回归分析

汪四水

2016.7.14

一般线性模型

# 1线性回归

## 1.1回归方程：

。（1）用于参数估计的模式：， ；（2）用于检验，

## 1.2回归中的方差分析：

**平方和分解**： SST = SSR + SSE







**自由度分解**： n-1 = p + (n-p-1)

，其计算利用对称幂等矩阵的秩等于迹，再利用迹的性质进行计算。

## 1.3最小二乘估计：

 ，, 

## 1.4检验：

**带约束的回归的参数估计**：

针对约束模型：，，

利用拉格郞日乘子法求参数估计：

**回归方程的检验：**（至少有一个分量不为0）

约束条件的作用：回归平方和的减少等于残差平方和的增加，即=



**检验参数：**

 分母中的标准差可通过回归系数方差矩阵中求得。

## 1.5衡量回归好坏的量：

,即：

注：

，即

## 1.6预测：

**点预测：**

1. 个体预测。用来预测。预测误差的方差为：
2. 均值预测。用来预测均值。预测误差的方差为

**区间预测：**

个体预测区间：

（注：， ，且相互独立）

均值预测区间：

（注：，，且相互独立）

例1：棉花红铃虫第一代产卵高峰日百株卵量x（粒）与百株累计卵量y（粒）的8组数据见程序，试建立一元线性回归模型，并进行显著性检验，并预测x=20时的y值，并求其置信区间（alpha=0.5）.

**data** reg;

input x y @@;

datalines;

14.3 46.3 14 30.7 69.3 144.6 22.7 69.2

7.3 16 8 12.3 1.3 2.7 7.9 26.3 20 .

run;

**proc** **gplot** data=reg;

plot y\*x;

symbol i=rl v=dot;

**run**;



**proc** **reg** data=reg;

model y=x/xpx i cli;

output out=a press=presidi;

**run**;

Model Crossproducts X'X X'Y Y'Y

Variable Intercept x y

Intercept 8 144.8 348.1

x 144.8 5899.66 13109.99

y 348.1 13109.99 29890.25

 , 

X'X Inverse, Parameter Estimates, and SSE

Variable Intercept x y

Intercept 0.2249182623 -0.005520346 5.9223664747

x -0.005520346 0.0003049915 2.0768029572

y 5.9223664747 2.0768029572 601.80822932

 , SSE=601.80822932

Analysis of Variance

Sum of Mean

Source DF Squares Square F Value Pr > F

Model 1 14142 14142 140.99 <.0001

Error 6 601.80823 100.30137(即MSE, )

Corrected Total 7 14744

Root MSE 10.01506 R-Square 0.9592

Dependent Mean 43.51250 Adj R-Sq 0.9524

Coeff Var 23.01651





Parameter Estimates

Parameter Standard

Variable DF Estimate Error t Value Pr > |t|

Intercept 1 5.92237 4.74970 1.25 0.2589

x 1 2.07680 0.17490 11.87 <.0001

回归系数的标准差(Standard Error).

 ,

，

The REG Procedure

Model: MODEL1

Dependent Variable: y

Output Statistics

Dep Var Predicted Std Error

Obs y Value Mean Predict 95% CL Predict Residual

1 46.3000 35.6206 3.6027 9.5773 61.6640 10.6794

2 30.7000 34.9976 3.6127 8.9459 61.0493 -4.2976

3 144.6000 149.8448 9.6297 115.8484 183.8412 -5.2448

4 69.2000 53.0658 3.6311 26.9988 79.1327 16.1342

5 16.0000 21.0830 4.0132 -5.3172 47.4833 -5.0830

6 12.3000 22.5368 3.9571 -3.8127 48.8863 -10.2368

7 2.7000 8.6222 4.6013 -18.3464 35.5908 -5.9222

8 26.3000 22.3291 3.9649 -4.0274 48.6856 3.9709

9 . 47.4584 3.5564 21.4532 73.4636 .

Sum of Residuals 0

Sum of Squared Residuals 601.80823

Predicted Residual SS (PRESS)  **5610.21596（以下程序是关于此值的计算）**

，其中表示去掉第i个观测值 后，用剩下的n-1个观测值建立线性回归模型后得到的第i个观测值的残差。

该值越小越好。（Predicted Residual SS (PRESS)预测残差平方和）

**proc** **print** data=a;

**run**;

**data** aa;

set a;

ss+(presidi\*\***2**);

**run**;

**proc** **print** data=aa;

**run**;

## 1.7偏F检验

的作用：，该值也等于

可以证明， ，其中为矩阵中对角线元素。

令 。若分子分母除以SST，则表示为

该检验可用于前进法，后退法，逐步回归法中挑选因子。

## 1.8标准回归分析（通径分析）

针对标准化的回归模型：， ，由正规方程,有：,

进一步有：， ，

称是对的直接通径系数，

间接通径系数： 

**注1**：标准化回归方程与非标准化时的回归方程中总平方和，回归平方和，残差平方和间的关系如下：



**注2**：复相关系数平方

**注3**：

注4：通径的图示：

**绘图path**

**例子：利用上面的数据(hald)进行通径分析**

**（step1）求直接通径系数**

**Data** hald;

Input x1-x4 y;

datalines;

7 26 6 60 78.5

1 29 15 52 74.3

11 56 8 20 104.3

11 31 8 47 87.6

7 52 6 33 95.9

11 55 9 22 109.2

3 71 17 6 102.7

1 31 22 44 72.5

2 54 18 22 93.1

21 47 4 26 115.9

1 40 23 34 83.8

11 66 9 12 113.3

10 68 8 12 109.4

**Run**;

\*ods trace on;

\*ods output Reg.MODEL1.Fit.y.ParameterEstimates=dp;

**proc** **reg** data=hald ;

model y=x1-x4/stb;

**run**;

\*ods output off;

\*ods trace off;

Parameter Estimates

Parameter Standard Standardized

Variable DF Estimate Error t Value Pr > |t| Estimate

Intercept 1 62.40537 70.07096 0.89 0.3991 0

x1 1 1.55110 0.74477 2.08 0.0708 0.60651

x2 1 0.51017 0.72379 0.70 0.5009 0.52771

x3 1 0.10191 0.75471 0.14 0.8959 0.04339

x4 1 -0.14406 0.70905 -0.20 0.8441 -0.16029

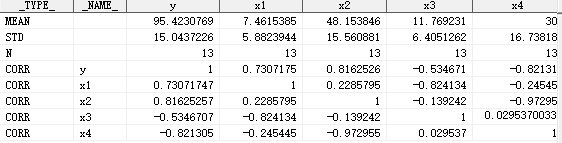
**(step 2)求相关系数**

**proc** **corr** data=hald noprob outp=corr;

var y x1-x4;

**run**;

**数据集corr中的结果如下：**



**（step 3）计算间接通径系数**

**data** tcorr;

set corr ;

