第十三章　GLIM

§13.1　GLIM 入门

§13.1.1　GLIM 简介

GLIM 第1版于70年代初问世，主要供专业统计人员使用。70年代中期3.22 版实现了商品化，1985年推出的3.77版最为流行，1993年又推出第4版。有别于其它通用统计软件包，使用GLIM要求用户对其建模过程有一个完整的概念。GLIM 提供了预分析的工具，如灵活的图形、优良的制表功能，回归、方差分析、列联表分析、生存分析，相应的误差分析等。其用于模型分析较之SAS、SPSS等的另一个优点是价格便宜。

GLIM 主要应用于三个方面。首先，它是一个强大的统计建模工具，供用户指定和拟合统计模型，评价拟合优度并给出估计量。建模过程不是预先设定的，这样用户就有了对建模过程最大程度的控制。其次，GLIM 可用于数据探索。最后，它是一个复杂的计算器，能够对单一的数或向量进行算术、逻辑、函数操作，宏操作。

§13.1.2　GLIM 系统组成

PC GLIM 3.77 系统组成：

EXAMPLES.GLM / EXAMPLES.LOG	样本程序及运行结果
EXMACLIB.GLM / EXMACLIB.LOG	宏调用样本程序及运行结果
GLIM.BAT / GLIM.LOG	系统运行批文件/运行记录文件
GLIMPROG.EXE / GLIMPROG.OVL	执行文件和覆盖文件
MACLIB.GLM	宏定义库
PROB.GLM	概率分布函数的宏
READ.ME	通道说明

§13.1.3　运行

设软件安装于$C:\backslash$ GLIM >，DOS 引导后，使用以下命令启动GLIM：

```
C:\> CD \GLIM <Enter>
C:\GLIM> GLIM <Enter>
```

或GLIM [参数1][参数2] $<$ *Enter* $>$

因GLIM.BAT 执行GLIMPROG.EXE，后者需要两个命令行参数，GLIM.BAT 则允许用户指定一个、两个或不指定，如：

```
C:\ GLIM>GLIMPROG %1 MACLIB.GLM
C:\ GLIM>GLIMPROG %1 %2
C:\ GLIM>GLIMPROG GLIM.LOG MACLIB.GLM
```

%1 是记录(transcript)文件，%2 表示宏定义文件。第三行即默认方式，代以GLIM.LOG，%2 代以MACLIB.GLM，因此只要对GLIM.BAT 进行适当的修改，可以使软件在任何工作盘上运行。

进入GLIM后，出现问号(?)提示，用户交互地以输入数据和指令。也可以先编好ASCII形式的命令文件，再读入执行。读取程序的语句是input/reinput，在该语句的后面要指定读入的通道号，通道号与DOS文件相联系。dinput 则用于读取data 语句中指示的量。return 或finish指令可以用于结束由input/reinput 或suspend引起的读取，skip 或exit 结束当前的栈也可以终止读取。

GLIM 4新增了manual命令，提供用户热线帮助。

【例13.1】例6.2问题的GLIM 程序如下：

```
$unit 8
$fact d 2 v 2 p 2
$data d v p count
$read
1 1 1 19  1 1 2 132
1 2 1  0  1 2 2   9
2 1 1 11  2 1 2  52
2 2 1  6  2 2 2  97
$yvar count
$erro poison
$fit d+v+p:+d.v:+v.p:+d.p $
$finish
```

GLIM 用$或 引导指令。很容易与其它软件类比，本例首先指定单元或记录个数、因素及水平，建立数据集。然后是模型部分，因本例是一个列联表分析，因变量是观察频数，误差为泊松分布。最后用fit命令把模型拟合出来，从主效应开始，依次增加交互项。

设程序存于文件loglin.glm，在GLIM中要运行它，使用以下命令：

$inp 7
File name? loglin.glm
$INP? $

运行指定通道号为7，第三行指示输入结束。

GLIM 作为计算器，如：

$calc 3.14159265/3 $

当运行结束时，使用命令$STop 返回至DOS系统下。

§13.1.4 GLIM 语言

GLIM 字符集：字母A—Z、数字0—9、符号+、-、*、/、**、()分别表示加、减、乘、除和括号，括号改变运算的优先级。特殊字符及其功能列表如下：

符号	名称	用途
$	美元符(dollar)	指令记号
:	冒号(colon)	重复符号
%	百分号(percent)	系统记号
#	井字号(hash)	替换符号
’	单引号(quote)	字串引号
&	与号(ampersand)	逻辑与
?	问号(query)	逻辑或
>	大于号(greater than)	大于号
<	小于号(less than)	小于号
_	下线(underline)	联接
[	左方括号(left-hand bracket)	左方括号
]	右方括号(right-hand bracket)	右方括号
;	分号(semi-colon)	定维记号
@	位置符号(at)	无效字符
|	竖线(modulus)	取模

变量的名字最多是四个字符，大小写字母意义相当。

令牌(tokens) 是输入字符的序列，其中包括指令名(directive names)、标识符(identifier)、值(values)、关键字(key words)、运算符号及以分隔符。由令牌可组成GLIM 的语句，一个语句由指令名及一系列项(item)组成，每个项都是令牌。GLIM 的一次运行(session) 是一套完整的语句，其定义是：[任务[$end 任务]] $stop，其中的任务是相关语句的组合，以end 结束。

标识符实际上可以指某种结构的数据或子文件(subfile)，有六种数据结构：

常量(scalar)	存单一的数
向量(vector)	存一列数
指针(pointer)	存向量名字
宏(macro)	存程序文本
函数(function)	影射实数
内部数据(internal)	系统变量值

