Lab06

May 9, 2023

1 Lab 06

1.1 Xinyu Chang

[2]:

```
# import the packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
```

1.1.1 1 Logistic Regression

1. load file titanic.csv, and do quick sanity checks. Note that age has 260 missing observations, the other relevant variables are pretty good.

[3]:

```
titanic = pd.read_csv("titanic.csv", sep=",")
titanic.head()
```

[3]:

```
   pclass  survived                                             name     sex  \
0       1         1                     Allen, Miss. Elisabeth Walton  female
1       1         1                    Allison, Master. Hudson Trevor    male
2       1         0                      Allison, Miss. Helen Loraine  female
3       1         0              Allison, Mr. Hudson Joshua Creighton    male
4       1         0   Allison, Mrs. Hudson J C (Bessie Waldo Daniels)  female

       age  sibsp  parch  ticket      fare    cabin embarked boat   body  \
0  29.0000      0      0   24160  211.3375       B5        S    2    NaN
1   0.9167      1      2  113781  151.5500  C22 C26        S   11    NaN
2   2.0000      1      2  113781  151.5500  C22 C26        S  NaN    NaN
3  30.0000      1      2  113781  151.5500  C22 C26        S  NaN  135.0
4  25.0000      1      2  113781  151.5500  C22 C26        S  NaN    NaN

                         home.dest
0                     St Louis, MO
1  Montreal, PQ / Chesterville, ON
2  Montreal, PQ / Chesterville, ON
3  Montreal, PQ / Chesterville, ON
4  Montreal, PQ / Chesterville, ON
```

[4]:
```
titanic.shape
```

[4]:
```
(1309, 14)
```

[5]:
```
titanic.isna().sum()
```

[5]:
```
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
```

There are 1309 rows and 14 columns in the dataset. The first 5 lines of data look reasonable and the data fits the type that the column defines. However, there are some missing values for the age, cabin, boat, body, and home.dest.

2. Based on the survivors' accounts, described above, which variables do you think are the most important ones to describe titanic survival? Based on survivors' accounts and general understanding of the event, the most important variables could be: pclass, sex, age.

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

[6]:
```
titanic['child'] = (titanic['age'] < 14).astype(int)
```

4. Estimate a multiple logistic regression model where you explain survival by these variables.Remember: pclass should be treated as categorical! Wrap it into C()

[7]:
```
m = smf.logit('survived ~ C(pclass) + sex + age + child', data=titanic).fit()
m.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.468021
         Iterations 6
```

[7]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                           Logit Regression Results
```

```
==============================================================================
Dep. Variable:               survived   No. Observations:                 1046
Model:                          Logit   Df Residuals:                     1040
Method:                           MLE   Df Model:                            5
Date:                Tue, 09 May 2023   Pseudo R-squ.:                  0.3079
Time:                        00:42:04   Log-Likelihood:                -489.55
converged:                       True   LL-Null:                       -707.31
Covariance Type:            nonrobust   LLR p-value:                 6.522e-92
================================================================================
==
                     coef    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
--
Intercept          3.1956      0.367      8.709      0.000       2.476
3.915
C(pclass)[T.2]    -1.2295      0.226     -5.437      0.000      -1.673
-0.786
C(pclass)[T.3]    -2.2406      0.226     -9.904      0.000      -2.684
-1.797
sex[T.male]       -2.5016      0.167    -15.014      0.000      -2.828
-2.175
age               -0.0266      0.008     -3.521      0.000      -0.041
-0.012
child              0.5858      0.320      1.828      0.068      -0.042
1.214
================================================================================
==
"""
```

(1)Pclass: The coefficients for C(pclass)[T.2] (-1.2295) and C(pclass)[T.3] (-2.2406) represent the difference in log-odds of survival for passengers in class 2 and class 3, respectively, compared to the reference class (class 1). Both coefficients are negative, indicating that passengers in class 2 and class 3 had lower chances of survival than those in class 1. The effect is stronger for class 3 passengers, as its coefficient is more negative.

(2)Sex: The coefficient for sex[T.male] (-2.5016) indicates that the log-odds of survival for males were lower than for females (reference category). This negative coefficient suggests that men had lower chances of survival compared to women.

(3)Age: The coefficient for age (-0.0266) shows that the log-odds of survival decrease as age increases. This negative coefficient implies that older passengers had lower chances of survival compared to younger passengers.

(4)Child: The coefficient for child (0.5858) represents the difference in log-odds of survival for children (age < 18) compared to non-children (reference category). The coefficient is positive but not statistically significant at the 0.05 level (p-value = 0.068), suggesting that there may not be a strong relationship between being a child and survival chances.

5. Interpret the results. Did men or women, old or young have larger chances to survive

