Project STAT632

Abhishek Sendil and Jessica Grover

Spring 2022

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
```

```
titanic<-read_csv("titanic.csv")
```

```
## Rows: 891 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): Name, Sex, Ticket, Cabin, Embarked
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Compute theSummary Statistics and removing variables

```
titanic3<-titanic %>%
  select(Survived,Pclass,Sex,Age)%>%
  mutate(Survived=factor(Survived,levels=c(0,1),labels=c("no","yes")))%>%
  mutate(Pclass=factor(Pclass))%>%
  drop_na()
```

```
summary(titanic3)
```

```
##  Survived  Pclass      Sex                  Age
##  no :424   1:186   Length:714         Min.   : 0.42
##  yes:290   2:173   Class :character   1st Qu.:20.12
##            3:355   Mode  :character   Median :28.00
##                                       Mean   :29.70
##                                       3rd Qu.:38.00
##                                       Max.   :80.00
```

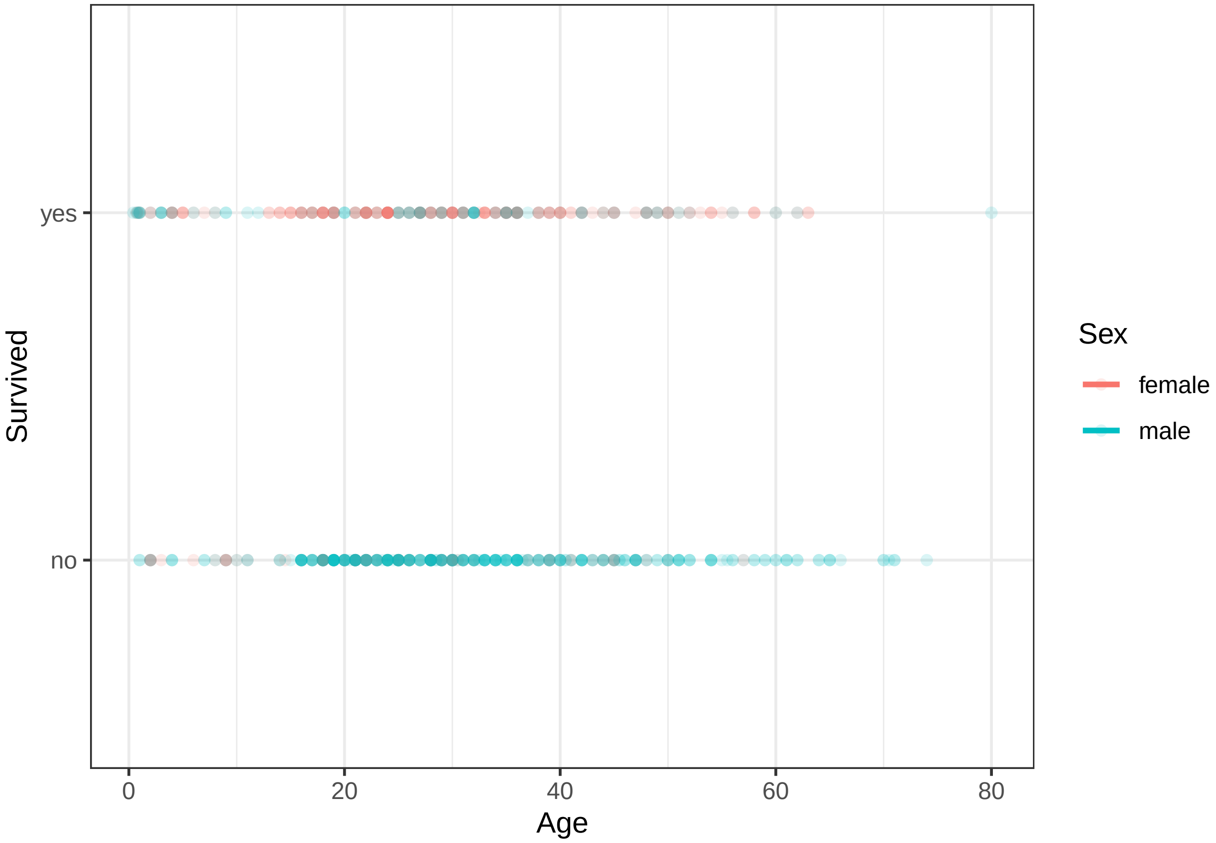
Logistic Regression Scatterplot as Age being the predictor

```
ggplot(titanic3,aes(x=Age,y=Survived,color=Sex))+
geom_point(alpha=0.15)+
geom_smooth(method="glm",method.args=list(family="binomial"),se=FALSE)+
theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```

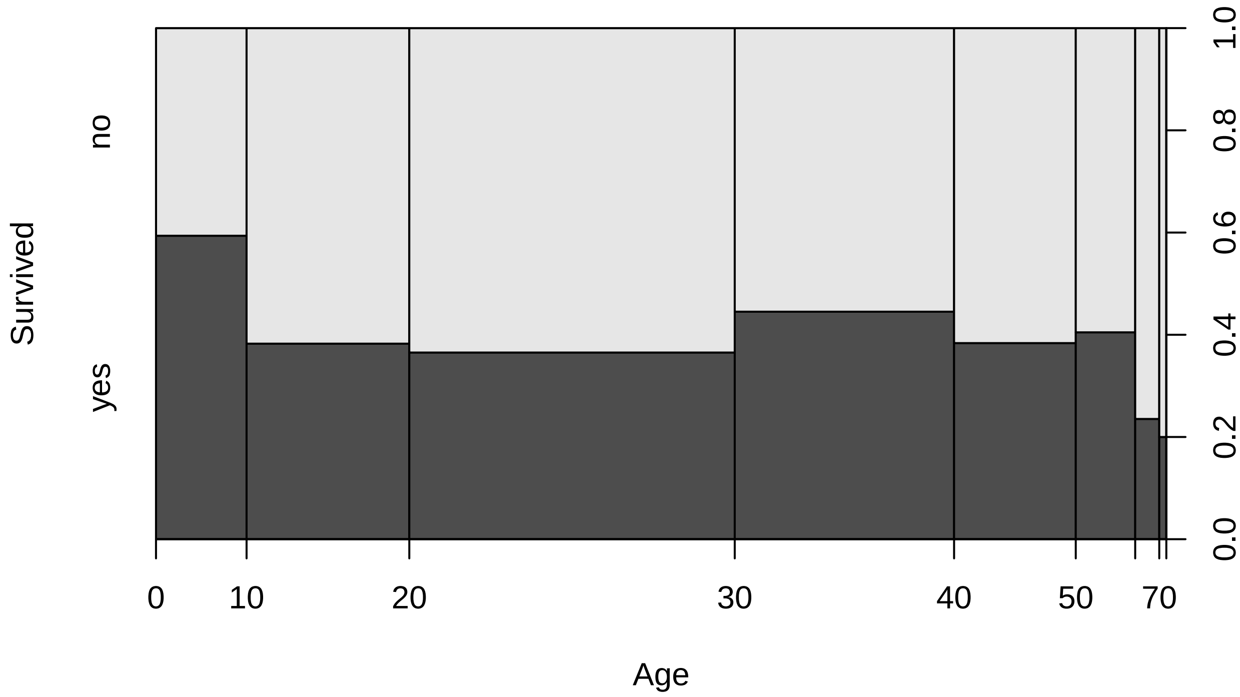
```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: Computation failed in ‘stat_smooth()‘:
## y values must be 0 <= y <= 1
```



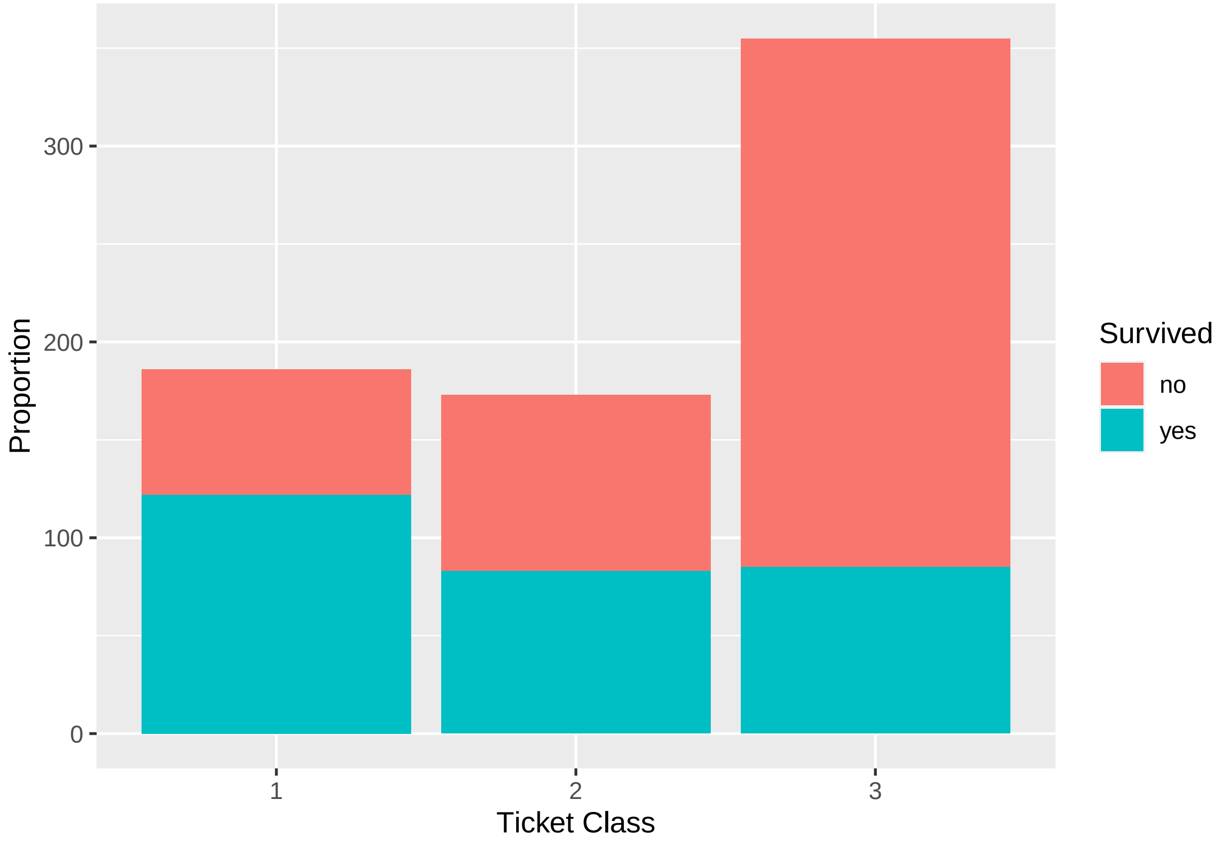
Bar Plots for the 3 predictors

