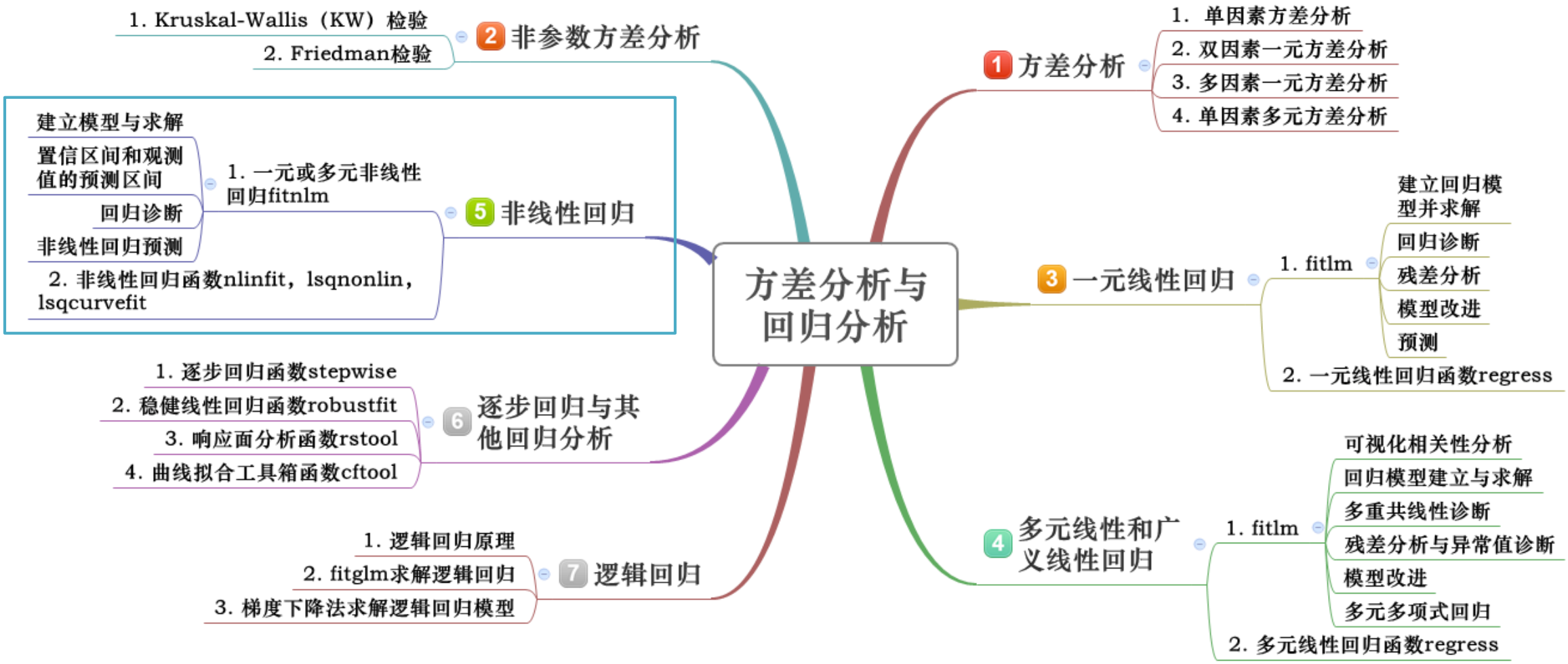
信陽師範学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

第11章 方差分析与回归分析

讲授人：牛言涛

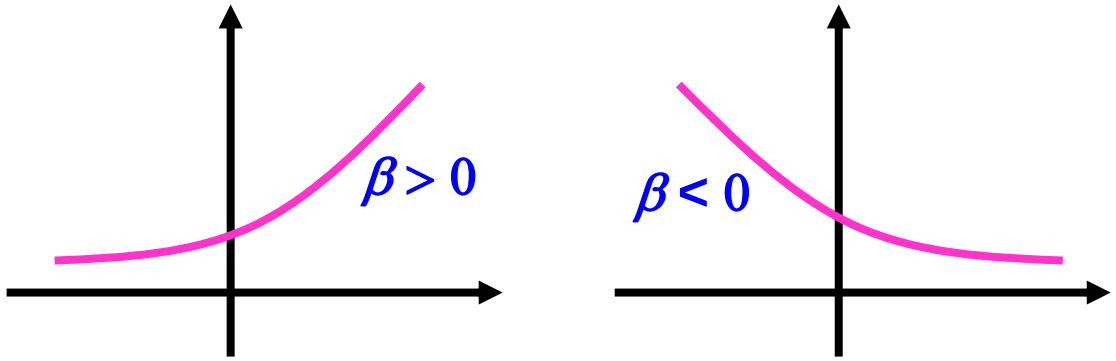
日期：2020年4月16日

第11章 方差分析与回归分析知识点思维导图

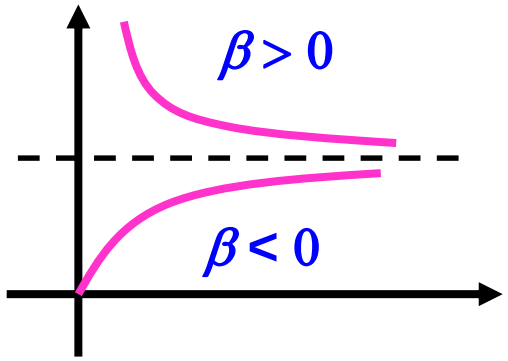


非线性回归函数

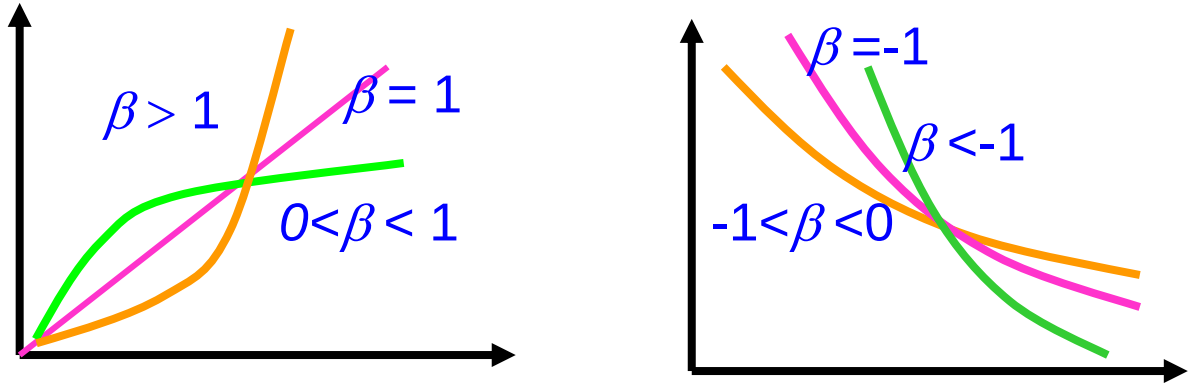
指数函数 $y = \alpha e^{\beta x}$



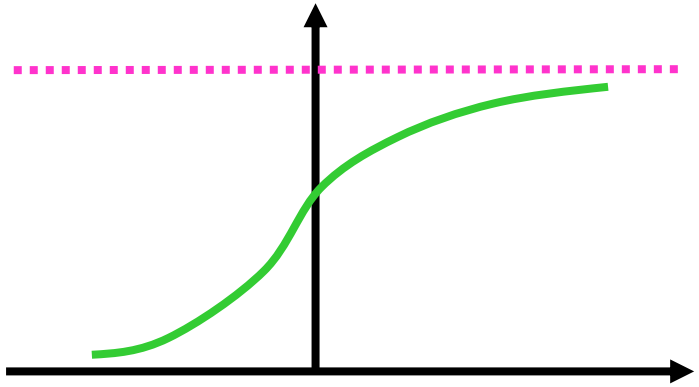
负指数函数 $y = \alpha e^{\frac{\beta}{x}}$



幂函数 $y = \alpha x^{\beta}$



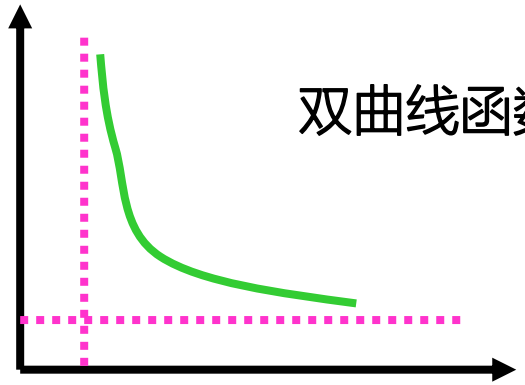
S型曲线 $y = \frac{1}{\alpha + \beta e^{-x}}$



