回归分析原理及其应用

参考书目：
1. 庞浩. 计量经济学 第 4 版[M]. 北京：科学出版社, 2019.01.
2. 司守奎，孙玺菁. 数学建模算法与应用（第3版）[M]. 北京：国防工业出版社,2021.04.
3. 赵静，但琦，严尚安，杨秀文.数学建模与数学实验（第5版）[M]. 北京：高等教育出版社，2020.08.

回归分析原理及其应用

1 一元回归参数的最小二乘估计

2 一元线性回归

3 多元回归参数的最小二乘估计

4 多元线性回归

1、一元回归参数的最小二乘估计

最小二乘估计

线性回归模型设定后，我们面临的第一个问题是如何得到样本回归模型中未知参数的估计，构造“合适的”样本回归模型，使得样本回归模型能够“较好地”估计总体回归模型。

1、一元回归参数的最小二乘估计

在回归分析中有很多种构造样本回归函数的方法，而最广泛使用的一种是普通最小二乘法（**method of ordinary least squares，简记OLS**）

德国数学家高斯(C.F.Gauss)

1、一元回归参数的最小二乘估计

一、普通最小二乘法（OLS）

普通最小二乘法是由德国数学家高斯(C.F.Gauss)最早提出和使用的。这种方法除计算比较方便外，在一定的假设条件下，得到的估计量还具有非常好的统计性质，（但这种方法对异常值非常敏感）。从而使它成为回归分析中最有功效和最为流行的方法之一。

1、一元回归参数的最小二乘估计

计量经济分析的目标是寻求总体回归函数：

$$E(Y|X_i) = \beta_1 + \beta_2 X_i$$

在实际分析中，由于很难得到总体的全部观测值，只能得到部分样本值，因此我们希望利用得到的部分样本值来估计出一个样本回归函数（**SRF**）：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad (2.9)$$

