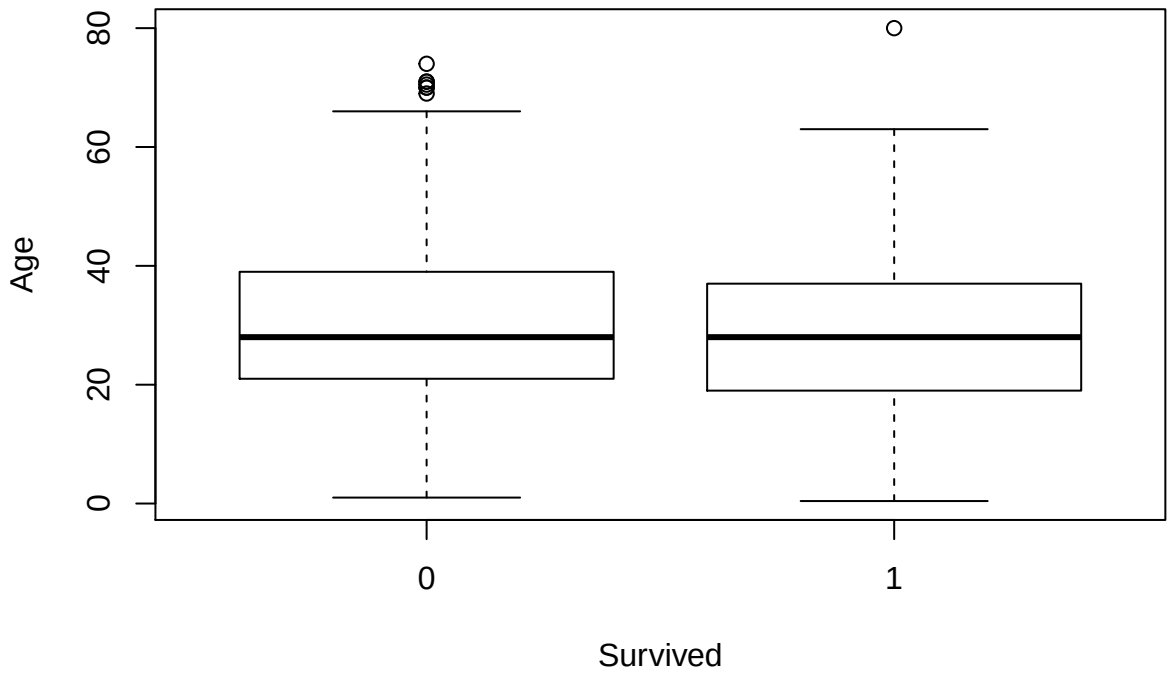
Homework 4

Q1

Q 1.1: 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"])

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

data = read.csv('titanic.csv', header=TRUE, sep = ',')
attach(data)
survived = as.factor(Survived)
# Age
boxplot(Age ~ Survived, xlab = "Survived", ylab = "Age")
```




