**28. Logistic回归**

**（一）Logistic回归**

**一、原理**

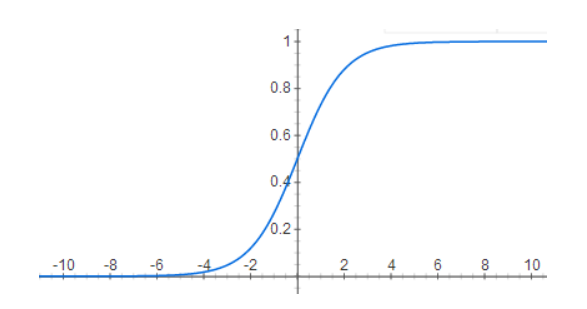
二元或多元线性回归的因变量都是连续型变量，若因变量是分类变量（例如：患病与不患病；不重要、重要、非常重要），就需要用Logistic回归。

Logistic回归分析可以从统计意义上估计出在其它自变量固定不变的情况下，每个自变量对因变量取某个值的概率的数值影响大小。

Logistic回归模型有“条件”与“非条件”之分，前者适用于配对病例对照资料的分析，后者适用于队列研究或非配对的病例-对照研究成组资料的分析。

对于二分类因变量，*y=*1表示事件发生；*y=*0表示事件不发生。事件发生的条件概率P{ *y*=1 | *xi* } 与 *xi* 之间是非线性关系，通常是单调的，即随着*xi* 的增加/减少，P{ *y*=1 | *xi* } 也增加/减少。

Logistic函数F(*x*)=，图形如下图所示：



该函数值域在(0,1)之间，*x*趋于-∞时，F(*x*)趋于0；*x*趋于+∞时，F(*x*)趋于1. 正好适合描述概率P{ *y*=1 | *xi* }. 例如，某因素*x*导致患病与否：*x*在某一水平段内变化时，对患病概率的影响较大；而在*x*较低或较高时对患病概率影响都不大。

记事件发生的条件概率P{ *y*=1 | *xi* } = *pi*，则

*pi* = =

记事件不发生的条件概率为

1- *pi* =

则在条件*xi*下，事件发生概率与事件不发生概率之比为

=

称为事件的发生比，简记为odds. 对odds取自然对数得到

上式左边（对数发生比）记为Logit(y), 称为y的Logit变换。可见变换之后的Logit(y)就可以用线性回归，计算出回归系数α和β值。

若分类因变量*y*与多个自变量*xi*有关，则变换后Logit(y)可由多元线性回归：



或 

**二、回归参数的解释**

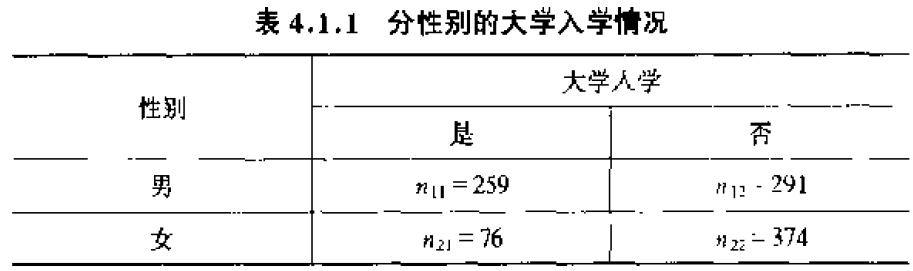
1. **三个名词**

**发生比（odds）= =**

例如，事件发生概率为0.6，不发生概率为0.4，则发生比为1.5（发生比>1，表示事件更可能发生）。

**发生比率（OR）= = = =**

即主对角线乘积/副对角线乘积，也称为交叉积比率，优势比。例如，



说明：大于1（小于1）的发生比率，表明事件发生的可能性会提高（降低），或自变量对事件概率有正（负）的作用；发生比率为1表示变量对事件概率无作用。

**相对风险（RR）= =**

用来进行两组概率之间的比较。当*p*1= *p*2时，相对风险为1，表明两组在事件发生方面没有差别。

2. **连续型自变量回归参数的解释**

截距α ： 基准发生比的对数，即当Logistic回归模型中没有任何自变量时（除常量外，所有自变量都取0值）所产生的发生比。

由于理解发生比，比理解对数发生比更容易，故将Logistic回归模型改写为：

odds = =

若βk>0（βk<0），则>1（<1）, 即*x*k每增加一个单位值时发生比会相应增加（减少）；若βk=0，则=1, 那么*x*k不论怎样变化发生比都不会变化。

反映了自变量*x*k增加一个单位时发生比所变化的倍数，即(*x*k+1)时的发生比与原发生比【(*x*k)时】之比。自变量*x*k每变化一个单位，发生比率的变化率为

注：由于βk是自变量*x*k的偏系数估计，故称为调整发生比率（AOR）的估计。实际中，往往更关心的不是自变量变化1个单位，而是变化一段水平b-a个单位，例如年龄每增加5岁，此时调整发生比率为

AOR =

3. **二分类自变量回归参数的解释**

二分类变量，例如性别，取值可以用0或1编码，也称为标志变量或虚拟变量。

若*x*k为取值0或1的二分类变量，则有





两式作差得



可见βk就是在控制其它变量条件下，*x*k=1与*x*k=0的对数发生比的差；也即是发生比率的对数，即调整发生比率的估计可表示为

= 

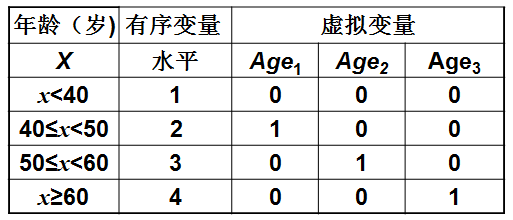
注意，发生比率是，而不是两者概率比

4. **多分类变量的处理与回归参数的解释**

当分类自变量多于两个类别时，需要建立一组虚拟变量来代表类型的归属性质。若一个分类变量包括m个类别，则可以产生m个相应的虚拟变量，但建模需要的虚拟变量的数目为m-1. 省略的那个类别作为参照类。

例如，年龄是有序变量，按年龄段分为四个类别：*x*<40，40≤*x*<50，50≤*x*<60, x≥60.

