相关与回归分析

**一、概述**

1、分类

按自变量的多少分为一元回归分析、多元回归分析。

按自变量与多变量之间的关系分为线性回归、非线性回归。

一元线性回归：回归分析中只有一个因变量和一个自变量，且二者的关系可用一条直线近似表示。

多元线性回归分析：回归分析中有两个或者两个以上的自变量，且因变量与自变量之间的关系是线性关系。

2、相关与回归

相关分析研究的是现象之间是否相关、相关的方向和密切程度，一般不区别自变量或因变量。

而回归分析则要分析现象之间相关的具体形式，确定其因果关系，并用数学模型来表现其具体关系。一般来说，回归分析是通过规定因变量和自变量来确定变量之间的因果关系，建立回归模型，并根据实测数据来求解模型的各个参数，然后评价回归模型是否能够很好的拟合实测数据；如果能够很好的拟合，则可以根据自变量作进一步预测。

3、统计量

R2，F检验值和T检验值。R2又称为方程的确定性系数（coefficient of determination），表示方程中变量X对Y的解释程度。R2取值在0到1之间，越接近1，表明方程中X对Y的解释能力越强。通常将R2乘以100％来表示回归方程解释Y变化的百分比。F检验是通过方差分析表输出的，通过显著性水平（significant level）检验回归方程的线性关系是否显著。一般来说，显著性水平在0.05以下，均有意义。当F检验通过时，意味着方程中至少有一个回归系数是显著的，但是并不一定所有的回归系数都是显著的，这样就需要通过T检验来验证回归系数的显著性。同样地，T检验可以通过显著性水平或查表来确定。

**回归分析 🡪 回归诊断 🡪 残差分析 🡪 结果分析**

**二、相关分析与回归过程**

**1、CORR过程**

Corr过程计算变量间的Pearson相关系数。corr过程一般由下列语句控制：

|  |
| --- |
| proc corr data=数据集 </选项列表> **;** |
| var 变量列表； |
| with 变量列表 ； |
| partial 变量列表； |
| weight 变量； |
| freq 变量； |
| by 变量； |
| run ; |

**proc corr**语句中的<选项列表>

Pearson——计算通常的Pearson积矩相关，是缺省值。

outp＝数据集名——产生含有Pearson相关的一个新数据集。

cov——计算协方差–方差矩阵。

sscp ——计算各变量未校正的平方和与交叉积和。

csscp ——计算各变量校正的平方和与交叉积和。

nocorr——抑制Pearson相关的计算及输出。

nomiss ——将带有某一变量缺失值的观测值从所有计算中除去。

proc corr过程中的主要语句**。**

var语句——指明要进行相关分析的变量名，缺省时，为在所有数值变量间计算相关系数。

with语句——为了得到变量的特殊组合的开关，该语句和var联合使用。Var语句列出的变量放在输出阵的上方，而with语句列出的变量竖排在相关阵的下方。

例： **proc** **corr** data=test;

var a b;

with x y z;

**run**;

\*输出以下变量组合: x和a、x和b、y和a、y和b、z和a、z和b;

partial语句——指明求偏相关时的偏变量名，同时激活nomiss选择项。结果输出包括简单统计数和相关系数及显著性。设定partial变量时进行偏相关分析。

weight语句——规定一个数值变量，作为观测值的权重。

freq语句——规定一个数值变量，它的值表示工作数据集中观测值出现的频数。

by语句——指定的变量值来分组处理某数据集。

**实例**

试验测定迟熟早籼广陆矮4号某年5月5日至5月8日播种时（每隔10天播一期），播种至齐穗的天数（x）和播种至齐穗的总积温（y，日•度）的关系，数据列于下表， ，分析播种至齐穗的天数与总积温两者之间的相关关系。

|  |  |
| --- | --- |
| 播种至齐穗的天数（x） | 播种至齐穗的总积温（y，日•度） |
| 70 | 1616.6 |
| 67 | 1610.9 |
| 55 | 1440.0 |
| 52 | 1400.7 |
| 51 | 1423.3 |
| 52 | 1471.3 |
| 51 | 1421.8 |
| 60 | 1547.1 |
| 64 | 1533.0 |

**data** abc;

input x y @@;

cards;

70 1616.6 67 1610.9 55 1440.0 52 1400.7

51 1423.3 52 1471.3 51 1421.8 60 1547.1

64 1533.0

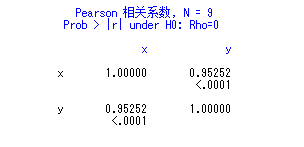
;

**proc** **corr** data=abc;

var x y;

**run**;

输出结果：



由输出结果得x与y的相关系数为0.95252，相关程度的显著水平小于0.0001，十分显著。说明播种至齐穗的总积温（y，日•度）对迟熟早籼广陆矮4号的抽穗有十分显著的影响。

**2、REG过程**

REG是一个通用的回归过程。它采用最小二乘法拟合线性回归模型。它还提供多种选择“最佳”回归模型的方法，是一个应用最广泛的回归过程。

回归分析的主要内容：

（1）由给定的数据确定回归方程。估计回归方程中的未知参数（常用最小二乘法）。

（2）检验回归方程的可信程度。

（3）如果是多自变量影响一个因变量，则可以找出对因变量影响较显著的自变量，剔除不显著的。（常用逐步回归法，向前回归法和向后回归法）。

（4）利用建立好的回归方程进行预测和控制。

REG过程语句格式：

|  |
| --- |
| **PROC REG 选择项1 ;**  **MODEL 因变量=自变量 / <选择项2> ;**  **Var 变量;**  **By 变量;**  **Freq 变量;**  **Output out=sas数据集 关键词=变量名1 变量名2;**  **Plot y轴变量\*x轴变量 /选择项3;**  **RUN;** |

* + 1. PROC REG语句中的<选项列表>
* OUTEST=SAS数据集——将有关模型的参数输出到指定的SAS数据集中
* OUTSSCP=SAS数据集——将相关矩阵输出到指定的SAS数据集中
* ALL——屏幕输出所有内容
* NOPRINT——不在屏幕输出任何内容
  + 1. MODEL语句中的选项

该语句定义建模用的因变量、自变量、模型的选择及结果输出的选择。

**与模型有关的选项有：**

* + - 1. SELECTION=选择合适的建立模型方法
* SELECTION=FORWARD SLENTRY=显著性水平

前进法（FORWARD）：对每一个尚不在方程内的自变量按一定的显著性水平，根据其一旦进入模型后对模型的贡献大小逐步引入方程，直至再没有对模型有显著贡献的自变量。

缺省SLENTRY=0.5

* SELECTION=BACKWARD SLSTAY=显著性水平

后退法（BACKWARD）：先建立包含全部变量的模型，然后按一定的显著性水平从模型中逐步剔除变量。

缺省SLSTAY =0.1

* SELECTION=STEPWISE SLENTRY =入选水平 SLSTAY=剔除水平

逐步法（STEPWISE）：按前进法进入变量，再对模型内所有变量检验，看是否有因新变量引入而对模型的贡献变得不显著的变量，若有就剔除，若无则保留，直至方程内所有的变量均显著，显然逐步法有两个水平，即选入水平和剔除水平，而且剔除水平应低于选入水平。

