Student Expectations - peer assessment grading and peer assessment comments

Adapted from Science Education Resource Center (2016): Guidelines for Students, Peer Review
https://serc.carleton.edu/introgeo/peerreview/tips.html

- Await the solutions to be published before you start your peer review activities
- Before you make your first score, read through the solutions document
- Make sure you allow enough time for you to read the assessment thoroughly and respond thoughtfully
- Be rigorous, point out the strengths as well as the weaknesses of the assessment
- When assigning scores, comment on the reasons why points were reduced. Be considerate and offer suggestions, not directives.
- Comments should be appropriate and constructive. There is no need to be rude. Be respectful and considerate of the writer's feelings (for example, terms such as “lack of effort” or “you don’t have much work to show for all the time you spent” are not constructive, and are oftentimes hurtful and offensive.)
- Be sure that your comments are clear and text-specific so that your peer will know what you are referring to (for example, terms such as "unclear" or "vague" are too general to be helpful).
- As a reader, raise questions that cross your mind, points that may have not occurred to your peer
- Be careful not to let your own opinions bias your review (for example, don't suggest that your peer resubmit the assignment just because you don't agree with his/her approach).
- Reread your comments before submitting your review. Make sure your comments make sense and are easy to follow.

ISYE6414 OAN HW4 peer review solutions

Mar 9, 2019

On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

The titanic.csv file contains data for 887 of the real Titanic passengers. Each row represents one person. The columns describe different attributes about the person including whether they survived, their age, their passenger-class, their sex and the fare they paid.

The response variable is defined by the following data input:

Survived: Whether the passenger survived (0=No, 1=Yes)

Explanatory variables are the following in the data file:

Pclass: ticket class of the passenger (1 = 1st, 2 = 2nd, 3 = 3rd)

Name: name of the passenger

Sex: sex of the passenger

age: age of the passenger

Sib_sp: # of siblings / spouses of the passenger aboard the Titanic

Par_ch: # of parents / children of the passenger aboard the Titanic

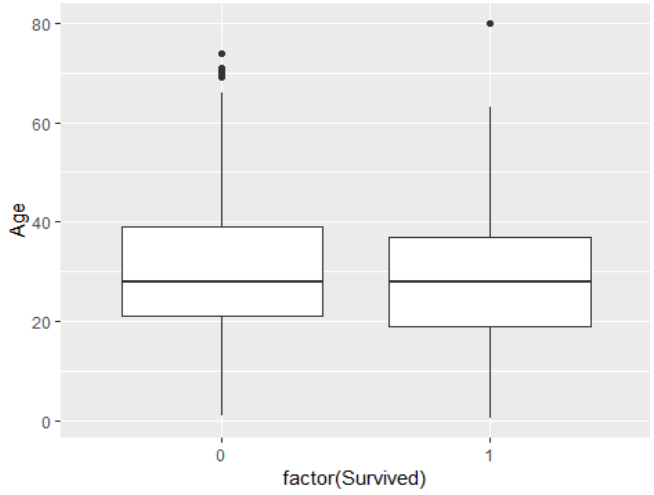
Fare: passenger fare

Question 1: Exploratory Data Analysis

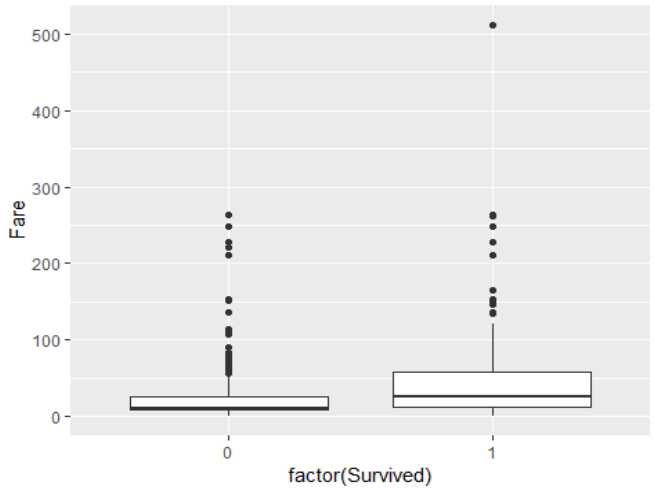
a) Using boxplots explore the relationship between survived and the numerical independent variables: Age and Fare . Can you observe differences in distribution of the predictors between the 2 classes? Please explain and interpret.If you cannot determine visually please observe the mean/median of the predictors by the 2 classes: for example:

