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;
```