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

Charudatta Deshpande

Due: Tuesday, December 5th, 2017

Problem Set 7

Collaborators: Charles Hemstreet, Robert Hinshaw, Ram Ganesan, Manjiri Kharkar

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.

```
# Load some libraries that we may use during this assignment.
library(ggplot2)
library(data.table)
library(tidyverse)
```

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

```
## v tibble  1.3.4     v purrr   0.2.4
## v tidyr   0.7.2     v dplyr   0.7.4
## v readr   1.1.1     v stringr 1.2.0
## v tibble  1.3.4     v forcats 0.2.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::between()   masks data.table::between()
## x dplyr::filter()    masks stats::filter()
## x dplyr::first()     masks data.table::first()
## x dplyr::lag()       masks stats::lag()
## x dplyr::last()      masks data.table::last()
## x purrr::transpose() masks data.table::transpose()
```

```
library(aod)
library(jtools)
```

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.

Answer -

Let x be the number of events per interval. In this case, x is number of downloads per second. The formula for the Poisson probability mass function with rate parameter λ is -

$P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$

x can take following values -
17, 8, 13, 11, 8, 11, 16, 7, 15, 13

To view the values calculated in terms of λ, replace x with above values in the formula. The values are calculated below -

second 1 - $P(17;\lambda) = \frac{e^{-\lambda}\lambda^{17}}{17!}$

second 2 - $P(8;\lambda) = \frac{e^{-\lambda}\lambda^{8}}{8!}$

second 3 - $P(13;\lambda) = \frac{e^{-\lambda}\lambda^{13}}{13!}$

second 4 - $P(11;\lambda) = \frac{e^{-\lambda}\lambda^{11}}{11!}$

second 5 - $P(8;\lambda) = \frac{e^{-\lambda}\lambda^{8}}{8!}$

second 6 - $P(11;\lambda) = \frac{e^{-\lambda}\lambda^{11}}{11!}$

second 7 - $P(16;\lambda) = \frac{e^{-\lambda}\lambda^{16}}{16!}$

second 8 - $P(7;\lambda) = \frac{e^{-\lambda}\lambda^{7}}{7!}$

second 9 - $P(15;\lambda) = \frac{e^{-\lambda}\lambda^{15}}{15!}$

second 10 - $P(13;\lambda) = \frac{e^{-\lambda}\lambda^{13}}{13!}$

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

Answer -

Refer to attached image file named 'Deshpande Charudatta ps7 Q1.2.pdf'.

Also calculated the formulae below -

In general, it is

$logP(x;\lambda) = -\lambda + xlog\lambda - log(x!)$

When calculated for each second, it is -

second 1 - $logP(17;\lambda) = -\lambda + 17log\lambda - log(17!)$

second 2 - $logP(8;\lambda) = -\lambda + 8log\lambda - log(8!)$

second 3 - $logP(13;\lambda) = -\lambda + 13log\lambda - log(13!)$

second 4 - $logP(11;\lambda) = -\lambda + 11log\lambda - log(11!)$

second 5 - $logP(8;\lambda) = -\lambda + 8log\lambda - log(8!)$

second 6 - $logP(11;\lambda) = -\lambda + 11log\lambda - log(11!)$

second 7 - $logP(16;\lambda) = -\lambda + 16log\lambda - log(16!)$

second 8 - $logP(7;\lambda) = -\lambda + 7log\lambda - log(7!)$

second 9 - $logP(15;\lambda) = -\lambda + 15log\lambda - log(15!)$

second 10 - $logP(13;\lambda) = -\lambda + 13log\lambda - log(13!)$

3. Compute the Maximum Likelihood estimate for λ, . Explain your result intuitively.

Answer -

Refer to attached image file named 'Deshpande Charudatta ps7 Q1.3.pdf'.

From the calculation, $\hat{\lambda} = x$.

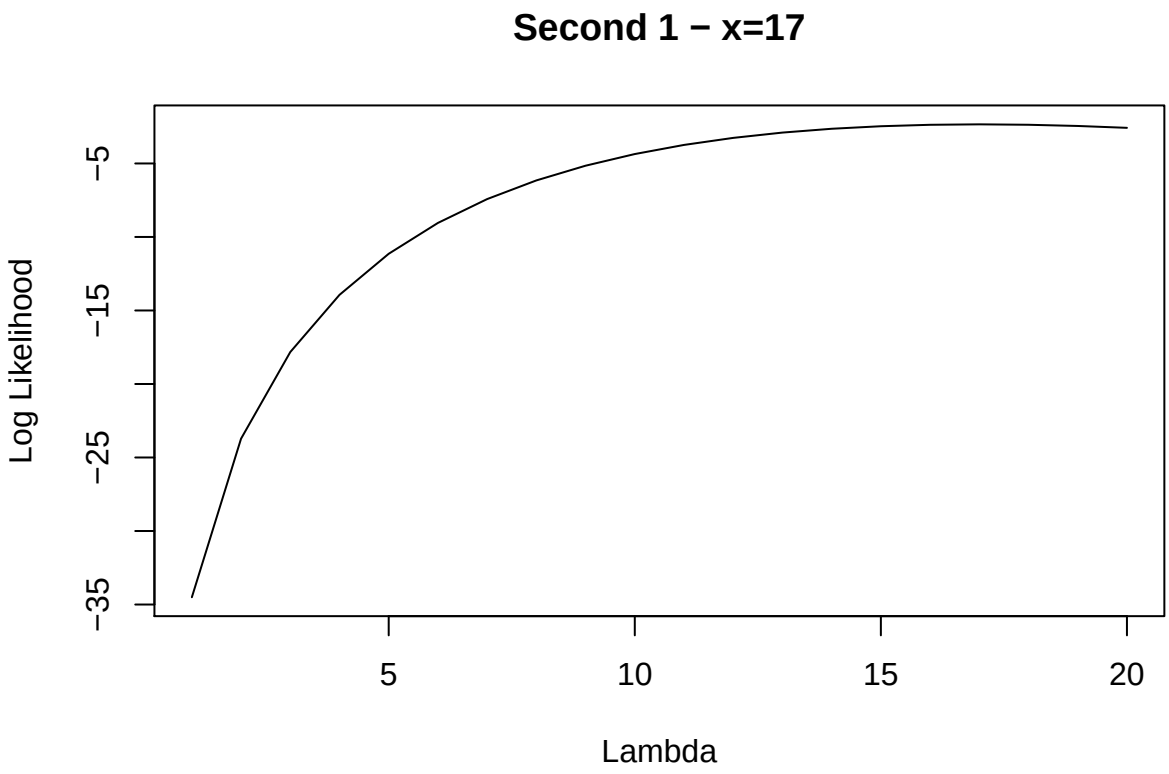
Intuitively, this is what I feel is the explanation -

λ is the rate parameter, or average number of events occuring per interval. The likelihood of observing x observations in an inetrval is the maximum, when $\lambda = x$, since that is the average. Any other value of x or λ decreases the liklelihood of observing that value. That is why we received $\hat{\lambda} = x$.

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

Answer -

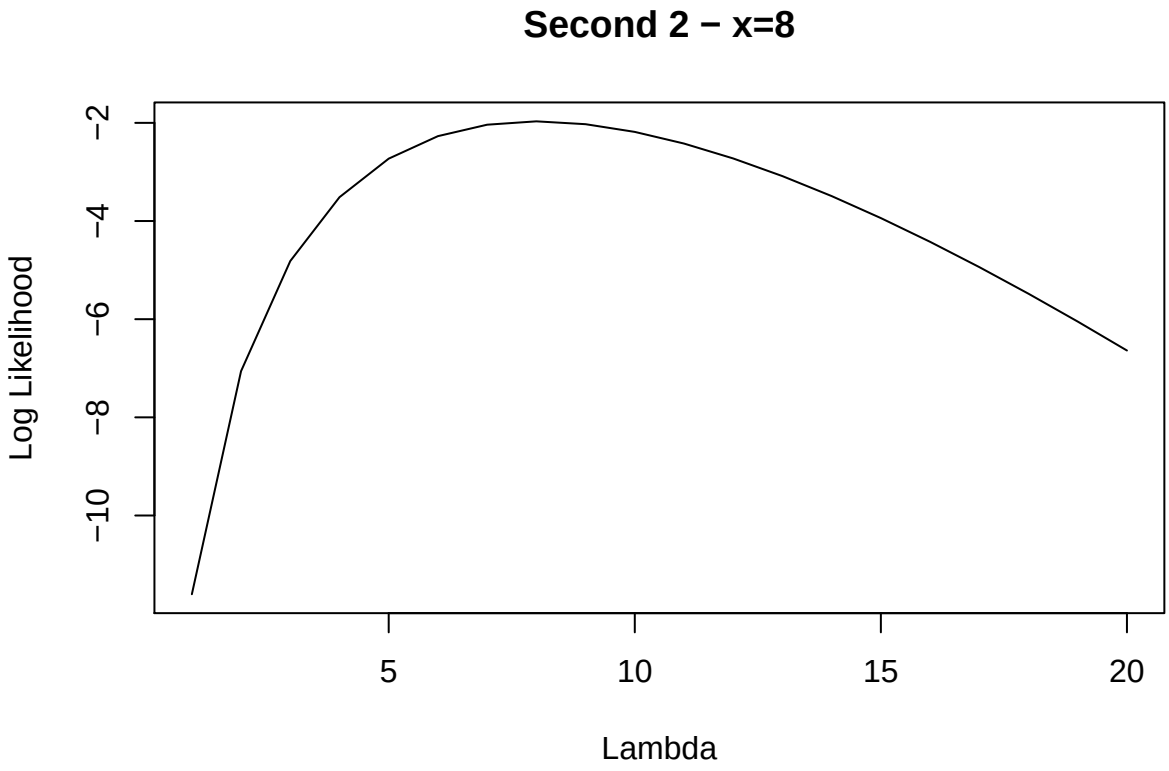
```
# Load values of x.
x <- c(17, 8, 13, 11, 8, 11, 16, 7, 15, 13)
#
#Let lambda be denoted by L from above equation.
log.likelihood <- function(x, L) -L + x*log(L) - log(factorial(x))
#For purpose of plotting, let Lambda be incremented by 1 from 1 to 20.
L <- (1:20/1)
#Input values for above 10 seconds.
s1 <- log.likelihood(17, L)
s2 <- log.likelihood(8, L)
s3 <- log.likelihood(13, L)
s4 <- log.likelihood(11, L)
s5 <- log.likelihood(8, L)
s6 <- log.likelihood(11, L)
s7 <- log.likelihood(16, L)
s8 <- log.likelihood(7, L)
s9 <- log.likelihood(15, L)
s10 <- log.likelihood(13, L)
#plot above 10 plots and verify results using which.max function.
plot(L, s1, type="l", main="Second 1 - x=17", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s1)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 17"
```