利用此样本回归函数来估计总体回归函数。

1、一元回归参数的最小二乘估计

那么，我们如何估计出 $\hat{\beta}_1, \hat{\beta}_2$ 呢？

对于给定的Y 和 X 的 n 对观测值

$$(X_i, Y_i) \qquad i = 1, 2, \cdots, n \qquad (2.10)$$

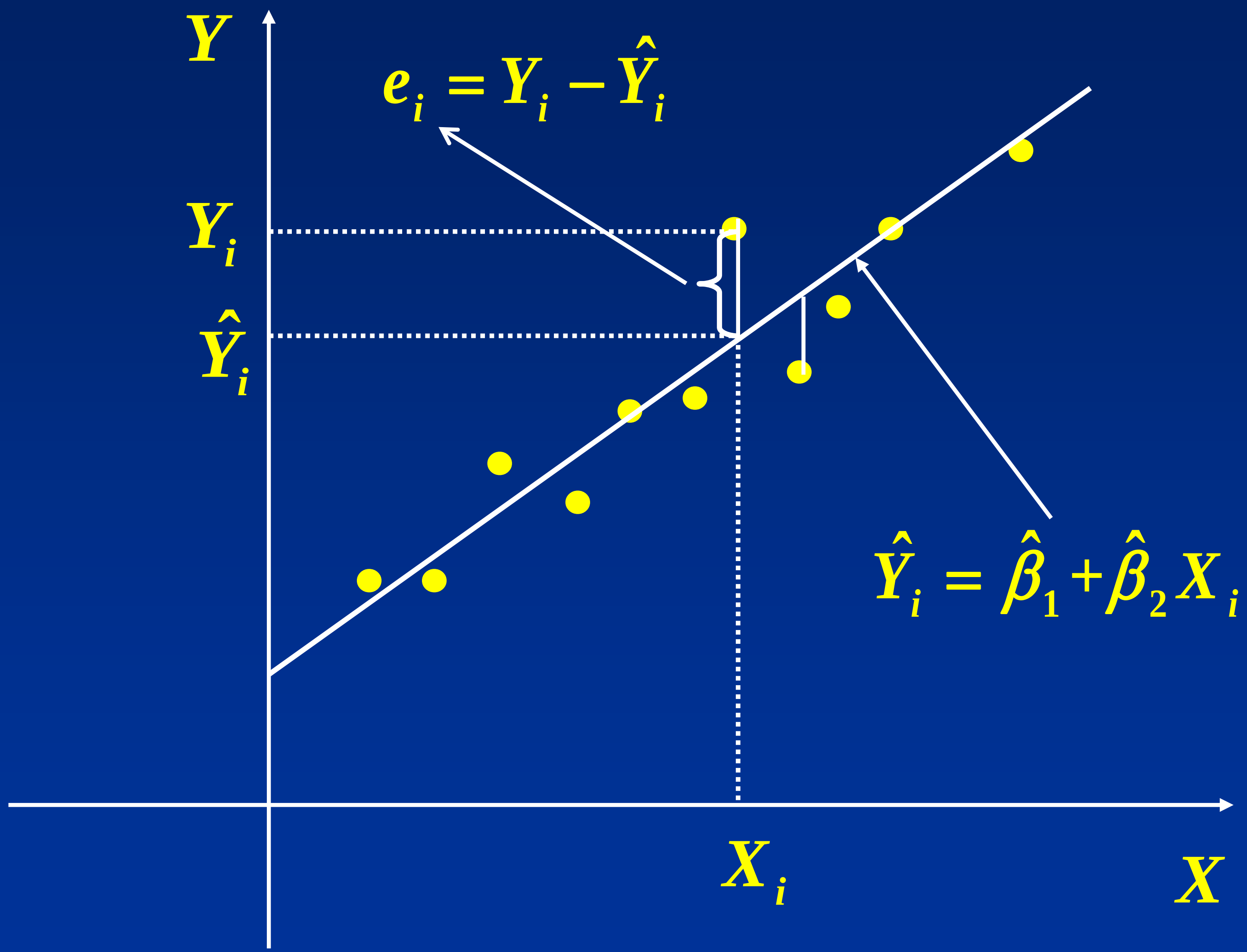
我们希望样本回归模型的估计值 $\hat{Y}_i$ 尽可能地

靠近观测值Y_i 。即残差：

$$e_i = Y_i - \hat{Y}_i, \qquad (2.11)$$

越小越好。

1、一元回归参数的最小二乘估计



1、一元回归参数的最小二乘估计

为了达到此目的，我们就必须使用最小二乘准则（残差平方和最小），使：

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \qquad (2.12)$$

尽可能地小，其中，e_i^2是残差的平方。

思考：为什么要使用残差平方和最小准则？

1、一元回归参数的最小二乘估计

记 $Q=\sum e_i^2 = f(\hat{\beta}_1,\hat{\beta}_2)$

$$=\sum (Y_i-\hat{\beta}_1-\hat{\beta}_2 X_i)^2$$

由多元函数的极值的求法知应有：

正规方程组

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_1} = -2\sum (Y_i-\hat{\beta}_1-\hat{\beta}_2 X_i)=0 & (2.14) \\ \dfrac{\partial Q}{\partial \hat{\beta}_2} = -2\sum (Y_i-\hat{\beta}_1-\hat{\beta}_2 X_i)X_i=0 & (2.15) \end{cases}$$

1、一元回归参数的最小二乘估计

上述方程组可化为：

$$
\begin{cases}
\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i & (2.16) \\
\sum X_i Y_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 & (2.17)
\end{cases}
$$

将上述方程组看成是以 $\hat{\beta}_1, \hat{\beta}_2$ 为未知量的线性方程组，求解该方程组，可得：

1、一元回归参数的最小二乘估计

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} \qquad (2.18)$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \qquad (2.19)$$

这里： $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$

上面得到的估计量 $\hat{\beta}_1$，$\hat{\beta}_2$ 是从最小二乘原理演算而得的。因此，称其为 β_1 和 β_2 最小二乘估计量。

1、一元回归参数的最小二乘估计

估计量（estimator）与估计值（ estimate)的区别：

估计值：由具体样本资料计算出来的结果就是估计值或点估计。是估计量 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 的一个具体数值。

估计量：$\hat{\beta}_1, \hat{\beta}_2$ 是样本 Y_i 的一个表达式，是 Y_i 的函数，而 Y_i 是随机变量，所以，$\hat{\beta}_1, \hat{\beta}_2$ 也是随机变量。

1、一元回归参数的最小二乘估计

实际上，这是样本的二重性。当我们要作统计推断时，需要用样本的观测值来计算估计量的观测值（估计值）。当我们需要了解估计量的性质时，需要将估计量作为随机变量来讨论。

1、一元回归参数的最小二乘估计

【注】 1. 最小二乘估计量 $\hat{\beta}_1,\hat{\beta}_2$ 的其它形式：

（1） $$\hat{\beta}_2=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}=\frac{l_{xy}}{l_{xx}}$$

$$\hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}$$

这里， $$l_{xx}=\sum(X_i-\bar{X})^2=\sum X_i^2-n(\bar{X})^2$$

$$l_{yy}=\sum(Y_i-\bar{Y})^2=\sum Y_i^2-n(\bar{Y})^2$$

$$l_{xy}=\sum(X_i-\bar{X})(Y_i-\bar{Y})=\sum X_iY_i-n\bar{X}\bar{Y}$$

1、一元回归参数的最小二乘估计

（2） $$\hat{\beta}_2=\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}=\sum\left(\frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}\right)Y_i$$

$$\hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}$$

$$=\sum_{i=1}^{n}\left(\frac{1}{n}-\left(\frac{(X_i-\bar{X})\bar{X}}{\sum(X_i-\bar{X})^2}\right)\right)Y_i$$

这里用到结论：$\sum(X_i-\bar{X})=0$

这说明 $\hat{\beta}_1,\hat{\beta}_2$ 均为 $Y_1,Y_2,\cdots,Y_n$ 的线性函数。

1、一元回归参数的最小二乘估计

（3）$\hat{\beta}_2 = r_{XY}\sqrt{\dfrac{\sum(Y_i-\bar{Y})^2}{\sum(X_i-\bar{X})^2}} = r_{XY}\sqrt{\dfrac{S_Y^2}{S_X^2}} = r_{XY}\dfrac{S_Y}{S_X}$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$$

其中，$r_{XY} = \dfrac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}$ 是 X，Y 的样本相关系数。

由此可以看出，当X，Y 正相关时，$\hat{\beta}_2 > 0$ 当X，Y 负相关时，$\hat{\beta}_2 < 0$。且 $\hat{\beta}_2$与相关系数 r_{XY} 成正比。

1、一元回归参数的最小二乘估计

【注】 2. 正规方程组的另一种表示形式：

$$\begin{cases} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = \sum_{i=1}^{n} e_i = 0 \\ \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = \sum_{i=1}^{n} X_i e_i = 0 \end{cases}$$

$$\Rightarrow \bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0$$

$$\Rightarrow \mathrm{cov}(X_i, e_i) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(e_i - \bar{e}) = 0$$

