**统计分析方法作业2**

姓名：孟衍璋 学号：16337183

# 题目1：回归分析

假设误差服从分布，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。

**线性回归拟合：**

首先使用textscan将txt文件中的内容读入矩阵中，再将三个定量变量存入数组之中，使用regress函数求出其线性回归的形式。



其中变量b表示回归系数；

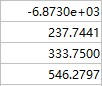
变量bint表示回归系数的区间估计；

变量r表示残差向量；

变量rint表示置信区间；

变量stats表示用于检验回归模型的统计量，有三个数值，相关系数r2、F值、与F对应的概率p。

其中回归系数如下：

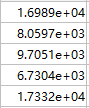


因此拟合出的曲线为：

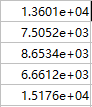
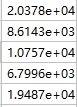


**预测：**

用上述公式验证后5个数据得到charges\_estimate：



置信度为95%的置信区间为（左为下界，右为上界）：

参考资料：

<https://wenku.baidu.com/view/139ec687b9d528ea81c779dd.html>

<https://www.mathworks.com/help/stats/regress.html>

# 题目2：方差分析

根据上例，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）。

首先使用textscan将txt文件中的内容读入矩阵中，再将性别和是否吸烟两项分别提取出来，将他们转化为字符串数组。

**单因素方差分析：**

进行单因素方差分析使用anova1函数：



得到检验p值：



小于给定显著性水平0.05，所以可以认为性别对个人医疗费用有显著影响。

**双因素方差分析：**

进行双因素方差分析使用anovan函数：



得到检验p值：



性别因素的检验p值大于显著性水平0.05，可认为没有显著影响；是否吸烟的检验p值远小于显著性水平0.05，所以有显著影响；两者的交叉影响的显著性水平小于0.05，所以可认为有显著影响。

参考资料：

<https://www.mathworks.com/help/stats/anova1.html>

<https://www.mathworks.com/help/stats/anova2.html>