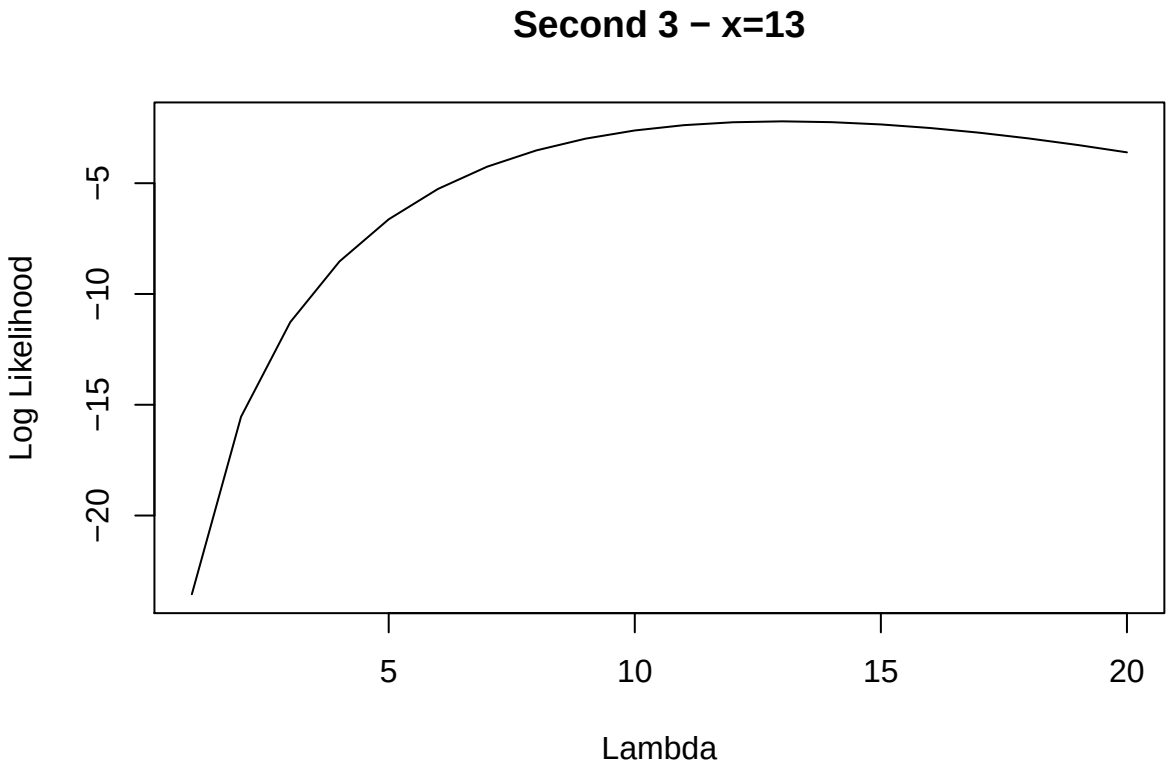
```
plot(L, s2, type="l", main="Second 2 - x=8", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s2)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 8"
```

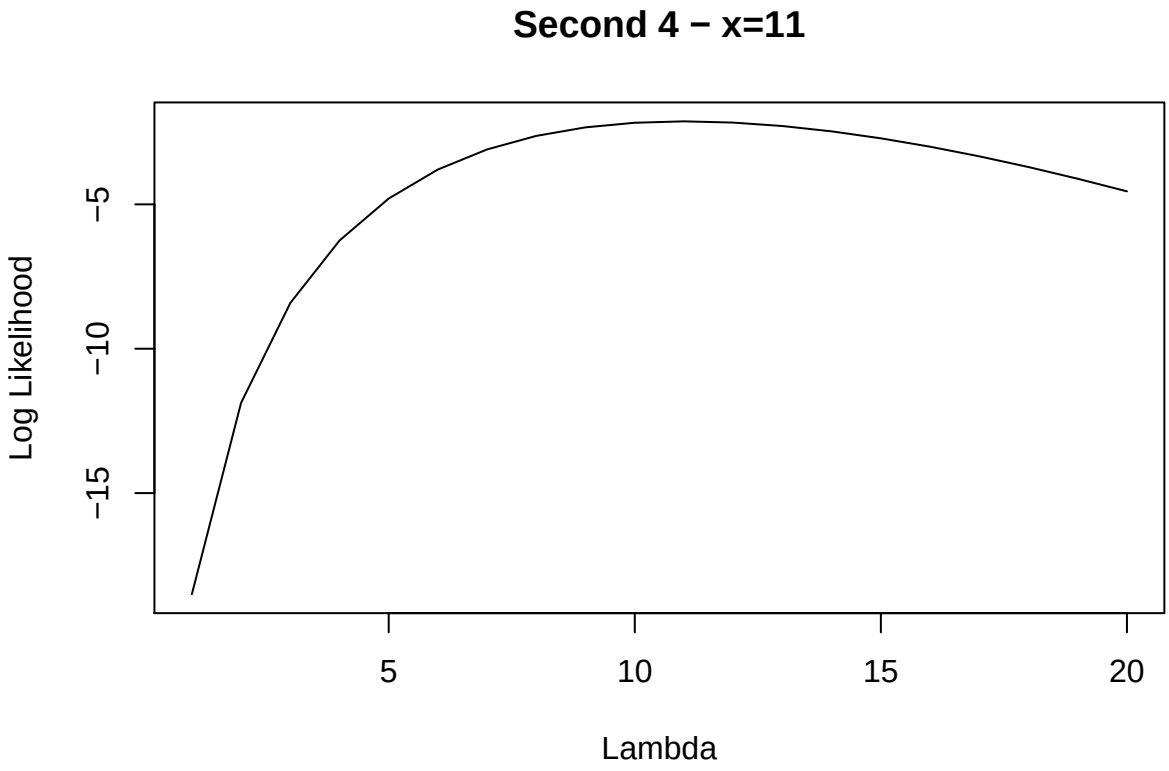
```
plot(L, s3, type="l", main="Second 3 - x=13", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s3)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 13"
```

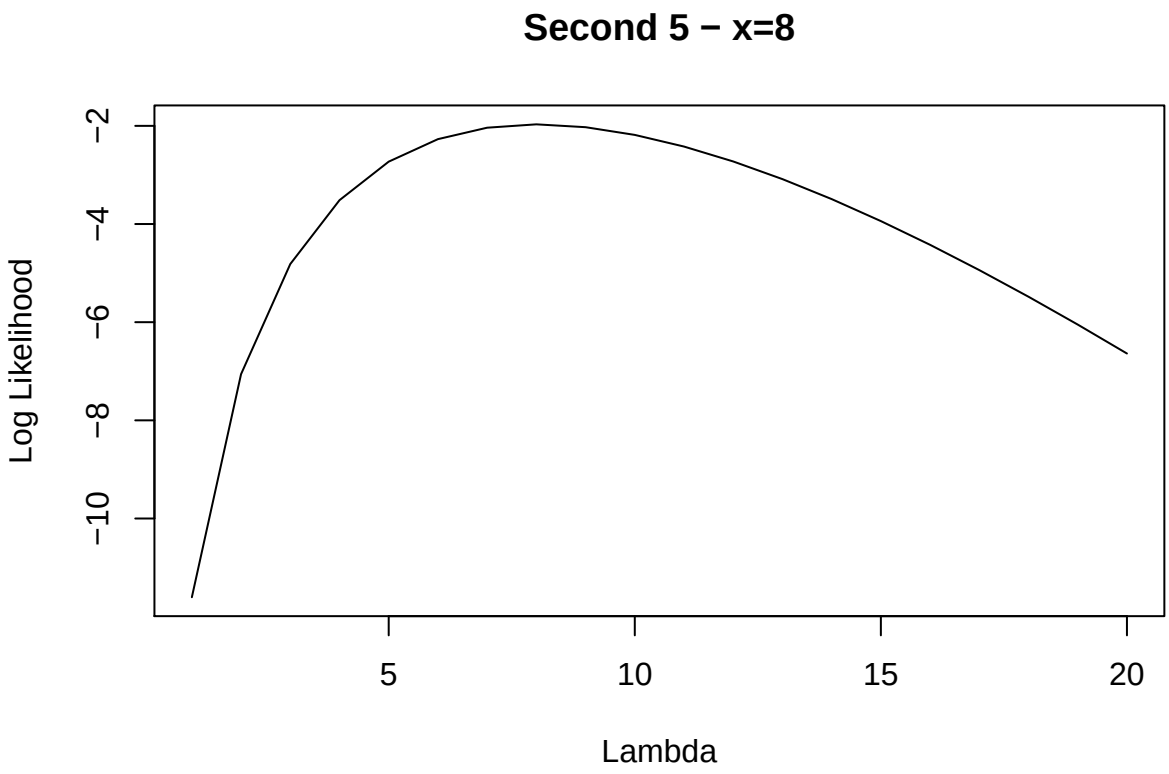
```
plot(L, s4, type="l", main="Second 4 - x=11", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s4)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 11"
```