summary(data[data$Survived==1,“Age”]).

```
data=read.csv("titanic.csv")

#--Boxplots----#
ggplot(data,aes(x=factor(Survived)))+geom_boxplot(aes(y=Age))
```




```
ggplot(data,aes(x=factor(Survived)))+geom_boxplot(aes(y=Fare))
```




```
#--Summary of fare--#
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

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
```

Obervations

Between Age and Fare, only Fare differs in mean and median by the 2 classes.Passengers who survived paid 26$ more on average than passengers who could not survive

b) 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.

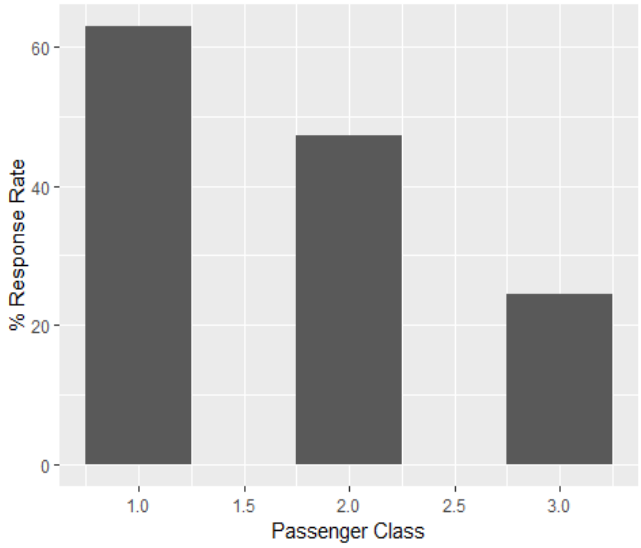
One way of doing this is a contingency table followed by a Chi-squared test. A more visual way would be to observe the % response rates w.r.t levels of the predictor. You can use the following code to plot a barchart of response rates vs a predictor:

P-Class

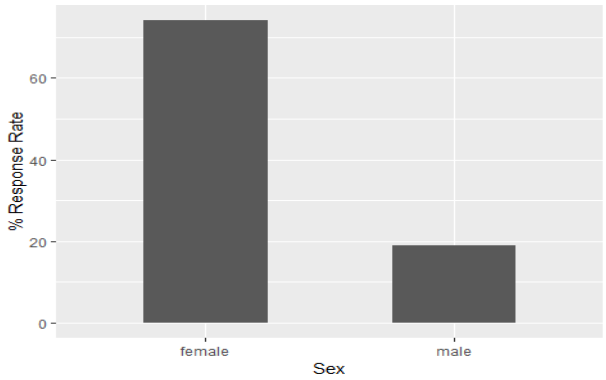
library(plyr)

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")

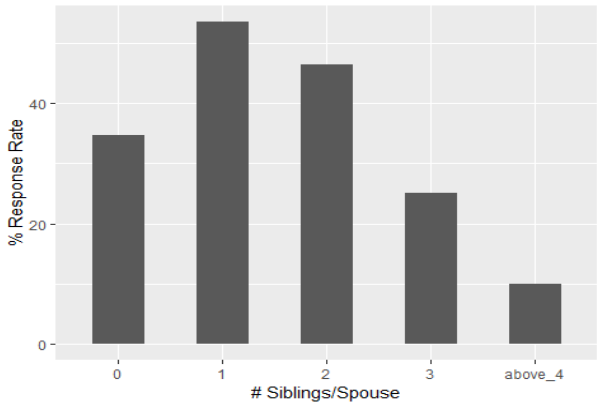
```
data$Sib_sp=ifelse(data$Sib_sp>=4,"above_4",data$Sib_sp)
data$Par_ch=ifelse(data$Par_ch>=4,"above_4",data$Par_ch)


#P-Class
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")
```



