数学建模模型算法精讲课——

多元线性回归分析

—— 江北老师

成名每在穷苦日，
败事多因得意时

多元线性回归

- 算法介绍
- 典型例题
- 逐步回归例题
- 具体代码


多元线性回归——算法介绍

➢ 多元线性回归模型

- 一般称由$y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$确定的模型：

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

为k元**线性回归模型**，并简记为$(Y\,,X\beta\,,\sigma^2I_n)$

- $Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}\,,X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}\,,\beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}\,,\epsilon=\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_n \end{bmatrix}$
- $y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$称为**回归平面方程**

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

- 对参数β和σ^2作点估计，建立y与$x_1\,,x_2\,,\cdots,x_k$之间的数量关系
- 对模型参数、模型结果等做检验
- 对y的值作预测，即对y作点（区间）估计

➢ 多元线性回归模型的参数估计

- 用**最小二乘法**求$\beta_0,\cdots,\beta_k$的估计量：作离差平方和

$$Q=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_kx_{ik})^2$$

- 选择$\beta_0,\cdots,\beta_k$使Q达到最小
- 解得估计值$\hat{\beta}=(X^TX)^{-1}(X^TY)$
- 得到的$\hat{\beta}_i$代入回归平面方程得：

$$y=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_kx_k$$

- 称为经验回归平面方程，$\hat{\beta}_i$称为经验回归系数

➢ 多元线性回归模型和回归系数的检验

1）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)

2）r检验法

- 定义

$$R=\sqrt{\frac{U}{L_{yy}}}=\sqrt{\frac{U}{U+Q_e}}$$

- 称为y与$x_1,\cdots,x_k$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

➢ 多元线性回归模型的预测

1）点预测

- 求出回归方程$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$，对于给定自变量的值$x_1^*, \cdots, x_k^*$，用$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*$来预测$y^* = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* + \epsilon$，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

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

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

多元线性回归——算法介绍

- **MATLAB实现：** $[b, bint, r, rint, stats] = regress(Y, X, alpha)$
  - 参数含义：
    - 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
    - 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
    - X、Y——因变量和自变量的样本值
    - $alpha$——显著性水平，默认为0.05
  - $b = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdots \\ \hat{\beta}_p \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$
- **画出残差及其置信区间：** $rcoplot(r, rint)$

➢ 例题1：某建筑材料公司的销售量因素分析

- 某建材公司对某年20个地区的建材销售量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进行了统计。试分析推销开支、实际帐目数、同类商品竞争数和地区销售潜力对建材销售量的影响作用。试建立回归模型，且分析哪些是主要的影响因素。
- 设：
  - ✓ 推销开支——x_1
  - ✓ 实际账目数——x_2
  - ✓ 同类商品竞争数——x_3
  - ✓ 地区销售潜力——x_4

序号	推销开支	实际账目数	同类商品竞争数	地区销售潜力	建材销售量
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

多元线性回归——典型例题

序号	推销开支	实际账目数	同类商品竞争数	地区销售潜力	建材销售量
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

➢ MATLAB求解

- 输入：

$x1 = [5.5\ 2.5\ 8\ 3\ \ldots\ldots 8\ 6\ 4\ 7.5\ 7]'$; (20维)

$x2 = [31\ 55\ 67\ \ldots\ldots\ 55\ 70\ 40\ 50\ 62\ 59]'$;

$x3 = [10\ 8\ 12\ \ldots\ldots\ 11\ 11\ 9\ 9]'$;

$x4 = [8\ 6\ 9\ 16\ \ldots\ldots\ 8\ 13\ 11]'$;

$y = [79.3\ 200.1\ \ldots\ldots\ 135.8\ 223.3\ 195]'$;

$X = [ones(size(x1)), x1, x2, x3, x4]$;