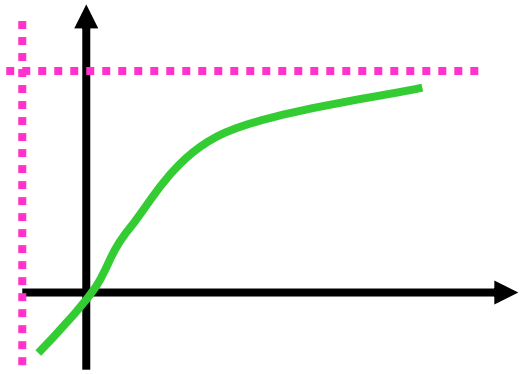
非线性回归函数

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

双曲线函数



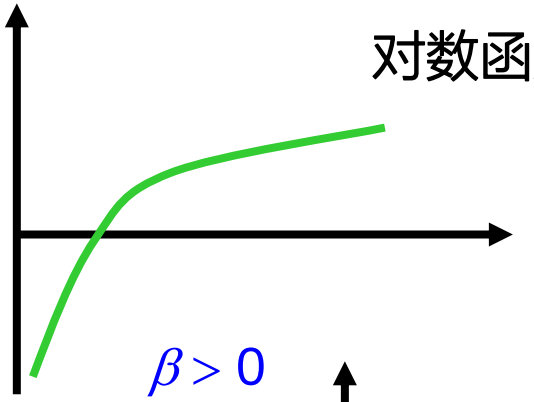
$\beta < 0$



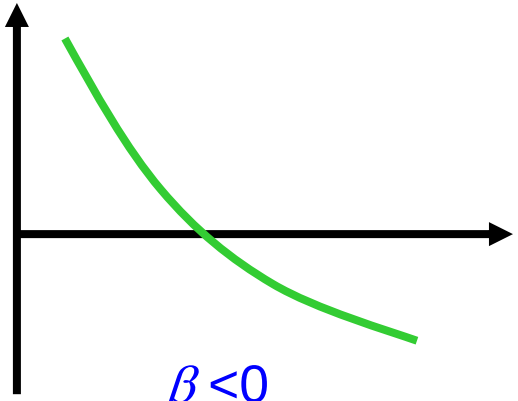
$\beta > 0$

$$y = \frac{x}{\alpha x + \beta}$$

对数函数

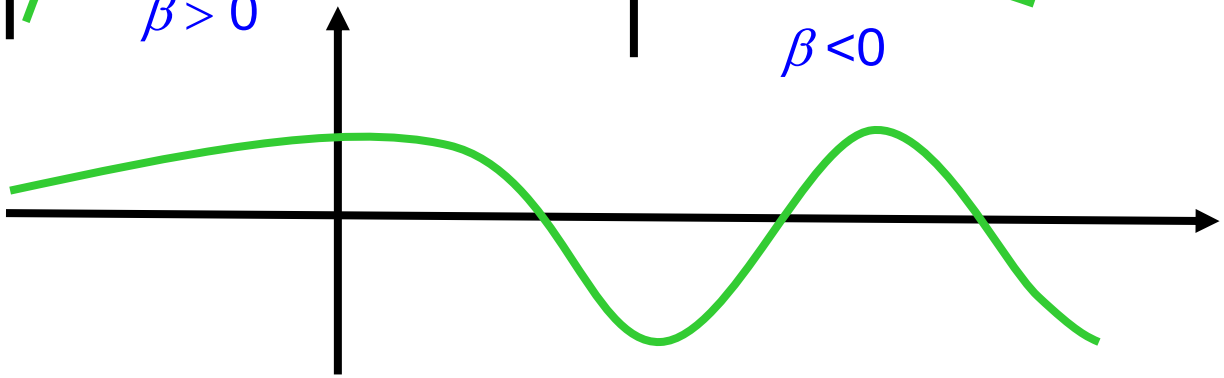


$\beta > 0$



$\beta < 0$

$$y = \alpha + \beta \lg x$$



多项式函数

$$y = a_1 x^n + a_2 x^{n-1} + \cdots + a_n x + a_0$$

一. 一元非线性回归

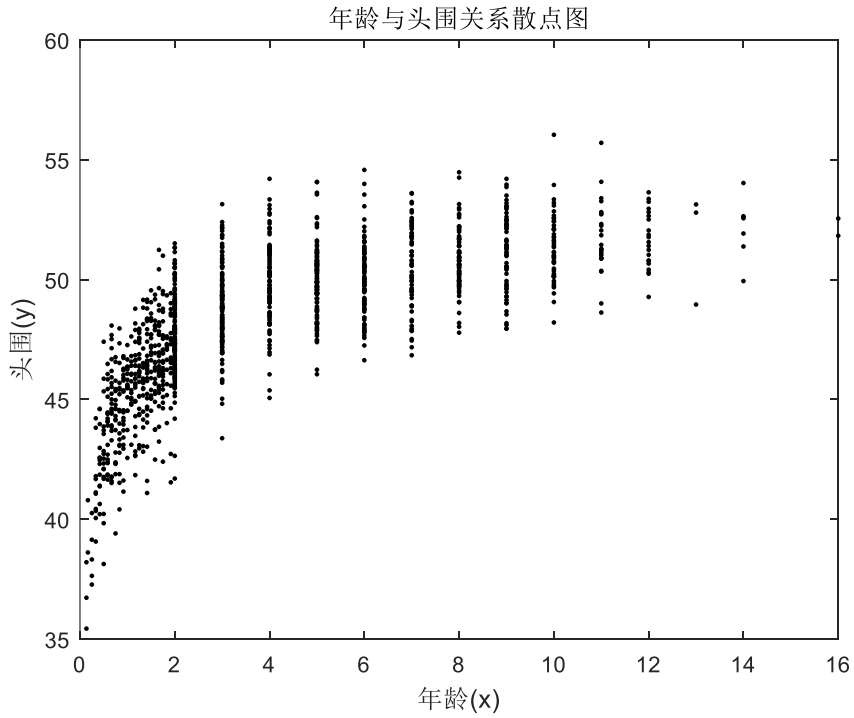
信阳师范学院
数学与统计学院

例1：头围(head circumference)是反映婴幼儿大脑和颅骨发育程度的重要指标之一，对头围的研究具有非常重要的意义。笔者研究了天津地区1281位儿童（700个男孩，581个女孩）的颅脑发育情况，测量了年龄、头宽、头长、头宽/头长、头围和颅围等指标，测得1281组数据，年龄跨度从7个星期到16周岁，试根据这1281组数据建立头围关于年龄的回归方程。

	A	B	C	D	E	F	G	H	I	J
1	序号	性别	年龄及标识	年龄	月龄	头宽	头长	头宽/头长	头围	颅围
2				单位：岁	单位：月	单位：毫米	单位：毫米		单位：厘米	单位：厘米
3	1	m	11Y	11	132	136.0476	168.7998	0.805970149	50.90952	48.3008
4	2	m	20M	1.666667	20	149.9043	161.2416	0.9296875	50.4282	49.01562
5	3	m	10Y	10	120	144.4456	156.6227	0.922252011	51.35181	48.14725
6	4	m	3Y	3	36	145.7053	163.761	0.88974359	50.27417	48.73305
7	5	m	3Y	3	36	139.8267	153.2635	0.912328767	48.52064	46.925
8	6	m	5Y	5	60	146.965	155.7829	0.943396226	50.30917	48.04408
9	7	m	4Y	4	48	151.164	163.761	0.923076923	52.34006	50.33477
10	8	m	8Y	8	96	139.8267	159.9819	0.874015748	49.05821	47.44784
11	9	m	8Y	8	96	138.9869	169.2197	0.82133995	51.17692	49.05429
12	10	m	11M	0.916667	11	133.1083	145.2854	0.916184971	46.21404	44.36468
13	11	m	4Y	4	48	146.965	163.3411	0.899742931	51.89274	49.36296

1. 绘制数据散点图

```
>> HeadData = xlsread('headcf.xls');
>> x = HeadData(:, 4);
>> y = HeadData(:, 9);
>> plot(x, y, 'k.');
>> xlabel('年龄(x)');  ylabel('头围(y)');
>> title('年龄与头围关系散点图')
```



- 备选方程

✓ 负指数函数： $y = \beta_1 e^{\frac{\beta_2}{x+\beta_3}}$

✓ 双曲线函数： $y = \dfrac{x+\beta_1}{\beta_2 x+\beta_3}$

✓ 幂函数： $y = \beta_1 (x+\beta_2)^{\beta_3}$

✓ Logistic曲线函数： $y = \dfrac{\beta_1}{1+\beta_2 e^{-(x+\beta_3)}}$

✓ 对数函数： $y = \beta_1 + \beta_2 \ln(x+\beta_3)$

2. 建立模型与求解

% 选择负指数函数作为理论回归方程。

```
>> HeadCir = @(beta, x)beta(1)*exp(beta(2)./(x+beta(3)));
>> beta0 = [53,-0.2604,0.6276];
>> opts = statset('Display','iter','TolFun',1e-10,'Robust','on');
% nlm1 = NonLinearModel.fit(x,y,HeadCir,beta0,'Options',opts) %不推荐使用
>> nlm1 = fitnlm(x,y,HeadCir,beta0,'Options',opts)  %推荐使用
```

