系统工程第 4 次作业

张博睿　自 75　2017011537

程序详见

$$./code/main.py$$

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**需要**进行显著性检验。并且应该是在自变量降维去线性**之后**。

理由：

（1）**需要进行检验的理由：**病态线性回归可以理解成两个步骤——特征提取和线性回归。显著性检验是在正态分布误差假设下对多元线性回归系数置信度的检验，由于病态线性回归本质上需要借助一般线性回归，因此有必要在最后进行显著性检验，验证模型的有效性。

（2）**降维之后检验的理由：**由于降维之前，特征之间可能线性相关，例如对于回归方程

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha$$

如果x_1和x_2线性相关（简化为$x_1 = x_2$），那么方程退化为

$$y = (\beta_1 + \beta_2) x_1 + \beta_3 x_3 + \alpha$$

从而，仅仅能够识别$(\beta_1 + \beta_2)$，单个β_1和β_2是不能被识别的（即回归得到的结果非常不稳定），因此单独对β_1和β_2进行显著性检验是没有意义的。

综上，只有通过降维，去除变量之间的线性相关情况后，才能确保回归系数的稳定，从而进行显著性检验。

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。

解：

实现功能
（1）能够根据特征值阈值（默认为0.95）进行病态回归； （2）能够选择是否设置常数项（默认为$False$）； （3）能够根据显著性水平（默认为0.05）进行假设检验； （4）能够输出回归方程以及置信区间等。

（1）算法流程

输入：有待进行病态回归的数据。
（1）输入数据，并进行归一化处理，同时保存均值和标准差信息。 $$\tilde{y}=\frac{y-\bar{y}}{std(y)},\qquad \tilde{x_i}=\frac{x_i-\bar{x_i}}{std(x_i)}$$ （2）计算样本特征的协方差矩阵，并利用特征分解对特征值排序，根据阈值选取变换矩阵，使得 $$z=Tx$$ （3）计算y与z之间的多元线性回归系数，其中Z矩阵中的样本呈行向量排列。 $$\omega=(Z^TZ)^{-1}Z^Ty$$ （4）计算 F 值，进行显著性检验，若$F>F_0$则满足线性性。 $$F=\frac{(N-n-1)ESS}{nRSS}$$ （5）根据之前保存的均值和标准差，恢复原始的系数。
输出：病态回归的系数。

（2）病态检验部分

设置病态问题的特征值阈值为0.95，通过矩阵特征分解，计算得到原始数据的特征值为

$$23.17,\qquad 0.024,\qquad 12.35,\qquad 8.45$$

根据筛选策略，程序会去除0.024对应的特征，将原来的 4 维特征降低为 3 维。

（3）多元回归部分

通过(2)中的变换矩阵，将原始的特征变换为

$$Z=XT$$

下面计算$Y=f(Z)$的线性回归系数，计算公式为

$$\omega=(Z^TZ)^{-1}X^TY$$

（4）假设检验部分

对(3)中计算的回归系数进行显著性检验

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

由于F服从$F(n,N-n-1)$分布，所以通过查表

$$F_0=4.3468(\alpha=0.05)$$

$$F=195\gg F_0$$

因此，在显著性水平$\alpha=0.05$前提下，可以认为存在线性性。

（5）实验结果

将上述所有结果整理如下：

项目	结果
显著性水平α	0.05
显著性检验F_0	4.3468
计算F	195
显著性检验	存在线性
病态特征值阈值	0.95
降维维度	3
回归直线	$y = 0.073x_1 + 0.599x_2 + 0.002x_3 + 0.105x_4 - 9.151$
置信区间	$y \in [\hat{y} - 1.1557, \hat{y} + 1.1557]$

其中，实际结果和预测结果如下

编号	0	1	2	3	4	5
预测值	15.80921	17.31053	18.29846	18.98922	18.75159	20.77049
真实值	15.9	16.4	19	19.1	18.88	20.4
编号	6	7	8	9	10	
预测值	22.41513	26.13458	27.61757	28.2663	26.51693	
真实值	22.7	26.5	28.1	27.6	26.3	

可以看到，预测结果和真实结果在置信区间内基本一致。