说明残差和所有解释变量不相关。

1、一元回归参数的最小二乘估计

【注】 3. OLS估计值受变量单位的影响：

注意到：$\hat{\beta}_2=\dfrac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$

$$\hat{\beta}_1=\bar{Y}-\hat{\beta}_2\bar{X}$$

（1） 若只有因变量 Y 的单位扩大 c 倍时，

则截距和斜率都扩大为原来的 c 倍。

1、一元回归参数的最小二乘估计

（2） 若只有自变量 X 的单位扩大 c 倍时，则斜率缩小为原来的 c 倍，而截距没有变化。

（3） 若因变量 Y 和自变量 X 的单位都扩大 c 倍时，则斜率没有变化，而截距扩大为原来的 c 倍。

1、一元回归参数的最小二乘估计

【注】 4. 评价估计量的标准：

评价由小样本构造的估计量的常用标准有：

（1）线性性

（2）无偏性

（3）有效性

1、一元回归参数的最小二乘估计

评价由大样本构造的估计量的常用标准有：

（4）渐进无偏性：样本容量无穷大时均值序列趋于总体真值。

（5）一致性：样本容量无穷大时估计量的均值依概率收敛于总体真值。

（6） 渐进有效性：样本容量无穷大时，它在所有的一致估计量中具有最小的方差。

1、一元回归参数的最小二乘估计

【例2.1】 下表示某地区10户家庭人均收入（X）和人均食物消费支出（Y）的数据

Y	X	Y	X
70	80	115	180
65	100	120	200
90	120	140	220
95	140	155	240
110	160	150	260

试用普通最小二乘法估计该地区居民人均食物消费支出的回归直线：$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

1、一元回归参数的最小二乘估计

【解】 $\sum_{i=1}^{10} x_i = 1700$ $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 170$

$\sum_{i=1}^{10} y_i = 1110$ $\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 111$

$$\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y}) = 16800$$

$$\sum_{i=1}^{10}(x_i - \bar{x})^2 = 33000$$

1、一元回归参数的最小二乘估计

又因为：

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{10}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{10}(x_i - \bar{x})^2} = \frac{16800}{33000} = 0.5091$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 111 - 0.5091 \times 170 = 24.4545$$

故所求回归方程是：

$$\hat{Y}_i = 24.4545 + 0.5091 X_i$$

2 一元线性回归

2.1 数学模型

2.2 模型参数估计

2.3 检验、预测与控制

2.4 可线性化的一元非线性回归（曲线回归）

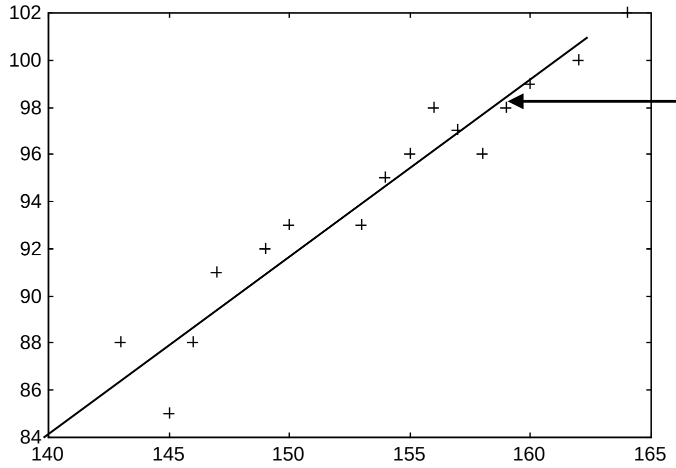
2.1数学模型

例1 测16名成年女子的身高与腿长所得数据如下：

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

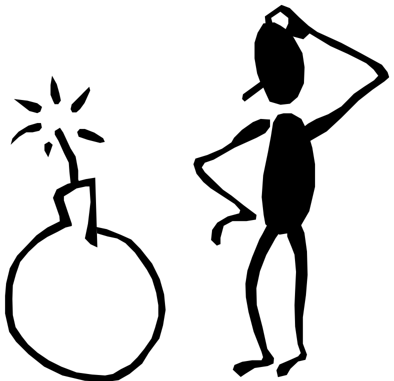
以身高x为横坐标，以腿长y为纵坐标将这些数据点（x_i，y_i）在平面直角坐标系上标出.

解答



散点图

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



2.1数学模型

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.

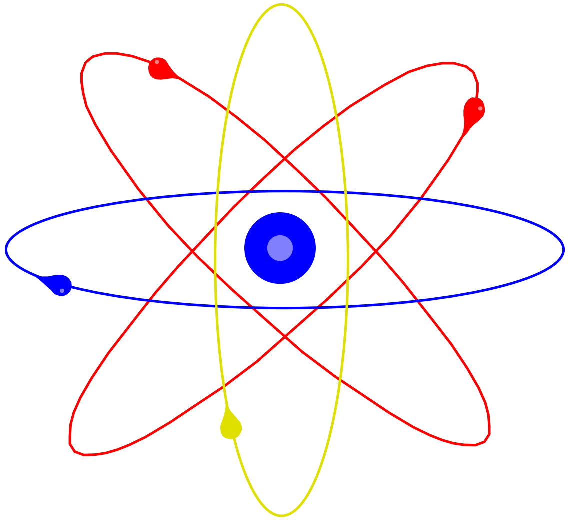
$Y=\beta_0+\beta_1 x$，称为 **y 对 x 的回归直线方程.**

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.