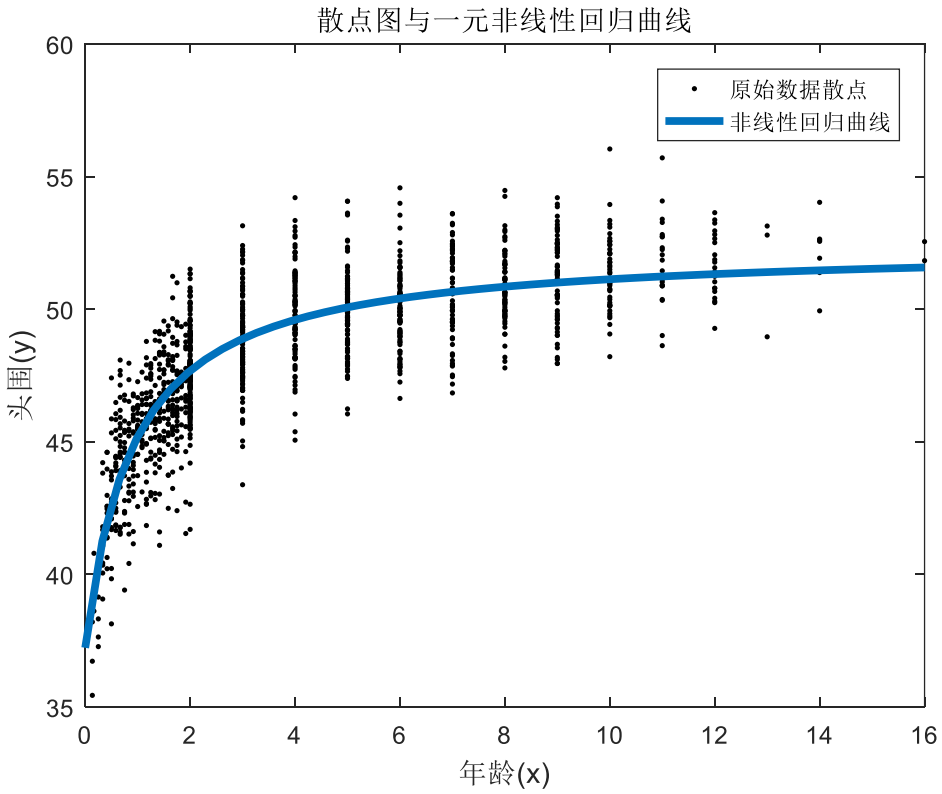
```
nlm1 =
Nonlinear regression model (robust fit):
    y ~ beta1*exp(beta2/(x + beta3))

Estimated Coefficients:
             Estimate       SE        tStat       pValue
             ________    ________    _______    __________
    beta1      52.377      0.1449     361.46             0
    beta2     -0.25951     0.016175   -16.044    6.4816e-53
    beta3      0.76038     0.072948    10.423    1.7956e-24

Number of observations: 1281, Error degrees of freedom: 1278
Root Mean Squared Error: 1.66
R-Squared: 0.747,  Adjusted R-Squared 0.747
F-statistic vs. zero model: 4.64e+05, p-value = 0
```

```
>> nlm1.plotSlice  %绘制图像
>> Alpha = 0.05;
%参数估计的置信区间
>> ci1 = nlm1.coefCI(Alpha)
ci1 =
   52.0923   52.6609
   -0.2912   -0.2278
    0.6173    0.9035
```

建立的模型为： $\hat{y} = 52.377e^{\frac{-0.25951}{x+0.76038}}$

非线性回归函数

```
>> xnew = linspace(0,16,50)';
>> ynew = nlm1.predict(xnew);
>> plot(x, y, 'k.');
>> hold on;
>> plot(xnew, ynew, 'linewidth', 3);
>> xlabel('年龄(x)');
>> ylabel('头围(y)');
>> legend('原始数据散点','非线性回归曲线');
>> title('散点图与一元非线性回归曲线')
```

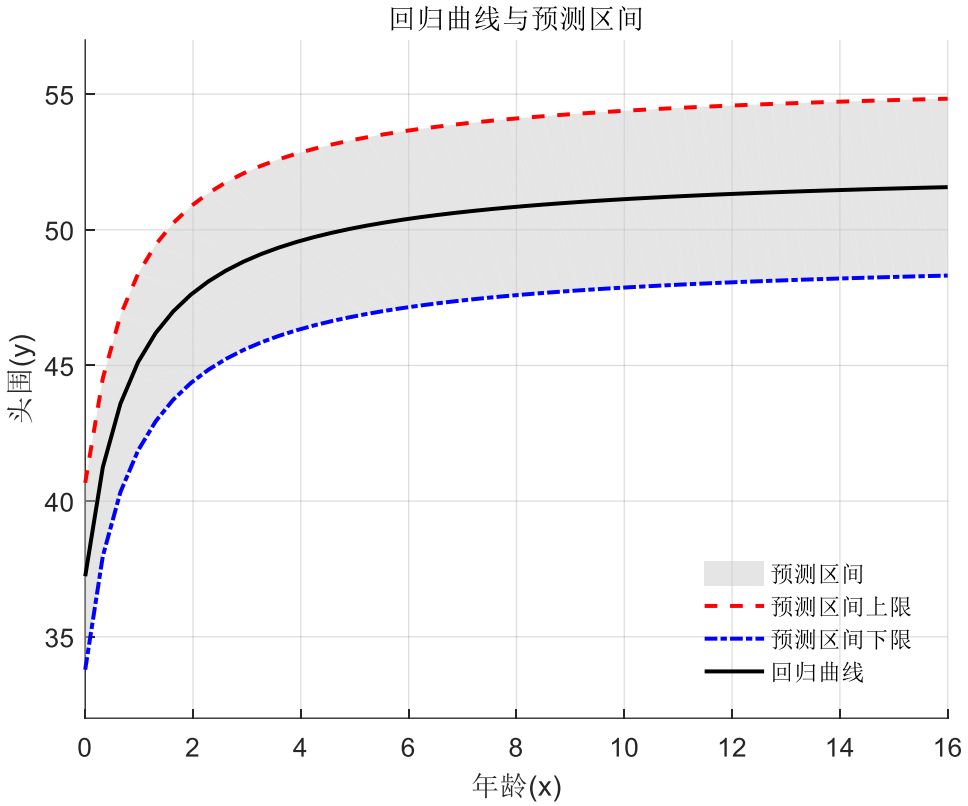


5. 置信区间和观测值的预测区间

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

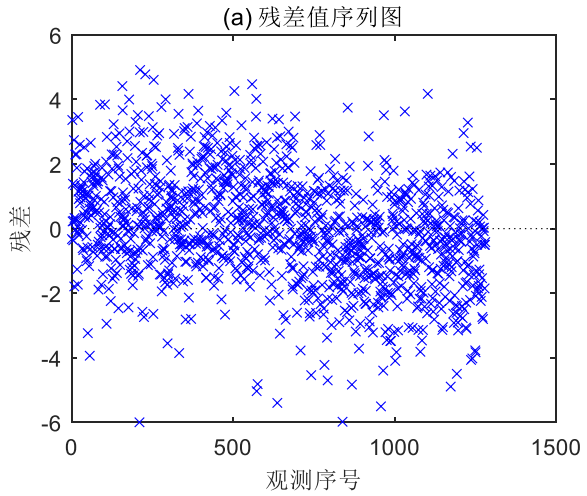
头围平均值的置信区间和观测值的预测区间

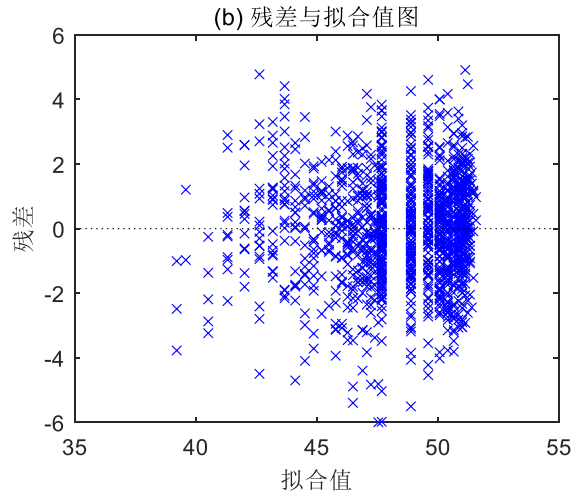
```
[yp,ypci] = nlm1.predict(xnew,'Prediction','observation');
yup = ypci(:,2);   %置信区间上限值
ydown = ypci(:,1);  %置信区间下限值
hold on; grid on;
h1 = fill([xnew;flipud(xnew)],[yup;flipud(ydown)],[0.8,0.8,0.8]);
set(h1,'EdgeColor','none','FaceAlpha',0.5);
plot(xnew,yup,'r--',xnew,ydown,'b-.',xnew, yp, 'k','LineWidth',1.5);
ylim([32, 57]);
xlabel('年龄(x)');ylabel('头围(y)');
legend('预测区间','预测区间上限','预测区间下限','回归曲线','Location', 'SouthEast');
legend('boxoff')
title('回归曲线与预测区间')
```

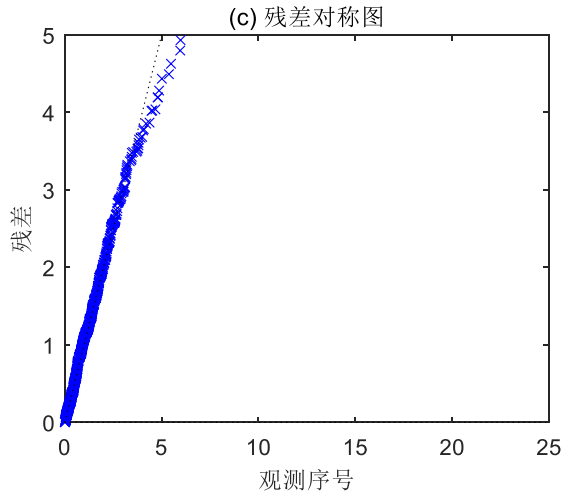


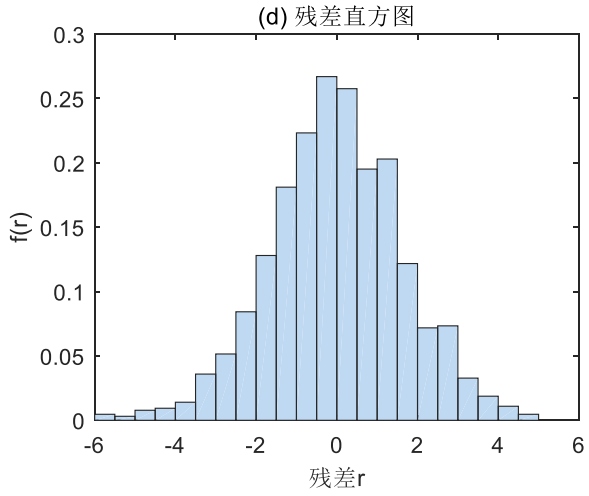
6. 回归诊断

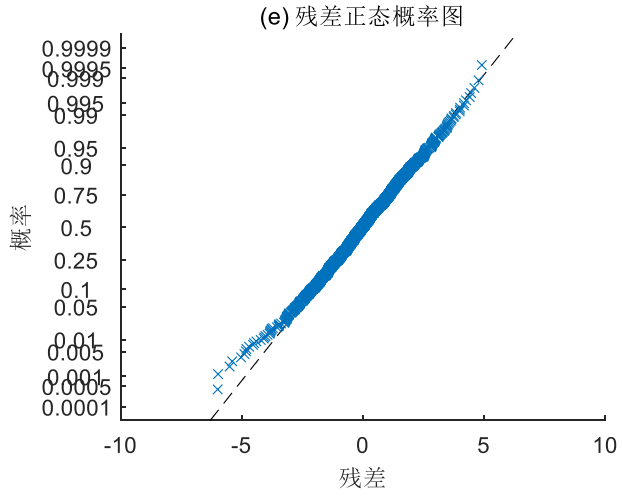
(1) 残差分析：>> plot_plotResiduals(nlm1)

