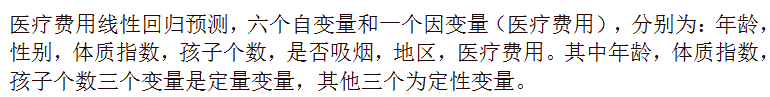
****

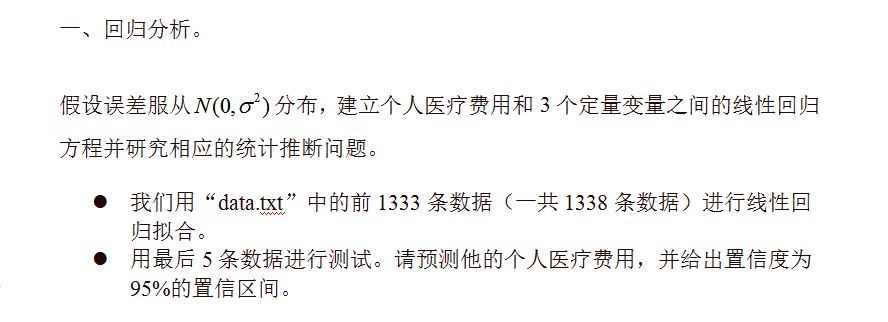
**《统计分析方法》**

**实验报告**

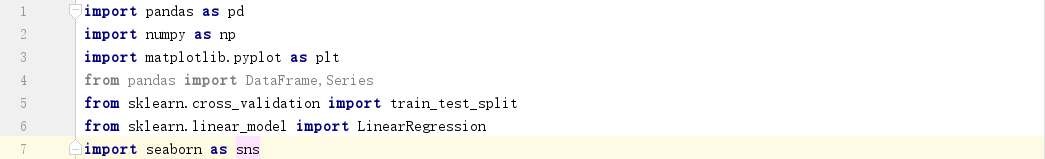
**（实验二）**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **学院名称** | **：** | 数据科学与计算机学院 | | | | | |
| **学生姓名** | **：** | 肖越 | | | | | |
| **学号** | **：** | 16338015 | | | | | |
| **专业（班级）** | **：** | 16级信息安全 | | | | | |
| **时间** | **：** | 2018 | 年 | 11 | 月 | 3 | 日 |

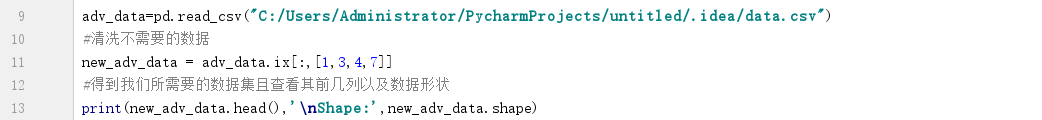




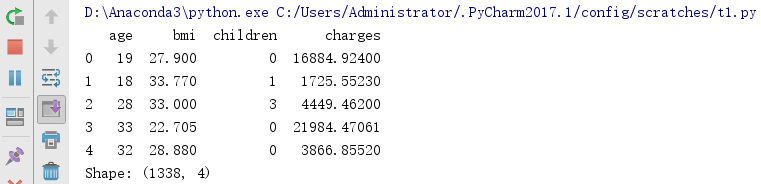
**1.引入要使用的包：**



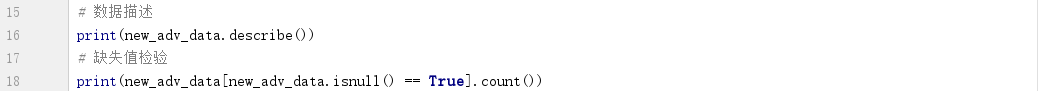
**2.读取表格（为方便起见，已经转换为excel(.csv)），并且我们本次研究的三个定量变量是年龄、bmi和孩子数量（第1、3、4列），并且展示前几行读取的数据视图：**

