



《统计分析方法实验》 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：陈镕希

学号：16337028

时间：2018 年 10 月 29 日

成绩：

实验二：医疗费用线性回归预测

一. 实验目的

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。

二. 实验内容

1、回归分析。

假设误差服从 $N(0, \sigma^2)$ 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。

2、方差分析。

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）。

三. 实验及算法原理

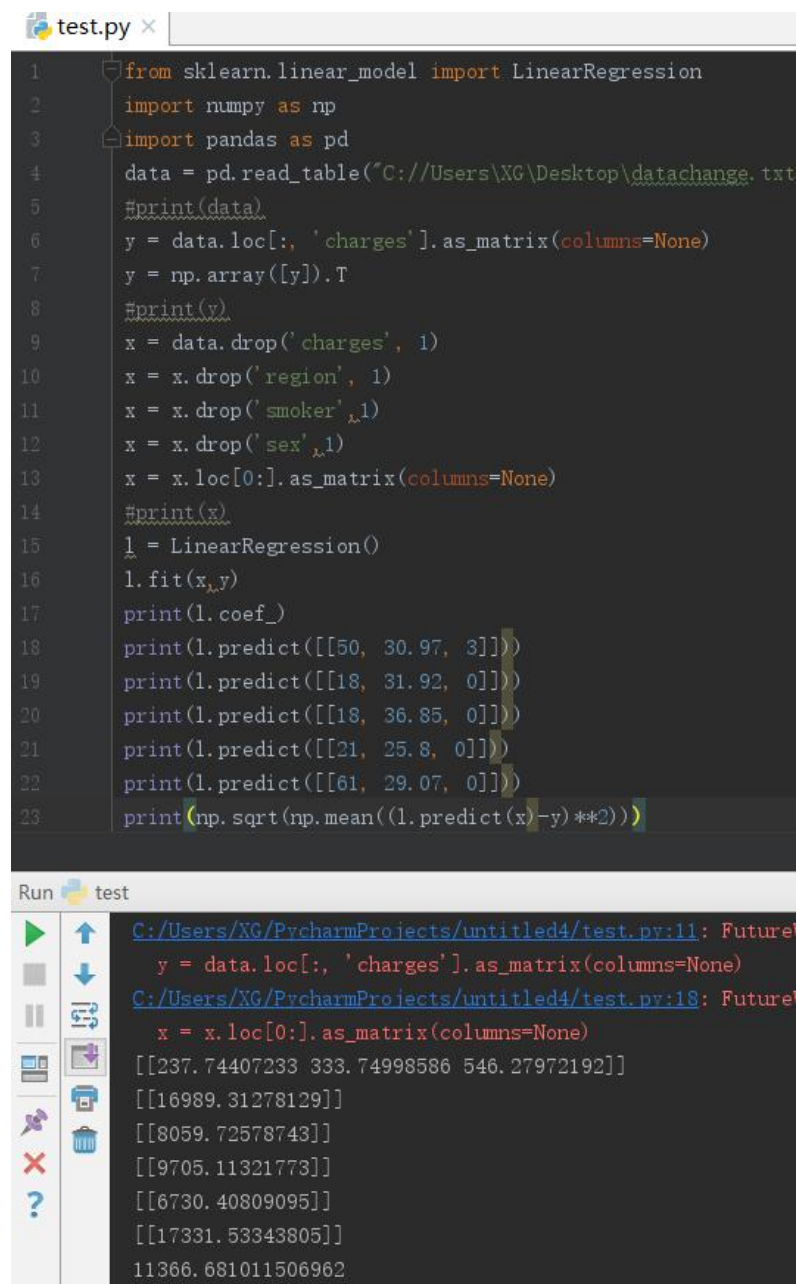
我是使用python语言来做的这一次实验,首先我利用pandas库中的read_table函数读入txt文件中的数据,第一题中用到了sklearn.linear_model库中的LinearRegression来作线性回归,其中运用最多的就是drop函数来去除掉不必要的列;第二题要使用statsmodels.formula.api库中的ols函数进行最小二乘法的线性回归以及statsmodels.stats.anova库中的anova_lm函数进行方差分析