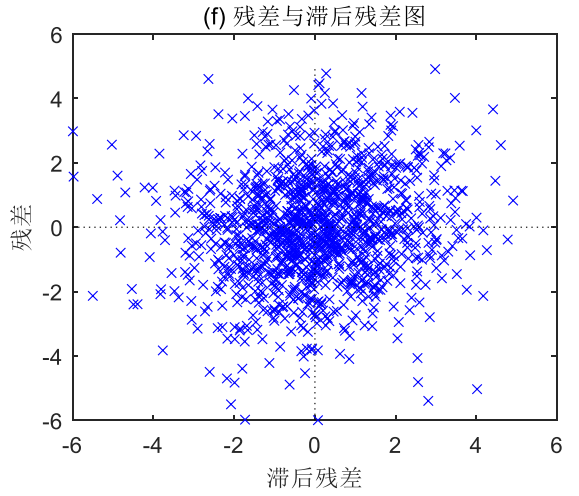






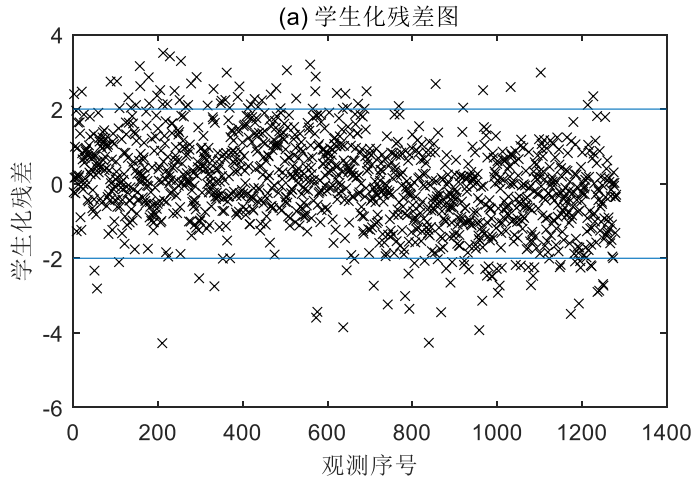


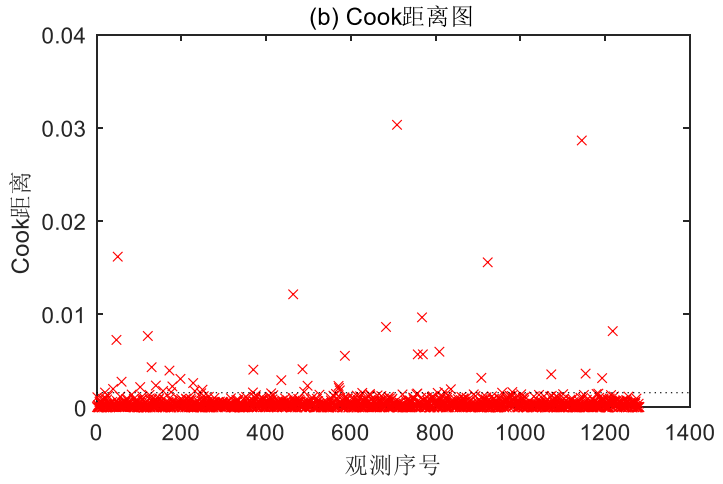


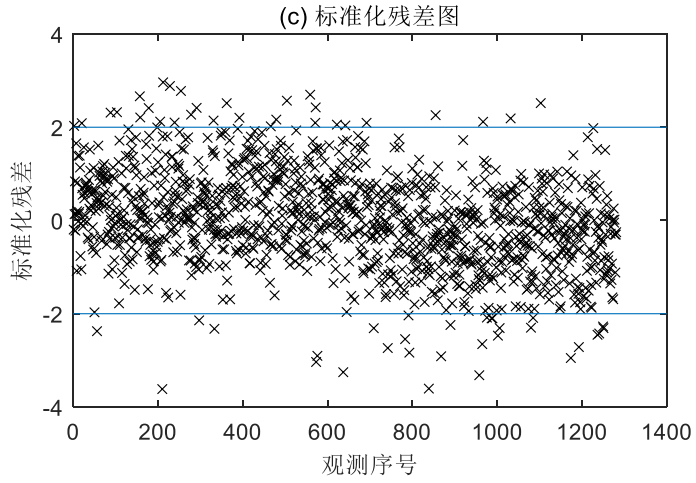


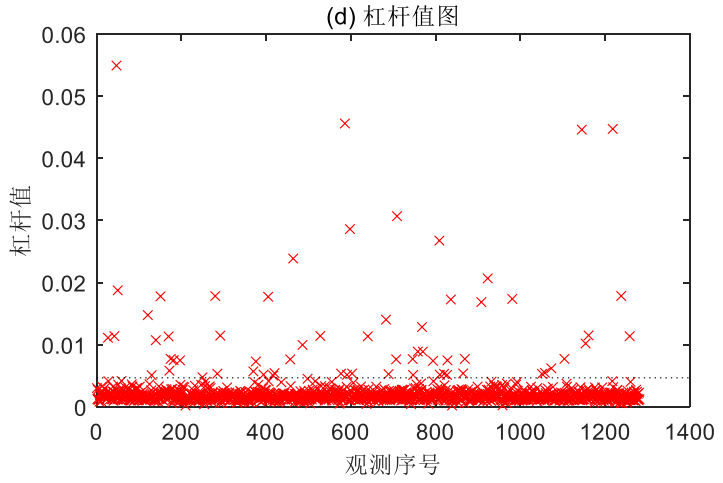
6. 回归诊断

（2）离群点诊断：>> [idout,idinf,idleve] = fitnlmplot_outliers(nlm1);



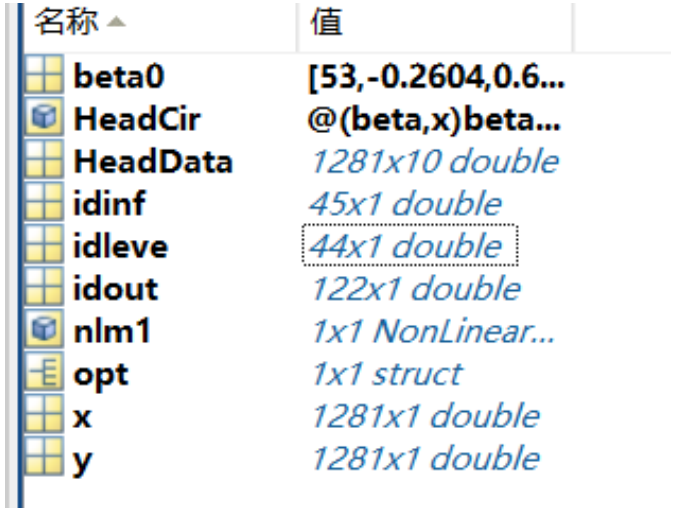






存在强影响点较多的异常值、高杠杆值和强影响点。

从数据的输出来看：异常值有122个、高杠杆值有44个，强影响点45个。



6. 回归诊断

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

（3）模型改进

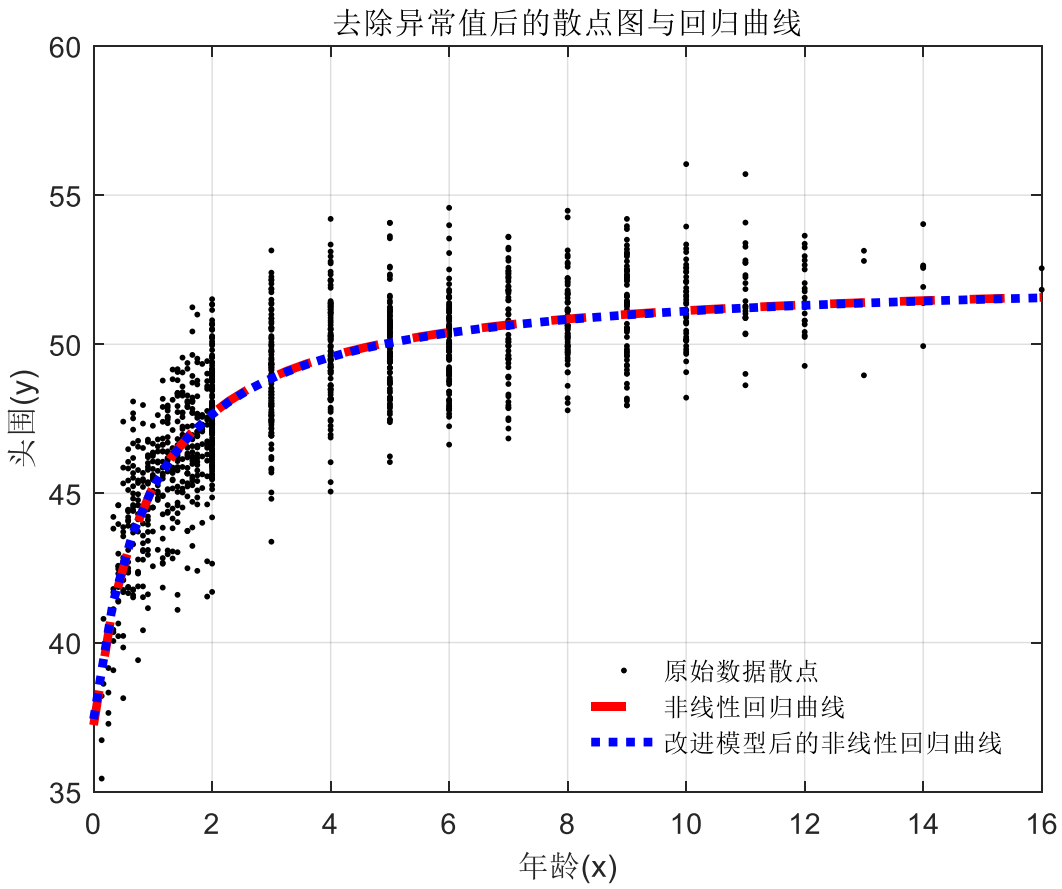
>> nlm2 = NonLinearModel.fit(x,y,HeadCir,beta0,'Exclude',idout,'Options',opt)

```
nlm2 =
Nonlinear regression model (robust fit):
    y ~ beta1*exp(beta2/(x + beta3))

Estimated Coefficients:
            Estimate       SE        tStat       pValue
            ________    ________    _______    __________
    beta1     52.369     0.12693     412.6            0
    beta2    -0.26243    0.014592   -17.984    5.9309e-64
    beta3     0.78167    0.067002    11.666    8.2311e-30

Number of observations: 1159, Error degrees of freedom: 1156
Root Mean Squared Error: 1.37
R-Squared: 0.807,  Adjusted R-Squared 0.807
F-statistic vs. zero model: 6.11e+05, p-value = 0
```

