Logistic 回归

王梦圆

2020-02

0.0.1 选择 Zelig 包里面的 turnout 数据集，这个数据集是为了确定投票率与被选举人的种族 (race)、年龄 (age)、受教育程度 (educate) 和收入 (income) 是否有关。因变量 vote 是 0-1 型变量，赞成为 1，反对为 0。种族 (race) 变量是一个分类型自变量，在进行回归分析时，可以将 “white” 记为 0，“others” 记为 1. 下面对该数据集进行二分类 Logistc 回归分析。

0.0.1.1 数据准备和模型建立

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.model_selection import train_test_split
df = sm.datasets.get_rdataset("turnout",package="Zelig",site="D:/github_
    repo/Rdatasets").data
```

```
df

#对数据进行处理：去空值，处理分类变量race
```

```
df.isnull().sum()#没有空值
```

```
df['race']=df['race'].replace("white",0)
df['race']=df['race'].replace("others",1)
df['race'].unique()
```

```
df.index=np.arange(df.shape[0])
X=df.iloc[:,:4]
X
```

```
Y=df['vote']
p=Y.sum()/len(Y)#投票比率为0.746

#将数据集随机划分为训练子集和测试子集，并返回划分好的样本和标签
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.2,random_
    state=0)
#x_train训练集特征值
#y_train训练集目标值
#x_test测试集特征值
#y_text测试集目标值，真实值

#拟合logistic回归方程
results=sm.Logit(y_train,X_train).fit()
```

```
print(results.summary())
```

Logistic 回归方程：

$$\log \frac{p}{1-p} = -0.5874 * \text{ race } + 0.0063 * \text{ age } + 0.0367 * \text{ educate } + 0.1559 * \text{ income}$$

根据输出的结果显示，在显著性水平为 0.05 下，四个变量的 P 值均小于 0.05，即四个自变量种族、年龄、受教育程度和收入都对是否决定投票都有显著影响。

0.0.1.2 模型准确率

```
#预测数据
y_predict = results.predict(X_test)
y_predict
```

```
y_predict = np.where(y_predict>0.5,1,0)
accuracy = (y_predict==y_test).sum()/len(y_test)
accuracy
```

以 0.5 作为阈值，预测准确率为 0.74，即用种族、年龄、受教育程度和收入这四个变量估计投票的概率是 74%，因为除了被选举人自身的优势之外，一些政治因素也是造成预测准确率不是很高的因素。

0.0.2 利用 KMsurv 包里面的 aids 数据集，利用 Logistic 回归分析法调查二分类变量 adult 与 infect、induct 的关系。infect 和 induct 变量都为连续型变量。

0.0.2.1 数据准备和建模

```

dat = sm.datasets.get_rdataset("aids",package="KMsurv",site="D:/github_
    repo/Rdatasets").data
dat

#对数据进行处理：去空值，处理分类变量race
```

```
dat.isnull().sum()#没有空值
```

```
dat.index=np.arange(dat.shape[0])
X=dat.iloc[:,:2]
Y=dat['adult']
p=Y.sum()/len(Y)

#将数据集随机划分为训练子集和测试子集，并返回划分好的样本和标签
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.2,random_
    state=0)

#拟合logistic回归方程
results=sm.Logit(y_train,X_train).fit()
```

```
print(results.summary())
```

回归方程：

$$\log \frac{p}{1-p} = 0.1037 * \text{ infect } + 0.7287 * \text{ induct}$$

又因为 infect 变量的 P 值为 0.103，在显著性水平 0.05 下，infect 变量对因变量 adult 率影响不显著，因此，除去 infect 变量后再次进行 Logistic 回归分析.

```
X_new=dat['induct']
X_train,X_test,y_train,y_test=train_test_split(X_new,Y,test_size=0.2,
    random_state=0)
results=sm.Logit(y_train,X_train).fit()
```

```
print(results.summary())
```

回归方程：

$$\log \frac{p}{1-p} = 0.9269 * \text{ induct}$$

变量 induct 的 P 值为 0，说明变量 induct 的影响是显著的。#### 模型准确率预测

```
predict = results.predict(X_test)
y_predict2 = np.where(predict>0.5,1,0)
accuracy2=(y_test==y_predict2).sum()/len(y_test)
accuracy2#0.847
```