设置3个虚拟变量Age1表示40≤*x*<50（属于该年龄段则Age1=1，否则Age1=0）；Age2表示50≤*x*<60（属于该年龄段则Age2=1，否则Age2=0）；Age3 表示x≥60（属于该年龄段则Age3=1，否则Age3=0）；另一个不指定虚拟变量的*x*<40作为参照类（Age1，Age2，Age3都为0；哪一类作为参照类是随意的，取决于偏好或解释的方便）。



则模型为



同二分类变量时一样，β1代表40≤*x*<50与参照类（Age0: *x*<40）在因变量上的差别，故

β1 = ln(odds)Age1 vs Age0

且40≤*x*<50对 *x*<40的发生比率为.

注：（1）名义变量直接就是分类变量；连续变量也可以改为分类变量，例如考试成绩按分数段分为高、中、低三档。

（2）另外，也可采用效应变量编码，三种取值：-1, 0(参照类), 1.

（3）用SAS中CLASS语句指定分类变量，可以自动进行效应编码，作为一组变量纳入模型，并对每一类别单独做显著性检验。

5. **用概率解释自变量的作用**

(1) 对事件发生概率的偏作用，可用Logit函数求该自变量的偏导数来刻画：



其中*p*为事件发生概率。对于二元Logistic回归，*p*=0.5时，斜率最陡，此时x=-α/β, 称为中位有效水平。

(2) 利用得到的Logistic回归方程，可以预测概率：



有预测概率就可以计算某自变量值发生离散变化时，预测概率的离散变化：



**三、标准化系数**

通常在线性回归模型中的连续型自变量是以不同尺度度量的，这就使得某自变量中一个单位的变化并不等价于另一自变量上一个单位的变化。因此，要使用标准化系数（使得因变量的作用具有可比性）。

标准化系数表示自变量的一个标准差的变化所导致的因变量上以其标准差为单位度量的变化。有两种标准化的方法：

（1）先回归再标准化；

（2）先标准化*x*，*y*再回归；

注：对于分类变量，例如性别、民族等，变量的标准化是没有意义的。但其尺度标准要一致，比如0,1,2,3都变成1,2,3,4.

**四、偏相关**

偏相关，用来刻画在控制其他变量下，某自变量对Logistic回归的贡献（依赖于其它变量）。偏相关系数计算公式如下：



其中，d.f.为自由度，分母为-2倍的截距模型（只有截距）的对数似然值。

**五、回归参数的估计**

Logistic回归参数的估计通常采用最大似然法，其基本思想是先建立似然函数与对数似然函数，再通过使对数似然函数最大求解相应的参数值，所得到的估计值称为参数的最大似然估计值。

假设有N个案例构成的总体，Y1, …, YN. 从中随机抽取*n*个案例作为样本，观测值标注为*y*1 *, …, y*n. 设*p*i=P(*y*i=1 | *x*i)为给定*x*i的条件下得到结果*y*i=1的条件概率，而在同样条件下得到结果为*y*i=0的条件概率为P(*y*i=0 | *x*i)=1-*p*i. 于是，得到一个观测值的概率为



其中，*y*i=1或*y*i=0. 当*y*i=1时，P(*y*i)= *p*i= P(*y*i=1 | *x*i).

由于各项观测相互独立，故它们的联合分布可表示为各边际分布的乘积：



称为*n*个观测的似然函数。对于Logistic回归，. 根据最大似然原理，估计参数α和β使得似然函数L(θ)最大，令



称为对数似然函数，也即让ln[L(θ)]最大。令





采用牛顿迭代法解出α和β的估计值。

**六、假设检验**

1. 似然比检验

H0：β1=β2=…=βp=0

统计量



服从自由度为变量个数该变量（从L0到L1）的卡方分布。若P值<0.05, 则拒绝原假设。

2. 比分检验

以未包含某个或几个变量的模型为基础，保留模型中参数的估计值，并假设新增加的参数为零，计算似然函数的一价偏导数（也称“有效比分”）及信息距阵，两者相乘便得比分检验的统计量S. 样本量较大时，S近似服从自由度为待检验因素个数的分布。

3. Wald检验

即广义的T检验，统计量为



其中为的标准误。H0：βk=0为真时，Z为标准正态分布，W服从自由度为1的渐近分布。

βk的95%置信区间估计为：.

发生比率的1-α置信区间为：.

注：上述三种方法中，似然比检验最可靠，比分检验一般与它相一致，但两者均要求较大的计算量；而Wald检验未考虑各因素间的综合作用，在因素间有共线性时结果不如其它两者可靠。为计算方便，通常向前选取变量用似然比或比分检验，而向后剔除变量常用Wald检验。