```
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
```

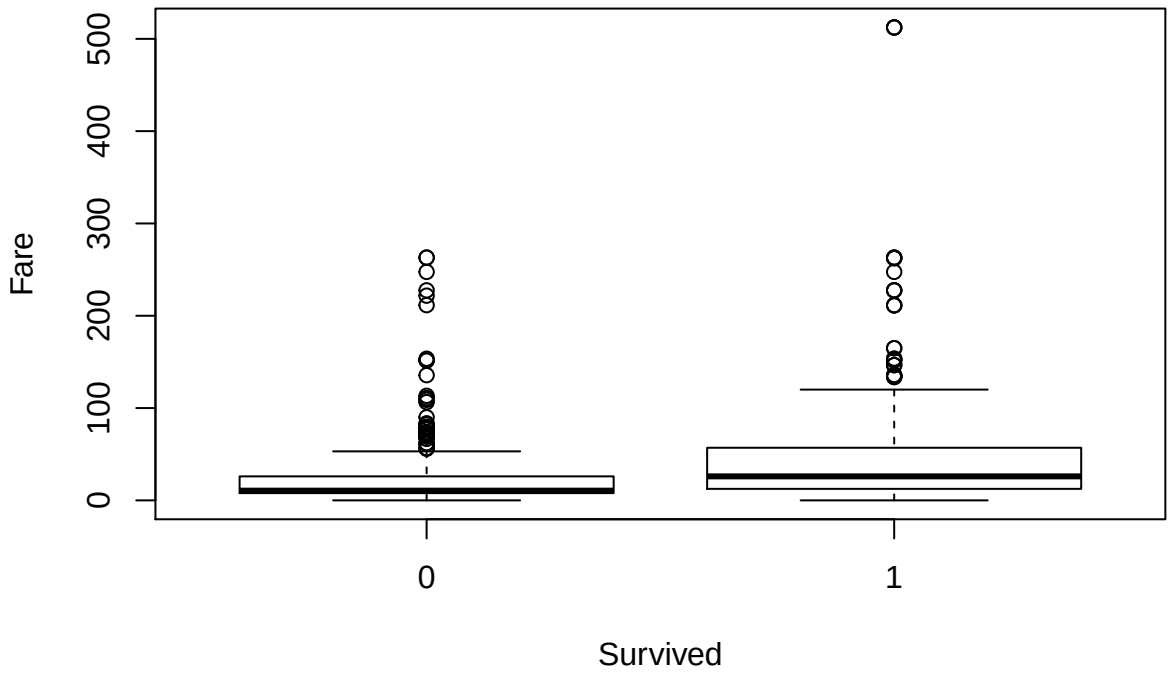
```
summary(data[data$Survived==0,"Age"])
```

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

```
#use Anova to determine differences in age survival
aovAge = aov(Age ~ survived)
TukeyHSD(aovAge)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Age ~ survived)
##
## $survived
##         diff       lwr       upr     p adj
## 1-0 -1.73014 -3.639805 0.1795246 0.0757237
```

```
# Fare
boxplot(Fare ~ Survived, xlab = "Survived", ylab = "Fare")
```




```
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
```

```r
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
```

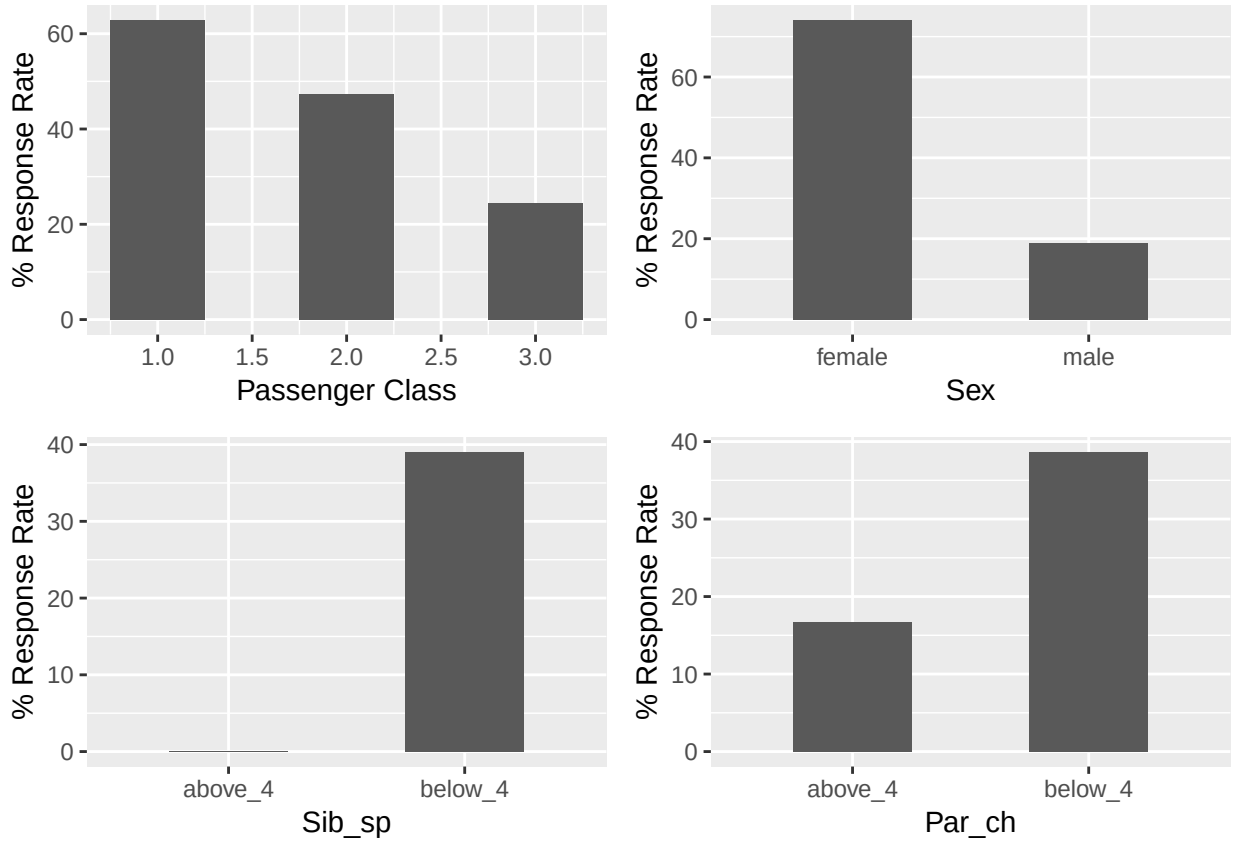
```r
#use Anova to determine differences in fare survival
aovFare = aov(Fare ~ survived)
TukeyHSD(aovFare)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Fare ~ survived)
##
## $survived
##         diff      lwr      upr p adj
## 1-0 26.18682 19.66797 32.70567     0
```

I did not see a clear difference between age and survuval rate or fare and survival rate with boxplots, or summarizing the data. Although an increase in fare did have a higher survival rate it was not clear cut. I performed an ANOVA to compare the means between survival and non-survival. THe results showed that there was not a significant difference in the mean age (p>.05), but there was a difference in fare (p=0).

Q 1.2: 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.