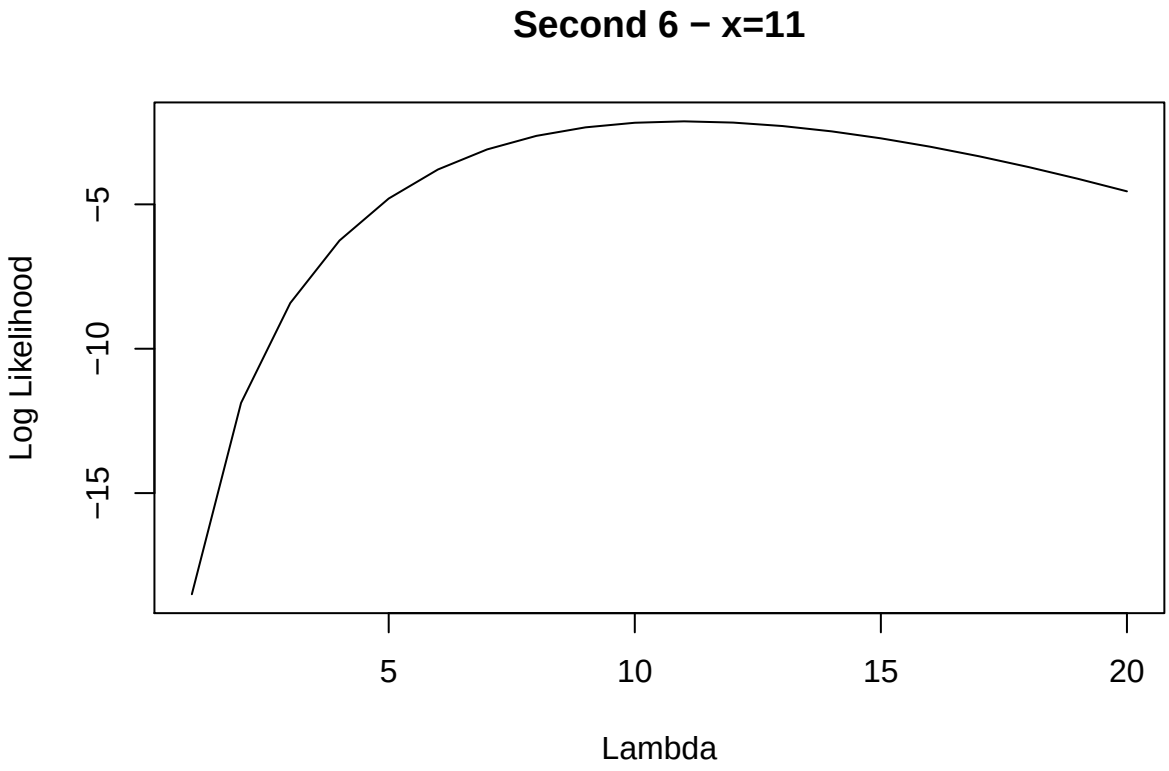
```
plot(L, s5, type="l", main="Second 5 - x=8", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s5)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 8"
```

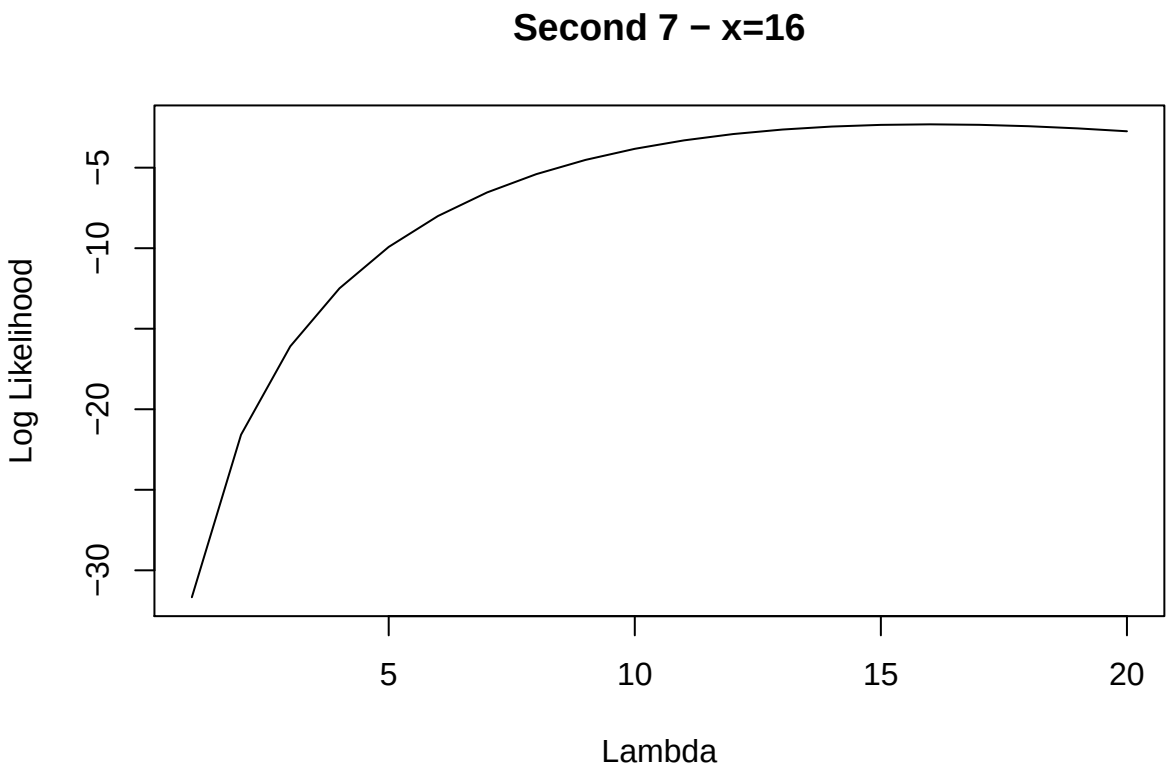
```
plot(L, s6, type="l", main="Second 6 - x=11", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s6)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 11"
```

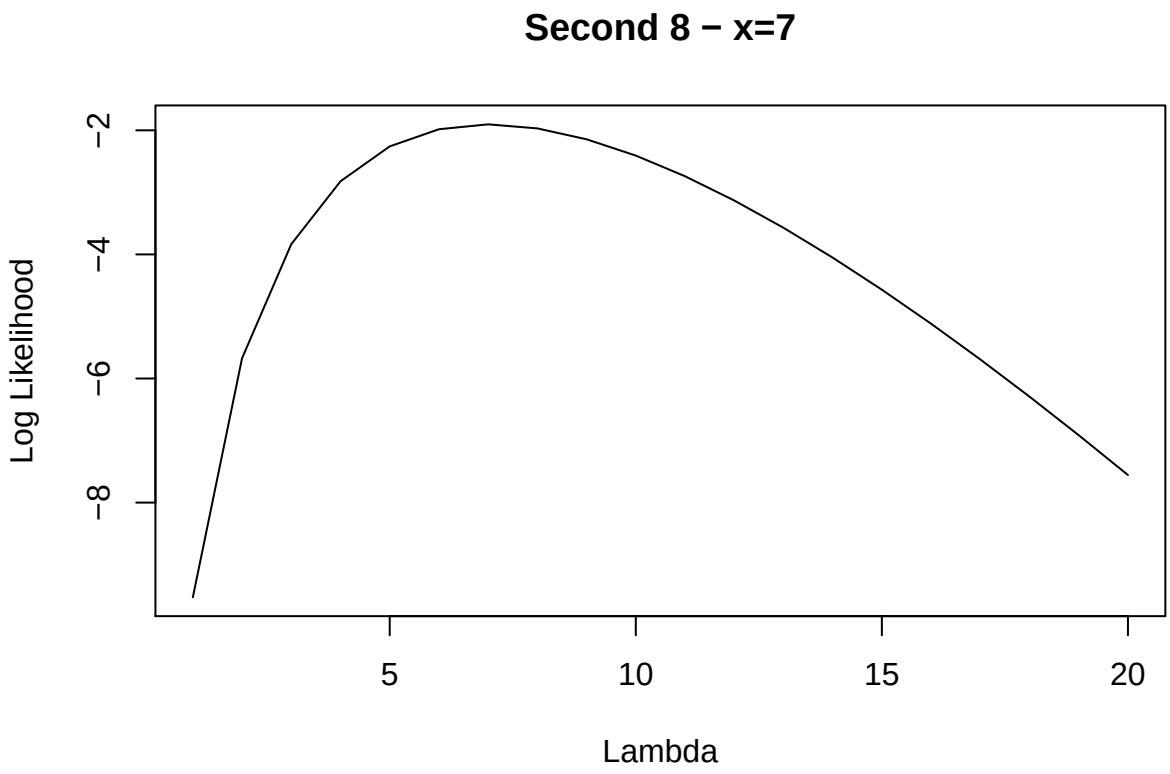
```
plot(L, s7, type="l", main="Second 7 - x=16", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s7)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 16"
```

