一、回归分析

1.使用前 1333 条数据进行拟合

考虑在 matlab 中使用函数 regress()来做此问

关键步骤为[b, bint, r, rint, stats]=regress(p1_y, p1_x);

左边的五个参数分别为

b：得到的线性回归方程的系数

bint：回归系数的置信区间

r：残差

rint：残差的置信区间

stats：相关系数 R^2, F 值，与 F 对应的 p 值，误差方差

右边的为个人医疗费用，三组自变量提出来得到的矩阵

得到拟合方程为：y=-6872.967063 + 237.744072*x1 + 333.749986*x2 + 546.279722*x3

其中 x1，x2，x3 分别为年龄，bmi，子女数

2.得到的预测值如图

```
第1334组数据: 原值10600.548300 预测值16989.312781
第1335组数据: 原值2205.980800 预测值8059.725787
第1336组数据: 原值1629.833500 预测值9705.113218
第1337组数据: 原值2007.945000 预测值6730.408091
第1338组数据: 原值29141.360300 预测值17331.533438
```

而要计算个人医疗费用置信度为 95%的置信区间，搜索不到相关函数，考虑从书上找

由书上 46 页可得置信区间为 $\hat{y}_0 \pm t_{0.975}(20)\sqrt{MSE[1+\boldsymbol{x}_0^t(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0]}$

其中 MES 由前面的实验数据得到$\mathrm{MSE}=\frac{\mathrm{SSE}}{\mathrm{n-p}}$，而其中$\mathrm{SSE}=\sum_{\mathrm{i}=1}^{\mathrm{n}} r_i^2$

即可计算得到

```
第1334组数据: 原值10600.548300 预测值16989.312781 置信区间[-6758.102748, 40736.728310]
第1335组数据: 原值2205.980800 预测值8059.725787 置信区间[-15687.160790, 31806.612365]
第1336组数据: 原值1629.833500 预测值9705.113218 置信区间[-14041.857743, 33452.084178]
第1337组数据: 原值2007.945000 预测值6730.408091 置信区间[-17016.419687, 30477.235869]
第1338组数据: 原值29141.360300 预测值17331.533438 置信区间[-6416.155347, 41079.222223]
```

二、方差分析

1.该问题中假设个人医疗费用服从方差分析模型，要比较不同性别对医疗费用是否有显著水平为 0.05 的差异

即为单因子的方差分析，只考虑性别这一个因子对医疗费用的影响，各个样本即为该因子的不同水平，假设一个性别的医疗费用是一个总体，两个总体独立地服从同方差正态分布

即为服从分布$\mathrm{N}(\mathrm{u_i}, \sigma^2), \mathrm{i} = 1, 2$(分别代表男性与女性)

实验即需检验假设:$H_0 : \mu_1 = \mu_2$是否成立，若是拒绝，即认为不同性别对个人医疗费用有显著差异，否则，就认为没有

使用的函数为 p=anova1(X,group)

其中 X 指的是医疗费用，此时为样本中的第七列，为列向量

其中 group 为医疗费用中元素的组别，为行向量

此时认为 male 为组别 1，female 为组别 2，

例如第 i 个元素为 male 性别，则 group(i)=1，为 female 则为 2

最后得到结果 p=0.0361<0.05

即拒绝假设，即认为对不同性别对个人医疗费用有显著差异

具体步骤在 problem_2.mlx 文件中

2.要比较性别、是否抽烟两个因素对个人医疗费用是否有显著水平为 0.05 的差异

从题目可知，因素性别有 male，female 两个水平，因素是否抽烟有 no，yes 两个水平

类似的, matlab 中有函数 anova2()来处理双因素方差分析, 但无法根据此处样本来实现，因为每个种类的样本数量不相同，而 anova2()函数要求严格的方阵

