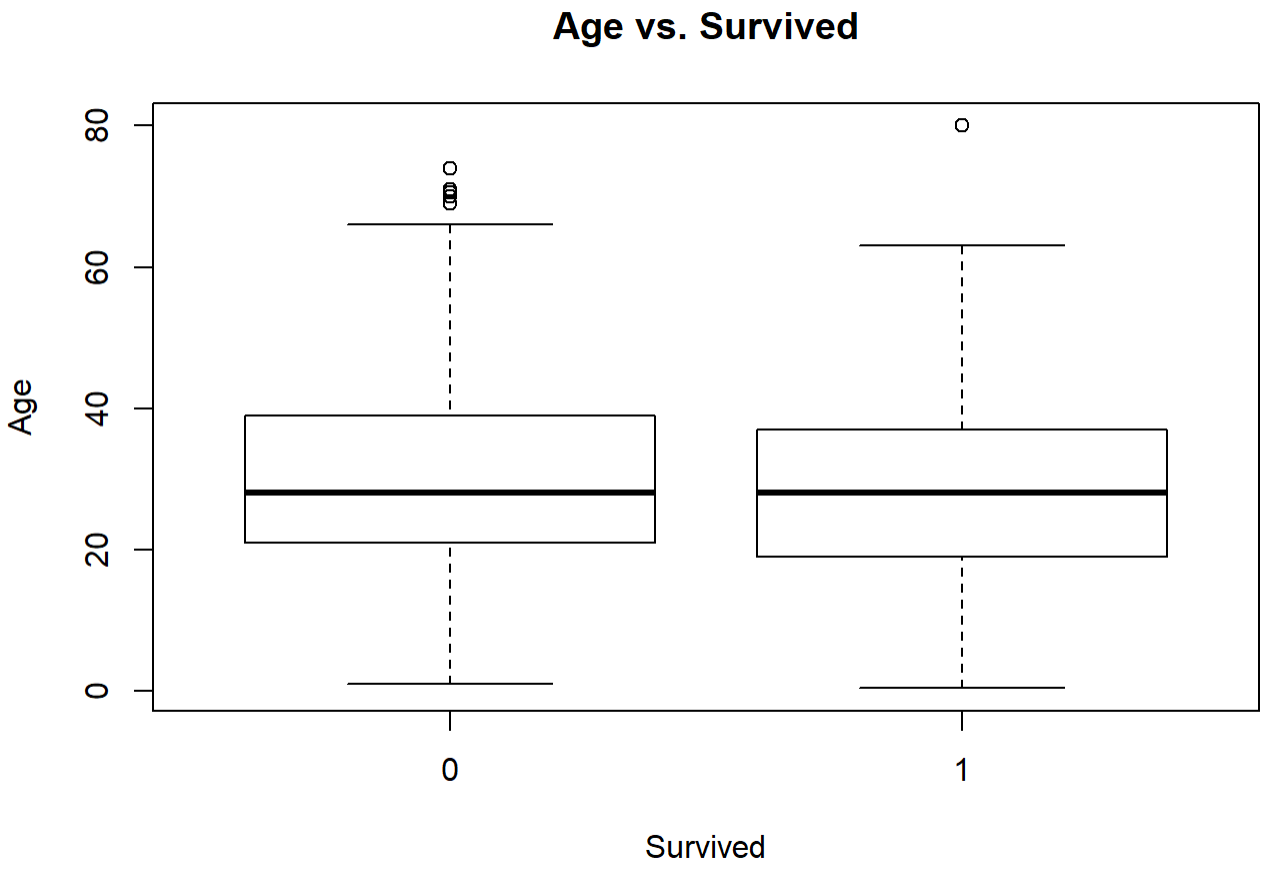
HW4 Peer Assignment

Question 1: Exploratory Data Analysis

1. Using boxplots explore the relationship between survived and the numerical independent variables: Age and Fare . Can you observe differences in distribution of the predictors between the 2 classes? Please explain and interpret.If you cannot determine visually please observe the mean/median of the predictors by the 2 classes: for example: summary(data[data$Survived==1,“Age”])

```
rm(list=ls())
data = read.csv('titanic.csv', header = T, sep = ',')
boxplot(Age ~ Survived, data, xlab = 'Survived', ylab = 'Age', main = 'Age vs. Surviv
ed')
```



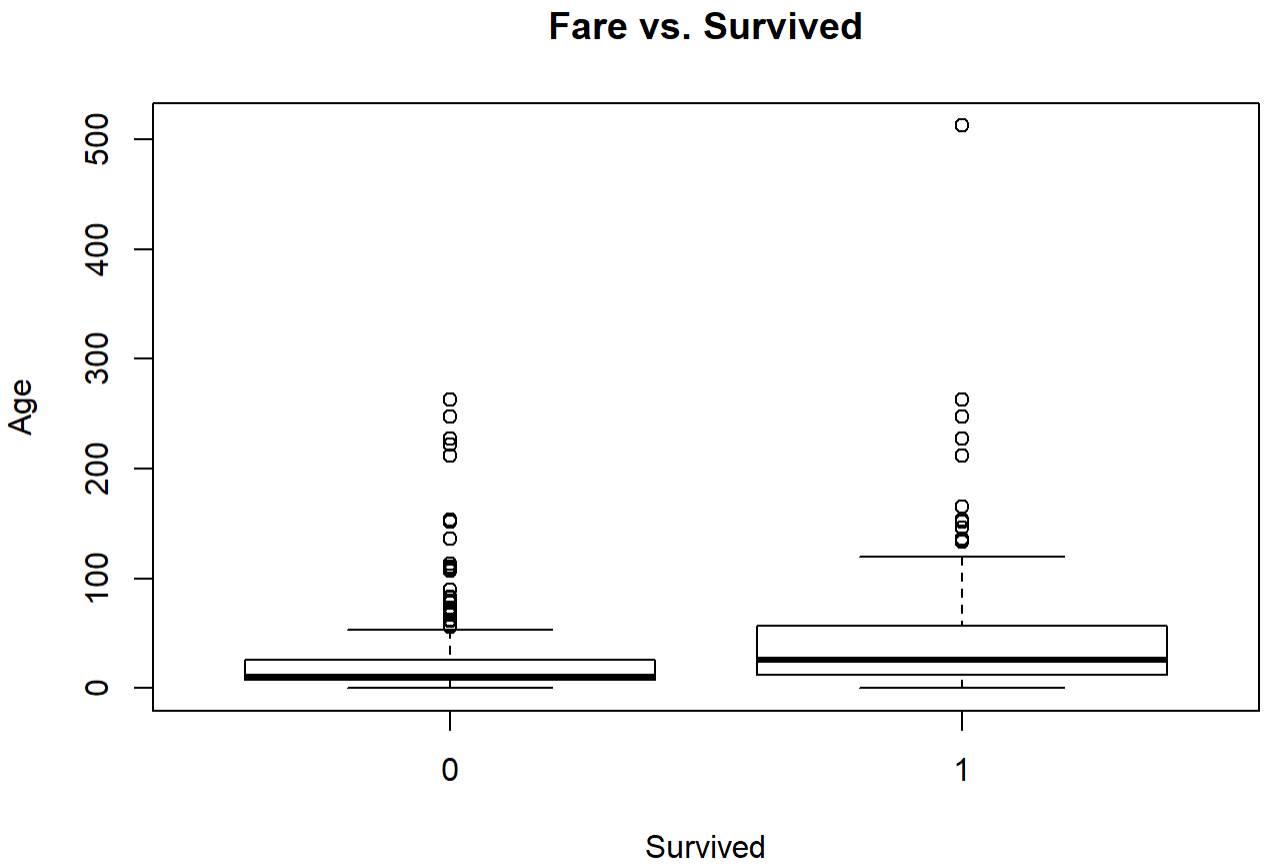

```
summary(data[data$Survived==1,'Age'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.42   19.00   28.00   28.41   36.75   80.00
```

```
summary(data[data$Survived==0,'Age'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   21.00   28.00   30.14   39.00   74.00
```

```
boxplot(Fare ~ Survived, data, xlab = 'Survived', ylab = 'Age', main = 'Fare vs. Surv
ived')
```




```
summary(data[data$Survived==1,'Fare'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.47   26.00   48.40   57.00  512.33
```

```
summary(data[data$Survived==0,'Fare'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.854  10.500  22.209  26.000 263.000
```