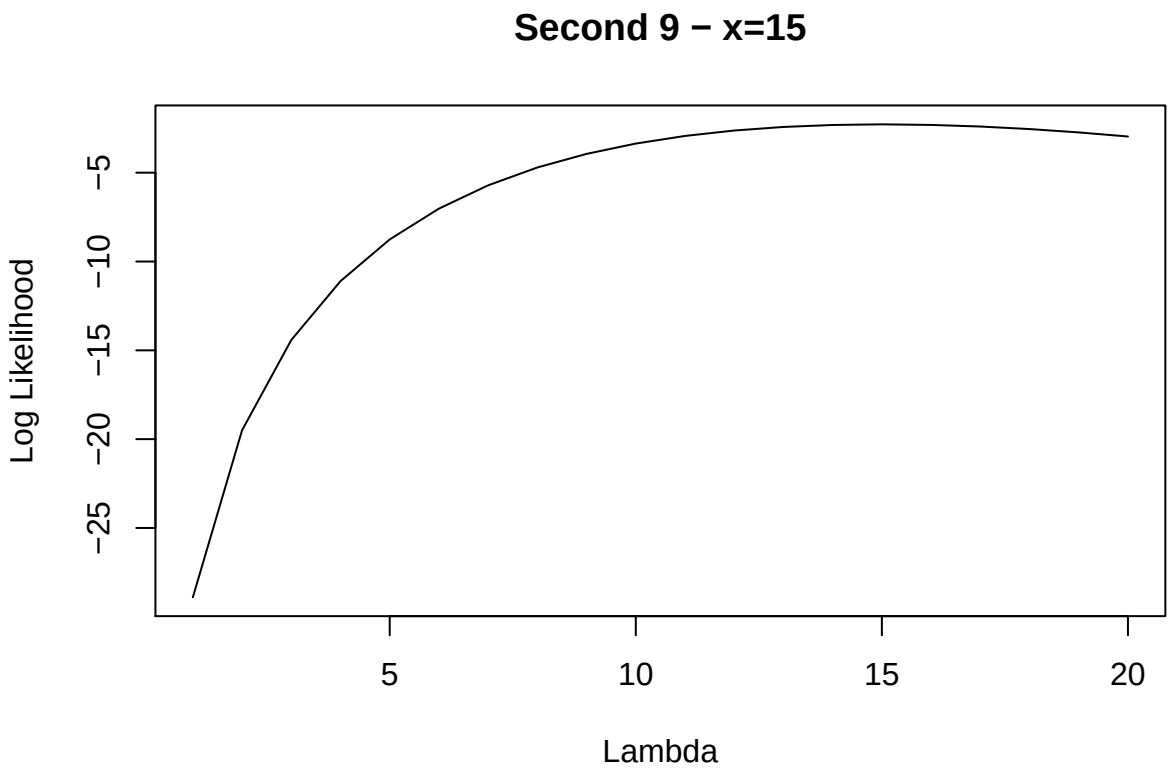
```
plot(L, s8, type="l", main="Second 8 - x=7", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s8)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 7"
```

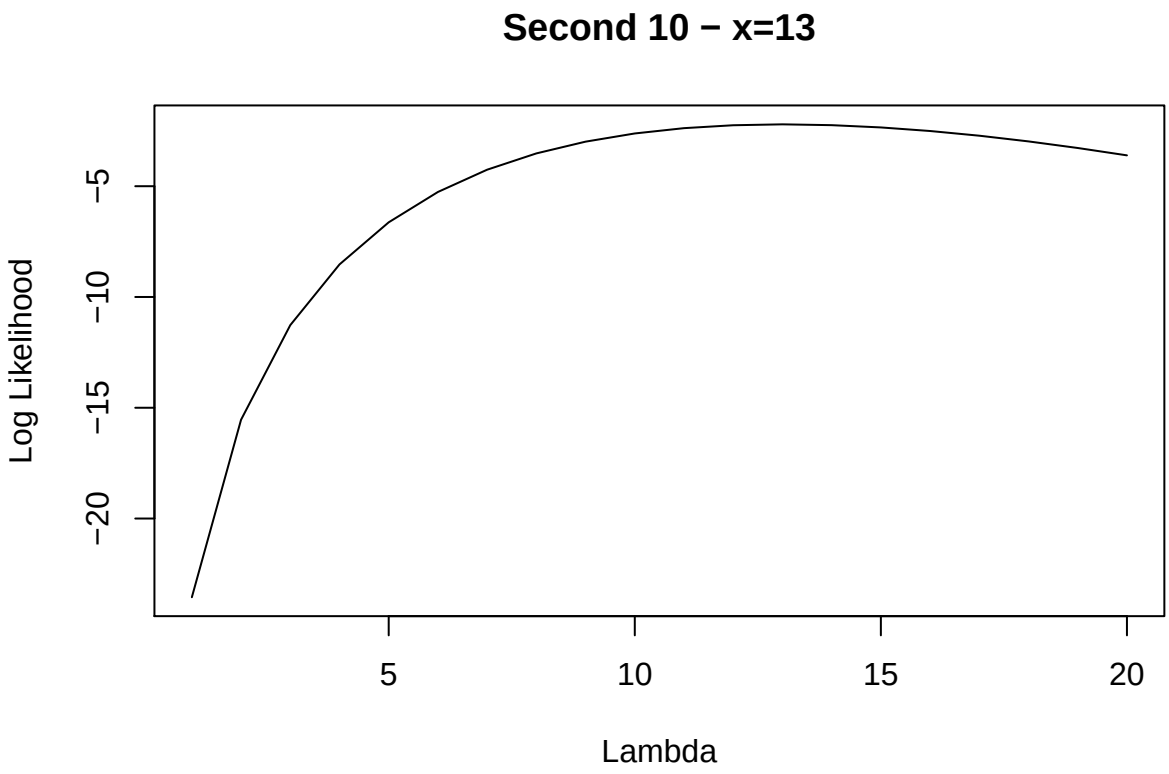
```
plot(L, s9, type="l", main="Second 9 - x=15", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s9)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 15"
```

```
plot(L, s10, type="l", main="Second 10 - x=13", xlab="Lambda", ylab="Log Likelihood")
```




```
print(paste0("log.likelihood is maximum at Lambda = : ", L[which.max(s10)]))
```

```
## [1] "log.likelihood is maximum at Lambda = : 13"
```

Explanation of results -

From above plots and which.max function results, it is verified that log likelihood is the maximum when $\lambda = x$. Thus we have verified that $\hat{\lambda} = x$.

All plots represent log curves which start at a low value, increase steadily, maximum value is at $\lambda = x$ and further increase in λ causes a decrease in log likelihood.

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.

```
# Load titanic data and convert to data.table
titanic <- read.csv("titanic.csv.bz2")
as.data.table(titanic)
```

```
##       pclass survived                                            name
##    1:      1        1                   Allen, Miss. Elisabeth Walton
##    2:      1        1                  Allison, Master. Hudson Trevor
```

```
##    3:      1        0                    Allison, Miss. Helen Loraine
##    4:      1        0            Allison, Mr. Hudson Joshua Creighton
##    5:      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels)
##   ---
## 1305:      3        0                           Zabour, Miss. Hileni
## 1306:      3        0                          Zabour, Miss. Thamine
## 1307:      3        0                      Zakarian, Mr. Mapriededer
## 1308:      3        0                            Zakarian, Mr. Ortin
## 1309:      3        0                             Zimmerman, Mr. Leo
##          sex     age sibsp parch ticket     fare   cabin embarked boat
##    1: female 29.0000     0     0  24160 211.3375      B5        S    2
##    2:   male  0.9167     1     2 113781 151.5500 C22 C26        S   11
##    3: female  2.0000     1     2 113781 151.5500 C22 C26        S
##    4:   male 30.0000     1     2 113781 151.5500 C22 C26        S
##    5: female 25.0000     1     2 113781 151.5500 C22 C26        S
##   ---
## 1305: female 14.5000     1     0   2665  14.4542               C
## 1306: female      NA     1     0   2665  14.4542               C
## 1307:   male 26.5000     0     0   2656   7.2250               C
## 1308:   male 27.0000     0     0   2670   7.2250               C
## 1309:   male 29.0000     0     0 315082   7.8750               S
##       body                        home.dest
##    1:   NA                      St Louis, MO
##    2:   NA Montreal, PQ / Chesterville, ON
##    3:   NA Montreal, PQ / Chesterville, ON
##    4:  135 Montreal, PQ / Chesterville, ON
##    5:   NA Montreal, PQ / Chesterville, ON
##   ---
## 1305:  328
## 1306:   NA
## 1307:  304
## 1308:   NA
## 1309:   NA
```

```
#transform 'survived', 'sex' and 'pclass' to categorical
titanic$survived <- factor(titanic$survived)
titanic$sex <- factor(titanic$sex)
titanic$pclass <- factor(titanic$pclass)
```

Answer -

Explanation of variables is as below -

survival - Survival. Values - 0 = No, 1 = Yes

pclass - Ticket class. Values - 1 = 1st, 2 = 2nd, 3 = 3rd. This is also a proxy for socio-economic status (SES) - 1st = Upper 2nd = Middle 3rd = Lower

sex - Sex/Gender of the passenger.

