



第四章 回归分析





趋向中间高度的回归

- ▶ 回归这个术语是由英国著名统计学家葛尔登（Francis Galton）在19世纪末期研究孩子及他们的父母的身高时提出来的。Galton发现身材高的父母，他们的孩子也高。但这些孩子平均起来并不像他们的父母那样高。对于比较矮的父母情形也类似：他们的孩子比较矮，但这些孩子的平均身高要比他们的父母的平均身高高。Galton把这种孩子的身高向中间值靠近的趋势称之为一种回归效应，而他发展的研究两个数值变量的方法称为回归分析。

Galton公式: $y = 33.73 + 0.516x$

其中 x 表示父亲身高, y 表示成年儿子的身高
(单位: 英寸, 1英寸=2.54厘米)。

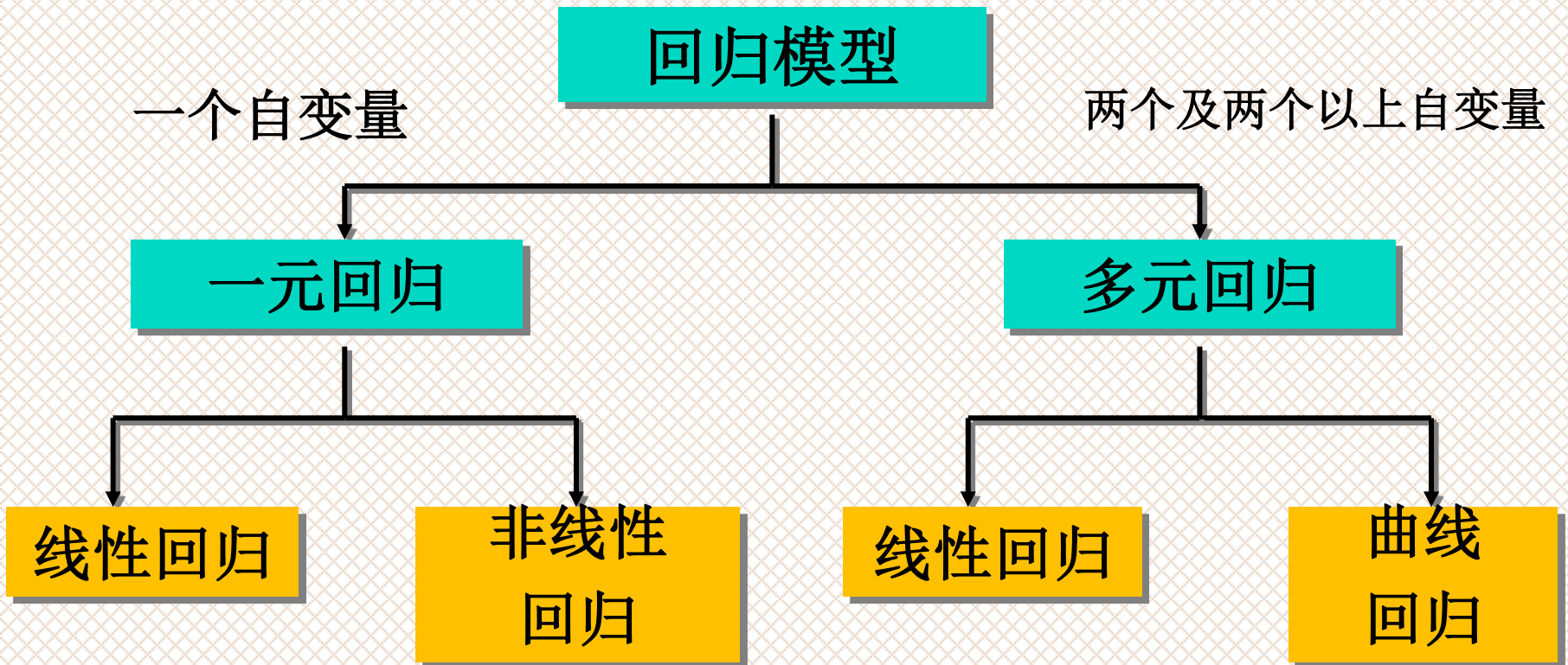
$y(\text{cm})$	$x(\text{cm})$
160.07	150
168.23	160
173.39	170
178.55	180
183.71	190
188.87	200
194.03	210

回归分析

回归分析在一组数据的基础上研究这样几个问题：

- (i) 建立因变量 y 与自变量 x_1, x_2, \dots, x_m 之间的回归模型（经验公式）；
- (ii) 对回归模型的可信度进行检验；
- (iii) 判断每个自变量 $x_i (i=1, 2, \dots, m)$ 对 y 的影响是否显著；
- (iv) 诊断回归模型是否适合这组数据；
- (v) 利用回归模型对 y 进行预报或控制。

回归模型的类型



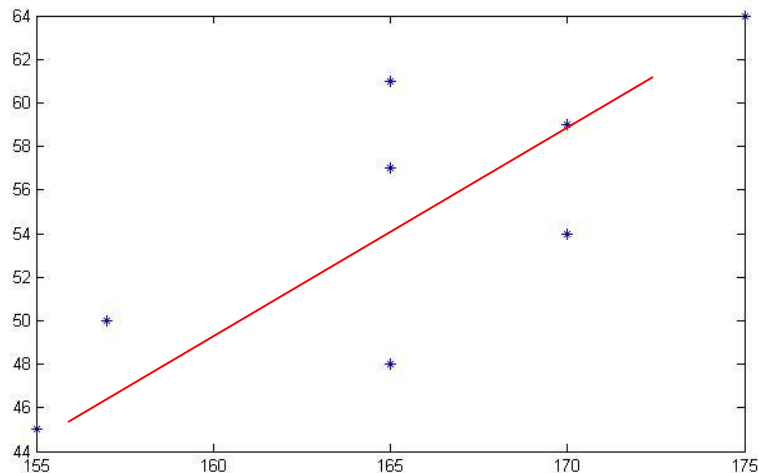
一、一元线性回归模型

例1 从某大学中随机选取8名女大学生，其身高和体重数据如表1-1所示。

编号	1	2	3	4	5	6	7	8
身高/cm	165	165	157	170	175	165	155	170
体重/kg	48	57	50	54	64	61	43	59

