作业 2 实验记录（概要）

智能科学与技术 16337174 吕高雅馨

实验目的：

利用第二章所学的知识，应用回归分析和方差分析对极其广泛的数据进行分析，揭示变量间的内在规律，并用于预测数据。

实验内容：

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。

（1）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。

（2）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）：

（1）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。

（2）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）。

实验过程：

第一题，根据对一元线性回归理解的基础，结合老师上课所讲内容，即可推广至多元线性回归（这里自变量有年龄、体质指数、孩子数量）。计算公式 ppt 上有给出：

4. 多元线性回归

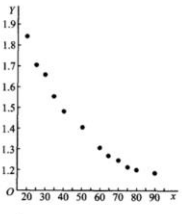
- 例 下面给出某种产品每件平均单价Y与批量x之间关系的一组数据

x	20	25	30	35	40	50	60	65	70	75	80	90
y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18

我们根据散点图取模型 $Y = b_0 + b_1 x + b_2 x^2 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

令$x_1=x$, $x_2=x^2$,则上式变为 $Y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

这是一个二元线性回归模型



$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} = \begin{pmatrix} 12 & 640 & 40\,100 \\ 640 & 40\,100 & 2\,779\,000 \\ 40\,100 & 2\,779\,000 & 204\,702\,500 \end{pmatrix}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1\,225 \\ 1 & 40 & 1\,600 \\ 1 & 50 & 2\,500 \\ 1 & 60 & 3\,600 \\ 1 & 65 & 4\,225 \\ 1 & 70 & 4\,900 \\ 1 & 75 & 5\,625 \\ 1 & 80 & 6\,400 \\ 1 & 90 & 8\,100 \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$$

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} = \frac{1}{\Delta} \begin{pmatrix} 4.857\,292\,5 \times 10^{11} & -1.957\,17 \times 10^{10} & 170\,550\,000 \\ -1.957\,17 \times 10^{10} & 848\,420\,000 & -7\,684\,000 \\ 170\,550\,000 & -7\,684\,000 & 71\,600 \end{pmatrix}$$

$$\vec{B} = (X^T X)^{-1} X^T Y = \begin{pmatrix} +2.19826629 \\ -0.02252236 \\ +0.00012507 \end{pmatrix}$$

Y=2.19826629-0.02252236x+0.00012507x^2

计算线性回归拟合方程的代码如下，beta 表示系数：

```
#读取数据
def loadDataSet():
    f  = open('data.txt','r')
    f.readline()
    for lines in f:
        readin = lines.strip('\n').split(',')
        age.append(float(readin[0]))
        bmi.append(float(readin[2]))
        children.append(float(readin[3]))
        charges.append(float(readin[6]))
    f.close()

def regr_analysis():
    xMat=[]
    for i in range(1333):
        xMat.append([1,age[i], bmi[i], children[i]])
    xMat = np.mat(np.array(xMat))
    yMat = np.mat(np.array([charges[0:1333]]).T)

    xTx = xMat.T * xMat
    beta = np.array(xTx.I * (xMat.T * yMat))
```

结合课本所给出的公式，计算置信区间：

对于给定的置信水平 α，由(2.37)式可得 Y 在 $(x_{01}, x_{02}, \cdots, x_{0,p-1})$ 处的取值 y_0 的置信度为 $1-\alpha$ 的置信区间为

$$\hat{y}_0 \pm t_{1-\frac{\alpha}{2}}(n-p)\sqrt{MSE}\sqrt{[1+x_0^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_0]} \qquad (2.38)$$

上式中的 MSE 即 σ^2，计算公式如下：

$$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{1}{n-p}Y^{\mathrm{T}}(I-H)Y \qquad (2.17)$$

为 σ^2 的无偏估计.

```
SSE = 0   #残差平方和
for i in range(1333):
    pred_y = beta[0] + beta[1]*age[i] + beta[2]*bmi[i] + beta[3]*children[i]
    SSE = SSE + (pred_y - charges[i])**2

sigma_2 = SSE/(1333-4)    #方差

con_interval = np.ones((1, 2))
t_value = 1.960

pred_y = 0
for i in range(1333, 1338):
    arr = np.array([1, age[i], bmi[i], children[i]])
    pred_y = beta[0] + beta[1]*age[i] + beta[2]*bmi[i] + beta[3]*children[i]
    #print(arr.T@np.array(xTx.I)@arr)
    temp1 = pred_y - t_value*np.sqrt(sigma_2*(1 + arr.T@np.array(xTx.I)@arr))
    temp2 = pred_y + t_value*np.sqrt(sigma_2*(1 + arr.T@np.array(xTx.I)@arr))
    con_interval[0][0] = temp1
    con_interval[0][1] = temp2
    print('预测值为', pred_y)
    print('置信区间为', con_interval)
```

对最后 5 条数据的预测结果及置信区间结果如下：

```
预测值为 [16989.31278129]
置信区间为 [[-5364.7161434  39343.34170598]]
预测值为 [8059.72578743]
置信区间为 [[-14295.69010234  30415.1416772 ]]
预测值为 [9705.11321773]
置信区间为 [[-12661.027248    32071.25368346]]
预测值为 [6730.40809095]
置信区间为 [[-15622.39926816  29083.21545006]]
预测值为 [17331.53343805]
置信区间为 [[-5027.04189257 39690.10876867]]
```

第二题，计算公式复杂，调用 statsmodels 库函数

```python
import pandas as pd
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

file = pd.read_table('data.txt', sep = ',')
data = file.drop('age', 1)
data = data.drop('bmi', 1)
data = data.drop('children', 1)
data0 = data.drop('region', 1)
data1 = data0.drop('smoker', 1)

def var_analysis(data, model):
    anova_results = anova_lm(ols(model, data).fit())
    print(anova_results)

#不同性别对个人医疗费用是否有显著差异
var_analysis(data, 'charges ~ C(sex)')

#不同性别、是否吸烟两个因素对个人医疗费用是否有显著差异
var_analysis(data0, 'charges ~ sex + smoker')
```

结果如下：

```
                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
                df        sum_sq       mean_sq            F         PR(>F)
sex            1.0  6.435902e+08  6.435902e+08    11.524608   7.069618e-04
smoker         1.0  1.208777e+11  1.208777e+11  2164.527244  1.190490e-281
Residual    1335.0  7.455290e+10  5.584487e+07          NaN            NaN
```

通过最后一列 PR（p 值）判断，性别和是否抽烟都对医疗费用有显著性差异。