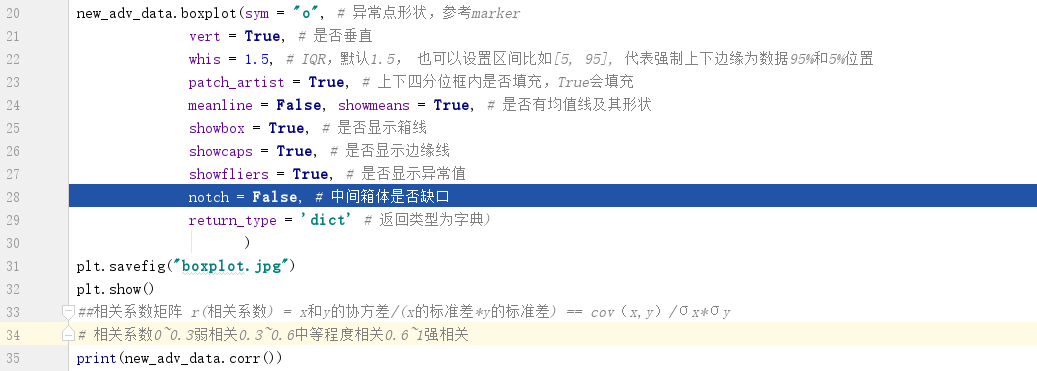


**结果：**

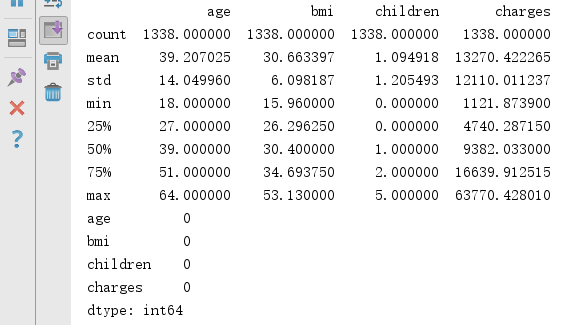


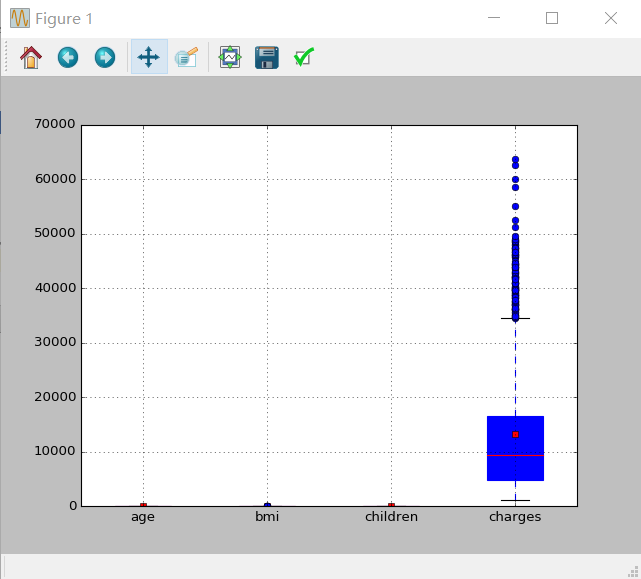
**3.展示数据的相关描述并且展示相关性的箱线图：**





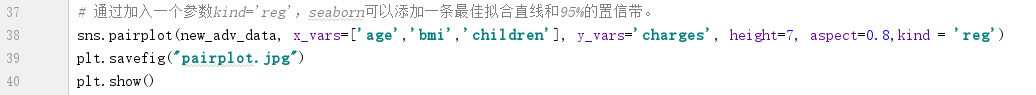