$[b, bint, r, rint, stats] = regress(y, X)$

如何分析四个因素x_1, x_2, x_3, x_4对试验指标Y的作用，都是必要因素吗？

- 计算结果（输出）：

$b = \ 191.9158 \quad -0.7719 \quad 3.1725 \quad -19.6811 \quad -0.4501$

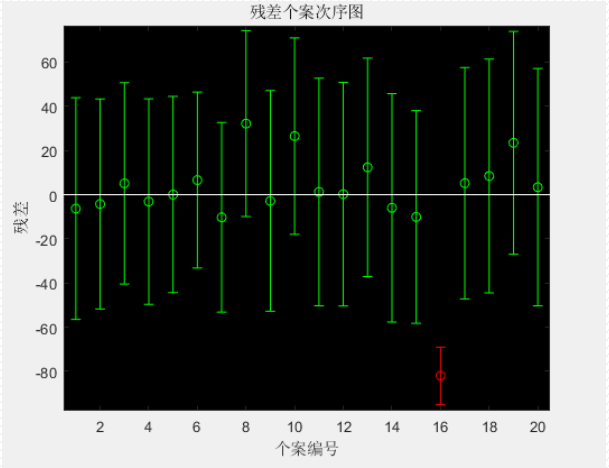
$bint = \ 103.1071 \quad 280.7245$ ……（系数的置信区间）

$r = [\ -6.3045 \quad -4.2215\ \ldots\ldots 8.4422 \quad 23.4625 \quad 3.3938]$

$rint =$ (略)

$stats = \ 0.9034(R^2) \quad 35.0509(F) \quad 0.0000(p)$

644.6510（σ的无偏估计，$r' * r/(n-5)$）



多元线性回归——逐步回归

➢ 多元线性回归的逐步回归

- “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
- 逐步回归分析法的思想:
  - ✓ 从一个自变量开始，根据自变量对Y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引入变量而变得不显著时，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。

➢ 逐步回归的MATLAB实现

- $stepwise(x, y, inmodel, alpha)$

  x——自变量数据，$n \times m$阶矩阵　　y——因变量数据，$n \times 1$阶矩阵

  $inmodel$——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）

  $alpha$——默认为0.05

多元线性回归——逐步回归

➢ **例题2：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：

$x1 = [7\ 1\ 11\ 11\ 7\ 11\ 3\ 1\ 2\ 21\ 1\ 11\ 10]';$

$x2 = [26\ 29\ 56\ 31\ 52\ 55\ 71\ 31\ 54\ 47\ 40\ 66\ 68]';$

$x3 = [6\ 15\ 8\ 8\ 6\ 9\ 17\ 22\ 18\ 4\ 23\ 9\ 8]';$

$x4 = [60\ 52\ 20\ 47\ 33\ 22\ 6\ 44\ 22\ 26\ 34\ 12\ 12]';$

$y = [78.5\ 74.3\ 104.3\ 87.6\ 95.9\ 109.2\ 102.7\ 72.5\ 93.1\ 115.9\ 83.8\ 113.3\ 109.4]';$