Rxy=y;

if \_type\_='CORR';

IF \_name\_='y' then delete;

drop \_type\_ \_name\_ y;

**run**;

**data** path(drop=x1-x4);

set tcorr;

P\_x1y=**0.75830**\*x1;

P\_x2y=**0.19319**\*x2;

P\_x3y=**0.33994**\*x3;

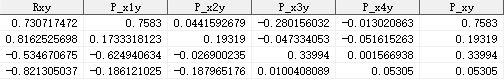
P\_x4y=**0.05305**\*x4;

input P\_xy@@;

datalines;

0.75830 0.19319 0.33994 0.05305

**run**;



**(验证**)

**Data** hald;

Input x1-x4 y;

datalines;

7 26 6 60 78.5

1 29 15 52 74.3

11 56 8 20 104.3

11 31 8 47 87.6

7 52 6 33 95.9

11 55 9 22 109.2

3 71 17 6 102.7

1 31 22 44 72.5

2 54 18 22 93.1

21 47 4 26 115.9

1 40 23 34 83.8

11 66 9 12 113.3

10 68 8 12 109.4

**Run**;

\*ods trace on;

\*ods output Reg.MODEL1.Fit.y.ParameterEstimates=dp;

**proc** **reg** data=hald ;

model y=x1-x4/stb;

**run**;

\*ods output off;

\*ods trace off;

输出的R2为：

**R-Square 0.9824**

**标准化回归系数为：**

Standardized

Estimate

0

0.60651

0.52771

0.04339

-0.16029

**proc** **corr** data=hald;

var x1-x4;

with y;

**run**;

相关系数为：

x1 x2 x3 x4

y 0.73072 0.81625 -0.53467 -0.82131

**data** c;

rr=**0.7307**\***0.60651** +**0.81625**\***0.52771** -**0.53467**\***0.04339** -**0.82131**\*(-**0.16029**);

**run**;

**proc** **print** data=c;

**run**;

**输出的RR为：**

rr

0.98237

**结果相同**

## 1.9选择模型的准则

设有p个变量的模型为全模型，q个变量的模型为选模型。其中.选取模型的准则有：

（1）均方误差越小越好



（2）修正的复决定系数越大越好



（3）Akaike信息量越小越好，BIC信息量越小越好





**注**：BIC准则倾向于选择较简单模型（变量较少的模型）

（4）Mallows的Cp量越小越好，（指预测均方误差越小越好）



## 1.10逐步回归

利用偏F检验，可进行前进法，后退法，逐步回归法进行模型的选择。

## 1.11复共线性

自变量有近似线性关系时称为复共线性关系，

共线性的危害：

1该现象的存在使得的逆不存在，导致回归系数的估计不存在。

2当存在不完全的共线性关系时，参数的方差会变大。因为。

3回归系数的估计值对样本数据敏感，回归系数的稳定性变差

4参数的检验也会出现困难。

5回归系数的专业意义变得难以解释

6回归模型是基于样本建立的，若样本数据中存在的多重相关问题在预测期发生了变化，则得到的预测结果则完全不能确定了

7对回归系数的解释不再适用（固定其他变量时，此话不再成立）

### 1.11.1方差膨胀因子

回归方程：， 的最小二乘估计有如下性质

由此可知各分量的方差为：，是在原来方差的基础上乘以得到，称为方差膨胀因子。该因子可以这样进行计算：

设变量y与自变量x1,x2,…,xp的复相关系数为;

变量xi与其余自变量的x1,…,xi-1,xi+1,…,xp的复相关系数为。则有

，因此有 ,



一般来说 方差膨胀因子大于10便认为有较强的共线性关系。

**容许度：**

### 1.11.2条件指数

将矩阵的最大特征值与最小特征值比值的平方根称为矩阵的条件指数。

一般来说（设计阵不含截距项），条件指数在10与30之间为弱共线性，在30与100之间为中等共线性，大于100为强共线性。

（SAS中的选项COLLINOINT—设计矩阵不含截距项； COLLIN---设计矩阵含截距项）

### 1.11.3方差比例

设的特征值为对应的特征向量矩阵为，可以证明：

，

称 为方差分量

记方差比例为：得到方差比例矩阵

一般来说，较大的条件指数所对应的方差分解的比例中，方差比例的数值超过0.5所对应的变量近似共线。

# 2方差分析

**§5.1 单因素方差分析**

设所考虑的因素为A，它有个水平，对第个水平得到一容量为的样本，记为（），设，且独立，其中表示因素A的第个水平下的理论均值。我们的目的是要知道这个水平的差异，即要检验的假设是。为了得到各水平的影响大小，将进行如下分解，，它称为因素第个水平的效应，，则得到单因素方差分析的数学模型：

 令为总平均，为第个水平下的样本均值。为总离差平方和，为误差平方和，为因素A的平方和（或称组间平方和），则有如下单因素方差分析表

表5.1 单因素方差分析表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F值 |  |
| 因素A |  |  |  |  |  |
| 误差 |  |  |  |  |  |
| 总和 |  |  |  |  |  |

若则否定，说明因素的水平间是有影响的。多重比较则进一步考察到底是哪些水平在真正起作用。

例1 设有三个小麦品种，经试种得第公顷产量的数据如下（单位：）

品种1：4350 4650 4080 4275

品种2：4125 3720 3810 3960 3930

品种3：4695 4245 4620

现问不同品种的小麦产量之间有无显著的差异？

其SAS程序如下：

data var1;

input kind$ yield@@;

cards;

1 4350 1 4650 1 4080 1 4275

2 4125 2 3720 2 3810 2 3960 2 3930

3 4695 3 4245 3 4620

;

proc glm;

class kind;

model yield=kind;

means kind/snk t alpha=0.05

means kind;

run;

此程序中先建立方差分析的数据集，然后调用GLM过程来进行方差分析，用CLASS语句指明分组变量为kind。MODEL语句指明以yield为因变量，以分组变量kind作为自变量，要求按分组变量进行方差分析。MEANS kind/snk t alpha=0.05语句要求对分组变量的各水平进行SNK、T（即LSD）检验，其显著性水平由alpha=0.05指定。means kind;要求给出分组变量kind的各水平的简单统计。其输出的结果为

General Linear Models Procedure

Dependent Variable: FIELD

Source DF Sum of Squares F Value Pr > F

Model 2 807311.250000 9.57 0.0059

Error 9 379488.750000

Corrected Total 11 1186800.000000

R-Square C.V. FIELD Mean

0.680242 4.883286 4205.00000

这部分给出了方差分析表，由Pr > F这一列可知，其概率值为0.0059比0.05小，故可判断品种这个因素对产量是有显著影响的。上面结果中还给出复相关系数的值，其值越大越好，因它表示由品种因素所引起的变异占总变异中比例。同时还给出变异系数和因变量的均值。

下面给出多重比较的结果：

T tests (LSD) for variable: FIELD

注释: This test controls the type I comparisonwise error rate

not the experimentwise error rate.