缺省SLENTRY =0.15 SLSTAY =0.1

在上述三种方法的使用中，若要求打印出每一次选入或剔除变量进行模型拟合时的所有统计量，可以加选DETAILS。

* + - 1. NOINT——表示拟合无常数项（截距）的回归模型

与屏幕输出有关的选项有：

* CORRB——输出参数估计的相关阵
* STB——输出标准化偏回归系数矩阵
* P——输出个体观测值、预测值及残差。若已选了CLI、CLM、R，则无需该选项
* R——输出每个个体观测值、残差及标准误差
* CLM——输出每个观测值因变量期望值的95%的上、下限
* CLI——输出每个个体观测值的95%的上、下限

与残差分析有关的选项有：

* VIF——输出变量间相关性的方差膨胀系数（Variance Inflation Factor）,VIF越大，说明由于共线性存在，使方差变大。
* COLLIN——输出条件数（Condition index）,它表示最大的本征性与每个自变量本征值之比的平方根。一般情况下，条件数越大越可能存在共线性。
* TOL——表示共线性水平的容许值，TOL（Tolerance Value）越小说明其可用别的自变量解释的部分多，自然可能与别的自变量存在共线性关系。
* DW——输出Durbin-Watson统计量
  + 1. 其他选择语句

注意，这部分的语句可以在REG过程被被激活后，以交互式方式运行。

* OUTPUT语句——建立SAS的输出结果数据集

语句格式为：

OUTPUT OUT=SAS数据集名 关键字名=输出数据集中的变量名

其中关键字名为需要的统计量名，它们有P（预测值）、R（残差）、L95M（期望值的95%的下限）、U95M（期望值的95%的上限）、L95（个体预测值的95%的下限）、U95（个体预测值的95%的上限）、STDP（期望值的标准误差）、STDR（残差的标准误差）、STDI（预测值的标准误差）、STUDENT（学生化残差）、COOKD（COOK氏D值）

* PLOT语句——绘制两变量的散点图

语句格式为：

PLOT X\*Y / 选项

* ADD 变量名列表——向模型中增加变量
* DELETE 变量名列表——删除原拟合模型中的有关变量
* REFIT——重新拟合模型
* PRINT——输出有关模型的相关信息

例1-2：

为测定某城市农田肥力水平土壤有机质与土壤含氮量的关系，从12块天地取土样，测得结果数据如表1-1所示，目的是建立有机质含量与全含氮量之间的回归方程。处理数据，找到两个变量之间的线性方程。

表1-1 某城市农田肥力水平土壤有机质与土壤含氮量的关系数据

|  |  |  |
| --- | --- | --- |
| Sn（土壤号） | Humus（有机质含量%） | N（全氮含量%） |
| 1 | 3.92 | 0.231 |
| 2 | 3.98 | 0.212 |
| 3 | 1.85 | 0.216 |
| 4 | 2.22 | 0.135 |
| 5 | 3.51 | 0.204 |
| 6 | 2.84 | 0.17 |
| 7 | 4.11 | 0.247 |
| 8 | 1.55 | 0.133 |
| 9 | 2.37 | 0.182 |
| 10 | 3.6 | 0.182 |
| 11 | 1.1 | 0.08 |
| 12 | 1.37 | 0.09 |

**data** test;

input sn humus n @@;

cards;

1 3.92 0.231 2 3.98 0.212

3 1.85 0.216 4 2.22 0.135

5 3.51 0.204 6 2.84 0.17

7 4.11 0.247 8 1.55 0.133

9 2.37 0.182 10 3.6 0.182

11 1.1 0.08 12 1.37 0.09

;

**run**;

option ps=**40** ls=**100**;

\*plot过程绘制humus与n两变量的散点图;

**proc** **plot** data=test;

plot humus\*n;

**run**;

\*使用corr相关分析过程分析两变量的相关性;

**proc** **corr** data=test;

var humus n;

**run**;

\*用REG过程建立直线回归模型;

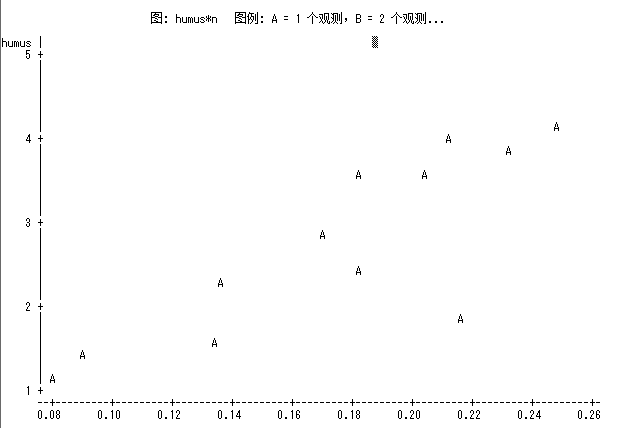
**proc** **reg** data=test;

model humus=n;\*改为model humus =n/noint r clm cli;

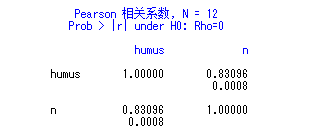
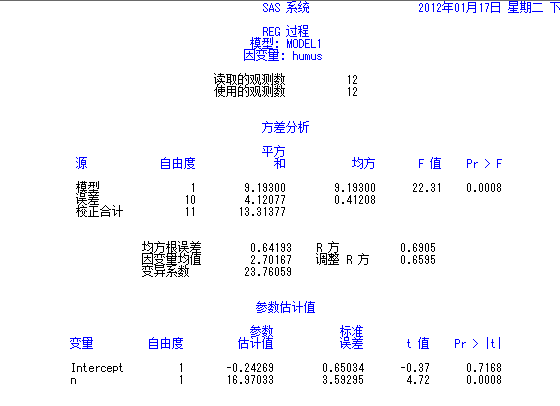
**run**;

**quit**;

输出结果：



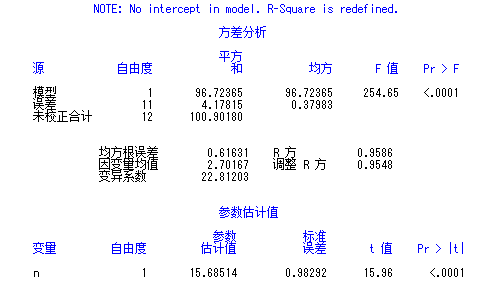
上图是用plot过程对原始数据humus与n两变量建立的散点图。由散点图可以初步验证两变量之间存在一定的线性关系。

由corr相关过程的输出可看到：变量humus和n的相关系数达0.83096.Pr>F = 0.0008小于0.05，所以humus与n正相关，可做回归分析。

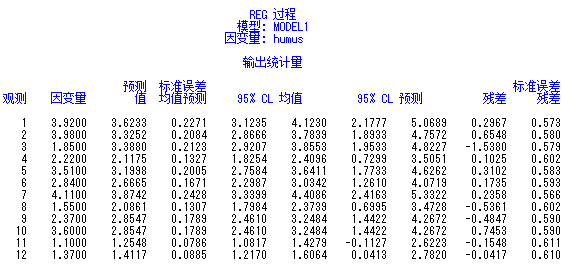
由REG过程分析得到的回归模型是显著的，Pr>F=0.0008，线性方程拟合优度为R方=0.6905（即此方程能解释humus变化的69.05%），。其中变量n在模型中是显著的，Pr>F = 0.0008.但截距项Intercept在此方程中不显著，可以将截距项取掉，再重新拟合。