返回

2.2模型参数估计

1. 回归系数的最小二乘估计

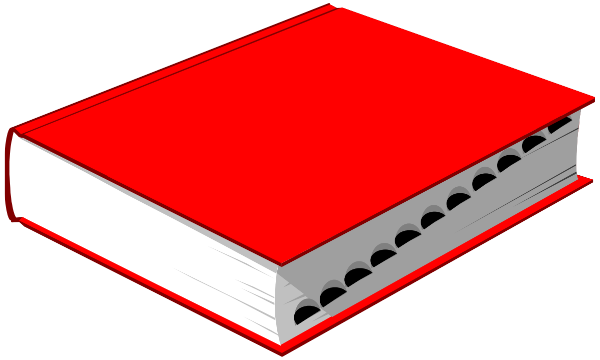
有 n 组独立观测值（x_1，y_1），（x_2，y_2），…，（x_n，y_n）

设 $\begin{cases} y_i = \beta_0 + \beta x_1 + \varepsilon_i, i = 1, 2, \dots, n \\ E\varepsilon_i = 0, D\varepsilon_i = \sigma^2 \quad \text{且} \varepsilon_1 \varepsilon_2, \dots, \varepsilon_n \text{相互独立} \end{cases}$

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 β_0 和 β_1 的估计 $\hat{\beta}_0$，$\hat{\beta}_1$ 使得

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



2.2模型参数估计

解得

或 $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

其中 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$， $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2, \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

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

2.2模型参数估计

2. σ^2 的无偏估计

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

称 Q_e 为残差平方和或剩余平方和.

σ^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.

返回

2.3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.

2.3检验、预测与控制

（Ⅰ）*F*检验法

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

其中 $U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ （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，$T=\dfrac{\sqrt{L_{xx}}\hat{\beta}_1}{\hat{\sigma}_e}$ ~t（n-2）

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

2.3检验、预测与控制

（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$|r|>\mathrm{r}_{1-\alpha}$时，拒绝 H_0；否则就接受 H_0.

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

2.3检验、预测与控制

2. 回归系数的置信区间

β_0 和 β_1 置信水平为 1-α 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 $$\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e/\sqrt{L_{xx}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e/\sqrt{L_{xx}}\right]$$

σ^2 的置信水平为 1-α 的置信区间为

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

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**.

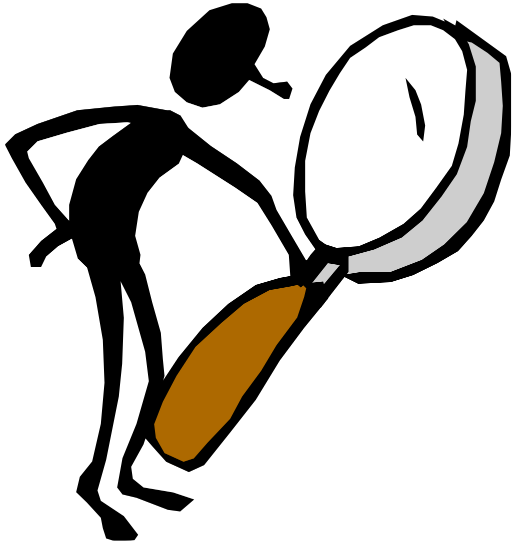
y_0 的置信水平为 $1-\alpha$ 的**预测区间**为

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

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**

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



2.3检验、预测与控制

（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 $\left(y', y''\right)$

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 $\left(x', x''\right)$ 就是所求的 x 的控制区间.

返回

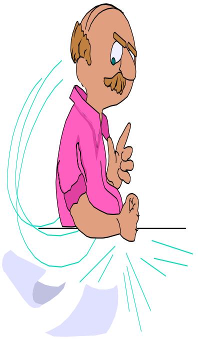
2.4可线性化的一元非线性回归（曲线回归）

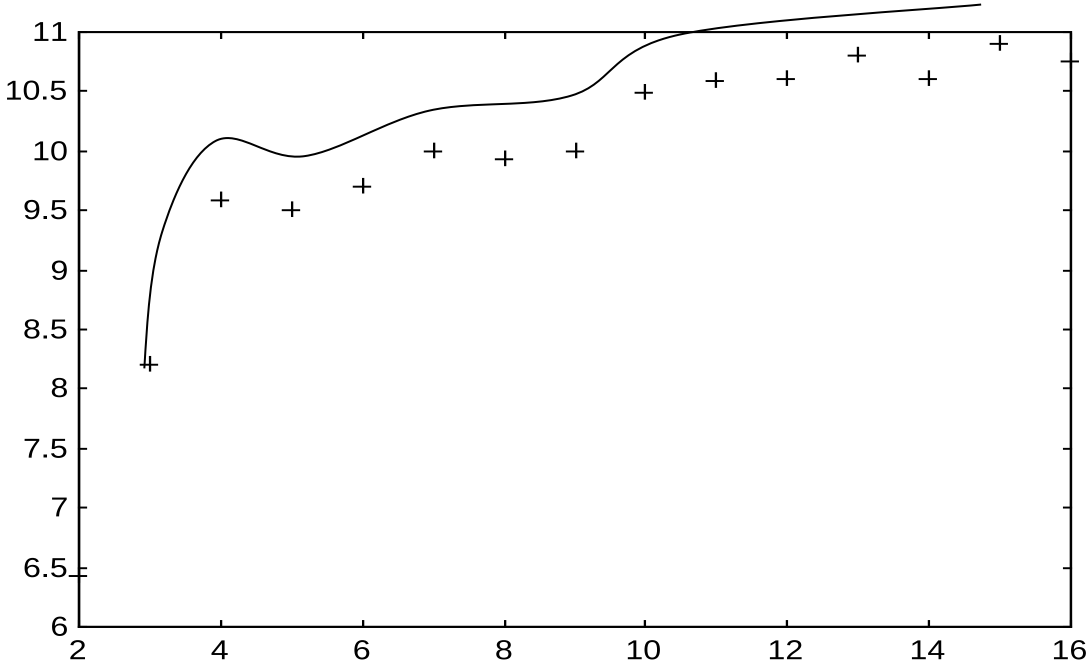
例2 出钢时所用的盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断增大.我们希望知道使用次数与增大的容积之间的关系.对一钢包作试验，测得的数据列于下表：

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

2.4可线性化的一元非线性回归（曲线回归）





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 $(x_i, y_i), i=1,2,\ldots,n$ 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.

