

第四节 多元线性回归

- 一、多元线性回归的数学模型
- 二、数学模型的分析与求解
- 三、*MATLAB*中回归分析的实现
- 四、小结



一、多元线性回归的数学模型

实际问题中的随机变量 Y 通常与多个普通变量 x_1, x_2, \dots, x_p ($p > 1$) 有关.

对于自变量 x_1, x_2, \dots, x_p 的一组确定值, Y 具有一定的分布, 若 Y 的数学期望存在, 则它是 x_1, x_2, \dots, x_p 的函数.

$$\mu_{Y|x_1, x_2, \dots, x_p} = \mu(x_1, x_2, \dots, x_p)$$

Y 关于 x 的回归函数



如果 $\mu(x_1, x_2, \dots, x_p)$ 是 x_1, x_2, \dots, x_p 的线性函数.

$$Y = b_0 + b_1 x_1 + \dots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$b_0, b_1, \dots, b_p, \sigma^2$ 是与 x_1, \dots, x_p 无关的未知参数.

多元线性回归模型



二、数学模型的分析与求解

设 $(x_{11}, x_{12}, \dots, x_{1p}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{np}, y_n)$
是一个样本。

用最大似然估计法估计参数：

1、类似与一元线性回归情况，由联合概率密度函数确定最大似然函数；

2、由于服从正态分布，似然函数值最大，即其指

数部分最小。 $Q = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2$



二、数学模型的分析与求解

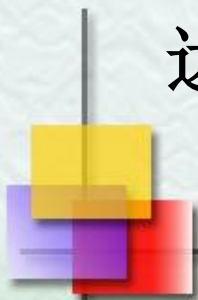
设 $(x_{11}, x_{12}, \dots, x_{1p}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{np}, y_n)$
是一个样本.

用最大似然估计法估计参数.

取 $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_p$, 当 $b_0 = \hat{b}_0, b_1 = \hat{b}_1, \dots, b_p = \hat{b}_p$ 时,

$$Q = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_p x_{ip})^2$$

达到最小.



$$Q = \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2,$$

$$\left. \begin{aligned} \frac{\partial Q}{\partial b_0} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}) = 0, \\ \frac{\partial Q}{\partial b_j} &= -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip}) x_{ij} = 0, \\ j &= 1, 2, \cdots, p. \end{aligned} \right\}$$

化简可得



$$\left. \begin{aligned}
 & b_0 n + b_1 \sum_{i=1}^n x_{i1} + b_2 \sum_{i=1}^n x_{i2} + \cdots + b_p \sum_{i=1}^n x_{ip} = \sum_{i=1}^n y_i, \\
 & b_0 \sum_{i=1}^n x_{i1} + b_1 \sum_{i=1}^n x_{i1}^2 + b_2 \sum_{i=1}^n x_{i1} x_{i2} + \cdots + b_p \sum_{i=1}^n x_{i1} x_{ip} \\
 & \quad = \sum_{i=1}^n x_{i1} y_i, \\
 & \quad \vdots \\
 & b_0 \sum_{i=1}^n x_{ip} + b_1 \sum_{i=1}^n x_{ip} x_{i1} + b_2 \sum_{i=1}^n x_{ip} x_{i2} + \cdots + b_p \sum_{i=1}^n x_{ip}^2 \\
 & \quad = \sum_{i=1}^n x_{ip} y_i.
 \end{aligned} \right\}$$

正规方程组



引入矩阵

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}.$$

正规方程组的矩阵形式

$$X'XB = X'Y$$



正规方程组 $X'XB = X'Y$

$$\hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = (X'X)^{-1} X'Y \quad \text{最大似然估计值}$$

$\mu(x_1, x_2, \dots, x_p) = b_0 + b_1 x_1 + \dots + b_p x_p$ 的估计是

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \dots + \hat{b}_p x_p$$

P 元经验线性回归方程



三、*MATLAB*中回归分析的实现

多元线性回归

1.确定回归系数的点估计值,用命令:

$b = \text{regress}(Y, X)$

2.求回归系数的点估计和区间估计,并检验回归模型,用命令:

$[b, bint, r, rint, stats] = \text{regress}(Y, X, \alpha)$

3.画出残差及其置信区间,用命令:

$\text{rcoplot}(r, rint)$



符号说明

(1)

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

一元线性回归, 取 $p = 1$.

$$b = \hat{B} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = (X'X)^{-1} X'Y.$$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p$$



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

概率论与数理统计

(2) α 为显著性水平, 默认为 0.05;

(3) $bint$ 为回归系数的区间估计;

(4) r 与 $rint$ 分别为残差及其置信区间;

(5) $stats$ 是用于检验回归模型的统计量, 有三个数值, 第一个是相关系数 r^2 , 其值越接近于 1, 说明回归方程越显著; 第二个是 F 值, $F > F_{1-\alpha}(p, n-p-1)$ 时拒绝 H_0 , F 越大, 说明回归方程越显著; 第三个是与 F 对应的概率 p , $p < \alpha$ 时拒绝, 回归模型成立.



