第六章　BMDP

§6.1　概要

§6.1.1　简介

BMDP 由美国加州大学研制，为生物医学领域的数据分析而设计的经典统计分析软件包，用于各种计算机系统。在PC、VAX 和UNIX 系统均有菜单驱动的界面，但它们都有一个MENTOR，帮助用户自动产生所需的程序；一个编辑器，用于产生或修改命令文件；数据输入和产出功能。部分版本还有特定的在线帮助、高分辩图形程序、数据录入的电子报表。

BMDP的程序用两个字符标识，统计功能分为八类：

D 系列数据描述系列，如数据描述、t-检验、缺失值处理及简单图表。

F 系列列联表分析，如对数线性模型分析。

L 系列寿命表和生存分析，如Cox 回归分析。

M 系列多元分析，如聚类分析、因子分析、典型相半分析及逐步判别分析等。

R 系列回归分析，如多元回归分析、逐步回归分析、多项式回归、非线性回归及logistic 回归等。

S 系列非参数统计。

T 系列时间序列分析。包括一维与二维谱分析及Box-Jenkins 模型。

V 系列各种方差分析过程。

§6.1.2　运行

有批处理和交互两种运行方式，前者运行预先编好的程序，后者是在程序开始后，通过系统提示而运行。不带参数直接打入命令BMDP 将进入菜单控制下的运行方式，设软件安装于D:\BMDP>，先用SET设置运行环境(如SET DNEWS=D:\BMDP)，则命令：

D:\BMDP> BMDP <Enter>

启动交互式操作。BMDP 当光带落在相应功能上时，打<Enter> 则进入相应的选择，以<ESC> 返回上一级菜单或由主菜单返回DOS 操作系统。若选择RUN，系统首先提示运行的模块名，然后询问是交互式还是批处理方式。批处理的输入输出文件名隐含以.INP 和.OUT 作为文件名。

每个BMDP 程序都是为特定类型的统计问题而设计的，所有BMDP 的程序都使用BMDP.EXE 文件运行，例：

D:\BMDP> BMDP xx <Enter>

D:\BMDP> BMDP xx OUT=文件名<Enter>

D:\BMDP> BMDP xx IN=文件名<Enter>

D:\BMDP> BMDP xx IN=文件1 OUT=文件2 <Enter>

第一行表示交互方式启动程序"xx" (1D, 2D, 等)。第二行表示交互方式启动程序"xx"，所有输出存入文件中，同时在屏幕上显示运行结果。第三行表示批处理方式启动程序"xx" 从文件中读取指令。第四行表示以批处理方式启动程序"xx"，从文件1 中读取BMDP 指令，结果存入文件2。

表 6.1 收入情况与工作满意的程度

收入	满意度(satisf)				
(income)	很不满意	不满意	一般	很满意	小计
<6	20	24	80	82	206
6-15	22	38	104	125	289
15-25	13	28	81	113	235
>25	7	18	54	92	171
小计	62	108	319	412	901

批处理和交互式处理的不同点：1. 错误处理。BMDP 指令出错时，批处理将返回DOS，而交互式回到编辑态。2. 运行：交互式运行时，一些逐步过程允许覆盖每步的变量筛选。3. 高分辨率绘图仅在交互式有效。字符图形及图形文件仅在批处理方式有效。4. 指令执行：在交互式中，以下程序在读入数据后可以解释方式运行。3D (t 检验), 1T (谱分析), 2T (Box-Jenkins 时序分析), 4V and 5V (方差分析)。DM 是一个特例。它总是一步一步地执行。

§6.1.3 BMDP 有关概念和编辑工具

§6.1.3.1 BMDP 文件

命令文件包含BMDP的指令，用其它软件或BMDP 的全屏幕编辑产生。数据文件，含有待分析的数据。数据也可以列在指令文件END的后面，对于较大的问题，指令和数据最好分开。所有程序输出、分析结果输出至屏幕(CON:)。你可在BMDP中使用DOS的再定向或屏幕打印命令(”CTRL/PrtSc”)。屏幕输出可以写入列表文件，屏幕菜单和高分辨图形除外。系统存贮文件是BMDP产生和使用的一种二进制文件。PLOTFILE 是能由BMDPLOT使用的图形文件。

BMDP最多能使用3个暂存中间文件，多数情况下只用1个，用户可以不去考虑它，在磁盘空间不够或分析一个大的数据集时会碰到问题。

BMDP 也使用一个文件，它在系统安装了BMDP后产生。全屏幕编辑把指令拷贝到文件SCRATCH.EDT，而文件BMDP.LOG包括运行程序的所有指令。

§6.1.3.2 命令文件

㈠文件格式

【例6.1】下表是一个社会调查的例子[2]，变量是收入、满意度。收入单位为千美元，满意度分级为：很不满意、不满意、一般、很满意。

分析程序4F.INP 的内容如下：

```
/input variables are 2.
       format is free.
       table is 4,4.
/variable names are satisf,income.
/category names(1) are 'v. dissat','lit sat',
                       'mod sat','v. sat'.
          codes(1) are 1,2,3,4.
```

```
          names(2) are '<6','6-15','15-25','>25'.
          codes(2) are 1,2,3,4.
/tables columns are satisf.
       rows are income.
/statistics chisq.
           gamma.
/end
20 24 80  82 22 38 104 125
13 28 81 113  7 18  54  92
/end
```

BMDP 程序好象通常的英语文章，由段落(paragraph)和句子语组成(sentences)。尽管每个程序有多种选择，但BMDP 能用一些预先指定的值或默认值，因而BMDP 运行并不需要太复杂的指令。

各段用斜杠(/)引导。此处INPUT 段指定变量数目为2、自由格式数据、4x4 表格数据；VARIABLE段指示变量名为SATISF 和INCODE；CATEGORY 段指定各变量的分类代码；TABLES 段指示将构成的表样；STATISTICS 段指示计算χ^2和伽马系数。

句子也就是命令，用关键字开头，再用合适的语法规则选择IS, ARE 或= 之一连接一个项目表，以句号(.) 结束。一组相关的句子就构成一个段。每个段有一个名字(如：INPUT、VARIABLE、GROUP、END)并且以反斜杠(/) 分开。自由格式书写的程序也是允许的，但不提倡。

多数段落和句子可以简写，如INP表示INPUT，VAR表示VARIABLE；一个单句可以分成任意数目的行，每段内只放属于自己的句子；一般来说，段的次顺和句子的顺序是任意的，仅少数例外；使用井号(#)引导程序的注释，但不要把它放在引号中；出错的句子在警告信息中显示。

通常，BMDP 要求所有指令被分成段落，以EMD 段结束，程序把它们作为一个单元处理，而数据管理(DM) 则是一次一个段落，称为解释式执行，少数其它程序也这样执行，只是在给定读取数据的指令之后，如：3D (t 检验), 1T (谱分析), 2T (Box-Jenkins 时序分析), 4V及5V (方差分析)。

在交互式运行方式下，BMDP 用反斜杠(\) 结束每一段的指令，它必须是本行最后一个非空字符，END 段结束一个运行。

BMDP 使用TRANSFORM 段中的USE 选择分析用的记录。TRANSFORM 段有几个保留的名字，当每个记录都使用时，USE 的值设为+1，而KASE 是记录的顺序号，若设置USE 的值为零，则相关的记录在分析中不被使用，如：USE = AGE >= 30 AND AGE <= 50 仅使用AGE 从30 到50 的数据。另外，省略第17、23 及37可以使用：OMIT = 17, 23, 37. BMDP 使用VARIABLE 段中的USE 句子选用变量，省略时处理所有变量。

使用TRANSFORM 段转换和生成变量，如：

```
LOGAGE = LOG(AGE).
AVE = (V1+V2+V3+V4)/4.0.
AVE = MEAN(V1,V2,V3,V4).
NEWVAR = OLDVAR.
IF(OLDVAR > 0 AND KASE < 10) THEN NEWVAR = 1/OLDVAR.
```

数据管理程序(DM)提供了更复杂的数据处理功能。

建议：1.使用BMDP 的文件将简化BMDP 指令数目和改善系统运行速度，2.在VARIABLE段中使用USE，若文件较大，生成关于这些变量一个专用的BMDP 文件，由SAVE 段中的KEEP完成。3.读取有格式数据时，省略INPUT 段中的CASE，可用程序算得的记录做比较。

