中山大學
SUN YAT-SEN UNIVERSITY

《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
专业（班级）：	16 信息安全
学 生 姓 名 ：	李默程
学　　号 ：	16338008
时　　间 ：	2018　年　11　月　4　日

成 绩 ：

实验二 ： 回归分析与方差分析

一. 实验目的

1. 掌握线性回归的使用与分析
2. 应用方差分析，掌握对数据方差分析模型的应用与理解

二. 实验内容

利用提供的病人数据集，首先进行回归分析，对三个定量变量拟合线性回归函数，并利用最后五个数据作为测试集进行测试。

对数据集的两个定量变量进行方差分析。利用anova过程，对单因素、双因素显著影响的置信度进行判断。

三. 实验过程与结果

引用的库文件和库函数

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn import linear_model
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import statsmodels.sandbox.regression.predstd as pred
```

第一题:

```
data = pd.read_csv('data.txt')#读取文件

#线性回归函数
train_data = data.iloc[:1333, :]#训练集
test_data = data.iloc[1333:, :]#测试集

clf = linear_model.LinearRegression()#回归函数
```

```
clf.fit(train_data.iloc[:, [0, 2, 3]], train_data.iloc[:, 6])#拟合
y_pred = clf.predict(test_data.iloc[:, [0, 2, 3]], ) #对测试集预测
print('回归系数，常量')
print(clf.coef_, clf.intercept_)#回归系数
print('预测结果：')
print(y_pred)
print('真实结果：')
print(np.array(test_data.iloc[:, 6]))
 #绘图
sns.pairplot(train_data, x_vars=['age', 'bmi', 'children'], y_vars='charges', size=7, aspect=0.8, kind='reg')
plt.show()

#计算置信区间
model = ols('charges~age+children+bmi', data).fit()#y, x
print('置信区间:')
_, pre_lower, pre_upper = pred.wls_prediction_std(model, alpha=0.05)
pre = np.array([pre_lower[1333:], pre_upper[1333:]]).T
print(pre)
```

回归系数，常量

[237.74407233 333.74998586 546.27972192] -6872.967063302547

预测结果:

[16989.31278129 8059.72578743 9705.11321773 6730.40809095

真实结果:

[10600.5483 2205.9808 1629.8335 2007.945 29141.3603]

置信区间:

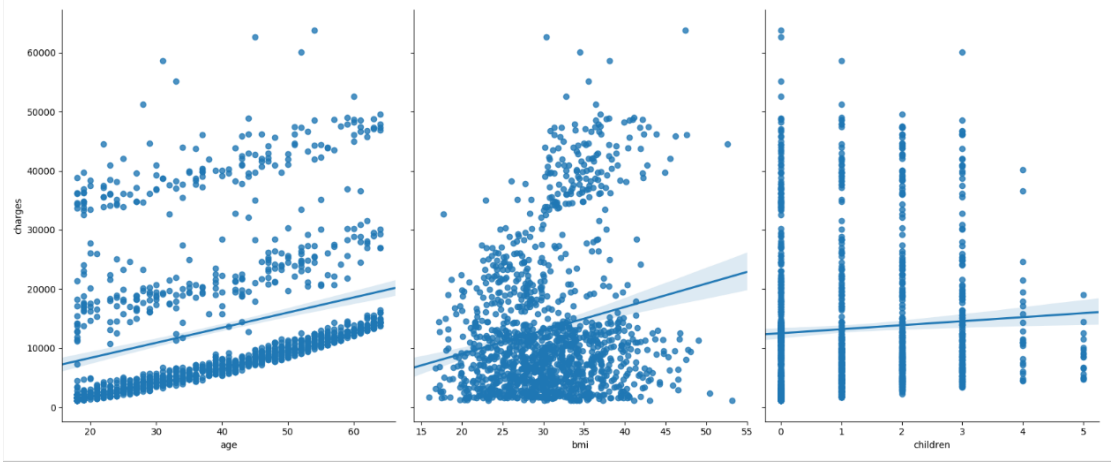
[[-5346.176902 39339.56914662]

 [-14340.32033631 30347.8367052]

 [-12713.80867656 31995.66701946]

 [-15650.17284686 29032.95567999]

 [-4970.33465636 39724.50063684]]



三个图分别显示了三个变量与回归函数之间的关系。其中散点是1333个数据，直线是回归函数的投影。

第一题分析:

从预测结果和实际结果比较来看，其差距非常明显。并且0.95置信区间给出的区间范围过大，失去了判断意义。不难从图中看出，仅有age这个向量与charges存在比较明显的正相关关系，但分布较散，没有明显的依赖关系。而另外两个变量则几乎不存在相关关系。

第二题:

```
#第二题
print('显著性分析:')
print('费用与性别的相关性')
model2 = ols('charges ~ sex', data).fit()
print(anova_lm(model2))

print('费用与性别和是否吸烟的相关性')
model3 = ols('charges ~ sex + smoker', data).fit()
print(anova_lm(model3))
```

显著性分析:

费用与性别的相关性

	df	sum_sq	mean_sq	F	PR(>F)
sex	1.0	6.435902e+08	6.435902e+08	4.399702	0.036133
Residual	1336.0	1.954306e+11	1.462804e+08	NaN	NaN

费用与性别和是否吸烟的相关性

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

[3 rows x 5 columns]

第二题分析：

从性别与费用相关性的proc anova过程可以看出，p值为0.03，小于0.05，可以认为性别与费用在水平0.05下显著相关。

性别、是否吸烟两个数据与费用的相关性分析中，看到p值分别在10的-4次方和10的-281次方级，可以认为费用在性别和是否抽烟上是存在显著差异的。

四. 实验心得

本次实验大量的时间都花在了寻找函数上。Python虽然拥有常用的数据分析库，但其函数库有点过于零散和庞大，以至于同一个功能有近十个函数可以实现，用法却都有一点细微的差别。比如本次对线性回归的操作，sklearn库提供了快速高效的线性回归拟合，却并不适用于高维回归的绘图；Seaborn提供了高效的数据可视化方案，却不提供具体的参数和预测接口。