INFX 573: Problem Set 7 - Maximum Likelihood, Logistic Regression

Rajendran Seetharaman

Due: Tuesday, December 5th, 2017

Problem Set 7

Collaborators:

Instructions:

Before beginning this assignment, please ensure you have access to R and RStudio.

1. Replace the "Insert Your Name Here" text in the `author:` field with your own name. List all collaborators on the top of your assignment.
2. Be sure to include well-documented (e.g. commented) code chucks, figures and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text.
3. Collaboration on problem sets is fun and useful but turn in an individual write-up in your own words and involving your own code. Do not just copy-and-paste from others' responses or code.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `YourLastName_YourFirstName_ps6.Rmd`, knit a PDF and submit the PDF file on Canvas.

1. Maximum Likelihood Solution

A website downloads per second can be approximated as a Poisson process with parameter λ. Assume that through a 10-second period, a website is downloaded 17, 8, 13, 11, 8, 11, 16, 7, 15, and 13 times. (This is your data).

1. Write down the Poisson probability to observe this number of visitors for each second, given the parameter value λ.

```
library(maxLik)
```

```
## Loading required package: miscTools
```

```
##
## Please cite the 'maxLik' package as:
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

```
#create data vector
dat <- c(17, 8, 13, 11, 8, 11, 16, 7, 15,13)
#number of observation
N <-10
#calculate lambda
```

```r
lamb <- mean(dat)
print(lamb)
```

```
## [1] 11.9
```

```r
#Poisson probabilities
pois_prob <- dpois(dat, lambda = lamb)
print(pois_prob)
```

```
##  [1] 0.03673884 0.06772528 0.10464677 0.11528068 0.06772528 0.11528068
##  [7] 0.05248406 0.04552960 0.07056681 0.10464677
```

2. Write down the log-likelihood of the same data.

```r
#create log-likelihood function
log_likelihood <- function(lam){dpois(dat,lambda = lam,log = TRUE)}
log_likelihood(lamb)
```

```
##  [1] -3.303921 -2.692296 -2.257165 -2.160385 -2.692296 -2.160385 -2.947246
##  [8] -3.089393 -2.651195 -2.257165
```

3. Compute the Maximum Likelihood estimate for λ, $\hat{\lambda}$. Explain your result intuitively.

Ans. The maximum likelihood estimate of lambda is computed as 11.9 which is the same as the mean of the values in the data set. This makes intuitive sense because the expected value of a Poisson random variable is equal to its parameter lambda 0, and the sample mean is an unbiased estimator of the expected value.

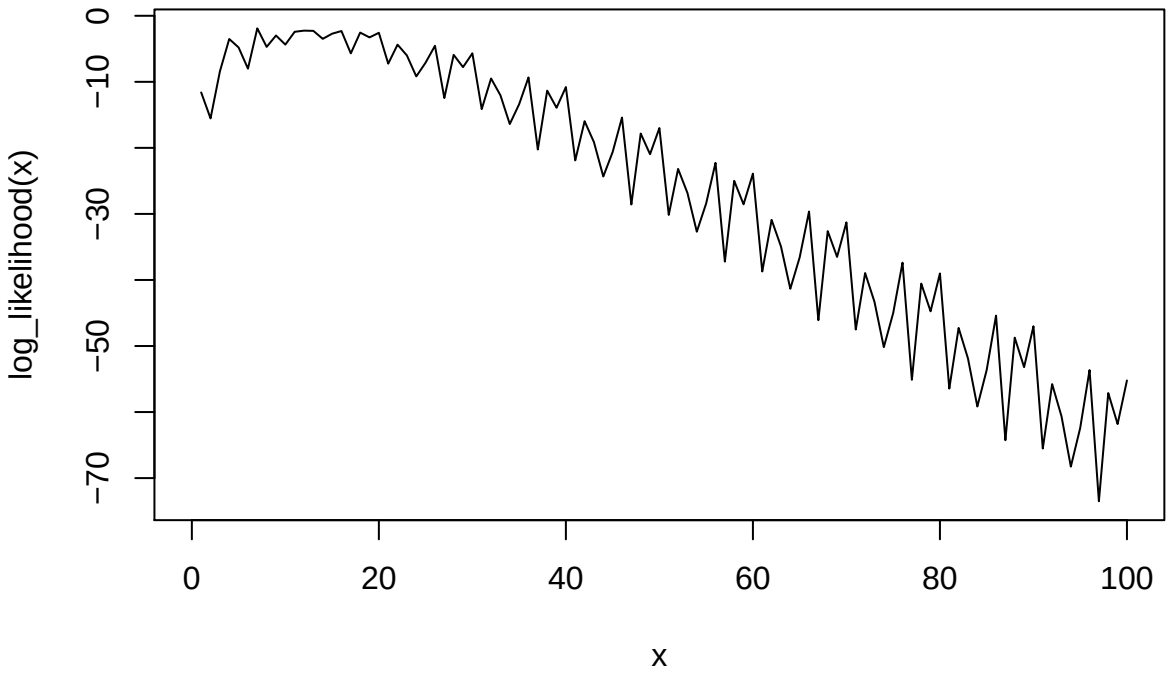
```r
#compute maximum likelihood estimate of lambda
lambda_hat <- maxLik(log_likelihood,start = 1)
#print summary
summary(lambda_hat)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 8 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: -26.21145
## 1  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]    11.90       1.09   10.92  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