```
plot(Survived~Age,data=titanic3)
```

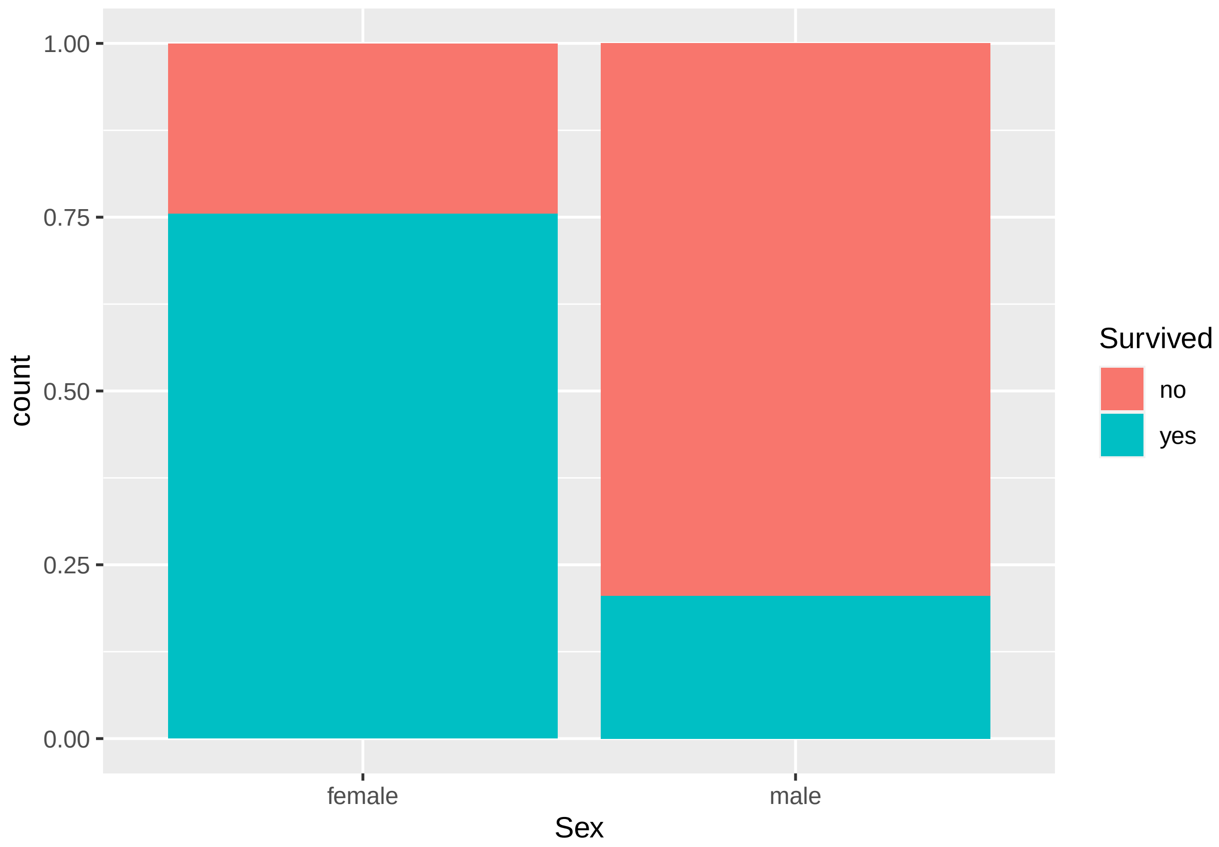


