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

TAPASVI BANSAL

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.

```r
# install.packages("tidyverse")
# install.packages("maxLik")
# install.packages("Amelia")
# install.packages("mfx")
library(mfx)
```

```
## Loading required package: sandwich

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: MASS

## Loading required package: betareg
```

```r
library(Amelia)
```

```
## Loading required package: Rcpp
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.4, built: 2015-12-05)
## ## Copyright (C) 2005-2017 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.4.2

## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --

## √ ggplot2 2.2.1     √ purrr   0.2.4
## √ tibble  1.3.4     √ dplyr   0.7.4
## √ tidyr   0.7.2     √ stringr 1.2.0
## √ readr   1.1.1     √ forcats 0.2.0

## Warning: package 'tidyr' was built under R version 3.4.2

## Warning: package 'purrr' was built under R version 3.4.2

## Warning: package 'dplyr' was built under R version 3.4.2

## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
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. Co
##
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
## https://r-forge.r-project.org/projects/maxlik/
```

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.

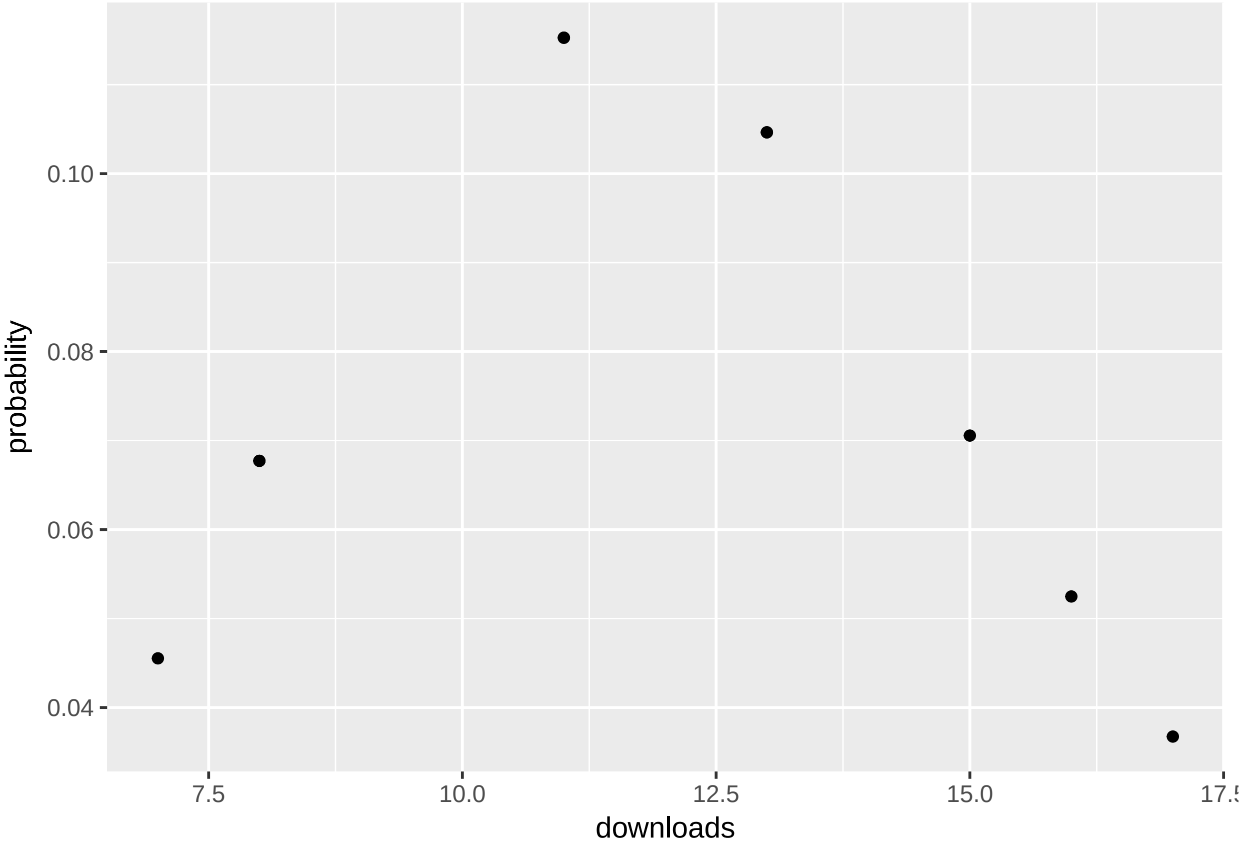
```
downloads <- c(17, 8, 13, 11, 8, 11, 16, 7, 15, 13)
downmean <- sum(downloads)/10

#sapply(downloads, function(x) dpois(x, lambda = downmean))

probability =list()
for(i in downloads)
  {
  probability <- dpois(downloads,lambda = downmean)
  }
```

```
ddf <- data.frame(downloads,probability)
ddf <- ddf[order(ddf$downloads),]
row.names(ddf) <- NULL


qplot(downloads, probability )
```



```
#ggplot(ddf, aes(x = downloads, y = probability)) + geom_bar(stat="identity") + ggtitle("Downloads per
```