常量分普通常量与系统常量，普通常量具有形式：%字母，故共有26 个，系统常量共有51 个。向量具有长度和水平数两种特征，用variate 和factor 指令产生的称为用户向量，与此对应的是系统向量，如：%fv 的长度由units 而定，存放的是fit 指令的拟合值。GLIM 的内部数据如%ssp，是平方和交叉乘积矩阵。

表达式：除了字符集中指示的以外，逻辑运算符：$<, <=, =, ==, /=, >=, >$ 以及&(AND)、?(OR)、/(NOT)，与一般高级语言相同。

GLIM 函数：由calculate 指令使用，或者由实参传至calculate 使用。

%ang(x)	方根的反正弦arcsin($\sqrt{x}$)
%exp(x)	自然指数exp(x)
%log(x)	自然对数ln(x)
%sin(x)	正弦函数sin(x)
%sqrt(x)	平方根函数$\sqrt{x}$
%np(x)	累积正态函数$\Phi(x)$
%nd	正态变量函数$N(x), N(\Phi(x)) = x; 0 < x < 1$
%tr	截尾函数
%gl(k,n)	产生一因素分组
%cu(x)	累积函数
%sr(0)/%sr(n)	标准伪随机函数，结果为$(0,1)$间的实数和$(0,n)$间的整数
%lr(0)/lr(n)	局部伪随机函数
%nd(sr(0))	随机正态偏差

函数%gl(最大值,重复数)根据指定的最大值进行若干次，很常用。

与广义线性模型相应，一个特定的模型可经误差的分布、线性部分构造和联系函数而确定。如对于正态误差分布，单位联系函数，是最简单的；列联表数据可看做来自poission分布，其联系函数为对数；量化的反应(quantal response) 结果r/n (n个对象中r个反应) 可视做具有二项分布，联系是probit 。通常可以使用GLIM默认的设置：误差为normal；连接函数为identity，尺度参量待估计；权为1，偏移量为零(偏移量出现于线性预报量中，没有参数，对每个观察有影响，如生存分析模型和稀释模型)，拟合为1 (常数项)。这些参数的关系可以组合成下表：

误差(error)	联系函数(link)	尺度参数
正态分布(Normal)	indentity	待估计
二项分布(Binomial)	logit	1
泊松分布(Poisson)	log	1
伽马分布(Gamma)	reciprocal	待估计

GLIM 用指令$link来指示联系函数，其后面的参数可以简写，如$link I 表示单位联系函数。除了上表给出的以外，还有S (平方根)、E(指数)、P(probit)、C(comp-log-log)。$是GLIM指令的分界符。

$error指令设定误差的分布形式，其参数也可以缩写，如：$error N 表示正态分布，除了上表给出的以外，还有G(gamma)即伽马分布。

尺度参数用指令$scale来设定。

现将GLIM 指令整理如下：

- $UNits n 表示标准向量的长度
- $DAta 变量名表表示待输入的变量名
- $Read 数据表读入数据

- $Yvariate 变量名指示因变量Y
- $Error 分布指示误差的类型
- $Fit 模型拟合模型

FIT 语句中效应的写法采用了Wilkinson与Rogers (1973) 的指定方法，这些记号如: $*, ., +, -, \backslash$的意义通常能从$DISP M 指令得到。

+ 表示效应的相加

. 表示简单交互

$*$ 表示交叉或层次交互

/ 表示嵌套

– 表示删除

因此，对于A,B,C三个效应的情况，可能有组合如：$A+B+A.B$、$A+B+A^2+B^2+A.B$。$A*B$与$A+B+A.B$等价，$A*B*C-A.B.C$与$A+B+C+A.B+B.C+A.C$等价，A/B与$A+A.B$等价，$(A+B)*C$与$A+B+C+A.C+B.C$等价，$(A+B)/C$与$A+B+A.B+A.B.C$等价。

GLIM 以美元符($) 做为语句分界符，如：

```
$units 18
$data freq
$read
15 11 14 17 5 11 10 4 8
10  7  9 11 3  6  1 1 4
$yvariate freq
$error p
$fit $
```

- $Display e r 显示拟合情况

这个语句使用重要，如使用M，L参数获得系统所分析的模型，E、A、U、V、C、S、T、R、W参数获得参数估计情况和残差，D则是离真度。

- $CAlculate 进行计算

$calculate mon=%gl(18,1) $fit mon
$display e r

- $INput 通道号文件读入程序文件
- $Argument 宏定义名参数表指示宏调用的信息
- $Use 宏定义名使用宏

- $Plot 纵横坐标画图

```
$calculate %a=50.84-%dv
$calculate %b=17-%df
$input 12 CHIT
$use CHIT %dv %df
$use CHIT %a %b
$plot %fv freq mon
$calculate f=%log(freq)
$calculate t=%log(%fv)
$plot t f mon
```

冒号(:) 用于命令的重复执行，如：calculate pw=1:x=2 $

- $STop 结束用于结束GLIM 的一次运行
- $FINish 指示程序文件结束
- $DINput 文件名通道号读数据文件
- $FOrmat Fortran 格式格式重定义 $data sex rev age
  $format (2x,f4.0,f1.0,5x,f2.0)
  $dinput 1
- $OUtput 文件名通道号结果写入外部文件
- $ECho 显示指示印出GLIM 所接受的所有信息
- $PRint 信息打印信息
- $LOok 常量/向量浏览常量/向量
- $Macro 宏定义名空格文本$End 宏的定义
- $ENVironment 代码获取系统信息

  它给出的信息包括C(通道分配)、D(数据)、E(外部PASS)、G(图形功能)、I(安装信息)、P(程序信息)、(随机数的种子)、S(系统)、U(可使用的空间)。
- $DElete 宏名表/变量取消宏定义/变量
- $EXTract 结构提供系统结构
- $Tabulate 变量;变量造表
- $TPrint 变量; 变量打印表的内容
- $End 结束一个分析
- $Assign 名=名,名=名变量赋值