实际分析时，使用一个专门的文件存放数据，开始使用简单的描述分析了解数据的性态并且生成后续分析的BMDP系统文件，最后使用BMDP进一步分析。

数据文件可以是文本(ASCII)文件或特殊的BMDP二进制文件。BMDP 也能产生不同计算机系统间传输的格式，参见SAVE段的有关说明。BMDP 等待使用的文件是一矩形的格式，可以想象成一个表。每行是一个case，每列是一个变量。但不要把case、observation与record或文件中的一行相混淆。BMDP的一个case 允许包含任何数目的物理记录或行，反过来一行也可以有几个case 。数据格式有自由格式(FREE), STREAM, SLASH, BINARY 及固定格式”fixed.”。

星号(”*”) 表示缺失值，对于固定格式空格作为缺失值。AM 程序对于缺失值的处理很有用。

㈡段落选项

1. PROBlem TITLE=’text’. 标题，最长160字符。LENgth=#. 贮数组的最大长度。INTeractive. 交互式运行，默认时非交互式。ERRlev=STRICT或NORMAL. 指示BMDP检查程序的方法。

2. INPUT 段选项

TITLE=’标题’.	(选项) 程序输出标题。
FILE=’文件名’.	数据文件名。
CASES=#.	读入的记录数，省略时读至数据尾。
VARiables=#.	每记录的变量数。
FORMAT=FREE,STREAM,	自由格式、连续格式等数据格式
SLASH,BINARY.	自由格式由空格或逗号分隔，每个case 在新的记录
FORMAT=’指示’(fixed).	或行开始。STREAM中行的界
	限被取消。SLASH 同STREAM 一样，只是case
	之间用斜线(/) 分隔。固定格式是FORTRAN 格式
CODE=name.	”名字” 是长度1-8 的字母代码，
	CODE=’ ’ 表示读取文件中的第一个数据集。
CASE=#.	指定在许多问题中，本分析读入的例数。
REWIND.	重绕命令，用于磁盘或磁带的重读。
DIRectory.	打印BMDP文件的CODE, CONTENT,LABEL等信息。
CONTent=参数名.	指示BMDP文件的格式，如DATA,CORR,MEAN,FREQ。
LABEL=’标号’.	用于两个或多个BMDP文件读取时。
RECLEN=#.	记录长度，默认为80或72。
MCHAR=’字符’.	指示缺失，缺省时为’*’。
ERRMAX=#.	最大允许出错的记录数。
MULTiple=#.	允许一行读入多个有格式记录。
BLEVEL=#.	空格的处理方法，默认取值为2。

3. VARIABLE 段选项

NAMES=变量列表.	变量名表，省略时命名为X(1), X(2), ...。
LABEL=变量列表.	记录标号。
USE=变量列表.	使用的变量，参阅HELP 及SELECT的有关内容。
MINIMUM=#-列表.	
MAXIMUM=#-列表.	每个变量取值的上下界及缺失值。
MISSING=#-列表	其中的# 表示与变量表相应。
BLANK=ZERO|MISSing.	空格的处理方法。
ADD=NEW|#.	命名经转换的变量或数目。
BREore|AFTer|CHECK.	变量转换前后的检查，默认为BEFORE。
GROUPING=变量.	分组变量。
RETAIN.	经转换的变量保持上一个记录的值，可定义滞后变量。

4. GROUP 段选项

RESET. 删除所有分组信息。
CODE(变量表)=#-列表. 为变量定义有效码表(数据值)
CUTPoint(变量表)=#-列表. 为变量定义分组间隔。
NAME(V-List)=Name-list. 每个码值或范围的名字。
BMDP 使用CODE 或者CUTP 与NAME 进行变量分组，如：
CODE(STYLE) = 1, 2, 4, 8.
NAME(STYLE) = NONE, SOME, OK, AMAZING.
CUTP(AGE) = 25, 35, 45.
NAME(AGE) = KID, YOUNG, MIDDLE, OLD.

设定STYLE 的有效码值为1,2,4,8; 其它任何码将作为“坏码”处理。AGE 被分成四个区间：¡=25, ¿25 及¡=35, ¿35 及¡=45, 及¿45。名字与码或区间是相配的相同名字下给予码值将使用不同的组合并，如：”NAME(AGE)=OUT, LOW, HI, OUT.” 把小于25 及大于45 的两组合并单一的组，其标记为”OUT”。

有必要对GROUP 段与TRANSFORM 段的操作做一区分，前者并不改变数据的值，后者则不同，以下句子改变AGE 的值。

```
IF(AGE <= 25) THEN TEMP = 1.
IF(AGE > 25 AND AGE <= 35) THEN TEMP = 2.
IF(AGE > 35 AND AGE <= 45) THEN TEMP = 3.
IF(AGE > 44) THEN TEMP = 4.
AGE = TEMP.
```

5. TRANSFORM 段选项

句式为：变量= 表达式。
IF (关系) THEN 句子.
IF (关系) THEN (句子-1，句子-2 ...).
OMIT = #-列表.

DELETE = #列表.
USE = 表达式.
表达式：+, -, *, /, MOD, <, <=, >, >=, ==, <>, OR, AND, NOT
系统保留字
USE 记录使用变量：+1 使用记录；0 不用；
-1 从数据中删除记录；-100 停止数据输入.
KASE 正处理的记录数.
XMIS 内部缺失值标记.
TOOLARGE 数据超出范围的内部标志- 太大.
TOOSMALL 数据超出范围的内部标志- 太小.
函数
LOG、LN、SQRT、EXP、ABS、SIN、COS、TAN、ATAN、ASIN、ACOS、INT、SIGN、CHAR。
综合统计函数
N、NMIS、MIN、MAX、SUM、SUMC、MEAN、MED、SD、SEM、T等的含义与SAS
相同。
TRIM(i,a1,...,an) 关于第i个最大值和第i个最小值的截尾均值
TT(i,a1,...,an) 关于第i个水平截尾均值的t-值
IQR() 四分差
RHO(y1,...,yn) (1,...,n) 与(y1,...,yn) 的相关
B(x1,y1,...,xn,yn) y=a+bx 的直线回归系数B
A(x1,y1,...,xn,yn) y=a+bx 的直线回归截距A
B(x1,y1,...,xn,yn) (x,y) 的直线相关系数R
TRND(y1,y2,...,yn) 是(y1,...,yn) 在(1,...,n) 上的趋势
TCON(y1,y2,...,yn) 是(y1,...,yn) 在(1,...,n) 上的回归截距
AREA(y1,...,yn) y 下的面积
TRAP(x1,y1,...,xn,yn) y 下的面积，采用梯形法则
LINT(x0,x1,y1,x2,y2) 线性插值y1+(x0-x1)*(y2-y1)/(x2-x1)
LIND() 最末一个可用值的位置
LVAL() 最末一个可用值
INDEX(a,b1,...,bn) 第一个等于a 的b 的位置
REC(a,b1,c1,...,bn,cn) 换码函数，当bj的值等于a时存于cj
日期函数
DAYS(mm,dd,yy) 1960年1月1日起的天数
DAYS(mmddyy) 1960年1月1日起的天数
MDY(#of days) 六位日期作为一个变量
MM(#of days) 月份
DD(#of days) 日期
YY(#of days) 年份
JULN(#of days) 五位数西历(Julian date)
DAYJ(Julian) 1960年1月以来的天数
命令
REPL(y1,...,yn) 使用这些值的线性插值替换缺失值

FILL(x1,y1,...,xn,yn) 用周围的(x,y) 变量对替换缺失的y值
TEXT('message') 打印记录号和标号的讯息
SHOW() 使用a1,...,an的值打印记录号和标号

6. SAVE 段选项

存贮段产生一个数据文件，所有读入或TRANSFORM 产生的变量可被存贮，许多程序产生的特殊变量如回归中的预测变量及残差也可以存贮。生成的文件亦有BMDP系统文件及ASCII 文件两种。两种文件都需要指示：

FILE='文件名'.	文件名
KEEP=变量表.	(可选) 要存贮的变量，隐含为自动行成的任何量
DELete=变量表.	不保存的变量列表
CODE=名字.	长度为1-8 的数据集标识名
COMPlete.	仅保存那些不缺失的记录
LABEL='说明'.	(可选) 至多为40 个字符组成的数据集说明
CONTENT=名称.	(可选)
NEW.	生成一个新的系统文件；若文件已存在，则数据集被追加到文件中，(即一个BMDP 文件可包括几个数据集，经其CODE 来区分)
PORT.	BMDP 传输文件，由其他计算机系统如VAX/VMS 使用
FORMAT=F.	FORTRAN 10F8.3 浮点格式
FORMAT=G.	FORTRAN 5G16.6 的可变格式
FORMAT='specifier'.	FORTRAN 的格式指示
MISSING=标志表.	(可选) 对内部缺失数据标志重编码为指定的值