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

sibsp - Number of siblings / spouses aboard the Titanic. The dataset defines family relations in this way - Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiances were ignored)

parch - Number of parents / children aboard the Titanic. The dataset defines family relations in this way - Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.

ticket - Ticket number. This seems to be same value for a family. E.g. if 4 family members were travelling together, they received the ticket with same number.

fare - Passenger fare. This is a sum of fares paid by all passengers if a family received the same ticket number.

cabin - Cabin number. This may include multiple values if a family occupied more than 1 cabin.

embarked - Port of Embarkation. Values - C = Cherbourg, Q = Queenstown, S = Southampton

boat - This variable is not specified on Kaggle. This is likely the number of the rescue boat utilized by the passengers. Many of these values are blank, indicating that either those passengers did not have access to a rescue boat, or the data could not be collected, likely due to the chaos that ensued.

body - This variable is not specified on Kaggle. This is likely the number assigned to the body of the passenger if it was recovered. For the passengers that survived, or the passenger died but the body could not be recovered, this is blank.

home.dest - This variable is not specified on Kaggle. This is likely the intended final destination of the passenger.

Observing the nature of the data, following four variables appear to be the most important in this dataset -
survived
pclass
sex
age

```
# Analyze important variables
summary(titanic$survived)
```

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

```
summary(titanic$pclass)
```

```
##   1   2   3
## 323 277 709
```

```
summary(titanic$sex)
```

```
## female   male
##    466    843
```

```
summary(titanic$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1667 21.0000 28.0000 29.8800 39.0000 80.0000     263
```

```
mean(titanic$age, na.rm=TRUE)
```

```
## [1] 29.88113
```

survived - this could be our response variable or variable of interest. There are no missing values.

pclass - this is important since it in a way defined a passenger's social status, influence and also the access to the emergency rescue services. There are no missing values.

sex - This seems important because it would be interesting to analyze the effect of gender on survival. There are no missing values.

age - This seems important because it would be interesting to analyze the effect of age on survival. Did the older passangers have same odds of survival as the younger ones? There are 263 missing values. And the mean passenger age is 29.88 years (excluding missing values). Age range is 0.1667 (infant) to 80 years.

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

```
# Estimate a logistic regression with 'survived' as response variable
# and pclass, sex and age as predictors.
# pclass and sex are discrete variables while age is continuous.
glm1 <- glm(survived ~ pclass + sex + age, data=titanic,
          family=binomial(link="logit"))
summary(glm1)
```

```
##
## Call:
## glm(formula = survived ~ pclass + sex + age, family = binomial(link = "logit"),
##     data = titanic)
##
## 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
```

```
# Based on summary, we can also run Wald test on individual predictor
# Wald test on pclass
wald.test(b = coef(glm1), Sigma = vcov(glm1), Terms = 2:3)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 103.3, df = 2, P(> X2) = 0.0
```

```
# Wald test on sex
wald.test(b = coef(glm1), Sigma = vcov(glm1), Terms = 4)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 226.3, df = 1, P(> X2) = 0.0
```

```r
# Wald test on age
wald.test(b = coef(glm1), Sigma = vcov(glm1), Terms = 5)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 29.5, df = 1, P(> X2) = 5.6e-08
```

```r
## odds ratios and 95% CI
exp(cbind(odds.ratio = coef(glm1), confint(glm1)))
```

```
## Waiting for profiling to be done...
```

```
##              odds.ratio       2.5 %     97.5 %
## (Intercept) 33.85457044 18.11476468 65.2744468
## pclass2      0.27787894  0.17763418  0.4303922
## pclass3      0.10130084  0.06453100  0.1565287
## sexmale      0.08226211  0.05906841  0.1133112
## age          0.96619149  0.95410521  0.9781055
```

Answer -

Based on summary and Wald test, we see that p-value for all above predictors are less than significant value of 0.05. So we reject null hypothesis for all above predictors and establish that all above predictors have a statistically significant relationship with the response variable.

Odds ratio analysis -

Looking at the odds ratios, pclass has significant effect on survival. The odds ratio of pclass2 is 0.2778 which means that a passenger in second class had about 27.78% odds of survival when compared to passenger in first class. The odds ratio of pclass3 is 0.1013 which means that a passenger in third class had about 10.13% odds of survival when compared to passenger in first class. These are significant effects.

sexmale has an odd ratio of 0.0822 or 8.22% when compared to females. This indicates that females have significantly higher odds of surviving the titanic sinking.

As age increases by an year, the odds of surviving become 96%. This indicates that younger passengers have a slightly better chance of surviving the incident.

Possible further analysis - not done for this question

We can conduct further research by predicting the model, and creating a plot of actual and predicted values. This is not being done for this assignment. We will stop at interpreting the results.

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?

Answer -

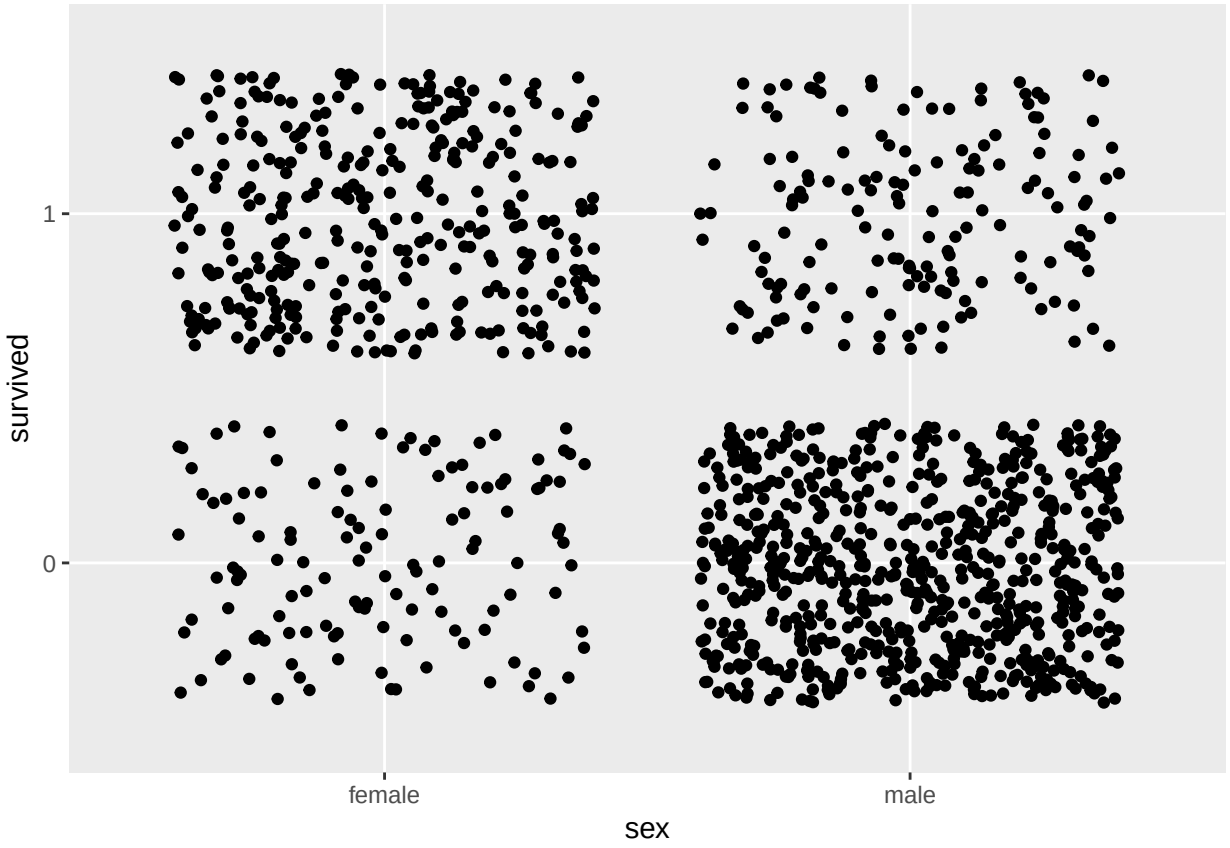
Our analysis from Q2.2 does indicate that females have very high odds of survival when compared to males. From above analysis -

sexmale has an odd ratio of 0.0822 or 8.22% when compared to females. This indicates that females have significantly higher odds of surviving the titanic sinking.

This is not surprising to me. It is very natural that women and children were the first to be rescued, followed by the men.

A simple plot can also confirm the same thing.

```
# Simple plot of survived vs sex.
ggplot(titanic, aes(sex, survived)) +
geom_jitter()
```



From this chart, we can easily conclude that females had much better survival rate than males.