**七、模型的评价——拟合优度检验**

检查模型估计与实际数据的符合情况。

检验统计量：1. 剩余差D; 2. 皮尔逊χ2

若统计量的P值＞0.05，则认为模型拟合较好。

**（二）PROC LOGISTIC过程步**

基本语法：

**PROC LOGISTIC data = 数据集** <可选项>**;**

**CLASS** 分类变量**;**

**FREQ** 频数变量;

**<WEIGHT** 权重变量;>

**<EFFECT** 效应名 = 效应类型(变量列表</可选项>);>

**MODEL** 因变量< (变量选项) > = 自变量列表</可选项>;

**<EXACT** 变量列表;>

**<CONTRAST** ’label’ 分类变量名 线性组合系数表;>

**<ODDSRATIO** < ’label’ > variable < / options > **;>**

**<OUTPUT OUT=**输出数据集</可选项>;>

**<ROC** < ’label’ > <变量> < /可选项>;**>**

**<SCORE** OUT=输出数据集<可选项>;**>**

<**TEST** 系数关系式;>

注：CLASS, EFFECT语句必须在MODEL语句之前；

CONTRAST, EXACT, ROC语句必须在MODEL语句之后。

说明：

（1）输入数据集可选项

DESCENDING —— 指定因变量按降序排序（“y=1”放前面）；

ORDER= —— 指定因变量的排序顺序；

PLOT —— 绘图选项；

（2）EFFECT语句

用原变量数据创建某种效应设计矩阵做对比用，例如LAG效应等。

（3）CLASS语句

对分类变量进行0-1化处理，变成虚拟变量；

（4）MODEL语句

是必不可少的，用来指定因变量和自变量。可以用可选项指定“y=1”，例如：

model remiss(event='1') = cell smear infil li blast temp;

可选项：

selection=stepwise /forward/ backward

sle或sls——指定变量进入或剔除出模型的显著水平；

Aggregate 和scale=n|p|d

计算偏差和pearson卡方拟和优度统计量，n表示对离差参数不进行校正；p规定离差参数的估计为pearson卡方统计量除以自由度；d规定离差参数的估计为偏差除以自由度；

alpha= ——设置置信限；

cl/clparm=WALD——估计所有参数/WALD参数的置信区间；

plrl——对自变量估计比数比的置信区间；

influence——做回归诊断；

RSQUARE——输出拟合的调整的R2；

EFFECTPLOT——输出模型拟合统计量；

（5）ESTIMATE语句——用来估计效应变量的线性组合的值；

（6）EXACT语句

用其它变量的充分统计量对变量的充分统计量做精确检验；

（7）CONTRAST语句

用来检验均值的线性组合关系的原假设。有三个基本参数，一是标签，二是分类变量名，三是效应均值线性组合的系数表（系数的次序是匹配分类变量按字母数字次序的水平值）。示例：

contrast 'US vs NON-U.S.' brand **1** **1** -**2**;

检验H0：β1+β2-2β3=0

（8）ROC语句——绘制ROC曲线；

（9）SCORE语句——输出若干结果到数据集。

（10）TEST语句

对系数关系式做检验，示例：（ai为自变量名）

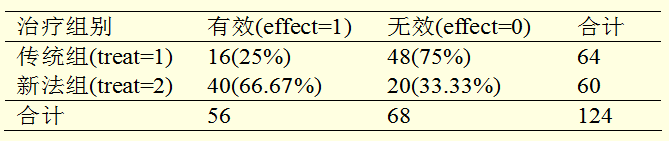
test1: test intercept + **.5** \* a2 = **0**;

test2: test intercept + **.5** \* a2;

test3: test a1=a2=a3;

test4: test a1=a2, a2=a3;

**例1** 不同治疗方法对某病疗效的影响研究：



用Logistic回归模型P{effect=1 | treat} = 拟合，即

Logit(p) = ln() =

**代码：**

**data** effects;

input treat effect count;

cards;

1 1 16

1 0 48

2 1 40

2 0 20

;

**proc** **logistic** data = effects DESCENDING;

freq count;

model effect=treat; /\*或者用 model effect(event='1') = treat; 前面就可以不用DESCENDING选项了\*/

**run**;

运行结果及说明：

| **响应概况** | | |
| --- | --- | --- |
| **有序 值** | **effect** | **总频数** |
| **1** | 1 | 56 |
| **2** | 0 | 68 |

建模的概率为 effect=1。

| **模型收敛状态** |
| --- |
| 满足收敛准则 (GCONV=1E-8)。 |

| **模型拟合统计量** | | |
| --- | --- | --- |
| **准则** | **仅截距** | **截距和协变量** |
| **AIC** | 172.737 | 152.361 |
| **SC** | 175.558 | 158.001 |
| **-2 L** | 170.737 | 148.361 |

| **检验全局零假设: BETA=0** | | | |
| --- | --- | --- | --- |
| **检验** | **卡方** | **自由度** | **Pr > 卡方** |
| **似然比** | 22.3768 | 1 | <.0001 |
| **评分** | 21.7087 | 1 | <.0001 |
| **Wald** | 20.2762 | 1 | <.0001 |

假设检验H0：β1=…=βk=0. 似然比检验的卡方值=-2ln*L*0 – (-2ln*L*k) = 170.737 – 148.361 =22.3768. 自由度为1-0=1（只有截距0个变量，到1个变量）， P值<0.0001 < α=0.05. 故拒绝H0.

| **最大似然估计值分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** | 1 | -2.8904 | 0.6390 | 20.4594 | <.0001 |
| **treat** | 1 | 1.7918 | 0.3979 | 20.2762 | <.0001 |

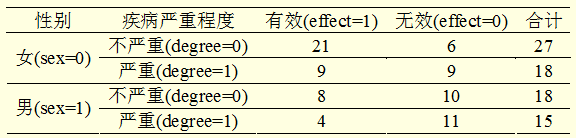
回归方程为：Logit(p)= -2.8904+1.7918\*treat

| **优比估计值** | | | |
| --- | --- | --- | --- |
| **效应** | **点估计值** | **95% Wald 置信限** | |
| **treat** | 6.000 | 2.751 | 13.087 |

发生比率（OR）=6.000.

| **预测概率和观测响应的关联** | | | |
| --- | --- | --- | --- |
| **一致部分所占百分比** | 50.4 | **Somers D** | 0.420 |
| **不一致部分所占百分比** | 8.4 | **Gamma** | 0.714 |
| **结值百分比** | 41.2 | **Tau-a** | 0.210 |
| **对** | 3808 | **c** | 0.710 |

**例2** 研究性别、疾病的严重程度对疾病疗效的影响，得数据如下：



拟合回归方程为Logit(p) = α + β1Sex +β2 Degree.