Alpha= 0.05 df= 9 MSE= 42165.42

Critical Value of T= 2.26

Least Significant Difference= 335.68

WARNING: Cell sizes are not equal.

Harmonic Mean of cell sizes= 3.829787

Means with the same letter are not significantly different.

T Grouping Mean N KIND

A 4520.0 3 3

A

A 4338.8 4 1

B 3909.0 5 2

这部分是参数LSD给出的结果。

Student-Newman-Keuls test for variable: FIELD

注释: This test controls the type I experimentwise error rate

under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 9 MSE= 42165.42

WARNING: Cell sizes are not equal.

Harmonic Mean of cell sizes= 3.829787

Number of Means 2 3

Critical Range 335.67055 414.3075

Means with the same letter are not significantly different.

SNK Grouping Mean N KIND

A 4520.0 3 3

A

A 4338.8 4 1

B 3909.0 5 2

这部分是参数SNK给出的结果。

多重比较结果在表示上有如下约定，标有相同字母的组，表示他们之间没有显著差异；对标有不同字母的组，则表示有显著差异。从二者比较的结果看所得的结论一致，也就是品种1和品种3的产量之间没有明显差异，而品种2与品种1和3的产量之间均有明显差异。

而means kind;语句输出的结果如下

Level of ------------FIELD------------

KIND N Mean SD

1 4 4338.75000 236.656396

2 5 3909.00000 154.288690

3 3 4520.00000 241.091269

其中SD表示标准差。

从上面也可看出方差分析的SAS过程的使用。一般地，对于非均衡设计用GLM过程，而对均衡设计可用ANOVA过程达到相同的结果，但它有比GLM过程快且占内存少的优点。其常用格式如下：

PROC ANOVA（GLM）选择项1；

CLASS 变量表； 必选项

MODEL 因变量=自变量表；

MEANS 效应/选择项2；

BY 变量表；

FREQ 变量表； 可选项

TEST H=效应项 E=误差项；

PROC ANOVA（GLM）的常用选择项为：

DATA=SAS数据集 指明ANOVA（GLM）过程要处理的数据集，缺省值为SAS最近产生的数据集。

OUTSTA=SAS数据集 将结果输出到指定的数据集中。

CLASS语句定义分组变量。

MODEL语句指定因变量和自变量，因变量为连续变量，自变量常常是分组变量。设A、B为分类变量，y为因变量，则以下几个MODEL语句的含义如下：

①MODEL y=A；单因素（A因素）方差分析。

②MODEL y=A B 主效应模型，无交互作用的两因素方差分析。

③MODEL y=A B A\*B 析因模型（带交互作用项），有交互作用的两因素方差分析

以上三个语句是必需的，其中CLASS语句必须出现在MODEL语句前面。

MEANS 语句要求ANOVA（GLM）过程计算出该语句后列出的每个水平所对应的因变量的均值。选择项2指明进行多重比较的方法，常用的TUKEY、SCHEFFE、LSD、DUNCAN和SNK等，同时还可规定检验的显著性水平的值（由ALPHA=值，缺省值为0.05）。不用选择项2时，可计算各水平对应因变量的均值。

TEST语句要检验的效应和误差项。其中H=后指定的是要检验的效应，E=后指定作为误差项的效应，此项必需指定，缺省是用MSE作为误差项。

BY语句和FREQ语句的使用同前，此处略。

**§5.2 两因素方差分析**

1 原理 设某实验影响的因素有两个A和B，并假定因素A有个水平，因素B有个水平，在因素A、B各水平的组合下均做次实验，其数值指标表示在A第水平B第水平下第次试验指标值，对固定的（），假定独立，且,,同样令，，，（），（），,则得到两因素的方差分析的数学模型： (I)，其中表示A因素的第水平的效应，表示因素B的第水平的效应，表示A因素的第水平和因素B的第水平的交叉效应。下面针对有无交互作用分别讨论。

（1）无交互作用 （I）式中=0时，表示无交互作用。为检验因素A和因素B的各水平是否有显著影响，我们需要检验假设和，同单因素方差分析一样，对总平方和及其自由度分别进行分解，得表2的方差分析表，

表5.2 无交互作用的两因素方差分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F值 |  |
| 因素A |  |  |  |  |  |
| 因素B |  |  |  |  |  |
| 误差 |  |  |  |  |  |
| 总和 |  |  |  |  |  |

若，则表示因素A有显著性影响；若，则表示因素B有显著影响；若二者均达到显著性水平，则说明因素A和因素B的水平均有显著影响。反之，则表明因素A、B没有显著性影响。

（2）有交互作用 对于要考虑交互作用的情况，我们最关心的检验是，当然还可以考虑因素A、B影响的显著性问题，即检验和，则有如下的方差分析表，见表3

表5.3 有交互作用的两因素方差分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 方差来源 | 平方和 | 自由度 | 均方 | F值 |  |
| 因素A |  |  |  |  |  |
| 因素B |  |  |  |  |  |
| 交互效应（A×B） |  |  |  |  |  |
| 误差 |  |  |  |  |  |
| 总和 |  |  |  |  |  |

只要，则表明因素A、B间存在交互作用，反之则没有交互作用。

例2 不考虑交互作用的两因素方差分析 为了考察蒸馏水的pH值和硫酸铜溶液浓度对化验血清中白蛋白与球蛋白的影响。pH值（A）取四种水平其值为5.40 5.60 5.70 5.80，记为A1、A2、A3和A4，硫酸铜浓度(B)取三种水平其值分别为0.04 0.08 0.10，记为B1、B2和B3，采用两因素的全面试验，所得结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A1 | A2 | A3 | A4 |
| B1 | 3.5 | 2.6 | 2.0 | 1.4 |
| B2 | 2.3 | 2.0 | 1.5 | 0.8 |
| B3 | 2.0 | 1.9 | 1.2 | 0.3 |

SAS程序如下:

ata var2;

do B=1 to 3;

do A=1 to 4;

input y@@;

output;

end;

end;

cards;

3.5 2.6 2.0 1.4

2.3 2.0 1.5 0.8

2.0 1.9 1.2 0.3

;

proc anova ;

class A B;

model y=A B;

means A B/snk;

run;

程序的第一部分建立SAS数据集var2，第二部分调用ANOVA过程进行两因素的方差分析，并用SNK法进行多重比较。结果如下：

Analysis of Variance Procedure

Dependent Variable: Y

Source DF Sum of Squares F Value Pr > F

Model 5 7.51083333 34.89 0.0002

Error 6 0.25833333

Corrected Total 11 7.76916667

这一部分给出的结果说明，模型是有效的，即这两种因素有影响。

R-Square C.V. Y Mean

0.966749 11.58130 1.79166667

Source DF Anova SS F Value Pr > F

A 3 5.28916667 40.95 0.0002

B 2 2.22166667 25.80 0.0011

这部分给出因素A、B效应的检验结果，从Pr > F列的概率值均比0.05小，说明这两种因素均有显著影响。其多重比较的结果为：

Analysis of Variance Procedure

Student-Newman-Keuls test for variable: Y

注释: This test controls the type I experimentwise error rate

under the complete null hypothesis but not under

partial null hypotheses.