```
ggplot(titanic3,aes(x=Pclass,fill=Survived))+
  geom_bar(posistion="fill") +
xlab("Ticket Class") + ylab("Proportion")
```

```
## Warning: Ignoring unknown parameters: posistion
```



```
ggplot(titanic3 ,aes(x=Sex,fill=Survived))+
 geom_bar(position="fill")
```



Mutiple Logistic Regression Model Summary

```
glm2<-glm(Survived~Age+Sex+Pclass,family ="binomial",data=titanic3)
summary(glm2)
```

```
## 
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass, family = "binomial", 
##     data = titanic3)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7303  -0.6780  -0.3953   0.6485   2.4657  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  3.777013   0.401123   9.416  < 2e-16 ***
## Age         -0.036985   0.007656  -4.831 1.36e-06 ***
## Sexmale     -2.522781   0.207391 -12.164  < 2e-16 ***
## Pclass2     -1.309799   0.278066  -4.710 2.47e-06 ***
## Pclass3     -2.580625   0.281442  -9.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 647.28  on 709  degrees of freedom
## AIC: 657.28
##
## Number of Fisher Scoring iterations: 5
```

Cross Validation

```
set.seed(243)
n<-nrow(titanic3)
train_index<-sample(1:n,round(0.7*n))
titanic_train<-titanic3[train_index, ]
titanic_test<-titanic3[-train_index, ]
```

Regression Summary on test model

```
glm1<-glm(Survived~Age+Sex+Pclass,family ="binomial",data=titanic_test)
summary(glm1)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass, family = "binomial",
##     data = titanic_test)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0597  -0.5818  -0.3372   0.5843   2.2824
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.40995    0.87264   6.200 5.66e-10 ***
## Age         -0.05677    0.01480  -3.835 0.000126 ***
## Sexmale     -2.81839    0.41472  -6.796 1.08e-11 ***
## Pclass2     -1.82661    0.58908  -3.101 0.001930 **
## Pclass3     -3.30289    0.60351  -5.473 4.43e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 295.15  on 213  degrees of freedom
## Residual deviance: 179.14  on 209  degrees of freedom
## AIC: 189.14
##
## Number of Fisher Scoring iterations: 5
```

Regression Summary on Train

```
glm3<-glm(Survived~Age+Pclass+Sex,family="binomial",data=titanic_train)
summary(glm3)
```

```
## 
## Call:
## glm(formula = Survived ~ Age + Pclass + Sex, family = "binomial", 
##     data = titanic_train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.5318  -0.7045  -0.4068   0.6353   2.4383  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  3.220455   0.454879   7.080 1.44e-12 ***
## Age         -0.028473   0.008976  -3.172 0.001514 ** 
## Pclass2     -1.210221   0.320886  -3.771 0.000162 ***
## Pclass3     -2.407876   0.322534  -7.465 8.30e-14 ***
## Sexmale     -2.451571   0.244165 -10.041  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 665.99  on 499  degrees of freedom
## Residual deviance: 458.45  on 495  degrees of freedom
## AIC: 468.45
## 
## Number of Fisher Scoring iterations: 5
```

Confusion Matrix

Accuracy Specificity Sensitivity

