16337152 林子杭 智能科学与技术

1、回归分析

问题

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。

代码及分析

首先，读入数据，仅读取年龄，体质指数，孩子个数这几个定量变量及医疗费用。前1333条数据用于建立线性回归模型，后5条数据用于测试。

```
import pandas as pd
import numpy as np
data_frame = pd.read_csv('data.txt', usecols=('age','bmi','children','charges'),
delimiter=',')
data = np.array(data_frame)
x_train, y_train = data[:-5, :-1], data[:-5, -1]
x_test, y_test = data[-5:, :-1], data[-5:, -1]
```

由公式$B=(X^TX)^{-1}X^TY$计算得到系数$B=(203.331,165.513,410.409)^T$，后五条数据的预测值分别为16523.70，8943.12，9759.10，8540.18，17214.64.

```
B = np.linalg.inv(x_train.T@x_train)@x_train.T@y_train
y_predict = x_test @ B
print(B, y_predict)
```

置信度为95%的置信区间为$\hat{y}\pm t_{0.025}(1333-3)\sqrt{MSE[1+x^T(X^TX)^{-1}x]}$

其中$MSE=\frac{SSE}{1333-3}=\frac{\sum_{i=1}^{1333}(\hat{y}-y)^2}{1330}$

实现如下：

```
from scipy import stats
n, p = x_train.shape
sse = np.sum(np.square(x_train @ B - y_train))
mse = sse / (n-p)
t = stats.t.interval(0.95, n-p)
for i in range(5):
    x = np.reshape(x_test[i], [-1, 1])
    tmp = np.sqrt(mse * (1 + x.T @ np.linalg.inv(x_train.T @ x_train) @ x))
    print(y_predict[i]+t[0]*tmp, y_predict[i]+t[1]*tmp)
```

5个置信区间计算结果如下：

```
(-5959.95452959, 39007.35809095)
(-13538.72905259, 31424.97186025)
(-12737.95470218, 32256.15224552)
(-13924.87208644, 31005.22591452)
(-5274.74063276, 39704.02539265)
```

观察结果可以发现，置信度为95%的置信区间跨度很大，几乎没有什么参考价值，这是因为医疗费用和年龄、体质指数、孩子个数这几个指标的线性关系不强，因而使用线性回归模型并不能得到很好的预测结果。

2、方差分析

问题

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）。

代码及分析

调用python中statsmodels包中的anova_lm函数可以方便地进行方差分析，代码实现如下：

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
data = pd.read_csv('data.txt', usecols=('sex', 'smoker','charges'), delimiter=',')
formula = 'charges~C(sex)'
anova_results = anova_lm(ols(formula, data).fit())
print(anova_results)
formula = 'charges~C(sex)+C(smoker)+C(sex):C(smoker)'
anova_results = anova_lm(ols(formula, data).fit())
print(anova_results)
```

输出结果如下：

```
               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  \
C(sex)              1.0  6.435902e+08  6.435902e+08    11.592531
C(smoker)           1.0  1.208777e+11  1.208777e+11  2177.284440
C(sex):C(smoker)    1.0  4.923397e+08  4.923397e+08     8.868165
Residual         1334.0  7.406056e+10  5.551766e+07          NaN

                        PR(>F)
C(sex)              6.818323e-04
C(smoker)          1.247285e-282
C(sex):C(smoker)    2.954255e-03
Residual                     NaN
```

结果分析：由输出的结果可以看到，对于性别这单个因素，检验假设H_0：”不同性别对个人医疗费用无显著差异“的p值为0.036<0.05，故拒绝原假设H_0，即认为性别对个人医疗费用有显著差异。对于性别和是否吸烟这两个因素，三个检验p值分别为$6.818 \times 10^{-4} < 0.05$,$1.247 \times 10^{-282} < 0.05$,$2.954 \times 10^{-3} < 0.05$.这说明：不同性别的个体的医疗费用存在显著差异，吸烟的个体和不吸烟的个体的医疗费用也存在显著关系。对于不同性别，是否吸烟对医疗费用也有显著的影响。