求根据一名女大学生的身高预报她的体重的回归方程，并预报一名身高为172cm的女大学生的体重。

解：1、选取身高为自变量 x ，体重为因变量 y ，作散点图：

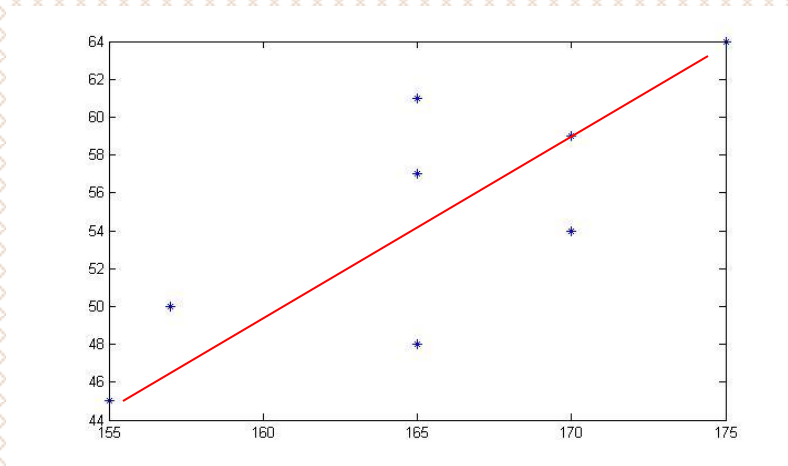


2.回归方程:

$$\hat{y} = 0.7831x - 74.6563$$

身高172cm的女大学生体重是

$$\hat{y} = 0.7831 \times 172 - 74.6563 = 60.0369 (kg)$$



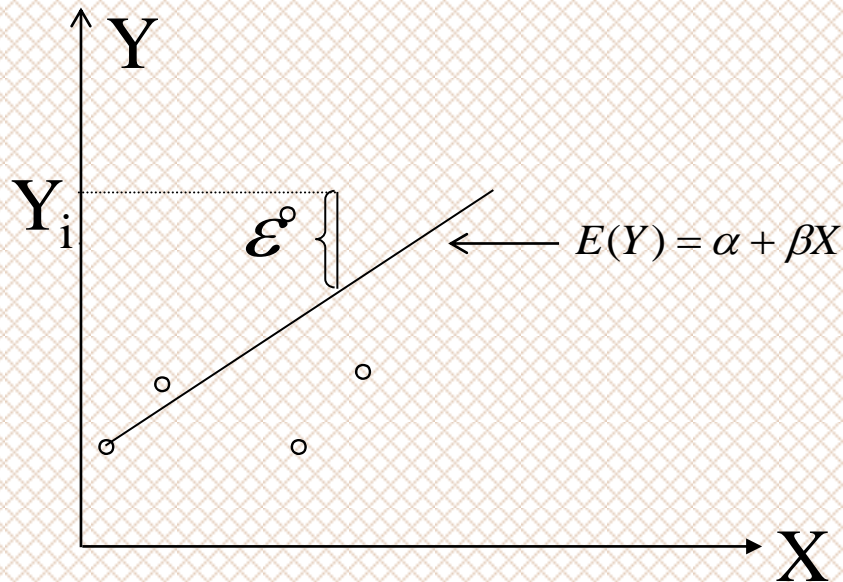
探究：身高为172cm的女大学生的体重一定是60.0369kg吗？如果不是，你能解析一下原因吗？

答：用这个回归方程不能给出每个身高为172cm的女大学生的体重的预测值，只能给出她们平均体重的估计值。因为散点图中的样本点散布在这条直线的附近，而不是在直线上，所以用一次函数求得的数据不是她体重的标准值。

由于所有的样本点不共线，而只是散布在某一直线的附近，所以身高和体重的关系可以用线性回归模型来表示：

$$y = bx + a + \varepsilon$$

其中 a 和 b 为模型的未知参数， ε 称为随机误差。





一般地，称由 $y = \beta_0 + \beta_1 x + \varepsilon$ 确定的模型为**一元线性回归模型**，记为

$$\begin{cases} y = \beta_0 + \beta_1 x + \varepsilon \\ E\varepsilon = 0, D\varepsilon = \sigma^2 \end{cases}$$

固定的未知参数 β_0 、 β_1 称为回归系数，自变量 x 也称为回归变量。

$Y = \beta_0 + \beta_1 x$ ，称为 **y 对 x 的回归直线方程**。

一元线性回归分析的主要任务是：

- 1、用试验值（样本值）对 β_0 、 β_1 和 σ 作点估计；
- 2、对回归系数 β_0 、 β_1 作假设检验；
- 3、在 $x = x_0$ 处对 y 作预测，对 y 作区间估计。

一元回归的Matlab实现

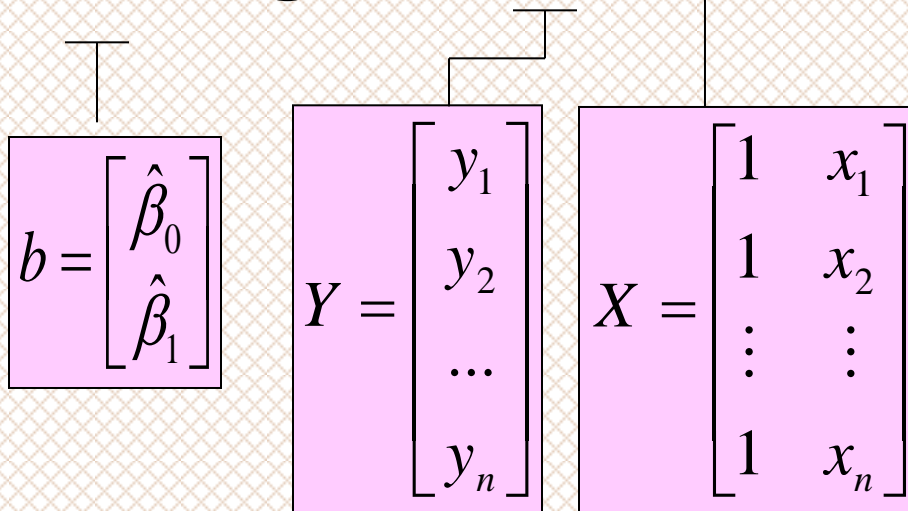
- 1、确定回归系数的点估计值： $\mathbf{b}=\text{regress}(\mathbf{Y}, \mathbf{X})$
- 2、求回归系数的点估计和区间估计、并检验回归模型：
 $[\mathbf{b}, \mathbf{bint}, \mathbf{r}, \mathbf{rint}, \mathbf{stats}] = \text{regress}(\mathbf{Y}, \mathbf{X}, \alpha)$
- 3、画出残差及其置信区间： $\text{rcoplot}(\mathbf{r}, \mathbf{rint})$