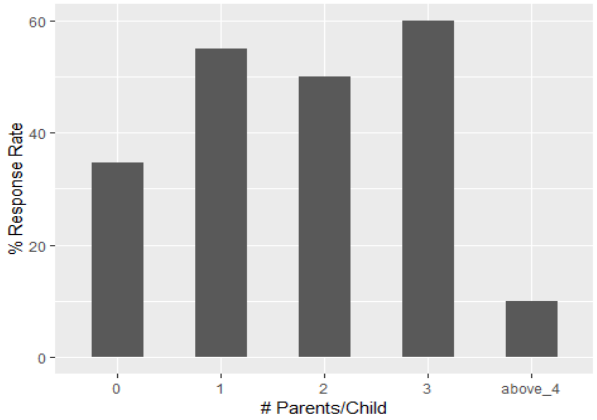

```
#Sex
ggplot(ddply(data,.(Sex),summarise,
             rr=100*sum(Survived)/length(Survived)),
       aes(x=Sex,y=rr))+geom_bar(stat = "identity",width=0.5)+
  labs(x="Sex", y="% Response Rate")
```




```
#Sib_sp
ggplot(ddply(data,.(Sib_sp),summarise,
             rr=100*sum(Survived)/length(Survived)),
       aes(x=factor(Sib_sp),y=rr))+geom_bar(stat = "identity",width=0.5)+
  labs(x="# Siblings/Spouse", y="% Response Rate")
```




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



Obervations

i) Survival rate differs sharply by passenger class and gender. Women and Class 1 passengers were the first to be evacuated

ii) Survival rate is slightly higher for passengers with 1 or 2 Siblings/Spouse

c) 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

Model:

$y_i \in \{0,1\}, y_i \; distributed \; i.i.d \; Bernoulli(1/(1+\exp(-\theta^T X_i))$

Or

$y_i \in \{0,1\}, y_i \; distributed \; i.i.d \; Bernoulli(pi) \; and \; log(odds(pi)) = \theta^T X_i$

Or

$y_i \in \{0,1\}, y_i \; distributed \; i.i.d \; Bernoulli(pi) \; and \; logit(pi) = \theta^T X_i$

X without multicollinearity, no outliers in X or y

Response Quality	Description	Points (out of 12)
Poor	Student did not answer question correctly	0
OK	Student partially analyses plots, does not correctly interpret/answer questions	4
Good	Student properly analyses plots, is partially correct interpreting/answering questions	8
Perfect	Student interprets and answers all questions correctly	12

Question 2: Fitting Regression Model

a) 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?

```
data$Pclass=as.factor(data$Pclass)
glmod=glm(Survived~Pclass+Sex+Age+Sib_sp,data,family=binomial)
summary(glmod)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Sex + Age + Sib_sp, family = binomial,
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9023  -0.6092  -0.3952   0.6148   2.6420
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.209261   0.426269   9.875  < 2e-16 ***
## Pclass2       -1.360179   0.271194  -5.016 5.29e-07 ***
## Pclass3       -2.497465   0.261558  -9.548  < 2e-16 ***
## Sexmale       -2.710295   0.197123 -13.749  < 2e-16 ***
## Age           -0.045823   0.007927  -5.781 7.44e-09 ***
## Sib_sp1        0.079219   0.211868   0.374 0.708472
## Sib_sp2       -0.206665   0.520185  -0.397 0.691153
## Sib_sp3       -2.356410   0.682434  -3.453 0.000554 ***
## Sib_spabove_4 -2.323146   0.686581  -3.384 0.000715 ***
## ---
## 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:  773.71  on 878  degrees of freedom
## AIC: 791.71
##
## Number of Fisher Scoring iterations: 5
```

