《系统工程导论》

第四章作业 2

多元线性回归和病态问题

姓名：卢志

班级：自 43 班

学号：2014011497

2016 年 4 月 4 日

1. 题目一

试说明：病态线性回归问题中，显著性检验是否需要？如果需要，是在自变量降维去线性之前，还是之后，还是前后都检验？给出理由证明你的结论。

解答：病态线性回归问题中，显著性检验是需要的。

因为显著性检验描述的是最终拟合回归线性方程与实验数据分布的线性程度。而无论问题是否为病态，其都存在这些数据“是否是线性”这样的问题。所以说显著性检验是一个较为重要的指标，需要通过假设检验来评估其置信度

显著性检验应该在自变量降维去线性之后进行。因为对于病态问题，在降维之前，XX^T接近奇异矩阵，所以其逆的特征值较大，从而导致剩余平方和较大，检验数 F 的值较小，可能会接受原假设，认为变量是非线性的。其实病态问题产生的本质就是存在了一些冗余的变量，如果把这些变量去除之后，其他变量本身的线性度较好，能够通过显著性检验。所以应该在降维之后（把这些冗余变量去除之后）进行显著性检验。

而且以 F 检验为例，其检验要求是各变量之间是线性无关的，才能使得变量 F 满足自由度为(n, N−n−1)的 F 分布。在降维前不满足条件，无法进行检验。

2. 题目二

2.1 题目描述

在前一次作业的基础上，编程实现多元线性回归。要求：

1）实现函数 function linear_regression(Y,X,alpha);

2）输入列向量因变量 Y，自变量 X;显著性水平 alpha；

3）能够自适应地进行多元、病态回归（特征值阈值自定）；

4）打印出显著性检验结果、回归直线方程和置信区间；

5）用你编写的程序处理下述回归问题。显著性水平取 0.05。

观测号	x1	x2	x3	x4	y
1	149. 3	4. 2	80. 3	108. 1	15. 9
2	161. 2	4. 1	72. 9	114. 8	16. 4
3	171. 5	3. 1	45. 6	123. 2	19. 0
4	175. 5	3. 1	50. 2	126. 9	19. 1
5	180. 8	1. 1	68. 8	132. 0	18. 88
6	190. 7	2. 2	88. 5	137. 7	20. 4
7	202. 1	2. 1	87. 0	146. 0	22. 7
8	212. 4	5. 6	96. 9	154. 1	26. 5
9	226. 1	5. 0	84. 9	162. 3	28. 1
10	231. 9	5. 1	60. 7	164. 3	27. 6
11	239. 0	0. 7	70. 4	167. 6	26. 3

2.2 最终结果

在作业中，代入数据，取显著性水平为 0.05，特征值阈值为 0.1，可得以下结果：

```
命令行窗口

  此问题是病态线性回归问题！由4维降至3维
  显著性水平 α = 0.050 条件满足
  回归方程为 y = -9.15145 + 0.07297*x1 + 0.59856*x2 + 0.00187*x3 + 0.10548*x4
  计算所得F=125.43203, Fa=4.53368, 由于F>Fa，否定原假设，认为存在线性关系。
  置信区间为 (y-1.24832 , y+1.24832)
fx >>
```

2.3 操作步骤

① 规范化自变量、因变量

按照定义，样本的均值和方差计算如下：

样本均值：$e(x_i)=\frac{1}{N}\sum_{i=1}^{N}x_i(t)$

样本方差：$\delta^2(x_i)=\frac{1}{N-1}\sum_{i=1}^{N}(x_i(t)-e(x_i))^2$

对应的归一化后数据$\overline{x}_i(t)=\frac{x_i(t)-e(x_i)}{\sqrt{\delta^2(x_i)}}$，同理$\overline{y}(t)=\frac{y(t)-e(y)}{\sqrt{\delta^2(y)}}$

② 判断问题是否病态

将上一步求取的规范化数据整理为(其中,输入 X 为 n 维向量,有 N 组数据):

$$Y=\left[y(1)...y(N)\right]_{1\times N},X=\left[x(1)...x(N)\right]_{n\times N}$$

对矩阵进行分解：$XX^T=Q\Lambda Q^T,XY^T$,其中$\Lambda=diag(\lambda_1,\lambda_2...\lambda_n)$，即为$XX^T$全部的特征值，将特征值按照从大到小的顺寻排序。