一元回归的Matlab实现

$$y = \beta_0 + \beta_1 x_1$$

1、确定回归系数的点估计值：

b=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：

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

回归系数的区间估计

残差

残差区间

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

显著性水平
(缺省时为0.05)

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 $p < \alpha$ 时拒绝 H_0 ，回归模型成立。

3、画出残差及其置信区间：

`rcoplot (r, rint)`

例2 某商场一年内每月的销售收入X(万元)与销售费用Y (万元)统计如表，试求销售费用Y关于销售收入X的线性回归方程。

X	187.1	179.5	157.0	197.0	239.4	217.8	227.1	233.4	242.0	251.9	230.0	271.8
Y	25.4	22.8	20.6	21.8	32.4	24.4	29.3	27.9	27.8	34.2	29.2	30.0

解：建立回归模型 $y=b_0+b_1x$

```
x1=[187.1 179.5 157.0 197.0 239.4 217.8 227.1 233.4 242.0 251.9 230.0 271.8]';
```

```
y=[25.4 22.8 20.6 21.8 32.4 24.4 29.3 27.9 27.8 34.2 29.2 30.0]';
```

```
x=[ones(12,1) x1];
```

```
[b, bint,r,rint,stats]=regress(y,x)
```

```
b = 3.4130 0.1081
```

```
bint = -7.0791 13.9050
```

```
0.0608 0.1554
```

```
stats = 0.7218 25.9430 0.0005
```

回归方程为: $y= 3.4130+0.1081x$

b0的置信区间: -7.0791 13.9050

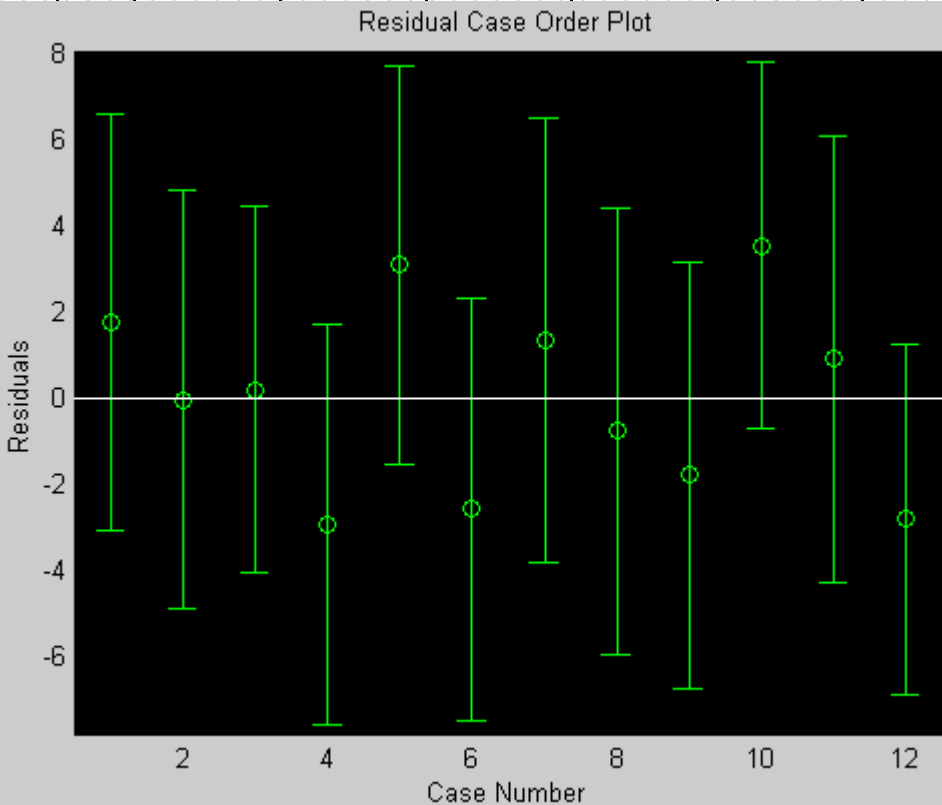
b1的置信区间: 0.0608 0.1554

复相关系数 $R=0.7218$,F统计量值为25.9430, 显著性概率 $P= 0.0005$

作回归残差图: `rcoplot(r,rint)`

例2 某商场一年内每月的销售收入X(万元)与销售费用Y (万元)统计如表，试求销售费用Y关于销售收入X的线性回归方程。

X	187.1	179.5	157.0	197.0	239.4	217.8	227.1	233.4	242.0	251.9	230.0	271.8
Y	29.3	27.9	27.8	34.2	29.2	30.0						



从残差图可以看出，所有数据的残差都包含零，且显著性概率 $P < 0.01$ ，回归效果显著。如果某个数据的残差不包含零，则常把它视为异常值，在回归中应把它剔除，再进行回归。

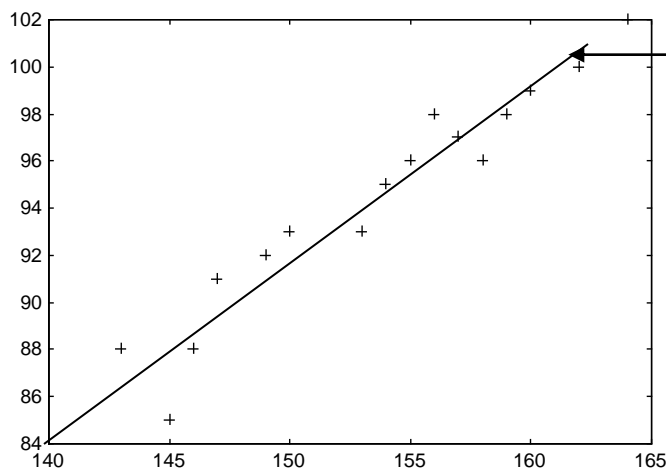
复相关系数 $R=0.7218$, F统计量值为25.9430，显著性概率 $P= 0.0005$

作回归残差图： `rcoplot(r,rint)`

例3 测16名成年女子的身高与腿长所得数据如下：

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

以身高 x 为横坐标，以腿长 y 为纵坐标将这些数据点 (x_i, y_i) 在平面直角坐标系上标出。



散点图

$$y = \beta_0 + \beta_1 x + \varepsilon$$

作图命令：

```
x=[143 145 146 147 149 150 153 154 155  
156 157 158 159 160 162 164];
```

```
y=[88 85 88 91 92 93 93 95 96 98 97 96 98  
99 100 102];
```

```
plot(x,y,'+')
```




