实验指导书： 回归分析模型

【实验目的】

1、通过本次试验掌握回归分析的基本思想和基本方法，理解最小二乘法的计算步骤，理解模型的设定 T 检验，并能够根据检验结果对模型的合理性进行判断，进而改进模型。理解残差分析的意义和重要性，会对模型的回归残差进行正态性和独立性检验，从而能够判断模型是否符合回归分析的基本假设。

2、掌握回归分析的几种求解方法。

【实验相关知识】

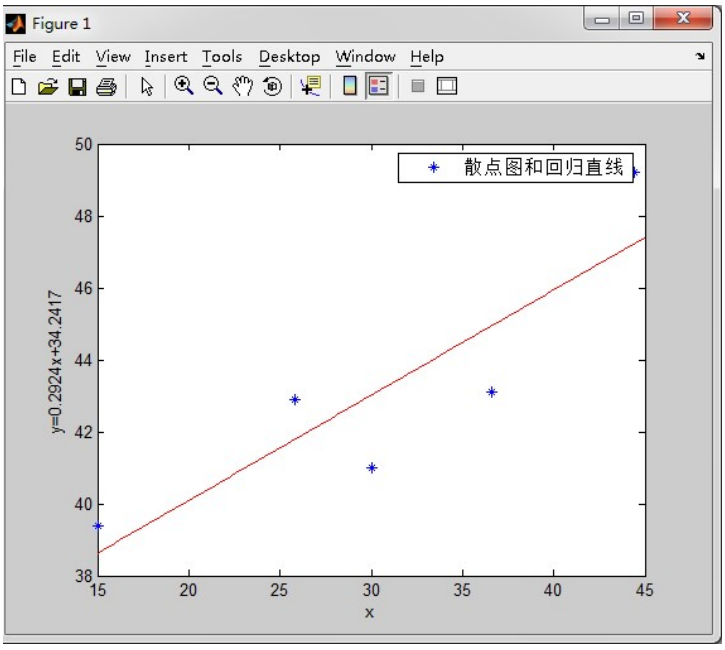
初识线性回归：

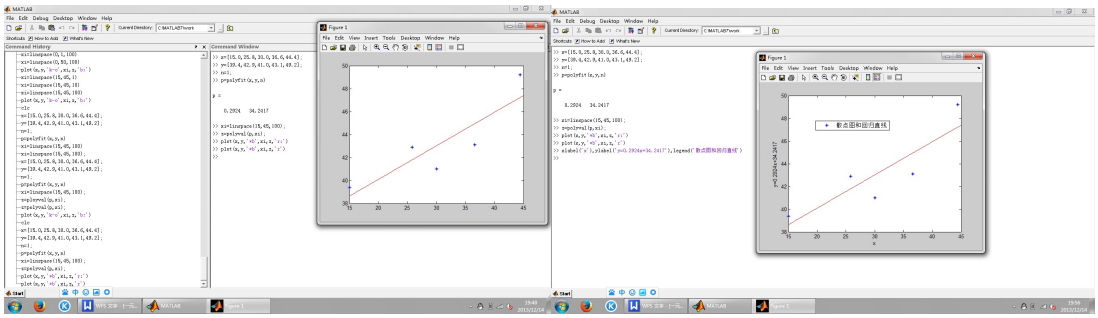
例 1：利用 matlab 计算一元回归分析函数如下所示，根据试验数据 x 和 y，可在 matlab 中利用一元拟合实现回归直线的计算、相关系数计算、同时得出一个散点图和回归直线图。

Matlab 程序代码如下：

```
x=[15.0,25.8,30.0,36.6,44.4]
y=[39.4,42.9,41.0,43.1,49.2]      x，y 以数组的形式输入
n=1;                              拟合阶数设置为 1
p=polyfit(x,y,n)                  计算拟合系数 a，b
xi=linspace(15,45,100);           规定拟合曲线横轴的分度
z=polyval(p,xi) ;                 虚拟出拟合曲线上的点
plot(x,y,'*b',xi,z,'r')           画出散点图和回归直线图
xlabel('x'),ylabel('y=0.2924x+34.2417'),legend('散点图和回归直线')
```