**结果：**



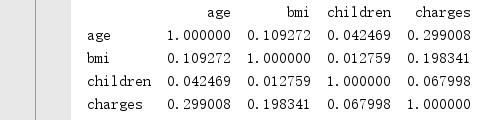


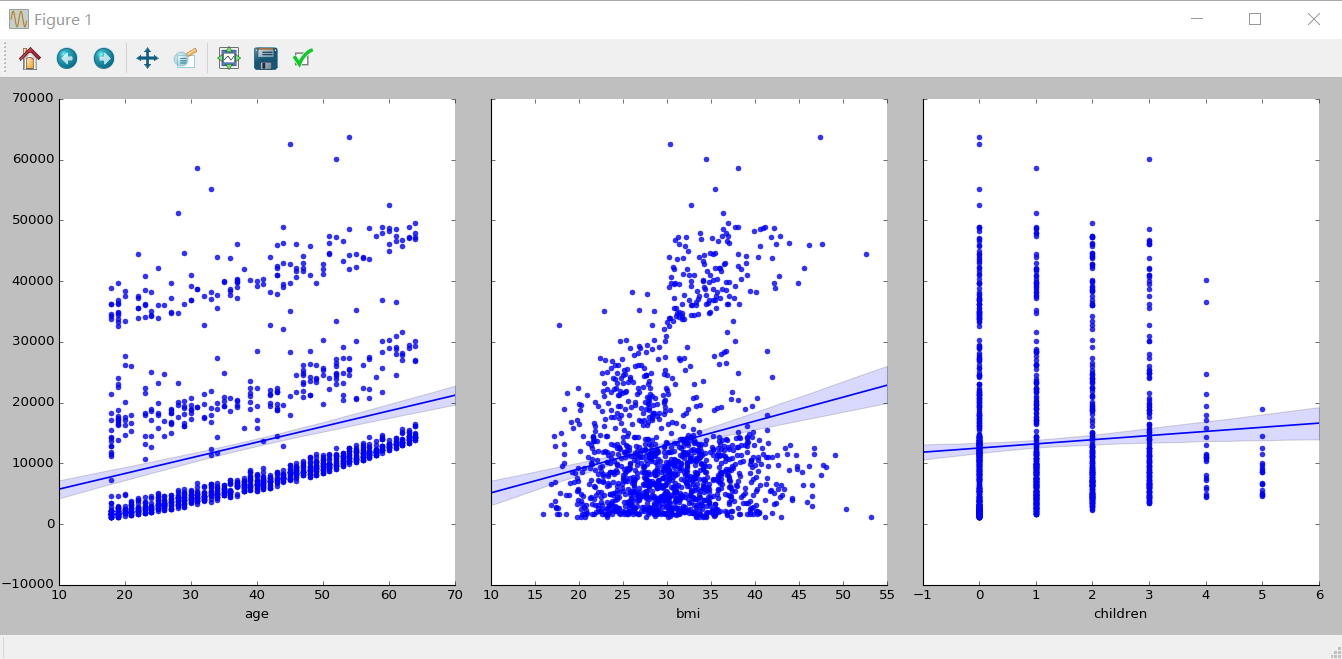
**以上箱线图是为了能够甄别是否存在不相关的定量变量，从图中可以看出charges在数值方面和定变量的差距较大，箱线图的可读性较低。**