例1 测得16名女子的身高和腿长如下(单位:cm):

身高	143	145	146	147	149	150	153	154
腿长	88	85	88	91	92	93	93	95
身高	155	156	157	158	159	160	162	164
腿长	96	98	97	96	98	99	100	102

试研究这些数据之间的关系.



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]';
```

回归分析及检验

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




```

MATLAB
File Edit View Web Window Help
[Icons] ? Current Directory: H:\概率论!

>>
b =

    -16.0730
         0.7194

bint =

    -33.7071    1.5612
         0.6047    0.8340

stats =

    0.9282    180.9531    0.0000
    
```

$$\hat{b}_0 = -16.0730,$$

$$\hat{b}_1 = 0.7194.$$

\hat{b}_0 的置信区间 $(-33.7071, 1.5612)$.

\hat{b}_1 的置信区间 $(0.6047, 0.834)$.

$$r^2 = 0.9282, F = 180.9531,$$

$$p = 0.0000.$$

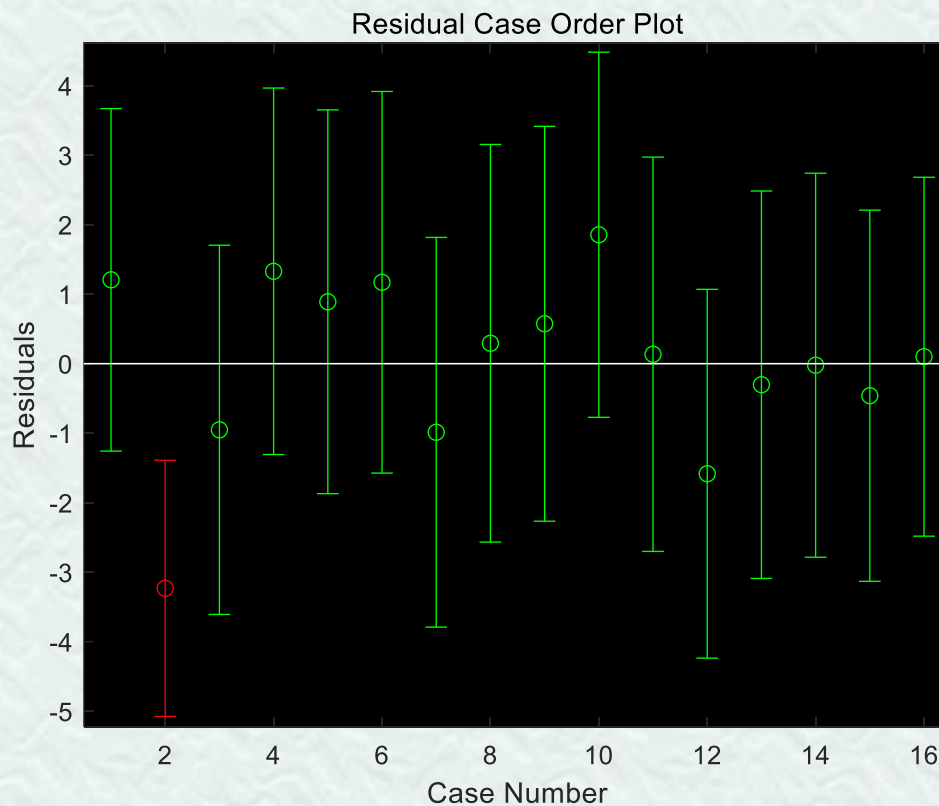
$p < 0.05$, 回归模型

$y = -16.0730 + 0.7194x$ 成立.

残差分析

`rcoplot(r,rint)`

画出残差和其置信区间

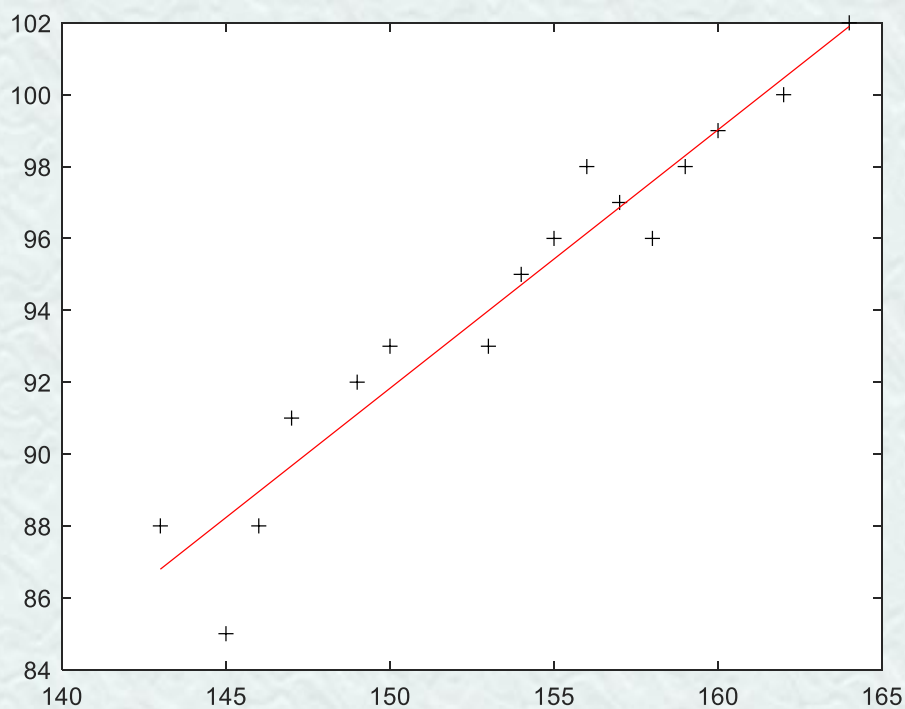


预测及作图

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

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

数据比较