**代码：**

**data** effects2;

input sex degree effect count @@;

cards;

0 0 1 21 0 0 0 6 0 1 1 9 0 1 0 9

1 0 1 8 1 0 0 10 1 1 1 4 1 1 0 11

;

**run**;

**proc** **logistic** data = effects2 DESCENDING;

freq count;

model effect = sex degree / scale=none aggregate; \*模型的拟合优度检验;

output out=predict pred=prob; \*output语句设置输出结果,这里结果存在predict数据中，预测值为prob;

**run**;

**proc** **print** data=predict;

**run**;

\*考虑两个自变量的交互作用;

**proc** **logistic** data = effects2 DESCENDING;

freq count;

model effect = sex degree sex\*degree;

**run**;

**运行结果：**

| **响应概况** | | |
| --- | --- | --- |
| **有序 值** | **effect** | **总频数** |
| **1** | 1 | 42 |
| **2** | 0 | 36 |

建模的概率为 effect=1。

| **模型收敛状态** |
| --- |
| 满足收敛准则 (GCONV=1E-8)。 |

| **偏差和 Pearson 拟合优度统计量** | | | | |
| --- | --- | --- | --- | --- |
| **准则** | **值** | **自由度** | **值/自由度** | **Pr > 卡方** |
| **偏差** | 0.2141 | 1 | 0.2141 | 0.6436 |
| **Pearson** | 0.2155 | 1 | 0.2155 | 0.6425 |

剩余差D和Pearson拟合优度检验的P值分别为0.6436和0.6425都远大于0.05，故拟合结果较好可以接受。

唯一轮廓数: 4

| **模型拟合统计量** | | |
| --- | --- | --- |
| **准则** | **仅截距** | **截距和协变量** |
| **AIC** | 109.669 | 101.900 |
| **SC** | 112.026 | 108.970 |
| **-2 L** | 107.669 | 95.900 |

| **检验全局零假设: BETA=0** | | | |
| --- | --- | --- | --- |
| **检验** | **卡方** | **自由度** | **Pr > 卡方** |
| **似然比** | 11.7694 | 2 | 0.0028 |
| **评分** | 11.2410 | 2 | 0.0036 |
| **Wald** | 10.0644 | 2 | 0.0065 |

假设检验H0：β1=…=βk=0. 似然比检验的P值=0.0028< α=0.05. 故拒绝H0.

| **最大似然估计值分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** | 1 | 1.1568 | 0.4036 | 8.2167 | 0.0042 |
| **sex** | 1 | -1.2770 | 0.4980 | 6.5750 | 0.0103 |
| **degree** | 1 | -1.0545 | 0.4980 | 4.4844 | 0.0342 |

拟合回归方程为：Logit(p)= 1.1568-1.2770\*Sex – 1.0545\*Degree

| **优比估计值** | | | |
| --- | --- | --- | --- |
| **效应** | **点估计值** | **95% Wald 置信限** | |
| **sex** | 0.279 | 0.105 | 0.740 |
| **degree** | 0.348 | 0.131 | 0.924 |

优势比（OR）: 男性（Sex=1）治愈与未治愈的比值为：

p1/(1-p1) = exp(1.1568-1.2770\*1 – 1.0545\*Degree)

女性（Sex=0）治愈与未治愈的比值为：

p0/(1-p0) = exp(1.1568-1.2770\*0 – 1.0545\*Degree)

两个比值的比即优势比为：

OR = [p1/(1-p1)]/[ p0/(1-p0)] = =e-1.2770 = 0.279

| **预测概率和观测响应的关联** | | | |
| --- | --- | --- | --- |
| **一致部分所占百分比** | 60.0 | **Somers D** | 0.419 |
| **不一致部分所占百分比** | 18.1 | **Gamma** | 0.536 |
| **结值百分比** | 21.9 | **Tau-a** | 0.211 |
| **对** | 1512 | **c** | 0.709 |

| **Obs** | **sex** | **degree** | **effect** | **count** | **\_LEVEL\_** | **prob** |
| --- | --- | --- | --- | --- | --- | --- |
| **1** | 0 | 0 | 1 | 21 | 1 | 0.76075 |
| **2** | 0 | 0 | 0 | 6 | 1 | 0.76075 |
| **3** | 0 | 1 | 1 | 9 | 1 | 0.52555 |
| **4** | 0 | 1 | 0 | 9 | 1 | 0.52555 |
| **5** | 1 | 0 | 1 | 8 | 1 | 0.46999 |
| **6** | 1 | 0 | 0 | 10 | 1 | 0.46999 |
| **7** | 1 | 1 | 1 | 4 | 1 | 0.23601 |
| **8** | 1 | 1 | 0 | 11 | 1 | 0.23601 |

输出预测概率。

| **最大似然估计值分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** | 1 | 1.2528 | 0.4629 | 7.3239 | 0.0068 |
| **sex** | 1 | -1.4759 | 0.6628 | 4.9587 | 0.0260 |
| **degree** | 1 | -1.2528 | 0.6607 | 3.5954 | 0.0579 |
| **sex\*degree** | 1 | 0.4643 | 1.0012 | 0.2151 | 0.6428 |