Alpha= 0.05 df= 6 MSE= 0.043056

Number of Means 2 3 4

Critical Range 0.41456 0.5198111 0.5864881

Means with the same letter are not significantly different.

SNK Grouping Mean N A

A 2.6000 3 1

B 2.1667 3 2

C 1.5667 3 3

D 0.8333 3 4

Student-Newman-Keuls test for variable: Y

注释: This test controls the type I experimentwise error rate

under the complete null hypothesis but not under

partial null hypotheses.

Alpha= 0.05 df= 6 MSE= 0.043056

Number of Means 2 3

Critical Range 0.3590195 0.4501696

Means with the same letter are not significantly different.

SNK Grouping Mean N B

A 2.3750 4 1

B 1.6500 4 2

B

B 1.3500 4 3

结果显示因素A的四个水平之间均有显著性差异，而因素B的B1水平与其它两个水平B2、B3间均有显著性差异，而水平B2和B3之间没有显著性差异。

例3 有交互作用的两因素方差分析（需要有重复）

考察合成纤维中对纤维弹性有影响的两个因素，收缩率A和总拉伸倍数B，A和B各取四个水平，整个实验重复一次，试验的结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0（A1） | 4（A2） | 8（A3） | 12（A4） |
| 460（B1） | 71，73 | 73，75 | 76，73 | 75，73 |
| 520（B2） | 72，73 | 76，74 | 79，77 | 73，72 |
| 580（B3） | 75，73 | 78，77 | 74，75 | 70，71 |
| 640（B4） | 77，75 | 74，74 | 74，73 | 69，69 |

对上述问题进行方差分析的SAS程序如下：

data var3;

do B=1 to 4;

do A=1 to 4;

do i=1 to 2;

input y@@;

output;

end;

end;

end;

cards;

71 73 73 75 76 73 75 73

72 73 76 74 79 77 73 72

75 73 78 77 74 75 70 71

77 75 74 74 74 73 69 69

;

proc anova;

class A B;

model y=A B A\*B;

means A B A\*B/snk;

run;

输出的结果如下：

Analysis of Variance Procedure

Dependent Variable: Y

Source DF Sum of Squares F Value Pr > F

Model 15 158.71875000 7.87 0.0001

Error 16 21.50000000

Corrected Total 31 180.21875000

R-Square C.V. Y Mean

0.880701 1.569804 73.8437500

Source DF Anova SS F Value Pr > F

A 3 70.59375000 17.51 0.0001

B 3 8.59375000 2.13 0.1363

A\*B 9 79.53125000 6.58 0.0006

上述结果表明，在下，因素A有显著性影响，因素B则没有显著性影响，同时还可看出它们的交互作用达到了显著性水平。

Analysis of Variance Procedure

Student-Newman-Keuls test for variable: Y

注释: This test controls the type I experimentwise error rate

under the complete null hypothesis but not under

partial null hypotheses.

Alpha= 0.05 df= 16 MSE= 1.34375

Number of Means 2 3 4

Critical Range 1.2286997 1.4955629 1.6582506

Means with the same letter are not significantly different.

SNK Grouping Mean N A

A 75.1250 8 3

A

A 75.1250 8 2

B 73.6250 8 1

C 71.5000 8 4

Student-Newman-Keuls test for variable: Y

注释: This test controls the type I experimentwise error rate

under the complete null hypothesis but not under

partial null hypotheses.

Alpha= 0.05 df= 16 MSE= 1.34375

Number of Means 2 3 4

Critical Range 1.2286997 1.4955629 1.6582506

Means with the same letter are not significantly different.

SNK Grouping Mean N B

A 74.5000 8 2

A

A 74.1250 8 3

A

A 73.6250 8 1

A

A 73.1250 8 4

Level of Level of --------------Y--------------

A B N Mean SD

1 1 2 72.0000000 1.41421356

1 2 2 72.5000000 0.70710678

1 3 2 74.0000000 1.41421356

1 4 2 76.0000000 1.41421356

2 1 2 74.0000000 1.41421356

2 2 2 75.0000000 1.41421356

2 3 2 77.5000000 0.70710678

2 4 2 74.0000000 0.00000000

3 1 2 74.5000000 2.12132034

3 2 2 78.0000000 1.41421356

3 3 2 74.5000000 0.70710678

3 4 2 73.5000000 0.70710678

4 1 2 74.0000000 1.41421356

4 2 2 72.5000000 0.70710678

4 3 2 70.5000000 0.70710678

4 4 2 69.0000000 0.00000000

多重比较的结果表明，因素A中水平A2和A3之间没有显著差异，而A1与A4，A1与A2和A3，A4与A2和A3之间均达到显著性水平；因素B各水平间均无明显区别；由于该法并不提供对每个组合项的检验，故MEANS A\*B/SNK中的SNK对A\*B无效，相当于无后面的参数项，因而只输出观测数N，均值和标准差等项。

# 3协方差分析

**1. 原理**

可归为分块线性模型的特殊情形。先将普通的线性模型分块,考虑分块模型时,两个参数估计之间的关系,然后将这一关系应用于协方差分析模型上.



是阵,是阵,记,,则的最小二乘解为:,

略去部分,得到,则参数的最小二乘解为:(可称此为子模型的最小二乘解)

下面看与的关系,假定的列满秩(),并且假定和的列线性无关.则有:





其中,

**注1**：称为偏回归：

该回归参数的估计为

上式左边为回归方程的残差;

上式右边为回归方程的残差.