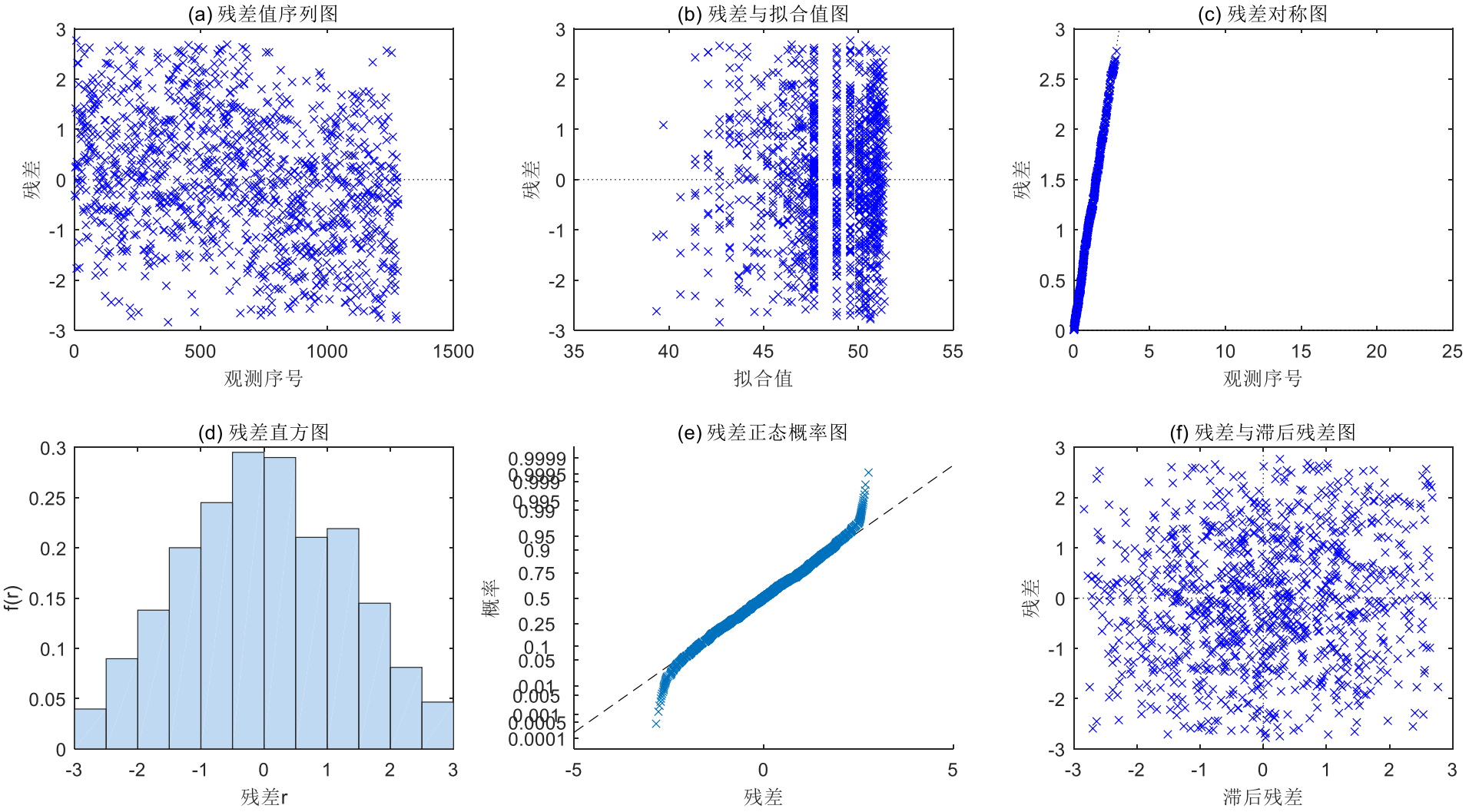
建立的模型为：$\hat{y} = 52.369e^{\frac{-0.26243}{x+0.78167}}$

改进模型后回归曲线

```
xnew = linspace(0,16,50)';
ynew = nlm1.predict(xnew);
plot(x, y, 'k.');
hold on; grid on
plot(xnew, ynew, 'r--','linewidth', 3);
ynew2 = nlm2.predict(xnew);
plot(xnew, ynew2, 'b:', 'linewidth', 3);
xlabel('年龄(x)');
ylabel('头围(y)');
legend('原始数据散点','非线性回归曲线','改进模型后的非线性回归曲线');
legend('boxoff')
title('去除异常值后的散点图与回归曲线')
```



模型改进后的残差图

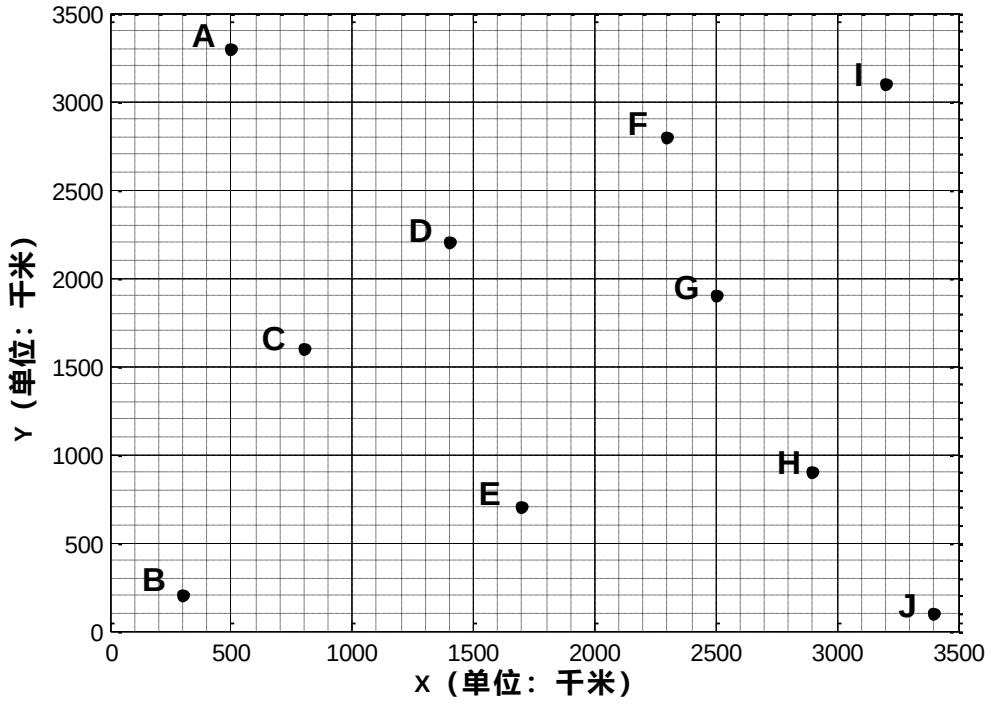


残差分析满足基本假定，独立、完美对称、不自相关、近似服从正态分布，但存在厚尾性。

二. 多元非线性回归

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS

例2：2011年4月1日某时在某一地点发生了一次地震，图中10个地震观测站点均接收到了地震波，观测数据如表所列。假定地震波在各种介质和各个方向的传播速度均相等，并且在传播过程中保持不变。请根据表中的数据确定这次地震的震中位置、震源深度以及地震发生的时间（不考虑时区因素，建议时间以分为单位）。



地震观测站	横坐标x(千米)	纵坐标y(千米)	接收地震波时间
A	500	3300	4月1日9时21分9秒
B	300	200	4月1日9时19分29秒
C	800	1600	4月1日9时14分51秒
D	1400	2200	4月1日9时13分17秒
E	1700	700	4月1日9时11分46秒
F	2300	2800	4月1日9时14分47秒
G	2500	1900	4月1日9时10分14秒
H	2900	900	4月1日9时11分46秒
I	3200	3100	4月1日9时17分57秒
J	3400	100	4月1日9时16分49秒

建立非线性回归模型

假设震源三维坐标为(x_0, y_0, z_0)，这里的z_0取正值，设地震发生的时间为2011年4月1日9时t_0分，地震波传播速度为v_0（单位：km/s）.