Answer: As for Age, we found the median ages are about 28.00 of the passengers who are survived or not survived, there are almost no differences of ages between

them based on boxplot or summarized analysis. In addition, we found most passengers had low fares, but there are some differences between survived and not survived, the numbers of survived passengers with high fares are more than passengers with low fares.

2. Modify the Sib_sp and par_ch variables so that any passenger having 4 or more of each variable is coded “above_4”(Hint: use ifelse).Describe the relationship between Survived and the categorical independent variables Pclass, Sex, Sib_sp and Par_ch. Does the survival rate vary with the categorical variables? Please interpret.

One way of doing this is a contingency table followed by a Chi-squared test. A more visual way would be to observe the % response rates w.r.t levels of the predictor. You can use the following code to plot a barchart of response rates vs a predictor:

```
data$Sib <- ifelse(data$Sib_sp >= 4, 'above_4', 'below_4')
tb_Sib = xtabs(~Survived + Sib, data)
tb_Sib
```

```
##          Sib
## Survived above_4 below_4
##        0      27     518
##        1       3     339
```

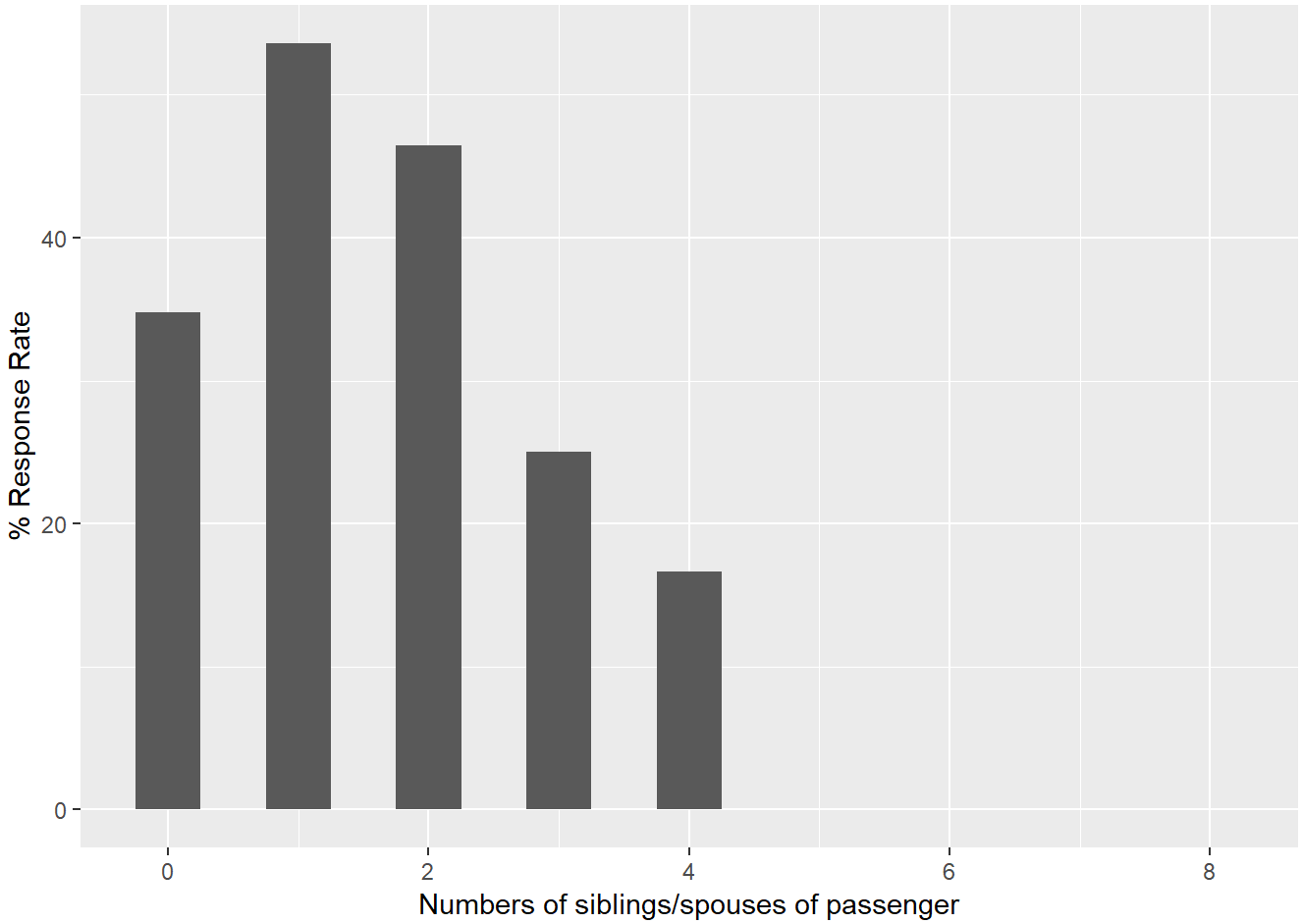
```
chisq.test(tb_Sib)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb_Sib
## X-squared = 9.4772, df = 1, p-value = 0.00208
```

```
library(plyr)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
ggplot(ddply(data,.(Sib_sp),summarise, rr=100*sum(Survived)/length(Survived)),
       aes(x=Sib_sp,y=rr))+geom_bar(stat = "identity",width=0.5)+
       labs(x="Numbers of siblings/spouses of passenger", y="% Response Rate")
```



Answer: From the Sib_sp data, we found that most passengers have the numbers of siblings/spouses below 4 people (857/887). Some of them are survived (339/(339+518)) if the numbers of siblings/spouses are less than 4. However, only a few of them are survived (3/(3+27)) if the numbers of siblings/spouses are above 4. In addition, by using chi-squred test, p-value is 0.00208, which is less than 0.05 significance level, we can reject the null hypothesis that survived passengers are independent of number of siblings/spouse. The conclusions are consistent with the visualized figure above.

```
data$Par <- ifelse(data$Par_ch >= 4, 'above_4', 'below_4')
tb_Par = xtabs(~Survived + Par, data)
tb_Par
```

