**27. 非线性回归**

现实世界中严格的线性模型并不多见，它们或多或少都带有某种程度的近似；在不少情况下，非线性模型可能更加符合实际。

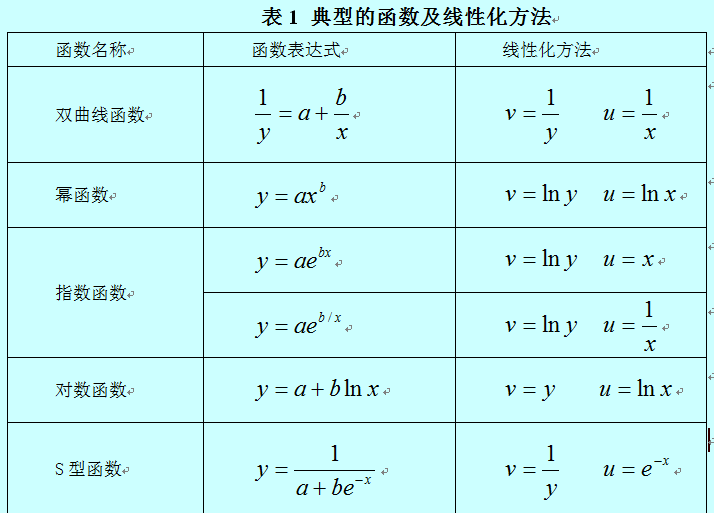
对变量间非线性相关问题的曲线拟合，处理的方法主要有：

（1）首先确定非线性模型的函数类型，对于其中可线性化问题则通过变量变换将其线性化，从而归结为前面的多元线性回归问题来解决；

（2）若实际问题的曲线类型不易确定时，由于任意曲线皆可由多项式来逼近，故常可用多项式回归来拟合曲线；

（3）若变量间非线性关系式已知（多数未知），且难以用变量变换法将其线性化，则进行数值迭代的非线性回归分析。

**（一）可变换为线性的非线性回归**



在很多场合，可以对非线性模型进行线性化处理，尤其是可变换为线性的非线性回归，运用最小二乘法进行推断，对线性化后的线性模型，可以应用REG过程步进行计算。

**例1** 有实验数据如下：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2 | 2.1 | 2.2 | 2.3 | 2.4 |
| Y | 109.95 | 40.45 | 20.09 | 24.53 | 11.02 | 7.39 | 4.95 | 2.72 | 1.82 | 1.49 | 0.82 | 0.3 | 0.2 | 0.22 |

试分别采用指数回归（*y*=*aebx*）方法进行回归分析。

代码：

**data** exam25\_1;

input x y;

cards;

1.1 109.95

1.2 40.45

1.3 20.09

1.4 24.53

1.5 11.02

1.6 7.39

1.7 4.95

1.8 2.72

1.9 1.82

2 1.49

2.1 0.82

2.2 0.3

2.3 0.2

2.4 0.22

;

**run**;

**proc** **sgplot** data = exam25\_1;

scatter x = x y = y;

**run**;

**proc** **corr** data = exam25\_1;

var x y;

**run**;

**data** new1;

set exam25\_1;

v = log(y);

**run**;

**proc** **sgplot** data = new1;

scatter x = x y = v;

title '变量代换后数据';

**run**;

**proc** **reg** data = new1;

var x v;

model v = x;

print cli;

title '残差图';

plot residual. \* predicted.;

**run**;

**data** new2;

set exam25\_1;

y1 = **14530.28**\*exp(-**4.73895**\*x);

**run**;

**proc** **gplot** data = new2;

plot y\*x=**1** y1\*x=**2** /overlay;

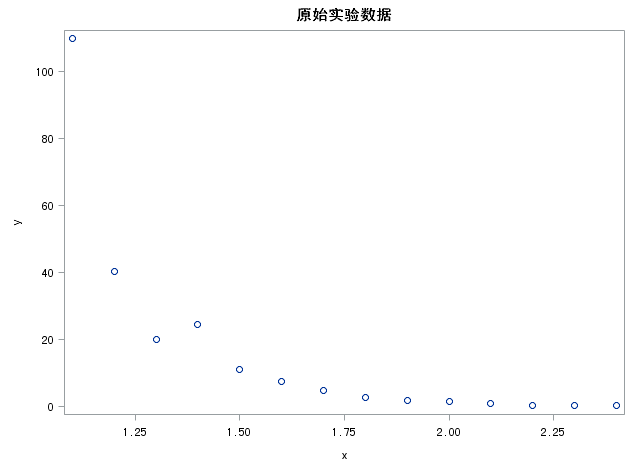
symbol v=dot i=none cv=red;

symbol2 i=sm color=blue;