1、输入数据：

解： $x=[143 \ 145 \ 146 \ 147 \ 149 \ 150 \ 153 \ 154 \ 155 \ 156 \ 157 \ 158$
 $159 \ 160 \ 162 \ 164]'$;

$X=[\text{ones}(16,1) \ x]$;

$Y=[88 \ 85 \ 88 \ 91 \ 92 \ 93 \ 93 \ 95 \ 96 \ 98 \ 97 \ 96 \ 98 \ 99 \ 100 \ 102]'$;

2、回归分析及检验：

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

$b, \text{bint}, \text{stats}$

得结果： $b = -16.0730$ $\text{bint} = \begin{matrix} -33.7071 & 1.5612 \\ 0.7194 & 0.8340 \end{matrix}$
 $\text{stats} = \begin{matrix} 0.9282 & 180.9531 & 0.0000 \end{matrix}$

即 $\hat{\beta}_0 = -16.073$, $\hat{\beta}_1 = 0.7194$; $\hat{\beta}_0$ 的置信区间为 $[-33.7017, 1.5612]$,

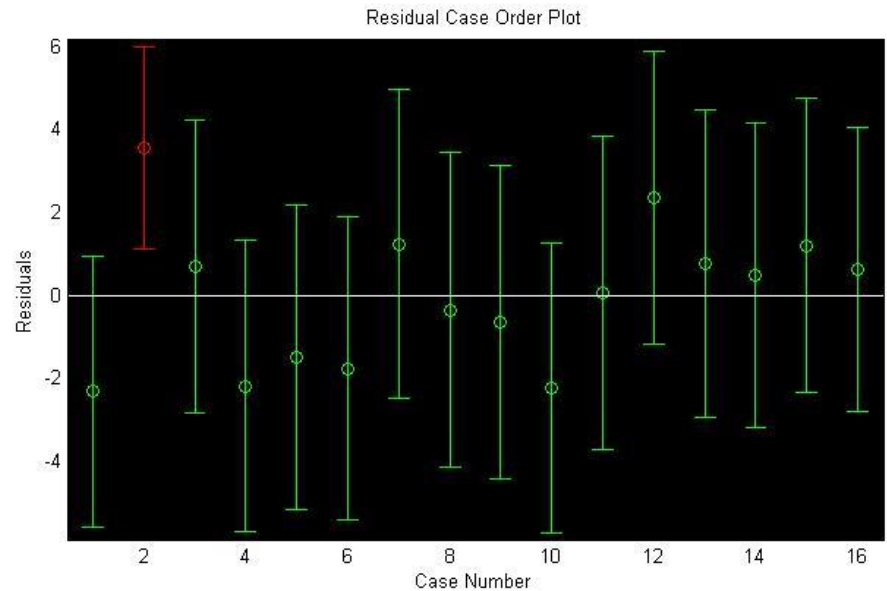
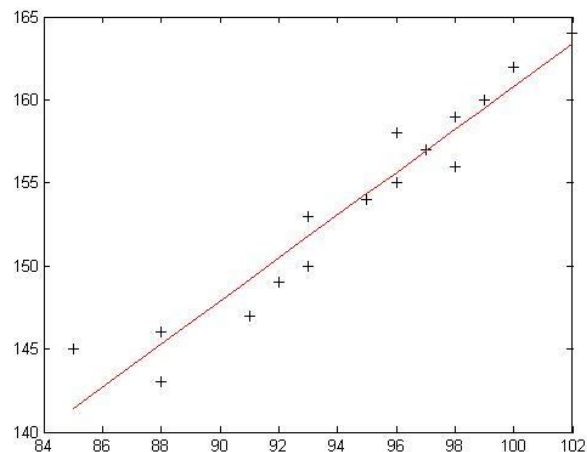
$\hat{\beta}_1$ 的置信区间为 $[0.6047, 0.834]$; $r^2=0.9282$, $F=180.9531$, $p=0.0000$

$p < 0.05$, 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。

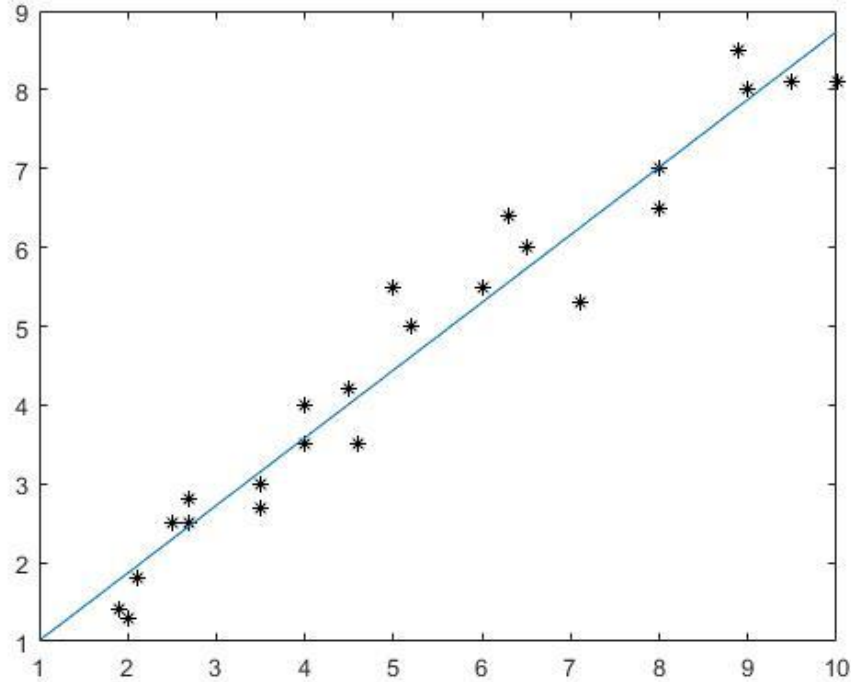
4、预测及作图： $z = b(1) + b(2) * x$ `plot(x,Y,'k+',x,z,'r')`