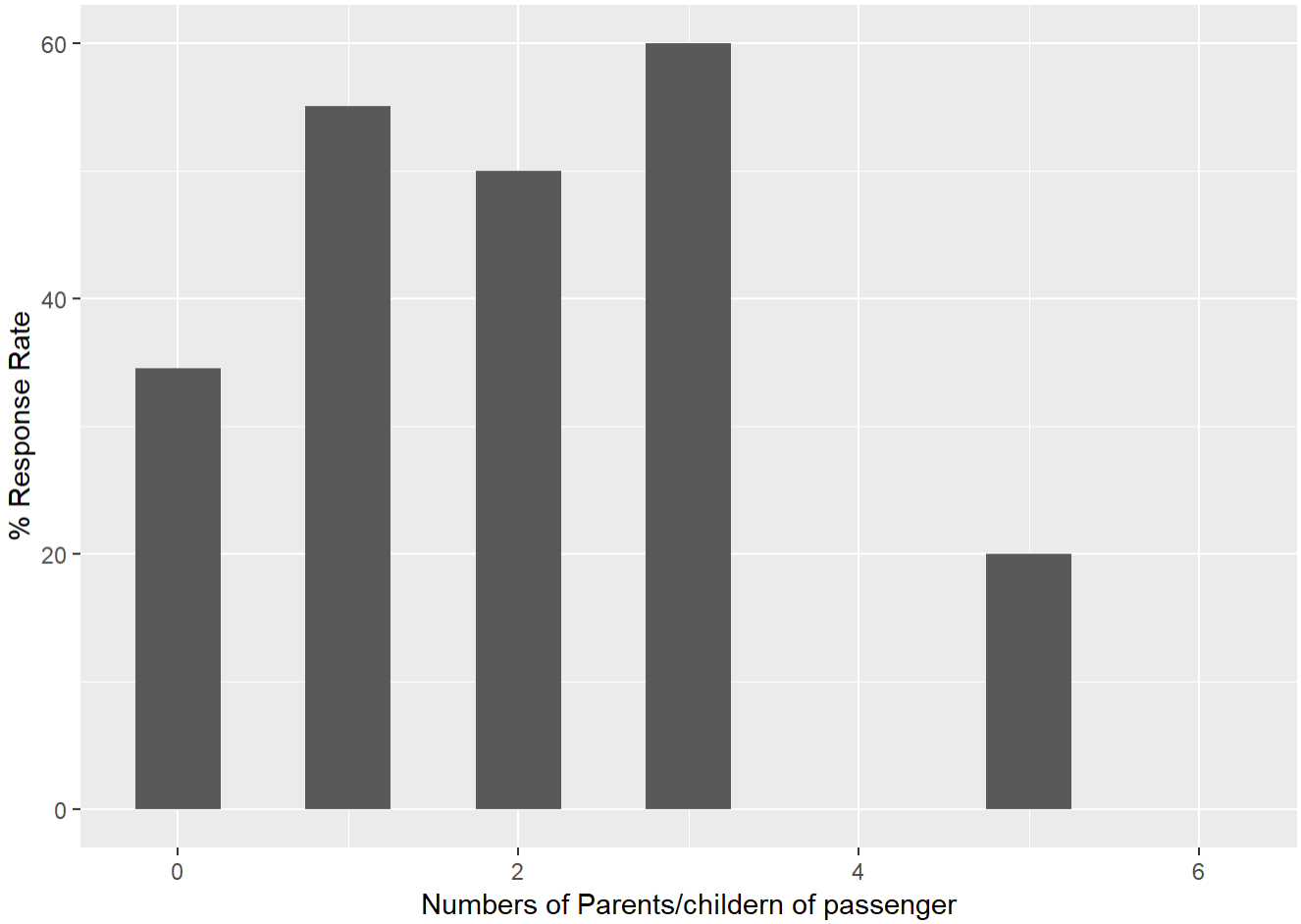
```
##          Par
## Survived above_4 below_4
##        0       9     536
##        1       1     341
```

```
chisq.test(tb_Par)
```

```
## Warning in chisq.test(tb_Par): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb_Par
## X-squared = 2.3691, df = 1, p-value = 0.1238
```

```
ggplot(ddply(data,.(Par_ch),summarise, rr=100*sum(Survived)/length(Survived)),
       aes(x=Par_ch,y=rr))+geom_bar(stat = "identity",width=0.5)+
       labs(x="Numbers of Parents/childern of passenger", y="% Response Rate")
```



Answer: From the Par_sp data, we found that most passengers have the numbers of parents/children less than 4 people (877 vs. 10), some of them are survived (341/(341+536)) if the number less than 4. However, if the numbers of parents/children are above 4, only few of them are survived (1/(1+9)). In addition, from the result of chi-squared test, the p-value is 0.1238, which is more than 0.05 significance level, thus we do not reject the null hypothesis that the survived passengers are independence of the number of parents/children. The conclusions are consistent with the visualized figure above.