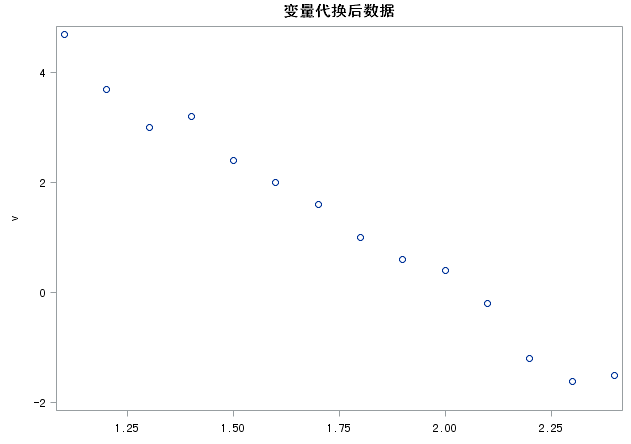
title '指数回归图';

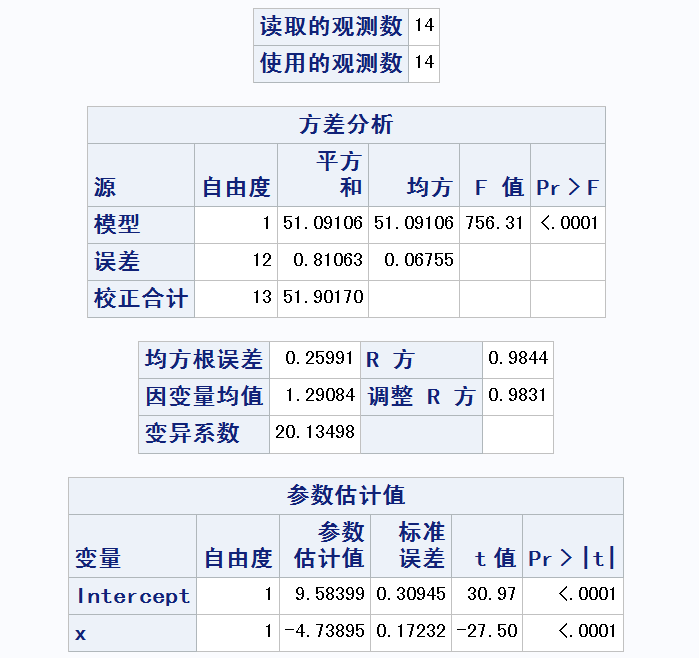
**run**;