设定阈值ε，当$\frac{\sum_{i=m+1}^{n}\lambda_i}{\sum_{i=1}^{n}\lambda_i}\geq\varepsilon$时，说明所需要保留的特征值只有(n-m)个，即为$\lambda_{m+1}...\lambda_n$。若 m>0,则说明问题是病态回归问题,需要降维。本题要求阈值ε自定，本次作业设置ε=0.1。

③ 降维处理

若前述步骤说明需要降维，降维处理选择较小的 n-m 个特征值:

$$\lambda_{m+1}...\lambda_n$$

设XX^T从大到小特征值所对应的特征向量为:

$$Q=\left[q(1),q(2)...q(n)\right]$$

选出较大 m 个特征值对应特征向量组成:

$$Q_m=[q(1),q(2),\dots,q(m)]$$

则

$$(ZZ^T)^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & 1/\lambda_m \end{bmatrix}$$

④ 病态回归方程求解

直接利用下述公式计算，其中 $\overline{y}, \overline{x}$ 都是经过了归一化处理后的结果：

$$\hat{d} = (ZZ^T)^{-1} ZY^T = \Lambda_m^{-1} Q_m^T XY^T$$

$$\hat{c} = Q_m \hat{d}$$

$$\overline{y} = \hat{c}^T \overline{x}$$

由上式可得到病态回归方程。

⑤ 反归一化

再将求得的回归参数代入原来的方程：

$$\frac{y(t) - e(y)}{\sqrt{\delta^2(y)}} = \sum \hat{c} \frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}}$$

$$\hat{b} = \delta(y) \sum_{i=1}^{n} \frac{c_i}{\delta x_i}, a = e(y) - \delta(y) \sum_{i=1}^{n} \frac{e(x_i)}{\delta(x_i)}$$

即可求得最终参数。

⑥ 显著性检验

利用 F 检验法进行统计性检验，根据讲义，统计量

$$F = \frac{ESS/f_E}{RSS/f_R}$$

服从自由度为(n,N-n-1)的 F 分布，对于给定的显著性水平，可查 F 分布表，做假设检验。可利用 finv 函数，用法如下：

x=finv(p,v1,v2)。分子自由度为 v1，分母自由度为 v2，求置信度（显著性水

平）为 1-p 的情况下得到的 F 的值即为 x。此时分子为解释平方和，自由度 1；分母为剩余平方和，自由度为 N-2（N 为数据总数）。

⑦ 预测精度（求取置信区间）

F检验的结果为x与y存在线性关系，因此求取置信区间。S_δ 为y的剩余均方差，表示变量y偏离回归直线的误差：

$$S_\delta = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N-2}} = \sqrt{\frac{RSS}{f_R}}$$

S_δ服从正态分布，可用norminv函数求取置信区间。norminv函数用法如下：

x=norminv(p,u,sigma)。norminv是正态分布的反函数，已知概率求百分位点（方差前的系数）。P为概率，u为正态分布的均值，sigma为正态分布的方差。由于本次中使用的是标准正态分布，故u=0，sigma=1。语句为：

z=norminv(1-alpha/2,0,1)%即得到正数的z进行置信区间计算。

2.4 原始代码

主函数 HW4.m

```
%%
%清除缓存区
clear all;
close all;

%%
%输入原始数据
N=11;n=4;
Y = zeros(N,1);
X = zeros(N,n)
%第一列为Y，后n列为X
Y(:,1) = [15.9 ,16.4 ,19.0 ,19.1 ,18.88, 20.4 ,22.7 ,26.5 ,28.1 ,27.6 ,26.3 ];
X(:,1) = [149.3,161.2,171.5,175.5,180.8,190.7,202.1,212.4,226.1,231.9,239.0];
X(:,2) = [4.2,4.1,3.1,3.1,1.1,2.2,2.1,5.6,5.0,5.1,0.7];
X(:,3) = [80.3,72.9,45.6,50.2,68.8,88.5,87.0,96.9,84.9,60.7,70.4];
X(:,4) = [108.1,114.8,123.2,126.9,132.0,137.7,146.0,154.1,162.3,164.3,167.6];
```

```
%%
%调用自己编写的病态多元线性回归拟合函数
%取显著性水平0.05
%取特征值阈值为0.1
alpha=0.05;
threshold=0.1;
linear_regression(Y,X,alpha,threshold);
```

