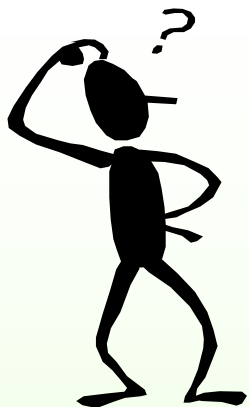


# 数学建模与数学实验

## 回归分析





## 实验目的

- 1、直观了解回归分析基本内容。
- 2、掌握用数学软件求解回归分析问题。

## 实验内容

- 1、回归分析的基本理论。
- 2、用数学软件求解回归分析问题。
- 3、实验作业。

# 回归分析



## 一元线性回归

## 多元线性回归

数学模型及定义

\*模型参数估计

\*检验、预测与控制

可线性化的一元非线性回归（曲线回归）

数学模型及定义

\*模型参数估计

检验与预测

\*多元线性回归中的

逐步回归分析

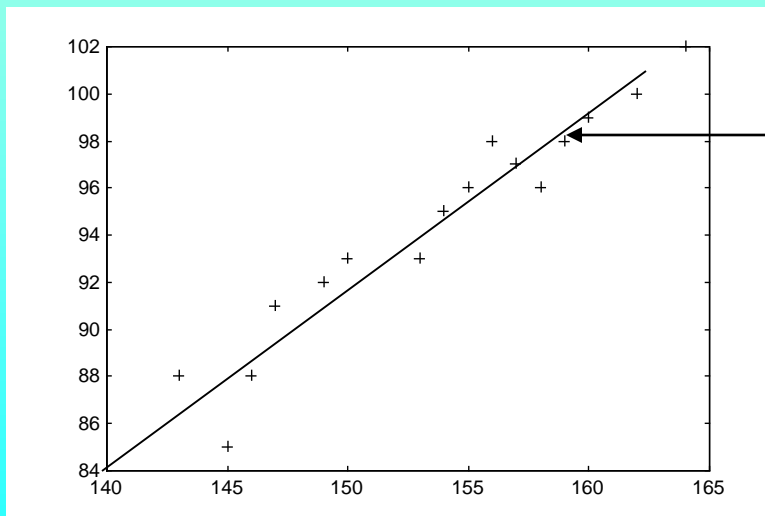
# 一、数学模型

例1 测16名成年女子的身高与腿长所得数据如下：

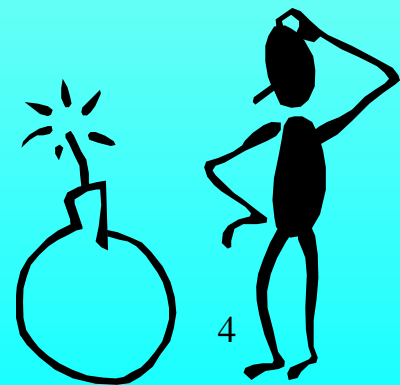
身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

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。

解答



散点图



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

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

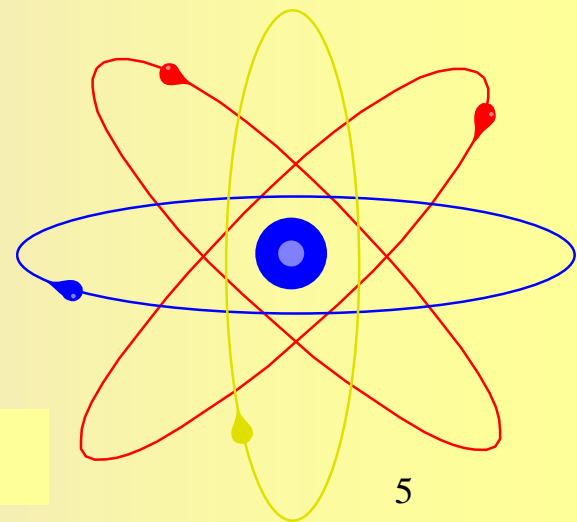
固定的未知参数  $\beta_0$ 、 $\beta_1$  称为回归系数，自变量  $x$  也称为回归变量。

$Y = \beta_0 + \beta_1 x$ ，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  $\beta_0$ 、 $\beta_1$  和  $\sigma$  作点估计；
- 2、对回归系数  $\beta_0$ 、 $\beta_1$  作假设检验；
- 3、在  $x=x_0$  处对  $y$  作预测，对  $y$  作区间估计。

返回



## 二、模型参数估计

### 1、回归系数的最小二乘估计

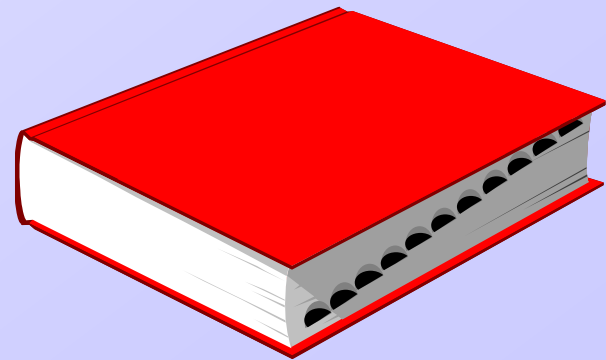
有  $n$  组独立观测值,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

$$\text{设 } \begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \dots, n \\ E\varepsilon_i = 0, D\varepsilon_i = \sigma^2 \text{ 且 } \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ 相互独立} \end{cases}$$

$$\text{记 } Q = Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

**最小二乘法**就是选择  $\beta_0$  和  $\beta_1$  的估计  $\hat{\beta}_0, \hat{\beta}_1$  使得

$$Q(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$$





解得 
$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \end{cases} \quad \text{或} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