例3：考察某种纤维的**强度**与其**拉伸倍数**的关系. 下表是实际测定的**24**个纤维样品的**强度**与相应的**拉伸倍数**的数据记录：

编号	拉伸倍数 x_i	强 度 y_i	编号	拉伸倍数 x_i	强 度 y_i
1	1.9	1.4	13	5	5.5
2	2	1.3	14	5.2	5
3	2.1	1.8	15	6	5.5
4	2.5	2.5	16	6.3	6.4
5	2.7	2.8	17	6.5	6
6	2.7	2.5	18	7.1	5.3
7	3.5	3	19	8	6.5
8	3.5	2.7	20	8	7
9	4	4	21	8.9	8.5
10	4	3.5	22	9	8
11	4.5	4.2	23	9.5	8.1
12	4.6	3.5	24	10	8.1

```
x=[1.9,2,2.1,2.5,2.7,2.7,3.5,3.5,4,4,4.5,4.6,5,5.2,6,6.3,6.5,7.1,8,8,8.9,9,9.5,10]';  
y=[1.4,1.3,1.8,2.5,2.8,2.5,3,2.7,4,3.5,4.2,3.5,5.5,5,5.5,6.4,6,5.3,6.5,7,8.5,8,8.1,8.1]';  
plot(x,y,'k*')  
xx=[ones(24,1) x];  
[b, bint,stats]=regress(y,xx);  
f=@(x)b(1)+b(2)*x;  
hold on  
fplot(@(x)f(x),[1,10])
```



实验题1

1: 1949年—1994年我国人口数据资料如下:

年 份 x_i	1949	1954	1959	1964	1969	1974	1979	1984	1989	1994
人口数 y_i	5.4	6.0	6.7	7.0	8.1	9.1	9.8	10.3	11.3	11.8

请分析我国人口增长的规律。

2. 从常识上理解, 一个家庭的消费支出主要受这个家庭收入的影响。一般而言, 家庭收入高的其家庭消费支出也高; 家庭收入低的其家庭消费支出也低。现有部分调查数据如下表, 请分析这些数据的关系。

表 家庭收入与消费支出

家庭编号	1	2	3	4	5	6	7	8	9	10
家庭收入	800	1200	2000	3000	4000	5000	7000	9000	10000	12000
家庭消费支出	770	1100	1300	2200	2100	2700	3800	3900	5500	6600



二、多元线性回归

如果根据经验和有关知识认为与因变量有关联的自变量不止一个，那么就应该考虑用最小二乘准则建立多元线性回归模型。

数学模型及定义

一般称

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{cases}$$

为高斯—马尔柯夫线性模型(**k 元线性回归模型**), 并简记为 $(Y, X\beta, \sigma^2 I_n)$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ 称为**回归平面方程**.

线性模型 $(Y, X\beta, \sigma^2 I_n)$ 考虑的主要问题是:

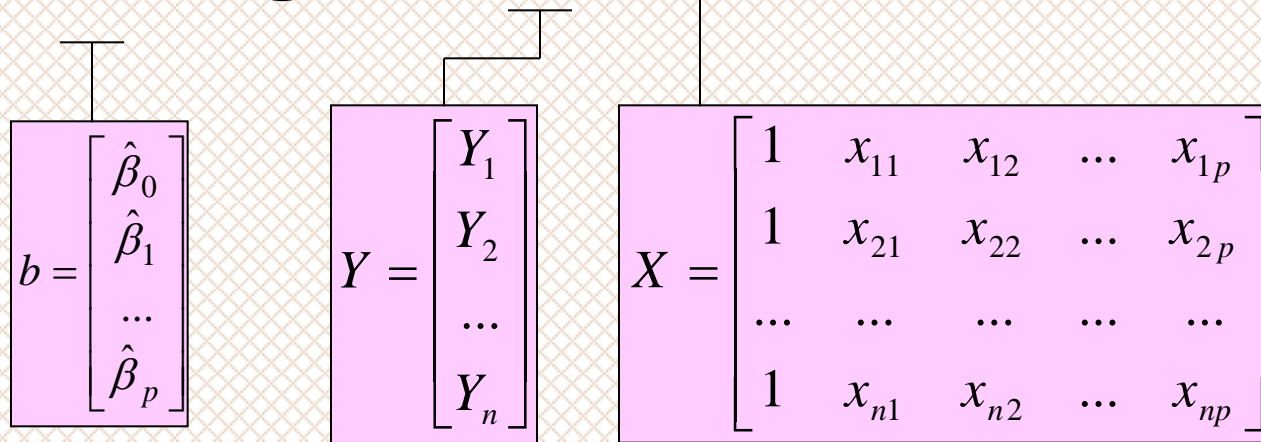
- (1) 用试验值(样本值)对未知参数 β 和 σ^2 作点估计和假设检验, 从而建立 y 与 x_1, x_2, \dots, x_k 之间的数量关系;
- (2) 在 $x_1 = x_{01}, x_2 = x_{02}, \dots, x_k = x_{0k}$, 处对 y 的值作预测与控制, 即对 y 作区间估计.

多元回归的Matlab实现

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

1、确定回归系数的点估计值：

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


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

2、求回归系数的点估计和区间估计、并检验回归模型：

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

回归系数的区间估计

残差

残差区间

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

显著性水平
(缺省时为0.05)

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 $p < \alpha$ 时拒绝 H_0 ，回归模型成立。

3、画出残差及其置信区间：

`rcoplot (r, rint)`



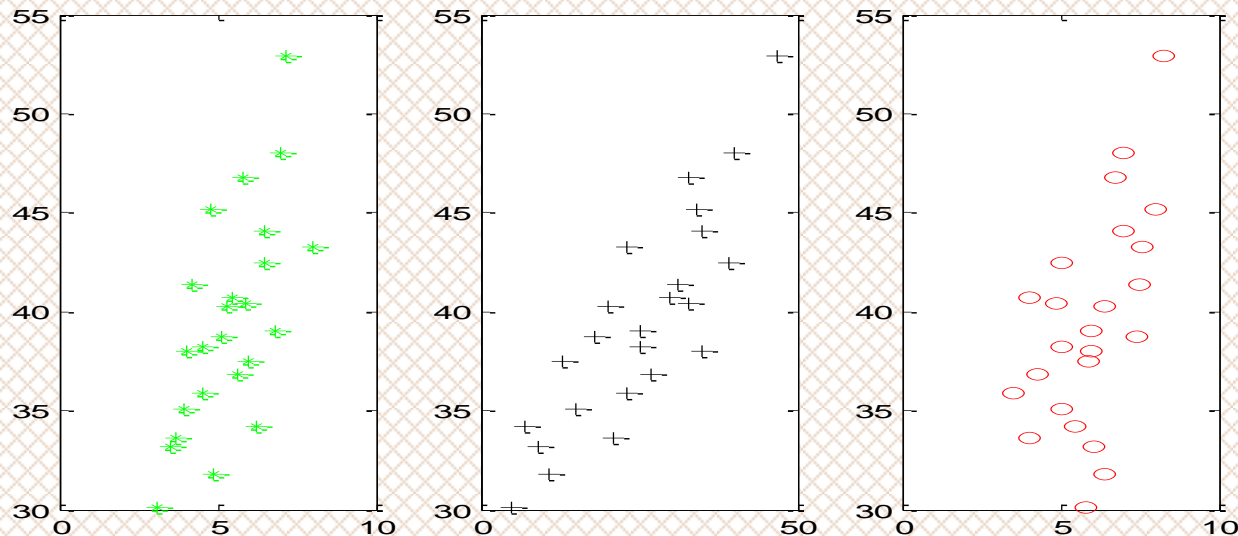
例2：某科学基金会希望估计从事某研究的学者的年薪 Y 与他们的研究成果(论文、著作等)的质量指标 x_1 、从事研究工作的时间 x_2 、能成功获得资助的指标 x_3 之间的关系，为此按一定的实验设计方法调查了24位研究学者，得到如下数据（ i 为学者序号）：