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

log(odds of survival)=4.21-1.36*Pclass2-2.50*Pclass3-2.71*Sexmale-0.05*age+0.08*Sibsp1-0.21*Sibsp2-2.36*Sibsp3-2.32*Sibspabove_4

c) Interpret the coefficients of Pclass, Sex and Age

With other predictors being constant, the estimated odds of survival for passengers in Class 2 is 75% (1-exp(-1.36)) lower than that in Class1

With other predictors being constant, the estimated odds of survival for passengers in Class 3 is 92% (1-exp(-2.50)) lower than that in Class1

With other predictors being constant, the estimated odds of survival for males is 93.3% (1-exp(-2.71)) lower than that for females

With other predictors being constant, the estimated odds of survival decreases by 4.5% (1-exp(-0.046)) for every one year increase in passenger age

Response Quality	Description	Points (out of 12)
Poor	Student did not answer the question correctly	0
OK	Student answers 1 question correctly	4
Good	Student answers 2 questions correctly	8
Perfect	Student answers 3 questions correctly	12

Question 3: Inference

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

```
confint(glmod,level = 0.95)
```

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

```
##                     2.5 %      97.5 %
## (Intercept)    3.39725885  5.06994752
## Pclass2       -1.89950769 -0.83517436
## Pclass3       -3.02186763 -1.99524718
## Sexmale       -3.10560467 -2.33196595
## Age           -0.06171406 -0.03059983
## Sib_sp1       -0.33881003  0.49275082
## Sib_sp2       -1.23828310  0.80592962
## Sib_sp3       -3.80781824 -1.08284917
## Sib_spabove_4 -3.86478271 -1.09329565
```

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

At the 0.05 significance level, Pclass, Sex, Age and Sib_sp3 and Sib_spabove_4 are significant. Sib_sp1 and Sib_sp2 are not significant and have p-values of 0.71 and 0.69 respectively. This means the there in no significant difference in survival rates between passengers having 1 or 2 siblings/spouses than passengers having none

Response Quality	Description	Points (out of 8)
Poor	Student did not answer the question correctly	0
Good	Student answers 1 question correctly	4
Perfect	Student answers 2 questions correctly	8

Question 4: Goodness of fit

a) 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?

```
agg_dat=ddply(data,.(Pclass,Sex,Sib_sp),summarise,
              total=length(Survived),
              Num_survived=sum(Survived))

glmod_rep=glm(cbind(Num_survived,total-Num_survived)~Pclass+Sex+Sib_sp,
              agg_dat,family=binomial)
summary(glmod_rep)
```

```
##
## Call:
## glm(formula = cbind(Num_survived, total - Num_survived) ~ Pclass +
##     Sex + Sib_sp, family = binomial, data = agg_dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8028  -0.5020   0.1604   1.0439   2.3182
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.3101     0.2391   9.662  < 2e-16 ***
## Pclass2        -0.8633     0.2472  -3.492 0.000479 ***
## Pclass3        -1.7943     0.2176  -8.246  < 2e-16 ***
## Sexmale        -2.7041     0.1912 -14.143  < 2e-16 ***
## Sib_sp1         0.1884     0.2074   0.909 0.363600
## Sib_sp2         0.1669     0.4871   0.343 0.731900
## Sib_sp3        -1.6137     0.6657  -2.424 0.015350 *
## Sib_spabove_4  -1.5870     0.6722  -2.361 0.018229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.538  on 24  degrees of freedom
## Residual deviance:  39.758  on 17  degrees of freedom
## AIC: 114.03
##
## Number of Fisher Scoring iterations: 4
```

```
1-pchisq(deviance(glmod_rep),glmod_rep$df.residual)
```

```
## [1] 0.001399696
```

