数模加油站

数学建模模型算法精讲课——

一元线性回归分析

—— 江北老师

成名每在穷苦日，
败事多因得意时

一元线性回归

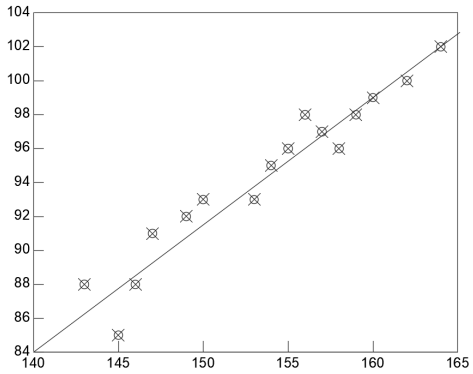
- MATLAB算法介绍
- MATLAB具体代码

一元线性回归——MATLAB算法介绍

➢ **某团队测了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

- 上节课我们已经介绍了一元线性回归方程的最小二乘法求解、误差、置信/预测区间、拟合优度、显著性检验的概念和解法



- 通过计算我们得到了
- $\hat{y} = 0.7194x - 16.08$
- $x = 170$, $\hat{y} = 106.218$
- 置信区间：$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}} = 106.218 \pm 2.007$
- 预测区间：$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}} = 106.218 \pm 3.472$
- 且模型通过了拟合度和显著性检验
- 那么如何使用MATLAB进行计算呢？

一元线性回归——MATLAB算法介绍

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

- 输入变量

  X、Y——自变量和因变量的样本值

  $alpha$——显著性水平，默认为0.05

- 输出变量:

  b——回归系数

  $bint$——回归系数的区间估计

  r——残差

  $rint$——置信区间

  $stats$——用于检验回归模型的统计量

  $stas$有四个数值：决定系数R^2、F值、与F对应的概率P、无偏估计σ^2

➢ regress命令：回归系数

$$y = \beta_0 + \beta_1 x \quad \Longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \cdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \Longleftrightarrow \quad Y = X\beta$$

- $b = regress(Y, X)$

$$b = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_p \end{bmatrix} \qquad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & \cdots \\ 1 & x_n \end{bmatrix}$$

- 具体代码:

$x = [143\ 145\ 146\ 147\ 149\ 150\ 153\ 154\ 155\ 156\ 157\ 158\ 159\ 160\ 162\ 164]';$
$X = [ones(16,1)\ \ x];$
$Y = [88\ 85\ 88\ 91\ 92\ 93\ 93\ 95\ 96\ 98\ 97\ 96\ 98\ 99\ 100\ 102]';$
$[b] = regress(Y, X)$

- 得：$b = -16.0730\ \ 0.7194$，即$\hat{\beta}_0$=-16.0730 ， $\hat{\beta}_1$=0.7194

➢ regress命令：回归系数的置信区间

- β_0和β_1置信水平为$1-\alpha$的置信区间分别为

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

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

- $[b\,,\ bint] = regress(Y\,,X)$
- 得：$b = -16.0730 \quad 0.7194$

$bint = -33.7071 \quad 1.5612$

$\quad\quad\quad\quad 0.6047 \quad 0.8340$

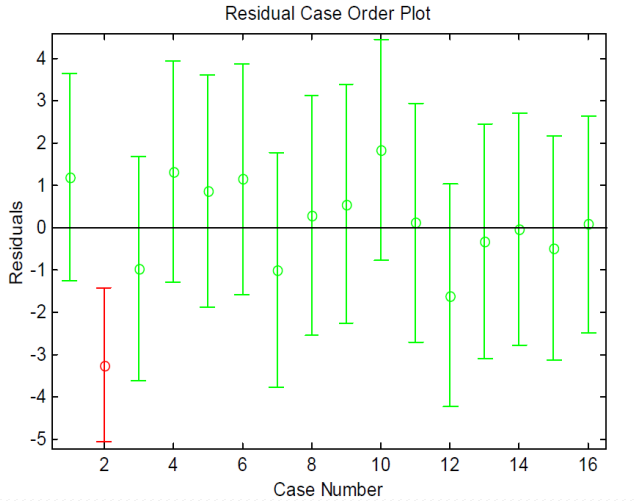
- β_0和β_1置信水平为$1-\alpha$的置信区间分别为$[-33.7071\,,1.5612]$和$[0.6047\,,0.8340]$