Sex\*Degree的卡方检验P值=0.6428远大于其它参数的P值，故不用考虑该交互作用的影响。

**例3** 多分类自变量的虚拟变量处理（CLASS语句将会自动完成自变量的“虚拟变量化”处理过程，因变量不需要用CLASS语句处理），以及对比检验（CONTRAST语句）。

**代码**：

**data** uti;

input diagnosis : $13. treatment $ response $ count @@;

datalines;

complicated A cured 78 complicated A not 28

complicated B cured 101 complicated B not 11

complicated C cured 68 complicated C not 46

uncomplicated A cured 40 uncomplicated A not 5

uncomplicated B cured 54 uncomplicated B not 5

uncomplicated C cured 34 uncomplicated C not 6

;

**run**;

ods select FitStatistics;

**proc** **logistic** data = uti;

freq count;

class diagnosis treatment /param=ref;

model response = diagnosis|treatment;

**run**;

ods select FitStatistics GoodnessOfFit TypeIII OddsRatios;

**proc** **logistic** data = uti;

freq count;

class diagnosis treatment;

model response = diagnosis treatment /

scale=none aggregate;

**run**;

\*clodds：计算似然比的置信区间 clparm: 计算参数的置信区间;

ods select ClparmPL CloddsPL;

**proc** **logistic** data = uti;

freq count;

class diagnosis treatment;

model response = diagnosis treatment / scale=none aggregate clodds=pl clparm=pl;

**run**;

\*contrast:定制假设检验的方式，变量需要是矩阵形式;

ods select ContrastTest ContrastEstimate;

**proc** **logistic** data = uti;

freq count;

class diagnosis treatment /param=ref;

model response = diagnosis treatment;

contrast 'B versus A' treatment -**1** **1** / estimate=exp;

contrast 'A' treatment **1** **0**;

contrast 'joint test' treatment **1** **0**, treatment **0** **1**;

**run**;

**运行结果及说明**：（部分）

| **偏差和 Pearson 拟合优度统计量** | | | | |
| --- | --- | --- | --- | --- |
| **准则** | **值** | **自由度** | **值/自由度** | **Pr > 卡方** |
| **偏差** | 2.5147 | 2 | 1.2573 | 0.2844 |
| **Pearson** | 2.7574 | 2 | 1.3787 | 0.2519 |

| **3 型效应分析** | | | |
| --- | --- | --- | --- |
| **效应** | **自由度** | **Wald 卡方** | **Pr > 卡方** |
| **diagnosis** | 1 | 10.2885 | 0.0013 |
| **treatment** | 2 | 24.6219 | <.0001 |

| **优比估计值** | | | |
| --- | --- | --- | --- |
| **效应** | **点估计值** | **95% Wald 置信限** | |
| **diagnosis uncomplicated-complicated** | 0.382 | 0.212 | 0.688 |
| **treatment C-A** | 1.795 | 1.069 | 3.011 |
| **treatment C-B** | 4.762 | 2.564 | 8.847 |

| **参数估计和剖面似然置信区间** | | | | |
| --- | --- | --- | --- | --- |
| **参数** |  | **估计值** | **95% 置信限** | |
| **Intercept** |  | 1.6528 | 1.3621 | 1.9751 |
| **diagnosis** | **complicated** | -0.4808 | -0.7897 | -0.1987 |
| **treatment** | **A** | -0.1304 | -0.4618 | 0.2052 |
| **treatment** | **B** | 0.8456 | 0.4756 | 1.2523 |

| **优比估计和剖面似然置信区间** | | | | |
| --- | --- | --- | --- | --- |
| **效应** | **单位** | **估计值** | **95% 置信限** | |
| **diagnosis uncomplicated-complicated** | 1.0000 | 0.382 | 0.206 | 0.672 |
| **treatment C-A** | 1.0000 | 1.795 | 1.074 | 3.031 |
| **treatment C-B** | 1.0000 | 4.762 | 2.615 | 9.085 |

| **对比检验结果** | | | |
| --- | --- | --- | --- |
| **对比** | **自由度** | **Wald 卡方** | **Pr > 卡方** |
| **B versus A** | 1 | 8.6919 | 0.0032 |
| **A** | 1 | 4.9020 | 0.0268 |
| **joint test** | 2 | 24.6219 | <.0001 |

| **对比估计和逐行检验结果** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **对比** | **类型** | **行** | **估计值** | **标准 误差** | **Alpha** | **置信限** | | **Wald 卡方** | **Pr > 卡方** |
| **B versus A** | **EXP** | **1** | 2.6539 | 0.8786 | 0.05 | 1.3870 | 5.0778 | 8.6919 | 0.0032 |

**例4** 自变量连续有序变量（例如，年龄）的Logistic回归。

**代码：**

**data** coronary;

input sex ecg age ca @@ ;

datalines;

0 0 28 0 1 0 42 1 0 1 46 0 1 1 45 0

0 0 34 0 1 0 44 1 0 1 48 1 1 1 45 1

0 0 38 0 1 0 45 0 0 1 49 0 1 1 45 1

0 0 41 1 1 0 46 0 0 1 49 0 1 1 46 1

0 0 44 0 1 0 48 0 0 1 52 0 1 1 48 1

0 0 45 1 1 0 50 0 0 1 53 1 1 1 57 1

0 0 46 0 1 0 52 1 0 1 54 1 1 1 57 1

0 0 47 0 1 0 52 1 0 1 55 0 1 1 59 1

0 0 50 0 1 0 54 0 0 1 57 1 1 1 60 1

0 0 51 0 1 0 55 0 0 2 46 1 1 1 63 1

0 0 51 0 1 0 59 1 0 2 48 0 1 2 35 0

0 0 53 0 1 0 59 1 0 2 57 1 1 2 37 1

0 0 55 1 1 1 32 0 0 2 60 1 1 2 43 1

0 0 59 0 1 1 37 0 1 0 30 0 1 2 47 1

0 0 60 1 1 1 38 1 1 0 34 0 1 2 48 1

0 1 32 1 1 1 38 1 1 0 36 1 1 2 49 0

0 1 33 0 1 1 42 1 1 0 38 1 1 2 58 1

0 1 35 0 1 1 43 0 1 0 39 0 1 2 59 1

0 1 39 0 1 1 43 1 1 0 42 0 1 2 60 1

0 1 40 0 1 1 44 1

;