2.4可线性化的一元非线性回归（曲线回归）

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** $y=a\,x^{b}$，其中 $x>0,a>0$

（3）**指数曲线** $y=a\,e^{bx}$ 其中参数 $a>0$.

（4）**倒指数曲线** $y=a\,e^{b/x}$ 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.由散点图我们选配到指数曲线 $y=a\,e^{b/x}$

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

3、多元回归参数的最小二乘估计

一、多元回归模型参数的最小二乘估计

利用样本观测值对模型进行估计，其多元回归模型的样本回归函数为：

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}$$

残差为:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}) \end{aligned}$$

3、多元回归参数的最小二乘估计

根据最小二乘准则，应选择使残差平方和最小的 $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ 作为参数 $\beta_1, \beta_2, \cdots, \beta_k$ 的估计。

要使残差平方和：

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}) \right]^2$$

达到最小，由多元函数的极值条件知：

$$\frac{\partial Q}{\partial(\hat{\beta}_j)} = \frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial(\hat{\beta}_j)} = 0, \qquad j = 1, 2, \cdots, k$$

3、多元回归参数的最小二乘估计

即

$$e_i$$

$$\begin{cases} \sum_{i=1}^{n}\boxed{(Y_i-\hat\beta_1-\hat\beta_2 X_{2i}-\cdots-\hat\beta_k X_{ki})}=0 \\ \sum_{i=1}^{n}X_{2i}\boxed{(Y_i-\hat\beta_1-\hat\beta_2 X_{2i}-\cdots-\hat\beta_k X_{ki})}=0 \\ \cdots\cdots \\ \sum_{i=1}^{n}X_{ki}\boxed{(Y_i-\hat\beta_1-\hat\beta_2 X_{2i}-\cdots-\hat\beta_k X_{ki})}=0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n}e_i=0 \\ \sum_{i=1}^{n}X_{2i}e_i=0 \\ \cdots\cdots \\ \sum_{i=1}^{n}X_{ki}e_i=0 \end{cases}$$

此方程组称为正规方程组。

3、多元回归参数的最小二乘估计

整理**正规方程组**可得：

$$
\begin{cases}
n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki} = \sum_{i=1}^{n} Y_i \\
\hat{\beta}_1 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki} X_{2i} = \sum_{i=1}^{n} X_{2i} Y_i \\
\quad \cdots\cdots \\
\hat{\beta}_1 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ki} X_{2i} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2 = \sum_{i=1}^{n} X_{ki} Y_i
\end{cases}
$$

3、多元回归参数的最小二乘估计

矩阵形式为：

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_{2i} & \cdots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{2i}^{2} & & \sum_{i=1}^{n} X_{ki} X_{2i} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki} X_{2i} & \cdots & \sum_{i=1}^{n} X_{ki}^{2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{2i} Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ki} Y_i \end{pmatrix}$$

3、多元回归参数的最小二乘估计

X 为样本观测矩阵

令

$$X=\begin{pmatrix}1 & X_{21} & \cdots & X_{k1}\\ 1 & X_{22} & \cdots & X_{k2}\\ \vdots & \vdots & & \vdots\\ 1 & X_{2n} & \cdots & X_{kn}\end{pmatrix}_{n\times k},\quad \hat{\beta}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \\ \hat{\beta}_k\end{pmatrix},\quad Y=\begin{pmatrix}Y_1\\ Y_2\\ \\ Y_n\end{pmatrix}$$

则

$$\begin{pmatrix}\sum_{i=1}^{n}Y_i\\ \sum_{i=1}^{n}X_{2i}Y_i\\ \vdots\\ \sum_{i=1}^{n}X_{ki}Y_i\end{pmatrix}=\begin{pmatrix}1 & 1 & \cdots & 1\\ X_{21} & X_{22} & \cdots & X_{2n}\\ \vdots & \vdots & & \vdots\\ X_{k1} & X_{k2} & \cdots & X_{kn}\end{pmatrix}\begin{pmatrix}Y_1\\ Y_2\\ \\ Y_n\end{pmatrix}=X^TY$$

3、多元回归参数的最小二乘估计

$$\begin{pmatrix} n & \sum_{i=1}^{n} X_{2i} & \cdots & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{2i}^{2} & & \sum_{i=1}^{n} X_{ki} X_{2i} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki} X_{2i} & \cdots & \sum_{i=1}^{n} X_{ki}^{2} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{pmatrix} = X^{T} X$$

3、多元回归参数的最小二乘估计

于是，正规方程组的矩阵形式为：

$$X^T X \hat{\beta} = X^T Y$$

由古典假设条件中解释变量无多重共线性

假定，可知矩阵 $X^T X$ 可逆，于是 β 的最小二乘估计量的矩阵表达式为：

$$\underbrace{\hat{\beta}}_{k\times 1} = (\underbrace{X^T X}_{k\times k})^{-1} \underbrace{X^T}_{k\times n} \underbrace{Y}_{n\times 1}$$

4 多元线性回归

4.1数学模型及定义

一般称 $\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, \mathrm{COV}(\varepsilon, \varepsilon) = \sigma^2 I_n \end{cases}$

为高斯-马尔可夫线性模型（k 元线性回归模型），并简记为$(Y, X\beta, \sigma^2 I_n)$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

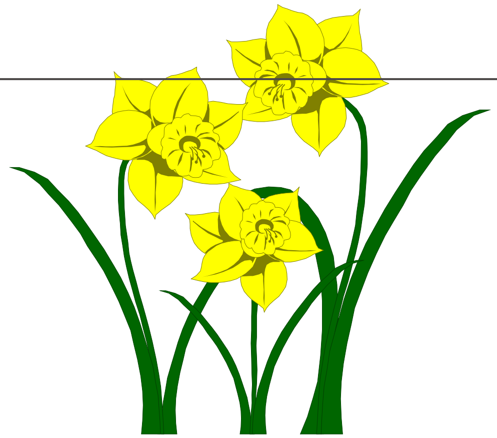
$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ 称为回归平面方程.