一元多项式回归

1.确定多项式系数,用命令:

$$[p,S]=polyfit(x,y,m)$$

$$x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n).$$

$p = (a_1, a_2, \dots, a_{m+1})$ 确定多项式

$$y = a_1 x^m + a_2 x^{m-1} + \dots + a_m x + a_{m+1},$$

S 是一个矩阵,用来估计预测误差.

也可使用命令: $polytool(x,y,m)$

结果产生一个交互式的画面,画面中有拟合曲线和 y 的置信区间,左下方的 *Export* 可以输出参数.



2.预测和预测误差估计用命令:

$$Y=polyval(p,x)$$

求回归多项式在 x 处的预测值 Y .

$$[Y,DELTA]=polyconf(p,x,S,alpha)$$

求回归多项式在 x 处的预测值 Y 以及预测值的显著性为 $1-\alpha$ 的置信区间 $Y \pm DELTA, \alpha$ 的默认值是 0.05.

一元多项式回归可化为多元线性回归求解.



例2 下面给出了某种产品每件平均单价 Y (元) 与批量 x (件) 之间的关系的一组数据 .

x	20	25	30	35	40	50
y	1.81	1.70	1.65	1.55	1.48	1.40
x	60	65	70	75	80	90
y	1.30	1.26	1.24	1.21	1.20	1.18

试用一元二次多项式进行回归分析.



输入数据

$x=[20,25,30,35,40,50,60,65,70,75,80,90];$

$y=[1.81,1.70,1.65,1.55,1.48,1.40,1.30,1.26,1.24,1.21,$
 $1.20,1.18];$

作二次多项式回归

$[p,S]=polyfit(x,y,2)$

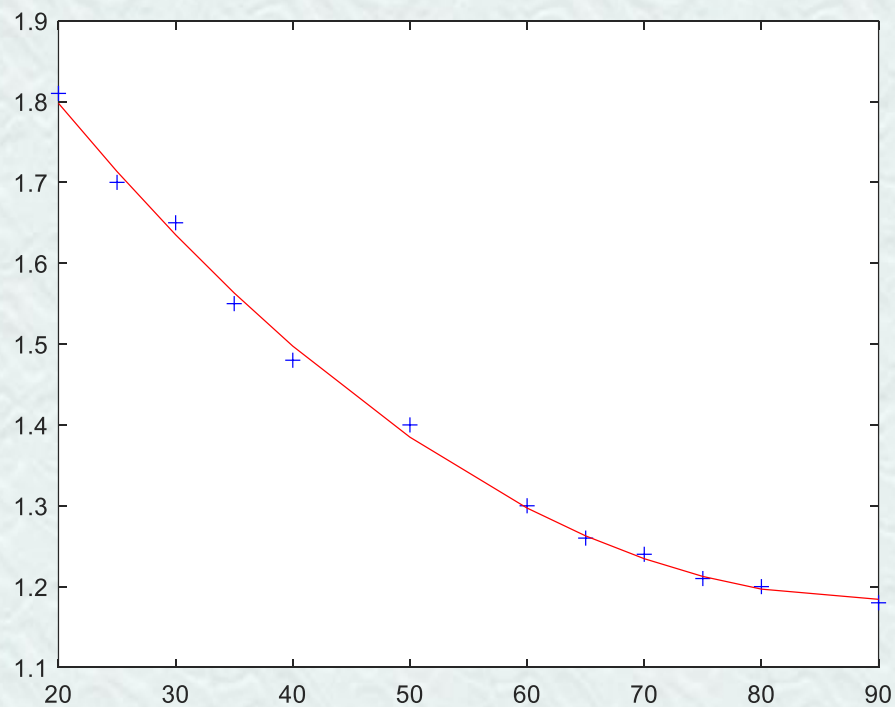
p =

0.0001 -0.0225 2.1983



预测及作图

$Y = \text{polyconf}(p, x, y)$
 $\text{plot}(x, y, 'b+', x, Y, 'r')$



化为多元线性回归