多元病态线性回归求解函数 linear_regression.m

```
function linear_regression(Y,X,alpha,threshold)
%自己编写的病态多元线性回归拟合函数

%%
%PART1: 规范化处理
[N,n]=size(X);%读取数据组数和变量数
old_X = X;old_Y = Y;
aver_x = mean(X,1);
aver_y = mean(Y);
sigma_x = sqrt(var(X,1));
sigma_y = sqrt(var(Y));
for i = 1:N
    for j = 1:n
        X(i,j) = (old_X(i,j) - aver_x(1,j)) / sigma_x(1,j);%规范化处理公式
    end
    Y(i,1) = (old_Y(i,1) - aver_y) / sigma_y;
end
%%
%PART2: 判定是否为病态问题
A = X' * X;
B = X' * Y;
[V,D] = eig(A);%进行分解，特征值由小到大排序
%取threshold为阈值
%保留和为特征值总和1-threshold的特征值
sum_D = 0;%特征值总和
for i = 1:n
    sum_D = sum_D + D(i,i);
end
temp_sum_D = 0;%从大到小前i项特征值之和
for i = n:-1:1
    temp_sum_D = temp_sum_D + D(i,i);
    if temp_sum_D / sum_D >= (1-threshold) %达到阈值，说明是病态问题
        break;
    end
```

```
end
%%
%PART3: 降维处理
%降至的维度m计算
m=n-i+1;
if m<n %发生了降维
    fprintf('此问题是病态线性回归问题！由%d维降至%d维\n',n,m);
end
inv_ZZ=zeros(m,m);
Qm=zeros(n,m);
for i = 1:m
    inv_ZZ(i,i) = 1/D(n+1-i,n+1-i);
    Qm(1:n,i:i) = V(1:n,n+1-i:n+1-i);%Qm即由特征向量构成
end
%%
%PART4: 病态回归方程参数求解
d=inv_ZZ * Qm' * B;
c=Qm * d;
%%
%PART5: 反归一化
real = zeros(1,n+1);
real(1,n+1) = aver_y;
for i = 1:n
    real(1,i) = c(i,1) / sigma_x(1,i) * sigma_y;
    real(1,n+1) = real(1,n+1) - c(i,1) * aver_x(1,i) * sigma_y / sigma_x(1,i);
end
%%
%PART6: 进行显著性检验
ESS = 0;RSS = 0;
esti_y=zeros(N,1);%y的拟合值
for i = 1:N
    esti_y(i,1) = esti_y(i,1) + real(1,n+1); %计算y的拟合值
    for j=1:n
        esti_y(i,1) = esti_y(i,1) + real(1,j)*old_X(i,j);%计算y的拟合值
    end
    ESS = ESS + (esti_y(i,1) - aver_y)^2; %解释平方和
    RSS = RSS + (old_Y(i,1) - esti_y(i,1))^2; %剩余平方和
end
F = (ESS/n) / (RSS/(N-n-1)); %F服从F分布
%x=finv(p,v1,v2)。分子自由度为v1，分母自由度为v2
%求置信度（显著性水平）为1-p的情况下得到的F的值即为x
Fa = finv(1 - alpha,n,N-n-1);
%检验显著性水平，若拒绝H0，则输出图像
if F > Fa
```

```
    fprintf('显著性水平 α = %.3f 条件满足\n',alpha);
    %%
    %PART7：预测精度
    S = sqrt(RSS/(N-n-1)); %S服从正态分布
    Z_half_alpha = norminv(1 - alpha/2,0,1);%置信半区间
    fprintf('回归方程为 y = %.5f',real(1,n+1));
    for i = 1:n
        fprintf(' + %.5f*x%d',real(1,i),i);
    end
    fprintf('\n计算所得F=%.5f, Fa=%.5f, 由于F>Fa, 否定原假设, 认为存在线性关
系。',F,Fa);
    fprintf('\n置信区间为 (y-%.5f , y+%.5f)\n',S*Z_half_alpha,S*Z_half_alpha);

else
   close all;
   warning = sprintf('多元线性回归曲线不满足α=%f)',alpha);
   warndlg(warning,'错误');
end
end
```