i	1	2	3	4	5	6	7	8	9	10	11	12
x_{i1}	3.5	5.3	5.1	5.8	4.2	6.0	6.8	5.5	3.1	7.2	4.5	4.9
x_{i2}	9	20	18	33	31	13	25	30	5	47	25	11
x_{i3}	6.1	6.4	7.4	6.7	7.5	5.9	6.0	4.0	5.8	8.3	5.0	6.4
y_i	33.2	40.3	38.7	46.8	41.4	37.5	39.0	40.7	30.1	52.9	38.2	31.8
i	13	14	15	16	17	18	19	20	21	22	23	24
x_{i1}	8.0	6.5	6.6	3.7	6.2	7.0	4.0	4.5	5.9	5.6	4.8	3.9
x_{i2}	23	35	39	21	7	40	35	23	33	27	34	15
x_{i3}	7.6	7.0	5.0	4.4	5.5	7.0	6.0	3.5	4.9	4.3	8.0	5.8
y_i	43.3	44.1	42.5	33.6	34.2	48.0	38.0	35.9	40.4	36.8	45.2	35.1

试建立 Y 与 X 之间关系的数学模型.

1. 作出因变量Y与各自变量的样本散点图

```
subplot(1,3,1), plot(x1,Y,'g*'),  
subplot(1,3,2), plot(x2,Y,'k+'),  
subplot(1,3,3), plot(x3,Y,'ro'),
```



从图可以看出这些点大致分布在一条直线旁边，因此，有比较好的线性关系，可以采用线性回归。

2. 利用MATLAB统计工具箱得到初步的回归方程

设回归方程为: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

建立m-文件输入如下程序数据:

```
x1=[3.5 5.3 5.1 5.8 4.2 6.0 6.8 5.5 3.1 7.2 4.5 4.9 8.0 6.5 6.5 3.7 6.2 7.0 4.0 4.5 5.9 5.6 4.8 3.9];  
x2=[9 20 18 33 31 13 25 30 5 47 25 11 23 35 39 21 7 40 35 23 33 27 34 15];  
x3=[6.1 6.4 7.4 6.7 7.5 5.9 6.0 4.0 5.8 8.3 5.0 6.4 7.6 7.0 5.0 4.0 5.5 7.0 6.0 3.5 4.9 4.3 8.0 5.0];  
Y=[33.2 40.3 38.7 46.8 41.4 37.5 39.0 40.7 30.1 52.9 38.2 31.8 43.3 44.1 42.5 33.6 34.2 48.0 38.0  
35.9 40.4 36.8 45.2 35.1];  
n=24; m=3;  
X=[ones(n,1),x1',x2',x3'];  
[b,bint,r,rint,s]=regress(Y',X,0.05);  
b,bint,s,
```



运行后即得到结果如表所示

回归系数	回归系数的估计值	回归系数的置信区间	
β_0	18.0157	[13.9052 22.1262]	
β_1	1.0817	[0.3900 1.7733]	
β_2	0.3212	[0.2440 0.3984]	
β_3	1.2835	[0.6691 1.8979]	
$R^2 = 0.9106$ $F = 67.9195$ $p < 0.0001$ $s^2 = 3.0719$			

得到回归方程为：

$$\hat{y} = 18.0157 + 1.0817x_1 + 0.3212x_2 + 1.2835x_3$$



实验题2

下表数据是某建筑材料公司去年20个地区的销售量（Y，千方），推销开支、实际帐目数、同类商品竞争数和地区销售潜力分别是影响建筑材料销售量的因素。试建立回归模型

地区 i	推销开支 (x1)	实际帐目数 (x2)	同类商品竞争数 (x3)	地区销售潜力 (x4)	销售量 Y
1	5.5	31	10	8	79.3
2	2.5	55	8	6	200.1
3	8.0	67	12	9	163.2
4	3.0	50	7	16	200.1
5	3.0	38	8	15	146.0
6	2.9	71	12	17	177.7
7	8.0	30	12	8	30.9
8	9.0	56	5	10	291.9
9	4.0	42	8	4	160.0
10	6.5	73	5	16	339.4
11	5.5	60	11	7	159.6
12	5.0	44	12	12	86.3
13	6.0	50	6	6	237.5
14	5.0	39	10	4	107.2
15	3.5	55	10	4	155.0
16	8.0	70	6	14	201.4
17	6.0	40	11	6	100.2
18	4.0	50	11	8	135.8
19	7.5	62	9	13	223.3
20	7.0	59	9	11	195.0

三、多项式回归

(一) 一元多项式回归 $y=a_1x^m+a_2x^{m-1}+\dots+a_mx+a_{m+1}$

1、回归：

(1) 确定多项式系数的命令： $[p, S]=\text{polyfit}(x, y, m)$

(2) 一元多项式回归命令： $\text{polytool}(x, y, m)$

2、预测和预测误差估计：

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

(2) $[Y, \text{DELTA}]=\text{polyconf}(p, x, S, \alpha)$ 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 $Y \pm \text{DELTA}$ ；alpha缺省时为0.5。

(二) 多元二项式回归

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

例 2 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

法一 直接作二次多项式回归：

$t=1/30:1/30:14/30;$

$s=[11.86 \ 15.67 \ 20.60 \ 26.69 \ 33.71 \ 41.93 \ 51.13 \ 61.49 \ 72.90 \dots$
 $85.44 \ 99.08 \ 113.77 \ 129.54 \ 146.48];$