Due to the low p-value of 0.001, we reject the Null Hypothesis of a good fit. Thus the data does not fit this new model well

b) Residual Analysis

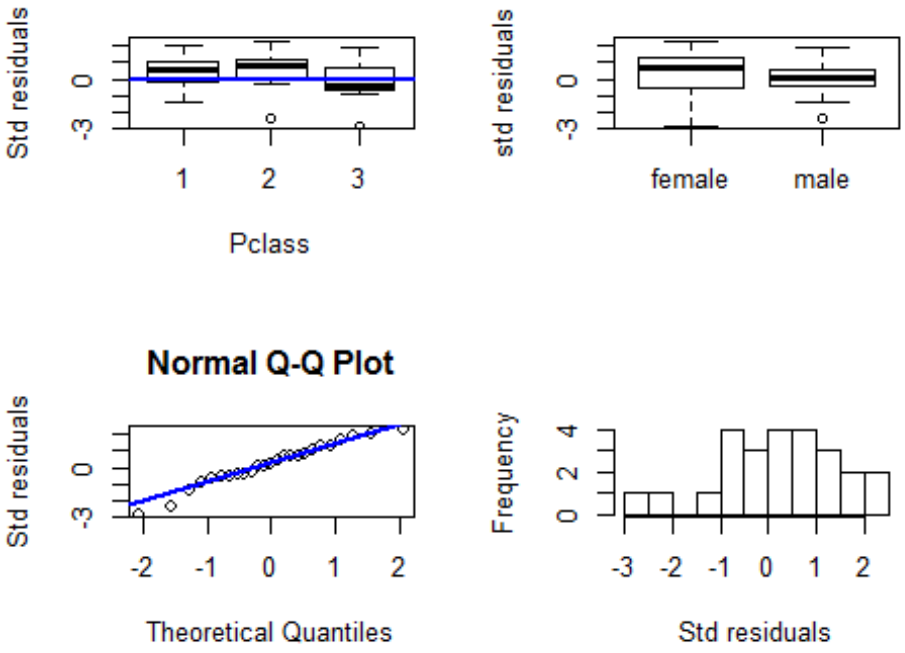
Produce the following deviance residual plots:
1) Boxplot of the residuals by Pclass

2) Boxplot of the residuals by Sex

3) QQPlot of the residuals

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

```
res=resid(glmod_rep,type="deviance")
par(mfrow=c(2,2))
plot(agg_dat$Pclass,res,ylab="Std residuals",xlab="Pclass")
    abline(0,0,col="blue",lwd=2)
    boxplot(res~agg_dat$Sex,ylab="std residuals")
    qqnorm(res,ylab="Std residuals")
    qqline(res,col="blue",lwd=2)
    hist(res,10,xlab="Std residuals",main="")
```



The plots on the first row show that the spread of the deviance residuals are fairly constant across 'Sex' and across 'Pclass' and there do not seem to be any patterns in the plots. The qq-plot shows an approximate normal distribution with skewness at the tails. This can be a possible reason for a lack of fit. Another reason for a lack of fit can be leaving out some important predictors (maybe Age)

Response Quality	Description	Points (out of 8)
Poor	Student did not answer the question correctly	0
Good	Student answers 1 question correctly	4
Perfect	Student answers 2 questions correctly	8

Question 5: Prediction

a) 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

```
x0=data.frame(Pclass='1',Sex="female",Age=20,Sib_sp='1')
predict(glmod,x0,type="response")

##         1
## 0.9668183
```

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

```
x0=data.frame(Pclass='3',Sex="male",Age=21,Sib_sp='above_4')
predict(glmod,x0,type="response")

##          1
## 0.01360073
```

c) Can you now infer which groups of people survived and which groups were left behind?

Higher class female passengers with 1-2 dependents were the first to survive. Class 3 male passengers with many dependents had the worst survival rates

Response Quality	Description	Points (out of 10)
Poor	Student did not answer the question correctly	0
OK	Student answers 1st question correctly	4
Good	Student answers 1st and 2nd questions correctly	8
Perfect	Student answers 1st and 2nd questions correctly and interprets 3rd question correctly	10