```r
require(gridExtra)
data$Par_ch = ifelse(data$Par_ch >4, "above_4", "below_4")
data$Sib_sp = ifelse(data$Sib_sp  >4, "above_4", "below_4")



#Pclass
p1 = ggplot(ddply(data,.(Pclass),summarise, rr=100*sum(Survived)/length(Survived)), aes(x=Pclass,y=rr))


#Sex
p2 = ggplot(ddply(data,.(Sex),summarise, rr=100*sum(Survived)/length(Survived)), aes(x=Sex,y=rr))+geom_


#Sib_sp
p3 = ggplot(ddply(data,.(Sib_sp),summarise, rr=100*sum(Survived)/length(Survived)), aes(x=Sib_sp,y=rr))


#Par_ch
p4 = ggplot(ddply(data,.(Par_ch),summarise, rr=100*sum(Survived)/length(Survived)), aes(x=Par_ch,y=rr))

grid.arrange(p1, p2, p3,p4, ncol=2)
```



Visual inspection of the categorical independent variables Pclass, Sex, Sib_sp and Par_ch indicates that survival rate does vary with the categorical variables. Lower classes, males, higher sibling/spouse ratio, and higher # of parent/ children ratio all had lower survival rates compared to higher classes, females, low sibling/spouse ratios and low #of parent / children ratios.

Q 1.3: 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

The model is the logistic regression model of the form:

$$g(p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

where $g(p)$ is a link function of the form:

$$g(p) = log(\frac{p}{1-p})$$

and p is the probability of success.

Model assumptions are:

1. Linearity, the relationshp between the link function, g, and the predicted variable is a linear function.
2. Independence in the response data.
3. The link function is a logit function.

Q2

Q 2.1: 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?