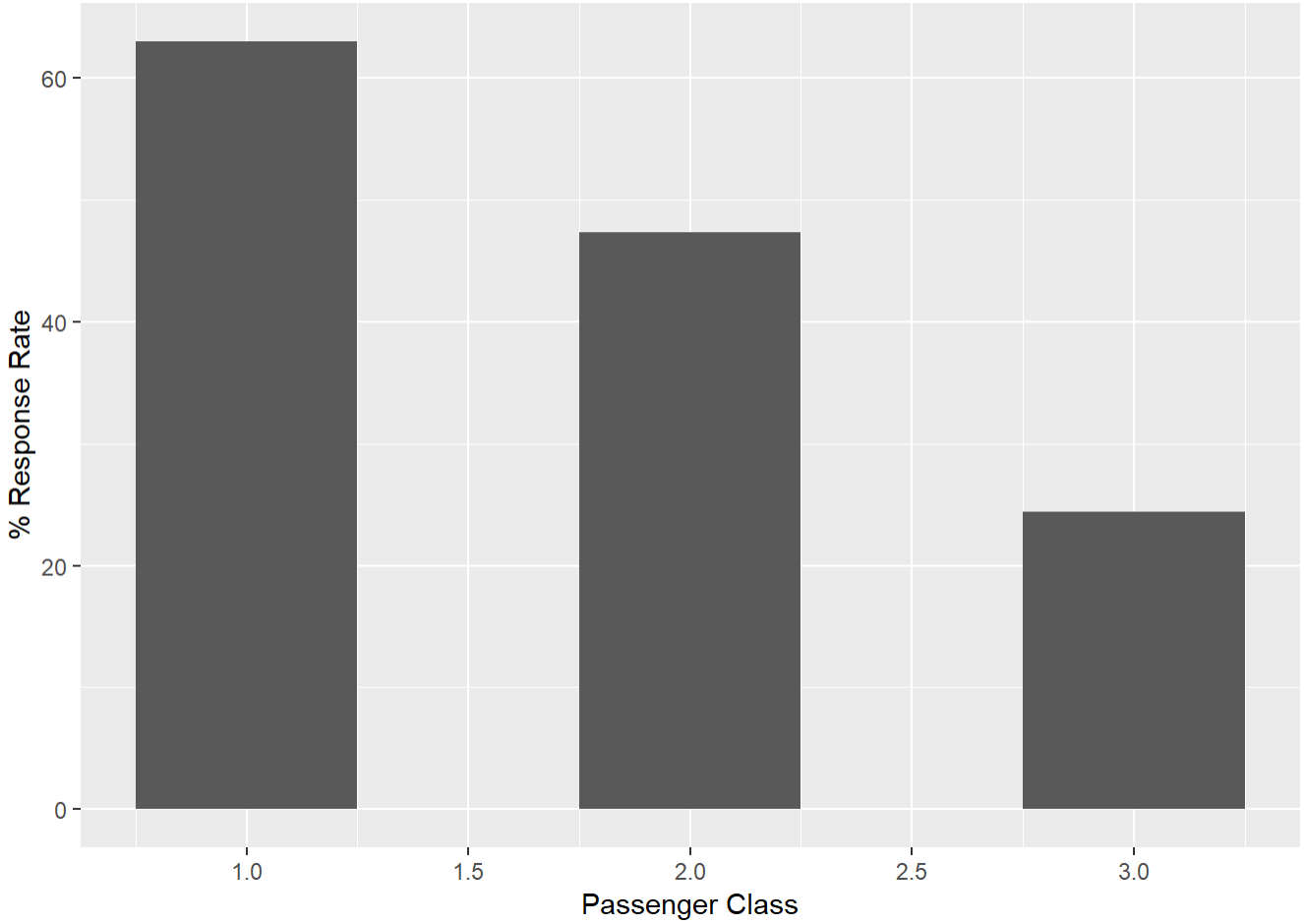
```
tb_Class = xtabs(~Survived + Pclass, data)
tb_Class
```

```
##          Pclass
## Survived   1   2   3
##        0  80  97 368
##        1 136  87 119
```

```
chisq.test(tb_Class)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tb_Class
## X-squared = 101.22, df = 2, p-value < 2.2e-16
```

```
ggplot(ddply(data,.(Pclass),summarise, rr=100*sum(Survived)/length(Survived)),
       aes(x=Pclass,y=rr))+geom_bar(stat = "identity",width=0.5)+
       labs(x="Passenger Class", y="% Response Rate")
```



Answer: From the Pclass data, we found that most of passengers have cheap ticket class ((368+119)/887). If the passengers have the cheap tickets (ticket class 3), the numbers of survived passengers are lower than the passenger who are not survived (119 vs. 368). However, if the passengers have expensive tickets (class 1), the number of survived passengers are more than not survived (136 vs. 80). In addition,

from the result of chi-squared test, the p-value is 2.2e-16, which is far lower than 0.05 significance level, so we do reject the null hypothesis that the number of survived passengers are independence of the ticket class. The conclusions are consistent with the visualized figure above.

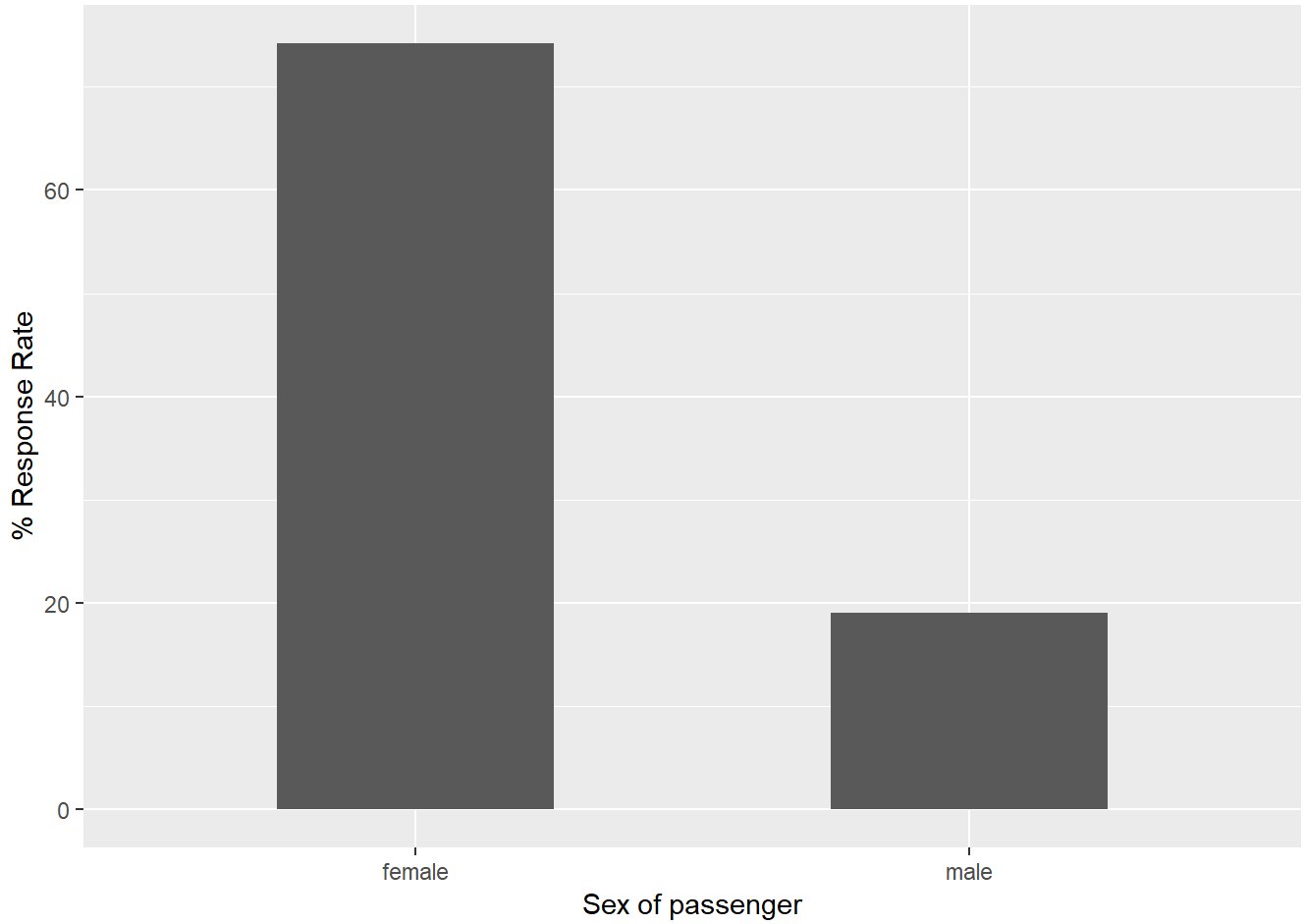
```
tb_Sex = xtabs(~Survived + Sex, data)
tb_Sex
```

```
##          Sex
## Survived female male
##        0     81  464
##        1    233  109
```

```
chisq.test(tb_Sex)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb_Sex
## X-squared = 258.39, df = 1, p-value < 2.2e-16
```

```
ggplot(ddply(data,.(Sex),summarise, rr=100*sum(Survived)/length(Survived)),
       aes(x=Sex,y=rr))+geom_bar(stat = "identity",width=0.5)+
       labs(x="Sex of passenger", y="% Response Rate")
```



Answer: In overview, we found male passengers are more than female. If the passengers are male, the numbers of survived passengers are less than not survived (109 vs. 464). but if the passengers are female, the survived passengers are more than not survived (233 vs. 81). In addtion, from the result of chi-squared test, the p-value is 2.2e-16, which is far lower than 0.05 significance level, so we do reject the null hypothesis tha the number of survived passengers are independence of the sex (gender). The conclusions are consistent with the visualized figure above.