```
X=[ones(12,1) x' (x.^2)'];
[b,bint,r,rint,stats]=regress(y',X);
b,stats
```

```
>>
b =

    2.1983
   -0.0225
    0.0001

stats =

    1.0e+003 *

    0.0010    1.7674    0.0000
```

与前面结果一致



多元二项式回归

rstool(x,y,'model',alpha)

其中,输入数据 x, y 分别为 $n \times m$ 矩阵和 n 维列向量; $alpha$ 为显著性水平,默认为 0.05; $model$ 为下列四种模型中的一种,输入相应的字符串,默认为线性模型.

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

purequadratic (纯二次):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^m \beta_{jj} x_j^2$$

interaction (交叉):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j \neq k \leq m} \beta_{jk} x_j x_k$$



quadratic(完全二次):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \leq j, k \leq m}^m \beta_{jk} x_j x_k$$

*rstool*的输出是一个交互式画面,画面中有 m 个图形,分别给出了一个独立变量 x_i 与 y 的拟合曲线,以及 y 的置信区间,此时其余 $m-1$ 个变量取固定值.可以输入不同的变量的不同值得到 y 的相应值.

图的左下方有两个下拉式菜单,一个用于传送回归系数、剩余标准差、残差等数据;另一个用于选择四种回归模型中的一种,选择不同的回归模型,其中剩余标准差最接近于零的模型回归效果最好.



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



选择纯二次模型,即

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

数据输入

收入 $x1=[1000,600,1200,500,300,400,1300,1100,1300,300];$
 价格 $x2=[5,7,6,6,8,7,5,4,3,9];$
 需求量 $y=[100,75,80,70,50,65,90,100,110,60]';$
 $x=[x1' \ x2'];$

回归、检验与预测

$\text{rstool}(x,y,\text{'purequadratic'})$

程序运行结果

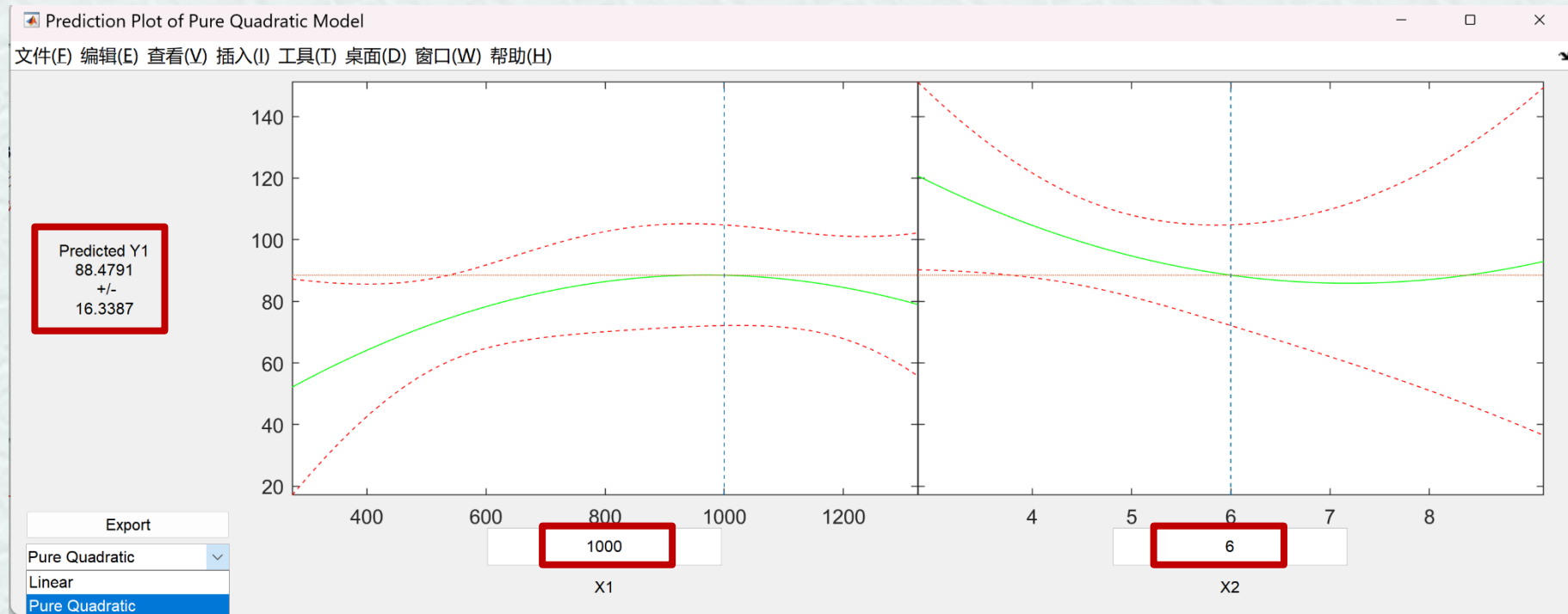
回归图形

回归结果

帮

助