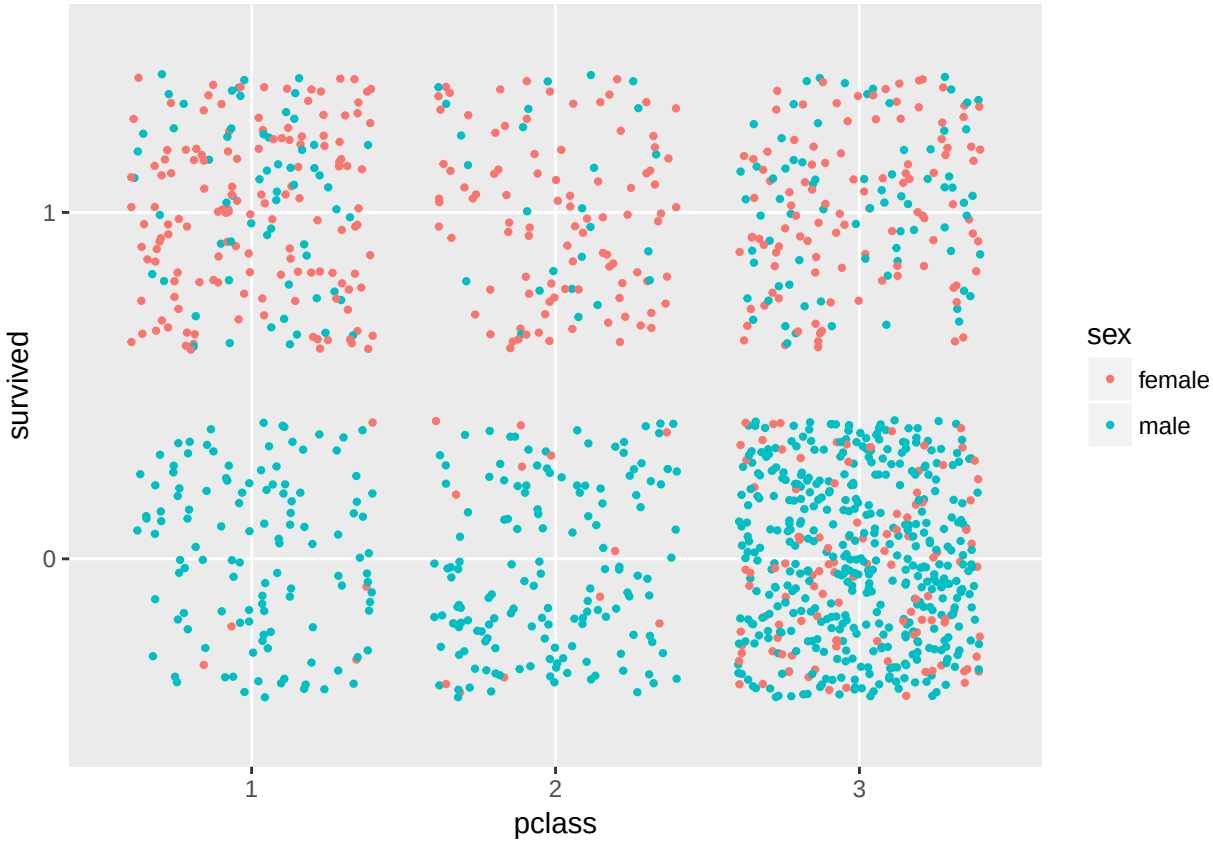
This indicates that the final moments of titanic were not as random as one would think. Probably there was chaos, but the rescue was systematic and clearly women were given preference. Order was followed even in those chaotic final moments.

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.

Answer -

One way of introducting interaction effects is to create a simple plot as shown below.

```
# Simple plot of pclass vs sex colored by survival.
ggplot(titanic, aes(pclass, survived, color = sex)) +
   geom_jitter(size = 0.8)
```



Based on this plot, we can draw following conclusions -

a. In all passanger classes, female survival rate is better than males for a specific passenger class. This indicates that women were rescued with priority.

b. Males in third class constitute the biggest share of the victims (not survived).

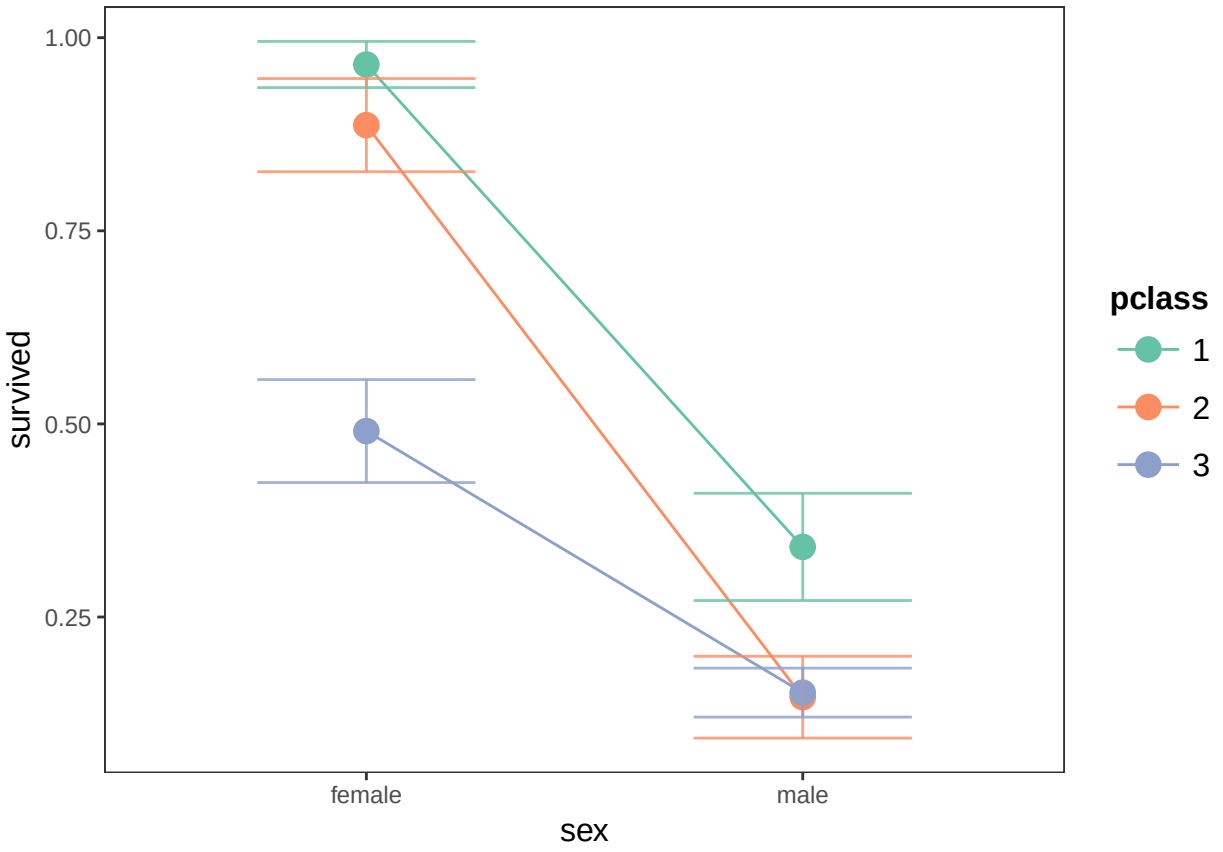
c. Almost all women in first class survived. The chances of survival are extremely high in this group.

Another way of introducing interaction effects is to create an interaction model with survival as response and sex and pclass as predictors.