四. 程序清单

1. test.py
2. test1.py
3. test2.py

五. 运行截图

第一题



```

test.py x
1  from sklearn.linear_model import LinearRegression
2  import numpy as np
3  import pandas as pd
4  data = pd.read_table("C://Users\XG\Desktop\datachange.txt")
5  #print(data)
6  y = data.loc[:, 'charges'].as_matrix(columns=None)
7  y = np.array([y]).T
8  #print(y)
9  x = data.drop('charges', 1)
10 x = x.drop('region', 1)
11 x = x.drop('smoker', 1)
12 x = x.drop('sex', 1)
13 x = x.loc[0:].as_matrix(columns=None)
14 #print(x)
15 l = LinearRegression()
16 l.fit(x,y)
17 print(l.coef_)
18 print(l.predict([[50, 30.97, 3]]))
19 print(l.predict([[18, 31.92, 0]]))
20 print(l.predict([[18, 36.85, 0]]))
21 print(l.predict([[21, 25.8, 0]]))
22 print(l.predict([[61, 29.07, 0]]))
23 print(np.sqrt(np.mean((l.predict(x)-y)**2)))

```

Run test

```

C:/Users/XG/PycharmProjects/untitled4/test.py:11: FutureWarning:
y = data.loc[:, 'charges'].as_matrix(columns=None)
C:/Users/XG/PycharmProjects/untitled4/test.py:18: FutureWarning:
x = x.loc[0:].as_matrix(columns=None)
[[237.74407233 333.74998586 546.27972192]]
[[16989.31278129]]
[[8059.72578743]]
[[9705.11321773]]
[[6730.40809095]]
[[17331.53343805]]
11366.681011506962

```

因为95%的置信区间即为 $[0.025, 0.975]$ 在标准正态分布中为 $[-1.96, 1.96]$

求得标准差为11366.681011506962因为假设误差属于正态分布，因此五名预测人员的95%置信区间分别为：

未命名4.cpp

```

1  #include<iostream>
2  #include<iomanip>
3  using namespace std;
4  int main(){
5      double a=16989.31278129,b=8059.72578743,c=9705.11321773,d=6730.40809095,e=17331.53343805;
6      double bzc=1.96*11366.681011506962;
7      cout<<"标准差乘1.96的结果为: "<<fixed<<setprecision(8)<<bzc<<endl;
8      cout<<"第一个预测人的置信区间为: ["<<a-bzc<<","<<a+bzc<<"]<<endl;
9      cout<<"第二个预测人的置信区间为: ["<<b-bzc<<","<<b+bzc<<"]<<endl;
10     cout<<"第三个预测人的置信区间为: ["<<c-bzc<<","<<c+bzc<<"]<<endl;
11     cout<<"第四个预测人的置信区间为: ["<<d-bzc<<","<<d+bzc<<"]<<endl;
12     cout<<"第五个预测人的置信区间为: ["<<e-bzc<<","<<e+bzc<<"]<<endl;
13 }

```

```

C:\Users\XG\Desktop\未命名4.exe
标准差乘1.96的结果为: 22278.69478255
第一个预测人的置信区间为: [-5289.38200126, 39268.00756384]
第二个预测人的置信区间为: [-14218.96899512, 30338.42056998]
第三个预测人的置信区间为: [-12573.58156482, 31983.80800028]
第四个预测人的置信区间为: [-15548.28669160, 29009.10287350]
第五个预测人的置信区间为: [-4947.16134450, 39610.22822060]

-----
Process exited after 1.208 seconds with return value 0
请按任意键继续. . .

```

第一个预测人的置信区间为: [-5289.38200126, 39268.00756384]

第二个预测人的置信区间为: [-14218.96899512, 30338.42056998]

第三个预测人的置信区间为: [-12573.58156482, 31983.80800028]

第四个预测人的置信区间为: [-15548.28669160, 29009.10287350]

第五个预测人的置信区间为: [-4947.16134450, 39610.22822060]