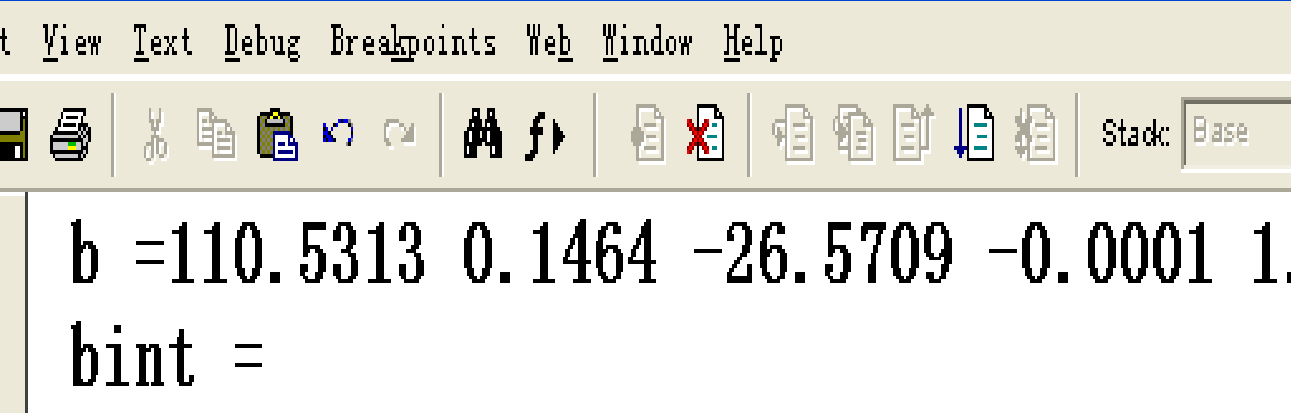
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

化为多元线性回归求解

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



回归系数的点估计以及区间估计





















The screenshot shows a debugger window titled "Untitled3*" with a menu bar (File, Edit, View, Text, Debug, Breakpoints, Web, Window, Help) and a toolbar. The main window displays assembly code and its output. The code defines a vector 'b' and a matrix 'bint'. The output shows the values of 'b' and 'bint'.

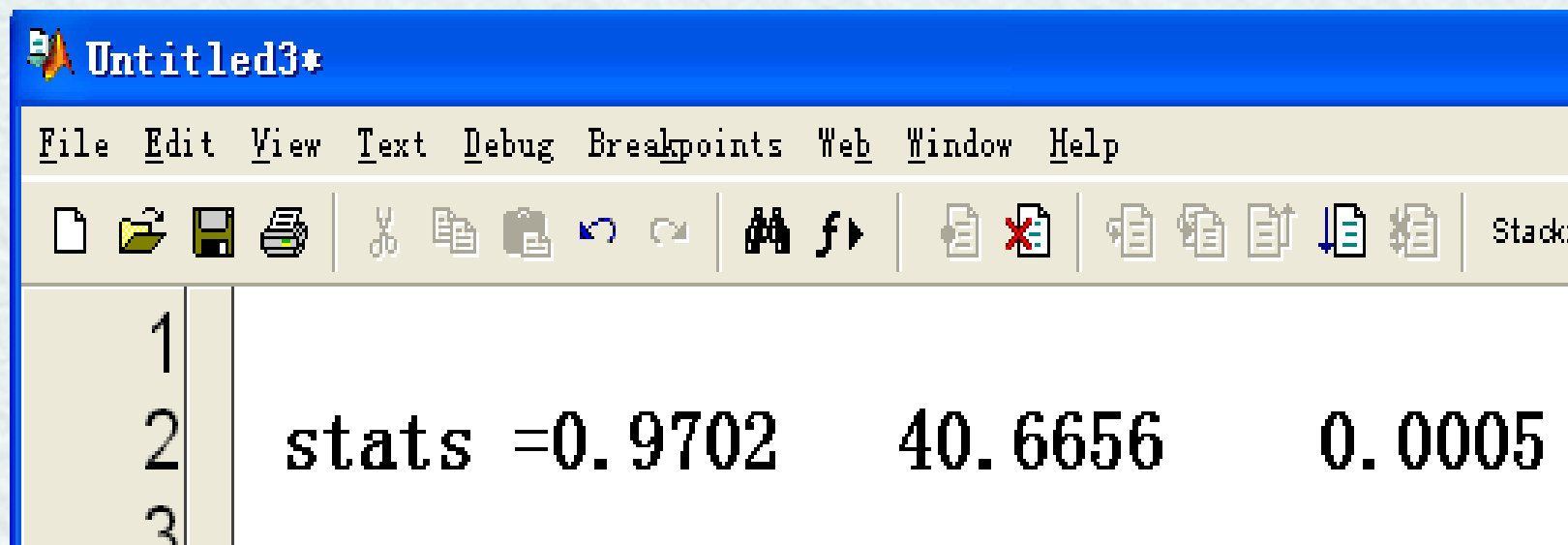
```
1  b =110.5313  0.1464 -26.5709 -0.0001  1.8475
2  bint =
3      57.2602  163.8024
4      0.0408   0.2521
5     -43.2247  -9.9171
6     -0.0001  -0.0000
7      0.3745   3.3205
```



残差及其置信区间

Untitled3*			
File Edit View Text Debug Breakpoints Web Window Help			
                 			
1	r =	rint =	
2	5.2724	-2.8991	13.4438
3	-0.7162	-10.7426	9.3103
4	-4.5158	-11.2788	2.2472
5	-1.9390	-11.3778	7.4997
6	-3.3315	-12.3214	5.6583
7	3.4566	-5.9980	12.9111
8	3.4843	-3.5514	10.5200
9	-3.4452	-13.0340	6.1437
10	-0.0976	-6.3831	6.1878
11	1.8320	-3.3221	6.9862

检验回归模型的统计量



The screenshot shows a software window with a menu bar (File, Edit, View, Text, Debug, Breakpoints, Web, Window, Help) and a toolbar. The main area displays the following statistics:

	stats	r^2	F	P
