实验使用工具：MATLAB2017a

* **第一题：**

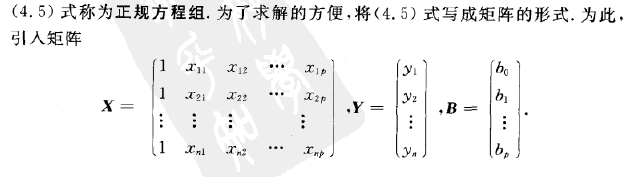
假设误差服从正态分布，建立个人医疗费用和3个定量变量之间的线性回归方程并研究相应的统计推断问题。

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。

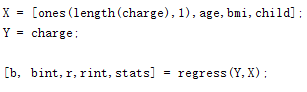
用最后5条数据进行测试。请预测他的个人医疗费用，并给出置信度为95%的置信区间。

**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。



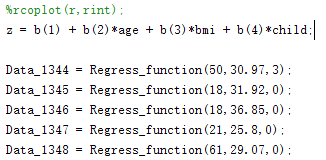
构造完之后，按照格式将参数写回。其中，b是B矩阵，也就是回归系数。Bint是区间估计的值，r为残差，rint是残差的置信区间，stats是检验会规模型的统计量。

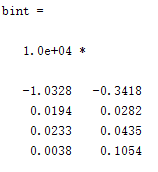
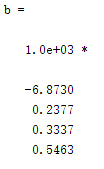


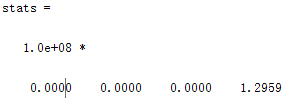
之后，我们就能根据结果输出这些值，我们可以看到b的值已经求出来了，所以最终的回归方程的拟合就是:

Y = -6873 + 237.7 X1 + 333.7 X2 + 546.3 X3

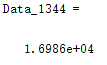
各个参数的输出如下，我们可以发现的问题是，stats的数值，相关系数是0，误差的方差也非常大。所以可以断定，这个回归方程并不显著。这些值并不符合线性回归。

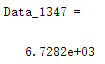




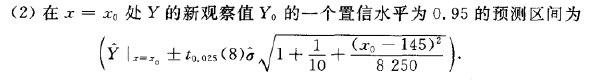


接下来是利用最后五条数据进行估计验证，得到的结果分别如下:





而预测区间，即置信区间，按照如下公式计算。



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。

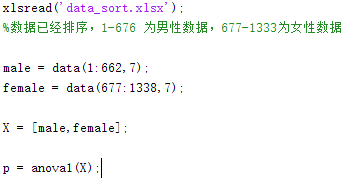
* **第二题：**

根据上例子，利用同样的数据集（1338条数据）：

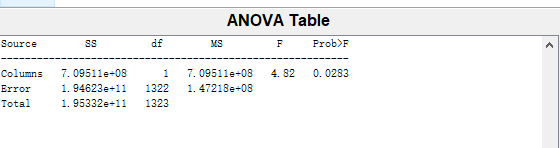
 利用方差分析知识，假设个人医疗费用服从方差分析模型，见（3.1）或（3.2）比较不同性别对个人医疗费用是否有显著（显著水平为0.05）差异。

 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）。

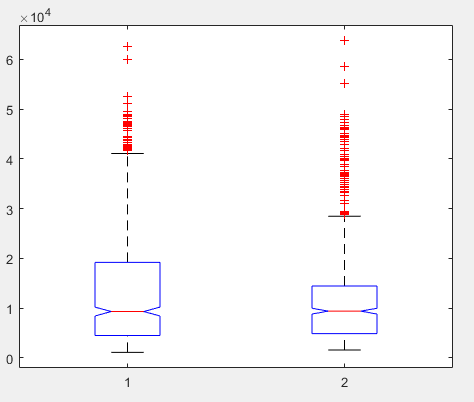
1. 首先使用方差分析模型。方差分析使用matlab的库函数anova1，ANOVA1返回一个p的值，判断这个p值和概率的大小，从而判断是否具有显著性差异



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。

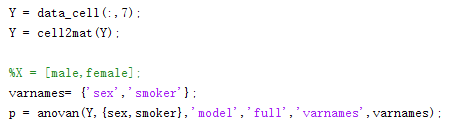


如图所示，算出的概率值P为0.0283，相比显著性水平0.05，P < 0.05，所以在显著性水平0.05下拒绝原假设。认为性别因素会具有非常显著的差异。

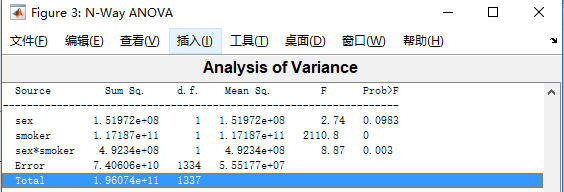


如图所示也可以看出分布的差异所在，男性的健康水平会比女性的健康水平的支出要高不少。或许是和吸烟有一定关系。

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。



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。