系统工程导论 第三次作业

来昆 自 72 2017011607

题目 1 使用 python/matlab 编程实现一元线性回归

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。
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。

代码及结果分析:

1. 要求 1&2

   其中 data 使用.mat 存储，直接读取进行处理。data 内部及 main 函数中的部分如下:

Data.data

	1	2
1	4	0.0090
2	3.4400	0.0130
3	3.6000	0.0060
4	1	0.0250
5	2.0400	0.0220
6	4.7400	0.0070
7	0.6000	0.0360
8	1.7000	0.0140
9	2.9200	0.0160
10	4.8000	0.0140
11	3.2800	0.0160
12	4.1600	0.0120
13	3.3500	0.0200
14	2.2000	0.0180

```
Data = load('data.mat');
data = Data.data;
linear_regression1(data, 0.05);
```

   函数定义及 X 和 Y 如下:

```
function linear_regression1(data, alpha)
X = data(:, 2);
Y = data(:, 1);
```

2. 一元线性回归（要求 3）

使用最小二乘法进行处理，公式如下：

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

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

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

代码如下：

```
% Average of X & Y
x_bar = mean(X);
y_bar = mean(Y);
% L_xy
L_xy = (X-x_bar)' * (Y-y_bar);
% L_xx
L_xx = (X-x_bar)' * (X-x_bar);
% b^
b_hat = L_xy / L_xx;
% a^
a_hat = y_bar - b_hat * x_bar;
% Print to screen
disp(['Linear Regression Outcome: y = ' num2str(b_hat) ' x + ' num2str(a_hat)]);

```

对 data 进行处理时，输出的结果为：

```
>> main
Linear Regression Outcome: y = -134.6066 x + 5.18
```

3. F 检验

思路：

按相关系数分解（平方和分解）可以将总共方和 TSS 分解为解释平方和 ESS 和剩余平方和 RSS，统计量 F 的定义如下：

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

$$L_{yy}=\sum_{i=1}^{N}(y_i-\bar{y})^2=\sum_{i=1}^{N}y_i^2-N\bar{y}^2=\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

（总平方和TSS）　（解释平方和ESS）　（剩余平方和RSS）

f_E 和 f_R 分别为二者的自由度.

F 检验的思路如下：

对于给定的显著性水平$\alpha(0 \le \alpha \le 1)$，以及自由度 (1，N-2)，查 F 分布表，得到相应的临界值 F_α，从而对Ho进行假设检验，即：

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

计算 F 分布的分位数，根据课件上给出的公式，

```
%   1. F_alpha
N = size(X,1);
p = 1 - alpha;
v1 = 1;
v2 = N - 2;
F_a = finv(p,v1,v2);
```

再求 F 的值(按照上面的公式)

```
%   2. F
Y_hat = b_hat * X + a_hat;
ESS = (Y_hat - y_bar)' * (Y_hat - y_bar);
RSS = (Y - Y_hat)' * (Y - Y_hat);
F = ((N - 2)*ESS) / RSS;
```

再根据 F 和 F_a 的大小关系判断是否符合线性，输出结果。
F 分布的计算及输出结果为:

```
F=21.9609, F_a=4.7472, F>F_a, X and Y have Linear relation.
```

