统计分析方法第二次作业

数据科学与计算机学院 16级信息安全 15952030 金永埈

实验内容：

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。

一、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。

* 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。

二、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）。

实验工具：Python

实验代码分析：

data=pd.read\_csv('D:/python/data.csv')

Data is a function required when reading data Excel.

x=a[['age','bmi','children']]

Within x, age and bmi also have check children function.

y=a[['charges']]

In y, there is a function to check the Charges data.

lrModel=LinearRegression()

Modeling.

lrModel.fit(x,y)

Training model.

print(np.c\_[np.array(['age','bmi','children']).reshape((3,1)),lrModel.coef\_.reshape((3,1))])

View coefficient.

print(lrModel.intercept\_)

View the intercept.

rint('Linear regression equation：')  
print('y=',lrModel.intercept\_[0],'+',lrModel.coef\_[0][0],'\* age','+',lrModel.coef\_[0][1],'\* bmi','+',lrModel.coef\_[0][2],'\* children')  
y\_pred=lrModel.predict([[50,30.970,3],[18,31.920,0],[18,36.850,0],[21,25.800,0],[61,29.070,0]])

Forecast.

plt.figure()   
plt.plot(range(len(y\_pred)),y\_pred,'b',label="predict")  
y\_test=[10600.54830,2205.98080,1629.83350,2007.94500,29141.36030]  
plt.plot(range(len(y\_pred)),y\_test,'r',label="test")  
plt.legend(loc="upper right")  
plt.xlabel('test cases')  
plt.ylabel('charges')  
plt.show()

Draw prediction and actual plans.

def ci\_t(data, confidence=0.95):   
 sample\_mean = np.mean(data)  
 sample\_std = np.std(data, ddof=1)  
 sample\_size = len(data)  
 alpha = 1 - confidence  
 t\_score = scipy.stats.t.isf(alpha / 2, df=(sample\_size - 1))  
  
 ME = t\_score \* sample\_std / np.sqrt(sample\_size)  
 lower\_limit = sample\_mean - ME  
 upper\_limit = sample\_mean + ME  
  
 return lower\_limit, upper\_limit  
  
X1=[10600.54830,2205.98080,1629.83350,2007.94500,29141.36030]  
print('95% confidence interval：')  
print(ci\_t(X1))

95% confidence interval.

ata=pd.read\_csv('D:/python/data.csv')   
model=ols('charges ~ sex',data).fit()  
anovat=anova\_lm(model)  
print(anovat)  
print()

The ability to distinguish between charges and gender in data.

formula = 'charges~ sex + smoker'  
anova\_results=anova\_lm(ols(formula,data).fit())  
print(anova\_results)

The ability to tell the difference between charges and sex and smokers.

源代码：

New1

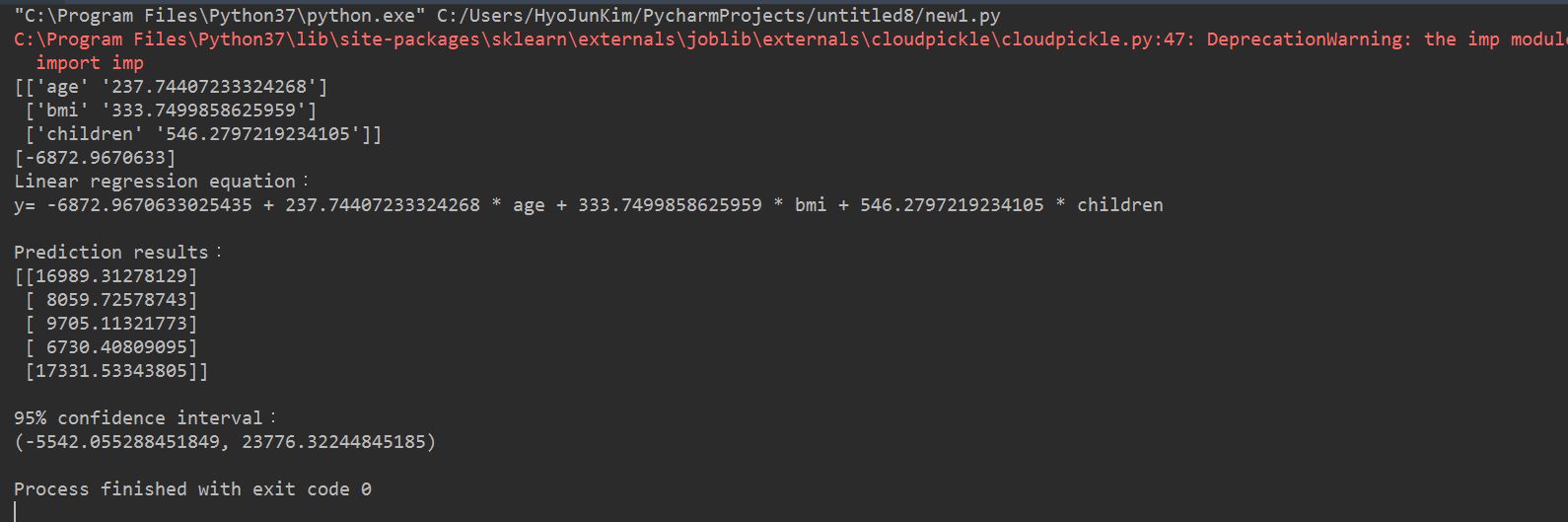
import pandas as pd  
import matplotlib.pyplot as plt  
import numpy as np  
import scipy  
from pandas import DataFrame,Series  
from sklearn.linear\_model import LinearRegression  
from scipy import stats  
  