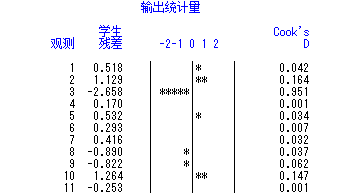
（残差是指观测值与预测值（拟合值）之间的差，即是实际观察值与回归估计值的差。利用残差所提供的信息，来考察模型假设的合理性及数据的可靠性称为残差分析。）

MODEL语句改为 model humus =n/noint r clm cli; 之后REG的输出



REG重新拟合的回归模型显著性提高，统计量F值为254.65，Pr > F<0.0001，模型拟合优度R方为0.9586，其中变量n在回归模型中的显著性<0.0001，说明截距项去掉后的拟合结果更好。回归方程为： humus=15.685\*n;



  
Cook’s D 统计量，用来度量因变量每个观测值对预测值的影响大小，此值越大，表明该观测值对预测值的影响越大，以此来发现原始数据中的强影响点。

从输出残差图看出，第三个观测点的残差较大，应检查原始数据是否有误，或试着去掉此点重新拟合模型，考察数据的变化情况。第三个观测点的COOK‘D值也较大，说明此点是一个强影响点。需要慎重考虑是否去掉此点进行重新拟合。

尝试删除第三个观测点，再进行回归模型拟合

**data** test1;

set test;

if sn=**3** then delete;

**run**;

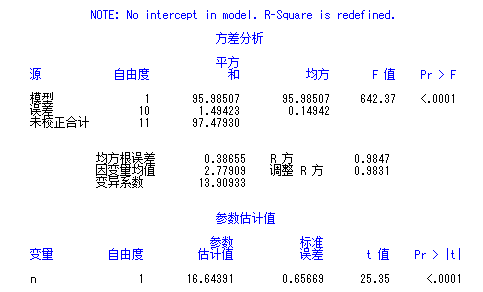
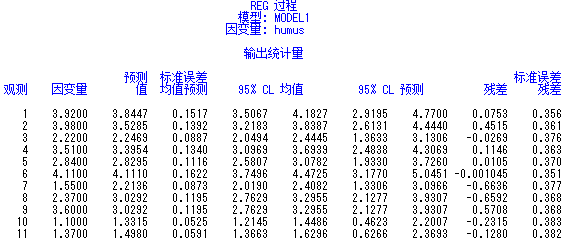
**proc** **reg** data=test1;

model humus= n/noint r cli clm;

**run**;

**quit**;

输出结果：

  
去掉第三个观测之后，F统计量的值为642.37，Pr >F < 0.0001，回归模型的显著性提高，模型拟合优度为0.9847.回归方程为humus ^= 16.64391。从统计学上表明用不含截距项的直线回归方程描述给定的数据是合适的。

去掉第三个观测点之后，其他点的残差出现增大的现象。可以初步推断，第三个观测值的测试结果对建立的模型有一定的影响。建议实际允许的情况下，重新测量第三个观测的值，以使建立的模型更加精确。

小结：corr相关过程用于分析中变量间的地位是平等的，不用分因变量和自变量，侧重分析变量之间的种种相关特征。

**三、分析**

共线性诊断：自变量之间如果有较强的相关关系，则较难找到理想的回归方程

1、回归诊断

点诊断：若干个观测点与多数观测点偏离很远或因过失误差，对回归方程的质量产生较坏的影响

1）用条件数和方差分量来进行共线性诊断

利用SAS中MODEL语句中的COLLIN 或COLLINOINT选项。二者都给出条件数和方差分量。

（1）条件数

条件数是最大特征根与每个特征根比值的平方根，其中最大条件数K称为矩阵|x’x|的条件数。条件数大，说明数据矩阵有较强的共线性，使结果不稳定，甚至使远离试验点的一些估计值或预测值毫无意义。

（2）方差分量

对大的条件数且同时又两个以上的方差分量超过50%，就意味着这些变量间有一定程度的相关性。

2）用容忍度和方差膨胀因子

（1）容忍度（TOL）

入选变量的TOL=1-R2 ，此处的是把该变量当做因变量时，其他自变量能解释的部分，tol表示此自变量不能被其他自变量解释的部分，因此这个值越小，其他自变量含有此自变量的信息越多，则共线性越强

（2）方差膨胀因子（VIF）

vif则为容忍度的倒数，因此它的值约大，则共线性越强。一般当VIF>5或10时，就认为存在严重的多重共线性。

3）用学生化残差对观测点中的强影响点进行诊断

用COOK’D统计量、HI统计量、STUDENT统计量等诊断那些点对因变量的影响大。其中最便于判断的是学生化残差STUDENT统计量，当STUDENT统计量的值大于2时，所对应的观测点可能是异常点。可认真核对数据看此观测点是否认为失误导致异常，并在剔除异常点之后再做回归分析。当COOK’S统计量大于2时，该点也属于强影响点，处理强影响点时要谨慎。

2、用各种回归方法筛选变量

判断一个回归模型是否较优，可以从以下两个方面考虑：一是整个回归模型及模型中各回归参数在统计学上有显著性意义，在专业上（特别是因变量的预测及回归方程的精度）有实际意义；二是在包含相同或相近信息的前提下，回归方程中的变量越少越好。

**实例**

数据来源：《植物营养研究方法》（北京农业大学出版社出版）。数据为土壤中三种形态的磷含量和植物吸磷量，数据见下表。X1为每种土壤同酸性氟化铵溶液浸提的无机磷含量ug/g，x2为用K2CO3溶液A法浸提的有机磷含量ug/g，x3为用K2CO3溶液B法浸提的有机磷含量ug/g，y为玉米幼苗吸磷量ug/g，研究玉米幼苗吸磷量（y）与x1、x2、x3之间的关系，分析哪些因素对幼苗吸磷量y的影响最大。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **X1** | **X2** | **X3** | **Y** | **X1** | **X2** | **X3** | **y** |
| **0.4** | **53** | **158** | **64** | **12.6** | **58** | **112** | **51** |
| **0.4** | **23** | **163** | **60** | **10.9** | **37** | **111** | **76** |
| **3.1** | **19** | **37** | **71** | **23.1** | **46** | **114** | **96** |
| **0.6** | **34** | **157** | **61** | **23.1** | **50** | **134** | **77** |
| **4.7** | **24** | **59** | **54** | **21.6** | **44** | **73** | **93** |
| **1.7** | **65** | **123** | **77** | **23.1** | **56** | **168** | **95** |
| **9.4** | **44** | **46** | **81** | **1.9** | **36** | **143** | **54** |
| **10.1** | **31** | **117** | **93** | **26.8** | **58** | **102** | **168** |
| **11.6** | **29** | **173** | **93** | **29.9** | **51** | **124** | **99** |

**data** test;

input x1 x2 x3 y @@;

x11 = x1\*x1;

x22 = x2\*x2;

x33 = x3\*x3;

cards;

0.4 53 158 64 12.6 58 112 51

0.4 23 163 60 10.9 37 111 76

3.1 19 37 71 23.1 46 114 96

0.6 34 157 61 23.1 50 134 77

4.7 24 59 54 21.6 44 73 93

1.7 65 123 77 23.1 56 168 95

9.4 44 46 81 1.9 36 143 54

10.1 31 117 93 26.8 58 102 168

11.6 29 173 93 29.9 51 124 99

;

**run**;

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

model y =x1-x3/selection=backward;

model y =x1-x3/selection=stepwise;

model y =x1-x3/selection=stepwise noint;

model y =x1-x3 x11 x22 x33/selection=stepwise noint;

model y =x1-x3 x11 x22 x33/selection=backward noint;