7. PRINT 段选项

多数程序具有专门的打印段选项；以下的说明适于所有程序。

NEWS.	打印程序信息(同NEWS 菜单选项)。
LEVEL=xxx.	MINIMAL, BRIEF, NORMAL, VERBOSE 控制程序输出的量
LINESIZE=#.	程序输出宽度80 或132
PAGESIZE=#.	每项行数，指定PAGE=0则不换页，默认为59
VNAME.	指定NO VNAME可以省略输入变量有关信息的打印
GNAME.	指定NO GNAME可以省略有关分类信息的打印
VUSE.	指定NO VUSE.可以省略/VARIABLE USE表的信息
DEBUG=xxx.	NONE,TEST,INFO,ALL指定调试的方法

§6.1.4 BMDP 模块用例

BMDP 的统计分析基本功能与SAS 及SPSS 相仿，现对它的列联表分析、生存分析和回归分析有关内容略作介绍，使用的有关记号如下，注意数据输入等段落总是必要的。

/ —指示段落	# —数字
. —句子结束	'c' —字符
—对于特定过程为必选项	VT —转换后的变量数
v —变量(名称或下标)	list —多于一个项目的列表
g —分组(名称或下标)	大写字母表示BMDP 的关键字

§6.2 4F、1L、2L

4F 语句格式：

```
/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/CATEGORY
/TABLE       ————┐
/PRINT           |   对子问题进行重复
/STATISTICS      |
/FIT         ————┘
/END
```

/INPUT 有选项TABLE=# list指示多维列联表的各个维数，CONTent= DATA 或TABLE 指定从BMDP文件中读入数据或表格，隐含是DATA类型。若使用TABLE，而且该表不是文件中的第一个，LABEL应指明表的顺序号，如：LABEL IS TABLE2。

/TABLE 是必选的，对两维表或多维表使用ROW=v list，COLumn=v list. 定义行列分类；对于多维列联表应使用INDices=v list或CATvar=v list. 若变量的分类的数目超过10个，则应在CATEGORY段指示CODES或CUTPOINTS。出现多个CATvar 时，第一个CATvar变化较第二个快，依次类推。PAIR 或CROSS 指示行列是配对或交叉。CONDition=v list. 指示对每个条件变量的每水平形成一个表。STACK=list. 作为单一指标，包含变量的所有可能的组合。COUNT=v list. 当输入为格子指标及频数时，指示频数若多于一个，则对每个变量分别进行。DELTA=#. 分析前每个格子加上的数，默认为 0 。EMPTY(#)=# list. 指示作为结构零处理的格子指标。INIT(#)=# list. 拟合矩阵的元素，默认值皆为1，当指示为零时即为结构零。SYMBols=c list. 指定模型识别各分类指标的符号，默认值为各变量的第一个字符。

/PRINT OBServed 打印观察频数表，除非指示NO OBS。EXCluded 删除某值的表，LIST=# 列出未进入分析的记录，LAMBda 对数线性模型参数估计值与其标准误，VARiance 对数线性模型参数间的方差协方差阵，PRECent=NONE,ROW,COL, TOT 即百分比，EXPected 为期望值，STANdardized. 打印标准化偏差。DIFFerence 指示观察值与期望值的差，FReeman为Freeman-Tukey 偏差统计量，CHISQuare. 对每个两维表或模型打印Pearson χ^2统计量，LRCHI 对每个两维列联表或模型打印似然比统计量。ADJusted 打印调整标准化偏差，MARGinal=# 打印# 阶的边缘合计表。LAMBda. 打印对数线性模型的估计参数，BETA. 打印相乘参数的估计即估计值取自然指数，VARIance. 打印参数的相关和协方差阵。BAR='c'. 指示两维或多维表中的竖分界线。

/STATISTICS 有CHISQuare 即列联表χ^2，CONTingency 即列联表统计量，LRCHI 即似然比χ^2，FISHer 即Fisher 精确概率(2x2表)，TETRAchoric 即四格表相关，CORRelation 积矩相

表 6.2 美国Florida 州1976-77 年度死刑的数据

Defendent	受害者的	死刑(P)		小计
种族(D)	种族(V)	是	否	
白人	白人	19	132	151
	黑人	0	9	9
黑人	白人	11	52	63
	黑人	6	97	103

关，SPEARman 即Spearman相关，GAMma 即Γ、Kendall, Sommer's D 等统计量，LAMBda 即λ，TAUS 即Goodman 和Kruskal τ，UNCertainty 是不确定系数，MCNemar 即McNemar 对称检验、kappa 可靠性检验，LINear 即2xC 或Rx2 表，进行趋势性检验，ALL 打印所有统计量，NO ALL 选项可以由其它选项覆盖，MINmum=# 指示表中期望值小于该值时则行合并。

/FIT ALL 对两维或三维表拟合所有的层次模型。SIMULtaneous 打印给定阶数所有效应的同时检验。MODEL=list 如：MODEL=vp,dp. 没有指定符号时使用名字的第一个字母。ALL. 拟合两维或三维表所有层次模型。ADD=SIMPle,MULT. 从特定的模型开始逐步法拟合模型，每步增加一个或多个效应，DELete=SIMPle, MULT. 逐步法每步剔除一个或多个效应。STEP=# 指示从用户指定的模型增加或剔除的最大步数或在逐步法中可以认为是极端值的格子数。INCLude=list. 拟合模型时必定包括的效应，用符号表示。CELL=NO,STAN,FR 指示在逐法中使用最大标准化偏差或最大FREEMAN偏差。PROBability=#. 决定模型拟合显著程度的概率水准，省略时为0. 05。STRATA= all 或list 依次删除每个指标的层，对仅有两层的变量无效。CONVERGE=#,#. 指示允许的最大绝对误差和偏差，省略时默认为0.01和0.00001。ITERation=# 指示最大迭代次数。

/SAVE CONTent=DATA 或/和TABLE 存贮频数表存于BMDP文件。

【例6.2】下面是Agrest, A.(1990) 分析Radelet(1981) 的数据，见表6.2。

```
/PROBLEM    TITLE IS 'Death Penalty'.
/INPUT      VARIATES ARE 3.
            FORMAT IS free.
            TABLE IS 2,2,2.
/VARIABLE   NAMES ARE penalty,victim,defendent.
/TABLE      INDICES ARE penalty,victim,defendent.
            SYMBOLS ARE p,v,d.
/CATEGORIES CODES(1) ARE 1,2.
            NAMES(1) ARE yes,no.
            CODES(2) ARE 1,2.
            NAMES(2) ARE white,black.
            CODES(3) ARE 1,2.
            NAMES(3) ARE white,black.
 /PRINT     MARGINAL IS 2.
            EXPECTED.
            STANDARDIZED.
```

```
          LAMBDA.
/FIT      ALL.
/FIT      MODEL IS vd,p.
/FIT      MODEL IS vd,vp.
/FIT      MODEL IS vd,vp,dp.
/FIT      MODEL IS pvd.
/END
```

标题是'Death Penalty'，读入变量数为3，转换增加的变量是0，变量总数为3。使用的变量是penalty，victim，defenden，格式是自由格式。表的形式是penalty (p) x victim (v) x defenden (d)。检验所有的模型，最大迭代次数为20，收敛准则是0.01000，0.000010000，显著水平为0.0500。

结果包括观测值、排除的值、期望值、对数线性模型估计参数及标准化偏差、二阶边缘表。

所有模型的结果。

MODEL	DF	LIKELIHOOD-RATIO CHISQ	PROB.	PEARSON CHISQ	PROB.	ITERATIONS
p.	6	170.50	0.0000	140.08	0.0000	1
v.	6	363.46	0.0000	399.45	0.0000	1
d.	6	395.80	0.0000	416.31	0.0000	1
p,v.	5	138.04	0.0000	122.58	0.0000	1
v,d.	5	363.35	0.0000	398.68	0.0000	1
d,p.	5	170.39	0.0000	142.02	0.0000	1
p,v,d.	4	137.93	0.0000	122.40	0.0000	1
pv.	4	131.79	0.0000	115.97	0.0000	1
pd.	4	170.16	0.0000	141.36	0.0000	1
vd.	4	233.55	0.0000	200.64	0.0000	1
p,vd.	3	8.13	0.0434	6.98	0.0726	1
v,pd.	3	137.71	0.0000	121.32	0.0000	1
d,pv.	3	131.68	0.0000	115.90	0.0000	1
pv,pd.	2	131.46	0.0000	115.75	0.0000	1
pd,vd.	2	7.91	0.0192	7.04	0.0296	1
vd,pv.	2	1.88	0.3903	1.43	0.4889	1
pv,pd,vd.	1	0.70	0.4025	0.38	0.5402	6

