Analyzing The Titanic Ship Wreck with Logistic Regression

Cicely Karas

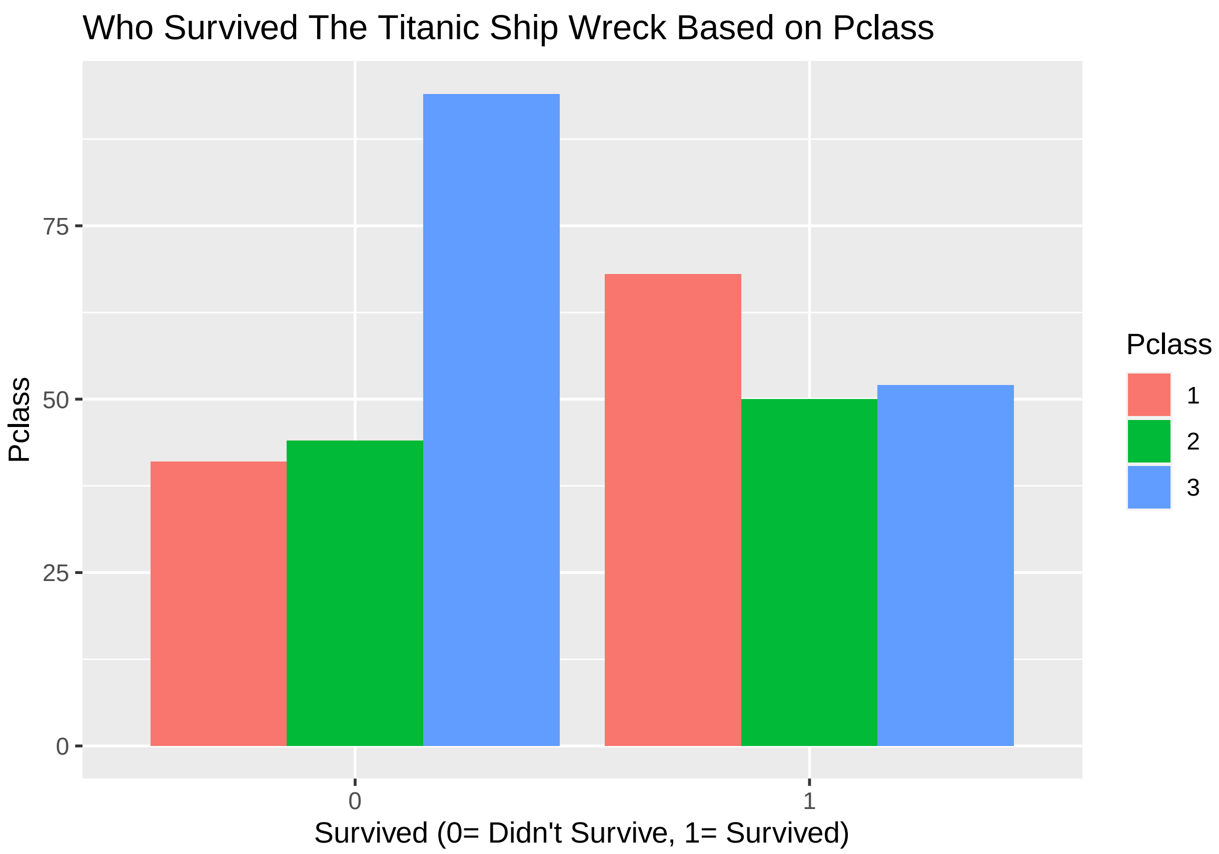
2022-11-27

My well defined problem: Can pclass, sex, and average age predict survival status the titanic ship wreck?

```
new_train = train%>%
  group_by(Survived,Pclass,Age,Sex)%>%
  summarize(Avg_age = mean(Age,na.rm=TRUE))%>%
  mutate(ifelse(Survived == 1, "Survivor","Non-Survivor"),
         Survived = as.factor(Survived))
```