3. Based on your findings, you want to build a logistic regression model to predict the probabilities of passenger survival given the attributes. Briefly state the model and its assumptions.

Answer: prop.surv = Survived / passengers (total)

log(prop.surv / 1- prop.surv) = b0 + b1 * Sib_sp + b2 * Pclass + b3 * Sex + b4 * Fare + b5 * Age

Based on our above results, the numbers of survived passenger are probably related to several categorical predicting variables such as Sib_sp, Par_ch, Pclass and Sex, and also related to numerical predicting variables such as Fare and Age. The odd is the ratio of probability of survived and the probability of not survived, and the logistic regression model is shown as above. The assumptions of losgistic regression model including Linearity assumption after logit transformation, response variables are independent random variables, and the logit link funciton g(p) = ln(p/(1-p)).

Question 2: Fitting Regression Model

1. Convert Pclass and Sib_sp to factor variables. Fit a logistic regression model on Survived as the response and Pclass, Sex, Age and Sib_sp as predictors. What are the model parameters and estimates?

```
rm(list=ls())
data = read.csv('titanic.csv', header = T, sep = ',')
str(data)
```

```
## 'data.frame':    887 obs. of  8 variables:
##  $ Survived: int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name    : Factor w/ 887 levels "Capt. Edward Gifford Crosby",..: 602 823 172 81
4 733 464 700 33 842 839 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 27 54 2 27 14 ...
##  $ Sib_sp  : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Par_ch  : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
table(data$Pclass)
```

```
##
##   1   2   3
## 216 184 487
```

```
table(data$Sib_sp)
```

```
##
##   0   1   2   3   4   5   8
## 604 209  28  16  18   5   7
```

```
data$Pclass = factor(data$Pclass, labels = c('1', '2', '3'))
data$Sib_sp = factor(data$Sib_sp, labels = c('0','1','2', '3', '4', '5', '8'))
model1 = glm(Survived ~ Pclass + Sex + Age + Sib_sp, data = data, family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Sib_sp, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8980  -0.5940  -0.3956   0.6135   2.4900
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.195506   0.425947   9.850  < 2e-16 ***
## Pclass2      -1.356275   0.271118  -5.003 5.66e-07 ***
## Pclass3      -2.492078   0.261428  -9.533  < 2e-16 ***
## Sexmale      -2.712272   0.197060 -13.764  < 2e-16 ***
## Age          -0.045449   0.007924  -5.735 9.73e-09 ***
## Sib_sp1       0.079540   0.211862   0.375 0.707338
## Sib_sp2      -0.204192   0.520051  -0.393 0.694586
## Sib_sp3      -2.351357   0.682375  -3.446 0.000569 ***
## Sib_sp4      -1.714913   0.743373  -2.307 0.021058 *
## Sib_sp5     -16.028132 958.557438  -0.017 0.986659
## Sib_sp8     -16.504894 750.841533  -0.022 0.982462
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1182.77  on 886  degrees of freedom
## Residual deviance:  770.72  on 876  degrees of freedom
## AIC: 792.72
##
## Number of Fisher Scoring iterations: 15
```

2. Write down the equation for the logarithm of odds of survival given the predicting variables.

Answer: log(odds) = 4.195506 - 1.356275 * Pclass2 - 2.492078 * Pclass3 - 2.712272 * Sex(male) -0.045449 * Age + 0.079540 * Sib_sp1 - 0.204192 * Sib_sp2 - 2.351357 * Sib_sp3 - 1.714913 * Sib_sp4 - 16.028132 * Sib_sp5 - 16.504894 * Sib_sp8

3. Interpret the coefficients of Pclass, Sex and Age

Answer: For Pclass, the coefficient of Pclass2 is -1.356275, the base case is the passengers with ticket class 1. It means that if passengers have ticket class 2, the log odds of survived decreased by 1.356275 (or the odds of survived decreased by 1-exp(-1.356275) = 0.7423814) versus passenger with ticket class 1 given that all other predicting variables fixed. If passengers have ticket class 3 (coefficient is -2.492078), the log odds of survived decreased by 2.492078 (or the odds of survived decreased by 1-exp(-2.492078) = 0.9172621) versus passengers with ticket class 1 given all other predicting variables fixed.

For sex, the coefficient of sex(man) is -2.712272, the base case is female passengers. The coefficient means that the log odds of survived decreased by 2.712272 (or the odds of survived decreased by 1-exp(-2.712272) = 0.9336142) versus female passengers given that all other predicting variables fixed.

For age, the coefficient of Age is -0.045449, which means that when age of passengers increase one unit, the log odds of survived decreased by 0.045449 (or the odds of survived decreased by 1-exp(-0.045449)= 0.04443166) given that all other predicting variables fixed.

Question 3: Inference

1. Find a 95% confidence interval for the parameters corresponding to all predictors plus the intercept.

Answer:

(1). **For Odds ratio**, the 95% confidencen intervals are below.

```
result1 <- exp(cbind("Odds ratio" = coef(model1), confint.default(model1, level = 0.9
5)))
result1
```

```
##               Odds ratio       2.5 %       97.5 %
## (Intercept) 6.638731e+01 28.80832448 152.98617524
## Pclass2     2.576186e-01  0.15142621   0.43828177
## Pclass3     8.273783e-02  0.04956508   0.13811233
## Sexmale     6.638580e-02  0.04511672   0.09768162
## Age         9.555686e-01  0.94084189   0.97052573
## Sib_sp1     1.082789e+00  0.71483652   1.64014033
## Sib_sp2     8.153057e-01  0.29420641   2.25937747
## Sib_sp3     9.523984e-02  0.02500231   0.36279155
## Sib_sp4     1.799793e-01  0.04192395   0.77265040
## Sib_sp5     1.094134e-07  0.00000000          Inf
## Sib_sp8     6.792283e-08  0.00000000          Inf
```

(2). **For log(Odds ratio)**, 95% confident intervals are below.

```
result2 <- cbind("Odds ratio" = coef(model1), confint.default(model1, level = 0.95))
result2
```

```
##               Odds ratio         2.5 %        97.5 %
## (Intercept)   4.19550597  3.360664e+00    5.03034756
## Pclass2      -1.35627504 -1.887657e+00   -0.82489325
## Pclass3      -2.49207834 -3.004469e+00   -1.97968793
## Sexmale      -2.71227213 -3.098502e+00   -2.32604182
## Age          -0.04544877 -6.098018e-02   -0.02991736
## Sib_sp1       0.07954020 -3.357014e-01    0.49478180
## Sib_sp2      -0.20419218 -1.223474e+00    0.81508932
## Sib_sp3      -2.35135691 -3.688787e+00   -1.01392685
## Sib_sp4      -1.71491328 -3.171898e+00   -0.25792859
## Sib_sp5     -16.02813247 -1.894766e+03 1862.70992273
## Sib_sp8     -16.50489364 -1.488127e+03 1455.11746887
```

2. Which variables are significant at the significance level ??=0.05? Give the p-value for any variable that is not significant.Please interpret.