- $Offset 向量引入一个偏移量
- $FActor 因素水平指示名义或因素变量
- $Variate 测量变量指示标识的长

不同的分析可以子文件的形式共存于文件中，即$subfile 文件名1,. . . , $subfile 文件名n $finish。GLIM 最多可以嵌套16 层，当前的层数存于系统量%CL 中，GLIM 每开始一项任务，其栈都重新初始化。GLIM 使用!引导注释。

控制指令有alias、cycle、recycle 指示控制拟合的计算。一次拟合的结果通过系统变量%fv、%lp、%wt、%wv、%dv、%
以及%dr、%va、%di 来观察。

指令tabulation, sort, look, tprint, print, plot, hist 进行向量制表、排序、按列或按表浏览、散点图和直方图，而最中心的内容是指定和拟合广义线性模型，对同一批数据指定不同的模型，增加或者减少包含的项。

输入/输出通道和宏调用：通道与DOS 的设备或文件联系，用$enviroment c 能观察到这些定义。

下例从6号通道读取外部文件test.dat数据矩阵：

```
$unit 9 $
$data x y $
$dinput 6 $ !若数据文件的宽度超于80列，使用$dinput 6 132 $
File name ? test.dat
$look x y $
```

若x与y的数据是先后次序排列，用以下的语句读取：

```
$unit 9 $
$data x $
$dinput 6
File name ? test.dat
$data y $
$dinput 6 $
$look x y $
```

第二次读取不需要继续给定文件名。

据READ.ME文件的提示，通道3指定为GLIM.LOG，通道5指定为MACLIB.GLM，如启用子文件TEST，使用命令：

```
? $INPUT 5 test $
```

屏幕显示：

```
****** Successful Macro Library Access *******
```

MACLIB.GLM 是以文本文件的形式提供给用户的宏义，其通道号存于系统变量

标识：GLIM 3.77 macro library, release 1.0, January 1985

子文件名	宏名	描述
数据描述与显示		
SUMM	SUMM	变量(variate)的综合统计量
STEM	STEM	茎叶图
SMOO	SMOO	变量(variate)的Tukey 平滑
统计工具		
CHIP	CHIP	χ^2概率
正态模型		
QPLOT	QPLOT	正态概率图
QPLOT	STAN	标准化残差
QPLOT	JACK	大折刀(Jackknife) 残差
TNOR	TNOR	使用适合度χ^2的正态性检验
TNOR	WDASH	Shapiro Francia W' 正态性检验
NORMAC	RSQ	R 平方统计量
NORMAC	TVAL	参数估计的t 值
LEV	LEV	杠杆值
BOXCOX	BOXCOX	关于y 变量的Box-Cox 转换族
BOXCOX	BOXFIT	固定λ的Box-Cox 转换
PRESS	PRESS	预测误差平方和
泊松模型和列联表		
二项分布模型		
伽马模型		
生存分析		
WEIB	WEIB	对于截尾数据拟合指数和威布尔分布
WEIB	RESP	使用WEIB宏后的残差图
其它		
TUNI	TUNI	变量是否为均匀分布的χ^2拟合度检验

MACLIB.GLM 说明了各个宏的入口参数和产出结果。

宏用于重复，专用的过程，循环和一些复杂的例行程序，如：

```
$macro n %nd($calc y= #n:...:z= #n $ !用于重复
$macro m :+a*b*c*d-a.b.c.d $endmac $
$fit #m $
```

可以借助于GLIM提供的宏功能进行专门的功能，如下面是一个结合系统变量进行残差图示的例子[5]。

```
$macro rplot$
$calc resid=%yv-%fv $
$sort resid $
$calc n=%cu(1) $
```

```
$calc norm=%nd((n-0.375)/(%(%nu+0.25)) $
$plot resid norm '*' $
$endmac $
$use rplot $
```

宏调用可以不限于一次，结合while指令可以进行多次调用，如：

$calc %a=1 $
$while %a update $
update 的结构是：
$macro update $
$calc %z1=%z1+1 $
... $calc %a=%if(%z1¡10,1,0) $
$endmac $

超于10次时停止，也可以根据条件进行切换和执行，如：

$calc %a=2 $
...
$switch %a one two thr$
$endmac $
$switch %a update $

根据%a的不同取值执行宏调用。
其它指令如：
$accuracy 4 $ 指定系统保留四位小数。
$calc x(8)=10.1 $ 改变x的第8个值。
$edit 2 3 x 0.2 0.1 $ 结果如同：$%calc x(2)=0.2:x(3)=0.1，其中的冒号表示重复最近一个指令。

对向量x进行排序只消使用指令：$sort x $!一个参数，按x的取值进行排序。
$sort y x $!两个参数，x次序不变，排序结果存于y。
$sort z y x $!三个参数，x不变，使用排序结果对y进行排序，结果存于z。
$units 10 $
$calc r=%sr(0)$
$sort s 1 r $

其好处是能够产生不重复的10个随机数。

$sort s 1 s $!记取排序的次序号。

产生滞后：

$assign a=3,9,4,6,5,1,8,2,10,7 $
$sort b a -2 $

结果是a的数据提前一行，因为最末一个数是无效的，故用以下语句：

$calc diff=b-a:wt=(%cu(1)/=10)$
$weight wt$
$look diff wt $
$sort b a 2 $!数组b含a的滞后值。
$look b a $

transrpit指令管理记录文件，很有用：

$units 9 $
$data x y $
$trans $
$dinput 8 $
File name: test.dat
$plot y x $
$trans i w f h o $
$plot y x $

先关闭记录，等结果满意后再存取。i，w，f，h，o 对应input(输入) ，warnin g messages(警告信息)，fault messages(错误信息)，help messages(帮助信息)，ordinary output(正常输出)。