第二题

```

test.py × test1.py ×
4  from statsmodels.formula.api import ols
5  from statsmodels.stats.anova import anova_lm
6
7  data = pd.read_table("C://Users\XG\Desktop\data.txt", sep=',')
8  x = data.drop('age', 1)
9  x = x.drop('region', 1)
10 x = x.drop('smoker', 1)
11 x = x.drop('children', 1)
12 x = x.drop('bmi', 1)
13 #print(x)
14 d1 = x[x['sex'] == 'male']['charges']
15 d2 = x[x['sex'] == 'female']['charges']
16 args = [d1,d2]
17 #levene test
18 #print(args)
19 w, p = stats.levene(*args)
20 #方差分析
21 #print(w, p)
22 f, p = stats.f_oneway(*args)
23 #print(f, p)
24 anova_results = anova_lm(ols('charges~C(sex)',x).fit())
25 print(anova_results)

```

Run test1

```

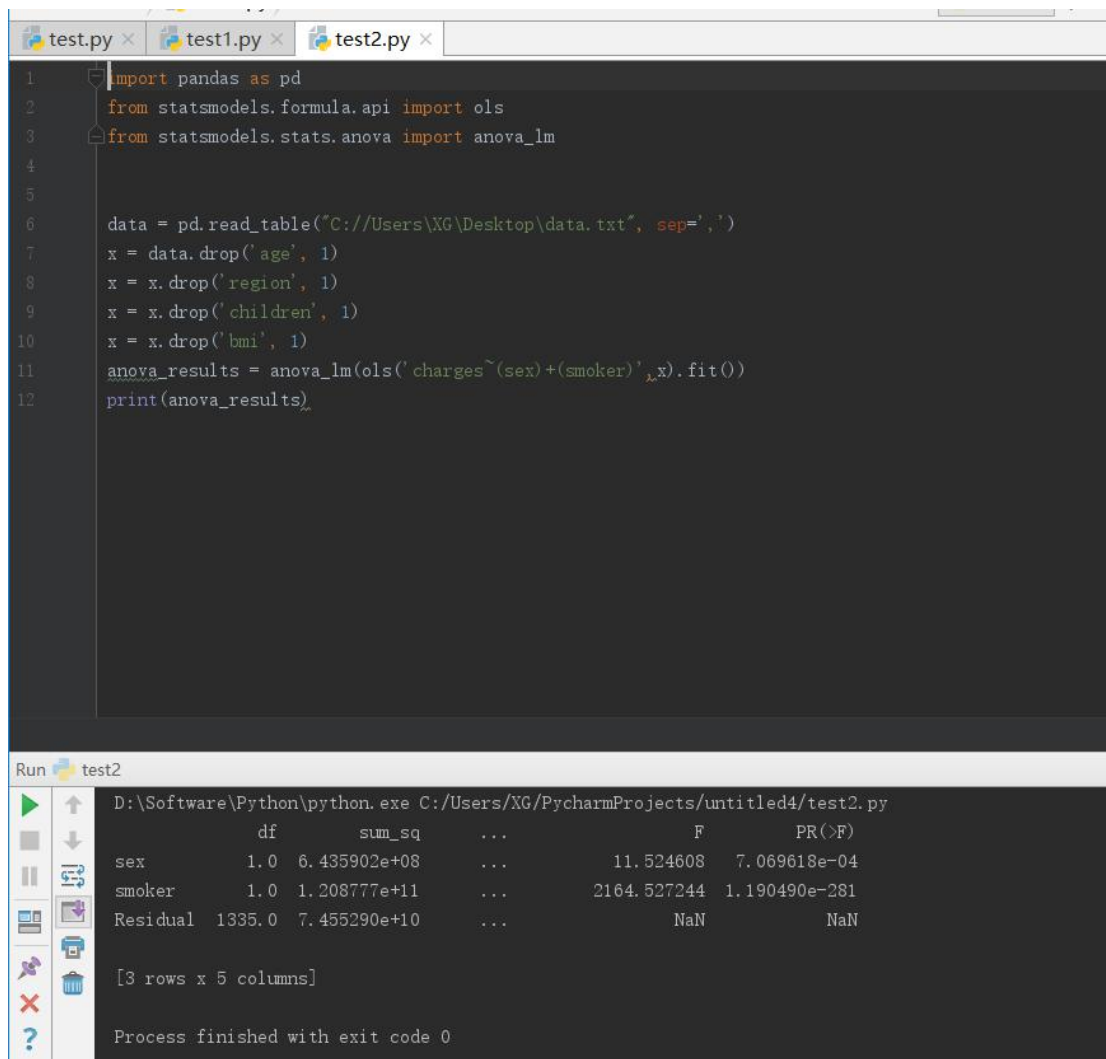
D:\Software\Python\python.exe C:/Users/XG/PycharmProjects/untitled4/test1.py