用$(x_i, y_i, 0), i = 1, 2, \ldots 10$分别表示地震观测站点A—J的三维坐标，用$T_i$表示观测到站点A—J接收到地震波的时刻，这里的$T_i$表示9时$T_i$分接收到地震波。建立$T_i$关于$x_0, y_0, z_0$的二元非线性回归模型如下：

$$T_i = t_0 + \frac{\sqrt{(x_i - x_0)^2 + (y_i - y_0)^2 + z_0^2}}{60v_0} + \varepsilon_i, \quad i = 1, 2, \cdots, 10$$

建立非线性回归模型

信阳师范学院
数学与统计学院

```
>> xyt = [500 3300 21 9;300 200 19 29;800 1600 14 51;1400 2200 13 17; 1700 700 11 46;
        2300 2800 14 47;2500 1900 10 14;2900 900 11 46; 3200 3100 17 57;3400 100 16 49];
>> modelfun = @(b,x)sqrt((x(:,1)-b(1)).^2+(x(:,2)-b(2)).^2+b(3).^2)/(60*b(4))+b(5);
% modelfun = 'y ~ sqrt((x1-b1)^2 + (x2-b2)^2 + b3^2)/(60*b4)+b5';
>> xy = xyt(:,1:2); Minutes = xyt(:,3); Seconds = xyt(:,4);
>> T = Minutes + Seconds/60;
>> b0 = [1000 100 1 1 1];
>> mnlm = fitnlm(xy,T,modelfun,b0)
```

$$\begin{cases} x_0 = 2200.5 \\ y_0 = 1399.9 \\ z_0 = 35.144 \\ v_0 = 2.9994 \\ t_0 = 6.9863 \end{cases}$$

也就是说地震发生的时间为2011年4月1日09时07分，震中位于(2200.5, 1399.9) 处，震源深度35.1公里。

```
mnlm =
Nonlinear regression model:
    y ~ sqrt((x1 - b1)^2 + (x2 - b2)^2 + b3^2)/(60*b4) + b5

Estimated Coefficients:
          Estimate       SE          tStat       pValue
          ________    _________    _______    __________
    b1     2200.5       0.53366     4123.5    1.5922e-17
    b2     1399.9       0.48183     2905.4     9.168e-17
    b3     35.144         61.893    0.56782        0.5947
    b4     2.9994     0.0041439     723.82     9.5533e-14
    b5     6.9863       0.02087     334.75     4.515e-12

Number of observations: 10, Error degrees of freedom: 5
Root Mean Squared Error: 0.00591
R-Squared: 1,  Adjusted R-Squared 1
F-statistic vs. constant model: 8.3e+05, p-value = 9.75e-15
```

三. 其他非线性回归函数

- 非线性回归函数nlinfit，lsqnonlin，lsqcurvefit

$$y = f\left(x_1, x_2, \cdots, x_p; a_1, a_2, \cdots, a_k\right)$$

未知参数

事先用m-文件定义的非线性函数

[beta, r, J, COVB,mse] = nlinfit(X, y, fun, b0, options)

$\begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \cdots \\ \hat{a}_k \end{bmatrix}$

残差

雅可比矩阵

$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$

回归系数初值

优化属性设置

6. 回归诊断

- 参数估计的置信区间
  - [beta,resid,J,Sigma]= nlinfit(X, y, fun, b0)
  - ci = nlparci(beta,resid, 'covar',Sigma)　　或
  - ci = nlparci(beta,resid,'jacobian',J)
- 预测值的置信区间
  - [beta,resid,J,Sigma]= nlinfit(X, y, fun, b0)
  - [ypred,delta] = nlpredci(fun,x,beta,resid,'covar',Sigma) 或
  - [ypred,delta] = nlpredci(fun,x,beta,resid,'jacobian',J)
  - 求nlinfit所得的回归方程在x处的预测值ypred及预测值的置信水平为1- alpha的置信区间ypred±delta；alpha缺省时为0.05.

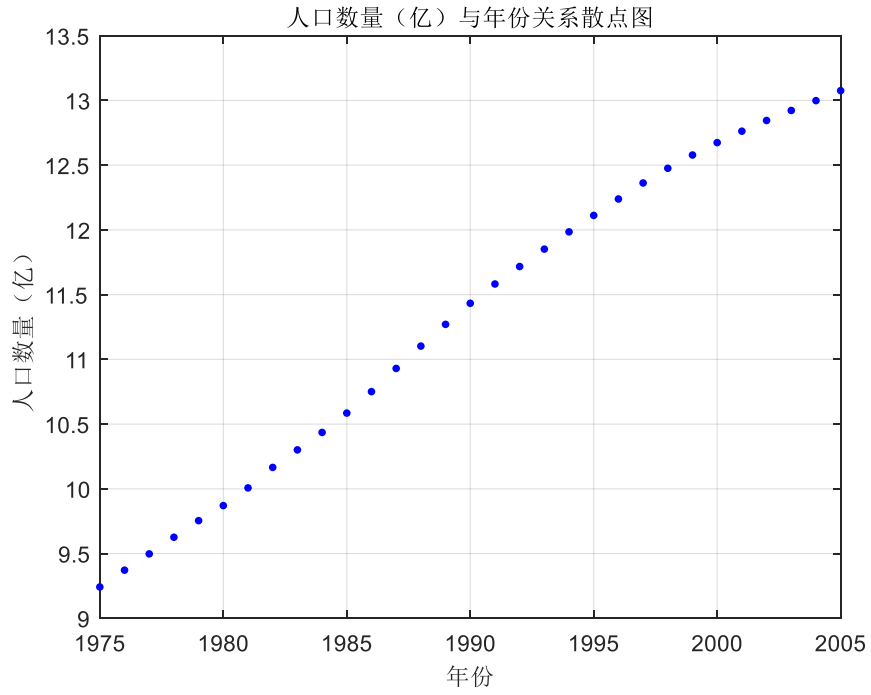
案例分析：人口增长模型

例：根据经验，人口增长的预测模型通常采用Logistic函数 $y(t)=\frac{A}{1+Be^{Ct}}$，其中$y(t)$为t 时刻人口数，A，B，C 为常数。试根据1975-2005年的中国人口数据，得出中国人口增长预测模型。

年份	时间	人口（亿）	年份	时间	人口（亿）	年份	时间	人口（亿）
1975	0	9.242	1986	11	10.751	1996	21	12.239
1976	1	9.3717	1987	12	10.93	1997	22	12.363
1977	2	9.4974	1988	13	11.103	1998	23	12.476
1978	3	9.6259	1989	14	11.27	1999	24	12.579
1979	4	9.7542	1990	15	11.433	2000	25	12.674
1980	5	9.8705	1991	16	11.582	2001	26	12.763
1981	6	10.007	1992	17	11.717	2002	27	12.845
1982	7	10.165	1993	18	11.852	2003	28	12.923
1983	8	10.301	1994	19	11.985	2004	29	12.999
1984	9	10.436	1995	20	12.112	2005	30	13.076
1985	10	10.585						

绘制散点图与模型回归

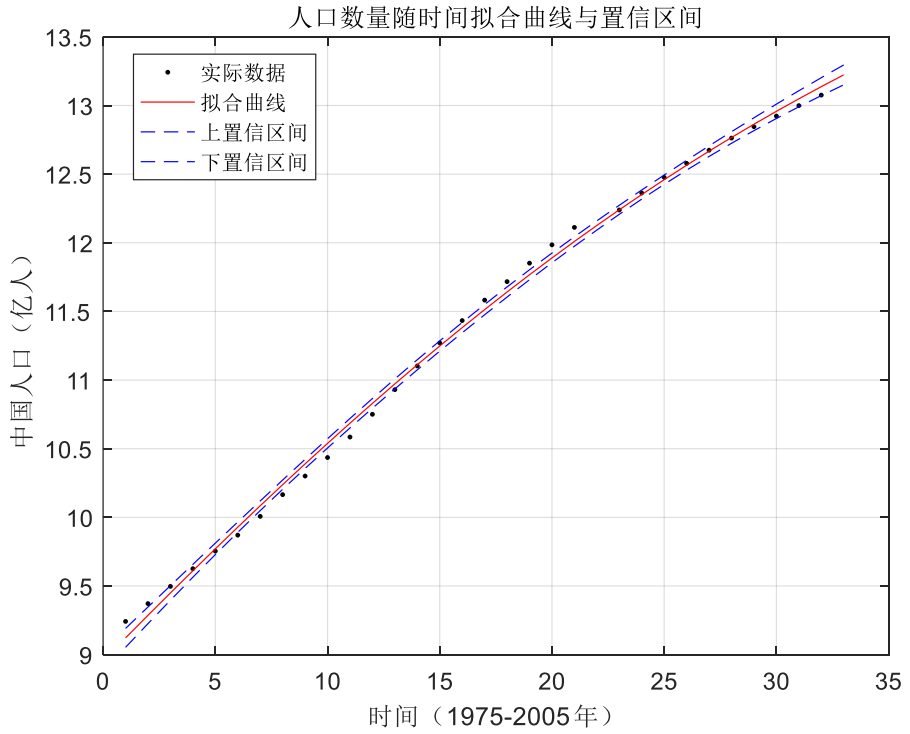
```
>> person = xlsread('renkou.xlsx');
>> ps = [person(:,3);person(:,6);person(:,9)];
>> year = [person(:,1);person(:,4);person(:,7)];
>> plot(year,ps,'b.','MarkerSize',10)
>> grid on;  xlabel('年份');  ylabel('人口数量（亿）')
>> title('人口数量（亿）与年份关系散点图')
```




```
%调用nlinfit函数作logistic回归的matlab程序
>> fun=@(beta,t)beta(1)./(1+beta(2)*exp(beta(3).*t));
>> t = 1:length(year);
>> [beta,resid,J,Sigma,mse] = nlinfit(t',ps,fun,[15,1,1]);
beta =
   15.4577    0.7259   -0.0442
>> yp=fun(beta,t');
>> ci = nlparci(beta,resid,'covar',Sigma)
ci =
   14.8318   16.0836
    0.6656    0.7863
   -0.0493   -0.0390
>> [ypred,delta] = nlpredci(fun,t,beta,resid,'covar',Sigma);
```