**run**;

\*selection用于选择逐步回归方法,包括forward,backward,stepwise

include:设定每个拟合模型中包含model语句中列的因子的个数.

units:可以设置自变量每次变化10个单位，计算的调整的发生比率AOR;

**proc** **logistic** data = coronary descending;

model ca = sex ecg age ecg\*ecg age\*age sex\*ecg sex\*age ecg\*age / selection=forward include=**3** details lackfit;

**run**;

**proc** **logistic** descending;

model ca=sex ecg age;

units age=**10**;

**run**;

**运行结果及说明**（部分）：

| **最大似然估计值分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** | 1 | -5.6418 | 1.8061 | 9.7572 | 0.0018 |
| **sex** | 1 | 1.3564 | 0.5464 | 6.1616 | 0.0131 |
| **ecg** | 1 | 0.8732 | 0.3843 | 5.1619 | 0.0231 |
| **age** | 1 | 0.0929 | 0.0351 | 7.0003 | 0.0081 |

经过筛选之后的3个自变量及截距的估计，回归方程略。

| **优比估计值** | | | | |
| --- | --- | --- | --- | --- |
| **效应** | **点估计值** | **95% Wald 置信限** | | |
| **sex** | 3.882 | 1.330 | 11.330 | |
| **ecg** | 2.395 | 1.127 | 5.086 | |
| **age** | 1.097 | 1.024 | 1.175 | |
| **残差卡方检验** | | | |
| **卡方** | **自由度** | **Pr > 卡方** | |
| 2.3277 | 5 | 0.8022 | |

| **对条目合格的效应分析** | | | |
| --- | --- | --- | --- |
| **效应** | **自由度** | **评分卡方** | **Pr > 卡方** |
| **ecg\*ecg** | 1 | 0.3766 | 0.5394 |
| **age\*age** | 1 | 0.7712 | 0.3798 |
| **sex\*ecg** | 1 | 0.0352 | 0.8513 |
| **sex\*age** | 1 | 0.0290 | 0.8647 |
| **ecg\*age** | 1 | 0.8825 | 0.3475 |

上述5个自变量的卡方检验P值>α=0.05，故不必加入回归方程。

| **Hosmer 和 Lemeshow 检验的分区** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **组** | **总计** | **ca = 1** | | **ca = 0** | |
| **观测值** | **期望值** | **观测值** | **期望值** |
| **1** | 8 | 2 | 1.02 | 6 | 6.98 |
| **2** | 8 | 1 | 1.80 | 7 | 6.20 |
| **3** | 8 | 3 | 2.59 | 5 | 5.41 |
| **4** | 8 | 3 | 3.42 | 5 | 4.58 |
| **5** | 8 | 4 | 4.07 | 4 | 3.93 |
| **6** | 9 | 6 | 5.38 | 3 | 3.62 |
| **7** | 9 | 4 | 5.97 | 5 | 3.03 |
| **8** | 8 | 7 | 5.99 | 1 | 2.01 |
| **9** | 8 | 7 | 6.98 | 1 | 1.02 |
| **10** | 4 | 4 | 3.77 | 0 | 0.23 |

| **Hosmer 和 Lemeshow 拟合优度检验** | | |
| --- | --- | --- |
| **卡方** | **自由度** | **Pr > 卡方** |
| 4.7766 | 8 | 0.7812 |

| **优比** | | |
| --- | --- | --- |
| **效应** | **单位** | **估计值** |
| **age** | 10.0000 | 2.531 |

自变量年龄按10岁为单位变化，得到的调整优势比为：AOR=2.531.

**例5** 回归诊断。

**代码**：

**data** uti2;

input diagnosis : $13. treatment $ response trials;

datalines;

complicated A 78 106

complicated B 101 112

complicated C 68 114

uncomplicated A 40 45

uncomplicated B 54 59

uncomplicated C 34 40

;

\*INFLUENCE诊断;

**proc** **logistic** data=uti2;

class diagnosis treatment / param=ref;

model response/trials = diagnosis treatment / influence;

**run**;

**proc** **logistic** data=uti2;

class diagnosis treatment / param=ref;

model response/trials = diagnosis / scale=none aggregate=(treatment diagnosis) influence iplots;

**run**;

**运行结果及说明**（略）

**例6** 精确检验。

代码：

**data** liver;

input time $ group $ status $ count @@;

datalines;

early antidote severe 6 early antidote not 12

early control severe 6 early control not 2

delayed antidote severe 3 delayed antidote not 4

delayed control severe 3 delayed control not 0

late antidote severe 5 late antidote not 1

late control severe 6 late control not 0

;

\*estimate=both,表示对第一个exact语句中指定的变量进行精确点估计. joint,表示对第二个exact中time、group进行联合检验;

**proc** **logistic** descending;

freq count;

class time(ref='early') group(ref='control') / param=ref;

model status = time group / scale=none aggregate clparm=wald;

exact 'Model 1' intercept time group / estimate=both;

exact 'Joint Test' time group / joint;