```

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

Process finished with exit code 0

可以看到p值为0.036133，由于题目中显著水平为0.05，因此可以判断不同性别对个人医疗费用有显著性差异



```

1 import pandas as pd
2 from statsmodels.formula.api import ols
3 from statsmodels.stats.anova import anova_lm
4
5
6 data = pd.read_table("C://Users\XG\Desktop\data.txt", sep=',')
7 x = data.drop('age', 1)
8 x = x.drop('region', 1)
9 x = x.drop('children', 1)
10 x = x.drop('bmi', 1)
11 anova_results = anova_lm(ols('charges~(sex)+(smoker)', x).fit())
12 print(anova_results)

```

Run test2

	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]

Process finished with exit code 0

可以看到关于性别的p值为7.069618e-4，由于题目中显著水平为0.05，因此可以判断不同性别对个人医疗费用有显著性差异；而关于是否吸烟的p值为1.190490e-281，这个数是非常非常的小了，可以判断是否吸烟对个人医疗费用有着极其明显的显著性差异。

六. 参考文献

1. Pandas 常用 I/O (一) -----read_csv(),read_table()
——https://blog.csdn.net/shener_m/article/details/81047669
2. pandas.read_table
——http://pandas.pydata.org/pandas-docs/version/0.20/generated/pandas.read_table.html
3. pandas 读取 txt 文件 (read_table 函数)
——https://blog.csdn.net/sinat_22659021/article/details/80881723
4. python 的 pandas 库中 read_table 的参数
——<https://blog.csdn.net/uvwxyzhao/article/details/80880735>

5. python Pandas 读取 txt 表格
——<https://blog.csdn.net/u011077672/article/details/50960580>
6. Python 科学计算：读取 txt, csv, mat 文件
——<https://blog.csdn.net/xierhacker/article/details/53201308>
7. Python 科学计算：读取 txt, csv, mat 文件
——https://blog.csdn.net/xierhacker/article/details/53201308?utm_source=blogxgwz1
8. Python 之读取 TXT 文件的三种方法
——https://blog.csdn.net/shandong_chu/article/details/70173952?utm_source=blogxgwz0
9. 显著性水平 置信度 置信区间 实例讲解
——<https://blog.csdn.net/bitcarmanlee/article/details/50911533>
10. 正态总体参数的置信区间及显著性检验一览表
——<https://wenku.baidu.com/view/b6270595dd88d0d233d46ac8.html>
11. Python 多元线性回归-sklearn.linear_model, 并对其预测结果评估
——https://blog.csdn.net/HHTNAN/article/details/78843722?utm_source=blogxgwz0
12. python 实现多元线性回归
——https://blog.csdn.net/cool_jia/article/details/79241070
13. 斯坦福机器学习笔记三 - 多变量线性回归
——<https://blog.csdn.net/u011221820/article/details/79141302>
14. statsmodels.stats.anova.anova_lm
——http://www.statsmodels.org/stable/generated/statsmodels.stats.anova.anova_lm.html
15. sklearn 实战-乳腺癌细胞数据挖掘
——<http://www.cnblogs.com/webRobot/p/6877283.html>
16. 【通俗向】方差分析--几种常见的方差分析
——https://blog.csdn.net/Yunru_Yang/article/details/68065070
17. python 方差分析
——<https://blog.csdn.net/yijiaobani/article/details/78113293>
18. Python 统计分析：[3]单因素方差分析
——<https://jingyan.baidu.com/article/cddddd41c6a2f2553cb00e13b.html>
19. Python 统计分析：[4]多因素方差分析
——<https://jingyan.baidu.com/article/a378c96090f550b328283020.html>
20. Scipy 教程 - 统计函数库 scipy.stats
——<https://blog.csdn.net/pipisorry/article/details/49515215>
21. scipy.stats.f_oneway
——https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html
22. Python 多因素方差分析
——https://blog.csdn.net/qq_38214903/article/details/82938612
23. 【通俗向】方差分析--几种常见的方差分析
——https://blog.csdn.net/Yunru_Yang/article/details/68065070