故模型(VD,P)、(VD,PV)与(PV,PD,VD)值得特别注意，从最后三个模型可见PD 没有什么影响，而PV的影响很大而以模型(VD,PV)较佳。模型(VD,P) 估计参数如下，其标准误估计使用 δ 方法直接估计，对数线性模型估计THETA(MEAN)= 2.8379。

下表是模型(VD,PV) 的期望值，括号内为标准化偏差，即：(观察值- 期望值)/$\sqrt{\text{期望值}}$。

defenden	victim	penalty

		yes	no	TOTAL
white	white	21.2	129.8	151.0
		(-0.5)	(0.2)	
	black	0.5	8.5	9.0
		(-0.7)	(0.2)	
	TOTAL	21.7	138.3	160.0
black	white	8.8	54.2	63.0
		(0.7)	(-0.3)	
	black	5.5	97.5	103.0
		(0.2)	(-0.0)	
	TOTAL	14.3	151.7	166.0

对数线性模型参数为：THETA(MEAN) = 2.7237，λ 的估计值如下，括号内数据为估计值与标准误的比值。

penalty		victim		defenden	
yes	no	white	black	white	black
-1.171	1.171	0.799	-0.799	-0.391	0.391
(-10.108)	(10.108)	(5.796)	(-5.796)	(-4.130)	(4.130)

victim	penalty		defenden	victim	
	yes	no		white	black
white	0.264	-0.264	white	0.828	-0.828
	(2.282)	(-2.282)		(8.748)	(-8.748)
black	-0.264	0.264	black	-0.828	0.828
	(-2.282)	(2.282)		(-8.748)	(8.748)

$\lambda_{vp} = 0.264$表示受害者是白人时，所接受的处罚要重些，$\exp(4 * \lambda_{vp})$是D各水平上的估计比数比。

模型(VD,PV,PD) 参数标准误用信息矩阵的逆而获。参数估计THETA(MEAN) = 2.6922，其它结果从略。

表 6.3 两种实验条件下的生存时间(Gehan白血病数据)

病人分组	生存时间(有加号为截尾)
6-MP	6+,6,6,6,7,9+,10+,10,11+,13,16,17+,19+,20+,22,23
	25+,32+,32+,34+,35+
Control	1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23

1L 语句格式：

/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/FORM
/GROUP
/ESTIMATE ┐ 对子问题重复 ┘
/PRINT
/END

/FORM 必选，指定生存变量的结构。共有三类即时间型、日期型和寿命表型。UNIT=c. c可以是DAY,WEEK,MONTH,YEAR；省略时使用MONTH。对时间型有TIME= v. 指示每记录的时间变量名或下标；STATus=v. 用作RESPonse 或LOSS 的示性变量，当所用的代码不出现时为截尾；RESPonse=# list，指示反应的代码，如RESP=1,6 表示1,6为失效，省略时取最小的代码；LOSS=# list 指示截尾的取值如LOSS= 2 TO 5。对日期型，输入包括ENTRY=v,v,v. 包含月、日、年的三个变量名或下标。TERMination=v,v,v. 指示截止的月、日、年变量，省略时系统处理为缺失。STATus=v.、RESPonse=# list.、LOSS=# list.与时间型类似，CENSOR=#,#,#. 用于研究或分析结束时的月、日、年标志，仅用于截尾观察的截止日期不清时。对寿命表类型的输入，每个记录包括了一个时间区间内的事件，记录应按时间排序。NENTer=v. 进入区间的数目。NDEAD=v. 区间内死亡数。NLOST=v. 区间内丢失数。NWITHdrawn=v. 区间退出数。INTERval=v. 每区间的下限。

/ESTimate 用于指示分析的方法、统计量和图示方法，为可选项并且可以重复。METHod=c. c 可以是LIFE或PROD，对于METHOD=LIFE.，PERIOD=# 是期望相等的区间数，WIDTH=#. 是时间间隔宽度，CUTPoint=# list. 划分间隔的时间点。对于METHOD=PROD，VARiable=v list. 指示与PL 估计一起打印的变量名表；PRINT. 打印生存分布的估计。PLOT=c list. 打印图示，包括SURV,LOG,CUM,HAZ,DEN 对应不同的生存估计。SIZE=#,#. 即横轴与纵轴的宽度与高度。GROUPing=v. 分组变量名。STATistic=MANTEL或BRESLOW 是可选项，指示生存曲线相等的检验。

【例6.3】生存分析。BMDP1L 用于寿命表分析，下面是Gehan [3] 有关白血病的数据，第一组用六腺嘌呤(6-MP) 和对照组(control) 的情况，数据列于下表：

相应的程序如下：

```
/INPUT     TITLE IS 'Kaplan-Meier test example'.
```

```
            VARIABLE=3.
            FORMAT=STREAM.
 /VARIABLE  NAMES=group,time,indica.
 /FORM      TIME=time.
            UNIT=weeks.
            STATUS=indica.
            RESPONSE=0.
 /GROUP     CODES(indica)=1,0.
            NAMES(group)='6-mp','control'.
 /ESTIMATE  METHOD=life.
            GROUPING=group.
            STATISTICS=Mantel,BR,TAR.
            PRINT.
 /END
  1 6   1  1 17  1    2  1  0  2  8  0
  1 6   0  1 19  1    2  1  0  2  8  0
  1 6   0  1 20  1    2  2  0  2  11 0
  1 6   0  1 22  0    2  2  0  2  11 0
  1 7   0  1 23  0    2  3  0  2  12 0
  1 9   1  1 25  1    2  4  0  2  12 0
  1 10  1  1 32  1    2  4  0  2  15 0
  1 10  0  1 32  1    2  5  0  2  17 0
  1 11  1  1 34  1    2  5  0  2  22 0
  1 13  0  1 35  1    2  8  0  2  23 0
  1 16  0             2  8  0
  /END
```

反应编码：0 (DEAD)，截尾编码：1(CENSORED)

检验统计量：

	STATISTIC	D.F.	P-VALUE
GENERALIZED SAVAGE (MANTEL-COX)	16.793	1	0.0000
TARONE-WARE	15.124	1	0.0001
GENERALIZED WILCOXON (BRESLOW)	13.458	1	0.0002

详细的寿命表，共有十个间隔，风险函数在每个时间间隔的中点计算 λ i= 2qi/[hi(1+pi)]，死亡密度函数为$p_iq_i/h_i, h_i$是i个区间的宽度。

表 6.4 与生存时间有关的描述统计量

	分位点	估计值	标准误
处理组	75TH	10.40	4.05
	MEDIAN (50TH)	23.40	3.13
对照组	75TH	3.72	1.57
	MEDIAN (50TH)	8.31	2.00
	25TH	12.91	1.92

6-MP 组

区间 weeks 等于 小于	进入	退出	失访	死亡	暴露	死亡比例	生存比例	区间开始时的累积生存频率	风险函数	死亡密度
0.00- 3.50	21	0	0	0	21.0	0.0000	1.0000	1.0000	0.0000	0.0000
3.50- 7.00	21	1	0	3	20.5	0.1463	0.8537	1.0000	0.0451	0.0418
7.00-10.50	17	2	0	2	16.0	0.1250	0.8750	0.8537	0.0381	0.0305
10.50-14.00	13	1	0	1	12.5	0.0800	0.9200	0.7470	0.0238	0.0171
14.00-17.50	11	1	0	1	10.5	0.0952	0.9048	0.6872	0.0286	0.0187
17.50-21.00	9	2	0	0	8.0	0.0000	1.0000	0.6217	0.0000	0.0000
21.00-24.50	7	0	0	2	7.0	0.2857	0.7143	0.6217	0.0952	0.0508
24.50-28.00	5	1	0	0	4.5	0.0000	1.0000	0.4441	0.0000	0.0000
28.00-31.50	4	0	0	0	4.0	0.0000	1.0000	0.4441	0.0000	0.0000
31.50-35.00	4	4	0	0	2.0	0.0000	1.0000	0.4441	0.0000	0.0000
对照组										
0.00- 3.50	21	0	0	5	21.0	0.2381	0.7619	1.0000	0.0772	0.0680
3.50- 7.00	16	0	0	4	16.0	0.2500	0.7500	0.7619	0.0816	0.0544
7.00-10.50	12	0	0	4	12.0	0.3333	0.6667	0.5714	0.1143	0.0544
10.50-14.00	8	0	0	4	8.0	0.5000	0.5000	0.3810	0.1905	0.0544
14.00-17.50	4	0	0	2	4.0	0.5000	0.5000	0.1905	0.1905	0.0272
17.50-21.00	2	0	0	0	2.0	0.0000	1.0000	0.0952	0.0000	0.0000
21.00-24.50	2	0	0	2	2.0	1.0000	0.0000	0.0952	0.5714	0.0272

