**统计分析方法第二章作业**

姓名：陈泽锋 学号：16337035

**背景**

医疗费用线性回归预测，六个自变量和一个因变量（医疗费用），分别为：年龄，性别，体质指数，孩子个数，是否吸烟，地区，医疗费用。其中年龄，体质指数，孩子个数三个变量是定量变量，其他三个为定性变量。

**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。

1.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。

2.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。

1. 方差分析。

根据上例子，利用同样的数据集（1338条数据）：

1.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。

2.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）。

**程序清单**

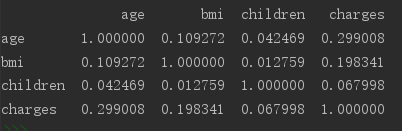
见Medical.py

**实验结果**

一．回归分析

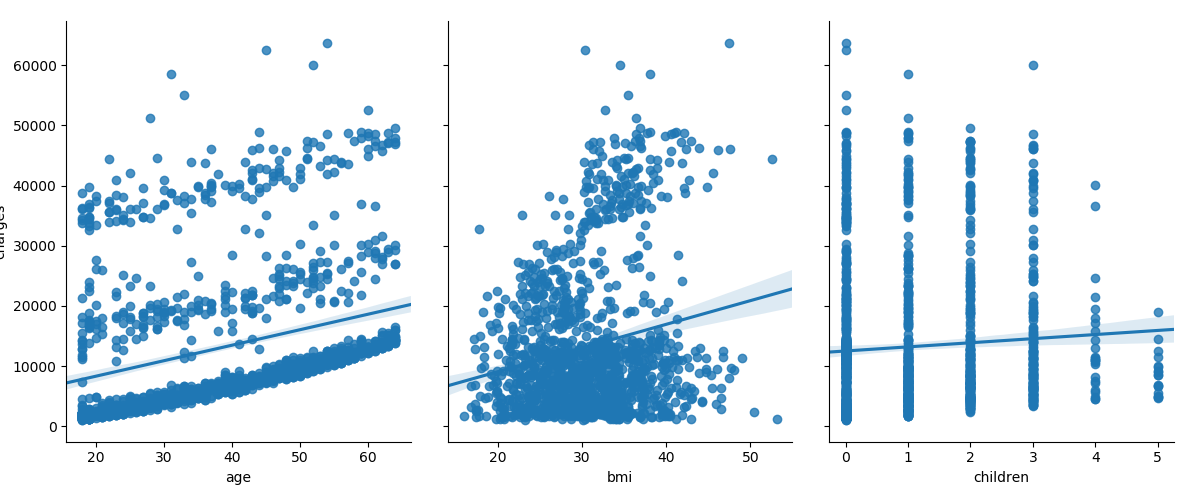
1.1333条数据的线性回归拟合

age、bmi、children、charges的皮尔森系数矩阵：



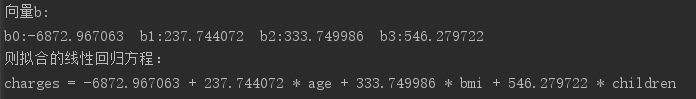
结果保存在Quantitative Pearson.csv中，由该皮尔森系数矩阵可以看出，charges与这三个定量的数值的相关性不高，拟合出来的结果应该会有较大的偏差。

不同因素对于charges的预测图（阴影部分为置信度95%的预测区间）：



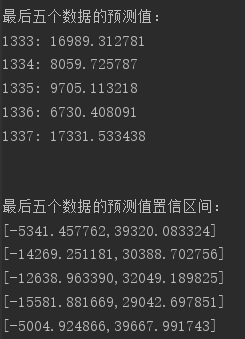
结果保存在Linear regression fitting.png中。

求得的线性回归拟合方程：



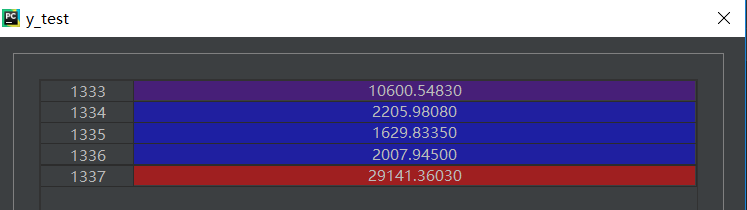
结果保存在Linear regression equation.png中。

2.预测剩余的5条数据及置信度95%的置信区间：



结果保存在Last five.png中。

对比5条数据实际的charges的值：

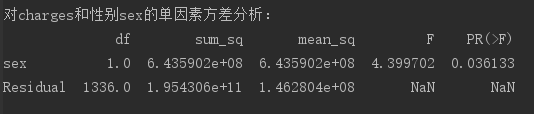


发现结果确实会有很大的偏差 ，说明这个用线性回归拟合模型不准确。

二．方差分析

1.charges与sex的单因素方差分析

利用ols和anova\_lm求得charges和sex的单因素方差分析相关矩阵：



结果保存在Charges\_Sex.png中。

Df表示自由度，sum\_sq表示平方和，mean\_sq表示均方和，F表示F检验统计量的值，PR（>F）表示检验p值，sex是因素sex，Residual是残差。

此时p值 = 0.036133 < 0.05,所以有理由拒绝假设H0（均值相等），则charges和sex显著相关。

2.charges与sex、smoker的双因素方差分析

利用ols和anova\_lm求得charges和sex、smoker的双因素方差分析相关矩阵：



结果保存在Charges\_Sex\_Smoker.png中。

Df表示自由度，sum\_sq表示平方和，mean\_sq表示均方和，F表示F检验统计量的值，PR（>F）表示检验p值，sex是因素sex，smoker是因素smoker，Residual是残差。

此时sex、smoker的p值都明显<0.05,所以有理由拒绝假设H0（均值相等），则charges和sex、smoker显著相关。