《系统工程导论》第四次作业

黑箱建模2

20170111010 杜澍洽 自71

题目1

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。

【答】病态线性回归问题中需要显著性检验。

显著性检验描述的是回归模型的有效性，而数据间“是否存在线性关系”这一问题无论病态与否都存在。因此在病态线性回归问题中仍然需要显著性检验。

显著性检验应该在自变量降维去线性之后进行。以F检验为例，要求各变量之间线性无关，对于病态问题，这需要通过降维来实现。此外降维前XX^T接近奇异矩阵，可能会导致得到“变量是间不存在线性关系”这样的检验结果，但实际上如果把病态问题中的冗余变量去除，剩余变量的线性度可能比较好，能够通过显著性检验。因此应该在自变量降维去线性之后进行显著性检验。

题目2

在前一次作业的基础上，使用python或者matlab编程实现多元线性回归

要求：

1. **实现函数linear_regression(Y, X, alpha)**
2. **输入为列向量因变量Y，自变量矩阵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. **完成作业报告，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

最终结果如下（阈值设为 0.1）：

此问题是病态线性回归问题，需要从4维降至3维
F=125.43203，Fa=4.53368，F>Fa，即存在线性关系
回归方程：y = -9.15145 + 0.07297x1 + 0.59856x2 + 0.00187x3 + 0.10548x4
置信区间为（y-1.24832，y+1.24832）

其中XX^T分解得到的Λ为：

	1	2	3	4
1	0.0236	0	0	0
2	0	8.4507	0	0
3	0	0	12.3501	0
4	0	0	0	23.1756

可见最小的特征值 0.0236 显著小于另外几个特征值并且接近 0，需要降维，和最终结果吻合。

下面介绍解答步骤：

（1） 规范化自变量、因变量

依据下面两个公式对自变量和因变量进行规范化：

$$\overline{x_i}(t) = \frac{x_i(t) - e(x_i)}{\sqrt{\delta^2(x_i)}}$$

$$\overline{y_i}(t) = \frac{y(t) - e(y)}{\sqrt{\delta^2(y)}}$$

其中$e(x_i)$和$e(y)$分别为x_i和y的均值，$\delta^2(x_i)$和$\delta^2(y)$分别为x_i和y的方差。

（2） 判断问题是否病态

将规范化后的数据整理为：

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

分解$XX^T = Q\Lambda Q^T$，其中$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$，$\lambda_i$为$XX^T$的特征值。

将特征值按从大到小的顺序排列，设定阈值ε，根据$\frac{\sum_{i=m+1}^{n}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$是否小于$\varepsilon$来确定$m$，我设定的$\varepsilon$为 0.1.

（3） 降维处理

若上一步说明了需要降维处理，则将XX^T从大到小的特征值所对应的特征向量也排好：

$$Q = [q(1), q(2), \dots, q(n)]$$

选出较大的m个特征值对应的特征向量组成：

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

然后得到：

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

（4） 病态回归方程求解

利用下面的公式求解相关参数：

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

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

$$y \approx \hat{c}^T x$$

（5） 还原

将求得的参数代入原方程即可得到最终参数：

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

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

（6） 显著性检验

利用下面的公式计算F：

$$F = \frac{(N-n-1)ESS}{nRSS}$$

参照上一次作业，仍然利用$finv$函数计算Fa，然后通过比较两者的大小判

断是否接受原假设。

（7） 预测精度

利用下面的公式计算S_δ：

$$S_\delta = \sqrt{\frac{RSS}{N-n-1}}$$

参照上一次作业，仍然利用$norminv$函数计算$Z_{\alpha/2}$.