回归分析

- **函数关系**：当一个或若干个变量x取一定值时，某一个变量y有确定的值与之相对应。

  例：圆面积S与圆半径的关系 $S=\pi r^2$

- **相关关系**：当一个或若干个变量x取一定值时，与之相对于的另一个变量y的值虽然不确定，但却按某种规律在一定范围内变化。

  例：居民的可支配收入x与居民的消费支出y之间的关系。

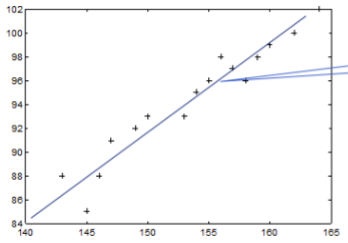
- **回归分析**：处理变量之间的相关关系的数学方法。

一元线性回归

例 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。



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

散点图

- **一元线性回归模型**

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

- 回归系数：β_0、β_1
- 回归变量：自变量 x
- y对x的回归直线方程：$Y=\beta_0+\beta_1 x$

- 一元线性回归分析的主要任务：
  - 用实验值（样本值）对β_0、β_1和σ作点估计；
  - 对回归系数β_0、β_1作假设检验；
  - 在 $x=x_0$ 处对 y 作预测，对 y 作区间估计。

➢ 回归系数的最小二乘估计

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

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

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

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

➢ 回归系数的最小二乘估计

由最小二乘法解得

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

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

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

➢ σ^2的无偏估计

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

$\boldsymbol{Q_e}$：残差平方和、剩余平方和

- σ^2的无偏估计：$\hat{\sigma}_e^2 = \frac{Q_e}{n-2}$
- $\hat{\sigma}_e^2$：剩余方差（残差的方差）
- $\hat{\sigma}_e^2$分别与$\hat{\beta}_0$、$\hat{\beta}_1$独立。

➢ 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，归结为如下假设检验。

$\mathbf{H_0}$： $\beta_1 = 0$ ； $\mathbf{H_1}$： $\beta_1 \neq 0$ ；

- 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，所得的回归方程无意义。

- 回归方程的显著性检验
- F检验法

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

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

- $F>F_{1-\alpha}(1,n\text{-}2)$，拒绝$H_0$，否则接受$H_0$。

- 回归方程的显著性检验
- t 检验法

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

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

- $|T|>t_{1-\alpha/2}(n-2)$，拒绝H_0，否则接受H_0

- 回归方程的显著性检验
- r 检验法

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

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

- $|r|>r_{1-\alpha}$，拒绝H_0，否则接受H_0

- 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

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

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

- σ^2的置信水平为1-α的置信区间为

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

➢ 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-α的预测区间： $[\hat{y}_0 - \delta(x_0),\ \hat{y}_0 + \delta(x_0)]$

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

当n很大且x_0在 $\bar{x}$ 附近取值时，

- y的置信水平为1-α的预测区间近似为

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

➢ 要求：

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

➢ 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。

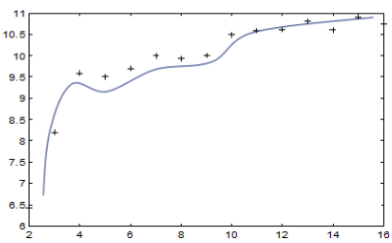
➢ **例** 出钢时所用的盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断增大。我们希望知道使用次数与增大的容积之间的关系。对一钢包作试验，测得的数据列于下表：

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		

➢ 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：

- 对变量x和y作n次试验观察得$(x_i, y_i)(i=1,2,\ldots,n)$。
- 画散点图，根据散点图确定须配曲线类型。



- 由n对试验数据确定每一类曲线的未知参数。
- 非线性回归线性化方法：通过变量代换把非线性回归化成线性回归。

➢ 通常选择的六类曲线

- 双曲线 $\frac{1}{y}=a+\frac{b}{x}$
- 幂函数曲线 $y=ax^b$ $(x>0, a>0)$
- 指数曲线 $y=ae^{bx}$ $(a>0)$
- 倒指数曲线 $y=ae^{\frac{b}{x}}$ $(a>0)$
- 对数曲线 $y=a+blog(x)$ $x>0$
- S型曲线 $y=\frac{1}{a+be^{-x}}$

$\boldsymbol{y=\beta_0+\beta_1x_1+\ldots+\beta_px_p}$

➢ 确定回归系数的点估计值

b=regress(Y, X)

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

$\boldsymbol{y=\beta_0+\beta_1x_1+\ldots+\beta_px_p}$

➢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：相关系数R^2；F值；与F对应的概率p；σ^2的估计值。

- 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，$p<\alpha$时拒绝H_0，回归模型成立；
- 模型误差项ε的方差σ^2的估计值。