其中  $\underline{x} = \frac{1}{J} \sum_{j=1}^J x^j, \underline{\lambda} = \frac{1}{J} \sum_{j=1}^J \lambda^j, \underline{x}_{\mathcal{S}} = \frac{1}{J} \sum_{j=1}^J x^j_{\mathcal{S}}, \underline{x\lambda} = \frac{1}{J} \sum_{j=1}^J x^j \lambda^j.$

(经验) 回归方程为: 
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

## 2、 $\sigma^2$ 的无偏估计

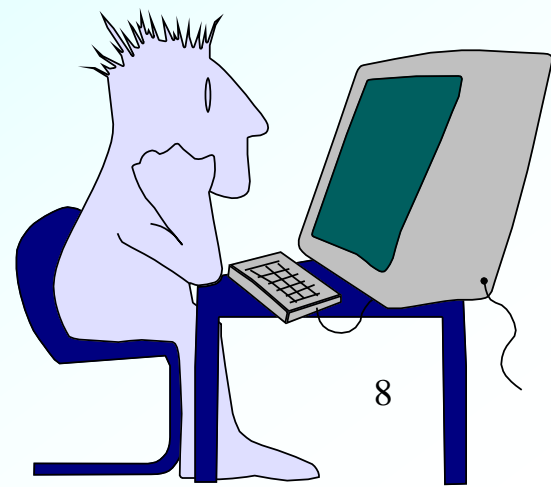
$$\text{记 } Q_e = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

称  $Q_e$  为残差平方和或剩余平方和.

$\sigma^2$  的无偏估计为  $\hat{\sigma}_e^2 = Q_e / (n - 2)$

称  $\hat{\sigma}_e^2$  为剩余方差（残差的方差）， $\hat{\sigma}_e^2$  分别与  $\hat{\beta}_0$ 、 $\hat{\beta}_1$  独立。

$\hat{\sigma}_e$  称为剩余标准差.





### 三、检验、预测与控制

#### 1、回归方程的显著性检验

对回归方程  $Y = \beta_0 + \beta_1 x$  的显著性检验，归结为对假设

$$H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$$

进行检验.

假设  $H_0 : \beta_1 = 0$  被拒绝，则回归显著，认为  $y$  与  $x$  存在线性关系，所求的线性回归方程有意义；否则回归不显著， $y$  与  $x$  的关系不能用一元线性回归模型来描述，所得的回归方程也无意义.

## (I) F检验法

当  $H_0$  成立时, 
$$F = \frac{U}{Q_e / (n-2)} \sim F(1, n-2)$$

其中 
$$U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (\text{回归平方和})$$

故  $F > F_{1-\alpha}(1, n-2)$ , 拒绝  $H_0$ , 否则就接受  $H_0$ .

## (II) t检验法

当  $H_0$  成立时, 
$$T = \frac{\sqrt{L_{xx}} \hat{\beta}_1}{\hat{\sigma}_e} \sim t(n-2)$$

故  $|T| > t_{1-\frac{\alpha}{2}}(n-2)$ , 拒绝  $H_0$ , 否则就接受  $H_0$ .

其中 
$$L_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

### (III) r检验法

记

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

当 $|r| > r_{1-\alpha}$ 时，拒绝  $H_0$ ；否则就接受  $H_0$ 。

其中  $r_{1-\alpha} = \sqrt{\frac{1}{1 + (n-2)/F_{1-\alpha}(1, n-2)}}$

## 2、回归系数的置信区间

$\beta_0$  和  $\beta_1$  置信水平为  $1-\alpha$  的置信区间分别为

$$\left[ \hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}} \right]$$

和  $\left[ \hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e / \sqrt{L_{xx}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e / \sqrt{L_{xx}} \right]$

$\sigma^2$  的置信水平为  $1-\alpha$  的置信区间为

$$\left[ \frac{Q_e}{\chi_{1-\frac{\alpha}{2}}^2(n-2)}, \frac{Q_e}{\chi_{\frac{\alpha}{2}}^2(n-2)} \right]$$



### 3、预测与控制

#### (1) 预测

用  $y_0$  的回归值  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$  作为  $y_0$  的预测值.

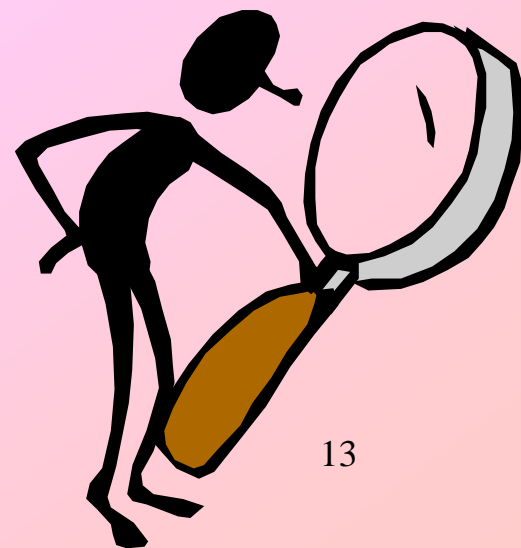
$y_0$  的置信水平为  $1-\alpha$  的预测区间为

$$[\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0)]$$

$$\text{其中 } \delta(x_0) = \hat{\sigma}_e t_{1-\frac{\alpha}{2}}(n-2) \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{L_{xx}}}$$