**注2**：  ，即，因此可认为是扣除协变量的影响后的估计值。

对任一可估函数,



在进行协方差分析时,将看成是协变量，只要针对作两两比较即可。

对任一可估函数，其BLU估计,其中

若服从正态分布，则可进行假设检验或进行区间估计等。

假设检验

检验统计量：，其中：

是指假设成立时的模型的残差平方和

，即含协变量时的模型残差平方和。

检验

检验统计量：

其中：是指假设成立时的模型残差平方和，即无协变量时的情况。

，即含协变量时的模型残差平方和。

* 研究牡蛎在不同温度的水中不同位置上的生长情况。

有人做了如下试验：分别在通向发电站的入口处（温度较低）不同位置（底部和表层）和出口处（温度较高）不同位置（底部和表层）及电站附近的深水处（底部和表层的中间）总共5个不同位置点上，随机地各放4袋牡蛎（每袋中有10个），共5×4=20袋。在将每袋牡蛎放入位置点之前，先洗干净称出每袋的初始体重，放在5个不同点一个月后再称出最后体重。试验结果数据见表1所示。

**表1 牡蛎在不同温度和位置上的生长数据**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 位置  *trt* | 重复数*rep*（*x*为初始体重，*y*为最后体重） | | | | | | | |
| 1 | | 2 | | 3 | | 4 | |
| *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 1（入口底部） | 27.2 | 32.6 | 32.0 | 36.6 | 33.0 | 37.7 | 26.8 | 31.0 |
| 2（入口顶部） | 28.6 | 33.8 | 26.8 | 31.7 | 26.5 | 30.7 | 26.8 | 30.4 |
| 3（出口底部） | 28.6 | 35.2 | 22.4 | 29.1 | 23.2 | 28.9 | 24.4 | 30.2 |
| 4（出口顶部） | 29.3 | 35.0 | 21.8 | 27.0 | 30.3 | 36.4 | 24.3 | 30.5 |
| 5（附近中部） | 20.4 | 24.6 | 19.6 | 23.4 | 25.1 | 30.3 | 18.1 | 21.8 |

data oysters;

input trt rep initial final;

cards;

1 1 27.2 32.6

1 2 32.0 36.6

1 3 33.0 37.7

1 4 26.8 31.0

2 1 28.6 33.8

2 2 26.8 31.7

2 3 26.5 30.7

2 4 26.8 30.4

3 1 28.6 35.2

3 2 22.4 29.1

3 3 23.2 28.9

3 4 24.4 30.2

4 1 29.3 35.0

4 2 21.8 27.0

4 3 30.3 36.4

4 4 24.3 30.5

5 1 20.4 24.6

5 2 19.6 23.4

5 3 25.1 30.3

5 4 18.1 21.8

;

proc print;

run;

proc glm;

class rep trt;

model final=rep trt initial/solution;

lsmeans trt / stderr tdiff;

means trt;

run;

proc glm;

class rep trt;

model final=trt initial trt\*initial;

run;

广义线性回归模型

# 1logistic回归

设因变量（或者响应变量）仅取0或1两个值，记事件发生的概率为，假设有个因素影响的取值，则称为Logistic线性回归模型，其中称为Logistic线性回归模型的协变量，是“事件发生”与“事件没有发生”的比值，称为优势（odds）。若odds值大于1，表明事件更容易发生；若odds值小于1，表明事件更不易发生；若odds值为1，则事件发生与不发生的可能性相同。

，又称为的Logistic变换，它是的单增函数，其值的范围为全体实数。

从Logistic线性回归模型可看出有，由此式可解释某预报因子的作用，如在其他变量固定的条件下，变化到时，,通常取，有。此外还可得出相应因变量发生与不发生的的概率值的表达式：即，。

对Logistic线性回归模型中参数的估计可用极大似然估计法，由于似然方程是一个关于参数的非线性方程，因而可通过迭代法求得其近似解。

# 2对数线性模型

设分类变量取值分别为：，分类变量取值分别为，其对应的总体概率分布律为：，，，列联表表示为表4.1.1。

表4.1.1 离散型随机变量，的总体分布及边缘分布表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | 随机变量 | | | | 合计 |
|  |  |  |  |
| 随机变量 |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 合计 | |  |  |  |  |  |

其中，；，

若利用抽样获得的数据整理成如下表4.1.2

表4.1.2离散型随机变量，的抽样数据表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | | 随机变量 | | | | 合计 |
|  |  |  |  |
| 随机变量 |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| 合计 | |  |  |  |  |  |

其中，；，

针对表4.1.2中的数据均为频数，对数线性模型是将联合单元的理论频数取对数后进行分解的一种统计建模方法，具体地有：



从而可写出对数线性模型写成一般形式：



其中的为常数项，为随机变量取时的主效应，为随机变量取时的主效应，为随机变量，的交互效应，=0则意味着随机变量，相互独立。

利用此特点，可利用对数线性模型方便判断变量间的关系。

对于二维列联表，共有三种模型

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 的分解式 | 随机变量，的关系 | 模型间的关系 |
| （I） |  | 饱和模型，，不独立 | 模型(I)包含模型（II） |
| （II） |  | ，相互独立 | 模型（II）包含模型（III） |
| （III） | 或 | ，独立，为离散均匀分布或为离散均匀分布 |  |

对数线性模型的检验：

模型（II）的独立性检验：



模型（III）的均匀性检验：



## 2.1三维表分析

对于三个离散随机变量，，分别取个不同类型的值，构成表，符号的表达与二维时相类似。对数线性模型的表示为：



其特点有当低阶的效应为0时，相应的高阶效应一定为0，如：，

|  |  |  |  |
| --- | --- | --- | --- |
| 类型 | 模型 | X,Y,Z关系 | 记号 |
| I |  | X，Y，Z相互独立 | （X，Y，Z） |
| II |  | Z与（X，Y）独立  Y与（X，Z）独立  X与（Y，Z）独立 | （Z，XY）  （Y，XZ）  （X，YZ） |
| III |  | 给定Z，X与Y独立  给定Y，X与Z独立  给定X，Y与Z独立 | （XZ，YZ）  （XY，YZ）  （XY，XZ） |
| IV |  | 各种优比分类取相同的值 | （XY，YZ，XZ） |
| V | 一般形式（饱和模型） | 一般模型 | （XYZ） |

对于三个离散随机变量，当其中一个固定时，其他两个变量取值的情况就是二维的，这相当于考虑条件分布。

# 3例子

例4.4某汽车保险公司某年有12299份保单，数据见表4.4.1。试判断变量间是否存在联系，并建立logistic回归模型和对数线性模型。

表4.4.1 某汽车保险公司的赔款记录

|  |  |  |  |
| --- | --- | --- | --- |
| 车辆类型（type） | 保险人年龄(age) | 保单类型(y) | |
| 有赔款记录的保单 | 无赔款记录的保单 |
| 普通车 | 年龄<25 | 741 | 2829 |
|  | 年龄≥25 | 882 | 4945 |
| 高性能车 | 年龄<25 | 453 | 1169 |
|  | 年龄≥25 | 248 | 1032 |

首先建立数据集：

**proc** **format**;

value typefmt **1**='普通车' **2**='高性能车';

value agefmt **1**='年龄<25' **2**='年龄超过25';

value yfmt **1**='有赔款记录的保单' **0**='无赔款记录的保单';

**run**;

**data** ex4\_4;

input type age y freq;

format type typefmt. age agefmt. y yfmt.;

datalines;

1 1 1 741

1 1 0 2829

1 2 1 882

1 2 0 4945

2 1 1 453

2 1 0 1169

2 2 1 248

2 2 0 1032

**run**;

（1）考察车辆类型，保险人年龄，保单类型间的关系。

即检验两个变量间是否独立，

**proc** **freq** data= ex4\_4;

weight freq;

table type\*y age\*y/norow nocol nocum nopercent chisq;