**run**;

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

model y=x1 x2 x11 x22 /noint selection=stepwise p r cli influence

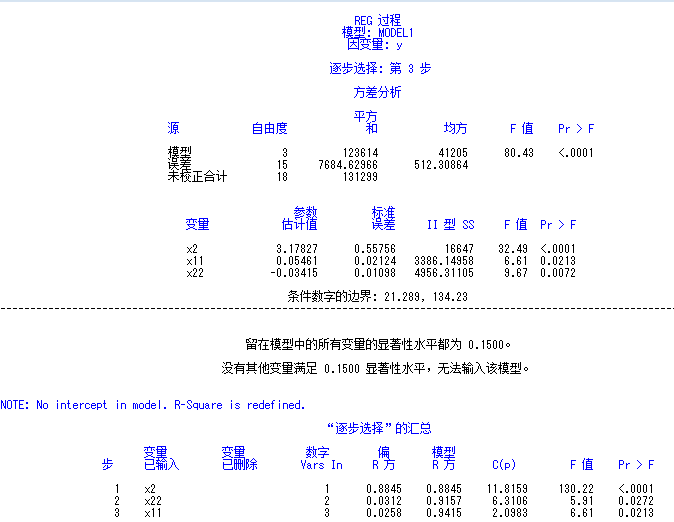
collin collinoint;

model y=x2 x11 x22 /noint p r cli influence collin collinoint;

**run**;

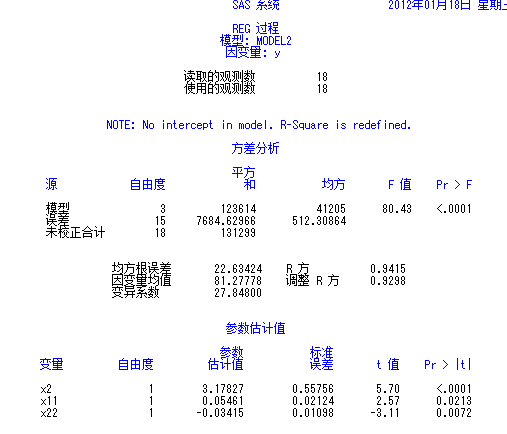
**quit**;

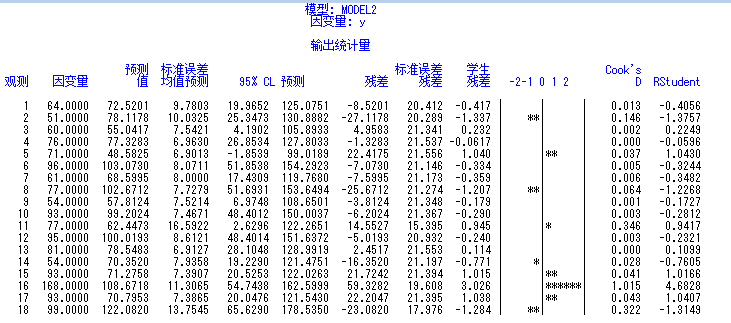
输出结果：

第一个REG过程输出结果太多，此处不列出，经过比较，我们选择对模型有显著作用的变量(x1、x2、x11、x22)，用第二个REG过程再进行筛选。

筛选出x2、x11、x22。无截距方程： y = 3.17827x2+0.05461x12-0.03415x22

再添加一条MODEL语句，对已筛选出的变量进行全回归法分析及预测和残差分析，根据此语句的结果进行最后的结论分析。



模型拟合优度为0.9415，x2、x11和x22在模型中的显著概率均表示为显著。F统计量为80.43，误差平方和为7684.62966，要说明此模型是否较好，仍需做残差分析。

由残差值和COOK’D统计量得出16号观测点为数据异常点。可将其去掉再进行残差分析，再看模型的精度是否会提高，若提高了，实际允许的情况下重新测量第三个数据点，以使模型更加精确。

**小结：**



**共线性诊断**

输出变量的特征值、条件数及相对于每个特征值这些故居的方差分解。

上图可得：第一个特征值的方差分解均分在x2、x11和x22上（即分在x1和x2两个变量上）。第二个特征值主要由x11承担。第三个特征值主要分解在x2和x22上，两个变量有较强的共线性。

3、直线回归分析：

通过拟合曲线方程来研究变量间的函数关系

常用对数函数，指数函数和幂函数。

步骤：先绘制散点图🡪观察各点趋势🡪选择适合的函数作变量变换。

（1）可直线化的拟合说明

**实例**

泵数据来源于2003年某研究人员所做的试验。观测11个水稻品种（03dh1、03dh2、…、03dh11）的各种性状：穗数x1 枝梗数x2、秕粒x3、200粒重y。每个水稻品种取5株。以5株为一个单位。研究水稻200粒重y与穗数x1、枝梗数x2、秕粒x3之间的关系，分析哪些因素对200粒重y的影响较大。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编号 | 株数 | 穗数x1 | 枝梗数x2 | 秕粒x3 | 200粒重y |
| 03dh1 | 5 | 14 | 16 | 59 | 5.87 |
| 03dh2 | 5 | 27 | 13 | 27 | 5.58 |
| 03dh3 | 5 | 31 | 11 | 94 | 5.83 |
| 03dh4 | 5 | 20 | 15 | 64 | 4.71 |
| 03dh5 | 5 | 24 | 14 | 167 | 5.59 |
| 03dh6 | 5 | 19 | 13 | 340 | 3.85 |
| 03dh7 | 5 | 30 | 13 | 40 | 5.52 |
| 03dh8 | 5 | 21 | 12 | 122 | 5.3 |
| 03dh9 | 5 | 29 | 13 | 90 | 5.65 |
| 03dh10 | 5 | 30 | 14 | 85 | 4.97 |
| 03dh11 | 5 | 41 | 13 | 120 | 5.31 |

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

model y=x1-x3 x11 x22 x33 /selection=stepwise;

model y=x1-x3 x11 x22 x33 /selection=b noint;

model y=x1-x3 x11 x22 x33 /selection=f noint;

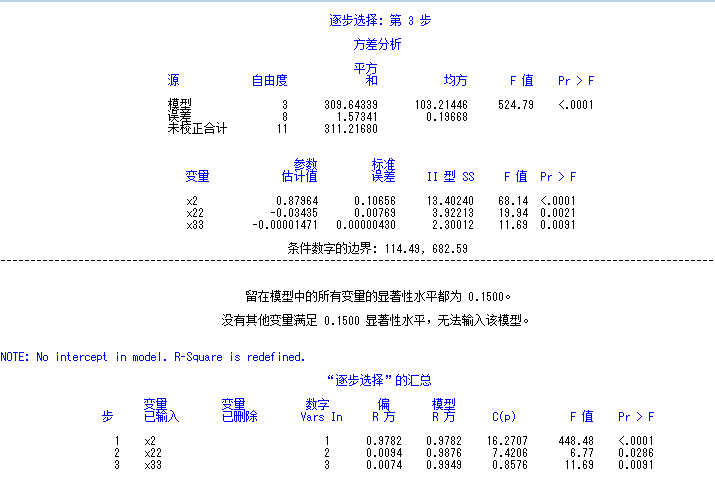
/\*经过三条模型筛选出变量x2,x22,x33，再加入下面的语句进行残差、共线性等分析\*/

model y=x1-x3 x11 x22 x33 /selection=stepwise noint r influence collin;