特别，当  $n$  很大且  $x_0$  在  $\bar{x}$  附近取值时，  
 $y$  的置信水平为  $1-\alpha$  的预测区间近似为

$$\left[ \hat{y} - \hat{\sigma}_e u_{1-\frac{\alpha}{2}}, \hat{y} + \hat{\sigma}_e u_{1-\frac{\alpha}{2}} \right]$$



## (2) 控制

要求:  $y = \beta_0 + \beta_1 x + \varepsilon$  的值以  $1 - \alpha$  的概率落在指定区间  $(y', y'')$

只要控制  $x$  满足以下两个不等式

$$\hat{y} - \delta(x) \geq y', \hat{y} + \delta(x) \leq y''$$

要求  $y'' - y' \geq 2\delta(x)$ . 若  $\hat{y} - \delta(x) = y'$ ,  $\hat{y} + \delta(x) = y''$  分别有解  $x'$  和  $x''$ , 即  $\hat{y} - \delta(x') = y'$ ,  $\hat{y} + \delta(x'') = y''$ .

则  $(x', x'')$  就是所求的  $x$  的控制区间.

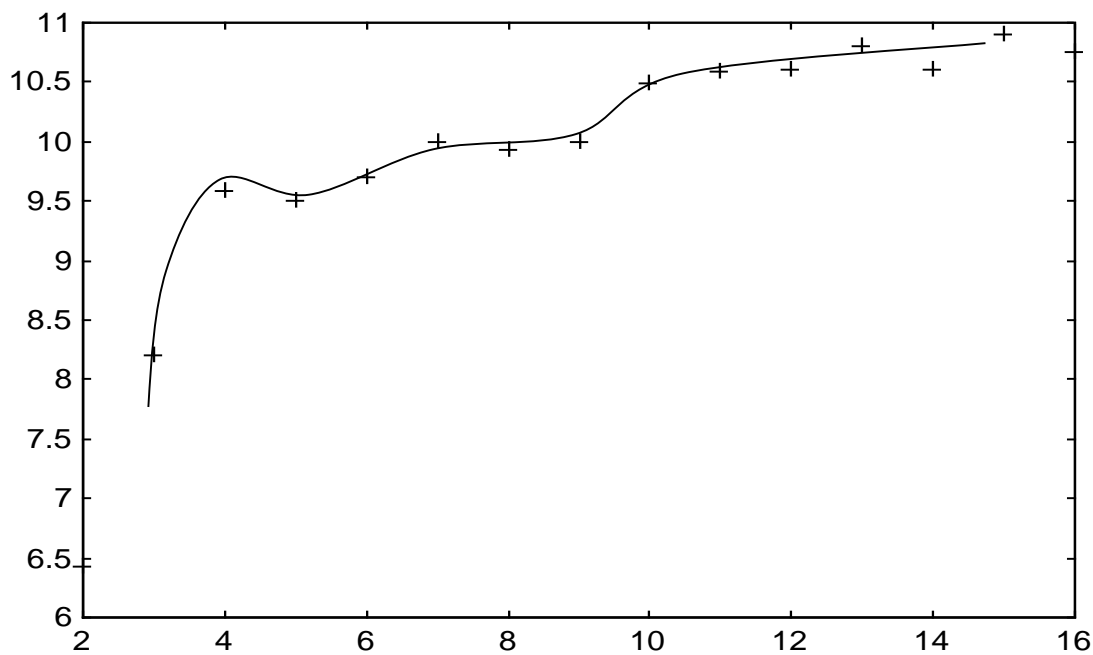
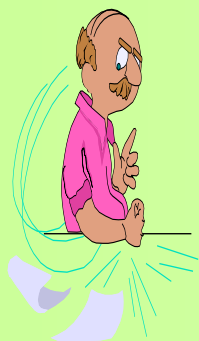
返回

## 四、可线性化的一元非线性回归 (曲线回归)

**例2** 出钢时所用的盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断增大.我们希望知道使用次数与增大的容积之间的关系.对一钢包作试验，测得的数据列于下表：

使用次数	增大容积	使用次数	增大容积
2	6.42	10	10.49
3	8.20	11	10.59
4	9.58	12	10.60
5	9.50	13	10.80
6	9.70	14	10.60
7	10.00	15	10.90
8	9.93	16	10.76
9	9.99		

解答



散点图

此即**非线性回归**或**曲线回归**问题（需要配曲线）

**配曲线的一般方法是：**

先对两个变量  $x$  和  $y$  作  $n$  次试验观察得  $(x_i, y_i), i = 1, 2, \dots, n$  画出散点图，根据散点图确定须配曲线的类型.然后由  $n$  对试验数据确定每一类曲线的未知参数  $a$  和  $b$ .采用的方法是通过变量代换把非线性回归化成线性回归，即采用非线性回归线性化的方法.



通常选择的六类曲线如下：

(1) 双曲线  $\frac{1}{y} = a + \frac{b}{x}$

(2) 幂函数曲线  $y = ax^b$ ，其中  $x > 0, a > 0$

(3) 指数曲线  $y = ae^{bx}$  其中参数  $a > 0$ .

(4) 倒指数曲线  $y = ae^{b/x}$  其中  $a > 0$ ,

(5) 对数曲线  $y = a + b \log x, x > 0$

(6) S 型曲线  $y = \frac{1}{a + be^{-x}}$

返回

解例 2. 由散点图我们选配倒指数曲线  $y = a^{b/x}$

根据线性化方法，算得  $\hat{b} = -1.1107, \hat{A} = 2.4587$

由此  $\hat{a} = e^{\hat{A}} = 11.6789$

最后得  $y = 11.6789e^{\frac{1.1107}{x}}$

# 一、数学模型及定义

一般称

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

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

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

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

返回

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  $\beta$  和  $\sigma^2$  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.