2L 语句格式

/PROBLEM
/INPUT
/VARIABLE
/TRANSFORM
/SAVE
/GROUP ┐
/PRINT │
/FORM │
/REGRESS │ 对子问题重复
/FUNCTION │
/TEST │
/PLOT │
/END ┘

在重复的子问题前应指定新的数据。FUNCTION 段对于时间协变量的情形是必需的，它可以使用TRANSFORM 段的句子，时间协变量应赋一个值，在FUNCTION 中只使用在REGRESSION 段句子COVARIATE, ADD 或AUXILIARY 中出现的变量名，TIME 是本段的保留字。

/FORM 与1L相仿，有时间型和日期型两种输入。

/REGRESS 指示回归模型，如COVAriates=v list. 指示固定协变量名称和下标。STATA=v. 指示分层变量名或下标。与变量筛选有关的选项有：STEPwise= MPLR 或PHH 指示最大偏似然比检验或Peduzzi-Hardi-Holford 统计量；REMOVE= #. 及ENTER=#. 指示逐步筛选变量p值的大小；START=IN,OUT 指示一个协变量在第一步是否在回归方程中；MOVE=# list. 指示一个协变量最多能被删除的次数，省咯为两次。RISK=LOGLINear, LINEAR,COMBINATION,USER 指示风险函数的形式为对数线性、线性、组合型及自定义，指定了LOGLIN以后对于每个固定协变量减去均值；与牛顿—拉弗森算法有关的选项有：CONVergence=#.与ITERation=#.指定收敛准则和迭代次数，默认值分别为0.00001和15；HALVing=#.指示每步最多使用的两分法次数，默认为5；TOLerance=# 矩阵求逆的容许限值，默认值为0.00001；INITial =# list. 指示对应于COVARIATES变量的初值。时间协变量的选项有：ADD=v list. 指示FUNCTION段中的时变协变量名；AUXIlliary=v list. 定义FUNCTION 中的时变协变量；PASS=# 指示时变协变量的层数。

/TEST 段是可选的，当变量筛选时忽略。ELIMinate=v list. 指示待检验的协变量回归系数；STATistics=c list. 指定计算WALD、LRATIO或SCORE 统计量，默认为WALD。

/PLOT 是可选的，指示STATA时对每层进行。TYPE=c list.对SURV,LOG或FIT 绘图。PATtern=# list. 定义生存函数的协变量值；SIZE=#,# 是横轴与纵轴字符的数目，默认值为100和50。

/PRINT CASE=#. 打印转换后原始数据的数目，默认为10。SURVIVAL. 打印排序的生存时间、Kaplan-Meier估计量、风险值、生存函数。CORRelation. 打印近似相关。COVAriance. 打印近似协方差。ITERations. 打印每步Newton-Raphson迭代的对数似然及参数估计值。

【例6.4】比例模式的检验。是1983年版BMDP引用Pike关于两组鼠接触某种致癌物数据进行生存分析的例子，拟合模型是$h_0(t)\exp(\beta_1 z_1+\beta_2 z_2(t))$，其中有一个时间协变量，可详见Kalbfleisch and Prentice(1980)，程序如下：

```
problem   title is 'checking the proportionality assumption'./
input     variables are 3. format is stream./
```

```
variable  names are survival, followup,group./
form      time=survival. status=followup. response=1. /
print     covariance./
regress   covariate=group. add=z2./
function  z2=group*(ln(time)-5.4)./
end
143 1 0  156 1 1  220 1 0  239 1 1
164 1 0  163 1 1  227 1 0  240 1 1
188 1 0  198 1 1  230 1 0  261 1 1
188 1 0  205 1 1  234 1 0  280 1 1
190 1 0  232 1 1  246 1 0  280 1 1
192 1 0  232 1 1  265 1 0  296 1 1
206 1 0  233 1 1  304 1 0  296 1 1
209 1 0  233 1 1  216 0 0  323 1 1
213 1 0  233 1 1  244 0 0  204 0 1
216 1 0  233 1 1  142 1 1  344 0 1
/end
```

共有36名失效，4名截尾，截尾占10%。自由变量名称及编码：3 group，4 z2。对数似然比：LOG LIKELIHOOD = -100.7113，χ^2：CHI-SQUARE = 3. 05，D.F.= 2，P-VALUE =0.2176。

参数估计结果如下，z2的回归系数相对于其标准误的值-0.1258很小，表明用一个固定协变量GROUP的模型不恰当。

VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
3 group	-0.5998	0.3484	-1.7216	0.5489
4 z2	-0.2295	1.8249	-0.1258	0.7949

渐近协方差阵：

```
ESTIMATED ASYMPTOTIC COVARIANCE MATRIX
---------------------------------------
                  group          z2
                      3             4
group      3         0.1214
z2         4         0.0563         3.3302
```

【例6.5】临床研究中，病人的预后可因为治疗过程中的一些事件而改变，我们可以把这些事件作为时间协变量引入。下面是著名的Stanford心脏移植数据，共有99例，数据放在文件HEART.DAT变量是生存天数(survival)，是否截尾(status)，等待移植的时间(waittime)，移植时的年龄(age)，以及排斥打分(mismatch)。当一个或多个协变量是生存时间的函数时，还需要更多的运行控制，下面程序也是上述BMDP 手册上的例子，固定协变量仍然用COVARIATE指示，时变协变量必须用REGRESSION段中的ADD指示，有必要用协变量以外的变量定义时变协变量时，应在AUXILIARY中引入。

```
problem   title is 'Heart Transplant Data with
          Time-dependent Covariates'./
input     variables are 6. file is '2L.DAT'. mult=4.
          format is '(4(F3.0,F5.0,F2.0,2F3.0,F5.2))'./
variable  names are id,survival,followup,waittime,age,mismatch.
          blanks are missing./
form      time=survival. status=followup. response=1. /
regress   add is xplant,xplntage,score.
          auxiliary=waittime,age,mismatch. /
function  xplant=0.0. xplntage=0.0. score=0.0.
          if (time GE waittime) then xplant=1.0.
          if (time GE waittime) then xplntage=age.
          if (time GE waittime) then score=mismatch./
print     cases are 99. covariance./
end
```

参与分析变量：1 id，2 survival，3 followup，4 waittime，5 age ，6 mismatch，格式为：(4(F3.0,F5.0,F2.0,2F3.0,F5.2))。时变协变量为：7 xplant，8 xplntage，9 score。失效为71，截尾28，占总例数的28.28%。自由变量：7 xplant，8 xplntage，9 score。

对数似然比：LOG LIKELIHOOD = -275.9557，χ^2 ：GLOBAL CHI- SQUARE=9.01，D.F.=3，P-VALUE =0.0291。

参数估计：

VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
7 xplant	-3.1780	1.1861	-2.6793	0.0417
8 xplntage	0.0552	0.0226	2.4423	1.0567
9 score	0.4442	0.2803	1.5851	1.5593

协方差阵：

		xplant	xplntage	score
		7	8	9
xplant	7	1.4069		
xplntage	8	-0.0246	0.0005	
score	9	-0.0870	-0.0003	0.0785

【例6.6】是Collett, D. (1991)[4] 的例子，资料是Smith, W.(1932) 关于一种保护血清对肺炎球菌的影响，第七天仍存活的鼠为生存，血清单位为cc 。利用LR 进行LOGIT 分析，程序如下：

```
/PROBLEM        TITLE = 'SERUM'.
/INPUT          VARIABLES = 3.
                FORMAT=FREE.
```

```
    /VARIABLE       NAMES = DOSE, Y, N.
    /TRANSFORMATION LOGDOSE=LN(DOSE).
    /REGRESS        COUNT=N.
                    SCOUNT=Y.
                    INTERVAL=LOGDOSE.
                    MODEL=LOGDOSE.
    /PRINT          CELLS=MODEL.
                    COVA.
    /END
    0.0028 35 40
    0.0056 21 40
    0.0112 9  40
    0.0225 6  40
    0.0450 1  40
```

结果：logit(p)=-9.19-1.83 log(dose)