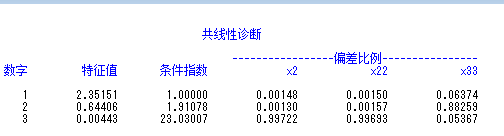
**run**;

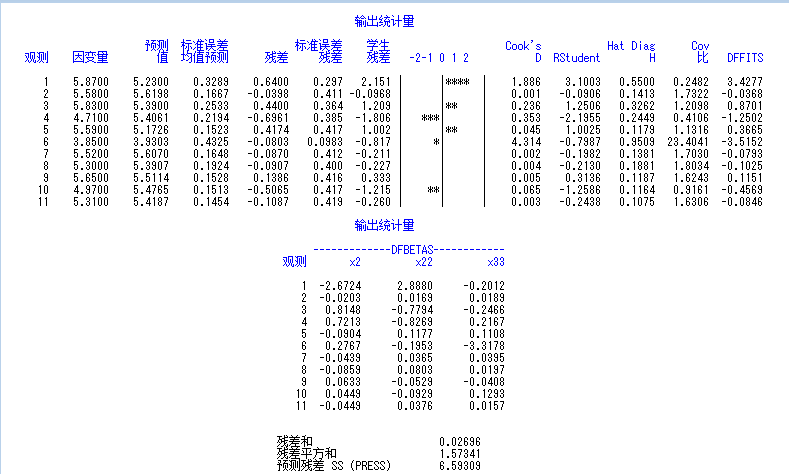
**quit**;





该模型的R方为0.9949，F值为524.79，误差平方和为1.57341，说明模型拟合得较好，但模型是否适用还要根据专业知识确定。模型为：0.87964x2-0.03435x22-0.00001471x32



对模型中的变量的共线性诊断，得x2和x22有较强的共线性，因为x22是x2的平方项。

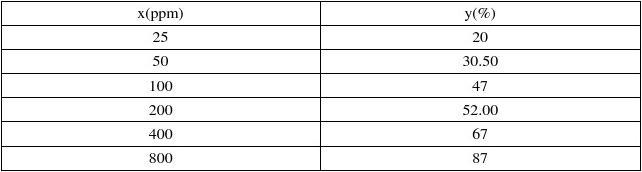
由RStudent值可看出，第一个观测点异常，检查是什么导致，是否可以剔除再重新拟合模型。第六个观测点的COOK’K统计量较大，是强影响点。

专业结论：通过对11个品种水稻的穗数、枝梗数、秕粒指标的研究，水稻200粒重y主要受x2枝梗数的影响，也受x3秕粒的影响，但影响不大，与x1穗数无关。

**小结：先筛选变量，选出对因变量具有较显著影响的变量。2. 拟合模型，判断好坏（是否去掉截距项、P值、COOK’S统计量、误差平方和）3. 做全回归分析和残差分析，找出异常的样本观测点，并检查是否可提出在拟合以提高精度。**

实例

农药的生测实验：农药的生测实验：农夫菊酯原药对甜菜夜蛾的生测，用的是点滴法。X指功夫原药的浓度（ppm）,y是四龄甜菜夜蛾的死亡率。研究并拟合甜菜夜蛾死亡率y与功夫原药浓度x之间的曲线方程.

******

**DATA** test;

input x y @@;

y1= log(y);

x1= log(x);

cards;

25 20 50 30.5 100 47 200 52.0 400 67 800 87

;

**run**;

**proc** **plot** ;

plot y\*x y\*x1 y1\*x y1\*x1;

**proc** **reg** ;

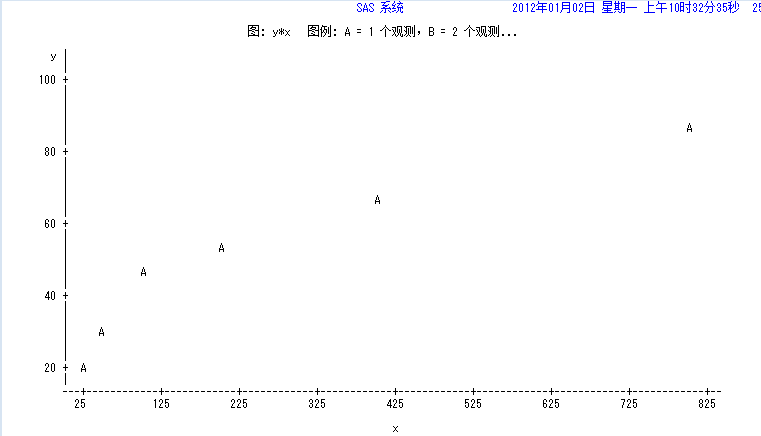
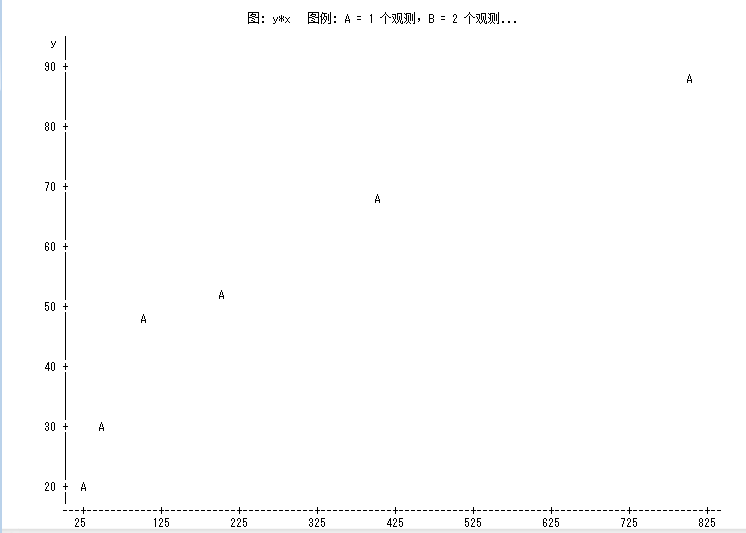
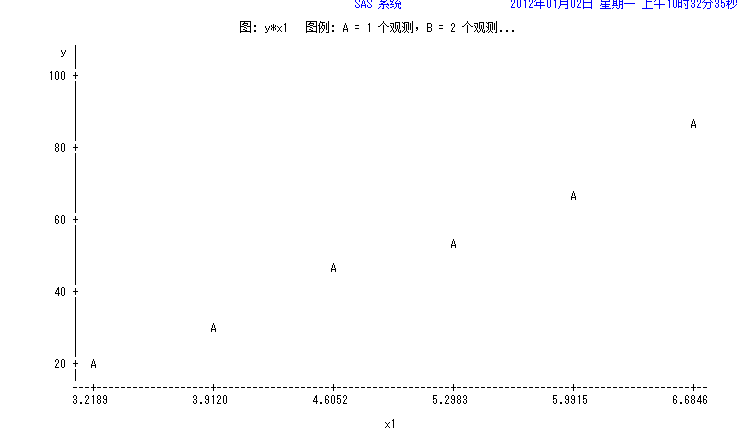
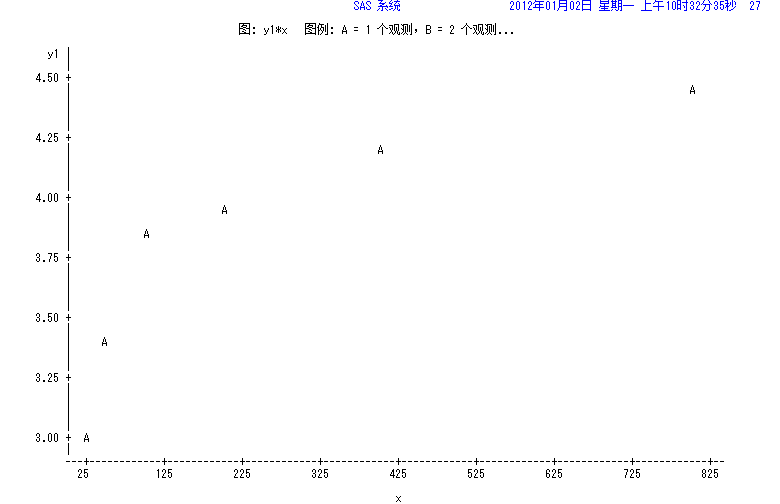
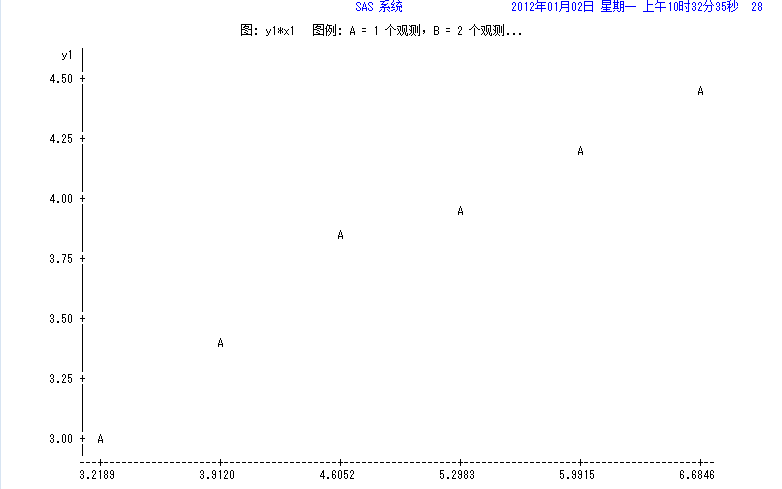
model y=x;