当运行中途停止时，可以使用dump和restore指令保存和恢复现场。

【例13.2】以下程序说明了宏的用法[1]，CHIT 对给定的卡方检验算出概率水平，有χ^2值及自由度两个参量，结果是用GLIM.LOG给出的。

```
$macro CHIT
$calc %p=(%2==1)*(2-2*%np(%sqr(%1)))+(%2==2)*(%exp(-%1/2))
      +(%2>2)*(1-%np(((%1/%2)**(1/3)-1
      +2/(9*%2))/%sqr(2/(9*%2))))
$print 'CHI2 P='%p' for CHI2='%1' WITH '*-4%2 'd.f.';
$$endma
$return
$macro UCHI
   $use CHIT $calc %dv=%d-%dv:%df=%e-%df $use CHIT
$endma
$units 4 $data FREQ $read
72 714 655 41
$yvar FREQ $error p
$assign clas=1,-1,-1,1
$calc c1=2*(%gl(4,1)-2.5) : c2=(c1/2)**2-1.25
$fit $use CHIT %dv %df
$calc %d=%dv :%e=%df $disp er
$fit clas  $use UCHI $disp er
$fit c1+c2 $use UCHI $disp er
$finish
```

运行结果：

```
[o] scaled deviance = 1266.8 at cycle  4
[o]            d.f. =    3
[i] $calc %d=%dv :%e=%df $disp er
[o] CHI2 P=  0.     for CHI2=   1267. WITH    3.d.f.
[o]          estimate        s.e.     parameter
```

```
[o]     1       5.915     0.02594     1
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72    370.50    -15.508
[o]      2       714    370.50     17.846
[o]      3       655    370.50     14.780
[o]      4        41    370.50    -17.118
[o] scaled deviance = 11.158 at cycle  3
[o]            d.f. =  2
[o] CHI2 P=  0.0038 for CHI2=   11.16 WITH    2.d.f.
[o] CHI2 P=  0.     for CHI2=   1256. WITH    1.d.f.
[o]          estimate        s.e.     parameter
[o]     1       5.281     0.04892     1
[o]     2      -1.247     0.04892     CLAS
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72     56.50      2.062
[o]      2       714    684.50      1.128
[o]      3       655    684.50     -1.128
[o]      4        41     56.50     -2.062
[i] $fit c1+c2 $use UCHI
    $disp er [o] scaled deviance = 1.4458 at cycle  3
[o]            d.f. = 1
[o] CHI2 P=  0.2292 for CHI2=   1.446 WITH    1.d.f.
[o] CHI2 P=  0.     for CHI2=   1265. WITH    2.d.f.
[o]          estimate        s.e.     parameter
[o]     1       5.271     0.04937     1
[o]     2    -0.06409     0.02066     C1
[o]     3      -1.255     0.04922     C2
[o]     scale parameter taken as  1.000
[o]   unit  observed    fitted   residual
[o]      1        72     67.23      0.582
[o]      2       714    728.31     -0.530
[o]      3       655    640.69      0.565
[o]      4        41     45.77     -0.705
```

【例13.3 】二分类数据分析常用probit、logit 和极值分布模型，这三者具有相同的模型形式：$\pi(x) = F(\alpha + \beta x)$。毒理实验中，许多毒物剂量对数的容许值(tolerance)分布通常近似正态分布，则$\pi(x) = \Phi[(x-\mu)/\sigma], \Phi(.)$ 是标准正态分布的累积分函数，故$F = \Phi, \alpha = -\mu/\sigma, \beta = 1/\sigma, \Phi^{-1}(\pi(x)) = \alpha + \beta x$, 即probit 模型；而对$\pi(x) = exp[-exp(\alpha + \beta x)]$，有$log[-log(\pi(x))] = \alpha + \beta x$，它与极值分布对应。$G(x) = exp(-exp[-(x-a)/b]), b > 0, -\infty < a < \infty$，均值为a+0.577b，标准差为$\pi_b/\sqrt{6}$。

下面是一个生物检测(bioassay) 的例子[2]，是一个甲壳虫接触气性二硫化炭 5 小时后的死亡情况，死亡与否是一个二分类数。程序依次对logit、probit、complementary log-log 计算估计值。

```
$c Fitting logit/probit/extreme-value models
$unit 8
$data dose kill number
$read
$1.691 6  59 1.724 13 60 1.755 18 62 1.784 28 56
$1.811 52 63 1.837 53 59 1.861 61 62 1.884 60 60
$yvar kill
$error bin number
$fit dose
$disp e r v$
$link p
$fit dose $
$disp e r v$
$link c
$fit dose $
$calc survive=number-kill $
$c the yvar is replaced with kill
$yvar kill
$fit dose $
$disp e r v$
$dele dose kill number survive$
$finish
```

link 语句中C 表示双对数联系，E 表示指数联系，G 表示logit 联系，I 表示单位联系，L 表示对数联系，P 表示probit 联系，R表示倒数联系，S表示方根联系。以上程序运行结果如下：

scaled deviance = 11.116 at cycle 4 d.f. = 6

scaled deviance = 9.987 at cycle 4 d.f. = 6

scaled deviance = 3.5143 at cycle 4 d.f. = 6

因为尺度参数均为1，则规格化deviance 与deviance 相同，同时还可以看出，三个模型以最后的complentary log-log 较佳。

估计值	标准误	参数的方差-协方差	
1 -60.74	5.181	26.84	
2 34.29	2.913	-15.09	8.484
1 -34.96	2.648	7.012	
2 19.74	1.487	-3.937	2.213
1 -39.52	3.234	10.46	
2 22.01	1.796	-5.806	3.226

