Logistic Regression

Langtao Chen

Nov 7, 2017

Contents

```r
# Clean the environment
rm(list = ls())
```

1. Read in Data

In this example, we'll use the titanic dataset to demonstrate how to conduct logistic regression analysis.

```r
# Read data
df <- read.csv("titanic.csv", na.strings=c(""),
               stringsAsFactors = FALSE)
str(df)
```

```
## 'data.frame':    1309 obs. of  7 variables:
##  $ pclass  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
##  $ sex     : chr  "female" "male" "female" "male" ...
##  $ age     : num  29 0.917 2 30 25 ...
##  $ sibsp   : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch   : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ fare    : num  211 152 152 152 152 ...
```

The titanic dataset contains 1309 passengers.

- pclass: ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- sibsp: the number of siblings and/or spouses aboard
- parsc: the number of parents and/or children aboard
- fare: passenger fare

```r
head(df)
```

```
##   pclass survived    sex     age sibsp parch     fare
## 1      1        1 female 29.0000     0     0 211.3375
## 2      1        1   male  0.9167     1     2 151.5500
## 3      1        0 female  2.0000     1     2 151.5500
## 4      1        0   male 30.0000     1     2 151.5500
## 5      1        0 female 25.0000     1     2 151.5500
## 6      1        1   male 48.0000     0     0  26.5500
```

```r
# Remove the index column
df$Index <- NULL
str(df)
```

```
## 'data.frame':    1309 obs. of  7 variables:
##  $ pclass  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
##  $ sex     : chr  "female" "male" "female" "male" ...
##  $ age     : num  29 0.917 2 30 25 ...
##  $ sibsp   : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch   : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ fare    : num  211 152 152 152 152 ...
```

2. Explore Data

```
# Frequency of survival
table(df$survived)
```

```
##
##   0   1
## 809 500
```

Among the 1309 passengers, 500 survived.

```
# Summary statistics
summary(df)
```

```
##      pclass          survived         sex                 age
##  Min.   :1.000   Min.   :0.000   Length:1309        Min.   : 0.1667
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   1st Qu.:21.0000
##  Median :3.000   Median :0.000   Mode  :character   Median :28.0000
##  Mean   :2.295   Mean   :0.382                      Mean   :29.8811
##  3rd Qu.:3.000   3rd Qu.:1.000                      3rd Qu.:39.0000
##  Max.   :3.000   Max.   :1.000                      Max.   :80.0000
##                                                     NA's   :263
##      sibsp            parch            fare
##  Min.   :0.0000   Min.   :0.000   Min.   :  0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:  7.896
##  Median :0.0000   Median :0.000   Median : 14.454
##  Mean   :0.4989   Mean   :0.385   Mean   : 33.295
##  3rd Qu.:1.0000   3rd Qu.:0.000   3rd Qu.: 31.275
##  Max.   :8.0000   Max.   :9.000   Max.   :512.329
##                                   NA's   :1
```

We notice that there are some missing values. Here we simply remove missing values.

```
# Remove missing data
df <- na.omit(df)
summary(df)
```

```
##      pclass          survived          sex                 age
##  Min.   :1.000   Min.   :0.0000   Length:1045        Min.   : 0.1667
##  1st Qu.:1.000   1st Qu.:0.0000   Class :character   1st Qu.:21.0000
##  Median :2.000   Median :0.0000   Mode  :character   Median :28.0000
##  Mean   :2.207   Mean   :0.4086                      Mean   :29.8518
##  3rd Qu.:3.000   3rd Qu.:1.0000                      3rd Qu.:39.0000
##  Max.   :3.000   Max.   :1.0000                      Max.   :80.0000
##      sibsp            parch             fare
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  8.05
##  Median :0.0000   Median :0.0000   Median : 15.75
##  Mean   :0.5033   Mean   :0.4211   Mean   : 36.69
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 35.50
##  Max.   :8.0000   Max.   :6.0000   Max.   :512.33
```