model y=x1;

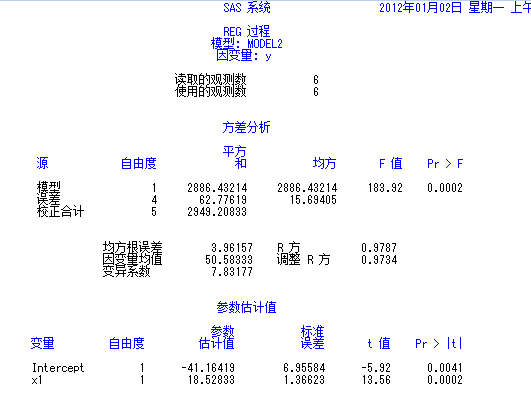
model y1=x;

model y1=x1;

**run**;

**quit**;    

用4个model语句建立了4个拟合模型，由图可得模型2和模型4拟合得较好，即y=a+b\*ln(x),x1=ln(x)和y1=a+b\*x1，y1=ln(y),x1=ln(x)。

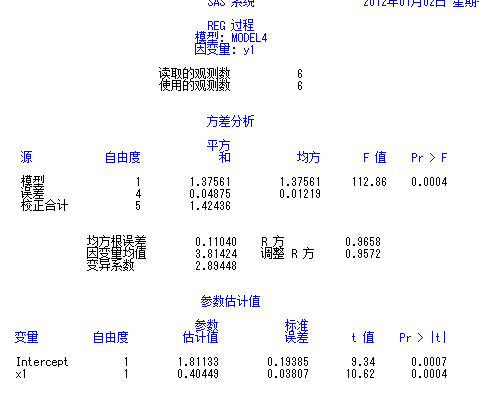


这是模型2拟合的输出，因为模型2的拟合相对于模型1来说，有较好的改善，所以此处不将模型1列出。模型的F值为189.92，P值=0.0002，十分显著，误差平方和为62.77619，R方为0.9787.截距项和x1的Pr>F都<0.01,对y的影响十分显著。此模型可表示为：

Y^=-41.16419+18.52833\*ln(X).

解释：四龄甜菜夜蛾的死亡率y与功夫原药的浓度（ppm）x呈线性关系。

模型3的效果较差，此处不列出。



模型4的拟合效果比模型2的拟合效果稍差。此模型可表示为：

Y^=EXP{1.81133+0.40449\*ln（X）}。

解释：四龄甜菜夜蛾的死亡率y与功夫原药的浓度（ppm）x呈幂函数关系。

可初步判断模型2最好，模型4次之，但仍需考察两模型各点的残差以及残差平方和来做进一步的判断。

**data** test;

t1=**0**; t2=**0**;

do i=**1** to **6**;

input x y @@;

x1=log(x); y1=log(y);

yhat1=-**41.16419**+**18.52833**\*x1; \*计算对数曲线的预测值;

resid1=y-yhat1; \*残差; t1=t1+resid1\*resid1; scrs1=t1; \*残差平方和;

yhat2=exp(**1.81133**+**0.40449**\*x1); \*计算幂函数曲线的预测值;

resid2=y-yhat2; t2=t2+resid2\*resid2; scrs2=t2;

output;

end; drop i t1 t2;

cards;

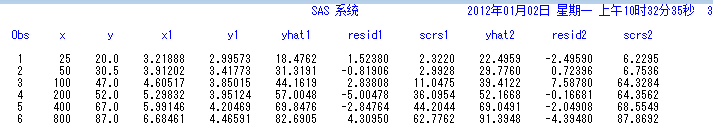
25 20 50 30.5 100 47 200 52.0 400 67 800 87

;

**proc** **print**;

**run**;

输出结果：



可以看出模型2拟合的较好，但第4和第6个样本点的残差值较大。用直线化拟合的方程的是否有实用价值，关键是看残差和残差平方和是否较小，同时看该方程的显著性和参数个数是否较少，即自由度，越小越好。

***小结：初步拟合曲线后得到的结果，还需要用因变量和预测值的残差以及残差平方和做进一步的判断。***

**四、非线性回归过程**

当通过相关分析得出变量间的关系是非线性的时候

但也可通过适当变换化为线性模型。对于上述这些可化为线性模型的回归问题，一般先将其化为线性模型，然后再用最小二乘法求出参数的估计值，最后再经过适当的变换，得到所求回归曲线。

NLIN过程采用迭代的方法求解方程，假设你要拟合的方程为model y=exp(a+bx),那么a、b就是参数名称。要用迭代法求解，需要指定参数初始值，为确保迭代收敛，可选用直线化求得的a和b值最为初始值。