散点图：





一、多元线性回归（含一元线性回归）

多元线性回归：$y=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$

1、确定回归系数的点估计值：

命令为：b=regress(Y, X)

①b 表示 $b=\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \ldots\\ \hat{\beta}_p\end{bmatrix}$，②Y 表示 $Y=\begin{bmatrix}Y_1\\ Y_2\\ \ldots\\ Y_n\end{bmatrix}$，③X 表示 $X=\begin{bmatrix}1 & x_{11} & x_{12} & \ldots & x_{1p}\\ 1 & x_{21} & x_{22} & \ldots & x_{2p}\\ \ldots & \ldots & \ldots & \ldots & \ldots\\ 1 & x_{n1} & x_{n2} & \ldots & x_{np}\end{bmatrix}$

2、求回归系数的点估计和区间估计、并检验回归模型：

命令为：[b, bint,r,rint,stats]=regress(Y,X,alpha)

①bint 表示回归系数的区间估计.

②r 表示残差.

③rint 表示置信区间.

④stats 表示用于检验回归模型的统计量,有三个数值：相关系数 r^2、F 值、与 F 对应的概率 p.

说明：相关系数 r^2 越接近 1,说明回归方程越显著；$F>F_{1-\alpha}(k,n-k-1)$ 时拒绝 H_0,F 越大,

说明回归方程越显著；与 F 对应的概率 p< α 时拒绝 H_0,回归模型成立.

⑤alpha 表示显著性水平(缺省时为 0.05)

3、画出残差及其置信区间.

命令为：rcoplot(r,rint)

例 2.如下程序.

解：(1)输入数据.

```
x=[143 145 146 147 149 150 153 154 155 156 157 158 159 160 162 164]';
X=[ones(16,1) x];
Y=[88 85 88 91 92 93 93 95 96 98 97 96 98 99 100 102]';
```

(2)回归分析及检验.

```
[b,bint,r,rint,stats]=regress(Y,X)
b,bint,stats
```

得结果：b =　　bint =

-16.0730　　-33.7071　　1.5612

0.7194　　0.6047　　0.8340

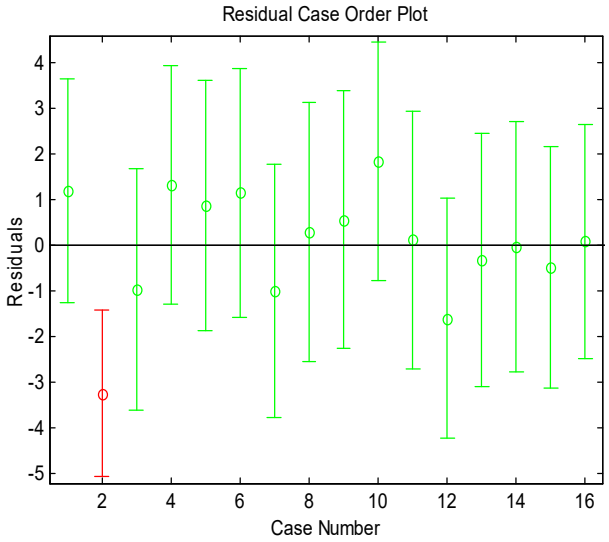
stats =

0.9282 180.9531 0.0000

即 $\hat{\beta}_0 = -16.073, \hat{\beta}_1 = 0.7194$；$\hat{\beta}_0$ 的置信区间为[-33.7017,1.5612]，$\hat{\beta}_1$ 的置信区间为[0.6047,0.834]; r^2=0.9282，F=180.9531，p=0.0000, 我们知道 p<0.05 就符合条件，可知回归模型 y=-16.073+0.7194x 成立.

(3)残差分析,作残差图.

rcoplot(r,rint)



从残差图可以看出,除第二个数据外,其余数据的残差离零点均较近,且残差的置信区间均包含零点,这说明回归模型 y=-16.073+0.7194x 能较好的符合原始数据,而第二个数据可视为异常点.