```
prob<-predict(glm1,newdata = titanic_test,type="response")
preds<-ifelse(prob>0.5,"yes","no")
```

```
cm<-table(predicted=preds,actual=titanic_test$Survived)
addmargins(cm)
```

```
##          actual
## predicted  no yes Sum
##       no  100  20 120
##       yes  16  78  94
##       Sum 116  98 214
```

Accuracy Percent correctly classified

```
(100+78)/214
```

```
## [1] 0.8317757
```

Sensitivity Percent of people survived correctly classified (1)

```
(78/98)
```

```
## [1] 0.7959184
```

Specificity Percent of people did not survived correctly classified (0)

```
(100/116)
```

```
## [1] 0.862069
```

Roc curve

```
library(pROC)
```

```
## Type ’citation("pROC")’ for a citation.
```

```
##
## Attaching package: ’pROC’
```

```
## The following objects are masked from ’package:stats’:
##
##     cov, smooth, var
```

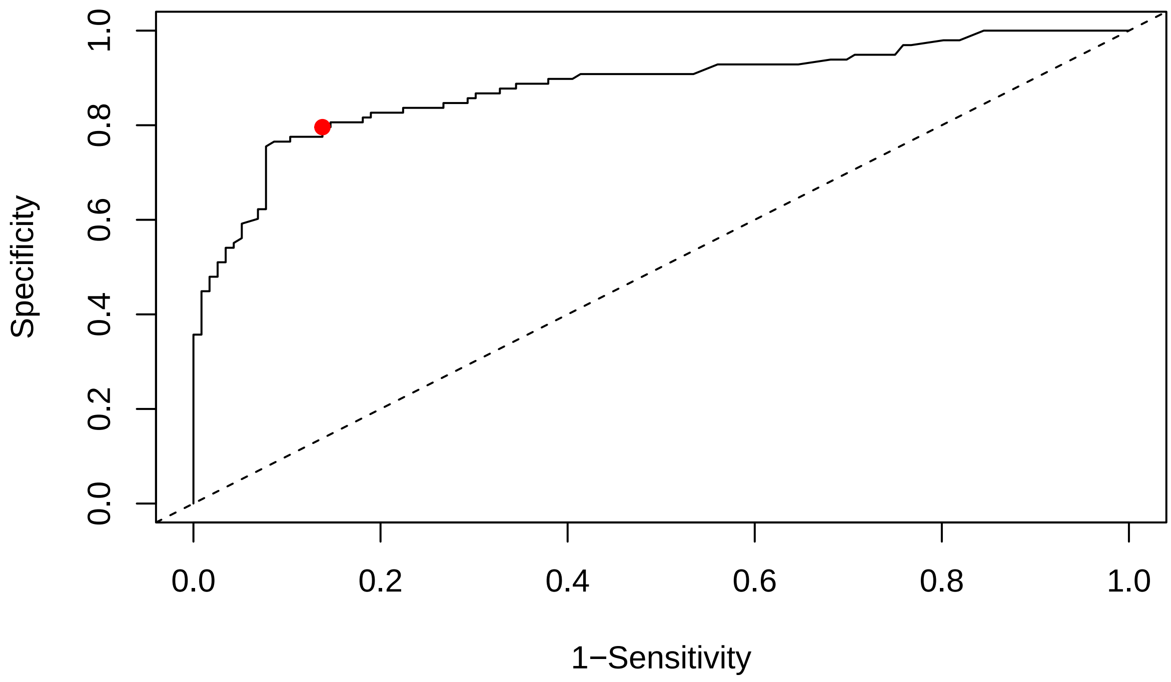
```
roc_obj<-roc(titanic_test$Survived,prob)
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
plot(1-roc_obj$specificities,roc_obj$sensitivities,type="l",
 xlab="1-Sensitivity",ylab="Specificity")

points(x=16/116,y=78/98,col="red",pch=19)
abline(0,1,lty=2)
```



```
auc(roc_obj)
```

```
## Area under the curve: 0.8805
```

Predict the Survial Rate

```
new_x<-data.frame(Age=25,Pclass='3',Sex='male')
predict(glm1,newdata = new_x,type="response")
```

```
##         1
## 0.1061612
```