ED50=(0.0054,0.0081)，LOG(ED50)=-5.021±1.96 x 0.1056

```
                        LOG LIKELIHOOD =   -87.062
GOODNESS OF FIT CHI-SQ   (2*O*LN(O/E))  =     2.809  D.F.=   3  P-VALUE= 0.422
GOODNESS OF FIT CHI-SQ (HOSMER-LEMESHOW)=     2.917  D.F.=   3  P-VALUE= 0.405
GOODNESS OF FIT CHI-SQ   ( C.C.BROWN )  =     1.871  D.F.=   2  P-VALUE= 0.392

                                      STANDARD
     TERM               COEFFICIENT       ERROR   COEFF/S.E.  EXP(COEFFICIENT)
LOGDOSE                   -1.8296      0.2545      -7.188      0.1605
CONSTANT                  -9.1894       1.255      -7.322      0.1021E-03

SUMMARY DESCRIPTION OF CELLS.
      CELLS ARE FORMED BY ALL COMBINATIONS OF VALUES OF VARIABLES IN THE MODEL.
-------------------------------------------------------------------------
                OBSERVED PREDICTED  S.E. OF  OBS-PRED  PRED.                    HAT
NUMBER NUMBER  PROPORTION PROB.OF  PREDICTED --------   LOG                   MATRIX
Y      FAILURE  Y         Y          PROB.   S.E.RES.  ODDS    CHI    DEVIANCE DIAGONAL INFLUENCE LOGDOSE
    1     39     0.0250   0.0289    0.0137   -0.1710  -3.5156 -0.1463  -0.1496   0.2686    0.011    -3.10
    6     34     0.1500   0.0956    0.0288    1.4926  -2.2474  1.1707   1.0901   0.3848    1.393    -3.79
    9     31     0.2250   0.2747    0.0423   -0.8797  -0.9710 -0.7039  -0.7186   0.3598    0.435    -4.49
   21     19     0.5250   0.5738    0.0501   -0.8116   0.2972 -0.6235  -0.6210   0.4098    0.457    -5.18
   35      5     0.8750   0.8271    0.0454    1.2314   1.5654  0.8008   0.8344   0.5771    2.069    -5.88

MINIMUM EXPECTED CELL FREQUENCY          =      1.15
NUMBER OF EXPECTED VALUES LESS THAN 5.0  =     2
```

与SAS 的程序进行比较:

```
    DATA SERUM1;
```

```
   INPUT DOSE Y N;
   LOGDOSE=LOG(DOSE);
CARDS;
PROC PROBIT;
   MODEL Y/N=LOGDOSE/D=LOGISTIC;
   OUTPUT OUT=SERUM2 PROB=PHAT;
PROC PRINT;
DATA SERUM3;
   SET SERUM2;
   YHAT=N*PHAT;
PROC PRINT;
```

§6.3 各系列模块功能概要

为了应用的方便，此处在上节的基础上，介绍一下BMDP各模块的功能。

§6.3.1 D 系列

1D 简单数据描述

对所有或部分记录提供常用的描述统计量，数据列表、排序等。

2D 详细的数据描述

计算许多描述统计量，并对每个变量绘直方图。2D 用于识别异常值，研究分布的形状，以及对样本数据初步描述。输出内容包括：均值，中位数，众数，标准差和均值和中位数的标准误，偏度与峰度，极值，Shapiro 与Wilks' W，截尾均值, Hampel 和biweight 估计量。

3D T 检验

提供三种不同的t 检验，结果输出包括直方图和描述统计量。TWOGROUP 是两组t 检验，方差等或不等，包括方差齐性Levene 检验及截尾t 检验(trimmed t)、非参Mann-Whitney 秩和检验、Hotelling's T-方及Mahalanobis D-方，同时给出每组内的变量相关值。MATCHED 进行配对t 检验，输出包括配对t 统计量和Pearson 相关系数，也能给出trimmed t，非参符号和Wilcoxon 符号秩次检验，Spearman 相关，Hotelling T-方和Mahalanobis D-方。ONEGROUP 提供单样本t 检验，类似于配对t 检验，不打印Spearman 相关。

4D 字符频率—数字的和非数字的

计算单列字段每个字符(数字、字母或符号) 的频率，产生的数据可以是原来数据列表或者用指定符号替换过的数据，这样易于找出不需要的符号或字母。当数据以固定格式排齐时，4D 可用于初步检查数据的类型，计量单列数据的频率并发现数据错误，所有数据均以A1 格式(宽度为一个字符) 读取，4D 不接受BMDP 文件输入。输出为一个频数表，显示每列中不同字符的频率。

5D 直方图和单变量图

多组数据可以画在同一图上，也可分组画在几个图上。图的大小、标度以及每个区间的名称都可以控制，5D 的直方图较其他程序如2D 详细，输出内容：直方图与累积直方图，

正态概率图，去趋势正态图，半正态图，图的标度和大小设定，用于数据分组的标号和符号，分组或未分组资料的描述统计量

6D 双变量(散点) 图

产生一个变量对另一个变量的散点图，并且计算最佳拟合直线。可把记录分组、使用符号区分不同的组，按组别绘制、控制图的大小和所用数据范围等。输出内容：参加绘图的点数，Pearson 相关系数及其p-值，均值和标准差，最小二乘回归曲线及其截距，剩余均方

7D 方差分析和数据筛检

7D 用于数据的筛检和方差分析，第一项功能有：分组变量不同水平上的复合直方图，分组及不分组描述统计量，分组方差相等的Levene 检验，选择方差稳定性转换的Box-Cox 图。第二项功能有：完全随机单方式和双方式方差分析，平衡或不平衡固定效应模型，对比和条件对比，多重比较(Bonferroni, Duncan,Dunnett,Newman-Keuls,Tukey, Scheffe 法)，分组方差不等时的稳健性检验(Welch 与Brown-Forsythe)，使用截尾均值的ANOVA。

8D 相关，不完全资料

当资料有缺失时，利用四种方法计算方差与协方差。ALLVALUE: 对每个变量所有的数据计算均值，使用均值的偏差用于计算方差和协方差。COVPAIR: 使用两个变量都可接受的值计算协方差，而使用每个变量所有可接受的值计算方差。CORPAIR: 仅对两个变量都可以接受的记录计算方差和协方差。COMPLETE: 仅使用完全的样本计算。输出包括：均值，标准差，方差，变异系数；变量对的频数表，权重和，均值和方差矩阵；相关的估计；与不完全资料有关的两两t 检验。

9D 分组的多方式描述

根据一个或多个分组变量计算每个类别的均值和描述统计量。9D 用于产生下述统计量的图示：两个或多个因素析因设计的格点均值，重复测量设计的均值，同时计算两个或多个变量的均值。本程序对评价各组的一致性，在方差分析中观察数据格子均值间存在的趋势和交互有用。其它的如：单向方差分析，卡方检验，各组间均值的变动情况，边缘合计图示(Plots of marginal subsets)。

AM 缺失资料的描述与估计

对于多变量资料描述缺失值的模式，并利用三种处理方法获得协方差阵或相关阵。把缺失值替换成均值或者关于被估变量和其它变量的回归。其它变量是与被估变量的相关最大的一个或者一组高度相关的变量，或者是所有其它变量。输出内容：单变量综合统计量及头五例数据列表，样本对变量的图示，显示缺失值和极值的位置或模式，相关阵和特征值，每变量与其它变量的复相关平方，回归显著性检验，对具有缺失值或超范围值估计的样本列表，估计变量与被估变量间的R-平方，样本到均值的Mahalanobis D-平方，完整样本与带有估计样本的图示。

§6.3.2 F 系列

4F 两维与多维频数表

使用4F 构造、分析及存贮两维、多维或多维表的分表的情况，使用4F 来产生两个或多个分类变量的log-linear 模型。分析内容有：两维表独立性检验：卡方，似然比，Fisher 精确

检验，列联表系数，phi, Cramer's V，Yule's Q 与Y, 交叉乘积比，Yates' 校正卡方，关联情况(Kendall's tau, Somers' D, 及其它)，预测情况(Goodman 与Kruskal's tau 及lambda, 不确定系数)，McNemar's 对称检验，kappa，比例的线性趋势检验；关于对数线性模型，4F 可以拟合：所有可能的模型(两维或三维)，所有饱和模型，每个交互的边缘和偏性关联检验，用户指定的模型，从用户指定的模型中增加或剃除效应；对于每个指定的模型，4F 提供：模型适合度检验，指定模型的预测频数，对数线性模型及其标准误，Freeman-Tukey 量，标化统计量，χ^2 统计量的组分。

§6.3.3 L 系列

1L 寿命表和生存函数

提供了两种估计生存率的方法，即寿命表法和积限方法。进行两组生存曲线的比较，将给定资料列成寿命表的形式。输出内容：Kaplan-Meier统计量，寿命表(Cutler-Ederer)，生存函数图，对数生存函数及累积风险函数，寿命表法的风险和密度函数图，检验生存曲线是否相同的Mantel-Cox, Breslow, 及Tarone-Ware 统计量。