**run**;

其中选项chisq要求进行卡方检验。

车辆类型与保单类型间的列联表为：

表 — type X y

type y

频数 |无赔款记|有赔款记| 合计

|录的保单|录的保单|

---------+--------+--------+

普通车 | 7774 | 1623 | 9397

---------+--------+--------+

高性能车 | 2201 | 701 | 2902

---------+--------+--------+

合计 9975 2324 12299

卡方统计量及概率值为

“type \* y”表的统计量

统计量 自由度 值 概率

----------------------------------------------------------

卡方 1 68.5686 <.0001

由于超过68.5686值，自由度为1的概率值要小于0.0001，故认为车辆类型与保单类型间有关系。

保险人年龄与保单类型间的列联表为：

表 — age X y

age y

频数 |无赔款记|有赔款记| 合计

|录的保单|录的保单|

-----------+--------+--------+

年龄<25 | 3998 | 1194 | 5192

-----------+--------+--------+

年龄超过25 | 5977 | 1130 | 7107

-----------+--------+--------+

合计 9975 2324 12299

卡方统计量及概率值为

“age \* y”表的统计量

统计量 自由度 值 概率

----------------------------------------------------------

卡方 1 98.6061 <.0001

从输出的概率值可知（比0.0001小），保险人的年龄与保单类型间也是有关系的。

车辆类型与保险人年龄间的列联表：

表 — type X age

type age

频数 |年龄<25 |年龄超过25 | 合计

---------+--------+-----------+

普通车 | 3570 | 5827 | 9397

---------+--------+-----------+

高性能车 | 1622 | 1280 | 2902

---------+--------+-----------+

合计 5192 7107 12299

卡方统计量及概率值为

“type \* age”表的统计量

统计量 自由度 值 概率

----------------------------------------------------------

卡方 1 291.2870 <.0001

结果说明车辆类型与保险人年龄间有关系。

（2）用对数线性模型来考察车辆类型，保险人年龄与保单类型间的关系

**proc** **catmod** data=ex4\_4;

weight freq;

model type\*age\*y=\_response\_/p=freq;

loglin type|age|y;

**run**;

其中loglin type|age|y;语句中“type|age|y”表示三个因素的主效应，交互效应，三因素的交互效应等，相当于loglin type age y type\*age type\*y age\*y type\*age\*y;。

参数线性模型输出：

Maximum Likelihood Analysis of Variance

Source DF Chi-Square Pr > ChiSq

--------------------------------------------------

type 1 1597.46 <.0001

age 1 0.00 0.9581

type\*age 1 191.92 <.0001

y 1 2662.83 <.0001

type\*y 1 42.85 <.0001

age\*y 1 66.93 <.0001

type\*age\*y 1 0.79 0.3746

从方差分析表可看出，年龄（age）的作用不显著（概率值为0.9581，比0.05大），三因素的交互效应（type\*age\*y）不显著（其概率值为0.3746）。其他因素如车辆类型，赔款保单类型，以及车辆类型和保险人的年龄间有交互效应（其概率值比0.001还要小），车辆类型与赔款保单类型间，以及保险人的年龄与赔款保单类型间均存在交互效应，概率值均比0.05小。

故对数线性模型可写表示为：

（3）用logistic回归模型来考察保单类型与车辆类型和保险人年龄间的关系

**proc** **logistic** data=ex4\_4 desc;

weight freq;

model y=type age;

**run**;

desc要求模型以y取值为1进行建模。从模型的输出可以看出来

Response Profile

Ordered Total Total

Value y Frequency Weight

1 有赔款记录的保单 4 2324.0000

2 无赔款记录的保单 4 9975.0000

Probability modeled is y='有赔款记录的保单'.

参数的假设检验结果：

Analysis of Maximum Likelihood Estimates

Standard Wald

Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 -1.2666 0.1069 140.3892 <.0001

type 1 0.3512 0.0520 45.5956 <.0001

age 1 -0.4100 0.0469 76.4425 <.0001

可以看出来type,age对y的作用均达到了极显著性水平（0.01水平），从而可写出logistic回归模型为：



模型说明了，普通车辆更容易出现理赔保单，另外保险人年龄超过25岁更容易出现理赔保单。下面的优势比也说明了同样的问题。

优势比及置信区间：

Odds Ratio Estimates

Point 95% Wald

Effect Estimate Confidence Limits

type 1.421 1.283 1.573

age 0.664 0.605 0.728

非线性回归

现实世界中常见为非线性关系。常用的处理的方法有：

（1）将非线性关系转化为线性关系，利用线性回归方法进行处理。

（2）非线性关系不能转化为线性关系时，利用迭代法进行求解（计算方法中的利用）

（3）若非线性的函数关系未知，则可用多项式回归来拟合曲线。

## 1可变换成线性的非线性回归

若非线性模型具有如下形式，均可线性化

其中，而则分别化为新的因变量、线性回归参数和自变量，即可归结为线性回归模型来解。常见的可线性化的非线性模型。

**表1 典型的函数及线性化方法**

|  |  |  |
| --- | --- | --- |
| 函数名称 | 函数表达式 | 线性化方法 |
| 双曲线函数 |  |  |
| 幂函数 |  |  |
| 指数函数 |  |  |
|  |  |
| 对数函数 |  |  |
| S型函数 |  |  |
| 生长曲线 (logistic) 模型 | *yt* = |  |
| 龚伯斯（Gompertz）曲线 | *yt* = |  |

注： 生长曲线 (logistic) 模型 ，一般，特别地当时，有：，其中*b* = 。

*a* > 0情形的图形分别见图1和图2。美国人口统计学家Pearl和Reed广泛研究了有机体的生长，得到了上述数学模型。生长模型（或逻辑斯谛曲线，Pearl-Reed曲线）常用于描述有机体生长发育过程。其中*k*和0分别为*yt*的生长上限和下限。= *k*, = 0。*a*, *b* 为待估参数。曲线有拐点，坐标为（,），曲线的上下两部分对称于拐点。

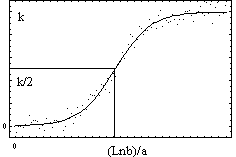
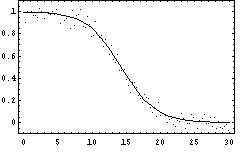
 

图1 *yt* = *k* / (1 +) 图2 *yt* = *k* / (1 +)

为能运用最小二乘法估计参数*a*, *b*，必须事先估计出生长上极限值*k*。线性化过程如下。当*k*给出时，作如下变换，

*k*/*yt* = 1 + ，移项， *k*/*yt* - 1 =  取自然对数，*Ln* ( *k*/*yt* - 1) = *Lnb* - *a t* + *ut* 令*yt*\* = *Ln* ( *k*/*yt* - 1), *b*\* = *Lnb*, 则*yt*\* = *b*\* - *a t* + *ut*

