《系统工程导论》第三次作业

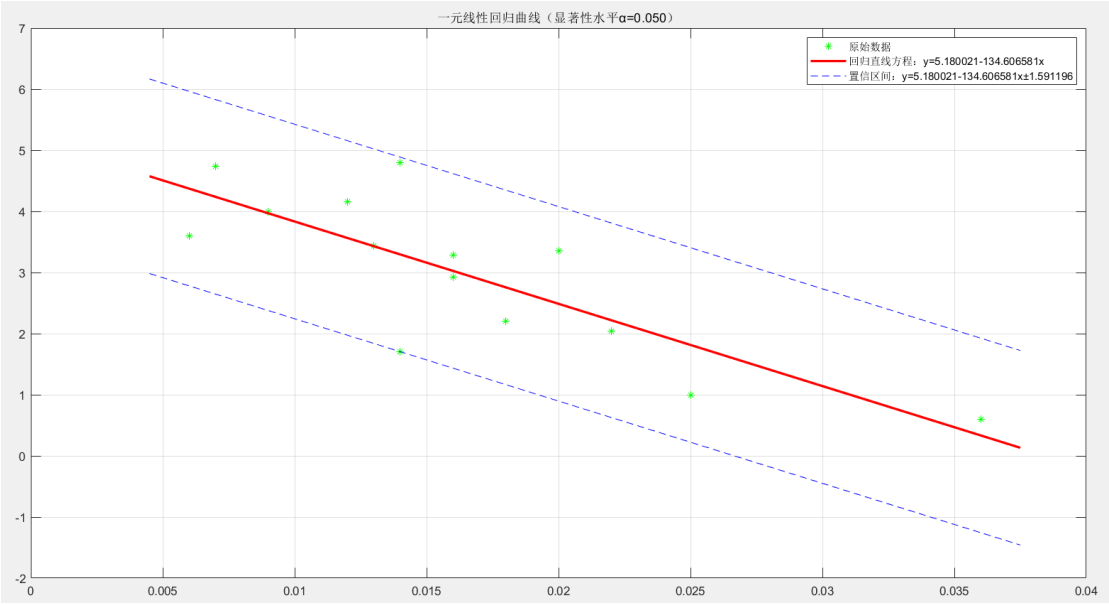
黑箱建模 1

要求：

1. 实现函数linear_regression1(data, alpha)
2. 输入为Nx2的矩阵data，第一列为Y，第二列为X；显著性水平alpha；
3. 打印出回归直线方程（也可以打印中间过程数据）
4. 用F检验进行统计检验，matlab中F分布对于给定显著性水平和自由度的分度数函数为finv，请大家自行学习使用该函数；python 请大家自己找合适的函数。输出检验结果，如果输入数据满足线性关系，那么继续做5和6，否则结束
5. 打印出置信区间，matlab 中标注正态分布相应的分位数函数是norminv，请大家自行学习使用该函数
6. 画出所有数据点、回归直线（y 为因变量，x 为自变量）和置信区间对应的两条边界线
7. 完成作业报告，显著性水平取 0.05

解答：

1. 结果图像：



由图可见，一元线性回归直线方程为

$$y = 5.18 - 134.61x$$

符合显著性水平$\alpha = 0.05$的要求，置信区间为

$$y = 5.18 - 134.61x \pm 1.59$$

2. 求解过程：

（1）原始数据如下：

编号	成分A(x)	成分B(y)	编号	成分A(x)	成分B(y)
1	0.009	4.0	8	0.014	1.7
2	0.013	3.44	9	0.016	2.92
3	0.006	3.6	10	0.014	4.8
4	0.025	1.0	11	0.016	3.28
5	0.022	2.04	12	0.012	4.16
6	0.007	4.74	13	0.020	3.35
7	0.036	0.6	14	0.018	2.2

（2）求解一元线性回归方程

原理：假设已经得到了x和y的若干数据对x_i和$y_i(i = 1,2,\dots,N)$，称为样本点，

如果x和y存在某种线性关系，则x和y可用$y = a + bx + \varepsilon$表示，其中a和b是待定系数，ε是随机变量，该模型为一元回归模型。利用最小二乘原理使目标误差平方和最小，可以得到：

$$X_i = [x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_N - \bar{x}], \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$Y_i = [y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_N - \bar{y}], \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\hat{b} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{L_{xy}}{L_{xx}}, \hat{a} = \bar{y} - \hat{b}\bar{x}$$

一元回归直线方程：

$$y = \hat{a} + \hat{b}x$$

（3）显著性检验

F检验中计算F的公式：

$$F = \frac{(N-2)ESS}{RSS}$$

对于给定的显著性水平α以及自由度$(1, N-2)$，可以用$finv(p, v1, v2)$来计算F_a，其中p取$1-\alpha$，$v1$和$v2$分别取1和$N-2$。比较F和F_a，当$F > F_a$时，否定原假设，认为x和y不存在线性关系，否则接受原假设。

（4）精度分析

通过F检验确定x和y存在线性关系，接下来求取置信区间。首先求取y的剩余均方差为：

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

再用$norminv(p, u, sigma)$求取标准正态分布上$\alpha/2$百分位点的值$Z_{\alpha/2}$，，其中p取$1-\alpha/2$，u取 0，$sigma$取 1. 这样得到：

$$L_1: y_1 = a + bx - Z_{\alpha/2}S_\delta$$

$$L_2: y_2 = a + bx + Z_{\alpha/2}S_\delta$$