置信区间绘图与模型F检验

```
>> plot(t,ps,'k.',t,ypred,'r-',t,ypred-delta,'b--',t,ypred+delta,'b--');
>> grid on
>> xlabel('时间（1975-2005年）'); ylabel('中国人口（亿人）')
>> legend('实际数据','拟合曲线','上置信区间','下置信区间')
>> title('人口数量随时间拟合曲线与置信区间')
```




```
ybar = mean(y);
n = length(t);
SSR1 = sum((ypred-ybar).^2);
MSR1 = SSR1/3;
SSE1 = sum((y-ypred).^2);
MSE1 = SSE1/(n-3) ;
r2 = SSR1/(SSR1+SSE1)
fvalue1 = MSR1/MSE1
falpha1 = finv(0.95,3,n-3)
pvalue1 = 1-fcdf(fvalue1,3,n-3)
```

```
%运行结果
r2 =
    0.9974
fvalue1 =
   3.5476e+03
falpha1 =
    2.9467
pvalue1 =
     0
```

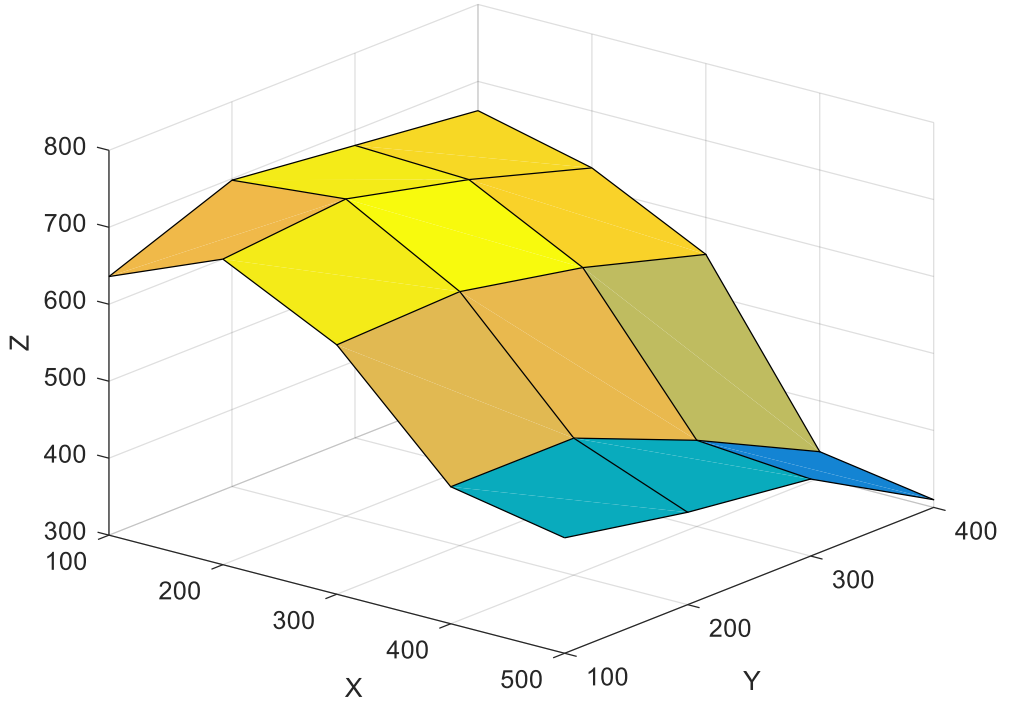
可知模型非常显著（$p = 0$），得到的Logistic函数表达式为：

$$y(t) = \frac{16.1634}{1 + 0.7712e^{-0.0408t}}$$

案例分析：曲面拟合

例：在一丘陵地带测量高程，x和y方向每隔100米测一个点，得高程如下表，试拟合一曲面，确定合适的模型，并由此找出最高点和该点的高程。

高程		x				
		100	200	300	400	500
y	100	636	697	624	478	450
	200	698	712	630	478	420
	300	680	674	598	412	400
	400	662	626	552	334	310



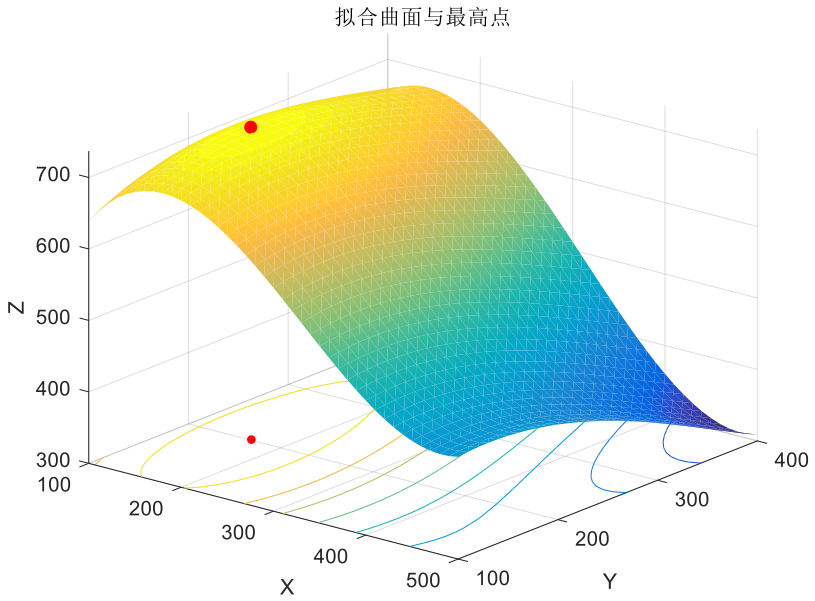

```
z=[636 697 624 478 450;698 712 630 478 420;680 674 598 412 400;662 626 552 334 310];
[x,y]=meshgrid(100:100:500,100:100:400);  %生成网格点
surf(x,y,z)  %对实际数据绘制网格面
view(39,26)  %改变视角
```

建立模型并预测

```
xy=[x(:),y(:)];  %组合成两列
zd=z(:); %列向量
fun_gc=@(b,t)b(1)*t(:,1) + b(2)*t(:,2) + b(3)*t(:,1).^2 + b(4)*t(:,1).*t(:,2) + ...
    b(5)*t(:,2).^2 + b(6)*t(:,1).^3 + b(7)*t(:,1).^2.*t(:,2) + ...
    b(8)*t(:,1).*t(:,2).^2 + b(9)*t(:,2).^3 + b(10);
[beta,resid,J,Sigma,mse] = nlinfit(xy,zd,fun_gc,0.1*ones(10,1))

syms x y
fh = vpa(beta(1)*x+beta(2)*y+beta(3)*x^2+beta(4)*x*y+beta(5)*y^2+...
    beta(6)*x^3+beta(7)*x^2*y+beta(8)*x*y^2+beta(9)*y^3+beta(10),6);
sol = solve(diff(fh,x)==0,diff(fh,y)==0,x,y);   %偏导并求方程组的解
for i = 1:4
    maxy(i) = subs(subs(fh,x,sol.x(i)),y,sol.y(i));
end
```

通过求偏导，然后由偏导等于0解得最高点为(167.2419 200.6160)，最高点处的高程为731.6817.

```
[xi,yi] = meshgrid(100:10:500,100:10:400);
xydat = [xi(:),yi(:)];
zi = reshape(fun_gc(beta,xydat),size(xi));  %求z并重排
surfc(xi,yi,zi); shading interp   %绘制曲面
view(39,26)  %改变视角
hold on
plot3(sol.x(1),sol.y(1),maxy(1)+5,'r.','MarkerSize',20)
plot3(sol.x(1),sol.y(1),300,'r.','MarkerSize',12)
```



lsqnonlin与lsqcurvefit

- nonlinfit, lsqnonlin, lsqcurvefit在功能上是类似的，但对于拟合过程的控制、输出参数的种类等有所不同，对于初学者而言，掌握三个函数的任意一个即可。
- [X,RESNORM,RESIDUAL,EXITFLAG,OUTPUT,LAMBDA] = lsqnonlin(fun,x0,lb,ub,options)
  - fun是事先用 m-文件定义的待拟合的非线性函数；
  - x0是回归系数的初值；lb, ub是回归参数的上下界
  - options是回归参数选项
- [X,RESNORM,RESIDUAL,EXITFLAG,OUTPUT,LAMBDA] = lsqcurvefit(fun,x0,xd,yd,lb,ub,options)

