中山大學
SUN YAT-SEN UNIVERSITY

# 项目报告

SUN YAT-SEN UNIVERSITY

院（系）数据科学与计算机学院 学 号 16337184 审批

专 业 计算机类 实验人 闵正铖

实验题目：统计分析方法项目二 2018 年 11 月 5 日

实验使用工具：MATLAB2017a

## ➢ 第一题：

假设误差服从正态分布，建立个人医疗费用和 3 个定量变量之间的线性回归方程并研究相应的统计推断问题。

我们用“data.txt”中的前 1333 条数据（一共 1338 条数据）进行线性回归拟合。

用最后 5 条数据进行测试。请预测他的个人医疗费用，并给出置信度为 95%的置信区间。

**分析题目：**题目是一个多元回归分析的问题。Matlab 中，可以通过 regress 函数求得多元线性回归，并且可以给出区间估计。

首先使用 xls 格式导入 txt 文件，在 matlab 中存储相应的列向量。根据实验原理，regress 函数接收的点估计，是使用 X 矩阵、Y 矩阵和 B 矩阵进行的求解。B 矩阵的最终求解方程 X 矩阵的二次型的逆，所以只要将数据构造成下列矩阵就可以了。

(4.5) 式称为正规方程组. 为了求解的方便，将(4.5) 式写成矩阵的形式. 为此，引入矩阵

$$\boldsymbol{X}=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1p}\\ 1 & x_{21} & x_{22} & \cdots & x_{2p}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{pmatrix},\boldsymbol{Y}=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\boldsymbol{B}=\begin{pmatrix}b_0\\ b_1\\ \vdots\\ b_p\end{pmatrix}.$$

构造完之后，按照格式将参数写回。其中，b 是 B 矩阵，也就是回归系数。Bint 是区间估计的值，r 为残差，rint 是残差的置信区间，stats 是检验会规模型的统计量。

```
X = [ones(length(charge),1),age,bmi,child];
Y = charge;

[b, bint,r,rint,stats] = regress(Y,X);
```

之后，我们就能根据结果输出这些值，我们可以看到 b 的值已经求出来了，所以最终的回归方程的拟合就是：

$Y = -6873 + 237.7 X_1 + 333.7 X_2 + 546.3 X_3$

各个参数的输出如下，我们可以发现的问题是，stats 的数值，相关系数是 0，误差的方差也非常大。所以可以断定，这个回归方程并不显著。这些值并不符合线性回归。

```
%rcoplot(r,rint);
z = b(1) + b(2)*age + b(3)*bmi + b(4)*child;

Data_1344 = Regress_function(50,30.97,3);
Data_1345 = Regress_function(18,31.92,0);
Data_1346 = Regress_function(18,36.85,0);
Data_1347 = Regress_function(21,25.8,0);
Data_1348 = Regress_function(61,29.07,0);
```

```
b =            bint =

   1.0e+03 *      1.0e+04 *

   -6.8730        -1.0328   -0.3418
    0.2377         0.0194    0.0282
    0.3337         0.0233    0.0435
    0.5463         0.0038    0.1054

stats =

   1.0e+08 *

    0.0000    0.0000    0.0000    1.2959
```

接下来是利用最后五条数据进行估计验证，得到的结果分别如下：

```
Data_1344 =     Data_1345 =

   1.6986e+04      8.0573e+03
```

```
Data_1346 =     Data_1347 =     Data_1348 =

   9.7024e+03      6.7282e+03      1.7327e+04
```

而预测区间，即置信区间，按照如下公式计算。

（2）在 $x = x_0$ 处 $Y$ 的新观察值 $Y_0$ 的一个置信水平为 0.95 的预测区间为

$$\left(\hat{Y}\mid_{x=x_0} \pm t_{0.025}(8)\hat{\sigma}\sqrt{1+\frac{1}{10}+\frac{(x_0-145)^2}{8\,250}}\right).$$

因为本例的N 为 1333，所以 t 分布的值为 1.960.但是本例为多元分析，而且不符合线性回归模型，所以无法求出结果。

## ➢ 第二题：

根据上例子，利用同样的数据集（1338 条数据）：

利用方差分析知识，假设个人医疗费用服从方差分析模型，见（3.1）或（3.2）比较不同性别对个人医疗费用是否有显著（显著水平为 0.05）差异。

利用方差分析知识（两因素等重复试验下），假设个人医疗费用服从两因素的方差分析模型,见教材（3.23）请对性别、是否吸烟两个因素，对方差进行分析（显著水平为 0.05）。

1. 首先使用方差分析模型。方差分析使用 matlab 的库函数 anova1，ANOVA1 返回一个 p 的值，判断这个 p 值和概率的大小，从而判断是否具有显著性差异