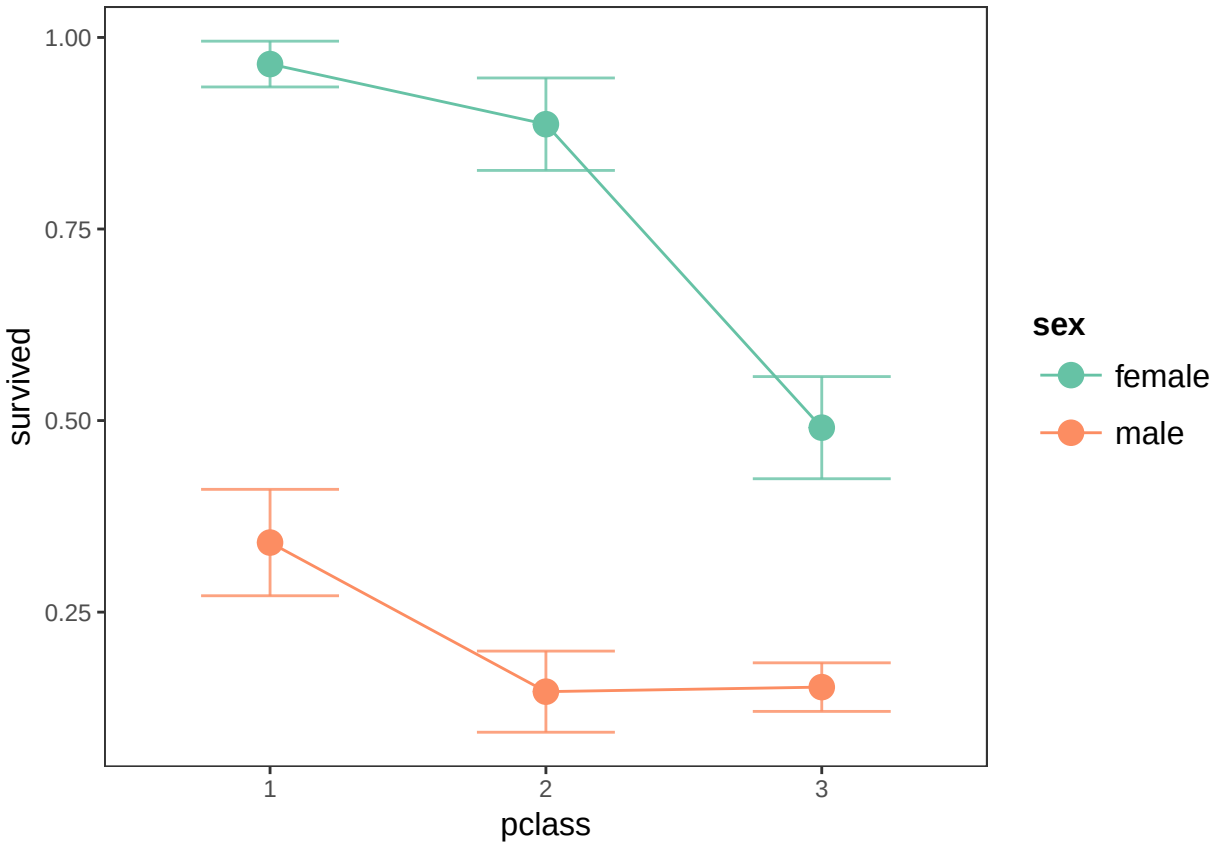
```
# Create an interaction model.
interaction.model <- glm(survived ~ sex * pclass, data=titanic,
          family=binomial(link="logit"))
coef(summary(interaction.model))
```

```
##                  Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)      3.325035  0.4548995  7.3093844 2.683691e-13
## sexmale         -3.984846  0.4814577 -8.2766265 1.267139e-16
## pclass2         -1.266647  0.5485482 -2.3090896 2.093861e-02
## pclass3         -3.362076  0.4748246 -7.0806701 1.434590e-12
## sexmale:pclass2  0.161727  0.6104281  0.2649403 7.910555e-01
## sexmale:pclass3  2.303894  0.5158024  4.4666212 7.946466e-06
```

```
# Create a cat_plot of survived, pclass and sex using interaction model
cat_plot(interaction.model, pred = sex, modx = pclass, geom="line")
```




```
# Create a second cat_plot of survived, pclass and sex using interaction model
cat_plot(interaction.model, pred = pclass, modx = sex, geom="line")
```




```
## odds ratios and 95% CI
exp(cbind(odds.ratio = coef(interaction.model), confint(interaction.model)))
```

```
## Waiting for profiling to be done...
```

```
##                  odds.ratio        2.5 %      97.5 %
## (Intercept)     27.79997633 12.672471481 78.42424430
## sexmale          0.01859531  0.006336619  0.04350195
## pclass2          0.28177482  0.087279914  0.78657939
## pclass3          0.03466321  0.011933127  0.07986704
## sexmale:pclass2  1.17553928  0.368905932  4.19992901
## sexmale:pclass3 10.01309646  3.946683443 30.96235468
```

Analysis of interaction model-

Looking at above plots, which are self explanatory, our conclusions from previous plot are further reinforced.

However, we learn few new insights from these plots. Males in second and third class had almost identical odds of survival. In fact, males from class 3 had very slight better odds of survival than males from class 2.

Also, we learn that female odds of survival decline sharply from second to third class. 3rd class females had significantly less odds of survival from females in second class.

Odds ratio analysis of interaction model confirms our conclusions.

sexmale odd ratio is significantly less compared to sexfemale (implied).
pclass2 and pclass3 odd ratios are significantly less compared to pclass1 (implied).

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?

Answer -

We will select following less obvious variables for this analysis -

fare - This is not a good choice for analysis, if fare is used as is. Fare may be same for all family members, and is not a passenger level attribute. For fare to be used in a better way, we need to calculate fare for each passenger using the ticket number. Also, the exact fare for individual passengers is unknown, so we would have to calculate mean price per passenger. But for the purpose of this assignment, we will use the fare as is and analyze the results. This is a continuous variable.

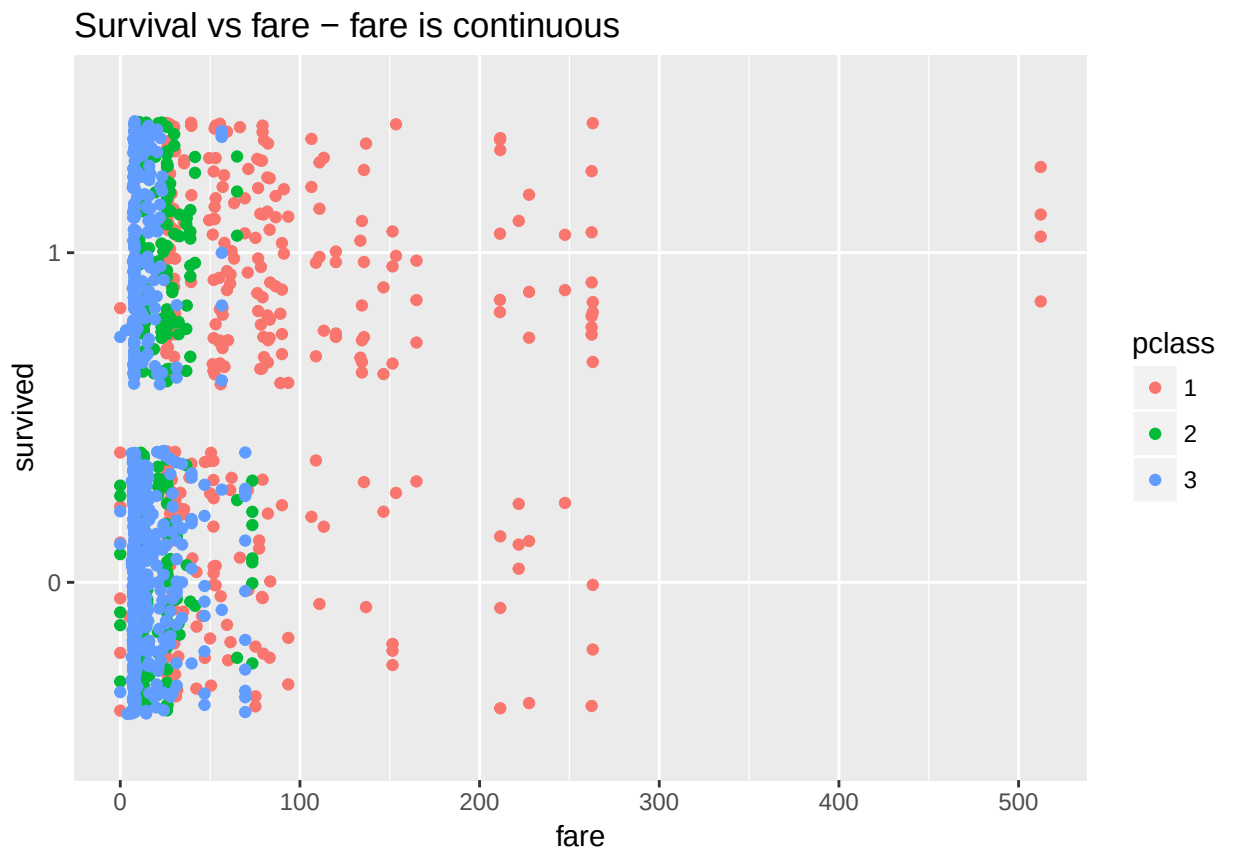
embarked - Discrete variable. Two records have blank values.

sibsp - Continuous variable.

parch - Continuous variable.

We will begin by creating simple plots, and carrying out Chi Square test individually on each variable.

```
# Plot of survival vs fare colored by pclass.
ggplot(titanic, aes(fare, survived, color=pclass)) +
      geom_jitter() +
      labs(title = "Survival vs fare - fare is continuous")
```