```
Pclass = as.factor(Pclass)
Sib_sp = as.factor(Sib_sp)

model_q2 = glm(Survived ~ Pclass + Sex + Age + Sib_sp, family=binomial)


summary(model_q2)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Sib_sp, family = binomial)
##
## 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
```

The above summary shows the model parameters (Pclass2, Pclass3, Sexmale, Age, Sib_sp1, Sib_sp2, Sib_sp3, Sib_sp4, Sib_sp5, Sib_sp8), with the estimates located in the first column "Estimate".

Q 2.2: Write down the equation for the logarithm of odds of survival given the predicting variables.

$$g(p) = 4.195506 - 1.356275Pclass2 - 2.492078Pclass3 - 2.712272Sexmale - 0.045449Age + 0.079540Sibsp1 - 0.204192Sibsp2 - 2.351357Sibsp3 - 1.714913Sibsp4Sibsp4 - 16.028132Sibsp5 - 16.504894Sibsp8$$

Q 2.3: Interpret the coefficients of Pclass, Sex and Age

In general a positive coefficient is associated with a higher survival rate, where a negative coefficient is associated with a lower survival rate.
More specifically, considering Pclass, Sex and Age, defining 1 as the highest class and 3 as the lowest class, survival rates decrease with lower classes because Pclass2 > Pclass3. For males, the log odds of survival decrease by 2.7 or the odds of survival decrease by .067 versus females given that all other predicting variables are fixed. The log odds of survival devrease with age by -0.045449, or the odds of survival decrease by 0.96 for every one year increase in age given that all other predicting variables are fixed.

Q 3

Q 3.1: Find a 95% confidence interval for the parameters corresponding to all predictors plus the intercept.

```
exp(confint.default(model_q2))
```

```
##                    2.5 %       97.5 %
## (Intercept) 28.80832448 152.98617524
## Pclass2      0.15142621   0.43828177
## Pclass3      0.04956508   0.13811233
## Sexmale      0.04511672   0.09768162
## Age          0.94084189   0.97052573
## Sib_sp1      0.71483652   1.64014033
## Sib_sp2      0.29420641   2.25937747
## Sib_sp3      0.02500231   0.36279155
## Sib_sp4      0.04192395   0.77265040
## Sib_sp5      0.00000000          Inf
## Sib_sp8      0.00000000          Inf
```

Q3.2 Which variables are significant at the significance level alpha=0.05? Give the p-value for any variable that is not significant.Please interpret.

Significant variables at alpha =0.05

- (Intercept)
- Pclass2
- Pclass3
- Sexmale
- Age

- Sib_sp3
- Sib_sp4

variables not significant at alpha = 0.05

variable	Pr(>abs(z))
Sib_sp1	0.707338
Sib_sp2	0.694586
Sib_sp5	0.986659
Sib_sp8	0.982462

The above table shows values that are not significant at the alpha = 0.05 level, as shown in the table all p-values are greater than 0.05 indicating that you cannot reject the null hypothesis that the coefficient is equal to zero.

Q 4

Q 4.1: 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?

```
newData = read.csv('titanic.csv', header=TRUE, sep = ',')
newData.agg.n = aggregate(Survived~Pclass+Sex+Sib_sp,FUN=length, data=newData)
newData.agg.y = aggregate(Survived~Pclass+Sex+Sib_sp,FUN=sum, data =newData)


Pclass.agg = factor(newData.agg.n$Pclass)
Sex.agg = factor(newData.agg.n$Sex)
Sib_sp.agg = factor(newData.agg.n$Sib_sp)