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

```
lambda = downmean
poissonloglike <-
  function(lambda)
    {
      n =length(downloads)
      loglike = (sum(downloads)*log(lambda) - n*lambda)
      return(loglike)
    }
```

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

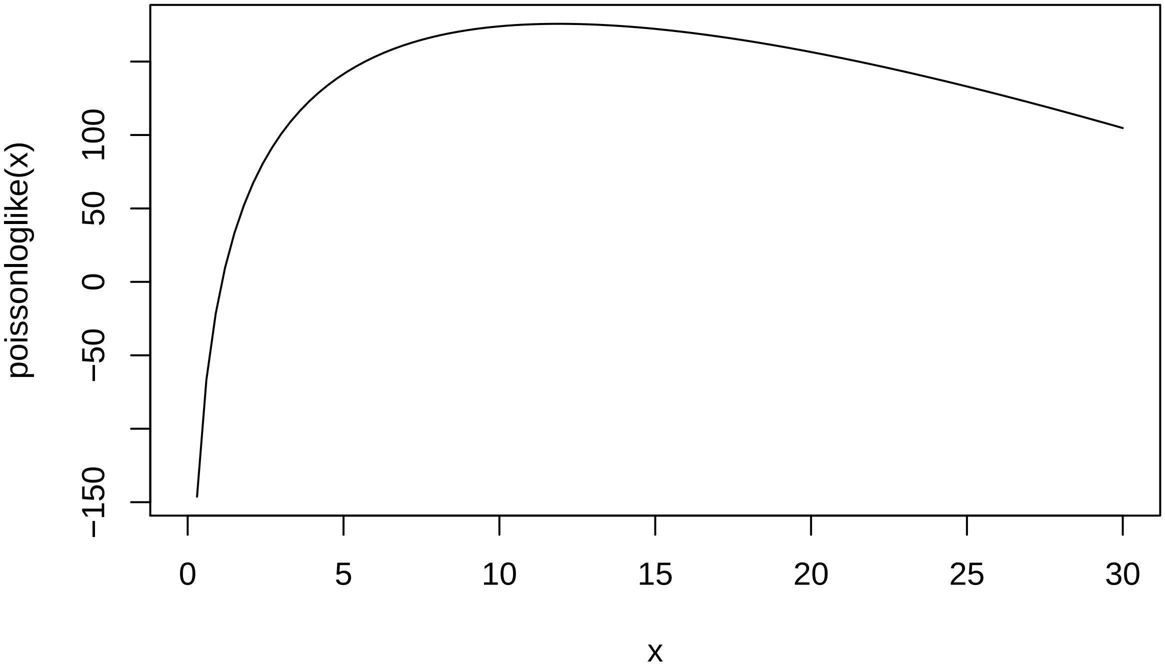
```
ml<- maxLik(poissonloglike, start =0.5)
summary(ml)
```

```
## --------------------------------------------
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 9 iterations
## Return code 2: successive function values within tolerance limit
## Log-Likelihood: 175.7081
## 1  free parameters
## Estimates:
##      Estimate Std. error t value Pr(> t)
## [1,]   11.900      1.065   11.17  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## --------------------------------------------
```

The Maximum Likelihood estimation for the given set of data in accordance with the Poisson distribution model is achieved at 11.9. For this estimate the log likelihood is 175.7081.

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

```
curve(poissonloglike,0,30)
```



```
?curve
```

The curve above the exhibits the poisson distribution curve nature. On increasing the limit of data points, we see the drop in the value of likelihood function. Hence, we understand that when the rate of occurrence of some event (downloads in this case) is small, the range of likely possibilities will lie near the zero line and as the occurrence of the independent events becomes more common (like the repetition of number of website downloads), the center of the curve moves toward the right.