拟合值：

unit	死亡数	总数	logit	probit	comp log-log
1	6	59	3.503	3.407	5.653
2	13	60	9.820	10.686	11.282
3	18	62	22.421	23.438	20.942
4	28	56	33.875	33.784	30.339
5	52	63	50.048	49.559	47.681
6	53	59	53.339	53.370	54.188
7	61	62	59.239	59.682	61.117
8	60	60	58.755	59.239	59.948

§13.2 广义线性模型简介

§13.2.1 一般理论

一个随机变量的统计模型隐含着这样一个思想：即被研究的变量有一个确定的结构，能够解释现有资料和进行预测。这也是广义线性模型的思想。它的三个组成部分是，随机部分、系统部分和联系部分。随机部分指示了因变量的概率分布，系统部分是自变量的线性函数，联系部分指明了系统部分与随机部分期望值间的关系。

广义线性模型的随机部分由指数族分布的独立观测组成，它们的分布形式为：

$$f(y,\theta,\phi)=\exp\{[y\theta-b(\theta)]/a(\phi)+c(y,\phi)\}$$

其中ϕ是尺度化参数或离散参数，在列联表分析中常常取值为1，当$\phi>1$ 时模型方差过大，称为过度离散。

如正态分布$\theta=\mu, b(\theta)=0.5\theta^2, a(\phi)=\sigma^2, c(y,\phi)=-0.5[log(2\pi\phi)+y^2/\phi]$。$\phi=1$ 时，上式可以化成如下的形式(Agresti, A. 1990)：

$f(y;\theta)=a(\theta)b(y)exp[yQ(\theta)], Q(\theta)$ 称为自然参数。

当$\eta=Q$时称作典型联系函数，很常用。

如对于logit 模型，有：

$f(y;\pi)=\pi^y(1-\pi)^{1-y}=(1-\pi)exp[ylog(\pi/(1-\pi)], Q(\pi)=log[\pi/(1-\pi)]$

表 13.1 几种分布所对应模型的各个部分

随机部分	联系函数	系统部分	模 型
正态分布	单位联系	连续的	回归
正态分布	单位联系	分类的	方差分析
正态分布	单位联系	混合的	协方差分析
伯努利分布	logit	混合的	logistic 回归
泊松分布	对数联系	混合的	对数线性模型
多项分布	广义的logit	混合的	多项反应模型

对于Poisson 模型，有：

$$f(n;m)=\exp(-m)(m^n)/n!=exp(-m)(1/n)exp(nlog(m)]$$

故其联系函数是log。与广义线性模型的定义$\mu=X\beta,\eta=g(\mu)$ 相比，有：$\eta=log(m)=X\beta$，是一个对数线性模型。列联表中的计数可以看成来自泊松分布。

一个观察y的标准线性模型是$y=X\beta+\varepsilon,\mu=X'\beta$构成系统部分，随机部分与系统部分的关联是$\eta=g(\mu),g$是$\mu$的任何单调可微函数。最常用的链接或联系函数为：

单位(identity) 联系 $\eta=\mu$

对数比数(logit) 联系 $\eta=ln[\mu/(1-\mu)],0<\mu<1$

概率单位(probit)联系 $eta=\Phi^{-1}(\mu),0<\mu<1,\Phi$ 为$N(0,1)$ 的分布函数

重对数(log-log) 联系 $\eta=ln[-ln(1-\mu)],0<\mu<1$

指数(power) 联系 $\eta=\begin{cases}\mu^r & r\neq 0\\ \log(\mu) & \text{r=0}\end{cases}$

指数分布族包括正态、伽马、泊松、二项、贝塔、负二项、卡方和逆正态分布等，它们关联的的分析模型列于下表。

广义线性模型的参数估计常用迭代加权最小二乘法。

表示模型拟合效果的量有偏差或离真度(deviance) 和广义Pearson χ^2统计量。设对给定模型的似然为$L(\hat{\mu};y)$，由数据而来的最大似然为$L(y;y)$，则规格化偏差$=2[L(y;y)]-L(\hat{\mu};y),\hat{\mu}=y$是估计值。

模型的检验常结合残差分析、数据诊断与图形分析等手段，类似通常的回归诊断问题。

许多统计软件可以进行广义线性模型分析，如GLIM、Genstat、S 以及SAS 6. 08 中的PROC GENMOD 等。第二章介绍了线性回归分析、logistic 回归和Cox回归，这里则对列联表的对数线性模型进行更详细的说明，模型中危险因子、混杂因子和反应均为离散性的数据，可整理成列联表格式。模型在SAS 、SPSS/PC+、BMDP、SYSTAT等软件包均可实现。

§13.2.2 列联表分析用例

我们知道$R\times C$表的独立性检验时，每个观察格子的的期望值边缘概率的积乘以总的格子数，将这一等式两取对数，就成为一个对数线性模型，即对数的线性和，联系函数也就是对数。根据列联表变量是有序或名义的类型，对应不同的对数线性模型。

最简单的是2×2 表，其比数比(odds ratio) 是$\theta = n_{11}n_{22}/n_{n_12}n_{21}$，$log(\theta)$ 的渐近标准误是$\sigma(log[\theta]) = \sqrt{\Sigma_i\Sigma_j 1/n_{ij}}, i,j = 1,2$。有时也在此式的每一个 n 值上加上0.5，标准误的公式也类似。两维列联表中对数线性模型参数λ_{12}可经$0.25ln(\theta)$而估计。

格子数为正值时的对数线性模型：$\log(m) = X\beta$。如在2×2表下，$\log(m_{ij}) = \mu + \lambda_i^x + \lambda_j^y$，约束为$\Sigma_i\lambda_i^x = \Sigma_j\lambda_j^y = 0$，此即：

$$\log\begin{pmatrix} m_{11}\\ m_{12}\\ m_{21}\\ m_{22}\end{pmatrix} = \begin{pmatrix} 1 & 1 & 1\\ 1 & 1 & -1\\ 1 & -1 & 1\\ 1 & -1 & -1\end{pmatrix}\begin{pmatrix}\mu\\ \lambda_1^x\\ \lambda_1^y\end{pmatrix}$$