Answer: (1). These variables are significant (a = 0.05) including: Pclass2 (5.66e-07), Pclass3 (2e-16), Sexmale (2e-16), Age (9.73e-09), Sib_sp3 (0.000569), Sib_sp4 (0.021058).These p-value < 0.05, so we reject the null hypothesis that the regression coefficients are zero (independent), therefore these predicting variables which will explain the variablity in survived.

(2). The p-values are not significant including the variables: Sib_sp1 (p-value = 0.707338), Sib_sp2 (p-value = 0.694586), Sib_sp5 (p-value = 0.986659), Sib_sp8 (p-value = 0.982462). These p-value > 0.05, so we can not reject the null hypothesis that the regression coefficients are zero, therefore these predicting variables which can not explain the variablity in survived.

Question 4: Goodness of fit

1. Aggregate the column “Survived” w.r.t the categorical predictors Pclass, Sex and Sib_sp. Fit a different Logistic Regression model with the number of successes as count of survived passengers as the new response vs Pclass, Sex and Sib_sp as predictors (follow the Obesity data example in the lecture). Perform a goodness of fit test for this new model? Does this model fit the data well?

```
model2 = glm(Survived ~ Pclass + Sex + Sib_sp, data = data, family = binomial)
#summary(model2)

sur.agg.n = aggregate(Survived~Pclass+Sex+Sib_sp,FUN=length, data = data)
sur.agg.y = aggregate(Survived~Pclass+Sex+Sib_sp,FUN=sum, data = data)
Pclass.agg = factor(sur.agg.n$Pclass, labels = c('1', '2', '3'))
Sex.agg = factor(sur.agg.n$Sex, labels = c('female', 'male'))
Sib.agg = factor(sur.agg.n$Sib_sp, labels = c('0','1','2', '3', '4', '5', '8'))

sur.agg = data.frame(Survived = sur.agg.y$Survived,
                     Total = sur.agg.n$Survived,
                     Pclass = Pclass.agg,
                     Sex = Sex.agg,
                     Sib_sp = Sib.agg)
model.agg = glm(cbind(Survived,Total-Survived)~Pclass+Sex+Sib_sp,
                data = sur.agg,family=binomial)
summary(model.agg)
```

```
##
## Call:
## glm(formula = cbind(Survived, Total - Survived) ~ Pclass + Sex +
##     Sib_sp, family = binomial, data = sur.agg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8067  -0.4196   0.0466   0.8426   2.3131
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.3128     0.2392   9.669  < 2e-16 ***
## Pclass2       -0.8638     0.2473  -3.493 0.000477 ***
## Pclass3       -1.7953     0.2177  -8.247  < 2e-16 ***
## Sexmale       -2.7073     0.1912 -14.158  < 2e-16 ***
## Sib_sp1        0.1880     0.2075   0.906 0.364923
## Sib_sp2        0.1665     0.4872   0.342 0.732622
## Sib_sp3       -1.6154     0.6658  -2.426 0.015254 *
## Sib_sp4       -0.8773     0.7295  -1.203 0.229137
## Sib_sp5      -18.0939  3532.9461  -0.005 0.995914
## Sib_sp8      -18.5927  2823.2957  -0.007 0.994746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 414.964  on 28  degrees of freedom
## Residual deviance:  39.565  on 19  degrees of freedom
## AIC: 117.63
##
## Number of Fisher Scoring iterations: 17
```

This is for overall regression test, this part just for reference, not required for assignment.

```
gstat = model.agg$null.deviance - deviance(model.agg)
cbind(gstat, 1-pchisq(gstat,length(coef(model.agg))-1))
```

```
##         gstat
## [1,] 375.3996 0
```

Answer: Because the p-value is approximately zero, thus we reject the null hypothesis that all regression coefficients are zero, the overall regression is statistically significant with probably explanatory power.

Test for GOF: Using deviance residuals.

```
deviances2 = residuals(model.agg,type="deviance")
dev.tvalue = sum(deviances2^2)
c(dev.tvalue, 1-pchisq(dev.tvalue,19))
```

```
## [1] 39.564798056  0.003730766
```

```
#OR
c(deviance(model.agg), 1-pchisq(deviance(model.agg),19))
```

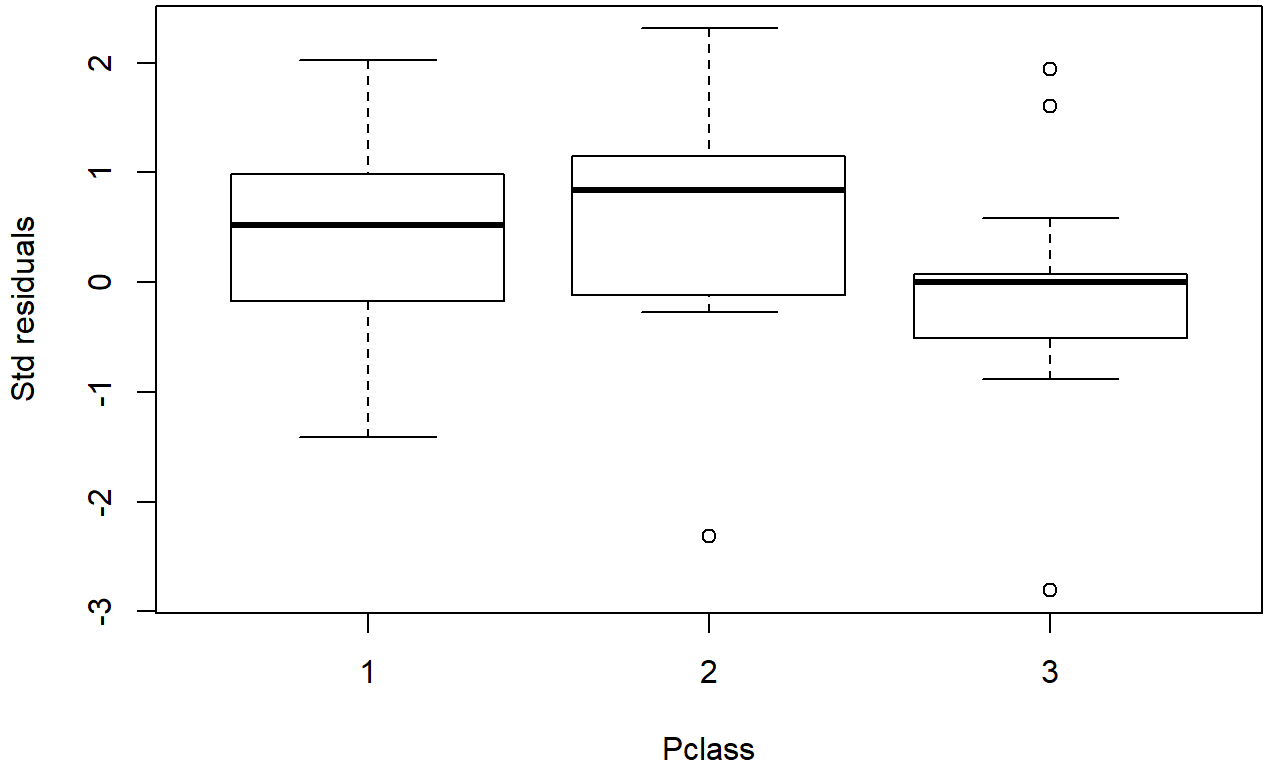
```
## [1] 39.564798056  0.003730766
```

Answer: The aggregate method allows to compute the number of observed successes for a different Logistic Regression model, we perform a goodness of fit test for this new model and find the p-value of the test is 0.003730766, which is very small, so it indicates possibly **not** a good model fit.