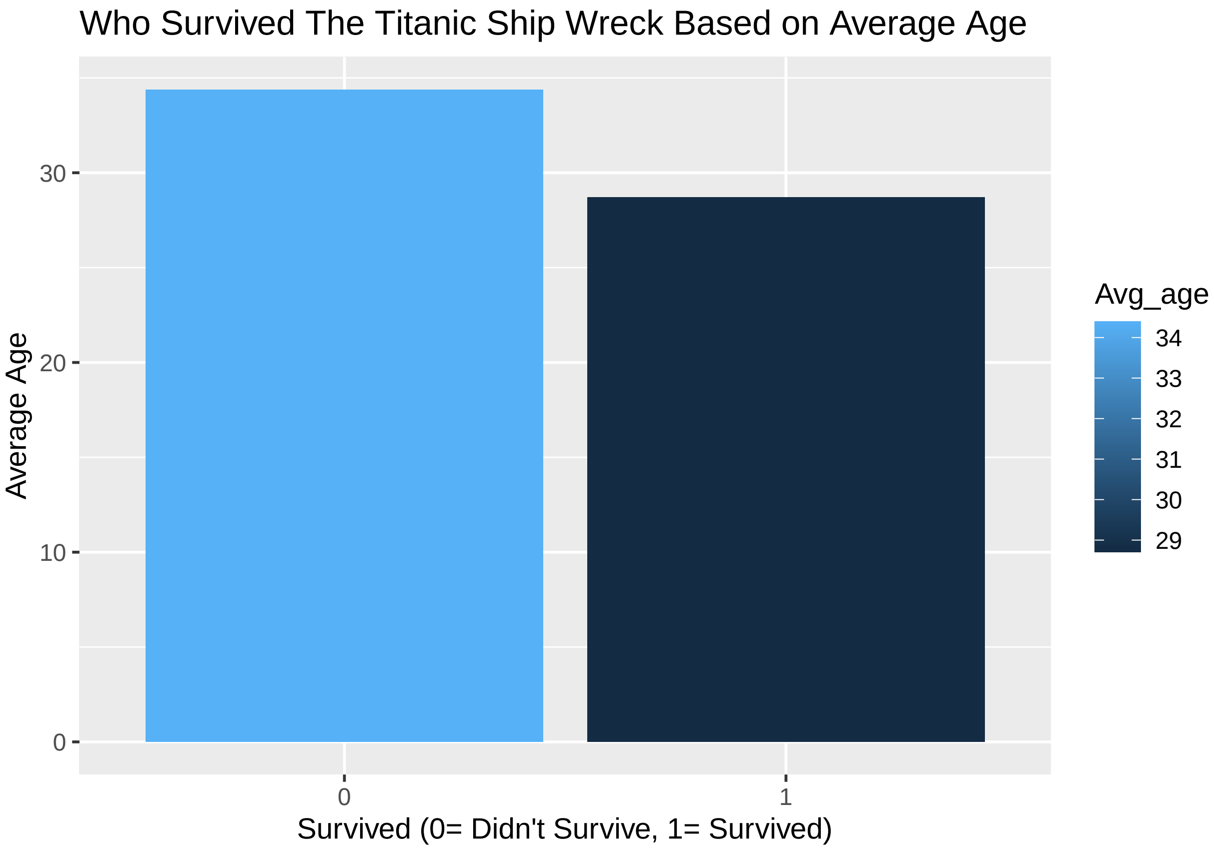
Plot of well defined problem:

```
#Pclass and Survived
new_train%>%
  group_by(Survived,Pclass)%>%
  mutate(ifelse(Survived == 1, "Survivor","Non-Survivor"),
         Survived = as.factor(Survived),
         Pclass = as.factor(Pclass))%>%
  summarize(count= n())%>%
  ggplot(aes(group = Pclass,x=Survived,y= count, fill= Pclass))+
  geom_bar(stat="identity",position = "dodge")+
  ggtitle("Who Survived The Titanic Ship Wreck Based on Pclass")+
  xlab("Survived (0= Didn't Survive, 1= Survived)")+
  ylab("Pclass")
```