(4)预测及作图.

z=b(1)+b(2)*x

plot(x,Y,'k+',x,z,'r')

二、多项式回归

(一)一元多项式回归.

1、一元多项式回归：$y = a_1 x_m + a_2 x_{m-1} + ... + a_m x + a_{m+1}$

(1)确定多项式系数的命令：[p,S]=polyfit(x,y,m)

说明：$x=(x_1,x_2,\ldots,x_n)$,$y=(y_1,y_2,\ldots,y_n)$；$p=(a_1,a_2,\ldots,a_{m+1})$是多项式 $y=a_1x^m+a_2x^{m-1}+\ldots+a_mx+a_{m+1}$ 的系数；S 是一个矩阵,用来估计预测误差.

(2)一元多项式回归命令：polytool(x,y,m)

2、预测和预测误差估计.

(1)Y=polyval(p,x)求 polyfit 所得的回归多项式在 x 处的预测值 Y；

(2)[Y,DELTA]=polyconf(p,x,S,alpha)求 polyfit 所得的回归多项式在 x 处的预测值 Y 及预测值的显著性为 1-alpha 的置信区间 Y±DELTA；alpha 缺省时为 0.5.

例 3. 观测物体降落的距离 s 与时间 t 的关系,得到数据如下表,求 s. (关于 t 的回归方程 $\hat{s} = a + bt + ct^2$)

t (s)	1/30	2/30	3/30	4/30	5/30	6/30	7/30
s (cm)	11.86	15.67	20.60	26.69	33.71	41.93	51.13
t (s)	8/30	9/30	10/30	11/30	12/30	13/30	14/30
s (cm)	61.49	72.90	85.44	99.08	113.77	129.54	146.48

解法一：直接作二次多项式回归.

```
t=1/30:1/30:14/30;
s=[11.86 15.67 20.60 26.69 33.71 41.93 51.13 61.49 72.90 85.44 99.08 113.77 129.54 146.48];
[p,S]=polyfit(t,s,2)
```

得回归模型为：

$$\hat{s} = 489.2946t^2 + 65.8896t + 9.1329$$

解法二：化为多元线性回归.

```
t=1/30:1/30:14/30;
s=[11.86 15.67 20.60 26.69 33.71 41.93 51.13 61.49 72.90 85.44 99.08 113.77 129.54 146.48];
T=[ones(14,1) t' (t.^2)'];
[b,bint,r,rint,stats]=regress(s',T);
b,stats
```

得回归模型为：

$$\hat{s} = 9.1329 + 65.8896t + 489.2946t^2$$

预测及作图：

```
Y=polyconf(p,t,S)
plot(t,s,'k+',t,Y,'r')
```

(二)多元二项式回归

多元二项式回归命令：rstool(x,y,'model', alpha)

说明：x 表示 n×m 矩阵；Y 表示 n 维列向量；alpha：显著性水平(缺省时为 0.05)；model 表示由下列 4 个模型中选择 1 个(用字符串输入,缺省时为线性模型)：

linear(线性)：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$

purequadratic(纯二次)：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{n} \beta_{jj} x_j^2$

interaction(交叉)：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \ne k \le m} \beta_{jk} x_j x_k$

quadratic(完全二次)：$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j,k \le m} \beta_{jk} x_j x_k$

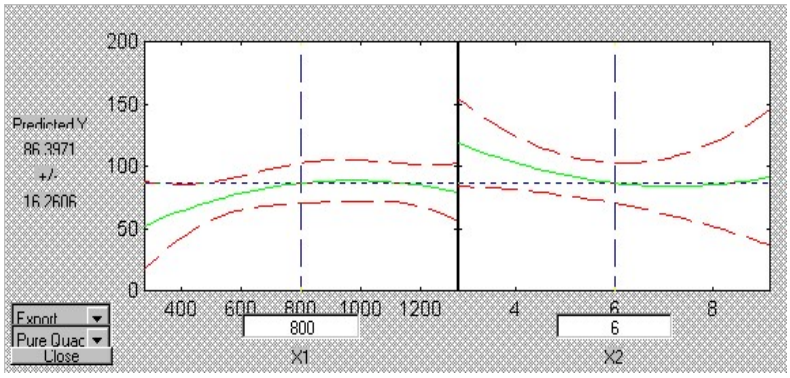
例 4. 设某商品的需求量与消费者的平均收入、商品价格的统计数据如下,建立回归模型,预测平均收入为 1000、价格为 6 时的商品需求量.