2L 带协变量的生存分析—COX 模型

分析影响生存时间的其它测量变量，分析使用Cox 的比例风险回归模型，模型假设风险函数能用协变量的线性函来表达。量化生存时间与协变量之间的关系，估计回归系数从而给出每个变量对风险函数的相对作用，可以检验表示处理效应的回归系数显著性。这些回归系数以比例风险模型中病人的基线特征为条件。程序提供了逐步方法，处理分组资料。输出包括：每个协变量的回归系数、渐近标准误、标化回归系数；对于极大偏对数似然函数的卡方显著性检验；生存函数图、对数累积风险函数图。

§6.3.4 M 系列

1M 变量的聚类分析

提供了四种相似性的测量方法，三种聚类方法。输出内容：所有类一览表，每步形成的聚类树，聚类过程解释，影子相关阵(Shaded correlation matrix)，相关阵。

2M 记录的聚类分析

根据几种距离的测量方法，进行样本的聚类。输出内容：树形聚类图，amalgamation 距离、每变量均值和新类均值列表，影子相关矩阵，原始标度或标化后的数据列表，样本距离矩阵。

3M 分块聚类(BLOCK CLUSTERING)

对于分类资料形成类别的块，结果把数据矩阵排成分块矩阵。输出内容：识别子矩阵的分块记号，分块的符号表，计数及其值，每种码的频数。

4M 因子分析

四种据相关或协方差阵抽取因子的方法，几种旋转方法。输入数据可以是相关阵或协方差阵、因子载荷或因子得分系数。输出内容：单变量综合统计量，旋转和未旋转因子载荷及其图示，排序和旋转后的因子载荷，因子得分系数及其图示，原始数据的Mahalanobis 距离，因子得分和差值，复相关平方，特征值，排序和影子相关阵，标准得分，协方差阵，相关阵的逆，偏相关，剩余相关。

6M 典型相关分析

两组变量的典型相关，及Bartlett 关于剩余特征值的检验。输入可以是协方差阵或相关阵。输出内容：一元综合统计量和头五个样本数据列表，各变量与其它变量复相关平方，典型相关及相应的特征值和典型变量载荷，典型变量得分和系数，任意变量或典型变量对其它变量或典型变量的双变量图。

7M 逐步判别分析

对于两组或多组资料进行判别分析，可以交互式地每一步指示那一个变量进入或剃除。每组采用刀切法和交叉识别(jackknife-validation) 方法减少偏性。输出内容：每步F 统计量，Wilks' Lambda 或U 统计量(具有近似F 值)和马氏距离，分类函数，矩阵，刀切法分类，正确分类比例，分类一览表，每样本分到各个组时的后验概率及马氏距离，典型判别函数系数，特征值，每个样本的典型得分，头两个典型变量图。

8M BOOLEAN 因子分析

对于二分类资料估计布尔型因子，与传统的方法不同，它在矩阵相乘时采用逻辑算法，因而得分和因子载荷是二分类的。输出内容：每步上偏差为正(# 乘以观察得分为1，估计值为0) 和为负(# 乘以观察得分为0，估计值为1) 的数目、每次循环的总偏差，因子得分，数据矩阵和偏差的Compact 显示。

9M PREFERENCE PAIRS 资料的线性得分

计算每个观察的得分，对观测变量按其评价的重要性进行加权。输出内容：进入线性函数中的变量系数及其t 值，每个样本对的preference matrix, 原始评价值和预期值、误差，在每步结束时每个样本的得分，多种评判下的得分及其相关，变量或得分的散点图。

KM K-MEANS 记录聚类

使用欧氏距离来度量每个样本与每类中心的距离。输出内容：每类中各变量的描述统计量，直方图示类均值到类内和类外样本的距离，样本到三个最大类形成平面的正交投影散点图，类间均方与类内均方的方差分析及F 比，类的轮廓，合并类内协方差和相关阵，类中心距离，类和用户指定变量的交叉表。

§6.3.5 R 系列

1R 多元线性回归

对全部样本、部分样本或多组样本估计多元线性回归方程，检验各组回归线是否相等。输出内容：单变量综合统计量；复相关系数及估计标准误；回归方差分析表；回归系数及其标准误，t 值及标准偏回归系数；相关和协方差阵；残差，预测值及针对每个记录的统计量；残差的散点图、正态概率图及偏残差图。

2R 逐步回归

用逐步方法估计多元线性回归方程，每次进入或剃除的变量可以是一个或者一组，进行向前法或向后法筛选。有四种准则进行逐步筛选，强迫一些变量留在方程中。提供回归诊断功能。输出内容：每步上的：R-平方，调整R-平方, 及估计标准误；回归方差分析表；回归系数及其标准误标准回归系数，容许值，选入方程的F 值；偏回归，容许值，选入F 值(对尚未选入的变量)。同时有：各步一览表；回归系数表；数据、预测值、残差；回归诊断结果；编相关一览表，进入与剃除变量的F 值。

3R 非线性回归

非线性函数的最小二乘参数估计，六种函数及其导数是系统提供的，在FUN 段落中可以指示其它的函数。对于参数可以施加上下界约束和线性等式约束。使用迭代重加权最小二乘法(iteratively reweighted least squares) 获得极大似然解。输出内容：描述统计量；每步上的参数估计，剩余平方和，incremental halvings 数；渐近相关阵及参数标准误，残差的序列相关；每记录因变量的观测值和预测值，残差，权；散点图及正态概率图。

4R 主成分回归

关于因变量和一组主成分进行回归分析，主成分逐个引入，回归系数用主成分或原始或标化变量的形式报告。进入的次序是因变量与主成分的相关大小，能进行岭回归计算。输出内容：特征值、特征向量、累积方差贡献；主成分与因变量间的相关；主成分的回归系数；每步：进入的主成分，剩余平方和，F 比，R-平方，回归系数；每例的主成分值；散点图和正态概率图；在ridge 选项下：R-平方及其剩余平方和，每组岭因子的回归系数；岭迹，R-平方和残差平方和图示。

5R 多项式回归

对因变量拟合关于一个自变量的多项式，采用正交项式计算方法。每个样本可以有自已的权。输出内容：每个正交多项式的t 值、回归系数及其标准误，剩余均方自变量每个幂次的回归系数和剩余平方和，拟合度统计量一览表，回归残差、拟合值、正交多项式的值及对应每记录的有关统计量，散点图和正态概率图。

6R 偏回归与多元回归

计算剃除一组变量的线性效应以后，另一组变量的偏回归。6R 还特别用于多个因变量下的回归，分析使用原始数据、协方差阵或相关阵。输出内容：单变量综合统计量，相关阵，每个自变量与所有其它自变量的R-平方，每个因变量与自变量的R-平方，除去自变量效应后因变量间的偏相关，协方差阵及偏协方差，对应每个因变量的偏回归系数，散点图和正态概率图。

9R 所有子集回归

对于预测变量计算最优子集回归，子集的数目可以指定。子集选择有三种准则：1. 样本R-平方，2. 调整R-平方，3. Mallows 氏Cp。输出内容：对每种子集：小于十个子集的R-平方、调整R-平方及Mallows 氏Cp；对于最优子集：R-平方，调整R-平方，Mallows 氏Cp，估计标准误，回归系数F 检验；对最优子集中的每个变量：回归系数及其标准误，标准回归系数，t 统计量及其p 值、容许值，标准化、删除、加权残差及预测值，散点图和正态概率图；Mahalanobis 距离和Cook 距离；学生化残差的直方图；Durbin-Watson 统计量和序列相关。

AR 不用导数的非线性回归

利用拟高斯—牛顿最小二乘法估计非线性函数的参数，AR 内存六种函数，使用FUN 指定其它函数，参数可以有上下界约束和线性等式约束。AR 能用极大似然法估计参数的函数及其标准误，用于差分方程组的参数估计。输出内容：描述统计量；每次迭代的剩余平方和，参数估计，对分的数目；渐近相关阵估计；渐近标准误估计；每个记录的残差、预测值用其标准误、权、自变量与因变量值；残差和预测值、变量的散点图；加权残差的正态概率图。

