统计分析方法第二次作业

16337207 石恬 智能科学与技术

一、 回归分析

（一）线性回归拟合

1. 题解

本题属于多元回归分析，即随机变量 charges，对应多个普通变量 age， bmi， children。利用多元回归分析解题。

2. 原理

利用最大似然估计来估计参数

即取 $\hat{b}_0, \hat{b}_1, \cdots, \hat{b}_p$ **使当** $b_0 = \hat{b}_0, b_1 = \hat{b}_1, \cdots, b_p = \hat{b}_p$ **时**

$$Q = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_p x_{ip})^2$$

达到最小.

在这里设

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{B} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}.$$

将 Q 分别对$b_0, b_1 \dots b_p$求偏导并使结果等于 0

最后式子可以写成： $$\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{B} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}$$

即： $$\hat{\boldsymbol{B}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{pmatrix} = (\boldsymbol{X}^{\mathrm{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}$$

3. 结果

B=[-6872.9670633

237.74407233

333.74998586

546.27972192]

（二）测试并给出置信区间

1. 题解

将 data 中的最后 5 条作为测试数据，检验预测效果

对每一个结果得出置信度为 95%的置信区间。

2. 原理

$$\hat{Y}_0 = \mathbf{X}_0\hat{\mathbf{B}}$$

由此可得预测结果

置信区间：

预测误差为：
$$\begin{aligned} e_0 &= Y_0 - \hat{Y}_0 \\ &= \mathbf{X}_0\mathbf{B} + u_0 - \mathbf{X}_0\hat{\mathbf{B}} \\ &= \mathbf{X}_0(\mathbf{B} - \hat{\mathbf{B}}) + u_0 \end{aligned}$$

$$\begin{aligned} E(e_0) &= E(\mathbf{X}_0(\mathbf{B} - \hat{\mathbf{B}}) + u_0) \\ &= \mathbf{X}_0 E(\mathbf{B} - \hat{\mathbf{B}}) + E(u_0) \\ &= 0 \end{aligned}$$

又由 $\hat{\mathbf{B}} = \mathbf{B} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}$

可得 $e_0 = u_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}$

$$\begin{aligned} \mathrm{var}(e_0) &= E\left((e_0 - E(e_0))^2\right) = E(e_0^2) \\ &= E\left(\boxed{u_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}}\right)^2 \\ &= E(u_0 - \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U})(u_0 - \mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0') \\ &= E(u_0^2 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0' \\ &\quad - 2\boxed{u_0}\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{U}) \\ &= \sigma_u^2(1 + \mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0') \end{aligned}$$

标量

$\mathrm{cov}(u_0, \mathbf{U}) = E(u_0\mathbf{U}) = 0$

用S^2代替σ_u^2得到

$$\hat{\operatorname{var}}(Y_0-\hat{Y}_0)=S^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0')$$

由于 $Y_0-\hat{Y}_0 \sim N(0, \sigma_u^2(1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'))$

所以 $\dfrac{Y_0-\hat{Y}_0}{S\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}} \sim t(n-k)$

记 $se(e_0)=S\sqrt{1+\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'}$

在给定了置信度$(1-\alpha)$之后，Y_0的$(1-\alpha)$置信区间为

$$Y_0\in\left(\hat{Y}_0-t_{\frac{\alpha}{2}}\cdot se(e_0),\ \hat{Y}_0+t_{\frac{\alpha}{2}}\cdot se(e_0)\right)$$

```
Q=np.sum([i * i for i in Ytrain-np.dot(Xtrain,B)])
sigma=np.sqrt(Q/(Xtrain.shape[0]-2))
se = sigma * math.sqrt(1+np.dot(np.dot(Xtest[i],
     np.linalg.inv(np.dot(Xtrain.T,Xtrain))),Xtest[i].T))
```

3. 结果

Y = [16989.31278129 8059.72578743 9705.11321773 6730.40809095 17331.53343805]

置信区间：

```
[[ -5795.19981456  39773.82537714]
 [-14726.20048304  30845.6520579 ]
 [-13091.74415768  32501.97059314]
 [-16052.85941497  29513.67559687]
 [ -5457.61311633  40120.67999244]]
```