2. Residual Analysis: Produce the following deviance residual plots:

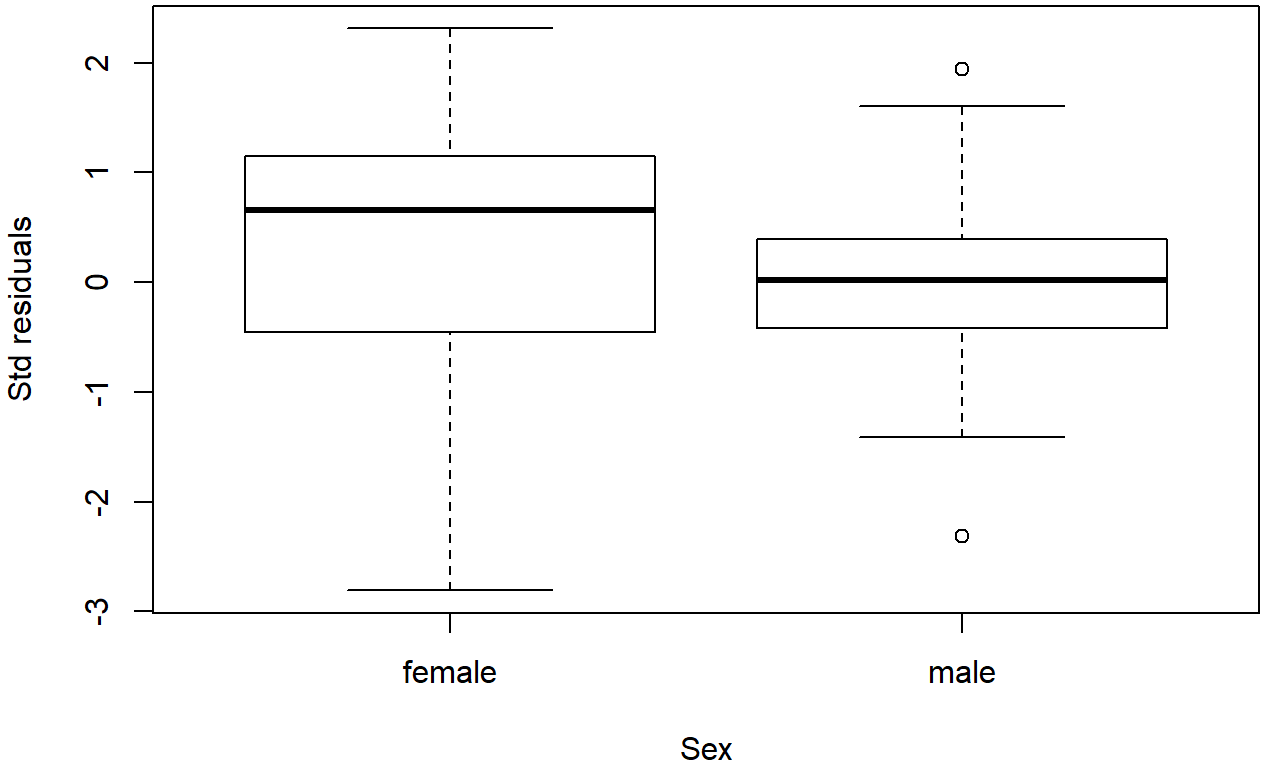
1). Boxplot of the residuals by Pclass.

```
res = resid(model.agg, type = 'deviance')
boxplot(res ~ Pclass.agg, xlab = 'Pclass', ylab = 'Std residuals', data = sur.agg)
```



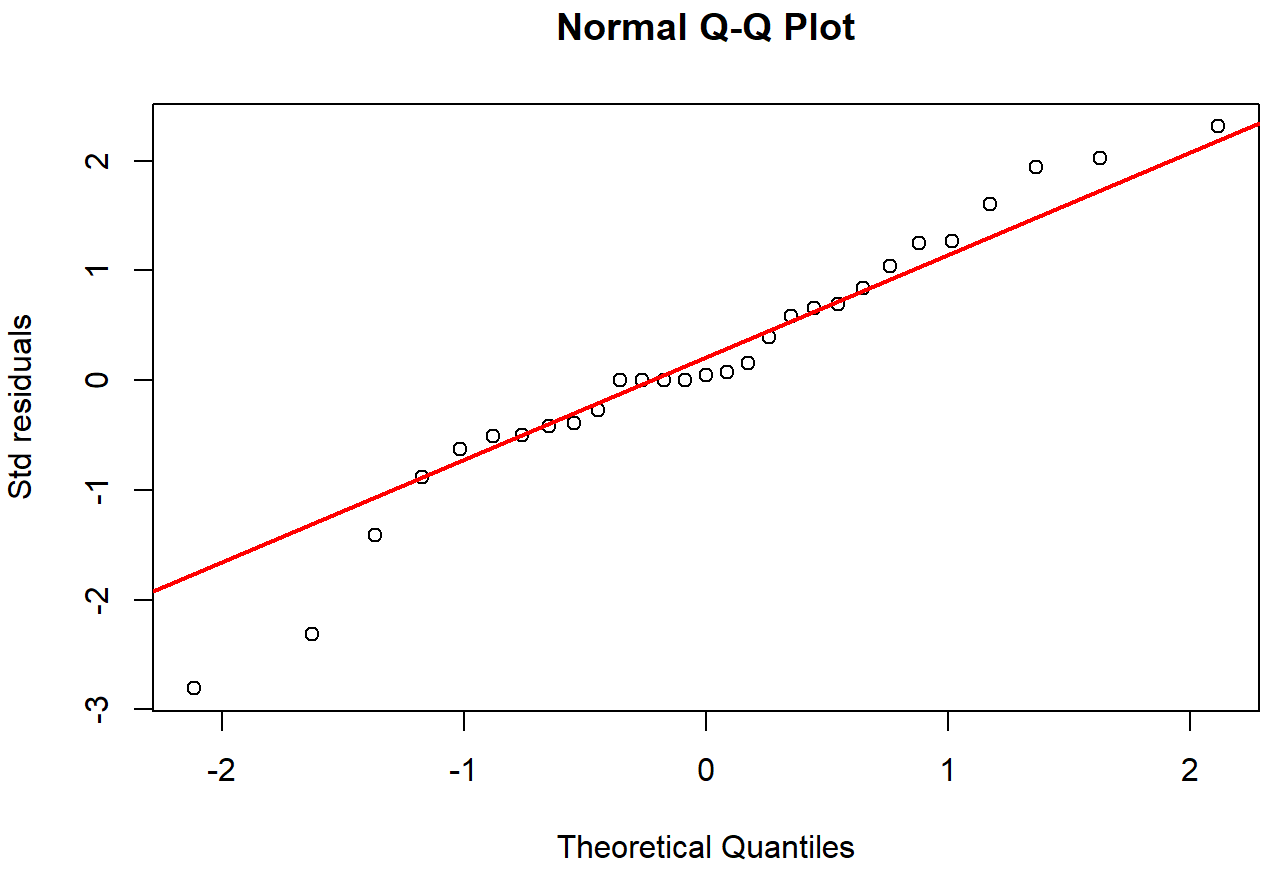
2). Boxplot of the residuals by Sex

```
boxplot(res ~ Sex.agg, xlab = 'Sex', ylab = 'Std residuals', data = sur.agg)
```