需求量	100	75	80	70	50	65	90	100	110	60
收入	1000	600	1200	500	300	400	1300	1100	1300	300
价格	5	7	6	6	8	7	5	4	3	9

解法一：选择纯二次模型,即 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$.

直接用多元二项式回归：

```
x1=[1000 600 1200 500 300 400 1300 1100 1300 300];
x2=[5 7 6 6 8 7 5 4 3 9];
y=[100 75 80 70 50 65 90 100 110 60]';
x=[x1' x2'];
rstool(x,y,'purequadratic')
```



在左边图形下方的方框中输入 1000,右边图形下方的方框中输入 6，则画面左边的“Predicted Y”下方的数据变为 88.47981,即预测出平均收入为 1000、价格为 6 时的商品需求量为 88.4791.

在画面左下方的下拉式菜单中选”all”, 则 beta、rmse 和 residuals 都传送到 Matlab 工作区中.

在 Matlab 工作区中输入命令：beta, rmse

得结果：beta =

110.5313
0.1464
-26.5709
-0.0001
1.8475

rmse =

4.5362

故回归模型为： $y = 110.5313 + 0.1464x_1 - 26.5709x_2 - 0.0001x_1^2 + 1.8475x_2^2$

剩余标准差为 4.5362, 说明此回归模型的显著性较好.

解法二：将 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$ 化为多元线性回归：

```
X=[ones(10,1) x1' x2' (x1.^2)' (x2.^2)'];
[b,bint,r,rint,stats]=regress(y,X);
b,stats
```

结果为: b =

110.5313
0.1464
-26.5709
-0.0001
1.8475

stats =

0.9702 40.6656 0.0005

三、非线性回归

1、非线性回归：

(1)确定回归系数的命令：[beta,r,J]=nlinfit(x,y,’model’, beta0)

说明：beta 表示估计出的回归系数；r 表示残差；J 表示 Jacobian 矩阵；x,y 表示输入数据 x、y 分别为矩阵和 n 维列向量,对一元非线性回归,x 为 n 维列向量；model 表示是事先用 m-文件定义的非线性函数；beta0 表示回归系数的初值.

(2)非线性回归命令：nlintool(x,y,'model', beta0,alpha)

2、预测和预测误差估计：

[Y,DELTA]=nlpredci('model', x,beta,r,J)

表示 nlinfit 或 nlintool 所得的回归函数在 x 处的预测值 Y 及预测值的显著性为 1-alpha 的置信区间 Y±DELTA.

例 5. 如下程序.

解：(1)对将要拟合的非线性模型 $y=a\,e^{b/x}$,建立 m-文件 volum.m 如下：

```
function yhat=volum(beta,x)
yhat=beta(1)*exp(beta(2)./x);
```

(2)输入数据：

```
x=2:16;
y=[6.42 8.20 9.58 9.5 9.7 10 9.93 9.99 10.49 10.59 10.60 10.80 10.60 10.90 10.76];
beta0=[8 2]';
```

(3)求回归系数：

```
[beta,r ,J]=nlinfit(x',y','volum',beta0);
      beta
```

(4)运行结果：

```
beta =
   11.6036
   -1.0641
```

即得回归模型为：

$$y=11.6036e^{-\frac{1.10641}{x}}$$

(5)预测及作图：

```
[YY,delta]=nlpredci('volum',x',beta,r ,J);
plot(x,y,'k+',x,YY,'r')
```

四、逐步回归

1、逐步回归的命令：stepwise(x,y,inmodel,alpha)

说明：x 表示自变量数据,$n\times m$ 阶矩阵；y 表示因变量数据,$n\times 1$ 阶矩阵；inmodel 表示矩阵的列数的指标,给出初始模型中包括的子集(缺省时设定为全部自变量)；alpha 表示显著性水平(缺省时为 0.5).