➢ regress命令：残差分析及置信区间

- $r_i = Y_i - \hat{Y}_i$ ⟹ $r = \begin{bmatrix} r_1 \\ r_2 \\ \cdots \\ r_n \end{bmatrix}$
- 置信区间：$\hat{y}_0 \pm t_{\frac{\alpha}{2}} \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}$
- $[b,\ bint,r,rint] = regress(Y,X)$
- 得：$r = 1.2056, -3.2331, \cdots, -0.4621, 0.0992$

$$rint = \begin{matrix} -1.2585 & 3.6697 \\ -5.0755 & -1.3907 \\ \cdots & \cdots \\ -2.4826 & 2.6810 \end{matrix}$$

- 残差图作图命令：$rcoplot(r,rint)$



- **regress命令：检验回归模型统计量**
  - $[b,\ bint,r,rint,stats] = regress(Y,X)$
  - 得：$stats = 0.9282$(决定系数) 180.9531(F值) 0.0000（F对应的概率p） 1.7437无偏估计σ^2
- 拟合优度
  - ✓ 总体平方和（***TSS***）：$TSS = \sum y_i^2 = \sum (y_i - \bar{y}_i)^2$
  - ✓ 回归平方和（**ESS**）：$ESS = \sum \hat{y}_i^2 = \sum (\hat{y}_i - \bar{y}_i)^2$
  - ✓ 残差平方和（***RSS***）：$RSS = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$
  - $TSS = ESS + RSS$
  - 拟合优度统计量：$R^2 = \frac{\mathrm{ESS}}{TSS} = 1 - \frac{\mathrm{RSS}}{TSS} = \frac{(\hat{y}_i - \bar{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} = 0.9282$
  - R^2判定系数/决定系数，取值范围为$[0,1]$，其越接近1，实际观测点离样本线越近，模型越好

一元线性回归——MATLAB算法介绍

- F统计量
  - 对回归方程$Y = \beta_0 + \beta_1 x$的显著性检验，归结为对下述假设进行检验

    $H_0 : \beta_1 = 0 ; H_1 : \beta_1 \neq 0$
  - 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，所得的回归方程也无意义.
  - 当H_0 成立时$F = \frac{ESS/1}{RSS/(n-2)} = \frac{MSR}{MSE} \sim F(1, n-2)$

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

    $F = 180.9531 > 4.6$，所以线性关系显著
- F统计量对应p值
  - $p = P(F(1, n-2) > F \mid H_0$成立$)$
  - p就是接受回归模型的风险，即犯错的概率
  - 本题$p = 0$，所以接受回归模型没有风险

- σ^2的无偏估计
  - σ^2的无偏估计

$$\hat{\sigma}_e^2 = \frac{RSS}{n-2}$$

  - RSS为残差平方和，$\hat{\sigma}_e^2$称为剩余方差(残差的方差)，与$\hat{\beta}_0$、$\hat{\beta}_1$独立，即是上节课说的均方残差（MSE）
- 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在置信水平为$1-\alpha$的预测区间为

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

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

一元线性回归——MATLAB具体代码

完整代码

```
% 1、输入数据
%输入X的样本值
x=[143 145 146 147 149 150 153 154 155 156 157 158 159 160 162 164]';
%插入\beta0对应列
X=[ones(16,1) x];
%输入Y的样本值
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);
%输出我们需要的数据
b,bint,stats
% 3、残差分析，做残差图
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')
```

欢迎关注数模加油站

THANKS