$[p,S]=polyfit(t,s,2)$

得回归模型为：

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

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);
```

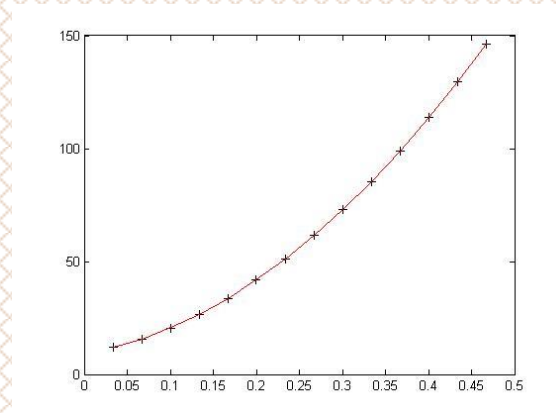
```
b,stats
```

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

命令: `rstool (x, y, 'model', alpha)`

n×m矩阵

n维列向量

显著性水平
(缺省时为0.05)

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

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

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	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

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

法一 直接用多元二项式回归：

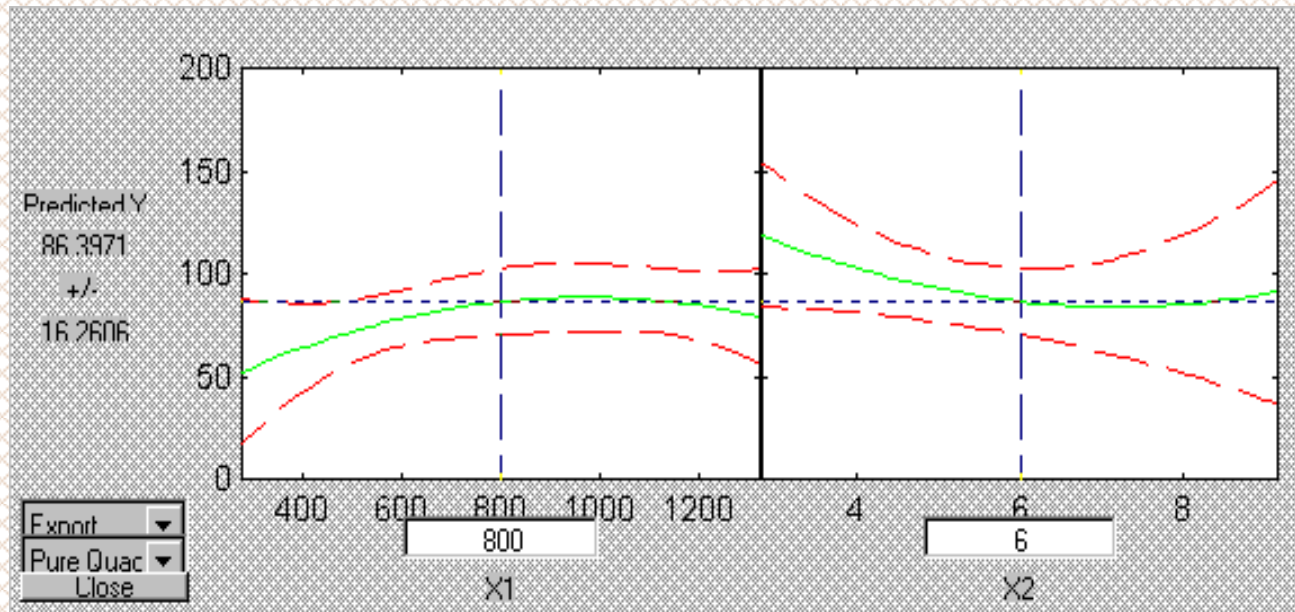
x1=[1000 600 1200 500 300 400 1300 1100 1300 300];

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： $y = 110.5313 + 0.1464 x_1 - 26.5709 x_2 - 0.0001 x_1^2 + 1.8475 x_2^2$

剩余标准差为 4.5362，说明此回归模型的显著性较好。

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

例 3 为了了解人口平均预期寿命与人均国内生产总值和体质得分的关系，我们查阅了国家统计局资料，北京体育大学出版社出版的《2000 国民体质监测报告》，表 5-5 是我国大陆 31 个省市的有关数据。我们希望通过这几组数据考察它们是否具有良好的相关关系，并通过它们的关系从人均国内生产总值（可以看作反映生活水平的一个指标）、体质得分预测其寿命可能的变化范围。

表 5-5 31 个省市人口预期寿命与人均国内生产总值和体质得分数据

序号	预期寿命	体质得分	人均产值	序号	预期寿命	体质得分	人均产值	序号	预期寿命	体质得分	人均产值
1	71.54	66.165	12857	12	65.49	56.775	8744	23	69.87	64.305	17717
2	73.92	71.25	24495	13	68.95	66.01	11494	24	67.41	60.485	15205
3	73.27	70.135	24250	14	73.34	67.97	20461	25	78.14	70.29	70622
4	71.20	65.125	10060	15	65.96	62.9	5382	26	76.10	69.345	47319
5	73.91	69.99	29931	16	72.37	66.1	19070	27	74.91	68.415	40643
6	72.54	65.765	18243	17	70.07	64.51	10935	28	72.91	66.495	11781
7	70.66	67.29	10763	18	72.55	68.385	22007	29	70.17	65.765	10658