2. Logistic Regression

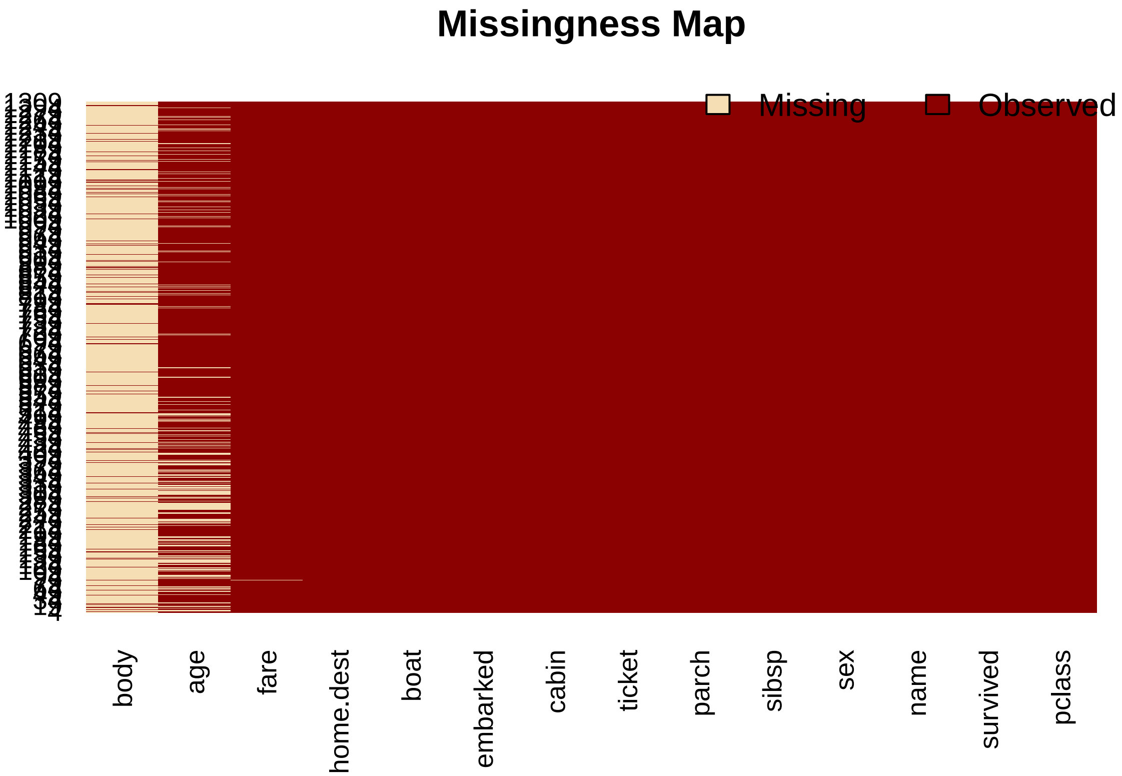
Download the Titanic survival data from canvas (files/data/titanic.csv.bz2). This is a long version of the survival where all passengers' data is observed individually.

```
btitanic <- read.csv("titanic.csv.bz2")
#head(btitanic)
```

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.

```r
missmap(btitanic)
```