**run**;

运行结果（部分）：

| **最大似然估计值分析** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **参数** |  | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** |  | 1 | 1.4132 | 0.7970 | 3.1439 | 0.0762 |
| **time** | **delayed** | 1 | 0.7024 | 0.8344 | 0.7087 | 0.3999 |
| **time** | **late** | 1 | 2.5533 | 1.1667 | 4.7893 | 0.0286 |
| **group** | **antidote** | 1 | -2.2170 | 0.8799 | 6.3480 | 0.0118 |

| **优比估计值** | | | |
| --- | --- | --- | --- |
| **效应** | **点估计值** | **95% Wald 置信限** | |
| **time early-delayed** | 2.019 | 0.393 | 10.359 |
| **time early-late** | 12.849 | 1.305 | 126.471 |
| **group control-antidote** | 0.109 | 0.019 | 0.611 |

| **参数估计和 Wald 置信区间** | | | | |
| --- | --- | --- | --- | --- |
| **参数** |  | **估计值** | **95% 置信限** | |
| **Intercept** |  | 1.4132 | -0.1489 | 2.9754 |
| **time** | **delayed** | 0.7024 | -0.9330 | 2.3378 |
| **time** | **late** | 2.5533 | 0.2666 | 4.8400 |
| **group** | **antidote** | -2.2170 | -3.9417 | -0.4924 |

**例7** 绘制ROC曲线。

ROC曲线是根据一系列不同的二分类方式（分界值或决定阈），以真阳性率（灵敏度）为纵坐标，假阳性率（1-特异度）为横坐标绘制的曲线。

1. ROC曲线能很容易地查出任意界限值时的对疾病的识别能力。

2. 选择最佳的诊断界限值。ROC曲线越靠近左上角,试验的准确性就越高。最靠近左上角的ROC曲线的点是错误最少的最好阈值，其假阳性和假阴性的总数最少。

3. 两种或两种以上不同诊断试验对疾病识别能力的比较。在对同一种疾病的两种或两种以上诊断方法进行比较时，可将各试验的ROC曲线绘制到同一坐标中，以直观地鉴别优劣，靠近左上角的ROC曲线所代表的受试者工作最准确。亦可通过分别计算各个试验的ROC曲线下的面积(AUC)进行比较，哪一种试验的 AUC最大，则哪一种试验的诊断价值最佳。

**代码**：

**data** Data1;

input disease n age;

datalines;

0 14 25

0 20 35

0 19 45

7 18 55

6 12 65

17 17 75

;

ods graphics on;

**proc** **logistic** data=Data1 plots(only)=roc(id=obs);

model disease/n=age / scale=none clparm=wald clodds=pl rsquare;

units age=**10**;

effectplot;

**run**;

ods graphics off;

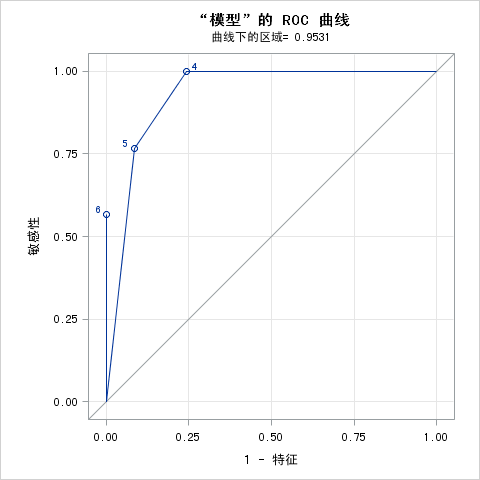
**运行结果**（部分）：

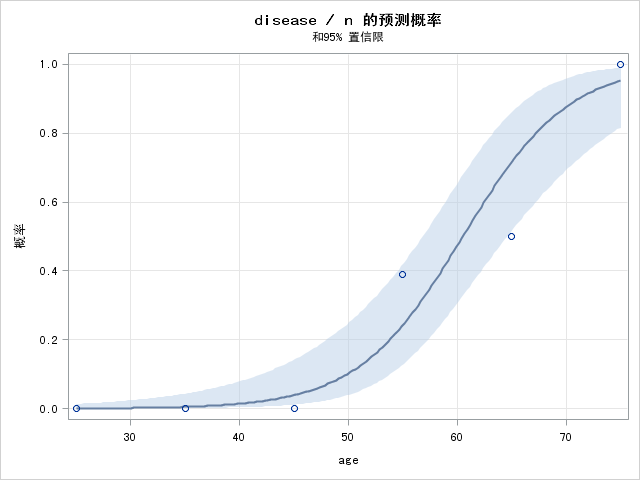
| **偏差和 Pearson 拟合优度统计量** | | | | |
| --- | --- | --- | --- | --- |
| **准则** | **值** | **自由度** | **值/自由度** | **Pr > 卡方** |
| **偏差** | 7.7756 | 4 | 1.9439 | 0.1002 |
| **Pearson** | 6.6020 | 4 | 1.6505 | 0.1585 |

| **最大似然估计值分析** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **参数** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** | 1 | -12.5016 | 2.5555 | 23.9317 | <.0001 |
| **age** | 1 | 0.2066 | 0.0428 | 23.3475 | <.0001 |

| **参数估计和 Wald 置信区间** | | | |
| --- | --- | --- | --- |
| **参数** | **估计值** | **95% 置信限** | |
| **Intercept** | -12.5016 | -17.5104 | -7.4929 |
| **age** | 0.2066 | 0.1228 | 0.2904 |

| **优比估计和剖面似然置信区间** | | | | |
| --- | --- | --- | --- | --- |
| **效应** | **单位** | **估计值** | **95% 置信限** | |
| **age** | 10.0000 | 7.892 | 3.881 | 21.406 |





**例8** 多分类Logistic回归，也称广义Logistic回归（因变量是多分类有序变量）。

**代码**：

**data** school;

length Program $ **9**;

input School Program $ Style $ Count @@;

datalines;