data=pd.read\_csv('D:/python/data.csv')  
a=data[0:1333]  
x=a[['age','bmi','children']]  
y=a[['charges']]  
lrModel=LinearRegression() #建模  
lrModel.fit(x,y) #训练模型  
print(np.c\_[np.array(['age','bmi','children']).reshape((3,1)),lrModel.coef\_.reshape((3,1))]) #查看系数  
print(lrModel.intercept\_) #查看截距  
print('Linear regression equation：')  
print('y=',lrModel.intercept\_[0],'+',lrModel.coef\_[0][0],'\* age','+',lrModel.coef\_[0][1],'\* bmi','+',lrModel.coef\_[0][2],'\* children')  
y\_pred=lrModel.predict([[50,30.970,3],[18,31.920,0],[18,36.850,0],[21,25.800,0],[61,29.070,0]]) #预测  
print()  
print('Prediction results：')  
print(y\_pred)  
print()  
  
plt.figure() #画出预测和实际图  
plt.plot(range(len(y\_pred)),y\_pred,'b',label="predict")  
y\_test=[10600.54830,2205.98080,1629.83350,2007.94500,29141.36030]  
plt.plot(range(len(y\_pred)),y\_test,'r',label="test")  
plt.legend(loc="upper right")  
plt.xlabel('test cases')  
plt.ylabel('charges')  
plt.show()  
#95%的置信区间  
def ci\_t(data, confidence=0.95):  
 sample\_mean = np.mean(data)  
 sample\_std = np.std(data, ddof=1)  
 sample\_size = len(data)  
 alpha = 1 - confidence  
 t\_score = scipy.stats.t.isf(alpha / 2, df=(sample\_size - 1))  
  
 ME = t\_score \* sample\_std / np.sqrt(sample\_size)  
 lower\_limit = sample\_mean - ME  
 upper\_limit = sample\_mean + ME  
  
 return lower\_limit, upper\_limit  
  
X1=[10600.54830,2205.98080,1629.83350,2007.94500,29141.36030]  
print('95% confidence interval：')  
print(ci\_t(X1))

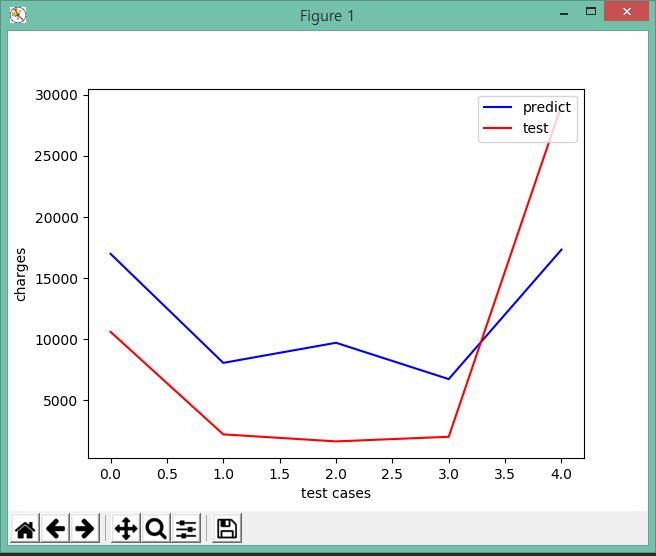
New2

import pandas as pd  
from statsmodels.formula.api import ols  
from statsmodels.stats.anova import anova\_lm  
  
#The ability to distinguish between charges and gender in data.  
data=pd.read\_csv('D:/python/data.csv')  
model=ols('charges ~ sex',data).fit()  
anovat=anova\_lm(model)  
print(anovat)  
print()  
  
#The ability to tell the difference between charges and sex and smokers.  
formula = 'charges~ sex + smoker'  
anova\_results=anova\_lm(ols(formula,data).fit())  
print(anova\_results)

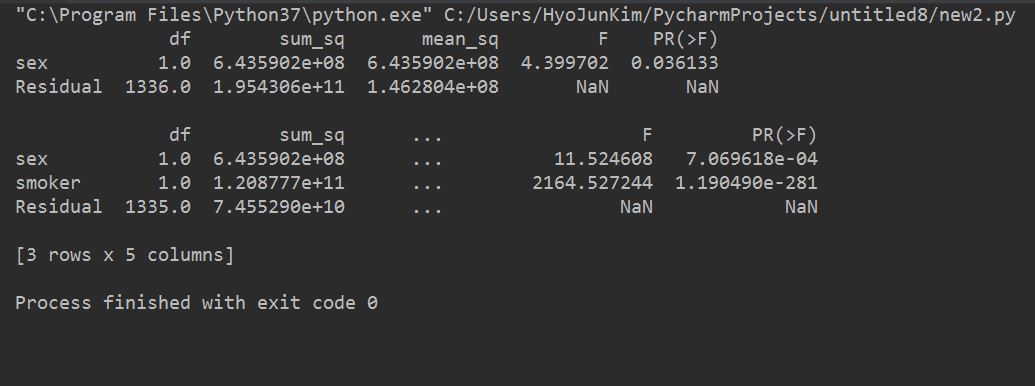
结果：

New1是回归分析结果





New2是方差分析结果



实验感想：

I wasted a lot of time reading Chinese books.

And one by one, I spent time analyzing and looking for Chinese.

It took me about a week to make and code a report.

Even while coding, I couldn't code myself, so I asked my classmates a lot.

The concepts were Regression analysis variance and analysismuch of the time needed to find Chinese books and Korean books.  
With this experiment, I wish I could understand the next experiment a little more.