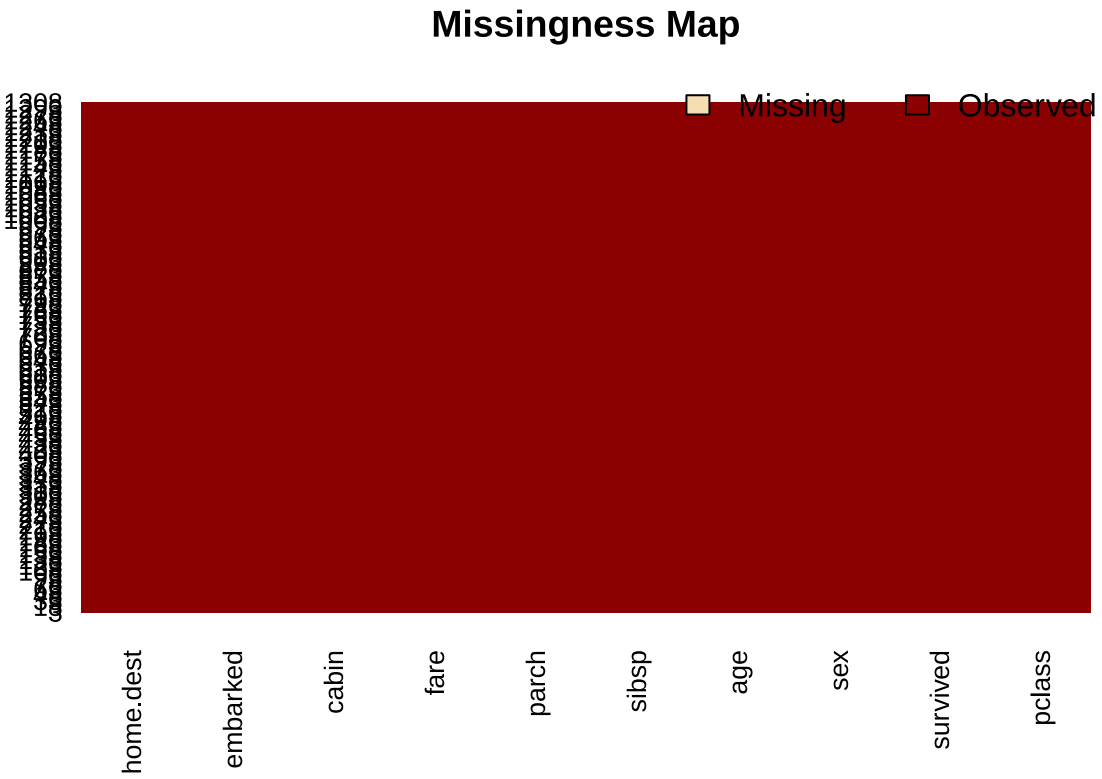


```r
#length(na.omit(btitanic$body))
#sapply(btitanic, function(x) length(na.omit(x)))

Otitanic <- btitanic[c(1,2,4,5,6,7,9,10,11,14)]
avgage <- mean(Otitanic$age, na.rm = T)
Otitanic$age[is.na(Otitanic$age)] <- avgage
Otitanic <- Otitanic[!is.na(Otitanic$fare),]

# Converting pclass, survived, sex variables into factors for analysis
Otitanic$pclass <- as.factor(Otitanic$pclass)
Otitanic$survived <- as.factor(Otitanic$survived)
Otitanic$sex <- as.factor(Otitanic$sex)

#sapply(Otitanic, function(x) length(unique(x)))
missmap(Otitanic)
```



```
## The following code is used for exploratory purposes

#Otitanic <- subset(btitanic,select=c(1:12,14))
#sapply(Otitanic, function(x) sum(is.na(x)))
#View(Otitanic)
```

Variables-

1) survival: Survival -> 0 = No, 1 = Yes
2) pclass: Ticket -> class 1 = 1st, 2 = 2nd, 3 = 3rd
3) sex: Sex
4) Age: Age in years

5) sibsp: Number of siblings / spouses aboard the Titanic
6) parch: Number of parents / children aboard the Titanic

7) ticket: Ticket number

8) fare: Passenger fare

9) cabin: Cabin number
   10)embarked: Port of Embarkation -> C = Cherbourg, Q = Queenstown, S = Southampton

pclass is a proxy for socio-economic status (SES), 1st is the Upper class, 2nd is Middle class and 3rd is the Lower class.

Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp in the dataset defines family relations, Sibling if it is brother, sister, stepbrother or stepsister and Spouse if husband or wife (mistresses and fiancés were ignored).

parch in the dataset defines family relations, when it is Parent it can be mother or father. When Child it can be any of daughter, son, stepdaughter or stepson. Also, for some children parch=0 who travelled only with a nanny.

Data Wrangling and Analysis: 1) From the dataset, the variable "body" can be ignored as it contains only 121 data points in the dataset where most variables have 1309 observations. With the variable age, as the second variable with the most missing data points (263).

2)Hence using mean age of the Titanic population to apply on the age column that are missing. Since the variable age is missing only one value, hence the particular record can be omitted from the data set, assuming no major loss of information.

3)The pcalss (passenger class) feature are converted into factor in the dataset because the value of 1,2,3 should not be treated as numerical but as category levels for analysis.