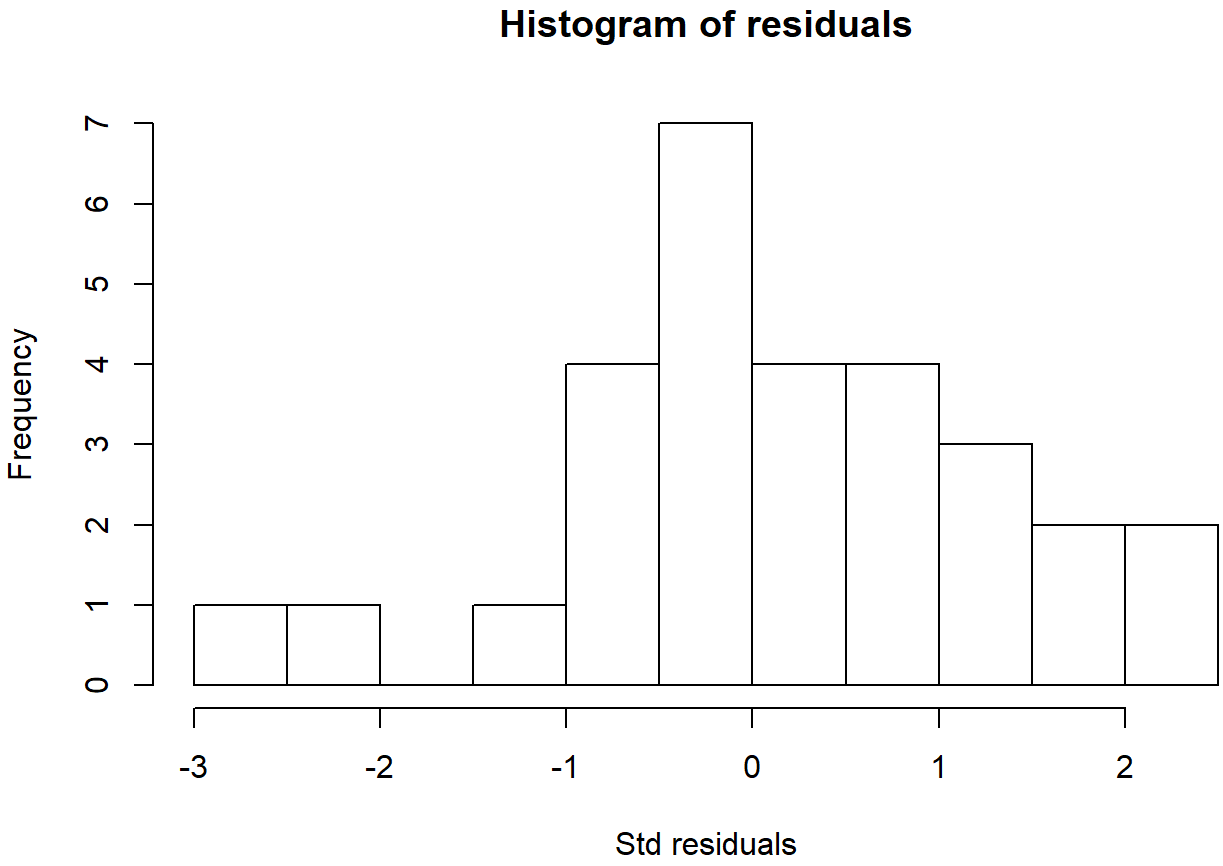
3). QQPlot of the residuals

```
{qqnorm(res, ylab = 'Std residuals')
qqline(res, col='red', lwd=2)}
```



4). Histogram of the residuals Comment on the plots.

```
hist(res, 10, xlab = 'Std residuals', main = 'Histogram of residuals' )
```



Answer: (1). For boxplot of the residuals by Pclass, we found that there is not a significant variability between the Pclass 1 and 2, but there have some different between Pclass 3 and the other two. (2). For boxplot of the residuals by Sex, we found that there is not a significant variability between female and male. (3). For QQPlot of the residuals, mojor points seems to have normality except left tail. (4). For histogram of the residuals, the distrubution appears to be right-skewed and not symmetric.

Question 5: Prediction

1. Now consider the original model in Question 2. Predict the probability of survival of a Class 1 female passenger of age 20 with 1 sibling/spouse.

```
rm(list=ls())
data = read.csv('titanic.csv', header = T, sep = ',')
data$Pclass = factor(data$Pclass, labels = c('1', '2', '3'))
data$Sib_sp = factor(data$Sib_sp, labels = c('0','1','2', '3', '4', '5', '8'))

model1 = glm(Survived ~ Pclass + Sex + Age + Sib_sp, data = data, family = binomial)
#summary(model1)
pred.data1 = data.frame(Pclass = '1', Sex = 'female', Age = 20, Sib_sp = '1')
predict.glm(model1, pred.data1, type = 'response')
```

```
##         1
## 0.9666272
```

2. Predict the probability of survival of a Class 3 male passenger of age 21 with "above_4" siblings/spouses.

```
rm(list=ls())
data = read.csv('titanic.csv', header = T, sep = ',')
data$Pclass = factor(data$Pclass, labels = c('1', '2', '3'))
data$Sib_sp <- as.factor(ifelse(data$Sib_sp >= 4, 'above_4', 'below_4'))
new_model1 = glm(Survived ~ Pclass + Sex + Age + Sib_sp, data = data, family = binomi
al)
#summary(new_model1)
pred.data2 = data.frame(Pclass = '3', Sex = 'male', Age = 21, Sib_sp = 'above_4')
predict.glm(new_model1, pred.data2, type = 'response')
```

```
##          1
## 0.01556882
```

3. Can you now infer which groups of people survived and which groups were left behind?

Answer: For categorial vairable of ticket class, we found the passengers with 1st class ticket, the probability of survival is large than 3st class ticket. For categorial vairable of sex, we found that the female passengers have more survivals than male passengers. For categorial variable of Sib_sp, we found that the passengers with few numbers of siblings/spouses have more survivals than passengers with large numbers of siblings/spouses. For numeric variable of Age, due to the predictor coefficient is negative, thus the passengers with old age will have low possibility of survivial.