1				
2	stats = 0.9702	40.6656	0.0005	
3				

相关系数 $r^2 = 0.9702$,接近于1,回归方程显著;

$F = 40.6656 > F_{0.95}(4,5) = 6.26$,回归方程显著;

$P = 0.0005 < \alpha = 0.05$,回归模型成立.

逐步回归分析

在实际问题中,影响因变量的因素很多,而这些因素之间可能存在多重共线性.为得到可靠的回归模型,需要一种方法能有效地从众多因素中挑选出对因变量贡献大的因素.

如果采用多元线性回归分析,回归方程稳定性差,每个自变量的区间误差积累将影响总体误差,预测的可靠性差、精度低;另外,如果采用了影响小的变量,遗漏了重要变量,可能导致估计量产生偏倚和不一致性.



“**最优**”的回归方程应该包含所有有影响的变量而不包括影响不显著的变量.

选择“**最优**”回归方程的方法

- 1.从所有可能的变量组合的回归方程中选择最优者;
- 2.从包含全部变量的回归方程中逐次剔除不显著因子;
- 3.从一个变量开始,把变量逐个引入方程;
- 4.“**有进有出**”的**逐步回归分析**.



逐步回归分析法在筛选变量方面比较理想,是目前较常用的方法. 它从一个自变量开始,根据自变量作用的显著程度,从大到小地依次逐个引入回归方程,但当引入的自变量由于后面变量的引入而变得不显著时,要将其剔除掉. 引入一个自变量或从回归方程中剔除一个自变量,为逐步回归的一步,对于每一步,都进行检验,以确保每次引入新的显著性变量前回归方程中只包含作用显著的变量.

反复进行上面的过程,直到没有不显著的变量从回归方程中剔除,也没有显著变量可引入到回归方程.



函数: *stepwise*

用法: *stepwise(x,y,inmodel,alpha)*

符号说明:

x—自变量数据,为 $n \times m$ 矩阵;

y—因变量数据,为 $n \times 1$ 矩阵;

inmodel—由矩阵*x*列的指标构成,表明初始模型中引入的自变量,默认为全部自变量;

alpha—判断模型中每一项显著性的指标,默认相当于对回归系数给出95%的置信区间.



例4 水泥凝固时放出的热量 y 与水泥中的四种化学成分 x_1, x_2, x_3, x_4 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
x_1	7	1	11	11	7	11	3
x_2	26	29	56	31	52	55	71
x_3	6	15	8	8	6	9	17
x_4	60	52	20	47	33	22	6
y	78.5	74.3	104.3	87.6	95.9	109.2	102.7
序号	8	9	10	11	12	13	
x_1	1	2	21	1	11	10	
x_2	31	54	47	40	66	68	
x_3	22	18	4	23	9	8	
x_4	44	22	26	34	12	12	
y	72.5	93.1	115.9	83.8	113.3	109.4	

输入数据