## 二、模型参数估计



### 1、对 $\beta_i$ 和 $\sigma^2$ 作估计

用最小二乘法求  $\beta_0, \dots, \beta_k$  的估计量：作离差平方和

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2$$

选择  $\beta_0, \dots, \beta_k$  使  $Q$  达到最小。

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

得到的  $\hat{\beta}_i$  代入回归平面方程得：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$$

称为经验回归平面方程。 $\hat{\beta}_i$  称为经验回归系数。

注意： $\hat{\beta}$  服从  $p+1$  维正态分布，且为  $\beta$  的无偏估计，协方差阵为  $\sigma^2 C$ 。  
 $C=L^{-1}=(c_{ij})$ ,  $L=X'X$

## 2、多项式回归



设变量  $x$ 、 $Y$  的回归模型为

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p + \varepsilon$$

其中  $p$  是已知的,  $\beta_i (i = 1, 2, \dots, p)$  是未知参数,  $\varepsilon$  服从正态分布  $N(0, \sigma^2)$ .

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$$

称为**回归多项式**.上面的回归模型称为**多项式回归**.

令  $x_i = x^i$ ,  $i=1, 2, \dots, k$  多项式回归模型变为多元线性回归模型.

返回

### 三、多元线性回归中的检验与预测

#### 1、线性模型和回归系数的检验

假设  $H_0 : \beta_0 = \beta_1 = \dots = \beta_k = 0$

##### (I) F检验法

当  $H_0$  成立时,  $F = \frac{U / k}{Q_e / (n - k - 1)} \sim F(k, n - k - 1)$

如果  $F > F_{1-\alpha}(k, n-k-1)$ , 则拒绝  $H_0$ , 认为  $y$  与  $x_1, \dots, x_k$  之间显著地有线性关系; 否则就接受  $H_0$ , 认为  $y$  与  $x_1, \dots, x_k$  之间线性关系不显著. 其中  $U = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  (回归平方和)  $Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  (残差平方和)

##### (II) r检验法

定义  $R = \sqrt{\frac{U}{L_{yy}}} = \sqrt{\frac{U}{U + Q_e}}$  为  $y$  与  $x_1, x_2, \dots, x_k$  的多元相关系数或复相关系数。

由于  $F = \frac{n-k-1}{k} \frac{R^2}{1-R^2}$ , 故用  $F$  和用  $R$  检验是等效的。

## 2、预测

### (1) 点预测

求出回归方程  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$ ，对于给定自变量的值  $x_1^*, \dots, x_k^*$ ，用  $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_k x_k^*$  来预测  $y^* = \beta_0 + \beta_1 x_1^* + \dots + \beta_k x_k^* + \varepsilon$ 。称  $\hat{y}^*$  为  $y^*$  的点预测。

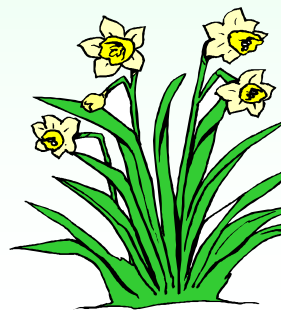
### (2) 区间预测

$y$  的  $1-\alpha$  的预测区间（置信）区间为  $(\hat{y}_1, \hat{y}_2)$ ，其中

$$\begin{cases} \hat{y}_1 = \hat{y} - \hat{\sigma}_e \sqrt{1 + \sum_{i=0}^k \sum_{j=0}^k c_{ij} x_i x_j} t_{1-\alpha/2}(n-k-1) \\ \hat{y}_2 = \hat{y} + \hat{\sigma}_e \sqrt{1 + \sum_{i=0}^k \sum_{j=0}^k c_{ij} x_i x_j} t_{1-\alpha/2}(n-k-1) \end{cases}$$

$$C=L^{-1}=(c_{ij}), L=X'X$$

$$\hat{\sigma}_e = \sqrt{\frac{Q_e}{n-k-1}}$$



返回

## 四、逐步回归分析

“最优”的回归方程就是包含所有对Y有影响的变量,而不包含对Y影响不显著的变量回归方程。

选择“最优”的回归方程有以下几种方法:

- (1) 从所有可能的因子(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。

以第四种方法,即逐步回归分析法在筛选变量方面较为理想.

## 逐步回归分析法的思想：

- 从一个自变量开始，视自变量 $Y$ 作用的显著程度，从大到地依次逐个引入回归方程。
- 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。
- 对于每一步都要进行 $Y$ 值检验，以确保每次引入新的显著性变量前回归方程中只包含对 $Y$ 作用显著的变量。
- 这个过程反复进行，直至既无不显著的变量从回归方程中剔除，又无显著变量可引入回归方程时为止。