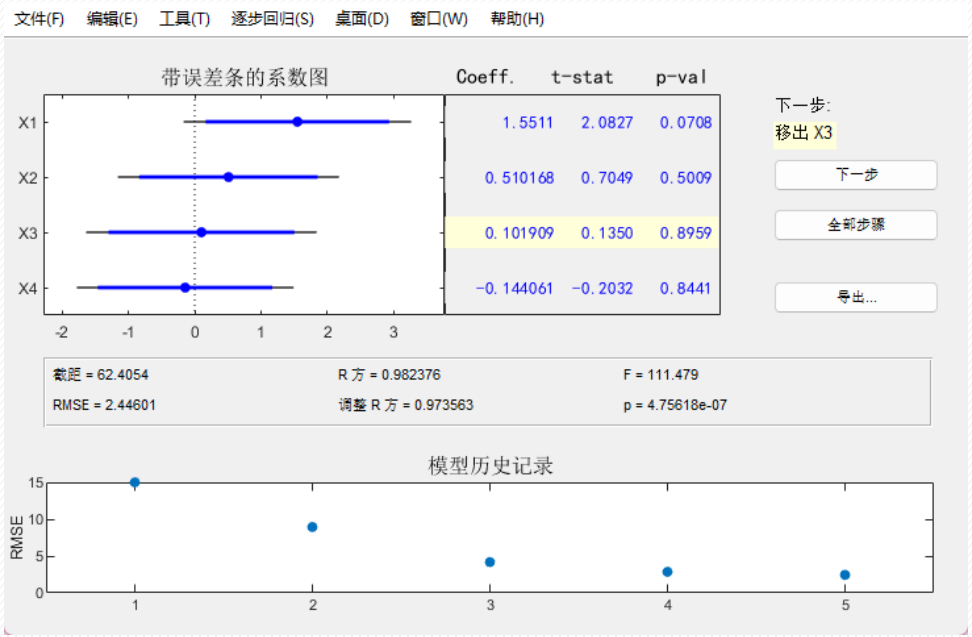
$x = [x1\ x2\ x3\ x4];$

多元线性回归——逐步回归

- 2、逐步回归：

  $stepwise(x, y)$

1）先在初始模型中取全部自变量：

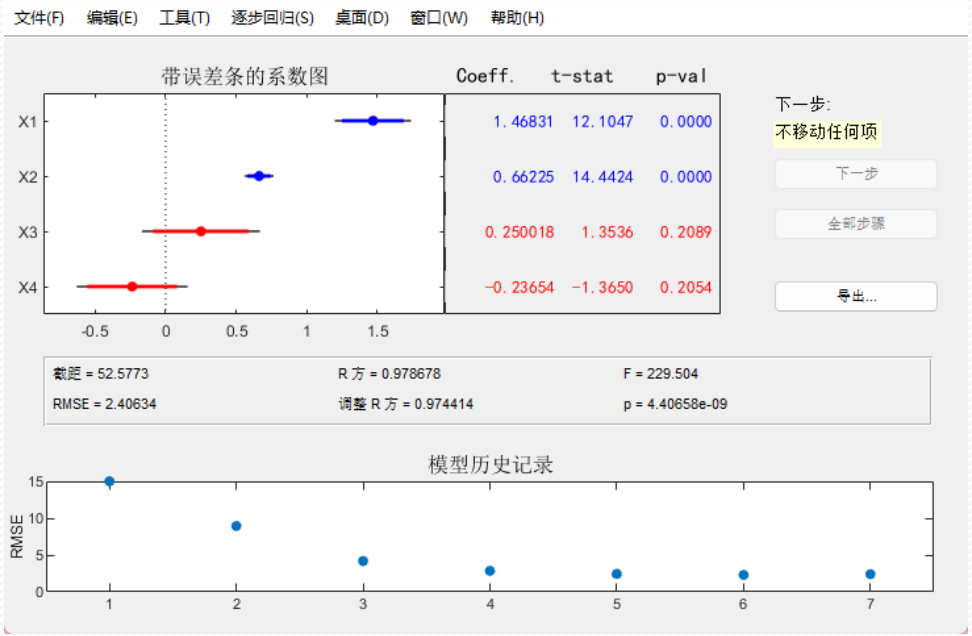


变量x_3和x_4对应的线段中心最靠近0点的显著性最差！！

多元线性回归——逐步回归

- 2、逐步回归：

  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

多元线性回归——逐步回归

- 2、逐步回归:

  3）对变量y和x_1、x_2作线性回归

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

  $b = regress(y, X)$

  得到结果:

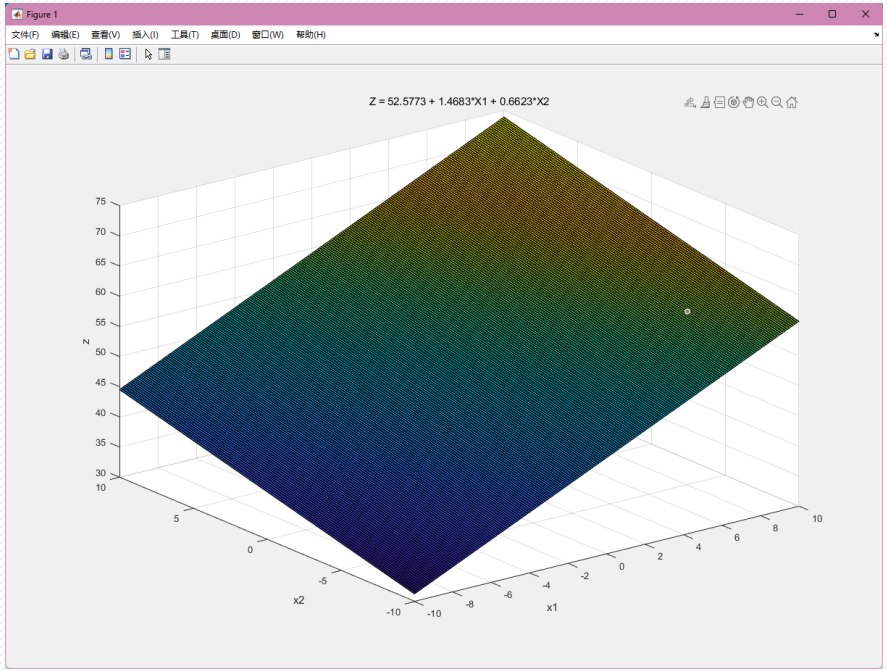
  $b = 52.5773$

  1.4683

  0.6623

- 故最终模型为:

  $y = 52.5773 + 1.4683x_1 + 0.6623x_2$



多元线性回归——具体代码

➢ **MATLAB代码**

```
% 1、某建筑材料公司的销售量因素分析
% （1）数据输入
x1=[5.5 2.5 8.0 3.0 3.0 2.9 8.0 9.0 4.0 6.5 5.5 5.0 6.0 5.0 3.5 8.0 6.0 4.0 7.5
7.0]';
x2=[31 55 67 50 38 71 30 56 42 73 60 44 50 39 55 70 40 50 62 59]';
x3=[10 8 12 7 8 12 12 5 8 5 11 12 6 10 10 6 11 11 9 9]';
x4=[8 6 9 16 15 17 8 10 4 16 7 12 6 4 4 14 6 8 13 11]';
y=[79.3 200.1 163.2 200.1 146.0 177.7 30.9 291.9 160.0 339.4 159.6 86.3 237.5 107.2
155.0 201.4 100.2 135.8 223.3 195.0]';
X=[ones(size(x1)),x1,x2,x3,x4]; % 自变量矩阵，包括常数项和四个自变量
% （2）求结果
[b,bint,r,rint,stats]=regress(y,X) % 进行多元线性回归分析
% （3）画残差图
rcoplot(r,rint) % 绘制残差图，用于评估回归模型的拟合情况
```

多元线性回归——具体代码

- **MATLAB代码**

```
% 2、水泥凝固时放热分析
% （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];  % 自变量矩阵，包括四个自变量
% （2）逐步回归
stepwise(x,y)  % 使用逐步回归分析方法，选取最佳模型
% 先在初始模型中取全部自变量，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
% (3) 线性回归
X=[ones(13 ,1)  x1 x2];
b=regress(y ,X)
```

多元线性回归——具体代码

➢ **MATLAB代码**

```
% 3、作图
% 定义自变量的范围
x1 = -10:0.1:10; % x1的取值范围
x2 = -10:0.1:10; % x2的取值范围

% 计算对应的因变量值
[X1, X2] = meshgrid(x1, x2); % 创建网格点坐标矩阵
Z = 52.5773 + 1.4683*X1 + 0.6623*X2; % 计算对应的因变量值

% 绘制三维图形
figure;
surf(X1, X2, Z);
xlabel('x1');
ylabel('x2');
zlabel('z');
title('Z = 52.5773 + 1.4683*X1 + 0.6623*X2');
```

欢迎关注数模加油站

THANKS