```
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];
```

逐步回归分析

`stepwise(x,y)`

程序运行结果

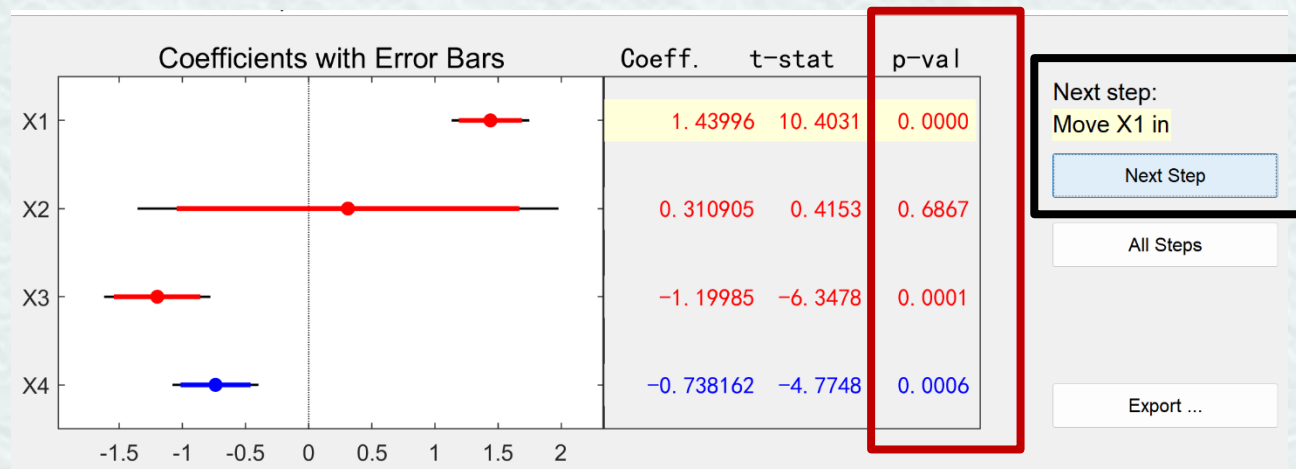
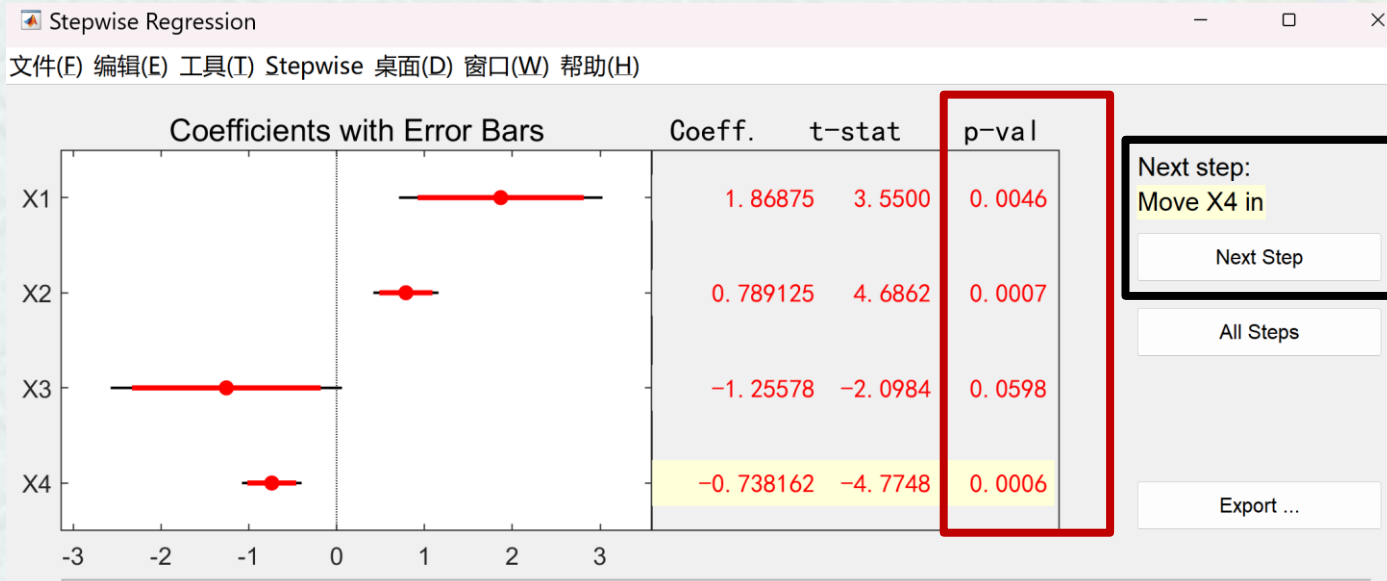
逐步回归

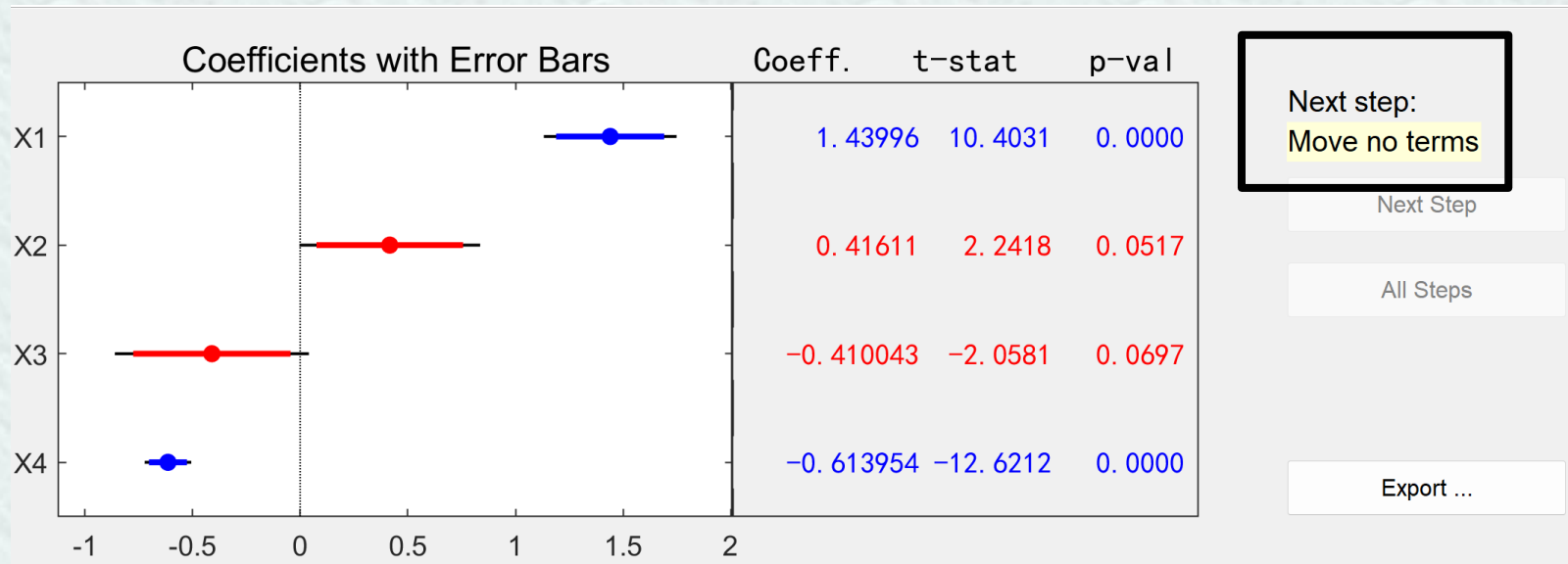
回归平面

帮

助







对变量 y 和 x_1, x_4 作线性回归.

$X=[\text{ones}(13,1),x1,x4];$

$[b,\text{bint},r,\text{rint},\text{stats}]=\text{regress}(y,X)$

```
>> b  
  
b =  
  
52.5773  
1.4683  
0.6623  
  
>> stats  
  
stats =  
  
0.9787 229.5037 0.0000
```

回归模型为 $y = 52.5773 + 1.4683x_1 + 0.6623x_4,$

三个统计量表明: 回归效果显著.



四、小结

1.多元线性回归的数学模型

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

$b_0, b_1, \cdots, b_p, \sigma^2$ 是与 x_1, \cdots, x_p 无关的未知参数.

2.数学模型的分析与求解

$$\hat{B} = (X'X)^{-1} X'Y,$$

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p.$$



3. *MATLAB*中回归分析的实现

(1) 多元线性回归

$$b = \text{regress}(Y, X)$$

(2) 一元多项式回归

$$[p, S] = \text{polyfit}(x, y, m)$$

(3) 多元二项式回归

$$\text{rstool}(x, y, 'model', \alpha)$$

(4) 逐步回归分析

$$\text{stepwise}(x, y, \text{inmodel}, \alpha)$$