3. Logistic Regression

With the visualization of linear relationship between shoe size and height, now let's formally use linear regression model to analyze this linear relationship.

```
# Regress survived on all other variables
model <- glm(survived ~ factor(pclass) + factor(sex) + age + sibsp + parch + fare,
             family=binomial(link='logit'),data=df)

summary(model)
```

```
##
## Call:
## glm(formula = survived ~ factor(pclass) + factor(sex) + age +
##     sibsp + parch + fare, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7163  -0.6638  -0.4221   0.6654   2.5220
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.800025   0.397340   9.564  < 2e-16 ***
## factor(pclass)2 -1.288689   0.260462  -4.948 7.51e-07 ***
## factor(pclass)3 -2.257549   0.271905  -8.303  < 2e-16 ***
## factor(sex)male -2.551596   0.173527 -14.704  < 2e-16 ***
## age             -0.039225   0.006645  -5.903 3.58e-09 ***
## sibsp           -0.358850   0.105897  -3.389 0.000702 ***
## parch            0.058585   0.102984   0.569 0.569443
## fare             0.001214   0.001942   0.625 0.531799
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1413.57  on 1044  degrees of freedom
## Residual deviance:  969.65  on 1037  degrees of freedom
## AIC: 985.65
##
## Number of Fisher Scoring iterations: 4
```

Let's use the stargazer package to report regression results in a more professional way.

```
# install.packages("stargazer") #Install stargazer package, do this only once
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

```
stargazer(model, type = "text",star.cutoffs = c(0.05, 0.01, 0.001),
          title="Logistic Regression", digits=4)
```

```
##
## Logistic Regression
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                survived
## ------------------------------------------------
## factor(pclass)2               -1.2887***
##                                (0.2605)
##
## factor(pclass)3               -2.2575***
##                                (0.2719)
##
## factor(sex)male               -2.5516***
##                                (0.1735)
##
## age                           -0.0392***
##                                (0.0066)
##
## sibsp                         -0.3589***
##                                (0.1059)
##
## parch                          0.0586
##                                (0.1030)
##
## fare                           0.0012
##                                (0.0019)
##
## Constant                      3.8000***
##                                (0.3973)
##
## ------------------------------------------------
## Observations                     1,045
## Log Likelihood                -484.8250
## Akaike Inf. Crit.             985.6501
## ================================================
## Note:              *p<0.05; **p<0.01; ***p<0.001
```

Interpretation of the logistic regression is:

- Parch and fare are not statistically significant;
- Positive coefficients indicate positive effects on probability of survival;
- Negative coefficients indicate negative effects on probability of survival:
    a. Being male reduces the log odds by 2.55 after controlling for other factors;
    b. A unit increase in age reduces the log odds by 0.039 after controlling for other factors;
    c. Having one more sibling and/or spouse aboard reduced the log odds by 0.359 after controlling for other factors.

We notice that logistic regression does not report R squared. We can calculate the McFadden pseudo R squared by sung the pscl package.

```
# McFadden R2
# install.packages("pscl")
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
pR2(model)
```

```
##          llh      llhNull           G2     McFadden         r2ML
## -484.8250406 -706.7852714  443.9204616    0.3140420    0.3461022
##         r2CU
##    0.4667857
```

Or you can manually calculate the pseudo R squared by using the following formula:

$R^2_{McFadden} = 1 - \frac{log(L_m)}{log(L_{null})}$

where $log(L_m)$ is the log likelihood of the model of interest, $log(L_{null})$ is the likelihood of the null model, that has only intercept without any independent variables.

```
# Fit the null model
nullmodel <- glm(survived ~ 1,family=binomial(link='logit'),data=df)

# Show the log likelihood of the null model
logLik(nullmodel)
```

```
## 'log Lik.' -706.7853 (df=1)
```

```
# Manually calculate McFadden pseudo R squared
cat("McFadden pseudo R2 = ", 1-logLik(model)/logLik(nullmodel))
```

```
## McFadden pseudo R2 =  0.314042
```

4. Use Logistic Regression Model to Predict New Data

Let's use the trained logistic regression model to predict the survival probability for Jack and Rose. The test data are created based on the plot of the movie: https://en.wikipedia.org/wiki/Titanic_(1997_film)

```
test <- data.frame(sex = c("male", "female"),
                   pclass = c("3","1"),
                   age = c(19,17),
                   sibsp = c(0,0),
                   parch = c(0,1),
                   fare = c(5,500))
test$sex <- factor(test$sex)
test$pclass <- factor(test$pclass,levels=c("1","2","3"))

print(test)
```

```
##      sex pclass age sibsp parch fare
## 1   male      3  19     0     0    5
## 2 female      1  17     0     1  500
```

Let's call the predict() function to do the prediction.

```r
test$pred <- predict(model,test, type="response")

print(test)
```

```
##      sex pclass age sibsp parch fare     pred
## 1   male      3  19     0     0    5 0.148259
## 2 female      1  17     0     1  500 0.978095
```

Jack's probability of survival was 0.15 whereas Rose's probability was 0.98. This makes sense as Rose is in the 1st class, female, and younger than Jack.