Kent, Washington: Aptech Systems, Inc.

§附录B.4 URL 地址

- ADE-4: J. Thioulouse, D. Chessel, S. Dolec, Universty of Lyon
  http://biomserv.univ-lyon1.fr/ADE-4.html
- Amos: SmallWaters Corporation
  http://www.smallwaters.com/amos

- Autobox: Automatic Forecasting Systems, Inc.
  http://darkstar.icdc.com// autobox
- CrossGraphs: Belmont Research Inc.
  http://www.belmont.com
- DataDesk: Data Description, Inc.
  http://www.lightlink.com/datadesk
- Epi Info: manuals and download sites
  http://mkn.co.uk/help/extra/people/Brixton_Books
  http://www.univ-lille2.fr/epiweb/epif/ en français For downloading Epi Info, Epi Map and SSS, see http://www.cdc.gov/epo/epi/epi.html
- EQS: Multivariate Software, Inc.
  http://www.mvsoft.com/
- Eviews, MicroTSP: Quantitative Micro Software
  http://www.eviews.com/
- GAUSS: Aptech Systems
  http://www.aptech.com
- Genstat, GLIM: NAG, Inc.
  http://www.nag.com
- GraphPad: GraphPad Software, Inc.
  http://www.graphpad.com
- IMSL, PV-WAVE: Visual Numerics, Inc.
  http://www.vni.com
- LIMDEP: Econometric Software, Inc.
  http://econwpa.wustl.edu/limdep/limdep.html
- Lisp-Stat: Luke Tierney, Univ. of Minnesota
  http://www.stat.umn.edu// luke/xls/xlsinfo/xlsinfo.html
- Maple: Waterloo Maple, Inc.
  http://www.maplesoft.com
- Mathematica: Wolfram Research, Inc.
  http://www.wri.com
- MATLAB: The Mathworks, Inc.
  http://www.mathworks.com
- MedCalc: Dept. Internal Medicine, University of Ghent
  http://allserv.rug.ac.be/f̃schoonj

- Minitab: Minitab Inc.
  http://www.minitab.com
- MLAB: Civilized Software, Inc.
  http://www.civilized.com
- MLn: Economic & Social Research Council
  http://www.ioe.ac.uk/multilevel
  or http://www.medent.umontreal.ca/multilevel/ mirror site for North America
- NCSS: NCSS Statistical software
  http://www.ncss.com
- Prophet: National Center for Research Resources
  http://www-prophet.bbn.com
- ProStat: Poly Software International
  http://www.polysoftware.com
- RATS: Estima
  http://www.estima.com
- Resampling Stats: Resampling Stats
  http://www.statistics.com/Welcome.html
- SAS, JMP: SAS Institute Inc.
  http://www.sas.com
- SC: Mole Software
  http://www.shsj.ulst.ac.uk
- Shazam: University of British Columbia, Canada
  http://shazam.econ.ubc.ca
- SigmaStat: Jandel Corporation
  http://www.jandel.com
- SIMSTAT: Provalis Research
  http://ourworld.compuserve.com/homepages/simstat
- SpaceStat: Regional Research Institute, West Virginia University
  http://www.rri.wvu.edu/spacestat.htm
- S-PLUS: MathSoft, Inc.
  http://www.mathsoft.com
- SPSS, SYSTAT, BMDP: SPSS Inc.
  http://www.spss.com

- Stata: Stata Corporation
  http://www.stata.com
- STATGRAPHICS: Manugistics, Inc.
  http://www.manugistics.com/statgraphics
- Statistica: StatSoft, Inc.
  http://www.statsoftinc.com
- Statit: Statware, Inc.
  http://ftp.statware.com// statware
- StatLib: an archive for distributing statistical software
  http://lib.stat.cmu.edu
- StatMost: DataMost Corporation
  http://www.datamost.com
- StatView: Abacus Concepts, Inc.
  http://www.abacus.com
- STPLAN: MD Anderson Cancer Center
  http://odin.mdacc.tmc.edu/anonftp/
- SUDAAN: Research Triangle Institute
  http://www.rti.org
- TSP: TSP International
  http://www.tspintl.com/
- UNISTAT: UNISTAT Ltd.
  http://www.unistat.com/
- ViSta: Prof. Forrest Young, Psychometrics, U. of North Carolina
  http://forrest.psych.unc.edu/research/ViSta.html
- WesVarPC: Westat Inc.
  http://www.westat.com

  其它URL地址：
  http://netlib.att.com (AT&T netlib)
  http://www.macsyma.com (macsyma)
  http://nr.harvard.edu (Numerical Recipes)

表格目录

插图目录