1、NLIN过程

PROC NLIN语句中的<选项列表>

outest=数据集 定义一个sas数据集。它包括每次proc nlin迭代所生成的估计参数。

best＝n 要求proc nlin只输出网格初始值组合中最好的n组和平方和，省略时，nlin显示所有参数初始值组合的残差平方和。

method＝gauss︱newton︱gradient︱dud︱marquardt 修正的高斯—牛顿法，牛顿法，最速下降法或梯度法，多元割线法，马科德特迭代法。如果不指定method=选择项，若有der语句，则使用gauss法，否则用dud法。

rho＝数值 指定一个数，用于控制寻找步长，默认值为0.1，而在method= marquardt时为10。

smethod=halve︱golden︱armgold︱cubic 指定nlin过程步长寻找方法，默认值为halve。

sigsq=数值 指定一个数值用以代替均方差，计算估计的标准误。

converge=c 指定收敛的准则。默认值是。

maxiter=i 设置迭代次数的限制。i的数值必须是正整数，默认值为100。

**pr**oc nlin过程中的主要语句

model语句——指定因变量和定义表达式来确定预测方程。表达式结果为数值，任何在数据步中可以用的操作符和函数同样可以用在model语句中。

parms语句——在每个“参数=数值”说明中，参数名代表一个被估计的参数，数值说明参数可能的起始值。通常给每一个参数指定一个数值。如果给每个参数指定n个数值，nlin在网格每个点上估计这个模型。下面的parms语句命名了5个参数，并设置它们的起始值：

parms 可能的起始值为：

b0=0 0

b1=4 to 8 4 5 6 7 8

b2=0 to 0.6 by 0.2 0.0 0.2 0.4 0.6

b3=1,10,100 1 10 100

b4=0,0.5,1 to 4 0.0 0.5 1 2 3 4

parms语句中，参数名必须是有效的sas名，而且必须不与nlin过程所用数据集中的变量同名。Nlin过程后只能有一个parms语句。

注意：在进行非线性回归模型拟合时，方程中各系数的取值区间不应过大，步长不应过小。可先尝试给定一个小范围，步长取大一些，然后根据迭代结果对各系数的取值进行反复调整。当找到一个合理数值后，可将步长减小，以使结果更精确。

der语句——使用语句给出的一阶领导数或二阶偏导数。对于大多数方法必须对所估计的每个参数指定一阶偏导数。而对newton方法则可指定一阶偏导数和二阶偏导数。

nlin过程与通常的过程步有些不同，一些data步中的程序语句可以在proc nlin语句后使用，这些语句可以出现在proc nlin中任何地方。

***实例***

某疾病防治站重复治钩虫病病人次数（times）与阳性率（rate）资料，画出其散点图并进行曲线拟合。

**程序如下：**

**data** nlin;

input times rate@@;

cards;

1 63.9 2 36.0 3 17.1 4 10.5 5 7.3 6 4.5 7 2.8 8 1.7

;

**proc** **plot** vpct=**50** hpct=**50**;

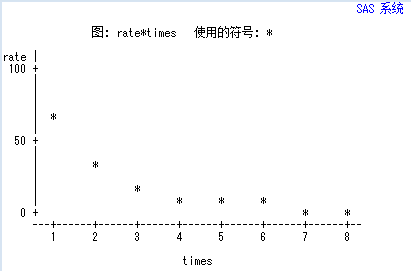
plot rate\*times='\*';

**run**;**proc** **nlin**;

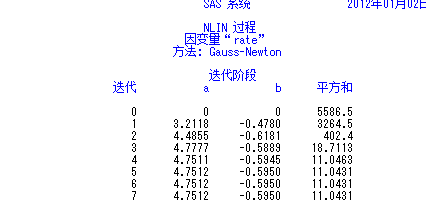
parms a=**0** b=**0**;

model rate=exp(a+b\*times);

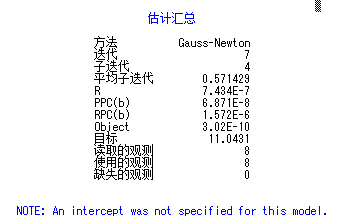
**run**;



如图所示，这是典型的指数衰减曲线🡪利用NLIN过程进行拟合n拟合模型



使用的是牛顿法进行迭代，a和b的初始值都赋为0.迭代阶段列出了参数a。b的值以及残差平方和，在第4次迭代之后，残差平方和已趋于稳定。当到第7次时，残差平方和为11.0431，满足系统的收敛标准，即停止迭代。



R是对参数的主收敛测量，它测量了残差的程度。PPC是预期参数改变测量。RPC是回顾参数改变测量。（不明白PPC和PRC）



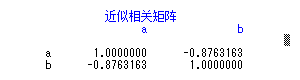
近似方差分析：

H0： 拟合模型的变量对因变量没有影响

由输出结果得：F统计量为1590.61，Pr>F<0.0001，十分显著，拒绝原假设，则说明模型中的一些变量对因变量是有影响的。



图中给出的是模型中未知参数的估计和区间估计值。参数估计值的标准误为近似标准误，近似的95%可信区间（Approximate 95% Confidence Limits）仅供参考，当可信区间的界值离0较近时，下结论应慎重。



A与b的近似相关矩阵，它们的相关系数为-0.8763163，存在负相关性。

综上可得：参数a^=4.7512，b^=-0.5950，拟合方程为：******

***小结：先通过散点图初步判断选择拟合曲线类型，用NLIN进行迭代求参数估计值，再根据残差、近似方差分析进一步分析模型拟合的效果，参数间的相关性用近似相关阵查看。***

**RSREG过程**

RSREG过程说明：

RSREG过程（二次响应面回归过程）用于拟合完全二次响应曲面的回归模型。并通过分析研究拟合曲面的形状的最佳响应的因子水平或范围。

对这样的数据进行分析一般有三项任务：

1. 模型拟合及对参数估计做方差分析
2. 为了调查预测响应曲面的形状而进行典型相关分析
3. 为了寻找最佳响应的范围而进行的岭嵴分析

五、**RSREG过程语句格式：**

|  |  |
| --- | --- |
| PROC RSREG | 选择项1； |
| MODEL | 响应变量=自变量/选择项2； |
| RIDGE | 选择项3； |
| WEIGHT | 变量； |
| ID | 变量； |
| BY | 变量； |
| RUN； |  |
|  |  |

PROC RSREG 和MODEL语句是必需的，其他语句用户可根据需要选择使用

1. PROC RSREG 选择项1
2. DATA=SAS数据集。
3. OUT=SAS数据集。
4. NOPRINT：当只要求创建输出数据集时，该项用于限制所有结果再Output窗口输出。
5. MODEL语句斜杠后的选择项2
6. LACKFIT：要求进行你和不足检验。若规定此选择项，则数据首先必须按自变量分类排序，以便让出现相同值的观测集中在一起。
7. OUT=SAS数据集中，若梅有规定统计量选择项，则输出数据集虽被创建，但没有观测。选择项的关键词将变成输出数据集中的特殊变量\_TYPE\_的值。
8. PREDICT：规定该模型的预测值。
9. RESIDUAL：规定残差，按实际值-预测值运算。
10. L95M或L95M：规定因变量的期望值或均值的95%置信区间的下限（或上限）。
11. L95和U95：规定各个预测值的95%置信区间的下限和上限。用来计算这个上下限的方差，包括误差的方差和参数估计的方差。
12. D：规定COOK’S D影响统计量。
13. RIDGE选择项3

该语句规定计算最佳响应的岭嵴。岭嵴从给定的点x0开始，且以x0为中心、半径为R的岭嵴上找到这样一个因子点，它使得这个岭嵴上预测响应为最佳。可认为在预测响应面上这个岭嵴是往上或往下寻找的最可能方向。于是岭嵴分析可用来帮助解释已存在的响应面或指示近一步的实验方向。