即考虑使用 python 来做这个问

在查询资料后决定使用 ols 回归模型，以及 anova_lm 函数来做这个问

详细步骤在 problem_2_2.ipynb 中，具体为读入数据，建立模型与方差分析三步

得到结果 p 为 7.069618e-04 与 1.190490e-281 都小于 0.05

即可认为因素性别和是否抽烟的各水平对因变量个人医疗费用有显著差异

实验截图如下：

第一题

实时编辑器 - C:\Users\76349\Desktop\2\problem_1.mlx

problem_1.mlx | problem_2.mlx | problem_2_2.mlx | +

```
%清空数据
clear; clc;

%读取表格
data=readtable("data.txt");

%读出三组自变量数据与因变量数据
%并构造出相应的数据
temp=size(data); length=temp(1)-5;
p1_x=[ones(length,1) table2array(data(1:1333,1)) table2array(data(1:1333,3)) table2array(data(1:1333,4))];
p1_y=table2array(data(1:1333,7));

%进行拟合使用的是regress()函数
%左边的五个参数分别为
%b：得到的线性回归方程的系数
%bint：回归系数的置信区间
%r：残差
%rint：残差的置信区间
%stats：相关系数R^2，F值，与F对应的p值，误差方差
[b, bint, r, rint, stats]=regress(p1_y, p1_x);

%输出得到的结果
fprintf('得到拟合方程为：y=%f + %f*x1 + %f*x2 + %f*x3\n其中x1，x2，x3分别为年龄，bmi，子女数\n',b(1),b(2),b(3),b(4));
```

```
得到拟合方程为：y=-6872.967063 + 237.744072*x1 + 333.749986*x2 + 546.279722*x3
其中x1，x2，x3分别为年龄，bmi，子女数
```

```
%用后五组数据测试
t1_x=[ones(5,1) table2array(data(1334:1338,1)) table2array(data(1334:1338,3)) table2array(data(1334:1338,4))];
t1_y=table2array(data(1334:1338,7));

%得到MSE的值来计算置信区间
SSE=0;
for i=1:1333
    SSE=SSE+r(i)^2;
end

MSE=SSE/(1333-4);

%输出测试结果
fprintf('')
for i=1:5
    fprintf('第%d组数据：原值%f 预测值%f 置信区间[%f, %f]\n',i+1333,t1_y(i),t1_x(i,:)*b,t1_x(i,:)*b-2.086*sqrt(MSE*[1+t1_x(i,:)*(b'*b)^(-1)*t1_x(i,:)']),t1_x(
end
```

```
第1334组数据：原值10600.548300 预测值16989.312781 置信区间[-6758.102748, 40736.728310]
第1335组数据：原值2205.980800 预测值8059.725787 置信区间[-15687.160790, 31806.612365]
第1336组数据：原值1629.833500 预测值9705.113218 置信区间[-14041.857743, 33452.084178]
第1337组数据：原值2007.945000 预测值6730.408091 置信区间[-17016.419687, 30477.235869]
第1338组数据：原值29141.360300 预测值17331.533438 置信区间[-6416.155347, 41079.222223]
```

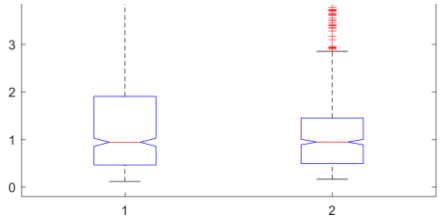
第二题

```
%清空数据
clear; clc;

%读取表格
%data_in_use即为要使用到的性别-医疗费用数据
data=readtable("data.txt");
data_in_use=table2cell([data(:,2) data(:,7)]);

%读出男性女性分别的数据
%用all来计数，用group来表示数据的类别
%得到以上数据的原因是使用的函数为p=anoval(X,group)形式
%其中X指的是医疗费用，此时为样本中的第七列，为列向量
%其中group为医疗费用中元素的组别，为行向量
all=1;
for i=1:1338
    if(data_in_use{i}=="male")
        if(all==1)
            group=1;
        else
            group=[group 1];
        end
    else
        if(all==1)
            group=2;
        else
            group=[group 2];
        end
    end
    all=all+1;
end

%使用anoval()函数来计算得到p值
p=anoval(cell2mat(data_in_use(:,2)),group);
```



ANOVA Table

Source	SS	df	MS	F	Prob>F
Groups	6.4359e+08	1	6.4359e+08	4.4	0.0361
Error	1.95431e+11	1336	1.4628e+08		
Total	1.96074e+11	1337			

In [8]:

```
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

#读入数据
data=pd.read_table("data.txt",sep=",")

#多因素方差分析
model=ols('charges ~ sex + smoker',data).fit()
anovat=anova_lm(model)
print(anovat)
```

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	6.435902e+08	6.435902e+08	11.524608	7.069618e-04
smoker	1.0	1.208777e+11	1.208777e+11	2164.527244	1.190490e-281
Residual	1335.0	7.455290e+10	5.584487e+07	NaN	NaN