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 $x_1, x_2, \ldots, x_k$ 之间的数量关系；

（2）在 $x_1 = x_{01}, x_2 = x_{02}, \ldots, x_k = x_{0k}$，处对 y 的值作预测与控制，即对 y 作区间估计.

4.2模型参数估计



1. 对 β_i 和 σ^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.

解得估计值 $\hat{\beta} = (X^T X)^{-1}(X^T Y)$

得到的 $\hat{\beta}_i$ 代入回归平面方程得：

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

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 β 的无偏估计，协方差阵为 $\sigma^2 C$.$C=L^{-1}=(c_{ij})$, $L=X^T X$

4.2模型参数估计

2. 多项式回归

设变量 x、Y 的回归模型为

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

其中 p 是已知的，$\beta_i \, (i = 1,2,\cdots,p)$ 是未知参数，ε 服从正态分布

$N(0,\sigma^2)$.

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

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，…，k 多项式回归模型变为多元线性回归模型.

4.3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 $$H_0: \beta_0 = \beta_1 = \cdots = \beta_k = 0$$

（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 $x_1,\cdots, x_k$ 之间显著地有线性关系；否则就接受 H_0，认为 y 与 $x_1,\cdots, x_k$ 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 $x_1,x_2,\ldots,x_k$ 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.

4.3多元线性回归中的检验与预测

2. 预测

（1）点预测

求出回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_kx_k$，对于给定自变量的值 $x_1^*,\cdots,x_k^*$，用 $\hat{y}^*=\hat{\beta}_0+\hat{\beta}_1x_1^*+\cdots+\hat{\beta}_kx_k^*$ 来预测 $y^*=\beta_0+\beta_1x_1^*+\cdots+\beta_kx_k^*+\varepsilon$.称 $\hat{y}^*$ 为 y^* 的点预测.

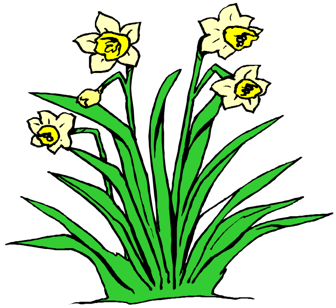
（2）区间预测

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

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

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

$C=L^{-1}=(c_{ij}),\ L=X^{\mathrm{T}}X$



4.4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.

4.4逐步回归分析

逐步回归分析法的思想：

- 从一个自变量开始，视自变量Y对作用的显著程度，从大到小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.

4.4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

4.4统计工具箱中的回归分析命令

多元线性回归

1. 确定回归系数的点估计值：

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

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

对一元线性回归，取 p=1 即可

4.4统计工具箱中的回归分析命令

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

置信区间

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）

4.4统计工具箱中的回归分析命令

例1 解： 1. 输入数据：

题目

```
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
```

To MATLAB(liti11)

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

p<0.05，可知回归模型 y=-16.073+0.7194x 成立.

4.4统计工具箱中的回归分析命令

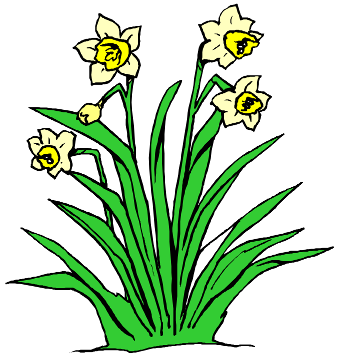
3. 残差分析，作残差图：

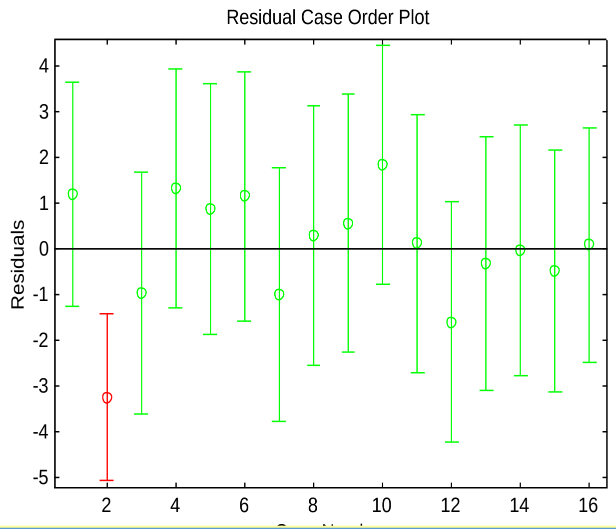
```
rcoplot(r,rint)
```

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





To MATLAB(liti12)

返回

4.4统计工具箱中的回归分析命令

多项式回归

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

1．回归：

（1）确定多项式系数的命令：[p，S]=polyfit（x，y，m）

其中 $x=(x_1, x_2, \cdots, x_n)$，$y=(y_1, y_2, \cdots, y_n)$；
$p=(a_1, a_2, \cdots, a_{m+1})$ 是多项式 $y=a_1x^m+a_2x^{m-1}+\cdots+a_mx+a_{m+1}$ 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.

