系统工程导论作业四

黑箱建模：多元线性回归

刘若涵 自 05 2020011126

1 病态线性回归显著性检验

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。

答：需要，且应该在自变量降维去线性之后进行显著性检验。显著性检验是为了验证之前所选取的总体分布假设是否成立，用检验结果来衡量真实情况与所选取的模型是否显著不同。因此病态线性回归也需要显著性检验，以验证所选取的分布形式是正确的。而自变量降维去线性之前回归分析误差很大，进行显著性检验可信度不高；应该先降维去线性以使回归分析结果可信。

2 方法原理

2.1 样本数据规范化

$$\bar{x}_i(t)=\frac{x_i(t)-e(x_i)}{\sqrt{\delta^2(x_i)}} \quad \forall i,t$$

$$\bar{y}(t)=\frac{y(t)-e(y)}{\sqrt{\delta^2(y)}} \quad \forall t$$

代码如下：

```
EX = mean(X, 1);
SX = std(X, 1, 1);
EY = mean(Y);
SY = std(Y);
X_bar = X;
for i = 1:N
    for j = 1:n
    X_bar(i,j) = (X(i,j) - EX(j))/SX(j);
    end
end
Y_bar = (Y - EY)/SY;
```

2.2 病态分析

病态分析步骤如下：

（1）计算出样本数据的 Gram 矩阵 XX^T 的特征值、特征向量并按照特征值从

大到小排列λ_1到λ_n。

（2）逐渐增加 m，计算 $\sum_{i=1}^{m}\lambda_i$，当 $\sum_{i=1}^{m}\lambda_i >= thresh$ 时停止（取定阈值 thresh = 0.99）。

（3）比较 m 与样本数据总维数 n 的大小，若 m < n 则认为样本数据存在病态，仅保留最大的 m 个特征值与其对应的特征向量。否则不存在病态。

代码如下：

```
A = X_bar'*X_bar;
B = X_bar'*Y_bar;
Lambda = eig(A);
[Q,V] = eig(A);
Lambda_sum = sum((Lambda'));
Lambda_tempt = 0;
thresh = 0.99;
for m = 1:n
    Lambda_tempt = Lambda_tempt + Lambda(n+1-m);
    if Lambda_tempt/Lambda_sum >= thresh
    break;
    end
end
```

2.3 回归方程求解

样本数据存在病态时：

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

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

$$y \approx \hat{d}^T z = \hat{d}^T Q_m^T x = \hat{c}^T x$$

去归一化：

$$\widehat{\beta}_i = \widehat{c}_i \cdot \frac{\sqrt{\delta^2(y)}}{\sqrt{\delta^2(x_i)}}$$

$$\widehat{\beta}_0 = e(y) - \sum_{i=1}^{n} \widehat{\beta}_i \, e(x_i)$$

样本数据不存在病态时：

$$\widehat{\beta} = (XX^T)^{-1} XY^T$$

回归方程为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

代码如下：

```
if m < n
    fprintf('为病态线性回归问题, m=%d\n',m);
    Vm = zeros(m,m);
    Qm = zeros(n,m);
    for i = 1:m
        Vm(i,i) = 1/V(n+1-i,n+1-i);
        Qm(1:n,i:i) = Q(1:n,n+1-i:n+1-i);
    end
    d = Vm * Qm' * B;
    c = Qm * d;
    beta = zeros(1,n+1);
    beta(1,n+1) = EY;
    for i = 1:n
        beta(1,i) = c(i,1)/SX(i) * SY;
        beta(1,n+1) = beta(1,n+1) - beta(1,i) * EX(i);
    end
else
    fprintf('为非病态线性回归问题');
    beta = (X' * X) / (X' * Y);
end
fprintf('回归方程为 y = %f',beta(1,n+1));
for i = 1:n
    fprintf('+%f*x%d',beta(1,i),i);
end
```

2.4 显著性检验

$$F = \frac{ESS / f_E}{RSS / f_R} = \frac{(N-n-1)\cdot ESS}{n \cdot RSS}$$

F_α 通过 `finv(1 - alpha, m, N-m-1)` 求得。

当 $F > F_\alpha$时，否定原假设H_0，认为 x 与 y 存在线性关系；

当 $F \le F_\alpha$时，接收原假设H_0，认为 x 与 y 不存在线性关系。

代码如下：

```
Y_pred = beta(1:n) * X' + beta(1,n+1);
TSS = sum((Y - EY).^2);
ESS = sum((Y_pred - mean(Y_pred)).^2);
RSS = TSS - ESS;
F = (ESS/m)/(RSS/(N-m-1));
Fa = finv(1 - alpha, m, N-m-1);
if(F>Fa)
```

```
    fprintf("F>Fα, x 与 y 存在线性关系\n");
else
    fprintf("F<=Fα, x 与 y 不存在线性关系\n");
end
```

2.5 精度分析

剩余均方差：

$$S_\sigma = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N-2}} = \sqrt{\frac{(1-r^2)L_{yy}}{N-2}}$$

$Z_{\alpha/2}$ 通过 norminv(1 - alpha/2, 0, 1) 求得。

置信区间为：

$$(\hat{y}_0 - Z_{\alpha/2}S_\delta, \hat{y}_0 + Z_{\alpha/2}S_\delta)$$

代码如下：

```
S = sqrt(RSS/(N-m-1));
Z = norminv(1 - alpha/2, 0, 1);
fprintf('置信区间为 [y - %f,y + %f]\n',S*Z,S*Z);
```

3 处理结果

为病态线性回归问题, m=3

Fa=4.346831, F=195.116494

F>Fa, x 与 y 存在线性关系

回归方程为 y = -9.151455+0.072966*x1+0.598562*x2+0.001872*x3+0.105482*x4

置信区间为 [y - 1.155715,y + 1.155715]