LR 逐步LOGISTIC 回归

用逐步法估计线性logistic模型的参数向量。对于分类变量及其交互产生设计变量，在逐步过程中视做一组。在逐步法中，连续变量或一组设计变量同时进入或。剃除。其层次规则是仅当低阶效应和主效应在模型时，高阶交互才进入模型。程序使用的数据既可以是每种不同协变量取值下的表格式数据，也可以是每个对象或样本的单一记录。输出内容：描述统计量(区间尺度的变量) ；不同的取值及其频数(分类变量)；每步的对数似然值及其改变量，拟合度卡方，Hosmer 和C. C. Brown拟合度卡方检验；回归系数及其标准误，它们的比值，系数的渐近相关阵；每步上的进入及剃除统计量；所有步骤的一览表；每组预测概率的直方图；正分与误分表；对分析变量的不同组合提供：成功与失败的频数，预测概率，观察比例，对数比数比，标准残差；第一组比率对其预测概率和对数比数的散点图。对分析变量的不同组合提供：综合描述和图示。

PR 多分类Logistic回归

多项和有序资料的处理，系数的逐步极大似然比估计和近似方差估计，对数似然值和拟合优度。

§6.3.6 S 系列

3S 非参统计

计算下面一个或几个非参统计量：符号检验，Wilcoxon 符号秩次检验，Kruskal-Wallis 单方式方差分析，Kendall 一致性系数，Friedman 两方式方差分析，Mann-Whitney 秩和检验，Kendall 和Spearman 秩次相关系数。

§6.3.7 T 系列

1T 一维与二维谱分析

图示、描述统计及单一或序列对的分析，1T 计算谱分解，绘谱密度图，显示每个频带对时间序列总方差的相对贡献。其选项与特点有：基于协方差的谱估计，谱分析之前的数据处理(Tapering and padding)，带宽和其它指示，Y 关于X 的滞后，协方差或周期图的权，部分时点的谱分析，指示分析的频率范围，可信带，缺失值处理，预滤波和再染色(Prefiltering and recoloring)。

2T BOX—JENKINS 时序分析

使用Box-Jenkins 自回归—积分移动平均方法建立时序模型和传递函数模型。估计模型参数，进行诊断检查或残差分析。提供的方法：时间序列图示，识别(自相关，偏自相关及互相关函数)，季节组分建模，估计(包括参数估计，t 检验，缺失值估计)，残差诊断(包括Ljung-Box Q 统计量)，预测，干预分析，多输入传递函数模型，包括“白化”(”prewhitening” 除去自相关的滤波)。

§6.3.8 V 系列

1V 单方式方差分析及协方差分析

各组间协变量回归系数的平行检验。指示组间或调整均值间的线性对照，并对每个对比进行t 检验。输出内容：每组均值及合并均值；方差分析表和两两t 检验；可选组内统计量：极值、协方差阵和相关阵。指示协变量时有：回归系数，标准误及t 值；各组均值，调整均值及其标准误；检验斜率为零及斜率相等的方差分析表；每个协变量的组内斜率；调整各组均值的两两t 检验；可选的散点图；回归系数及调整均值间的相关。

2V 重复测量资料的方差协方差分析

对各种固定效应和重复测设计进行方差协方差分析，每格子数等或不等。固定效应设计包括完全和不完全析因设计如拉丁方、不完全区组设计、部分析因设计，输出内容：变量的格子均值及标准差，方差分析，每个记录的的预测值及其剩余，调整协变量的格子均值。重复测量设计允许组合重复因素和分组因素，但必须交叉不能嵌套。每个对象必须在重复测量因素的任何组合下有一个反应的取值。输出包括：格子均值，标准差、方差分析表，平方和以及正交组分的相关阵，球型条件的检验，组内因素的正交分解，重复测量因素的保守检验。

3V 一般的混合模型方差分析使用极大似然或约束极大似然方法估计固定和随机模型。混合模型可以很任意，而不需要2V 或8V 那样需要平衡。3V 允许针对特定假设进行检验。输出内容：单变量统计量；模型参数估计及其渐近标准误，t 统计量和p-值；2* 对数似然函数值；参数协方差阵估计；哑变量；固定效应所定义的格子均值、预测均值及其标准误；剩余；指定假设的检验，对数似然值及对数似然比检验，相应的自由度和概率。

4V 重复测量数据的单变量和多变量方差协方差分析

是一个通用程序，处理平衡或不平衡设计和重复测量，裂区和交叉(change over) 设计。输出包括：因素水平上的权或格子的权，格子为空时的分析，多变量分析，同时单变量、重复测量以及多变量分析，针对性处理某些形式的数据缺失，检验用户指定的关于因子水平或格子均值间的对比，用户指定的正交化效应。

5V 有结构协方差阵的不平衡重复测量模型

针对一大类实验设计和模型进行重复测量分析，包括那些协方差阵为特定形式的设计和不完全资料。使用ML 或REML 方法得到回归和协方差估计。实验设计：Longitudinal studies 以及重复测量实验；平衡或不平衡设计，包括由于缺失观察引起的不平衡，时变协变量。协方差结构，完全的指定包括：复合对称(Compound symmetry)，一阶自回归，Banded 或一般自回归，结构未定义(完全参数化)。需要附加输入的有：因子分析，随机效应，线性模型，用户定义的FORTRAN 子程序。

8V 一般混合模型方差分析– 格子大小相同

对任何格子大小相同的完整设计进行方差分析，如区套、交叉或部分嵌套和交叉设计进行方差分析，效应可以是固定的、混合的(包括重复测量) 和随机的。8V 不使用分组变量区分分组，不存在GROUP 段。输出内容：含期望均值的方差分析表，方差分量估计，格子均值、边缘均值、剩余，以及其它可选的统计量。

CA 对应分析

是一个多变量探索性数据分析程序，用于把频数表转成图示。CA 对频数表关联度的分分解类似于连续变量的主成分分析。CA 使用的数据格式可以是记录、标记格子频数、频数表。

DM 数据管理

交互式数据处理程序，与BMDP 各个程序是兼容的，DM 使用BMDP 文件和ASCII 文件，读取多记录类型和层次文件。其功能有：三个过程合并文件，二十四种函数抽取数据信息，压缩和不压缩单记录和多记录数据互换过程，排序、转换。打印、数据存贮，显示记录结构，计算综合统计量。

【附】例6.5的数据。

```
 1   49 1             26 1400 0             54    2 1             79   95 1 66 54 1.08
 2    5 1             27  262 1             55   60 1  9 52 1.51 80  481 0 25 46 1.41
 3   15 1  0 54 1.11 28   71 1 70 54 0.47 56  941 0 66 38 0.98 81  444 0  5 52 1.94
 4   38 1 35 40 1.66 29   34 1             57  148 1             82  427 0
 5   17 1             30  851 1 15 44 1.58 58  342 1 20 48 1.82 83   79 1 31 53 3.05
 6    2 1             31   15 1             59  915 0 77 41 0.19 84  333 1 36 42 0.60
 7  674 1 50 51 1.32 32   76 1 16 64 0.69 60   52 1  2 49 0.66 85    4 1
 8   39 1             33 1586 0 50 49 0.91 61    1 1             86  396 0  7 48 1.44
 9   84 1             34 1571 0 22 40 0.38 62   68 1             87  109 1 59 46 2.25
10   57 1 11 42 0.61 35   11 1             63  841 0 26 32 1.93 88  369 0 30 54 0.68
11  152 1 25 48 0.36 36   99 1 45 49 2.09 64  583 1 32 48 0.12 89  206 1138 51 1.33
12    7 1             37   65 1 18 61 0.87 65   77 1 11 51 1.12 90  185 1159 52 0.82
13   80 1 16 54 1.89 38    4 1  4 41 0.87 66   31 1             91  339 1
14 1386 1 36 54 0.87 40 1407 0 40 48 0.75 67  284 1 56 19 1.02 92  339 0309 45 0.16
15    0 1             41 1321 0 57 45 0.98 68   67 1  2 45 1.68 93  264 0 27 47 0.33
16  307 1 27 49 1.12 42    2 1             69  669 0  9 48 1.20 94  164 1  3 43 1.20
17   35 1             43    1 1             70   29 1  4 53 1.68 96  179 0 12 26 0.46
18   42 1 19 56 2.05 44   39 1             71  619 0 30 47 0.97 97  130 0 20 23 1.78
19   36 1             45   44 1  0 36 0.0  72  595 0  3 26 1.46 98  108 0 95 28 0.77
20   27 1 17 55 2.76 46  995 1  1 48 0.81 73   89 1 26 56 2.16 99   20 1
21 1031 1  7 43 1.13 47   71 1 20 47 1.38 74   16 1  4 29 0.61100   38 0 37 35 0.67
22   50 1 11 42 1.38 48    8 1             75    1 1           101   30 0
23  732 1  2 58 0.96 49 1141 0 35 36 1.35 76  544 0 45 52 1.70102   10 0
24  218 1 82 52 1.62 51  284 1 31 48 1.08 77   20 1           103    5 1
25 1799 0 24 33 1.06 52  101 1             78  514 0209 49 0.81
```