4. Plot the log-likelihood as a function of λ in a suitable range around the $\hat{\lambda}$. Explain the result.

Ans. Looking at the graph we can see that the maximum value of this function is around lambda =11.9 which conforms with our computed maximum likelihood estimate of lambda which is also 11.9.

```r
#plot log likelihood function
curve(log_likelihood,0,100)
```



2. Logistic Regression

Download the Titanic survival data from canvas (files/data/titanic.csv.bz2). This is a long version of the survival where all passengers' data is observed individually.

Your task is to predict the survival in the Titanic's sinking.

1. Explore the dataset. What are the variables? What are the values/ranges/means of the more important ones? How many values are missing? Consult Kaggle Titanic Data for what the variable names mean.

Ans. The titanic dataset consists of information pertaining to the people who were aboard the RMS Titanic ship like names, passenger class, sex, age and whether or not they survived or not when the ship sank.

The dataset contains the following variables-

survived- If the passenger survived or not. 0 indicates No and 1 indicates Yes.

pclass- Indicates ticket class. 1= First class, 2= Second class, 3= Third class

sex- Sex of the passenger

Age- Age of passenger in years

sibsp- Number of siblings / spouses of passenger aboard the Titanic

parch- Number of parents / children of passenger aboard the Titanic

ticket- Ticket number of Passenger

fare- Fare paid by passenger for ticket in british pounds

cabin- Cabin number

embarked- Port of Embarkation for the passenger C = Cherbourg, Q = Queenstown, S = Southampton

name-Name of passenger

boat- Lifeboat number of passenger if he survived

body- Body identification number

home.dest- Home/Destination for passenger

The values/ range of values/ means for the important variables are the following- survived- 809 passengers did not survive and 500 passengers surived

pclass- 1st class: 323 passenger records 2nd class: 277 passenger records 3rd class: 709 passenger records

sex- Records for 466 females and 843 males are present in the dataset.

Age- Min age for passengers in the dataser was 0.1667 and max age was 80 years. Mean age was around 29 years.

sibsp- min value of 0 and max of 8 siblings/spouses with a mean of around 0.49 indicating that not many people had siblings or spouses on board.

parch-min value of 0 and max of 9 parents/children with a mean of around 0.49 indicating that not many people had parents/children on board

fare- Min value of 0 and max of 512.329 pounds with a mean fare of 33.295.

embarked- No data: 2 passengers C: 270 passengers Q: 123 passengers S: 914 passengers

The data has the following missing values- Age- 263 missing values (NA's) fare- 1 missing value (NA's) body- 1188 missing values (NA's)

source- https://www.kaggle.com/c/titanic/data

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
#read data
titanicdata <- read.csv('titanic.csv.bz2')
#convert survived and pclass to factors
titanicdata <- titanicdata %>%
  mutate(survived=as.factor(survived),pclass=as.factor(pclass))
#summarize data
str(titanicdata)
```

```
## 'data.frame':    1309 obs. of  14 variables:
##  $ pclass   : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...
##  $ name     : Factor w/ 1307 levels "Abbing, Mr. Anthony",..: 22 24 25 26 27 31 46 47 51 55 ...
##  $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
```