返回



# 统计工具箱中的回归分析命令

1、多元线性回归

2、多项式回归

3、非线性回归

4、逐步回归

返回

## 多元线性回归

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

1、确定回归系数的点估计值：

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

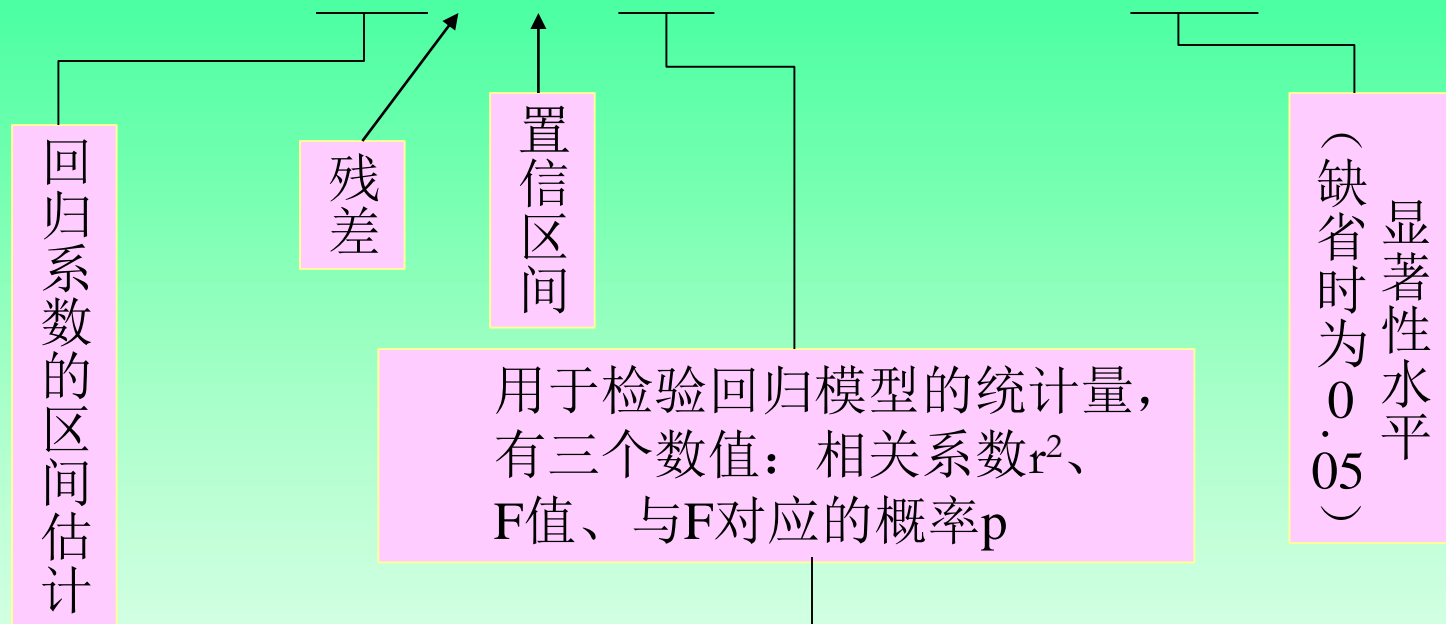
Diagram illustrating the components of the  $\text{regress}(\mathbf{Y}, \mathbf{X})$  function:

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

对一元线性回归，取  $p=1$  即可

## 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)`



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)`



## 题目

## 2、回归分析及检验:

## To MATLAB(liti11)

bint =

-33.7071      1.5612

0.6047      0.8340

0.9282	180.9531	0.0000
--------	----------	--------

$$r^2=0.9282, F=180.9531, p=0.0000$$

$p < 0.05$ , 可知回归模型  $y = -16.073 + 0.7194x$  成立.

### 3、残差分析，作残差图：

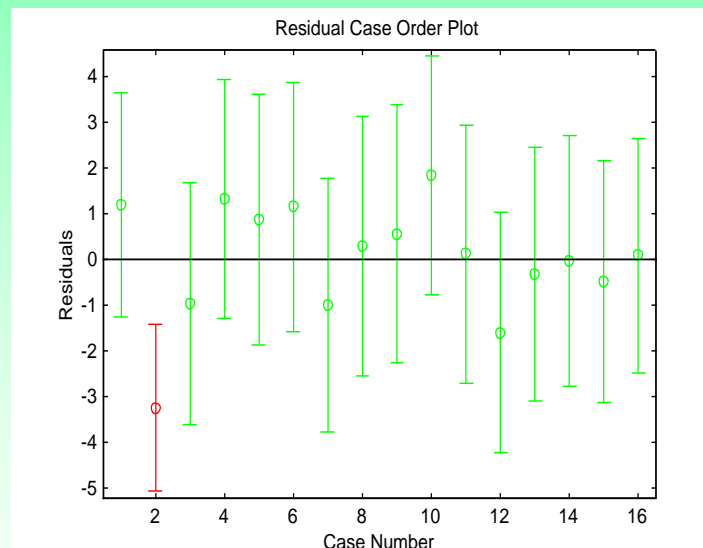
`rcoplot(r,rint)`

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：

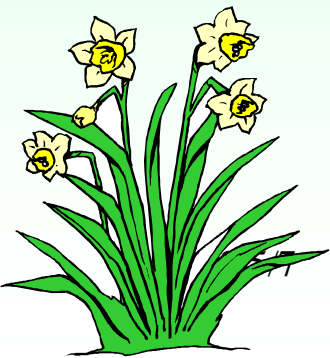
$z = b(1) + b(2) * x$

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



To MATLAB(liti12)

返回



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

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

$p=(a_1, a_2, \dots, a_{m+1})$  是多项式  $y=a_1x^m+a_2x^{m-1}+\dots+a_mx+a_{m+1}$  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:  $\text{polytool}(x, y, m)$

### 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)  $[Y, \text{DELTA}]=\text{polyconf}(p, x, S, \text{alpha})$  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 $1-\text{alpha}$ 的置信区间 $Y \pm \text{DELTA}$ ;  $\text{alpha}$ 缺省时为0.5.