此时可用最小二乘法估计*b*\*和*a*。

龚伯斯（Gompertz）曲线

英国统计学家和数学家最初提出把该曲线作为控制人口增长的一种数学模型，此模型可用来描述一项新技术，一种新产品的发展过程。曲线的数学形式是，

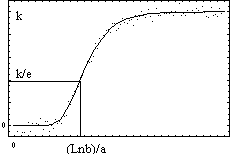


图3 *yt* =

曲线的上限和下限分别为*k*和0，= *k*, = 0。*a*, *b* 为待估参数。曲线有拐点，坐标为（,），但曲线不对称于拐点。一般情形，上限值*k*可事先估计，有了*k*值，龚伯斯曲线才可以用最小二乘法估计参数。线性化过程如下：当*k*给定时，

*yt* / *k* = ， *k*/*yt* = 

*Ln* (*k*/*yt*) = ， *Ln*[*Ln*(*k*/*yt*)] = *Lnb* - *a t*

令*y*\*= *Ln*[*Ln*(*k*/*yt*)], *b*\* = *Lnb*，则

*y*\* = *b*\* - *a t*

上式可用最小二乘法估计*b*\* 和 *a*。

Cobb-Douglas生产函数

下面介绍柯布−道格拉斯（Cobb-Douglas）生产函数。其形式是*Q* = *k Lα C* 1*- α*

其中*Q*表示产量；*L*表示劳动力投入量；*C*表示资本投入量；*k*是常数；0 < *α* < 1。这种生产函数是美国经济学家柯布和道格拉斯根据1899-1922年美国关于生产方面的数据研究得出的。*α*的估计值是0.75，*β*的估计值是0.25。更习惯的表达形式是

*yt* =

上式两边同取对数，得：

*Lnyt* = *Lnβ*0 + *β*1 *Lnxt* 1+ *β*2 *Lnxt* 2 + *ut*

取 *yt\** =*Lnyt*, *β*0*\** = *Ln β*0, *xt* 1*\** = *Ln xt* 1, *xt* 2*\**= *Ln xt* 2，有

*yt*\*= *β*0*\* +β*1 *xt* 1\* + *β*2 *xt* 2\* + *ut*

上式为线性模型。用OLS法估计后，再返回到原模型。若回归参数

*β*1 + *β*2 = 1，称模型为规模报酬不变型（新古典增长理论）；

*β*1 + *β*2 > 1，称模型为规模报酬递增型；

*β*1 + *β*2 < 1，称模型为规模报酬递减型。

对于对数线性模型，*Lny* = *Lnβ*0 + *β*1 *Lnxt*1+ *β*2 *Lnxt*2 + *ut* ，*β*1和*β*2称作弹性系数。以*β*1为例，

*β*1 = = = = 

可见弹性系数是两个变量的变化率的比。注意，弹性系数是一个无量纲参数，所以便于在不同变量之间比较相应弹性系数的大小。

## 2多项式回归

对观测数据（*xt*，*yt*）（*t*= 1，…，*N*），多项式回归模型为

，*t*=1,2,,*N*

令

，，，

则模型可表示为



当*X*列满秩时，由前面的讨论知，其最小二乘估计为



由此即可求得其多项式回归方程。但由于的计算既复杂又不稳定，故我们一般采用正交多项式法来进行多项式回归。

可用SAS的REG过程或者GLM过程进行回归计算。

## 3不可变换成线性的非线性回归分析

假设因变量*y*与自变量（*x*1，*x*2,…，*xp*）之间满足非线性模型，其中 为未知参数，*F*为已知表达式，仅未知的非线性函数，*ε* 为误差项。

现将观察数据， *t*=1,2,,*N*代人该非线性函数得非线性回归模型

其中为*y*的观察向量，为系数，*E* =为观察误差向量，非线性回归分析就是利用最小二乘准则来估计回归系数，即求 使得残差平方和



在  处达到最小。

非线性回归分析一般来用数值迭代法来进行，其共同特点是：由选定的初值出发，通过逐步迭代

即选择适当的步长*t* ( >0 ) 及确定搜索方向向量Δ＝（Δ1，Δ2，…，Δm），使得

再由取代，重复上述迭代过程，直至 *Q*(*β*)可认为达到最小值为止，即可将所得的作为其最小二乘估计，从而得到非线性回归方程



**下降方向和步长的选择**

首先考察的梯度向量（即导数）



其中为*F*的梯度矩阵。

现考虑一元函数，它从出发以 Δ为方向的射线上取值。由复合求导公式得



可以证明，当 *d*<0 时，在以 Δ为方向向量的射线上可以找到，使得。我们将满足 *d*<0 的Δ称为**下降方向**，Bard于1974年给出了Δ为下降方向的充要条件为，其中*P*为对称正定阵，由此我们可得下降算法的迭代公式为

其中*P*为任意正定阵，*G*为*F*的梯度，*t*为满足的正实数，即步长。

如何计算Δ以便修改参数向量有五种常用的非线性回归迭代方法：高斯-牛顿法（Gauss-Newton）、最速下降法（梯度法，Gradient）、牛顿法（Newton）、麦夸特法（Marquardt）、正割法（DUD）。以下我们介绍其中高斯-牛顿法。

**Gauss－Newton法**

首先选取的一初始近似值，令，则只要确定Δ的值即可确定。为此，考虑在处的Taylor展开式，并略去二次以上的项得



其中为*F*的梯度。此时其残差平方和



由，得其Δ的正则方程为

故

对给定精度、 ，当或时，即得的最小二乘法估计；否则用所得的代替，重复上述步骤，直至或*Q*(*β*)满足精度要求为止。该法称为Gauss－Newton法，其一般迭代公式为

其中：Δ为的解，*ti*为的最小值点。

Gauss-Newton法在初值选取适当，且可逆时非常有效，但在其他情形，其求解较为困难，对此，Marguardt对Δ的正则系数阵作适当修正，得到了改进算法。

## 4nlin非线性回归过程

proc nlin过程

proc nlin采用最小误差平方法（Least Squares Method）及迭代估计法（Iterative Estimation Method）来建立一个非线性模型。一般而言，用户必须自定参数的名字、参数的初始值（starting va1ue）、非线性的模型与迭代估计所用的准则。若用户不指明，则nlin 程序自动以高斯-牛顿迭代法（Gauss-Newton iterative procedure）为估计参数的方法。另外此程序也备有扫描（Grid search）的功能来帮助读者选择合适的参数启动值。由于非线性回归分析十分不易处理，nlin程序不保证一定可以算出符合最小误差平方法之标准的参数估计值。

nlin过程一般由下列语句控制：

|  |
| --- |
| **proc nlin data=数据集 <选项1>;** |
| **parameters 参数名=数值 ；** |
| **model 因变量＝表达式 </选项列表>；** |
| **bounds 表达式 ；** |
| **der.参数名｛．参数名｝= 表达式；** |
| **id 变量列表；** |
| **output out=数据集 </选项列表>；** |
| **by 变量列表；** |
| **run ;** |