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

```
#sapply(Otitanic, function(x) length(unique(x)))
#View(Otitanic)

## For exploratory purpose:
#model<- glm(survived ~ fare, data = Otitanic, family = binomial(link = 'logit'))
#summary(model)

## The above generalized linear model fails to explain relation clearly.
```

```
## Observing estimates for a binary logistic regression model and corresponding marginal effects.


blrmodel <- mfx::logitmfx(survived ~ fare + pclass + sex + age + sibsp + parch + embarked, data= Otitani
#blrmodel

blrmodel1 <- mfx::logitmfx(survived ~ pclass, data= Otitanic)
blrmodel1
```

```
## Call:
## mfx::logitmfx(formula = survived ~ pclass, data = Otitanic)
##
## Marginal Effects:
##             dF/dx Std. Err.        z     P>|z|
## pclass2 -0.166985  0.033128  -5.0405 4.642e-07 ***
## pclass3 -0.354205  0.030312 -11.6853 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "pclass2" "pclass3"
```

```
blrmodel2 <- mfx::logitmfx(survived ~ sex, data= Otitanic)
blrmodel2
```

```
## Call:
## mfx::logitmfx(formula = survived ~ sex, data = Otitanic)
##
## Marginal Effects:
##            dF/dx Std. Err.       z     P>|z|
## sexmale -0.53626   0.02468 -21.728 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "sexmale"
```

```
blrmodel3 <- mfx::logitmfx(survived ~ age, data= Otitanic)
blrmodel3
```

```
## Call:
## mfx::logitmfx(formula = survived ~ age, data = Otitanic)
##
## Marginal Effects:
##          dF/dx  Std. Err.       z   P>|z|
## age -0.0018632  0.0010551 -1.7659 0.07741 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Marginal effects show the change in probability when the predictor or independent variable increases by one unit. Here, in the case of titanic data, we see the closest "marginal" effects with variables like class, sex and age on the survival rate of the passengers.

In case of the factor of class, we see that in class alone there is drop of survival rate by approximately 16% in second class (for pclass =2) and it drops to approx. 35% for third class (for pclass =3). It can be inferred that survival rate of passengers who were lower class (less fare) was less than the passengers who were in upper class (high/more fare). It can also be said that the choice of/ being in class improves/reduces chances of survival.

Also, according to binary logistic regression model in this case, being male reduces your chances of survival by approx. 53%. It can also be said women were up to 54 per cent more likely to have lived than men.

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?

The idea of saving 'women and children first' has been described as 'the unwritten law of the sea'. Men actually have a distinct survival advantage,but in case of titanic Men stood back, while women and children were given priority to board the lifeboats. This 'chivalry' was also helped by the fact that the captain threatened to shoot men who got into the lifeboats before women.

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.

```
crossmodel <- mfx::logitmfx(survived ~ sex + pclass + sex*pclass, data= Otitanic)
crossmodel
```

```
## Call:
## mfx::logitmfx(formula = survived ~ sex + pclass + sex * pclass,
##     data = Otitanic)
##
## Marginal Effects:
##                      dF/dx Std. Err.        z     P>|z|
## sexmale         -0.758576  0.048695 -15.5782 < 2.2e-16 ***
## pclass2         -0.265413  0.098725  -2.6884  0.007179 **
## pclass3         -0.677908  0.066824 -10.1446 < 2.2e-16 ***
## sexmale:pclass2  0.038989  0.148879   0.2619  0.793413
## sexmale:pclass3  0.517927  0.097087   5.3347 9.573e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## dF/dx is for discrete change for the following variables:
##
## [1] "sexmale"         "pclass2"         "pclass3"         "sexmale:pclass2"
## [5] "sexmale:pclass3"
```

On introducing interactions (cross effects) between gender and passenger class, we witness that for male passengers from second class had approx. 3% chance of survival even if the overall chances of survival drops by approx. 26% in the second class.

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?

```
blrmodel4 <- mfx::logitmfx(survived ~ fare, data= Otitanic)
blrmodel4
```

```
## Call:
## mfx::logitmfx(formula = survived ~ fare, data = Otitanic)
##
## Marginal Effects:
##           dF/dx  Std. Err.      z     P>|z|
## fare 0.00294888 0.00038607 7.6382 2.203e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
blrmodel5 <- mfx::logitmfx(survived ~ embarked, data= Otitanic)
blrmodel5
```

```
## Call:
## mfx::logitmfx(formula = survived ~ embarked, data = Otitanic)
##
## Marginal Effects:
##              dF/dx Std. Err.       z  P>|z|
## embarkedC -0.90728   6.62652 -0.1369 0.8911
## embarkedQ -0.70214   7.56726 -0.0928 0.9261
## embarkedS -0.99160   0.95960 -1.0333 0.3014
##
## dF/dx is for discrete change for the following variables:
##
## [1] "embarkedC" "embarkedQ" "embarkedS"
```

```
crossmodel1 <- mfx::logitmfx(survived ~ fare + embarked + fare*embarked, data= Otitanic)
```

With respect to fare, an increase in a unit of fare improves the chances of survival by 0.2%, this is also supported by the analysis of class where the survival chances for 1st class is more than the survival chances for 3rd class.

With respect to port of embarkation, there seems to be no analysis that suggest there is differences in chances of survival.

3. How much work?

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

For the amount of work done, I spent roughly 10-12 hours on this assignment.