**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：

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

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

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

得回归模型为：

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

To MATLAB (liti21)

## 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')
```

To MATLAB(liti22)

To MATLAB(liti23)





## (二) 多元二项式回归



命令: `rstool (x, y, 'model', alpha)`

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

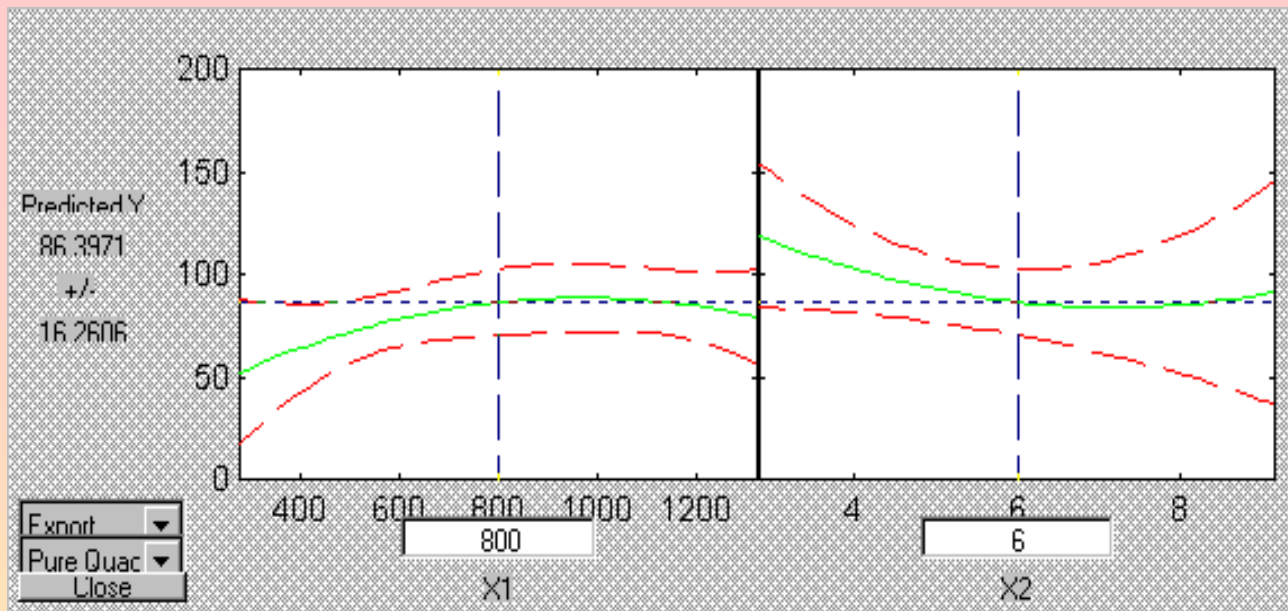
### 法一

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 =

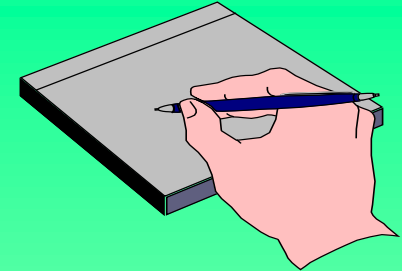
4.5362

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.

To MATLAB(liti31)

## 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

[To MATLAB\(liti32\)](#)

返回

# 非线性回归

是事先用m-文件定义的非线性函数

## 1、回归：

(1) 确定回归系数的命令：

**[beta, r, J]=nlinfit (x, y, ' model', beta0)**

估计出的  
回归系数

残差

Jacobian矩阵

输入数据x、y分别为  
 $n \times m$ 矩阵和n维列向  
量，对一元非线性回  
归，x为n维列向量。

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)**

求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 $Y \pm DELTA$ .

例 4 对第一节例2，求解如下：

题目

1、对将要拟合的非线性模型  $y=ae^{b/x}$ ，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
```

得结果： beta =

11.6036

-1.0641

To MATLAB(liti41)

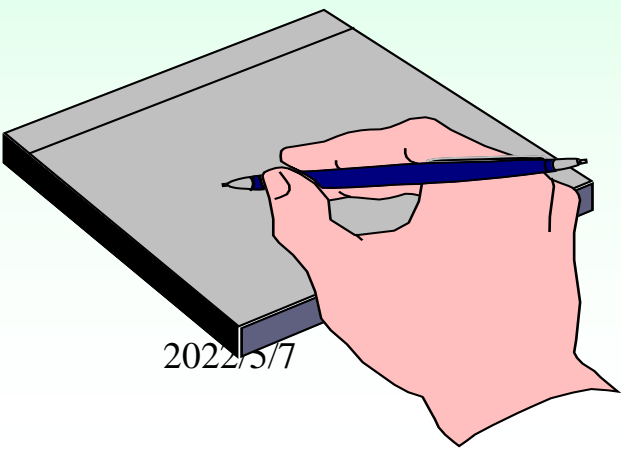
即得回归模型为：

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

#### 4、预测及作图：

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

To MATLAB(liti42)