将结果与原数据比较后发现，线性回归拟合的效果比较差，误差较大。只能预测准确方向，但是大小还有较大偏差。置信区间的大小也说明拟合的误差较大。

二、 方差分析

（一）单因素方差分析

1. 设性别为 male 的医疗费用均值为μ_1，性别为 female 的医疗费用均值为μ_2

假设：H0：$\mu_1 = \mu_2$ 即不同性别对个人医疗费用无显著差异

H1：$\mu_1 \neq \mu_2$ 即不同性别对个人医疗费用有显著差异

计算目标：

表 9－5 单因素试验方差分析表

方差来源	平方和	自由度	均方	F 比
因素 A	S_A	$s-1$	$\overline{S}_A = \frac{S_A}{s-1}$	$F = \frac{\overline{S}_A}{\overline{S}_E}$
误差	S_E	$n-s$	$\overline{S}_E = \frac{S_E}{n-s}$	
总和	S_T	$n-1$		

其中 $$S_E = \sum_{j=1}^{s}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_{\cdot j})^2,$$

$$S_A = \sum_{j=1}^{s}\sum_{i=1}^{n_j}(\overline{X}_{\cdot j} - \overline{X})^2 = \sum_{j=1}^{s} n_j(\overline{X}_{\cdot j} - \overline{X})^2 = \sum_{j=1}^{s} n_j\overline{X}_{\cdot j}^2 - n\overline{X}^2.$$

若 F<$F_{0.05}(1,\infty) = 3.84$ 则接受 H0，否则拒绝 H0

2. 实现

调用了 python 中的 statsmodels 库

```
model = ols('charges ~ sex', df).fit()
anovat = anova_lm(model)
print(anovat)
```

3. 结果

	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

明显可以看出 F=4.399702 > $F_{0.05}(1,\infty) = 3.84$ ，所以拒绝 H0，则认为不同性别对个人医疗费用有显著差异

（二）双因素等重复试验方差分析

1. 因素 A 性别具有 male 和 female 两个水平，因素 B 是否吸烟具

有 yes 和 no 两个水平，共进行了 1338 次试验

设性别为 male 不吸烟的医疗费用均值为μ_{11}，性别为 male 吸烟的医疗费用均值为μ_{12}，性别为 female 不吸烟的医疗费用均值为μ_{21}，性别为 female 吸烟的医疗费用均值为μ_{22}

进行三个假设：

$$\begin{cases} H_{01}: & \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \\ H_{11}: & \alpha_1, \alpha_2, \cdots, \alpha_r \text{ 不全为零}, \end{cases}$$

$$\begin{cases} H_{02}: & \beta_1 = \beta_2 = \cdots = \beta_s = 0, \\ H_{12}: & \beta_1, \beta_2, \cdots, \beta_s \text{ 不全为零}, \end{cases}$$

$$\begin{cases} H_{03}: & \gamma_{11} = \gamma_{12} = \cdots = \gamma_{rs} = 0, \\ H_{13}: & \gamma_{11}, \gamma_{12}, \cdots, \gamma_{rs} \text{ 不全为零}. \end{cases}$$

计算目标：

$$S_E = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{t}(X_{ijk} - \overline{X}_{ij\cdot})^2,$$

$$S_A = st\sum_{i=1}^{r}(\overline{X}_{i\cdot\cdot} - \overline{X})^2,$$

$$S_B = rt\sum_{j=1}^{s}(\overline{X}_{\cdot j\cdot} - \overline{X})^2,$$

$$S_{A\times B} = t\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} + \overline{X})^2.$$