4. 置信区间计算

设 S_δ 为 y 的剩余均方差，它表示变量 y 偏离回归直线的误差

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

给定显著性水平 α，对某一 x_0，相应的 y_0 将以 $(1-\alpha)$ 的概率落在下述区间（称为置信区间）

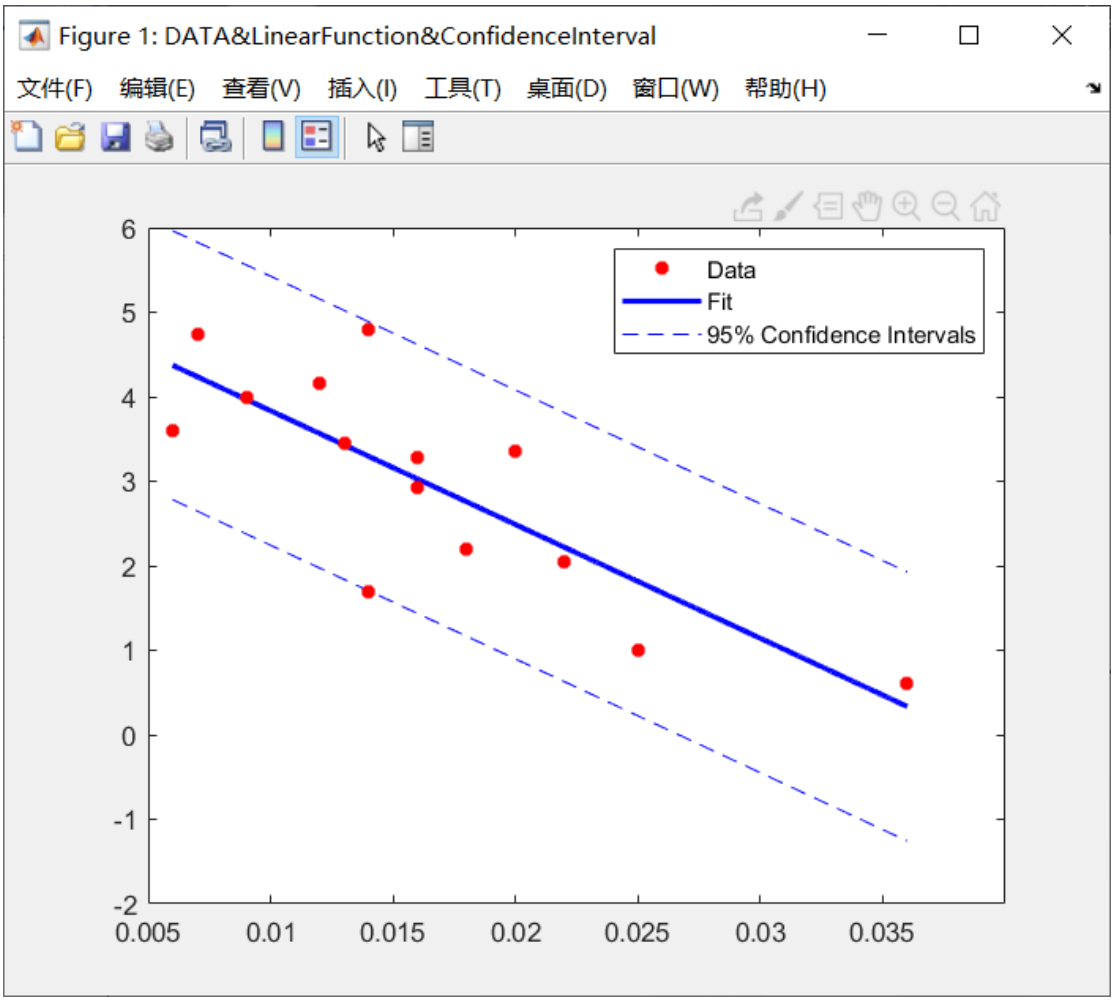
$$(\hat{y}_0 - Z_{\alpha/2}S_\delta, \hat{y}_0 + Z_{\alpha/2}S_\delta)$$

式中，$\hat{y}_0$ 是对应于 x_0 的 y_0 的预测值，$Z_{\alpha/2}$ 是标准正态分布上 $\alpha/2$ 百分位点的值。

```
Z_a = abs(norminv(alpha/2));
S_sigma = sqrt((Y-Y_hat)'*(Y-Y_hat) / (N-2));
disp(['Half width of Confindence Interval:' num2str(Z_a*S_sigma)]);
disp(['Confidence Interval: [y_hat-' num2str(Z_a*S_sigma) ',y_hat+' num2str(Z_a*S_sigma) ']']);
```

```
Half width of Confindence Interval:1.5912
Confidence Interval: [y_hat-1.5912,y_hat+1.5912]
```

5. 画出包括原始数据，拟合直线，置信区间的图像。



其中左下角的点不在置信区间内部。

代码:

```
figure('name','DATA&LinearFunction&ConfidenceInterval')
h_fit = plot(x_fit,y_fit,'b-','LineWidth',2);
hold on
h_cfivl_low = plot(x_fit,y_cfivl_low,'b--');
hold on
plot(x_fit,y_civl_up,'b--');
hold on
h_data = plot(X,Y,'r*','MarkerSize',5,'LineWidth',2);
hold on
legend([h_data,h_fit,h_cfivl_low],'Data','Fit','95% Confidence Intervals');
```