如：$\log(m_{12}) = \mu + \lambda_1^x + \lambda_2^y = \mu + \lambda_1^x - \lambda_1^y$

对于独立Poisson模型

$$L(m) = \Sigma_i n_i\log(m_i) - \Sigma_i m_i = \Sigma_i n_i(\Sigma_j x_{ij}\beta_j) - \Sigma_i\exp(\Sigma_j x_{ij}\beta_j)$$

$$\frac{\partial L(m)}{\partial\beta_j} = \Sigma_i n_i x_{ij} - \Sigma_i m_i x_{ij}, X'n = X'\hat{m}$$

$$\frac{\partial^2 L(m)}{\partial\beta_j\beta_k} = -\Sigma x_{ij}\frac{\partial m_i}{\partial\beta_k} = -\Sigma_i x_{ij}x_{ik}m_i$$

设有k个2×2 表，设定各表的边缘合计为$\{n_{+1k}, n_{+2k}, n_{1+k}, n_{2+k}\}$ 时服从超几何分布，而且只用n_{11k}应能够确定$\{n_{+1k}, n_{+2k}, n_{1+k}, n_{2+k}\}$，

$$m_{11k} = E(n_{11k}) = n_{1+k}n_{+1k}/n_{++k}$$

$$V(n_{11k}) = n_{+1k}n_{+2k}n_{+1k}, n_{+2k}/n_{++k}^2(n_{++k} - 1)$$

因有条件独立，它们可以相加，故有Mantel & Haenszel 统计量：

$$M^2 = (|\Sigma n_{11k} - \Sigma m_{11k}| - 0.5)^2/\Sigma V(n_{11k}) \sim \chi_{(1)}^2$$

假设$\{n_i, i = 1, ..., n\}$ 是一个列联表中的观察，n_i非负，其最简单的情形是泊松分布，方差与均值为m_i。它具有性质$n = \Sigma n_i$ 仍为泊松分布，参数为Σm_i。泊松分布用于时空上随机发生的事件数，如某地某年某月n_1 (自然流产数)、n_2 (引产数)、n_3 (活产数) 具有泊松分布。由于这样的泊松抽样是随机样本，若$n = \Sigma n_i$ 而每个n_i 以n 为条件，n_i 不再独立，它们服从多项分布。在流行病学前瞻性研究中，对应于研究因素各水平的边缘合计视为固定，而把各水平上反应变量视为独立的多项分布样本；在回顾性研究中，反应变量的各水平的边缘合计值视为固定而把研究因素的各个水平视为多项分布的样本；在横断面研究中，则可以把总的计数视为固定。

与析因分析类似，层次对数线性模型分析可以用于研究因素的交互，高层的交互影响隐含了低一级的效应，三维的饱和模型是：

$$log(m_{ijk}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{ik}^{xz} + \lambda_{jk}^{yz} + \lambda_{ijk}^{xyz} \qquad (XYZ)$$

对独立模型，上式即是：

$$log(m_{ijk}) = \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z \qquad (X, Y, Z)$$

相当于：

$$l(m) = n\mu + \Sigma_i n_{i++}\lambda_i^x + \Sigma_j n_{+j+}\lambda_j^y + \Sigma_k n_{++k}\lambda_k^z$$

其它形式的模型有:

$$\begin{aligned}
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} && (XY, Z)\\
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{jk}^{yz} && (XY, YZ)\\
\log(m_{ijk}) = &\quad \mu + \lambda_i^x + \lambda_j^y + \lambda_k^z + \lambda_{ij}^{xy} + \lambda_{jk}^{yz} + \lambda_{ik}^{zx} && (XY, YZ, XZ)
\end{aligned}$$

自由度的分解：$\Sigma\Sigma\Sigma\pi_{ijk} = 1$，即共有$IJK - 1$ 个线性无关参数。对独立模型，直接估计使用$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$，使用$\pi_{i++}\pi_{+j+}\pi_{++k}$共有$I - 1 + J - 1 + K - 1$ 个参数，故自由度为$(IJK - 1) - (I + J + K - 3) = IJK - I - J - K + 2$。一般的情况如下，第二列是自由度。

(X,Y,Z)	IJK-I-J-K+2
(XY,Z)	(k-1)(IJ-1)
(XZ,Y)	(J-1)(IK-1)
(YZ,Z)	(I-1)(JK-1)
(XY,YZ)	J(I-1)(K-1)
(XZ,YZ)	K(I-1)(J-1)
(XY,XZ)	I(J-1)(K-1)
(XY,XZ,YZ)	(I-1)(J-1)(K-1)
(XYZ)	0

模型的估计采用极大似然法(ML)和迭代比例拟合(IPF) 方法。SAS/IML 还有专门的IPF函数。相比于ML，IPF 总有解并收敛至极大似然解。模型的拟合优度检验可以用Pearson χ^2 和似然比统计量，它们的公式是：

$$\chi^2 = \Sigma_i\Sigma_j\Sigma_k(n_{ijk} - \hat{m}_{ijk})^2/\hat{m}_{ijk}$$

对(XYZ) 自由度=$IJK - I - J - K + 2$。对于2×2表，当n_{ij}较小时可用修正的公式。

$$G^2 = 2\Sigma_i\Sigma_j\Sigma_k n_{ijk} log(n_{ijk}/\hat{m}_{ijk})$$

利用方差稳定变换，使每个格子的分布基本接近标准正态分布，如Freeman- Turkey变换：$\sqrt{n} + \sqrt{n+1} - \sqrt{4m+1}$，它的平方和符合$\chi^2$分布，自由度与理论数$m$ 有关。