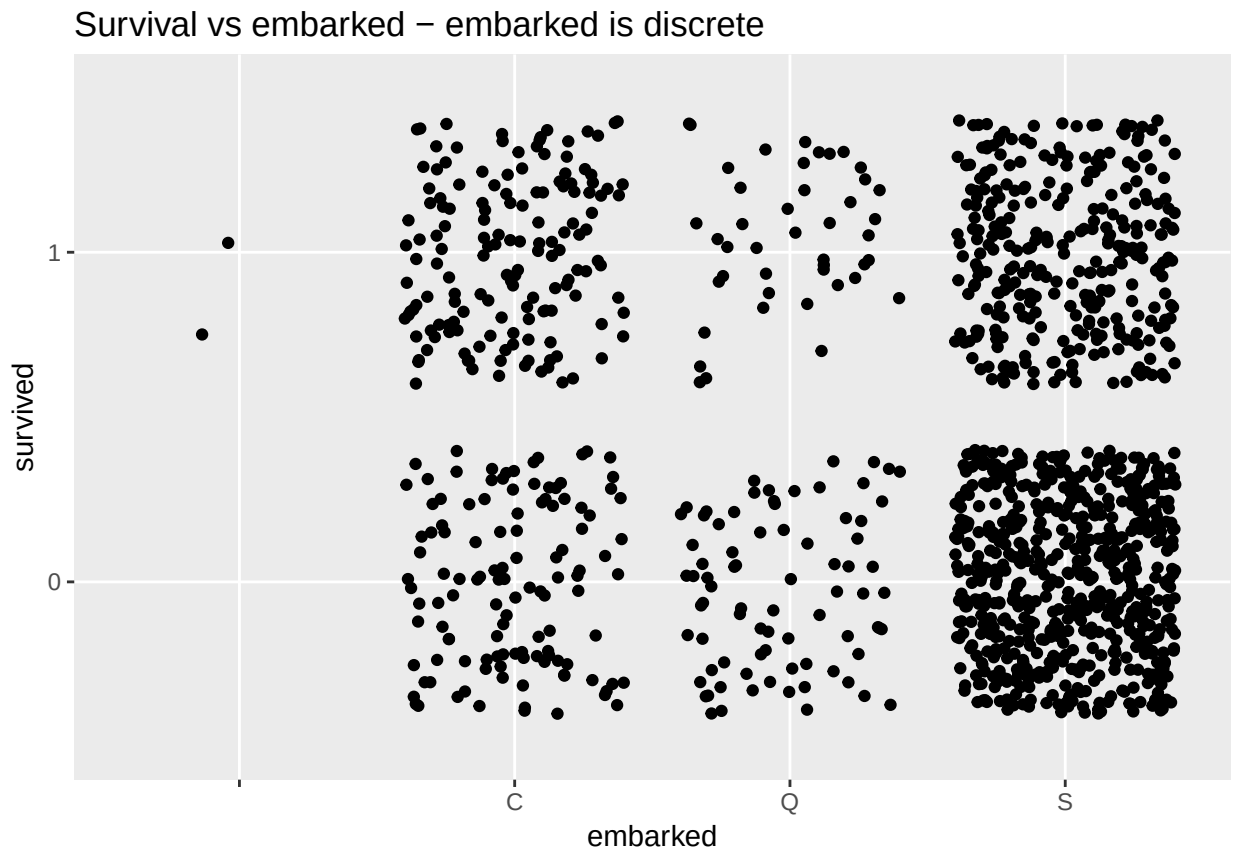



```
# Fare chi-square test
chisq.test(titanic$fare, titanic$survived)
```

```
##
##  Pearson's Chi-squared test
##
## data:  titanic$fare and titanic$survived
```

```
## X-squared = 564.05, df = 280, p-value < 2.2e-16
```

```
# Plot of survival vs embarked.
ggplot(titanic, aes(embarked, survived)) +
      geom_jitter() +
      labs(title = "Survival vs embarked - embarked is discrete")
```




```
# Embarked chi-square test
chisq.test(titanic$embarked, titanic$survived)
```

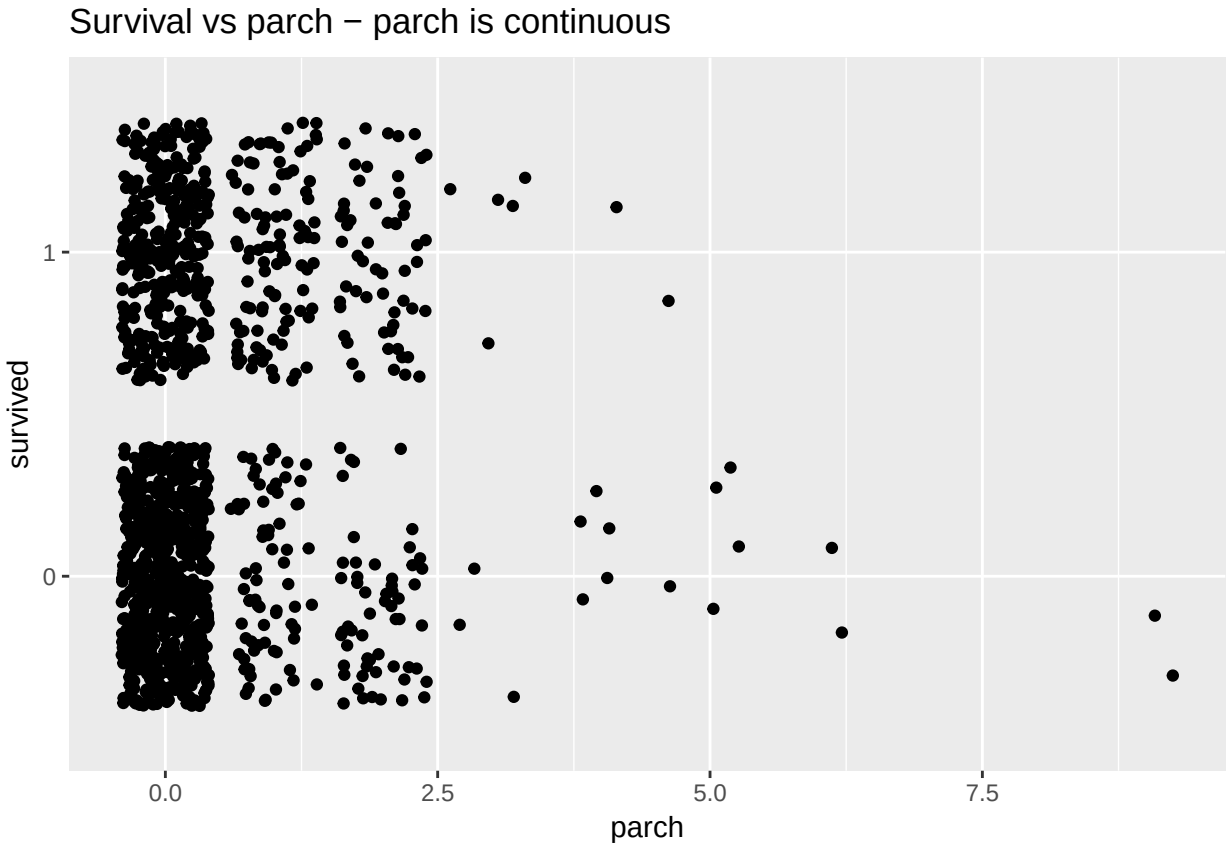
```
##
##  Pearson's Chi-squared test
##
## data:  titanic$embarked and titanic$survived
## X-squared = 47.441, df = 3, p-value = 2.801e-10
```

```
# Plot of survival vs sibsp.
ggplot(titanic, aes(sibsp, survived)) +
      geom_jitter() +
      labs(title = "Survival vs sibsp - sibsp is continuous")
```


```
# sibsp chi-square test
chisq.test(titanic$sibsp, titanic$survived)
```

```
##
##  Pearson's Chi-squared test
##
## data:  titanic$sibsp and titanic$survived
## X-squared = 43.499, df = 6, p-value = 9.289e-08
```

```
# Plot of survival vs parch.
ggplot(titanic, aes(parch, survived)) +
      geom_jitter() +
      labs(title = "Survival vs parch - parch is continuous")
```




```
# parch chi-square test
chisq.test(titanic$parch, titanic$survived)
```

```
##
##  Pearson's Chi-squared test
##
## data:  titanic$parch and titanic$survived
## X-squared = 53.879, df = 7, p-value = 2.485e-09
```

We can also create a logistic regression for above variables and analyze the results.

```
# Logistic regression - survived vs fare, embarked
# sibsp and parch
glm2 <- glm(survived ~ fare + embarked + sibsp + parch,
            data=titanic, family=binomial(link="logit"))
summary(glm2)
```

```
##
## Call:
## glm(formula = survived ~ fare + embarked + sibsp + parch, family = binomial(link = "logit"),
##     data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4432  -0.8883  -0.8269   1.2673   1.7906
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  12.654816 378.592894   0.033  0.97333
## fare          0.011391   0.001731   6.580 4.72e-11 ***
## embarkedC   -13.008957 378.592898  -0.034  0.97259
## embarkedQ   -13.337512 378.592936  -0.035  0.97190
## embarkedS   -13.638241 378.592891  -0.036  0.97126
## sibsp        -0.191820   0.067235  -2.853  0.00433 **
## parch         0.163388   0.075223   2.172  0.02985 *
## ---
## 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: 1621.5  on 1301  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 1635.5
##
## Number of Fisher Scoring iterations: 12
```

```
## odds ratios and 95% CI
exp(cbind(odds.ratio = coef(glm2), confint(glm2)))
```

```
## Waiting for profiling to be done...
```

```
##               odds.ratio        2.5 %       97.5 %
## (Intercept) 3.132684e+05 4.852677e-21           NA
## fare        1.011456e+00 1.008163e+00 1.015040e+00
## embarkedC   2.240174e-06           NA 1.442849e+20
## embarkedQ   1.612843e-06           NA 1.016922e+20
## embarkedS   1.193953e-06           NA 7.721417e+19
## sibsp       8.254553e-01 7.193820e-01 9.374175e-01
## parch       1.177494e+00 1.017707e+00 1.368893e+00
```

Based on above analysis -

a. Fare - P-value indicates that this is a statistically significant relationship. Looking at color plot, lower fare is mostly associated with 2nd and 3rd class, while higher fare is mostly associated with higher class. This creates a statistically significant relationship with survival.

b. embarked - P-value indicates that this is not a statistically significant relationship with survival.

c. sibsp - P-value indicates that this is a statistically significant relationship with survival. This may be a little surprising. The plot also confirms the finding. Based on odds ratio, an increase of 1 sibling causes the odds of survival to decrease by about 17.5%. I cannot think of a rational explanation. But it is possible that for larger families with a lot of siblings, the chaos caused more deaths.

d. parch - P-value indicates that this is a statistically significant relationship with survival, though very marginally. Addition of a parent/caretaker seems to increase the odds of survival by 17.7%. This could be due to the fact that parents preferred their children's survival over their own.

3. How much work?

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

Answer -

Expected time - 10 hours

Actual time - **21 hours**