其中，parameters语句和model语句是必需的，而其余语句供用户根据需要选择。

proc nlin语句中**选项1**的主要选择项。

***outest*＝数据集名——指定存放参数估计的每步迭代结果的数据集名。**

***best*＝*n*——要求过程只输出网格点初始值可能组合中最好的*n*组残差平方和。**

***method*＝gauss | marquardt | newton| gradient| dud |——设定参数估计的迭代方法。缺省时为gauss，除非没有der.语句。**

***eformat*——要求所有数值以科学记数法输出。**

***nopoint*——抑制打印输出。**

***noinpoint*——抑制迭代结果的输出。**

**parameters（parms）语句：**

用于对所有参数赋初值，参数之间以空格分隔。例如，parms b0=0 b1=1 to 10 b2=1 to 10 by 2 b3=1,10,100；

**model语句：**

表达式可以是获得数值结果的任意有效SAS表达式。这个表达式包括参数名字、输入数据集中的变量名以及在nlin过程中用程序设计语句创建的新变量。例如，model y=b0\*(1－exp(-b1\*x))；

**bounds语句：**

用于设定参数的约束，主要是不等式约束，约束间用逗号分隔。例如，bounds a<=20，b>30，1<=c<=10；

**der.语句：**

除非在proc nlin语句中指明所用的迭代法是dud，使用选择项method＝dud，否则der.语句是必需的。der.语句用于计算模型关于各个参数的偏导数，相应的格式为：

一阶偏导数 der．参数名＝表达式 ；

二阶偏导数 der．参数名．参数名＝表达式；

例如，对于model y=b0\*(1－exp(-b1\*x))；der.语句的书写格式为der.b0=1－exp(-b1\*x)；der.b1=b0\*x\*exp(-b1\*x)；

对于多数算法，都必须对每个被估计的参数给出一阶偏导数表达式。对于newton法，必须给出一、二阶偏导数表达式。例如，二阶偏导数表达式为，der.b0.b0.=0；der.b0.b1=x\*exp(-b1\*x)；der.b1.b1=-der.b1\*x；

**output语句**

用于把一些计算结果输出到指定的数据集中。有关的关键字及其意义如表2所示：

**表2 nlin过程中output语句的关键字**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 关键字 | 意义 | 关键字 | 意义 | 关键字 | 意义 |
| *predicted* | *p* | 预测值 | *stdp* | clm的标准差 | *u*95 | 95％cli上限 |
| *residual* | *r* | 残差 | *stdr* | 残差的标准差 | *l95* | 95％cu下限 |
| *parms* | 参数估计值 | *l95m* | 95％clm下限 | *student* | 学生氏残差 |
| *sse* | *ess* | 残差平方和 | *u95m* | 95％clm上限 | *h* | 杠杆点统计量hi |

## 5实例分析

**例34.1**  负指数增长曲线的非线性回归。根据对已有数据的XY散点图的观察和分析，发现*Y*随*X*增长趋势是减缓的，并且*Y*趋向一个极限值，我们认为用负指数增长曲线来拟合模型较为合适。程序如下：

|  |
| --- |
| **data** expd ; |
| input x y @@;; |
| cards; |
| 020 0.57 030 0.72 040 0.81 050 0.87 060 0.91 070 0.94 |
| 080 0.95 090 0.97 100 0.98 110 0.99 120 1.00 130 0.99 |
| 140 0.99 150 1.00 160 1.00 170 0.99 180 1.00 190 1.00 |
| 200 0.99 210 1.00 |
| ; |
| **proc** **nlin** data=expd best=**10** method=gauss; |
| parms b0=**0** to **2** by **0.5** b1=**0.01** to **0.09** by **0.01**; |
| model y=b0\*(**1**-exp(-b1\*x)); |
| der.b0=**1**-exp(-b1\*x); |
| der.b1=b0\*x\*exp(-b1\*x); |
| output out=expout p=ygs ; |
| **run**; |
| goptions reset=global gunit=pct cback=white border |
| htitle=**6** htext=**3** ftext=swissb colors=(back); |
| **proc** **gplot** data=expout; |
| plot y\*x ygs\*x /haxis=axis1 vaxis=axis2 overlay; |
| symbol1 i=none v=plus cv=red h=**2.5** w=**2**; |
| symbol2 i=join v=none l=**1** h=**2.5** w=**2**; |
| axis1 order=**20** to **210** by **10**; |
| axis2 order=**0.5** to **1.1** by **0.05**; |
| title1 'y=b0\*(1-exp(-b1\*x)'; |
| title2 'proc nlin method=gauss'; |

程序说明：

1. nlin过程中使用选项method=gauss，即指定用高斯-牛顿迭代算。

2.parms语句设置了参数的初始值，其中取0，0.5，1，1.5，2共5个值，取0.01，0.02，…，0.09共9个值，所有共有5×9=45个组合初始值。

3．选项best=10要求输出残差平方和最小的前10种初值的组合

4. der.语句表示两个一阶偏导数表达式。由模型知对参数和的一阶偏导数有：



5.output语句输出一个新数据集expout，包括原数据集和非线性回归模型的预测值ygs。

6. gplot过程的主要作用是绘制输出数据集expout中的原始数据的散点图及回归曲线的平滑线。



输出结果分析：首先列表输出了10个最好的迭代初始值组合，按回归模型的残差平方和从小到大排列。因此，最好的迭代初始值为B0=1.000000，B1=0.040000，此时回归模型的ESS=0.001404，与其他初始值组合的ESS相比是最小的。紧接着输出迭代过程分析表，从这个最好迭代初始值开始经过4次迭代误差平方和的变化就满足收敛准则，停止迭代，例如，经过第1次迭代，B0值从1修正为0.996139，B1值从0.040000修正为0.041857，而ESS从0.001404减小到0.000580，而从第2次迭代开始，B0 的值只有在小数点四位以后才有变化，B1的值只有在小数点五位以后才有变化，而ESS值几乎不变了，因此是满足了收敛性标准而停止迭代的。方差分析表给出，回归平方和为17.671723189，残差平方和为0.000576811，总平方和为17.672300000，总偏平方和为0.243855000。参数估计表给出了，B0和B1的渐近估计值，因此得到的非线性回归模型为

同时还给出B0和B1参数估计的渐近有效标准差（Asymptotic Std. Error）和渐近95%置信区间（Asymptotic 95 % Confidence Interval）。最后还给出了参数的渐近相关阵（Asymptotic Correlation Matrix）。