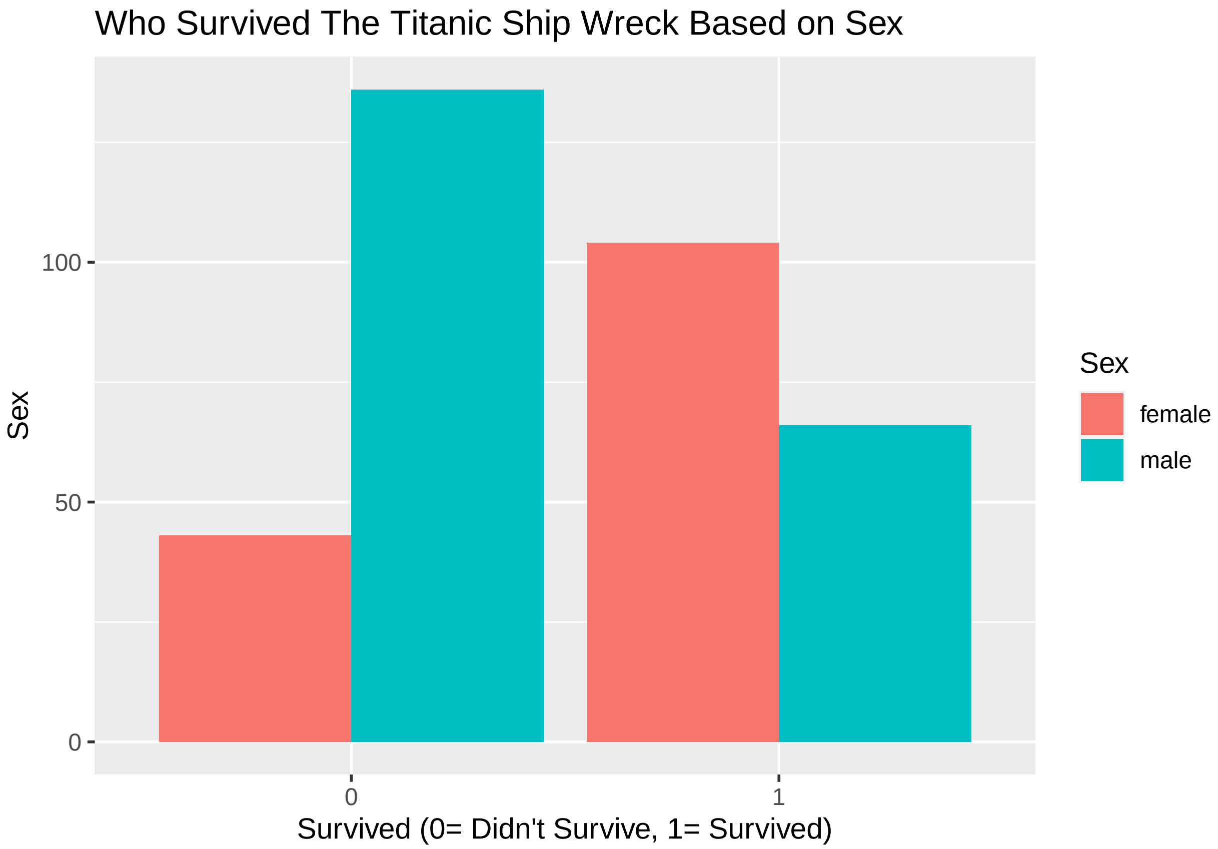
```
#Average Age and Survived

new_train%>%
  group_by(Survived)%>%
  summarize(Avg_age = mean(Age,na.rm=TRUE))%>%
  mutate(ifelse(Survived == 1, "Survivor","Non-Survivor"),
         Survived = as.factor(Survived))%>%
  ggplot(aes(x=Survived,y=Avg_age, fill= Avg_age))+
  geom_bar(stat="identity",position = "dodge")+
  ggtitle("Who Survived The Titanic Ship Wreck Based on Average Age")+
  xlab("Survived (0= Didn't Survive, 1= Survived)")+
  ylab("Average Age")
```



```
#Sex and Survived

new_train%>%
  group_by(Survived,Sex)%>%
  mutate(ifelse(Survived == 1, "Survivor","Non-Survivor"),
         Survived = as.factor(Survived),
         Sex = as.factor(Sex))%>%
  summarize(count= n())%>%
  ggplot(aes(x=Survived,y=count, fill= Sex))+
  geom_bar(stat="identity",position = "dodge")+
  ggtitle("Who Survived The Titanic Ship Wreck Based on Sex")+
  xlab("Survived (0= Didn't Survive, 1= Survived)")+
  ylab("Sex")
```