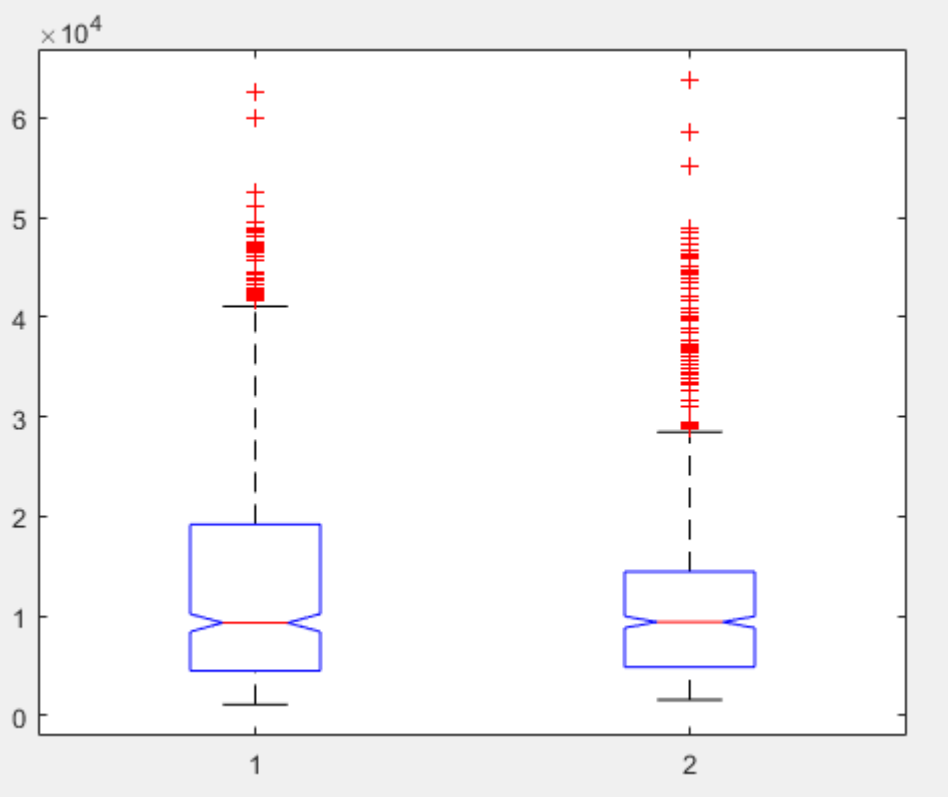
```
xlsread('data_sort.xlsx');
%数据已经排序，1-676 为男性数据，677-1333为女性数据

male = data(1:662,7);
female = data(677:1338,7);

X = [male,female];

p = anova1(X);
```

我将数据处理成如下形式，首先将表格排序，对于男性和女性分开来。接下来，分别截取 662 条数据进行方差分析。将这些数据放到一个 X 矩阵中，p 会自动帮我们算出单因素实验的方差分析表格。

**ANOVA Table**

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Columns | 7.09511e+08 | 1 | 7.09511e+08 | 4.82 | 0.0283 |
| Error | 1.94623e+11 | 1322 | 1.47218e+08 | | |
| Total | 1.95332e+11 | 1323 | | | |

如图所示，算出的概率值 P 为 0.0283，相比显著性水平 0.05，P < 0.05，所以在显著性水平 0.05 下拒绝原假设。认为性别因素会具有非常显著的差异。

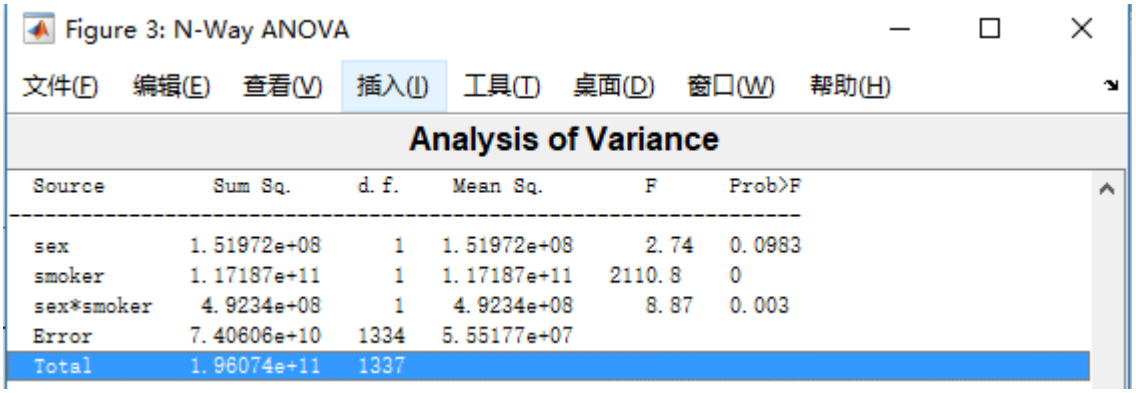



如图所示也可以看出分布的差异所在，男性的健康水平会比女性的健康水平的支出要高不少。或许是和吸烟有一定关系。

接下来是双因素的实验分析：双因素我们直接使用 anovan 的函数，函数接收一个 y 值还有几个 cell 的值，这些 cell 的值供我们分析相应的因素。现在的因素是性别和是否吸烟，分别命名为 sex 和 smoker。之后的 model 和 full 都是使用的模型参数，设置变量名为 sex 和 smoker。

```
Y = data_cell(:,7);
Y = cell2mat(Y);

%X = [male,female];
varnames= {'sex','smoker'};
p = anovan(Y,{sex,smoker},'model','full','varnames',varnames);
```

我们同样可以得到一个图，这个图里面的第一行是源，第二行是均方和，第三行是自由度，第四行是均值，后边就是 F 分布的值以及大于 F 的概率，即显著性水平。




在这个图里，我们可以明显看出，性别的值的显著性是 0.09，是大于 0.05 的，所以相对而言性别没有显著差异。然而吸烟的 F 值非常小，所以能够判断吸烟是有显著性差异的。而两者的交互因素也是非常显著的，小于 0.5。但是最显著的还是吸烟的差异。