当 $H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ 为真时，可以证明

$$F_A = \frac{S_A/(r-1)}{S_E/(rs(t-1))} \sim F(r-1, rs(t-1)).$$

取显著性水平为 α，得假设 H_{01} 的拒绝域为

$$F_A = \frac{S_A/(r-1)}{S_E/(rs(t-1))} \geqslant F_\alpha(r-1, rs(t-1)).$$

类似地，在显著性水平 α 下，假设 H_{02} 的拒绝域为

$$F_B = \frac{S_B/(s-1)}{S_E/(rs(t-1))} \geqslant F_\alpha(s-1, rs(t-1)).$$

在显著性水平 α 下，假设 H_{03} 的拒绝域为

$$F_{A\times B} = \frac{S_{A\times B}/((r-1)(s-1))}{S_E/(rs(t-1))} \geqslant F_\alpha((r-1)(s-1), rs(t-1)).$$

$$F_{0.05}(r-1, rs(t-1)) = F_{0.05}(s-1, rs(t-1))$$

$$= F_{0.05}((r-1)(s-1), rs(t-1)) = F_{0.05}(1, \infty)$$

$$= 3.84$$

若$F_A < F_{0.05}(1,\infty)$则接受 H01，否则拒绝 H11

若$F_B < F_{0.05}(1,\infty)$则接受 H02，否则拒绝 H12

若$F_A < F_{0.05}(1,\infty)$则接受 H03，否则拒绝 H13

2. 实现

调用了 python 中的 statsmodels 库

```
formula = 'charges ~ sex + smoker + sex:smoker'
anova_results = anova_lm(ols(formula, df).fit())
anova_results.to_csv('anova_result.csv')
print(anova_results)
```

3. 结果

结果存在 anova_result.csv 中

	df	sum_sq	mean_sq	F	PR(>F)
sex	1	6.44E+08	6.44E+08	11.59253	0.000682
smoker	1	1.21E+11	1.21E+11	2177.284	1.25E-282
sex:smoker	1	4.92E+08	4.92E+08	8.868165	0.002954
Residual	1334	7.41E+10	55517659		

由数据我们可以看出：

$$F_A > F_{0.05}(1,\infty)$$

$$F_B > F_{0.05}(1,\infty)$$

$$F_{AXB} > F_{0.05}(1,\infty)$$

因此，我们认为性别和是否吸烟都对个人医疗费用有显著差异。

三、 附录

```
import numpy as np
import math
from scipy import stats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
import pandas as pd
```

```
df=pd.read_csv('data.txt',header=0)
df=df.dropna()
age=np.array(df['age'])
bmi=np.array(df['bmi'])
children=np.array(df['children'])
Y=np.array(df['charges']).astype(float)

X=np.zeros((1338,1))
X=age
X=np.column_stack((X,bmi))
X=np.column_stack((X,children))

def problem1(X1,Y1):
    Xtrain=X1[0:1333]
    a = np.ones((Xtrain.shape[0],1))
    Xtrain=np.column_stack((a,Xtrain))
    Xtest=X1[1333:]
    b = np.ones((Xtest.shape[0], 1))
    Xtest = np.column_stack((b, Xtest))
    Ytrain=Y1[0:1333]


    B=np.zeros((4,1))

B=np.dot(np.dot(np.linalg.inv(np.dot(Xtrain.T,Xtrain)),Xtrain.T),
Ytrain)
    print(B)

    Ytest=np.dot(Xtest,B)
    print(Ytest)

    Q=np.sum([i * i for i in Ytrain-np.dot(Xtrain,B)])
    sigma=np.sqrt(Q/(Xtrain.shape[0]-2))

    #confidence interval
    result=np.zeros((5,2))
    for i in range(5):
        se = sigma *
math.sqrt(1+np.dot(np.dot(Xtest[i],np.linalg.inv(np.dot(Xtrain.T,
Xtrain))),Xtest[i].T))
        result[i,0]=Ytest[i]-2*se
        result[i,1]=Ytest[i]+2*se
    print(result)
```

```python
def problem2():
    model = ols('charges ~ sex', df).fit()
    anovat = anova_lm(model)
    print(anovat)

    formula = 'charges ~ sex + smoker + sex:smoker'
    anova_results = anova_lm(ols(formula, df).fit())
    anova_results.to_csv('anova_result.csv')
    print(anova_results)

if __name__=='__main__':
    problem1(X,Y)
    problem2()
```