**例5** 财政收入预测问题：财政收入与国民收入、工业总产值、农业总产值、总人口、就业人口、固定资产投资等因素有关。  
下表列出了1952-1981年的原始数据，试构造预测模型。

**解** 设国民收入、工业总产值、农业总产值、总人口、就业人口、固定资产投资分别为 $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$ 、 $x_5$ 、 $x_6$ ，财政收入为 $y$ ，设变量之间的关系为：

$$y = ax_1 + bx_2 + cx_3 + dx_4 + ex_5 + fx_6$$

使用非线性回归方法求解。

# 1. 对回归模型建立M文件model.m如下:

```
function yy=model(beta0,X)
a=beta0(1);
b=beta0(2);
c=beta0(3);
d=beta0(4);
e=beta0(5);
f=beta0(6);
x1=X(:,1);
x2=X(:,2);
x3=X(:,3);
x4=X(:,4);
x5=X(:,5);
x6=X(:,6);
yy=a*x1+b*x2+c*x3+d*x4+e*x5+f*x6;
```

## 2. 主程序liti6.m如下:

```
X=[598.00 349.00 461.00 57482.00 20729.00 44.00
.....
2927.00 6862.00 1273.00 100072.0 43280.00 496.00];

y=[184.00 216.00 248.00 254.00 268.00 286.00 357.00 444.00 506.00 ...
271.00 230.00 266.00 323.00 393.00 466.00 352.00 303.00 447.00 ...
564.00 638.00 658.00 691.00 655.00 692.00 657.00 723.00 922.00 ...
890.00 826.00 810.0]';

beta0=[0.50 -0.03 -0.60 0.01 -0.02 0.35];

betafit = nlinfit(X,y,'model',beta0)
```

To MATLAB(liti6)

结果为:

```
betafit =  
0.5243  
-0.0294  
-0.6304  
0.0112  
-0.0230  
0.3658
```

即  $y = 0.5243x_1 - 0.0294x_2 - 0.6304x_3 + 0.0112x_4 - 0.0230x_5 + 0.3658x_6$

返回

## 逐步回归

显著性水平（缺省时为0.5）

逐步回归的命令是：

`stepwise (x, y, inmodel, alpha)`

自变量数据,  
 $n \times m$ 阶矩阵

因变量数据,  
 $n \times 1$ 阶矩阵

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）

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.

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 $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$ 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	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

1、数据输入：

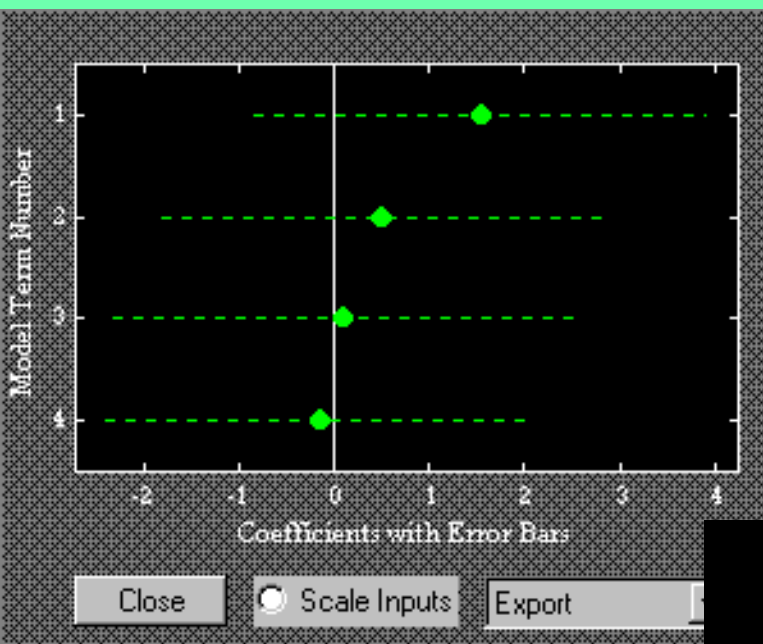
```
x1=[7 1 11 11 7 11 3 1 2 21 1 11 10]';
x2=[26 29 56 31 52 55 71 31 54 47 40 66 68]';
x3=[6 15 8 8 6 9 17 22 18 4 23 9 8]';
x4=[60 52 20 47 33 22 6 44 22 26 34 12 12]';
y=[78.5 74.3 104.3 87.6 95.9 109.2 102.7 72.5 93.1 115.9 83.8 113.3
109.4]';
x=[x1 x2 x3 x4];
```

## 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

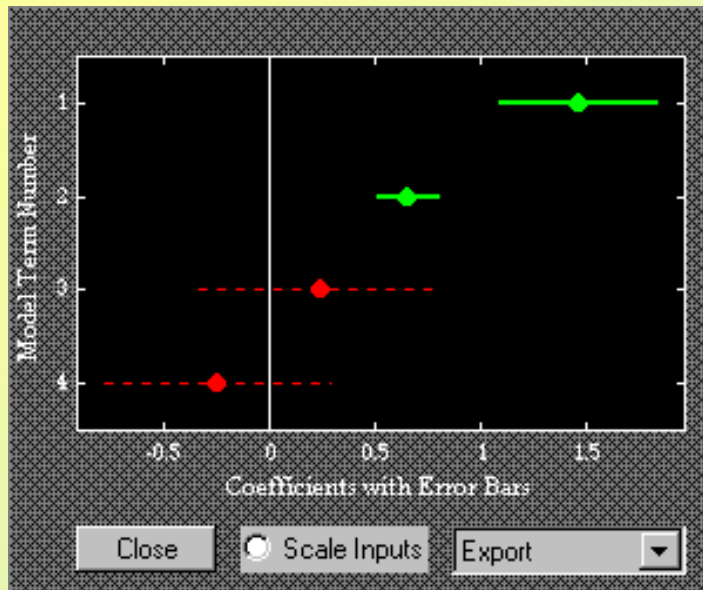


图Stepwise Plot中四条直线都是虚线，说明模型的显著性不好

从表Stepwise Table中看出变量 $x_3$ 和 $x_4$ 的显著性最差.