4.4统计工具箱中的回归分析命令

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：

To MATLAB（liti21）

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

4.4统计工具箱中的回归分析命令

法二

化为多元线性回归：

To MATLAB(liti22)

```
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：

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

预测及作图

To MATLAB(liti23)

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

4.4统计工具箱中的回归分析命令

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

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

4.4统计工具箱中的回归分析命令

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

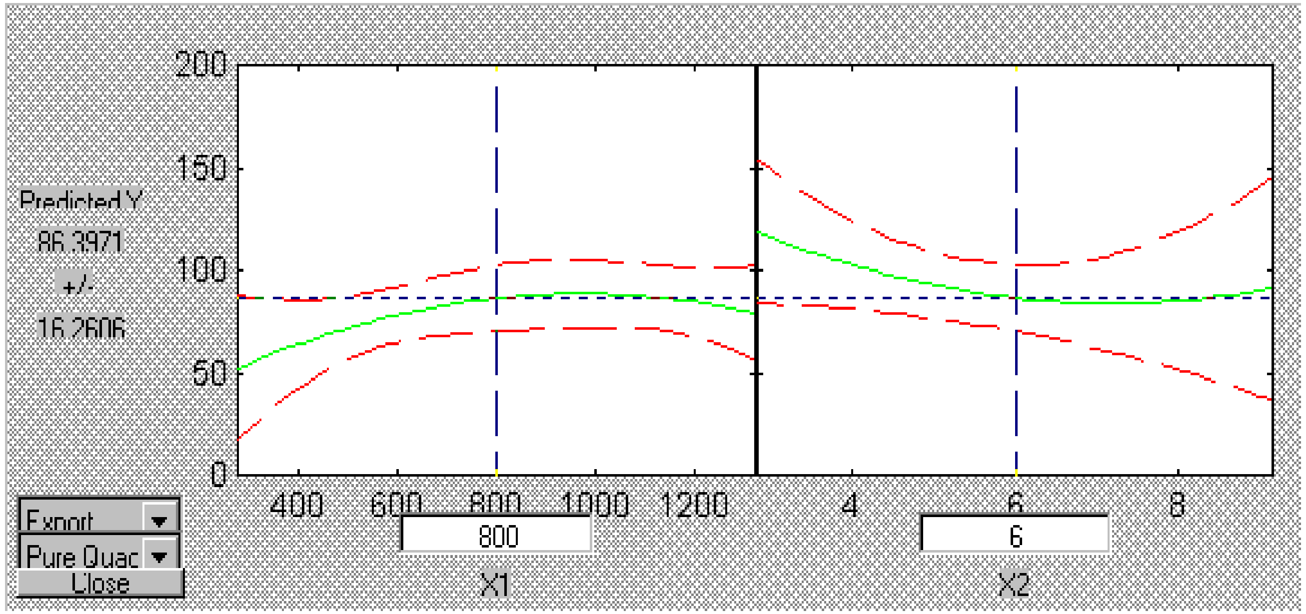
选择纯二次模型，即 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$

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')
```

4.4统计工具箱中的回归分析命令



将左边图形下方方框中的“800”改成1000，右边图形下方的方框中仍输入6. 则画面左边的“Predicted Y”下方的数据由原来的“86.3791”变为88.4791，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.

4.4统计工具箱中的回归分析命令

在MATLAB工作区中输入命令：beta, rmse

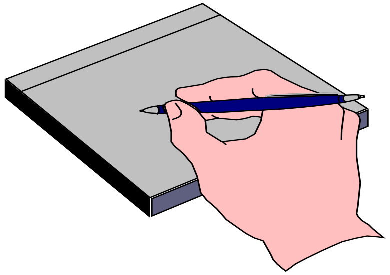
得结果：
```
beta =
        110.5313
        0.1464
        -26.5709
        -0.0001
        1.8475
rmse =
        4.5362
```

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.

To MATLAB(liti31)

4.4统计工具箱中的回归分析命令



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
```

To MATLAB(liti32)

结果为: b =

```
110.5313
0.1464
-26.5709
-0.0001
1.8475
```

```
stats =
   0.9702    40.6656    0.0005
```

返回

4.4统计工具箱中的回归分析命令

非线性回 归

1. 回归:

（1）确定回归系数的命令：

[beta，r，J]=nlinfit（x,y,′model′,beta0）

事先用M文件定义的非线性函数

估计出的回归系数

残差

Jacobi矩阵

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.

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lintool所得的回归函数在x处的预测值Y及预测值的显著性水平为1-alpha的置信区间Y ± DELTA.

4.4统计工具箱中的回归分析命令

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 $y=a\,e^{b/x}$，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]';
```

To MATLAB(liti41)

3. 求回归系数：

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

得结果：beta =

```
11.6036
-1.0641
```

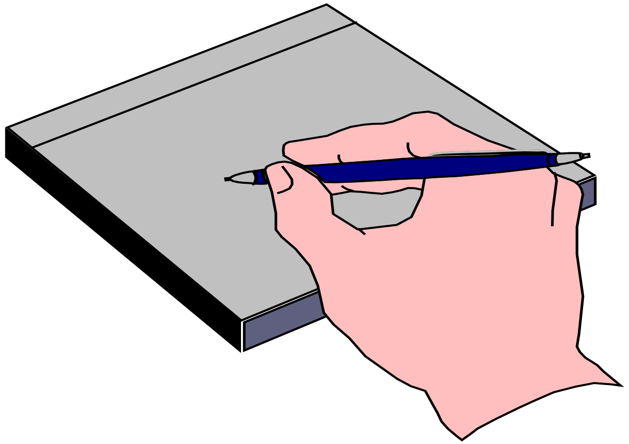
即得回归模型为：

$$y = 11.6036e^{-\frac{1.0641}{x}}$$

4.4统计工具箱中的回归分析命令

4. 预测及作图:

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



To MATLAB(liti42)

4.4统计工具箱中的回归分析命令

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.表中列出了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.