A.

Is there a relationship between the predictors and the response? Why or why not?

Pclass and Survived: We can see a pretty strong relationship between Pclass and Survived. We can see a big flip in pattern between Pclass 3 in survived and and didn't survive (more people didn't survive than survived). Pclass 2 stays about the same between both groups (survived and didn't survive). And Pclass 1 has an increase with more people survived than didn't survive.

Average age and Survived: With the variables average age and survived, it's harder to see a relationship here. There is only a two year difference in the average age between the groups "survived" and "didn't survived" so the relationship between average age and survived is quite weak.

Sex and Survived: With the variables sex and survived, we can see a very distinct flip in pattern between the two groups on the bar graph. There is a significant amount of males in the "didn't survive" category and much less in the "survived" category. With females on the bar graph, there are a few in the "didn't survive" category but much more in the "survived" category. In short, more males didn't survive than females and more females survived than males.

B. How strong is the relationship between the predictors and the response?

```
mod_mult <- glm(Survived ~ Avg_age + Sex + Pclass, data = new_train, family = "binomial")
summary(mod_mult)
```

```
##
## Call:
## glm(formula = Survived ~ Avg_age + Sex + Pclass, family = "binomial",
##     data = new_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4482  -0.9163  -0.4076   0.9079   2.1001
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  3.949344   0.581419   6.793 1.10e-11 ***
## Avg_age     -0.034075   0.008527  -3.996 6.44e-05 ***
## Sexmale     -1.697571   0.259837  -6.533 6.44e-11 ***
## Pclass      -0.935577   0.168163  -5.564 2.64e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 469.60  on 338  degrees of freedom
## Residual deviance: 379.12  on 335  degrees of freedom
##   (10 observations deleted due to missingness)
## AIC: 387.12
##
## Number of Fisher Scoring iterations: 4
```

To find how strong the relationship is, we look at null and residual deviance. The bigger reduction in residual deviance, the stronger the relationship. (469.60-379.12)/469.60 x 100= about 19% which means the relationship between the predictors and response is pretty weak.

C. What is the effect of the predictors on the response?

```
mod_mult$coefficients
```

```
## (Intercept)     Avg_age     Sexmale      Pclass 
##   3.9493440  -0.0340747  -1.6975712  -0.9355771
```

For the variable Age the log odds is -0.03692902 which means that 0 < odds, therefore as Age increases, the likeliness of survival decreases.

For the variable pclass, the log odds is -1.28854507 which means that 0 < odds, therefore as pclass decreases the likeliness of survival decreases.

For the intercept (Sex Female) log odds is 5.05600619 which means that odds > 1, therefore passengers who are female have a higher chance of survival than passengers who are male. And for Sex Male the log odds is -2.52213086 which means that 0 < odds, therefore passengers who are male are less likely to survive compared to those who are female.

D. What is the value of the measure that is used to assess the accuracy of predictions based on your model?

```
tr<- trainControl(method = "LOOCV")
tr2<-train(
 form = Survived ~ Avg_age + Sex + Pclass ,
   data = new_train,
   trControl = tr,
   method = "glm",
   family = "binomial", na.action = na.pass
 )
 tr2
```

```
## Generalized Linear Model
##
## 349 samples
##   3 predictor
##   2 classes: '0', '1'
##
```

```
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 348, 348, 348, 348, 348, 348, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7168142  0.4332486
```

The accuracy from the generalized linear model is about 72%