例：已知数据，拟合如下函数形式的曲线：$y = a_1 e^{a_2 x}$

X	1	2	3	4	5	6	7	8
Y	15.3	20.5	27.4	36.6	49.1	65.6	87.8	117.6

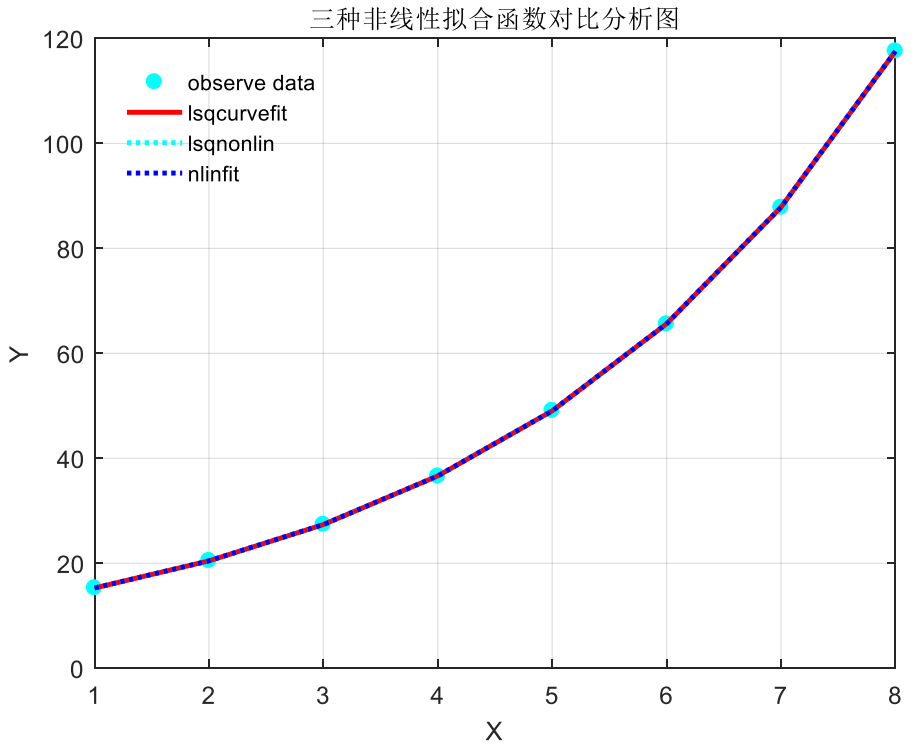
调用nlinfit、lsqnonlin、lsqcurvefit函数作非线性回归

lsqnonlin与lsqcurvefit

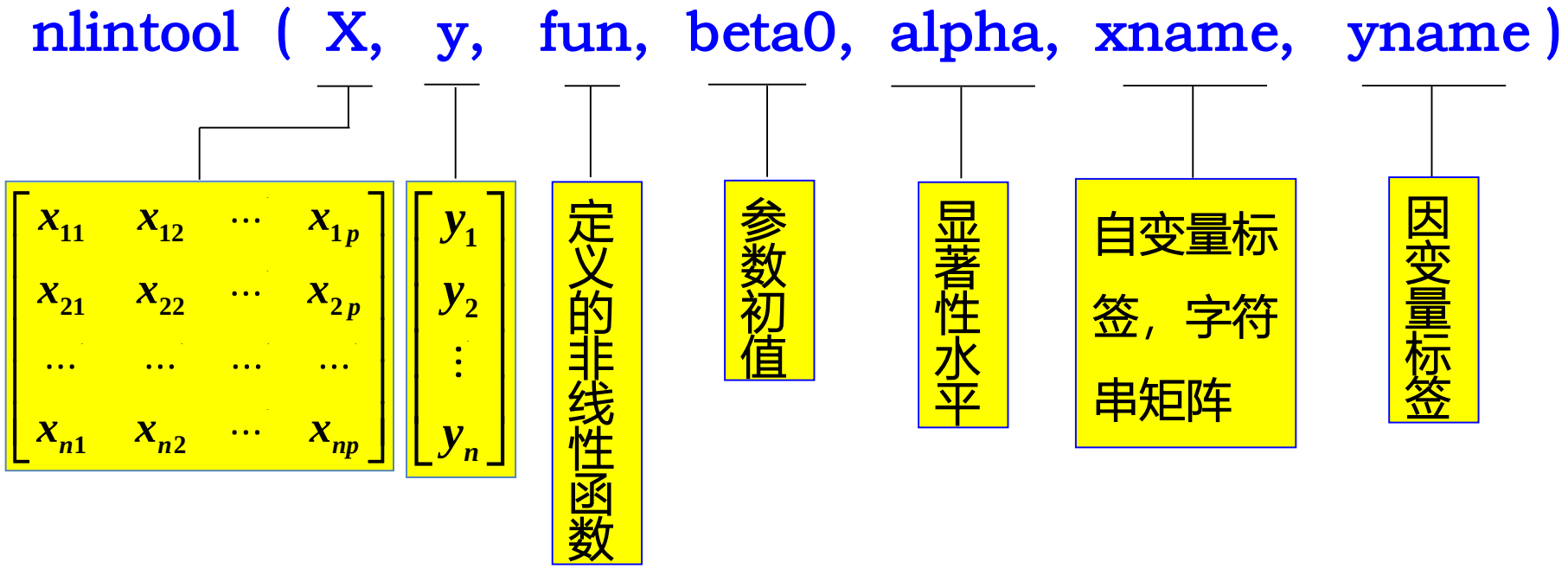
```
x=1:8;  %实际数据
y=[15.3 20.5 27.4 36.6 49.1 65.6 87.8 117.6];
objfun1=@(a,x)a(1)*exp(a(2)*x);  %模型
objfun2=@(a)a(1)*exp(a(2)*x)-y;
a0=[1,1];  %初始估计值
a1=lsqcurvefit(objfun1,a0,x,y)
a2=lsqnonlin(objfun2,a0)
a3=nlinfit(x,y,objfun1,a0)
```

```
a1 =  11.4241    0.2914
a2 =  11.4241    0.2914
a3 = 11.4241    0.2914
```

```
yp1=objfun1(a1,x);  yp2=objfun2(a2)+y;  yp3=objfun1(a3,x);
plot(x,y,'co','MarkerFaceColor','c');  hold on;  grid on
plot(x,yp1,'r',x,yp2,'c:',x,yp3,'b:','LineWidth',2);
legend('observe data','lsqcurvefit','lsqnonlin','nlinfit')
```



非线性拟合交互式工具nlintool

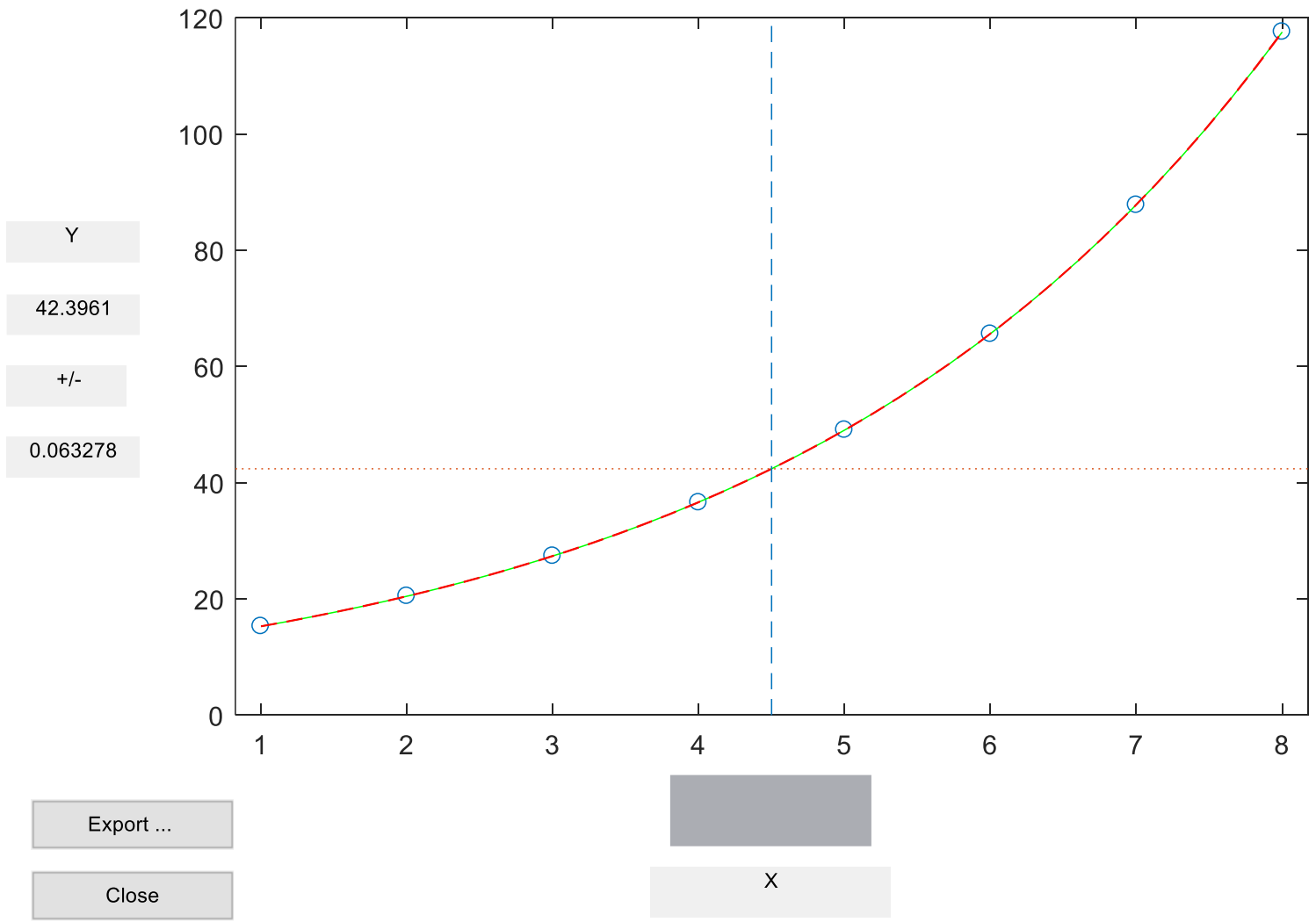


对于上例可以利用交互式工具nlintool进行拟合

nlintool(x,y, objfun1,a0,0.05,'X','Y');

非线性拟合交互式工具nlintool

信阳师范学院
数学与统计学院
SCHOOL OF MATHEMATICS AND STATISTICS




01
IDEA
02
LEARNING
03
CREATIVE
INSPIRATION
EDUCATION
BUSINESS
PHOTO BOOK

感谢聆听