4.4统计工具箱中的回归分析命令

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;
```

4.4统计工具箱中的回归分析命令

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)

4.4统计工具箱中的回归分析命令

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

4.4统计工具箱中的回归分析命令

逐步回归

显著性水平（缺省时为0.05）

逐步回归的命令是：

stepwise（x，y，inmodel，alpha）

自变量数据，N*M 阶矩阵

因变量数据，N*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**命令时产生图形窗口：**Stepwise Regression**

显示出各项的回归系数及其置信区间.

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**.

4.4统计工具箱中的回归分析命令

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];
```

4.4统计工具箱中的回归分析命令

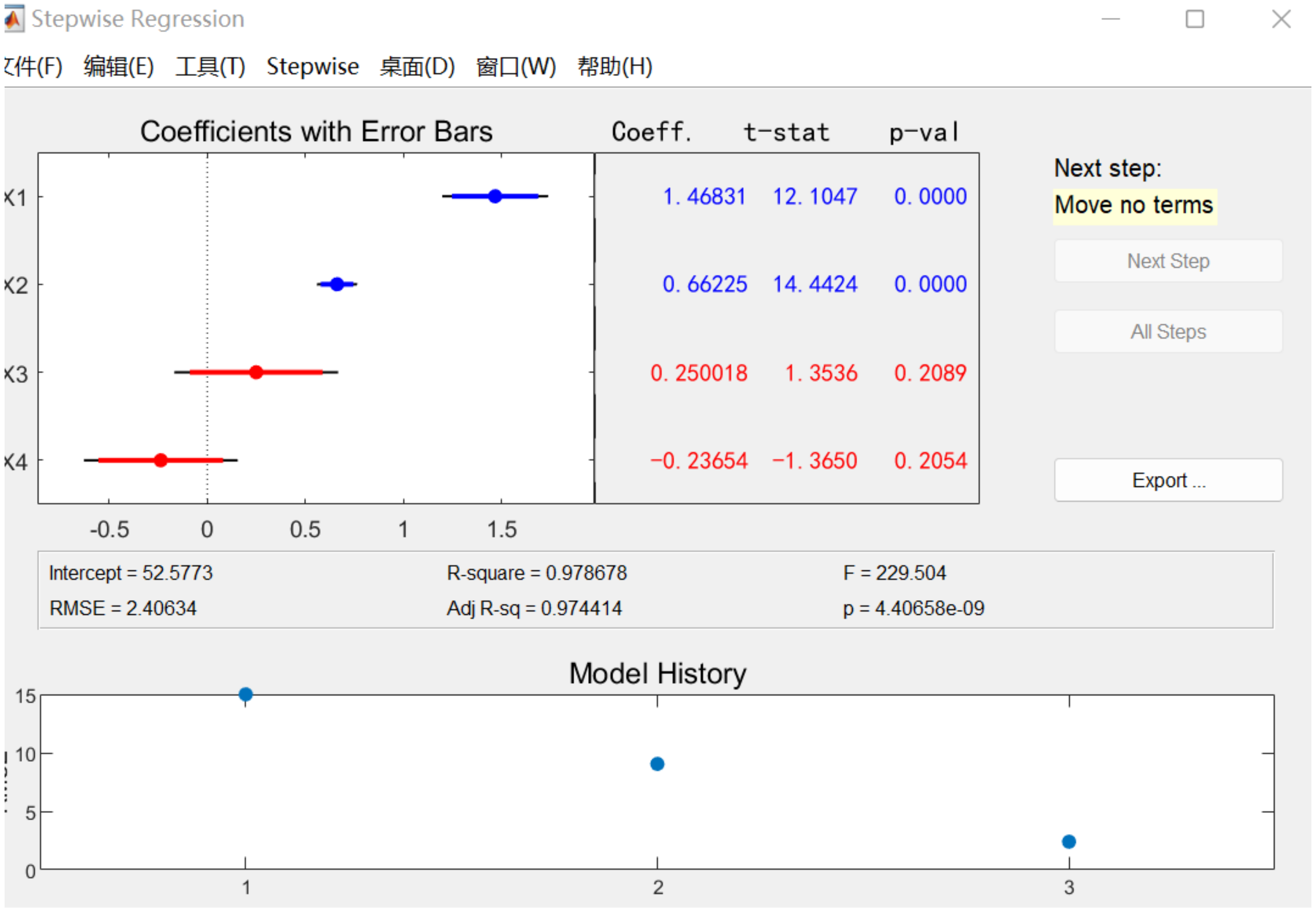
2. 逐步回归:

（1）先在初始模型中取全部自变量： stepwise(x,y)

得图Stepwise Regression

4.4统计工具箱中的回归分析命令

（2）在图Stepwise Plot中点击直线1和直线2，进入变量x_1和x_2



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.

To MATLAB(liti51）

4.4统计工具箱中的回归分析命令

（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

小结

回归分析

- 一元线性回归
  - 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）
- 多元线性回归
  - 数学模型及定义
  - *模型参数估计
  - *多元线性回归中的检验与
    测
  - 逐步回归分析

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（℃）	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℃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.

作业

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 $y=\dfrac{\beta_1 x_2-\dfrac{x_3}{\beta_5}}{1+\beta_2 x_1+\beta_3 x_2+\beta_4 x_3}$

其中 $\beta_1,\cdots,\beta_5$ 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.今测得一组数据如下表，试由此确定参数 $\beta_1,\cdots,\beta_5$，并给出置信区间. $\beta_1,\cdots,\beta_5$ 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

作业

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（kg/cm²）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