列联表分析，常要进行χ^2的分解，这时应当保证：①、分表自由度应与原表相同；②、原表中的每一个格子当且仅当在一个分表中；③、原表的边缘合计须为一个小表的边缘合计。在经验上,可以看一下分表的G^2 是否与总表相同相同。

以第 6 章表6.2死刑的资料分析为例。不计受害者的种族，执行死刑的白人为12%，黑人为10%，黑人较白人要低；但控制了受害者的民族，黑人要高。可见对其边缘合计表分析，会出现结论不一致的现象，称作Simpson's paradox。又如在不平衡的关于疗效的研究中，一种疗法可能对男性病人和女性病人都是好的，但对于所有病人就不一定好。

表 13.2 表6.2的边缘合计表

	是	否	小计
白人	19	141	160
黑人	17	149	166
总计	36	290	326

表 13.3 死刑例子的估计结果

模型	G^2	差值	自由度
(D,V,P)	137.93		4
		129.80	
(DV,P)	8.13		3
		6.25	
(DV,VP)	1.88		2
		1.18	
(DV,VP,DP)	0.70		1
(DVP)	0.00		0

利用对数线性模型分析得下表：

可见，(DV,VP)模型是可以接受的。

与尺度模型相比，有序资料分析具有优点：①、模型参数易于解释；②、名义模型会出现饱和模型而此时不会出现；③、检验利于寻找关联交互的类型。如两维列联表两个变量均为有序的线性关系(linear by linear association) 模型是：

$$\log(m_{ij}) = \mu + \lambda_i^y + \lambda_j^y + \beta u_i \upsilon_j$$

u_i, υ_j 是行列的分数，$\beta = 0$则为独立模型。

对于uniform association 仅有一个参数β，模型是：

$$\log(m_{hj}m_{ik}/m_{hk}m_{ij}) = \beta(u_i - u_h)(\upsilon_k - \upsilon_j), h < i, j < k$$

使用局部比数比$\theta_{ij} = m_{ij}m_{i+1,j+1}/m_{i,j+1}m_{i+1,j}$是有用的，对上述线性x线性模型有$\log(\theta_{ij}) = \beta(u_{i+1} - u_i)(\upsilon_{j+1} - \upsilon_j)$，在等距分数时，所有比数比是一样的，故称均匀关联。有序模型的一种构造方法是使用相邻两个概率的比。在线性x 线性关联下为：

$$log(\pi_{j+1|i}/\pi_{j|i}) = log(m_{i,j+1}/m_{ij}) = (\lambda_{j+1}^y - \lambda_j^y) + \beta(\upsilon_{j+1} - \upsilon_j)u_i$$

对于单位间隔$\{\upsilon_j\}$，可以简写为$\alpha_j + \beta u_i$。

表 13.4 工作满意度分析结果

模型	G^2	df	P
独立(I)	12.03	9	0.211
均匀关联($L \times L$)	2.39	8	0.967
条件独立($I\|L \times L$)	9.64	1	0.035

记$(\pi_1(x), ..., \pi_J(x))$ 是反应概率，相邻两组的logits 是$L_j = log[\pi_j(x)/\pi_{j+1}(x)], j = 1, ..., J-1$。若要拟合模型$L_j = \alpha_j + \beta' x$，可用关系式

$$L_j^\star = log[\pi_j(x)/\pi_J(x)] = \Sigma_{k=j}^{J-1}\alpha_k + \beta'(J-j)x = \alpha_j^\star + \beta' u_j, j = 1, ..., J-1$$

仍然用$\upsilon_1 \le \upsilon_2 \le ... \le \upsilon_j$ 做分数，模型$M(x) = \Sigma_j \upsilon_j \pi_j(x) = \alpha + \beta x$ 称作平均响应，表示了条件均值与自变量之间的线性关系。

使用有序反应资料的另一种方式是利用$F_j(x) = \pi_1(x) + ... + \pi_j(x), j = 1, ..., J$ 累积logit是$L_j = logit[F_j(x)] = log[F_j(x)/(1-F_j(x))]$，最简单的是$L_j(x) = \alpha_j, j = 1, , J-1, \alpha_j$ 称做断点，因为L_j是$F_j(x)$的增函数，所以α_j 是不减的，进一步，括入解释变量时，使用模型$L_j(x) = \alpha_j + \beta' x, j = 1, ..., J-1$。因为$L_j(x_1) - L_j(x_2) = \beta(x_1 - x_2)$ 从而称做比例比数比模型(proportional odds model)。对于单个自变量的情形，可以将函数用图表示出来，固定j时其图象类似于logistic 回归线。为了使$\beta_j > 0$有习惯上的解释，常常把模型写作$L_j(x) = \alpha_j - \beta' X, j = 1, ..., J-1$。相对于累积比数比模型和比例比数比模型，有累积链接模型(cumulative link model)：

$$\begin{aligned} G^{-1}[F_j(x)] &= \alpha_j - \beta' X \\ G^{-1}(u) &= \log(u/(1-u)) \\ G^{-1}(u) &= \log[-log(1-u)] \\ G^{-1}(u) &= \Phi^{-1}(n) \end{aligned}$$

等等。另外，常把多分类反应中的一个水平作为基线，如所有分类与最后一个水平比较。SAS CATMOD 提供了ALOGTS, CLOGITS, MEAN, LOGIT 几种选项来处理上述几种情况。

现对第 6 章工作满意度的资料进行分析，结果如下：

独立模型结果$G^2 = 12.03$, 自由度=9，关联是很弱的，但却忽略了“对工作满意的程度随工资增加”，使用均匀关联后，拟合得到改善$G^2 = 2.39$, 自由度=8，若用单位赋分，关联参数的估计为0.112(标准误0.036)，正值表示满意度随着工资的增加而增加，局部比数比是exp(0.112)=1.12，可信区间为exp(0.112±1.96*0.036)，即(1.04,1.20)，在端点的比数比则为exp(0.112(4-1)(4-1))=2.74 即高收入是低入的2.74倍。