```
[8]: m.get_margeff().summary()
```

```
[8]: <class 'statsmodels.iolib.summary.Summary'>
"""
        Logit Marginal Effects
=====================================
Dep. Variable:               survived
Method:                          dydx
At:                           overall
================================================================================
==
                    dy/dx    std err          z      P>|z|      [0.025
0.975]
--------------------------------------------------------------------------------
--
C(pclass)[T.2]    -0.1855      0.033     -5.679      0.000      -0.250
-0.121
C(pclass)[T.3]    -0.3380      0.029    -11.560      0.000      -0.395
-0.281
sex[T.male]       -0.3774      0.013    -29.342      0.000      -0.403
-0.352
age               -0.0040      0.001     -3.586      0.000      -0.006
-0.002
child              0.0884      0.048      1.836      0.066      -0.006
0.183
================================================================================
==
"""
```

In conclusion, based on the logistic regression results, women had larger chances of survival compared to men, and younger passengers had larger chances of survival compared to older passengers. Passengers in higher classes (class 1) also had better chances of survival than those in lower classes (classes 2 and 3). The relationship between being a child and survival chances is not statistically significant at the 0.05 level.

C(pclass)[T.2]: This represents the marginal effect of being in passenger class 2 (second class) compared to the reference category (passenger class 1). A negative effect (-0.1855) indicates that being in the second class reduces the probability of survival by 18.55% pt compared to first class, holding all other variables constant. This effect is statistically significant at the 0.1% level (p-value < 0.001).

C(pclass)[T.3]: This represents the marginal effect of being in passenger class 3 (third class) compared to the reference category (passenger class 1). A negative effect (-0.3380) indicates that being in the third class reduces the probability of survival by 33.80% pt compared to first class, holding all other variables constant. This effect is statistically significant at the 0.1% level (p-value < 0.001).

sex[T.male]: This represents the marginal effect of being male compared to the reference category (female). A negative effect (-0.3774) indicates that being male reduces the probability of survival by 37.74% pt compared to being female, holding all other variables constant. This effect is statistically significant at the 0.1% level (p-value < 0.001).

age: This represents the marginal effect of age on the probability of survival. A negative effect (-0.0040) indicates that a one-year increase in age reduces the probability of survival by 0.40% pt, holding all other variables constant. This effect is statistically significant at the 0.1% level (p-value < 0.001).

child: This represents the marginal effect of being a child on the probability of survival. A positive effect (0.0884) indicates that being a child increases the probability of survival by 8.84% compared to not being a child, holding all other variables constant. However, this effect is not statistically significant at the 5% level (p-value = 0.066), but it is significant at the 10% level.

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? While it is not possible to reconstruct the exact events of the last hours on the Titanic based solely on the logistic regression results, we can infer some general patterns that are in line with the survivors' accounts and historical records.

(1)Women and children first: The logistic regression results show that women had higher chances of survival compared to men. This is consistent with the widely reported policy of "women and children first" during the evacuation. However, the results regarding children are not statistically significant, which may suggest that the policy may not have been implemented consistently across all passengers.

(2)Socio-economic status: The results show that passengers in higher classes (class 1) had better chances of survival than those in lower classes (classes 2 and 3). This could indicate that there was a certain level of social stratification during the evacuation, where wealthier passengers were given priority or had easier access to lifeboats.

(3)Age factor: The analysis indicates that younger passengers had higher chances of survival compared to older passengers. This could be due to a variety of reasons, including younger individuals being more physically able to reach lifeboats or withstand the harsh conditions after the sinking.

(4)Order and breakdown: The results do not directly address whether the order broke down during the last hours on the Titanic. However, the fact that certain groups (e.g., women and passengers in higher classes) had higher survival chances might suggest that there was some level of organization during the evacuation process.