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

summary(model.agg)
```

```
##
## Call:
## glm(formula = cbind(Survived, Total - Survived) ~ Sex + Pclass +
##     Sib_sp, family = binomial, data = newData.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 ***
## Sexmale       -2.7073     0.1912 -14.158  < 2e-16 ***
## Pclass2       -0.8638     0.2473  -3.493 0.000477 ***
## Pclass3       -1.7953     0.2177  -8.247  < 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
```

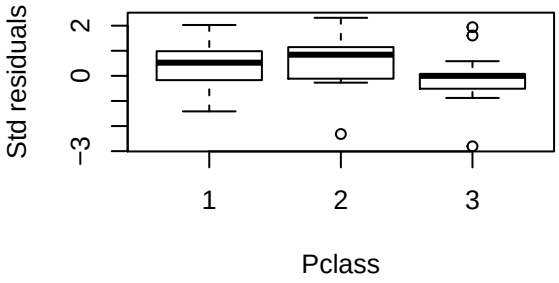
```
## Test for GOF: Using deviance residuals
deviances2 = residuals(model.agg,type="deviance")
dev.tvalue = sum(deviances2^2)
c(dev.tvalue, 1-pchisq(dev.tvalue,19))
```

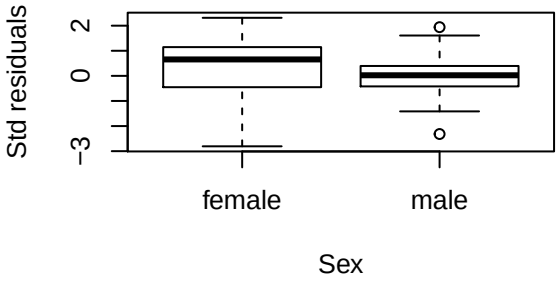
```
## [1] 39.564798056  0.003730766
```

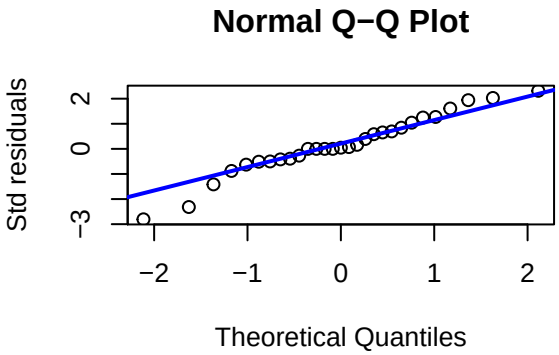
Based on th low p-value above ($p<0.05$), we reject the null hypothesis of good fit and conclude that the model is not a good fit.

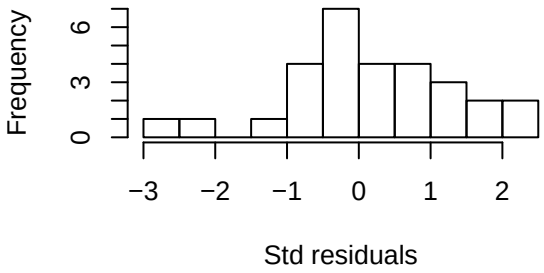
Q 4.2:Residual Analysis

```
## Residual Analysis
res = resid(model.agg,type="deviance")
par(mfrow=c(2,2))
boxplot(res~Pclass,xlab="Pclass",ylab = "Std residuals",data = newData.agg)
boxplot(res~Sex,xlab="Sex",ylab = "Std residuals",data = newData.agg)
qqnorm(res, ylab="Std residuals")
qqline(res,col="blue",lwd=2)
hist(res,10,xlab="Std residuals", main="")
```









There is somewhat of a mixed result between the goodness of fit test using the p-value and the graphical residual analysis. The above plots show that the model fits reaonably well. The boxplots show some skew between the sexes with a larger spread of deviances among the females than maes, the classes seem to fit nicely. The qq plot deviates on the left tail, but both the qq and histogram are reasonable.

Q5

Q5.1: 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

```
new = data.frame(Pclass=1, Sex = "female", Age = 20.0, Sib_sp = 1)
new$Pclass = factor(new$Pclass)
new$Sib_sp = factor(new$Sib_sp)
pred1 = predict(model_q2, new, type="response")
pred1
```

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

Q5.2: Predict the probability of survival of a Class 3 male passenger of age 21 with "above_4" siblings/spouses

```
new = data.frame(Pclass=3, Sex = "male", Age = 21.0, Sib_sp = 5)
new$Pclass = factor(new$Pclass)
new$Sib_sp = factor(new$Sib_sp)
pred2 = predict(model_q2, new, type="response")
pred2
```

```
##           1
## 1.53615e-08
```

Q 5.3: Can you now infer which groups of people survived and which groups were left behind?

The difference between the above predictions demonstrates that young, upperclass females with a low siblings/spouse ratio had a very good chance at survival, and young lower classed men with a high siblings/spouse ratio had a very poor chance at survival.