```
##  $ 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 ...
##  $ ticket   : Factor w/ 929 levels "110152","110413",..: 188 50 50 50 50 125 93 16 77 826 ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : Factor w/ 187 levels "","A10","A11",..: 45 81 81 81 81 151 147 17 63 1 ...
##  $ embarked : Factor w/ 4 levels "","C","Q","S": 4 4 4 4 4 4 4 4 4 2 ...
##  $ boat     : Factor w/ 28 levels "","1","10","11",..: 13 4 1 1 1 14 3 1 28 1 ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: Factor w/ 370 levels "","?Havana, Cuba",..: 310 232 232 232 232 238 163 25 23 230 ...
```

```
summary(titanicdata)
```

```
##  pclass  survived                                 name          sex
##  1:323   0:809    Connolly, Miss. Kate              :   2   female:466
##  2:277   1:500    Kelly, Mr. James                  :   2   male  :843
##  3:709            Abbing, Mr. Anthony               :   1
##                   Abbott, Master. Eugene Joseph     :   1
##                   Abbott, Mr. Rossmore Edward       :   1
##                   Abbott, Mrs. Stanton (Rosa Hunt):   1
##                   (Other)                           :1301
##       age              sibsp            parch           ticket
##  Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   CA. 2343:  11
##  1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   1601    :   8
##  Median :28.0000   Median :0.0000   Median :0.000   CA 2144 :   8
##  Mean   :29.8811   Mean   :0.4989   Mean   :0.385   3101295 :   7
##  3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000   347077  :   7
##  Max.   :80.0000   Max.   :8.0000   Max.   :9.000   347082  :   7
##  NA's   :263                                        (Other) :1261
##       fare                    cabin      embarked      boat
##  Min.   :  0.000                 :1014    :  2           :823
##  1st Qu.:  7.896   C23 C25 C27   :   6   C:270    13     : 39
##  Median : 14.454   B57 B59 B63 B66:   5   Q:123    C      : 38
##  Mean   : 33.295   G6            :   5   S:914    15     : 37
##  3rd Qu.: 31.275   B96 B98       :   4            14     : 33
##  Max.   :512.329   C22 C26       :   4            4      : 31
##  NA's   :1         (Other)       : 271            (Other):308
##       body                       home.dest
##  Min.   :  1.0                        :564
##  1st Qu.: 72.0   New York, NY         : 64
##  Median :155.0   London               : 14
##  Mean   :160.8   Montreal, PQ         : 10
##  3rd Qu.:256.0   Cornwall / Akron, OH:  9
##  Max.   :328.0   Paris, France        :  9
##  NA's   :1188    (Other)              :639
```

```
#find na values in data
sapply(titanicdata,function(x) sum(is.na(x)))
```

```
##    pclass  survived      name       sex       age     sibsp     parch
##         0         0         0         0       263         0         0
##    ticket      fare     cabin  embarked      boat      body home.dest
##         0         1         0         0         0      1188         0
```

2. Estimate a logistic regression model where you introduce the most important explanatory variebles. Interpret the results.

Ans.

According to me the most important explanantory variables for predicting the survival of a passenger on Titanic are sex, pclass, and age.

Interpretation for coefficients pertaining to passenger class (pclass):

Being a passenger travelling in the second class (pclass=2), versus travelling in first class (pclass=1), changes the log odds of survival by -1.280570. (Decreases odds of survival)

Being a passenger travelling in the third class (pclass=3), versus travelling in first class (pclass=1), changes the log odds of survival by -2.289661. (Decreases odds of survival)

Interpretation for coefficient pertaining to passenger sex (sex):

Being a male passenger, versus a female passenger, changes the log odds of survival by -2.497845. (Decreases odds of survival)

Interpretation for coefficient pertaining to passenger age (Age):

For a one unit increase in passenger age, the log odds of survival change by -0.034393. (Decreases odds of survival)

All p values of included variables in the model are below the critical p value of 0.05, indicating that all of them are statistically significant.