运行结果：

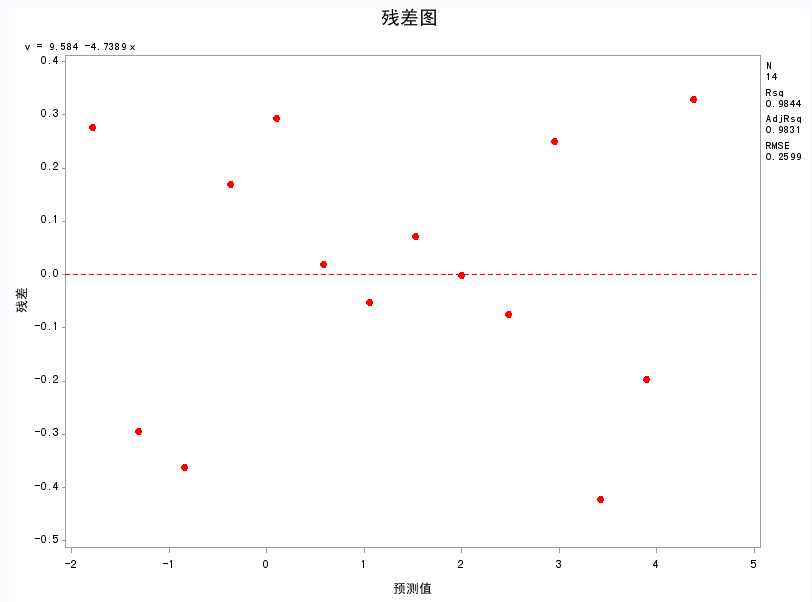


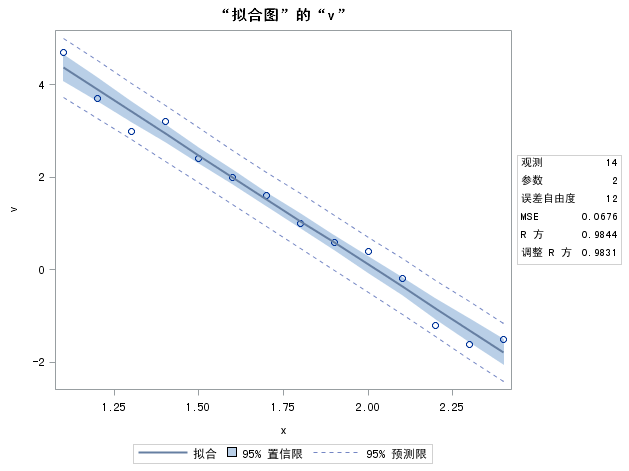


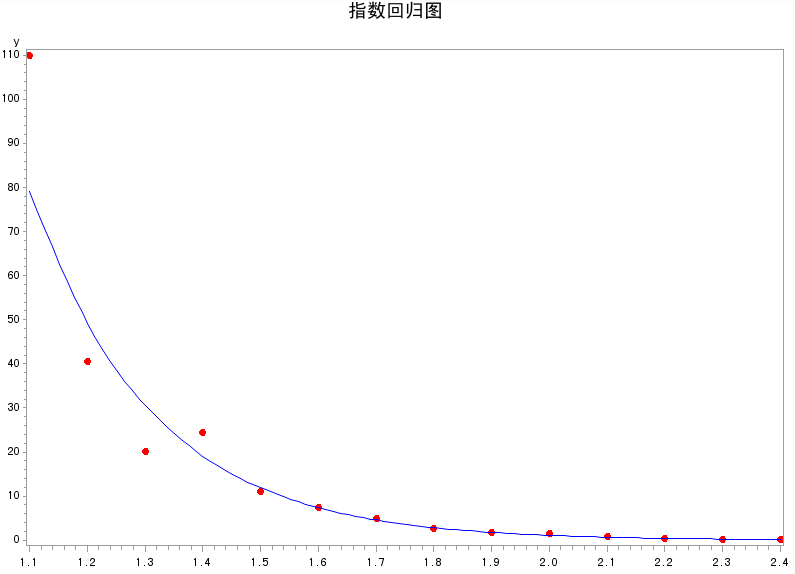












**程序说明：**

（1）调整后的R2=0.9831，说明拟合程度很好；F检验的P值=0.0001<α=0.05，拒绝原假设，故直线回归的斜率不为0；

（2）将线性回归系数代入，得到原回归方程

y =14530.28\*e−4.73895x

（3）残差图趋势，符合残差随机正态分布的假设（不带其它明显趋势）。

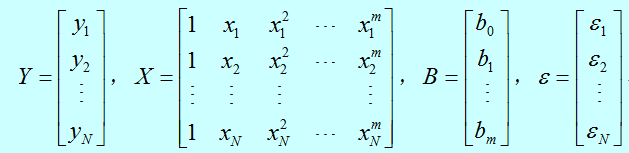
**二、多项式回归**

一般函数都可用多项式来逼近，故多项式回归分析可用来处理相当广泛的非线性问题。

对观测数据 (xt, yt), t=1,…,N. 多项式回归模型为：



令



则模型可写为：

Y=XB+ε

当X列满秩时，用最小二乘估计可求得其多项式回归方程。但由于的计算既复杂又不稳定，故一般采用正交多项式法来进行多项式回归。

多项式模型可以直接应用GLM（广义线性模型）求解。

**例2**重庆市种畜场奶牛群1—12月份(x1)，产犊母牛平均产奶量(y)

的资料如程序数据步中，试对该资料配置一个合适的回归方程。

**代码：**

**data** exam25\_2;

input x1 y @@;

x2=x1\*x1;

datalines;

1 3833.43

7 3476.76

2 3811.58

8 3466.22

3 3769.47

9 3395.42

4 3565.74

10 3807.08

5 3481.99

11 3817.03

6 3372.82

12 3884.52

;

**run**;

**proc** **sgplot** data = exam25\_2;

scatter x = x1 y = y;

title '原始数据散点图';

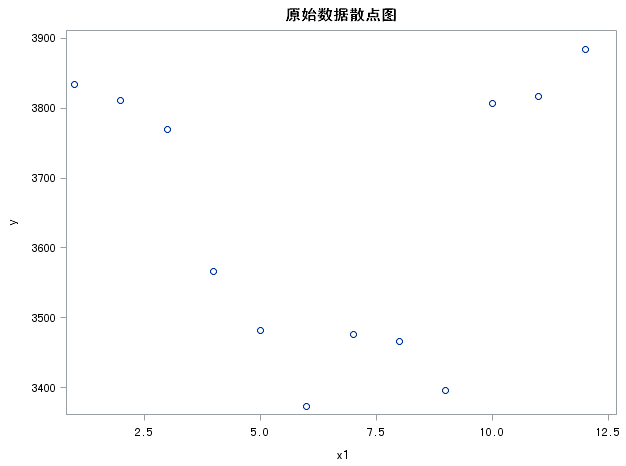
**run**;