$\boldsymbol{y=\beta_0+\beta_1x_1+\ldots+\beta_px_p}$

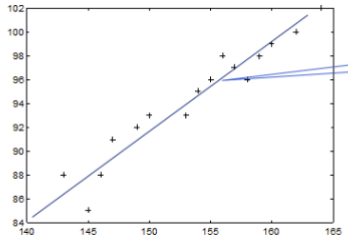
➢残差分析，画出残差及其置信区间：

rcoplot(r, rint)

例 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。



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

散点图

- **输入数据：(regress_eg_1.m)**

```
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
```

- **结论**

$\hat{\beta}_0 = -16.0730$ ；

$\hat{\beta}_1 = 0.7194$ ；

$\hat{\beta}_0$ 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。

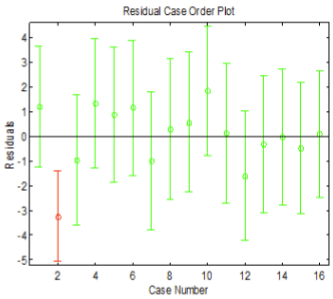
- 回归模型 y=-16.073+0.7194x 成立.

- **输出：**

```
b =
  -16.0730
    0.7194
bint =
  -33.7071    1.5612
    0.6047    0.8340
stats =
    0.9282  180.9531
    0.0000    1.7437
```

➢ 残差分析，作残差图：

rcoplot(r,rint)



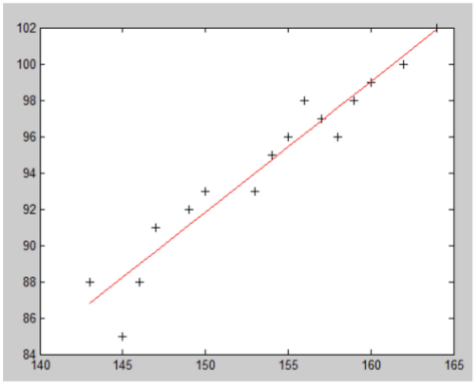
➢ 结论：

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。

➢ 预测及作图：

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

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



```
z =
   86.7944
   88.2331
   88.9524
   89.6718
   91.1105
   91.8298
   93.9879
   94.7073
   95.4266
   96.1460
   96.8653
   97.5847
   98.3040
   99.0234
  100.4621
  101.9008
```

一元多项式回归

$$y=a_1x^m+a_2x^{m-1}+\ldots+a_mx+a_{m+1}$$

➢ 回归

- 确定多项式系数：**[p，S]=polyfit(x, y, m)**

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

  $p=(a_1, a_2, \ldots, a_{m+1})$是多项式的系数；

  S是一个矩阵，用来估计预测误差。

- 多项式回归命令：polytool(x, y, m)

一元多项式回归

$$y=a_1x^m+a_2x^{m-1}+\ldots+a_mx+a_{m+1}$$

➢ 预测和预测误差估计

- **Y=polyval(p, x)**

  求polyfit所得的回归多项式在x处的预测值Y；

- **[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, Y+DELTA]。

➢ 例 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

➢方法一：二次多项式回归

- 输入：(regress_eg_2.m)

```
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,a]=polyfit(t,s,2)
```

- 回归模型：$s = 489.2946t^2+65.8896t+9.1329$

➢方法一：二次多项式回归

- 预测及作图

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

➢方法二：多元线性回归

- 输入：(regress_eg_2.m)

```
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
```

```
b =
    9.1329
   65.8896
  489.2946
stats =
  1.0e+007 *
   0.0000
   1.0378
        0
   0.0000
```

- 回归模型：$s = 489.2946t^2+65.8896t+9.1329$

➢ 确定回归系数的命令：

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

- beta：估计出的回归系数
- r：残差
- J：Jacobian 矩阵
- x, y：输入数据x, y分别为n×m矩阵和n维列向量，对一元非线性回归，x为n维列向量。
- 'model'：事先用m-文件定义的非线性函数
- beta0：回归系数的初值

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

➢ 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1-alpha的置信区间[Y-DELTA, Y+DELTA] 。

➢ **例** 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		

➢方法一：化为线性回归

$$y = ae^{\frac{b}{x}} \Rightarrow \ln y = \ln a + \frac{b}{x}$$

$$v = \ln y,\quad u = \frac{1}{x} \Rightarrow v = \ln a + bu$$

➢方法一：化为线性回归

- 输入：(regress_steel_1.m)

```
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];
u=1./x;   X=[ones(15,1) u'];
v=log(y); Y=v';
[b,bint,r,rint,stats]=regress(Y,X);
b,bint,stats
a=exp(b(1)), b(2)
```

- 回归模型：$y = 11.6791e^{-1.1107/x}$

➢ 方法二：非线性回归

- 对将要拟合的非线性模型$y=ae^{b/x}$，建立m-文件如下：

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

- 输入：(regress_steel_2.m)

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

- 回归模型：$y = 11.6037e^{-1.0641/x}$

➢ 方法二：非线性回归

- 预测及作图

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