8	71.85	67.71	9907	19	71.65	66.205	13594	30	66.03	63.28	11587
9	71.08	66.525	13255	20	71.73	65.77	11474	31	64.37	62.84	9725
10	71.29	67.13	9088	21	73.10	67.065	14335				
11	74.70	69.505	33772	22	67.47	63.605	7898				

模型的建立和求解 作表 5-5 数据 $(x_1, y), (x_2, y)$ 的散点图如图 5.3

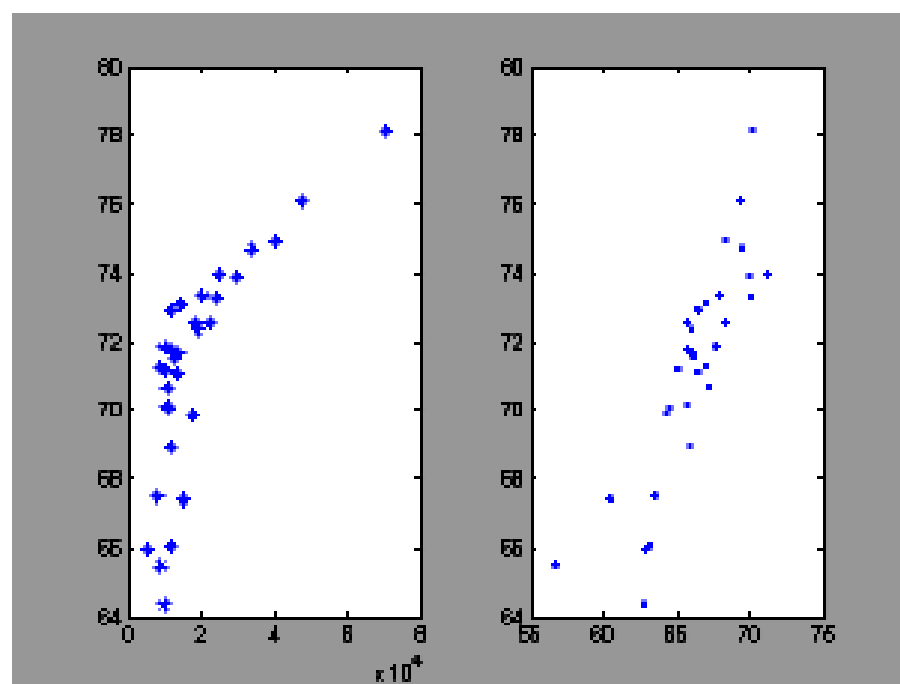


图 5.3 预期寿命与人均国内生产总值和体质得分的散点图

从图 5.3 可以看出人口预期寿命 y 与体质得分 x_2 有较好的线性关系, y 与人均国内生产总值 x_1 的关系难以确定, 我们建立二次函数的回归模型。

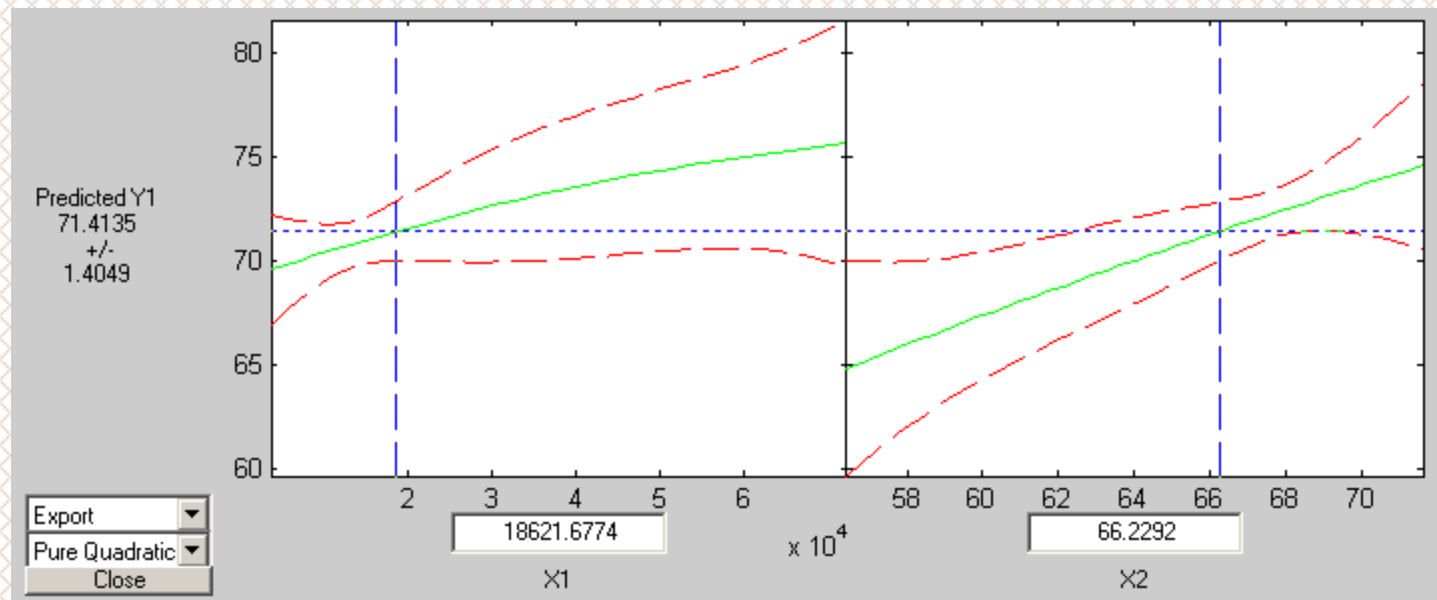
$y = [71.54 \ 73.92 \ 73.27 \ 71.20 \ 73.91 \ 72.54 \ 70.66 \ 71.85 \ 71.08 \ 71.29, 74.70 \ 65.49 \ 68.95$
 $73.34 \ 65.96 \ 72.37 \ 70.07 \ 72.55 \ 71.65 \ 71.73, 73.10 \ 67.47 \ 69.87 \ 67.41 \ 78.14 \ 76.10 \ 74.91$
 $72.91 \ 70.17 \ 66.03 \ 64.37];$

$x_1 = [12857 \ 24495 \ 24250 \ 10060 \ 29931 \ 18243 \ 10763 \ 9907 \ 13255 \ 9088 \ 33772 \ 8744 \ 11494$
 $20461 \ 5382 \ 19070 \ 10935 \ 22007 \ 13594 \ 11474 \ 14335 \ 7898 \ 17717 \ 15205 \ 70622 \ 47319$
 $40643 \ 11781 \ 10658 \ 11587 \ 9725];$

$x_2 = [66.165 \ 71.25 \ 70.135 \ 65.125 \ 69.99 \ 65.765 \ 67.29 \ 67.71 \ 66.525 \ 67.13, 69.505 \ 56.775$
 $66.01 \ 67.97 \ 62.9 \ 66.1 \ 64.51 \ 68.385 \ 66.205 \ 65.77, 67.065 \ 63.605 \ 64.305 \ 60.485 \ 70.29$
 $69.345 \ 68.415 \ 66.495 \ 65.765 \ 63.28 \ 62.84];$

$x = [x_1', x_2'];$

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





实验题3

在生产中测得 13 组数据：

序号	X1	X2	X3	X4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

求出关系式 $Y = f(X)$ 。



总 结

回归分析方法是处理变量间相关关系的有力工具.它不仅为建立变量间关系的数学表达式(经验公式)提供了一般的方法,而且还能判明所建立的经验公式的有效性,从而达到利用经验公式预测、控制等目的.因此,回归分析方法的应用越来越广泛,其方法本身也在不断丰富和发展.

Q & A

- 有什么问题吗？