**proc** **reg** data = exam25\_2;

model y=x1 x2;

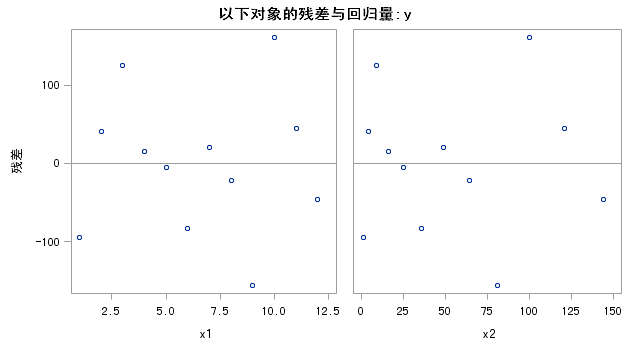
**run**;

运行结果：



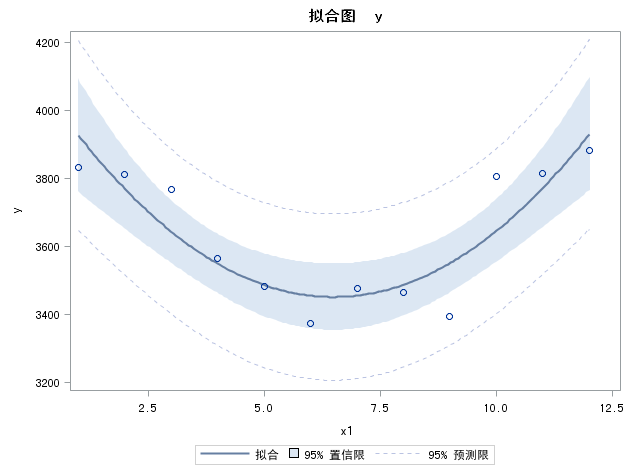












**程序说明：**

（1）观察数据的散点图，更适合二次多项式拟合，也可以测试几种不同次数的多项式拟合选择其中最优的；

（2）将回归系数代入多项式方程得到：

y= 4117.20136 -204.93668x1+ 15.78570x12

**三、不能变换为线性的非线性回归**

该类非线性回归分析就是利用最小二乘准则来估计回归系数β，使得残差平方和最小。

一般来用数值迭代法来进行，先选定回归系数的初值β0，按照给定的步长和搜索方向逐步迭代，直到残差平方和达到最小。

有5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）、牛顿法（Newton）、麦夸特法（Marquardt）、正割法（DUD）。

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

**（二）PROC NLIN过程步**

对于不能线性化的非线性模型。其估计不能直接运用经典的最小二乘法，而需要运用其他估计方法，如加权最小二乘法、直接搜索法、直接最优法与Taylor级数展开法进行线性逼近。此时，可以利用NLIN过程步实现相应的计算，它是采用最小误差平方法及迭代推测法来建立一个非线性模型，估计参数默认采用高斯-牛顿迭代法。NLIN过程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。

基本语法：

PROC NLIN data = 数据集 </可选项>;

PARMS 参数名=数值;

MODEL 因变量=表达式 </可选项>;

< BOUNDS 不等式;>

< CONTROL 变量;>

< der. 参数名=表达式;>

<OUTPUT out = 输出数据集</可选统计量>;>

说明：

（1）NLIN的可选项包括：

outest＝输出数据集——输出每步迭代的结果；

best＝n——只输出最好的n组残差平方和；

method＝gauss | marquardt | newton| gradient| dud |——设定参数估计的迭代方法，默认为gauss（没有der.语句）；

（2）PARMS语句指定参数并赋值，一般包括参数名、初始值（Grid Search可以帮助选择合适的初始值）、迭代准则；例如：

parms b0=**0** b1=**1** to **10** b2=**1** to **10** by **2** b3=**1**,**10**,**100**;

（3）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**;

（4）der.语句用于计算模型关于各参数的偏导数，相应格式为：

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，对于模型model y=b0\*(**1**－exp(-b1\*x)); 二阶偏导数表达式：der.b0.b1=x\*exp(-b1\*x);