**4.展示相关系数，通过加入一个参数kind='reg'，seaborn可以添加一条最佳拟合直线和95%的置信带。**

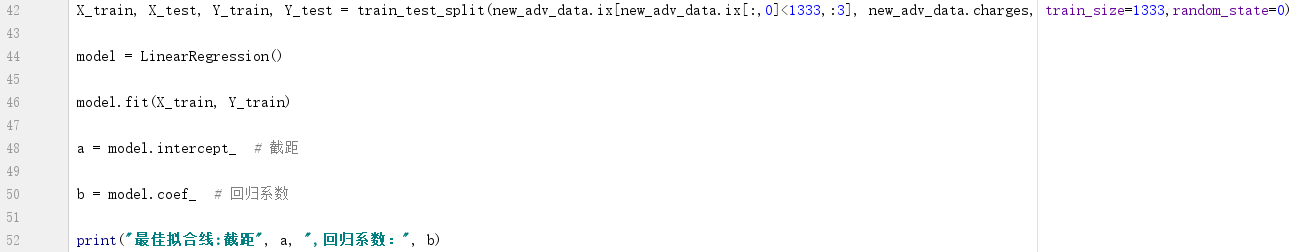


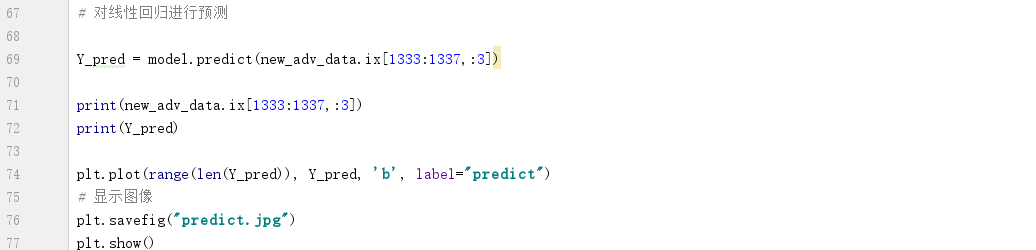
**结果：**

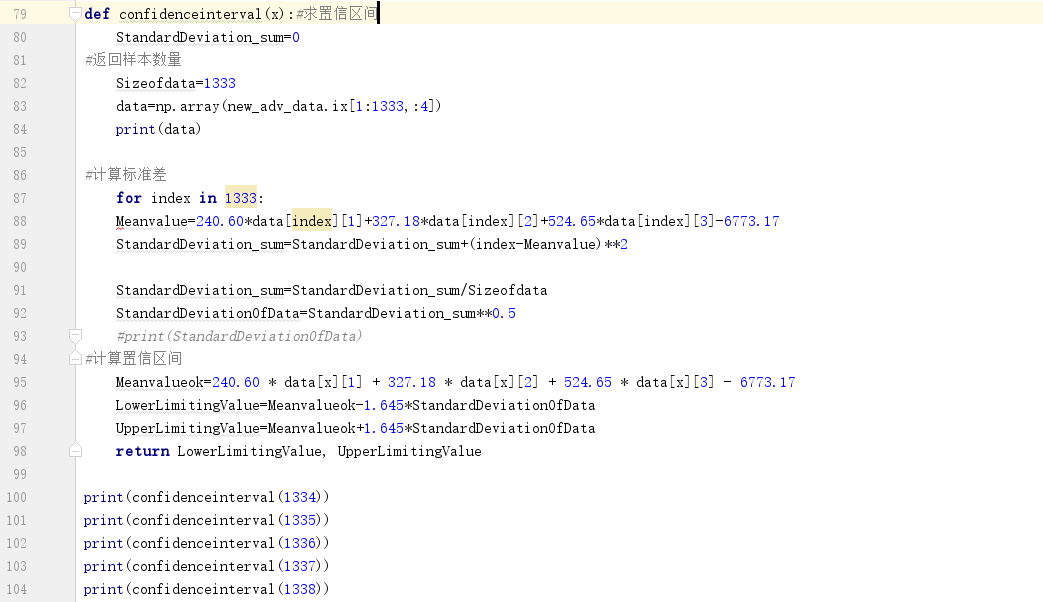


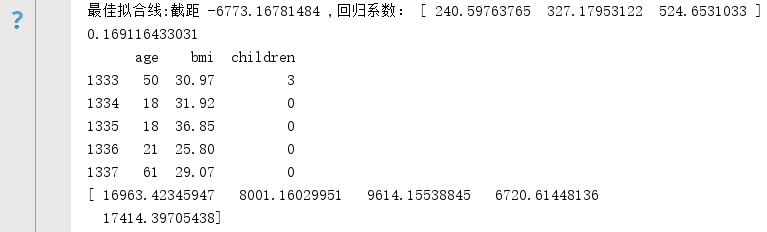


**5.通过前1333组对后五组数据的预测并计算置信区间：**





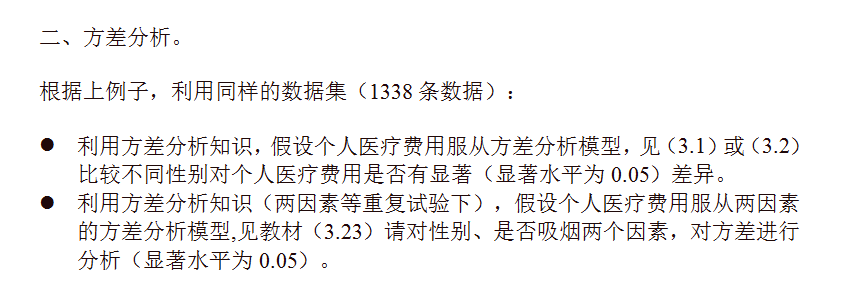
**结果：**



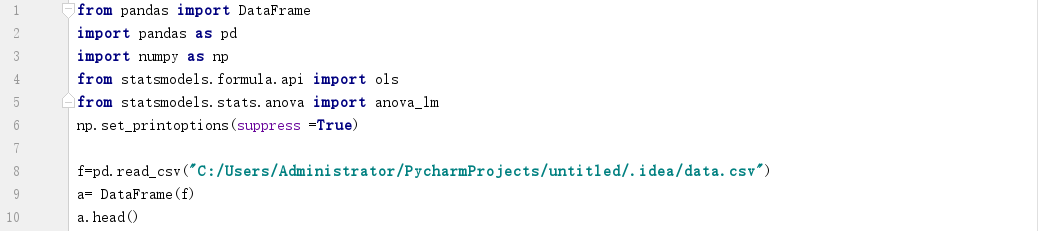
**因此，以上我们可以得到最佳拟合函数：y=240.598\*x1+327.180\*x2+524.653\*x3-6773.16**