```
#fit model with predictors as pclass, sex, age to variables
model <- glm(survived ~ pclass+sex+age, data=titanicdata, family
         = binomial(link = "logit"))
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age, family = binomial(link = "logit"),
##     data = titanicdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6399  -0.6979  -0.4336   0.6688   2.3964
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.522074   0.326702  10.781  < 2e-16 ***
## pclass2     -1.280570   0.225538  -5.678 1.36e-08 ***
## pclass3     -2.289661   0.225802 -10.140  < 2e-16 ***
## sexmale     -2.497845   0.166037 -15.044  < 2e-16 ***
## age         -0.034393   0.006331  -5.433 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  982.45  on 1041  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 992.45
##
## Number of Fisher Scoring iterations: 4
```

3. In general, women had much larger chance of survival. Is this surprising to you? Does this tell you anything about the Titanic's final hours?

Ans. Yes, the data shows a clear indication that the chances of survival for a woman was more than that for a man on titanic. One possible reasoning behind this might be that men might have helped the women to evacuate the ship before the men could leave i.e the ship crew were giving preference to women while evacuating the ship probably because the women might be taking care of children aboard.

4. Introduce interactions (cross effects) between gender and passenger class. Interaction effects mean you are allowing the result for men and women to differ for each different class. Interpret the results.

Ans.

Introducing the interaction effects between gender and passenger, the following are the interpretations of the model coefficients:

Being a female passenger travelling in the second class (pclass=2), versus a female passenger travelling in first class (pclass=1), changes the log odds of survival by -1.529875. (Decreases odds of survival) Statistically significant

Being a female passenger travelling in the third class (pclass=3), versus a female passenger travelling in first class (pclass=1), changes the log odds of survival by -4.064965. (Decreases odds of survival) Statistically significant

Interpretation for coefficient pertaining to passenger sex (sex): Being a male passenger in first class, versus a female passenger in first class, changes the log odds of survival by -3.886389. (Decreases odds of survival) Statistically significant

For a one unit increase in passenger age, the log odds of survival change by -0.038401. (Decreases odds of survival)

Being a male passenger travelling in the second class (pclass=2), versus a female passenger travelling in first class (pclass=1), changes the log odds of survival by -0.070404. (Decreases odds of survival) Not statistically significant

Being a male passenger travelling in the third class (pclass=3), versus a female passenger travelling in first class (pclass=1), changes the log odds of survival by 2.488808. (Increases odds of survival) Statistically significant

```
#fit model to variables with interactin effect between sex and pclass
model_interaction <- glm(survived ~ pclass+sex+age+sex:pclass,
                         data=titanicdata, family = binomial(link = "logit"))
summary(model_interaction)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age + sex:pclass, family = binomial(link = "logit"),
##     data = titanicdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0778  -0.6604  -0.4943   0.4263   2.4935
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.804345   0.546937   8.784  < 2e-16 ***
## pclass2         -1.529875   0.566481  -2.701  0.00692 **
## pclass3         -4.064965   0.510661  -7.960 1.72e-15 ***
## sexmale         -3.886389   0.492375  -7.893 2.95e-15 ***
## age             -0.038401   0.006743  -5.695 1.23e-08 ***
```

```
## pclass2:sexmale -0.070404   0.630978  -0.112  0.91116
## pclass3:sexmale  2.488808   0.540042   4.609 4.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.62  on 1045  degrees of freedom
## Residual deviance:  931.99  on 1039  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 945.99
##
## Number of Fisher Scoring iterations: 5
```

5. Do less obvious variables, such as fare (given we already control for class) and port of embarkation help explaining survival? Can you explain the outcome?

Ans.

The following is the interpretations of the coefficients of the logistic regression model- 1. For a one unit increase in passenger fare, the log odds of survival increase by 0.01103. 2. There does not seem to be a mojor difference in passenger survival rates having different ports of embarkation. The categorical variable coefficients are very close to each other. Also, with a high p-value, port of embarkation not seem to be statistically significant.

```
#use fare and embarked as model predictors
model1 <- glm(survived ~ fare + embarked,family=
                binomial(link="logit"), data= titanicdata)
summary(model1)
```

```
##
## Call:
## glm(formula = survived ~ fare + embarked, family = binomial(link = "logit"),
##     data = titanicdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2853  -0.8931  -0.8194   1.2660   1.6235
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  12.68370  378.59289   0.034    0.973
## fare          0.01103    0.00163   6.768  1.3e-11 ***
## embarkedC   -13.04249  378.59289  -0.034    0.973
## embarkedQ   -13.40782  378.59293  -0.035    0.972
## embarkedS   -13.68990  378.59289  -0.036    0.971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1740.1  on 1307  degrees of freedom
## Residual deviance: 1632.2  on 1303  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1642.2
##
```

```
## Number of Fisher Scoring iterations: 12
```

3. How much work?

Tell us, roughly how many hours did you spend on this homework.

Ans. About 5 hours.