1 regular self 10 1 regular team 17 1 regular class 26

1 afternoon self 5 1 afternoon team 12 1 afternoon class 50

2 regular self 21 2 regular team 17 2 regular class 26

2 afternoon self 16 2 afternoon team 12 2 afternoon class 36

3 regular self 15 3 regular team 15 3 regular class 16

3 afternoon self 12 3 afternoon team 12 3 afternoon class 20

;

ods graphics on;

**proc** **logistic** data=school;

freq Count;

class School Program(ref=first);

model Style(order=data) = School Program School\*Program / link=glogit;

oddsratio program;

**run**;

ods graphics off;

运行结果：

| **响应概况** | | |
| --- | --- | --- |
| **有序 值** | **Style** | **总频数** |
| **1** | self | 79 |
| **2** | team | 85 |
| **3** | class | 174 |

| **分类水平信息** | | | |
| --- | --- | --- | --- |
| **分类** | **值** | **设计变量** | |
| **School** | **1** | 1 | 0 |
|  | **2** | 0 | 1 |
|  | **3** | -1 | -1 |
| **Program** | **afternoon** | -1 |  |
|  | **regular** | 1 |  |

SAS系统自动为因变量的值做虚拟编码（效应变量）。

| **模型收敛状态** |
| --- |
| 满足收敛准则 (GCONV=1E-8)。 |

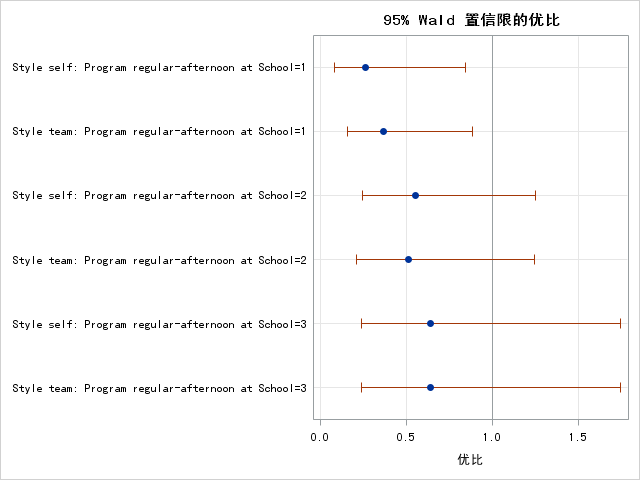
| **模型拟合统计量** | | |
| --- | --- | --- |
| **准则** | **仅截距** | **截距和协变量** |
| **AIC** | 699.404 | 689.156 |
| **SC** | 707.050 | 735.033 |
| **-2 L** | 695.404 | 665.156 |

| **检验全局零假设: BETA=0** | | | |
| --- | --- | --- | --- |
| **检验** | **卡方** | **自由度** | **Pr > 卡方** |
| **似然比** | 30.2480 | 10 | 0.0008 |
| **评分** | 28.3738 | 10 | 0.0016 |
| **Wald** | 25.6828 | 10 | 0.0042 |

| **3 型效应分析** | | | |
| --- | --- | --- | --- |
| **效应** | **自由度** | **Wald 卡方** | **Pr > 卡方** |
| **School** | 4 | 14.5522 | 0.0057 |
| **Program** | 2 | 10.4815 | 0.0053 |
| **School\*Program** | 4 | 1.7439 | 0.7827 |

| **最大似然估计值分析** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **参数** |  |  | **Style** | **自由度** | **估计值** | **标准 误差** | **Wald 卡方** | **Pr > 卡方** |
| **Intercept** |  |  | **self** | 1 | -0.8097 | 0.1488 | 29.5989 | <.0001 |
| **Intercept** |  |  | **team** | 1 | -0.6585 | 0.1366 | 23.2449 | <.0001 |
| **School** | **1** |  | **self** | 1 | -0.8194 | 0.2281 | 12.9066 | 0.0003 |
| **School** | **1** |  | **team** | 1 | -0.2675 | 0.1881 | 2.0233 | 0.1549 |
| **School** | **2** |  | **self** | 1 | 0.2974 | 0.1919 | 2.4007 | 0.1213 |
| **School** | **2** |  | **team** | 1 | -0.1033 | 0.1898 | 0.2961 | 0.5863 |
| **Program** | **regular** |  | **self** | 1 | 0.3985 | 0.1488 | 7.1684 | 0.0074 |
| **Program** | **regular** |  | **team** | 1 | 0.3537 | 0.1366 | 6.7071 | 0.0096 |
| **School\*Program** | **1** | **regular** | **self** | 1 | 0.2751 | 0.2281 | 1.4547 | 0.2278 |
| **School\*Program** | **1** | **regular** | **team** | 1 | 0.1474 | 0.1881 | 0.6143 | 0.4332 |
| **School\*Program** | **2** | **regular** | **self** | 1 | -0.0998 | 0.1919 | 0.2702 | 0.6032 |
| **School\*Program** | **2** | **regular** | **team** | 1 | -0.0168 | 0.1898 | 0.0079 | 0.9293 |

| **优比估计和 Wald 置信区间** | | | |
| --- | --- | --- | --- |
| **标签** | **估计值** | **95% 置信限** | |
| Style self: Program regular-afternoon at School=1 | 0.260 | 0.080 | 0.841 |
| Style team: Program regular-afternoon at School=1 | 0.367 | 0.153 | 0.883 |
| Style self: Program regular-afternoon at School=2 | 0.550 | 0.242 | 1.253 |
| Style team: Program regular-afternoon at School=2 | 0.510 | 0.208 | 1.247 |
| Style self: Program regular-afternoon at School=3 | 0.640 | 0.234 | 1.747 |
| Style team: Program regular-afternoon at School=3 | 0.640 | 0.234 | 1.747 |



注：多分类无序变量Logistic回归，用proc catmod过程步（略）。