**医药费和相关置信区间的预测可见上图。**

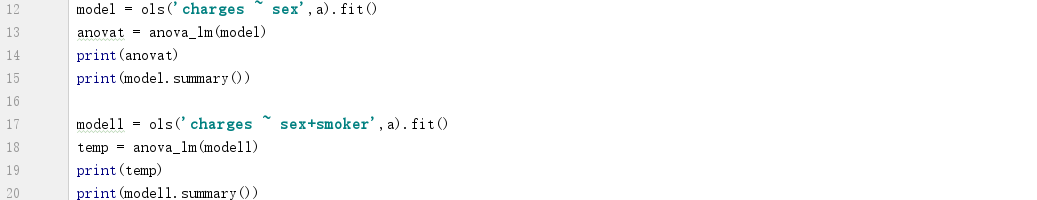
**反思：本曲线的拟合程度的得分为0.1691（远远小于1），那么我们认为其实三个定量变量的相关性并不是很强，我们通过最开始的分组相关性拟合图像应该可以看出，最不相关的应该就是孩子数，结合实际情况我们可以考虑将孩子数进行舍去，进而提升拟合相关系数，相信可以得到更为接近的预测结果。**



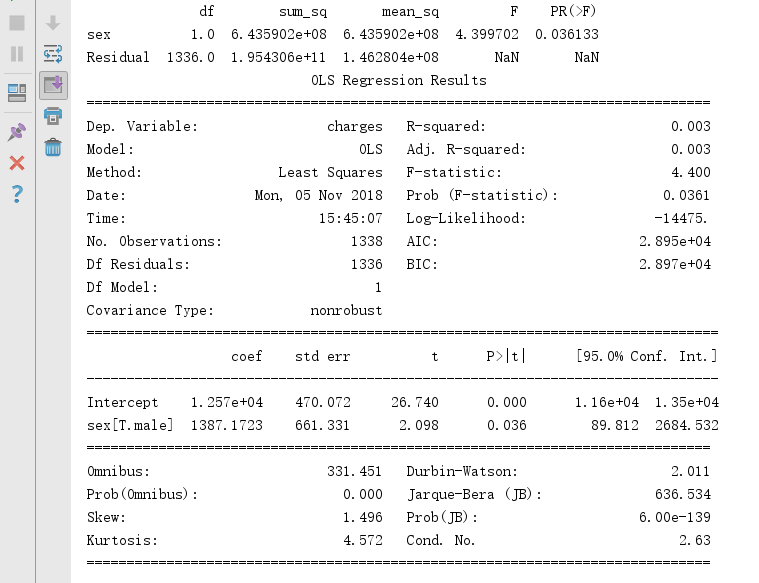
**1.调入引用包并且读取文件：**

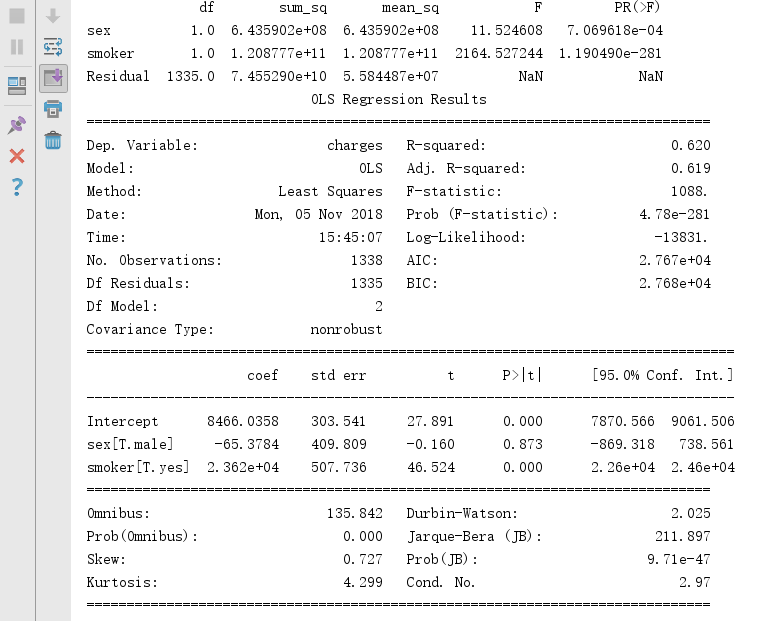


**2.通过ols.fit()和ols.summary()进行方差和显著性分析：**



**结果：**





**分析：**

1. **如图可知拒绝假设即性别和个人医疗费用有显著性差异（f=0.03<0.05）.**
2. **如图可知拒绝假设即性别和抽烟和个人医疗费用有显著性差异应该被拒绝（f=0.873>0.05）。**