Column #	Parameter	Confidence Intervals	
		Lower	Upper
1	1.551	-0.8319	3.934
2	0.5102	-1.806	2.826
3	0.1019	-2.313	2.517
4	-0.1441	-2.413	2.125
RMSE	R-square	F	P
2.446	0.9824	111.5	4.756e-0
Close		Help	

(2) 在图Stepwise Plot中点击直线3和直线4，移去变量 $x_3$ 和 $x_4$



移去变量 $x_3$ 和 $x_4$ 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.

Column #	Parameter	Confidence Intervals	
		Lower	Upper
1	1.468	1.1	1.836
2	0.6623	0.5232	0.8013
3	0.25	-0.3235	0.8236
4	-0.2365	-0.7746	0.3015
RMSE	R-square	F	P
2.406	0.9787	229.5	4.407e-0
Close		Help	

[To MATLAB\(liti51\)](#)



(3) 对变量 $y$ 和 $x_1$ 、 $x_2$ 作线性回归:

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

```
b=regress(y,X)
```

To MATLAB(liti52)

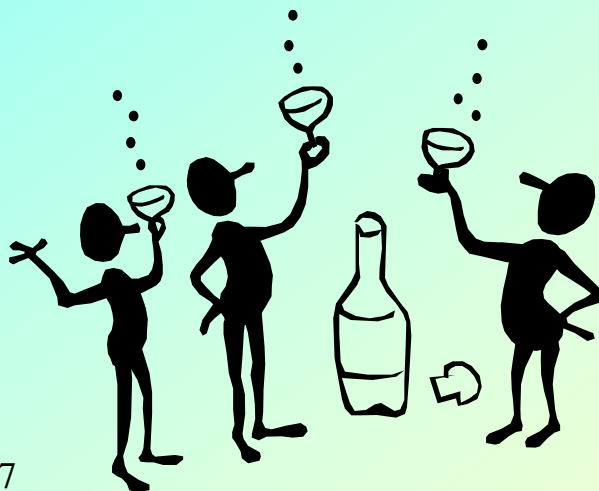
得结果:  $b =$

52.5773

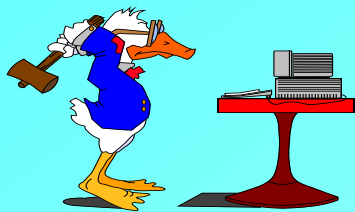
1.4683

0.6623

故最终模型为:  $y=52.5773+1.4683x_1+0.6623x_2$



返回



# 作业

1、考察温度 $x$ 对产量 $y$ 的影响，测得下列10组数据：

温度 ( $^{\circ}\text{C}$ )	20	25	30	35	40	45	50	55	60	65
产量 (kg)	13.2	15.1	16.4	17.1	17.9	18.7	19.6	21.2	22.5	24.3

求 $y$ 关于 $x$ 的线性回归方程，检验回归效果是否显著，并预测 $x=42^{\circ}\text{C}$ 时产量的估值及预测区间（置信度95%）。

2、某零件上有一段曲线，为了在程序控制机床上加工这一零件，需要求这段曲线的解析表达式，在曲线横坐标 $x_i$ 处测得纵坐标 $y_i$ 共11对数据如下：

$x_i$	0	2	4	6	8	10	12	14	16	18	20
$y_i$	0.6	2.0	4.4	7.5	11.8	17.1	23.3	31.2	39.6	49.7	61.7

求这段曲线的纵坐标 $y$ 关于横坐标 $x$ 的二次多项式回归方程。

3、在研究化学动力学反应过程中，建立了一个反应速度和反应物

含量的数学模型，形式为 
$$y = \frac{\beta_1 x_2 - \frac{x_3}{\beta_5}}{1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3}$$

其中  $\beta_1, \dots, \beta_5$  是未知参数， $x_1, x_2, x_3$  是三种反应物（氢，n 戊烷，异构戊烷）的含量， $y$  是反应速度. 今测得一组数据如表 4，试由此确定参数  $\beta_1, \dots, \beta_5$ ，并给出置信区间.  $\beta_1, \dots, \beta_5$  的参考值为  $(1, 0.05, 0.02, 0.1, 2)$  .

序号	反应速度 $y$	氢 $x_1$	n 戊烷 $x_2$	异构戊烷 $x_3$
1	8.55	470	300	10
2	3.79	285	80	10
3	4.82	470	300	120
4	0.02	470	80	120
5	2.75	470	80	10
6	14.39	100	190	10
7	2.54	100	80	65
8	4.35	470	190	65
9	13.00	100	300	54
10	8.50	100	300	120
11	0.05	100	80	120
12	11.32	285	300	10
13	3.13	285	190	120

2022/5/7

4、混凝土的抗压强度随养护时间的延长而增加，现将一批混凝土作成12个试块，记录了养护日期 $x$ （日）及抗压强度 $y$ （ $\text{kg/cm}^2$ ）的数据：

养护时间 $x$	2	3	4	5	7	9	12	14	17	21	28	56
抗压强度 $y$	35	42	47	53	59	65	68	73	76	82	86	99

试求  $\hat{y} = a + b \ln x$  型回归方程.

谢谢大家

