lab6

May 10, 2023

[1]:
```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
```

0.1 1. Logistic Regression

1.1. load file titanic.csv, and do quick sanity checks.

[2]:
```
titanic = pd.read_csv('titanic.csv.bz2', sep=",")
print(f"Rows, Columns: {titanic.shape}\n")
print("Number of NaN values in each column")
print(titanic.isna().sum())
print("\nSample")
print(titanic.sample(3))
```

```
Rows, Columns: (1309, 14)

Number of NaN values in each column
pclass          0
survived        0
name            0
sex             0
age           263
sibsp           0
parch           0
ticket          0
fare            1
cabin        1014
embarked        2
boat          823
body         1188
home.dest     564
dtype: int64

Sample
      pclass  survived                             name     sex   age  sibsp  \
1058       3         0  Nieminen, Miss. Manta Josefina  female  29.0      0
345        2         0      Berriman, Mr. William John    male  23.0      0
1048       3         1     Nakid, Miss. Maria ("Mary")  female   1.0      0
```

```
      parch   ticket     fare cabin embarked boat  body  \
1058      0  3101297   7.9250   NaN        S  NaN   NaN
345       0    28425  13.0000   NaN        S  NaN   NaN
1048      2     2653  15.7417   NaN        C    C   NaN

                           home.dest
1058                             NaN
345   St Ives, Cornwall / Calumet, MI
1048                             NaN
```

Data looks fine, age as 263 missing cabin has 101 boat has 823 body has 1188 and home.dest has 564

1.2. Based on the survivors' accounts, described above, which variables do you think are the most important ones to describe titanic survival? It would be class, sex, and age, because women and children of the 1st and 2nd class got to get on the lifeboat first.

1.3. Create a new dummy variable child, that is 1 if the passenger was youger than 14 and 0 otherwise.

[3]:
```
child = np.where(titanic.age < 14, 1, 0)
titanic['child']=child
```

1.4. Estimate a multiple logistic regression model where you explain survival by these variables.

[4]:
```
m = smf.ols("survived ~ C(pclass) + C(sex) + C(child)", data = titanic).fit()
m.summary()
```

[4]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:               survived   R-squared:                       0.354
Model:                            OLS   Adj. R-squared:                  0.352
Method:                 Least Squares   F-statistic:                     178.5
Date:                Wed, 10 May 2023   Prob (F-statistic):          4.92e-122
Time:                        06:02:49   Log-Likelihood:                -626.67
No. Observations:                1309   AIC:                             1263.
Df Residuals:                    1304   BIC:                             1289.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
==
                     coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--
```

```
Intercept          0.8900      0.025     35.282      0.000       0.841
0.940
C(pclass)[T.2]    -0.1732      0.032     -5.374      0.000      -0.236
-0.110
C(pclass)[T.3]    -0.3100      0.027    -11.615      0.000      -0.362
-0.258
C(sex)[T.male]    -0.4957      0.023    -21.704      0.000      -0.541
-0.451
C(child)[T.1]      0.2077      0.041      5.024      0.000       0.127
0.289
==============================================================================
Omnibus:                       48.471   Durbin-Watson:                   1.754
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.276
Skew:                           0.494   Prob(JB):                     2.70e-12
Kurtosis:                       3.021   Cond. No.                         5.44
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

1.5. Interpret the results. Did men or women, old or young have larger chances to survive? Based on the results, those who were male and not in the first class has the least chance of survival. While female children in the first class has the highest chance of survival.

1.6. 6. Based on the results above, explain what can you tell about the last hours on Titanic. Are the survivors' accounts broadly accurate? Did the order break down? Can you find anything else interesting? The order didn't break down, mostly women and children and first class got on the life boats which means the accounts were accurate. Interesting is a female children in the first class has over a 100% chance of making it on the life boat.