2、运行 stepwise 命令时产生三个图形窗口：Stepwise Plot,Stepwise Table,Stepwise History.

在 Stepwise Plot 窗口,显示出各项的回归系数及其置信区间.

(1)Stepwise Table 窗口中列出了一个统计表,包括回归系数及其置信区间,以及模型的统计量剩余标准差(RMSE)、相关系数(R-square)、F 值、与 F 对应的概率 P.

例 6 水泥凝固时放出的热量 y 与水泥中 4 种化学成分 x1、x2、x3、 x4 有关,今测得一组数据如下,试用逐步回归法确定一个线性模型.

序号	1	2	3	4	5	6	7	8	9	10	11	12	13
x_1	7	1	11	11	7	11	3	1	2	21	1	11	10

x_2	26	29	56	31	52	55	71	31	54	47	40	66	68
x_3	6	15	8	8	6	9	17	22	18	4	23	9	8
x_4	60	52	20	47	33	22	6	44	22	26	34	12	12
y	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

解：(1)数据输入：

```
x1=[7 1 11 11 7 11 3 1 2 21 1 11 10]';
x2=[26 29 56 31 52 55 71 31 54 47 40 66 68]';
x3=[6 15 8 8 6 9 17 22 18 4 23 9 8]';
x4=[60 52 20 47 33 22 6 44 22 26 34 12 12]';
y=[78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5 93.1 115.9 83.8 113.3 109.4]';
x=[x1 x2 x3 x4];
```

(2)逐步回归.

①先在初始模型中取全部自变量：stepwise(x,y)

得图 Stepwise Plot 和表 Stepwise Table.

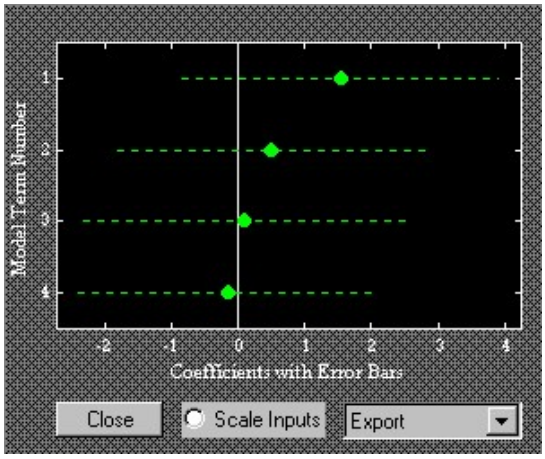
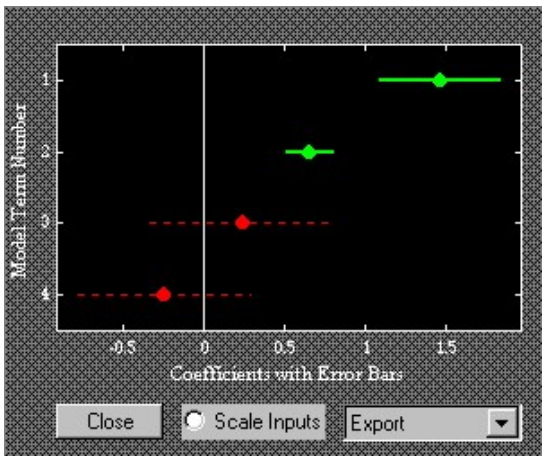


图 Stepwise Plot 中四条直线都是虚线,说明模型的显著性不好.

从表 Stepwise Table 中看出变量 x3 和 x4 的显著性最差.

②在图 Stepwise Plot 中点击直线 3 和直线 4,移去变量 x3 和 x4.



移去变量 x3 和 x4 后模型具有显著性

虽然剩余标准差(RMSE)没有太大的变化,但是统计量 F 的值明显增大,因此新的回归模型更好.

(3)对变量 y 和 x1、x2 作线性回归.

```
X=[ones(13,1) x1 x2];
b=regress(y,X)
```

得结果：b =

52.5773

1.4683

0.6623

故最终模型为：y=52.5773+1.4683x1+0.6623x2