例3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。

**代码：**

**data** expd;

input x y @@;

datalines;

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**;

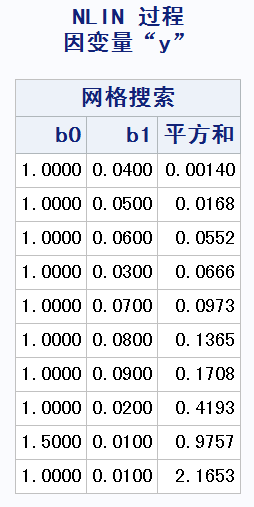
axis2 order=**0.5** to **1.1** by **0.05**;

title1 'y=b0\*(1-exp(-b1\*x)';

title2 'proc nlin method=gauss';

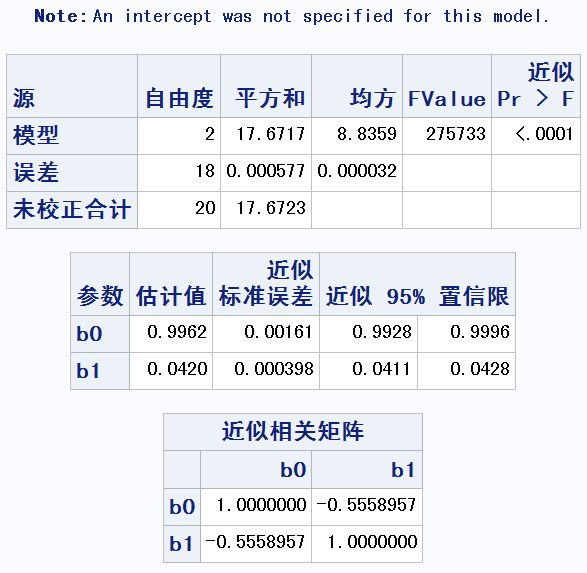
**run**;

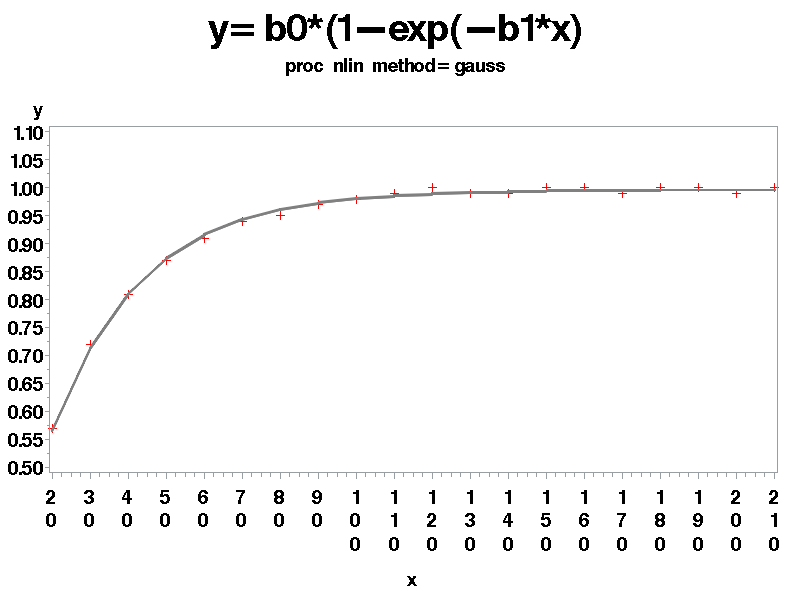
运行结果：









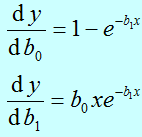


**程序说明：**

（1）parms语句设置了初始值网格值为b0取0, 0.5, 1, 1.5, 2共5个值，b1取0.01, 0.02, …, 0.09共9个值，所有可能组合为5×9=45,选项best=10要求输出残差平方和最小的前10种组合；

（2）最好的迭代初始值为b0=1.0000，b1=0.0400，此时回归模型残差为ESS=0.00140; 从该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（ESS值几乎不变），停止迭代；

（3）高斯-牛顿迭代算法要求给出模型对参数b0和b1的一阶偏导数表达式：



der.语句用来表示上面两个一阶偏导数表达式；

（4）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；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；

（5）方差分析表，给出了回归平方和为17.6717, 残差平方和为0.000577, 总平方和为17.6723.

（6）参数估计表，给出了b0和b1的渐近估计值，得到的非线性回归模型为

y = 1.0000000\*[1-exp(0.5558957x)]

同时还给出b0和b1参数估计的渐近有效标准差和渐近95%置信区间。