对数线性模型与logit 的区别：在参数的解释方面，它不区分反应变量和原因变量，这会影响我们对于模型的选择；对于层次模型来说，高阶效应的存在就意味着组成它的低阶效应的存在。在存在反应变量时，对数线性模型相应于该反应的logit 模型，在反应量具有两种以上的分类时，与广义logit 模型相应。

工作满意度分析的程序如下：

```
$unit 16
$factor income 4 satisf 4
$data income satisf count
$read
 1 1 20  1 2 24 ... 1 4 82
 ...     ...    ... ...
 4 1  7  4 2 18 ... 4 4 92
$calculate uv=income+satisf$
$yvar count
$error pois
$fit income=satisf+uv$
$calculate v=satisf$
$fit income+satisf+income v$
$finish
```

现以精神损害(well，mild，moderate，impaired)与生活事件(X_1)及社会经济状况(X_2，1=high，0=low)的关系研究为例说明比例比数比模型(Agresti, A 1990)。

```
well 1 1  mild 1 5  moderate 0 0 impaired 1 8
well 1 9  mild 0 6  moderate 1 4 impaired 1 2
well 1 4  mild 1 3  moderate 0 3 impaired 1 7
well 1 3  mild 0 1  moderate 0 9 impaired 0 5
well 0 2  mild 1 8  moderate 1 6 impaired 0 4
well 1 0  mild 1 2  moderate 0 4 impaired 0 4
well 0 1  mild 0 5  moderate 0 3 impaired 1 8
well 1 3  mild 1 5               impaired 0 8
well 1 3  mild 1 9               impaired 0 9
well 1 7  mild 0 3
well 0 1  mild 1 3
well 0 2  mild 1 1
```

使用stata的ologit命令拟合模型：$L_j(a) = \alpha_j - \beta_1 X_1 - \beta_2 X_2, L_j$是累积分布取logit。

```
. label define aa 0 well 1 mild 2 moderate 3 impaired
. infile a:aa x1 x2 using table98.raw
. ologit a x1 x2,table
```

对数似然值为-49.55。

```
Ordered Logit Estimates                          Number of obs =     40
                                                 chi2(2)       =   9.94
                                                 Prob > chi2   = 0.0069
```

```
Log Likelihood = -49.548948                              Pseudo R2     = 0.0912

------------------------------------------------------------------------------
       a |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |  -1.111234   .6108775     -1.819   0.069      -2.308532     .086064
      x2 |   .3188611   .1209918      2.635   0.008       .0817216    .5560006
---------+--------------------------------------------------------------------
   _cut1 |  -.2819054   .6422652             (Ancillary parameters)
   _cut2 |   1.212789   .6606523
   _cut3 |   2.209368   .7209676
------------------------------------------------------------------------------
       a |      Probability            Observed
---------|--------------------------------------
    well |   Pr(      xb+u<_cut1)        0.3000
    mild |   Pr(_cut1<xb+u<_cut2)        0.3000
moderate |   Pr(_cut2<xb+u<_cut3)        0.1750
impaired |   Pr(_cut3<xb+u)              0.2250
    . ologitp p0-p3
```

使用label语句对对因变量反序编码，可以直接对回归的系数进行解释。当生活事件分数增加时精神损害程序加大，在高的社会经济水平下减低。给定生活事件分数下，低于任何水平的精神损害的比数比，在高的或低的社会经济状况下均为exp(1.111)=3.04倍。最后的ologitp给出了该模型下的概率预测值。

为了适应数据的特定结构，对数线性模型有很多特殊的处理方法。如配对资料的对称模型、Bradley-Terry 模型、准独立(Quasi-independent) 模型等，它们的含义和实现方法可参考Agresti, A.(1990) 和Lindsay, J.K.(1989)，BMDP 手册。

GLIM4 提供了许多新的功能，如包括了逆高斯(IG)分布以及指数和负二次连接。模型定义语句包括了正交多项式，用户自定义矩阵，以及连续变量的交叉乘积。这些功能由一系新的指令和函数来完成，如ELIMINATE 指令可以使分析和计算大大简化，用于配对病例对照研究、多项反应模型、Cox 比例风险模型。函数包括了卡方、t、F、贝塔、二项、泊松分布的概率和分位点，不完全伽马函数、对数伽马、双伽马、三伽马函数。与早期版本的Gauss-Jordan算法相比，新版本增加了Givens算法，后者更稳定和更精确。数据结构中增加了表类型，GRAPH 命令用于高分辨图形，GLIM4提供了宏编辑器。

GLIM4 的MANUAL命令提供了在线帮助功能，包括用例说明。

广义线性模型的理论意义决定了GLIM的应用价值。有关GLIM在随机过程中的用法可见[11]，其中介绍了Markov链、点过程和更新过程、生存曲线包括Cox 模型、生长曲线、时间序列、重复测量等方面的应用，并附有相应的GLIM程序。除了GLIM 以外，SAS 6.08 PROC GENMOD 和Genstat 5 也能进行广义线性模型分析，这些软件在应用时各有特色，只有对广义线性模型有一定的了解才能应用自如。

广义模型有许多推广，如Hastie和Tibshirani 提出的广义和模型(GAM) ，Liang 和Zeger 提出的广义估计模型(GEE)。S-Plus 的glm和gam 分别用于拟合广义线性模型和广义和模型。

广义线性模型的一般介绍及其推广可以参考有关文献，SAS 6.12 GENMOD可以进行GEE模型分析。