若起始点x0缺省，取设计因子的最大和最小值的中点坐标作为x0，若半径也缺省，那么在0,0.1…0.9上计算岭嵴。通常，岭嵴分析根据对这些因子变量的编码值来拟合响应面，那么岭嵴的结果是对每个坐标用编码零值点开始不断扩大，但不能超出试验范围。不过，岭嵴的中心店和半径可以由用户规定。

**选择项3**

* 1. CENTER=

规定岭嵴起点x0的坐标列表。这个列表可以采用选择项RADIUS=描述的任何形式。坐标用原始的因子变量值（非编码）给出。这里必须为模型中的这些因子规定坐标，而且坐标的次序必须同MODEL语句中使用的次序相同。在列表中的坐标个数若超过模型中因子的个数，多于的坐标被忽略。起始点试验范围的内部。缺省时置每个坐标为相应因子的最大和最小值间的中间值。

* 1. MINIMUM|MIN

MAXIMUM|MIN

规定所计算的岭嵴类型。选择项MAX和MIN亮泽可以都规定；但至少规定一个。

* 1. OUTR=SAS数据集

命名存放所计算的最佳岭嵴的输出数据集。

* 1. RADIUS=（编码半径）

可有以下几种形式：

m1 ,m2,m3,…. 几个不同的值；

m to n 以m为起始值，n为终止值，增量为1的一系列值；

m to n by i 以m为起始值，n为终止值，增量为i的一系列值；

以上形式也可以混合使用，但混合形式要用逗号分隔开。

1. ID语句

ID语句设置一些变量，他们被输出到由 PROC RSREG语句的选择项OUT=创建的数据集中

1. BY语句

同其他过程中的用法一样。

**具有简单最优的响应面（二次曲面）**

**实例**

此例使用三因子的二次模型，试验的目的是想法使一种难闻的化学气味达到最小。考虑的相应变量（因变量）y表示臭气。与臭气有关的因子有三个：x1为温度、x2为气体-液体比、x3为容器的高度。共有15组试验数据。

**data** rsreg6\_19;

input y x1-x3@@;

label y="odor" x1="温度" x2="气体-液体比" x3="容器的高度";

cards;

66 -1 -1 0 39 1 -1 0 43 -1 1 0 49 1 1 0

58 -1 0 -1 17 1 0 -1 -5 -1 0 1 -40 1 0 1

65 0 -1 -1 7 0 1 -1 43 0 -1 1 -22 0 1 1

-31 0 0 0 -35 0 0 0 -26 0 0 0

;

**proc** **sort**;

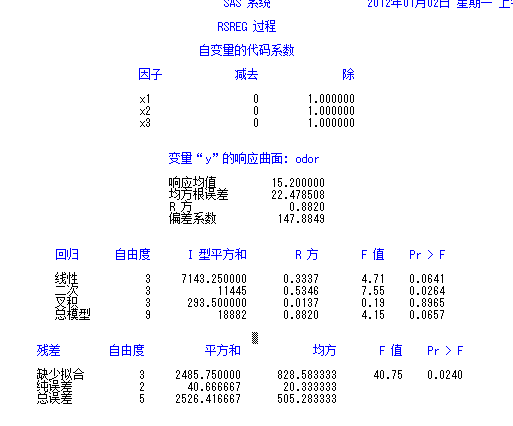
by x1-x3; \*变量x1 x2 x3由小到大排列;

**proc** **rsreg**;

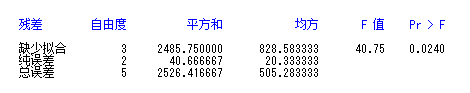
model y=x1 x2 x3/lackfit;/\*要求进行拟合的不合适度检验\*/

**run**;

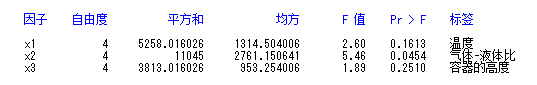
输出结果：



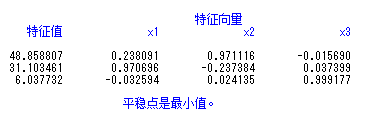
模型中线性项（一次项）、平方项（二次项）和交叉项的分析，对应的P值为各项再模型中的概率，一次项的P值为0.0641，二次项的P值为0.0264,基本可以认为它们在模型中是显著的。二次响应面回归模型是显著的（显著概率为0.0657）；其中的线性项和二次项在0.1水平也是显著的（显著水平分别为0.0641<0.1和0.0264<0.1）；模型的总拟合系数R-Square=0.8820，其中二次项和线性项分别为0.5346和0.3337。



模型拟合的不适合度的检验，Pr>F为0.0240，说明不足是显著的，此模型不能很好的解释变量间的关系，需要再做进一步的分析。

二次响应面模型各项的参数估计以及各项在模型中的显著概率水平，按0.05的水平则只有x3、x1的平方和x2的平方是显著的。

因子检验：只有x2（气体-液体比）是显著的。

响应面的典型分析是对原始数据进行标准化处理后的分析。由于原始数据就是标准的数据，所以输出没有变化稳定点的坐标：x1=0.121913,x2=0.199575,x3=1.770525。

对应于最大特征值（48.858807）的特征向量方向的最大分量同x2有关，次大特征值（31.103461）与x1有关，最小特征值（6.037732）同x3有关，而第三个特征值比前两个小得多，这表明响应面对于该因子x3（容器的高度）的改变不灵敏。

为了对因子变量中的某两个因子画出响应面趋势图，首先固定最不显著的因子变量x3=1.77，然后生成x1和x2的格子点图。

**data** test;

set test end=eof; \*得到实测值;

output;

if eof then

do;

y=**.**;

x3=**1.77**;

do x1=-**1.5** to **1.5** by **0.1**;

do x2= -**2** to **2** by **0.1**;

output;

end;

end;

end;

**proc** **rsreg** data=test out=cc noprint;

model y=x1 x2 x3/predict;

**run**;

**data** dd;

set cc;

if x3=**1.77**;

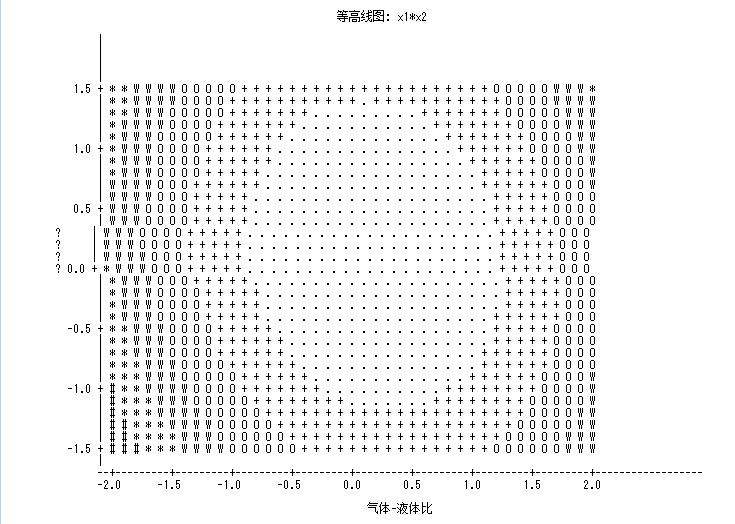
**proc** **plot** data=dd;

plot x1\*x2=y/contour=**6** hpos=**100** vpos=**36** hspace=**10**

haxis=-**2** to **2** by **0.5**

vaxis=-**1.5** to **1.5** by **0.5**;

**run**;



由输出的层次图可以看出取值在0附近。