IMT 573: Problem Set 8 - Regression Part III

LEE CHEN HSIN

Due: Tuesday, Dec 7, 2021

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset8.Rmd` file from Canvas. Open `problemset8.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset8.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object dont' exist
# if you run this on its own it with give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the R Markdown file to `ps8_YourLastName_YourFirstName.rmd`, knit a PDF and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

Setup In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(dplyr)
library(MASS) # Modern applied statistics functions
library(fst)
library(titanic)
library(bayesQR)

titanic_dataset<-read.csv(file = '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/titanic.csv')
```

Problem 1

Data: In this problem set we will use the Titanic dataset. The Titanic text file contains data about the survival of passengers aboard the Titanic. Table 1 contains a description of this data.

Variable	**Description**
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survived	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Table 1: Description of variables in the Titanic Dataset

Problem 1: Part a

1). Load the data and do a quick sanity check. That is, inspect the data using your usual inspect data functions to get a sense of how the variables are encoded and what values they typically take on.

```
data(Titanic)
summary(Titanic)
```

```
## Number of cases in table: 2201
## Number of factors: 4
## Test for independence of all factors:
##  Chisq = 1637.4, df = 25, p-value = 0
##  Chi-squared approximation may be incorrect
```

```
head(titanic_dataset)
```

```
##   pclass survived                                            name    sex
## 1      1        1                   Allen, Miss. Elisabeth Walton female
## 2      1        1                  Allison, Master. Hudson Trevor   male
## 3      1        0                    Allison, Miss. Helen Loraine female
## 4      1        0            Allison, Mr. Hudson Joshua Creighton   male
## 5      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
## 6      1        1                             Anderson, Mr. Harry   male
##       age sibsp parch ticket     fare   cabin embarked boat body
## 1 29.0000     0     0  24160 211.3375      B5        S    2   NA
## 2  0.9167     1     2 113781 151.5500 C22 C26        S   11   NA
## 3  2.0000     1     2 113781 151.5500 C22 C26        S        NA
## 4 30.0000     1     2 113781 151.5500 C22 C26        S       135
## 5 25.0000     1     2 113781 151.5500 C22 C26        S        NA
## 6 48.0000     0     0  19952  26.5500     E12        S    3   NA
##                         home.dest
## 1                    St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                    New York, NY
```

2). Are there missing values for any of the important variables? Find and list those. Based on missing values, reflect whether they are going useful for downstream modeling tasks.

```
#Yes, there are missing values in the column of age,fare,body

summary(titanic_dataset)
```

```
##      pclass         survived                                  name     
##  Min.   :1.000   Min.   :0.000   Connolly, Miss. Kate           :   2  
##  1st Qu.:2.000   1st Qu.:0.000   Kelly, Mr. James               :   2  
##  Median :3.000   Median :0.000   Abbing, Mr. Anthony            :   1  
##  Mean   :2.295   Mean   :0.382   Abbott, Master. Eugene Joseph  :   1  
##  3rd Qu.:3.000   3rd Qu.:1.000   Abbott, Mr. Rossmore Edward    :   1  
##  Max.   :3.000   Max.   :1.000   Abbott, Mrs. Stanton (Rosa Hunt):   1  
##                                  (Other)                        :1301  
##      sex           age              sibsp            parch      
##  female:466   Min.   : 0.1667   Min.   :0.0000   Min.   :0.000  
##  male  :843   1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000  
##               Median :28.0000   Median :0.0000   Median :0.000  
```

```
##                Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##                3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##                Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##                NA's   :263
##       ticket          fare                     cabin      embarked
##  CA. 2343:  11   Min.   :  0.000                  :1014    :  2
##  1601    :   8   1st Qu.:  7.896   C23 C25 C27    :   6   C:270
##  CA 2144 :   8   Median : 14.454   B57 B59 B63 B66:   5   Q:123
##  3101295 :   7   Mean   : 33.295   G6             :   5   S:914
##  347077  :   7   3rd Qu.: 31.275   B96 B98        :   4
##  347082  :   7   Max.   :512.329   C22 C26        :   4
##  (Other) :1261   NA's   :1         (Other)        : 271
##       boat          body                      home.dest
##         :823   Min.   :  1.0                       :564
##  13     : 39   1st Qu.: 72.0   New York, NY        : 64
##  C      : 38   Median :155.0   London              : 14
##  15     : 37   Mean   :160.8   Montreal, PQ        : 10
##  14     : 33   3rd Qu.:256.0   Cornwall / Akron, OH:  9
##  4      : 31   Max.   :328.0   Paris, France       :  9
##  (Other):308   NA's   :1188    (Other)             :639
```

```
#the reason of why the body is missing might because that if someone is survived,
#then his/her body may not be found.Besides, if someone is not survived, then
#his/her body may also not be found.
```

Problem 1: Part b (Categorical output)

1). Our goal is to determine the survival of passengers that takes into account the socioeconomic status of the passengers. What model would you fit? Explain the choice of your model and then fit the model.

```
#I would utilize simple liner regression model to determine the survival of
#passengers since this model can take into account of the relation between
#pclass and whether survived together.

model <- glm(survived ~ pclass, data =titanic_dataset,family=binomial)
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = titanic_dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3909  -0.7683  -0.7683   0.9780   1.6518
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26802    0.16792   7.551 4.31e-14 ***
## pclass      -0.77900    0.07096 -10.978  < 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: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1307  degrees of freedom
```

```
## AIC: 1617.3
##
## Number of Fisher Scoring iterations: 4
```

```
#according to the model, since the p-value of the fare is less than 0.05
#so it shows a regression between pclass and survived.
```

2). What might you conclude based on this model about the probability of survival for lower class passengers?

```
#I would utilize simple liner regression model to determine the survival of
#passengers since this model can take into account of the relation between
#pclass and whether survived together.

model <- glm(survived ~ pclass, data =titanic_dataset,family=binomial)
summary(model)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = titanic_dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3909  -0.7683  -0.7683   0.9780   1.6518
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26802    0.16792   7.551 4.31e-14 ***
## pclass      -0.77900    0.07096 -10.978  < 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: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1307  degrees of freedom
## AIC: 1617.3
##
## Number of Fisher Scoring iterations: 4
```

```
#according to the model, since the p-value of the fare is less than 0.05
#so it shows a regression between pclass and survived.
```

3). Create a new variable child, that is 1 if the passenger was youger than 14 years old. Check to make sure you have the new variable added in your dataframe.

```
titanic_dataset$child <-ifelse(titanic_dataset$age<14, "1", "0")
```

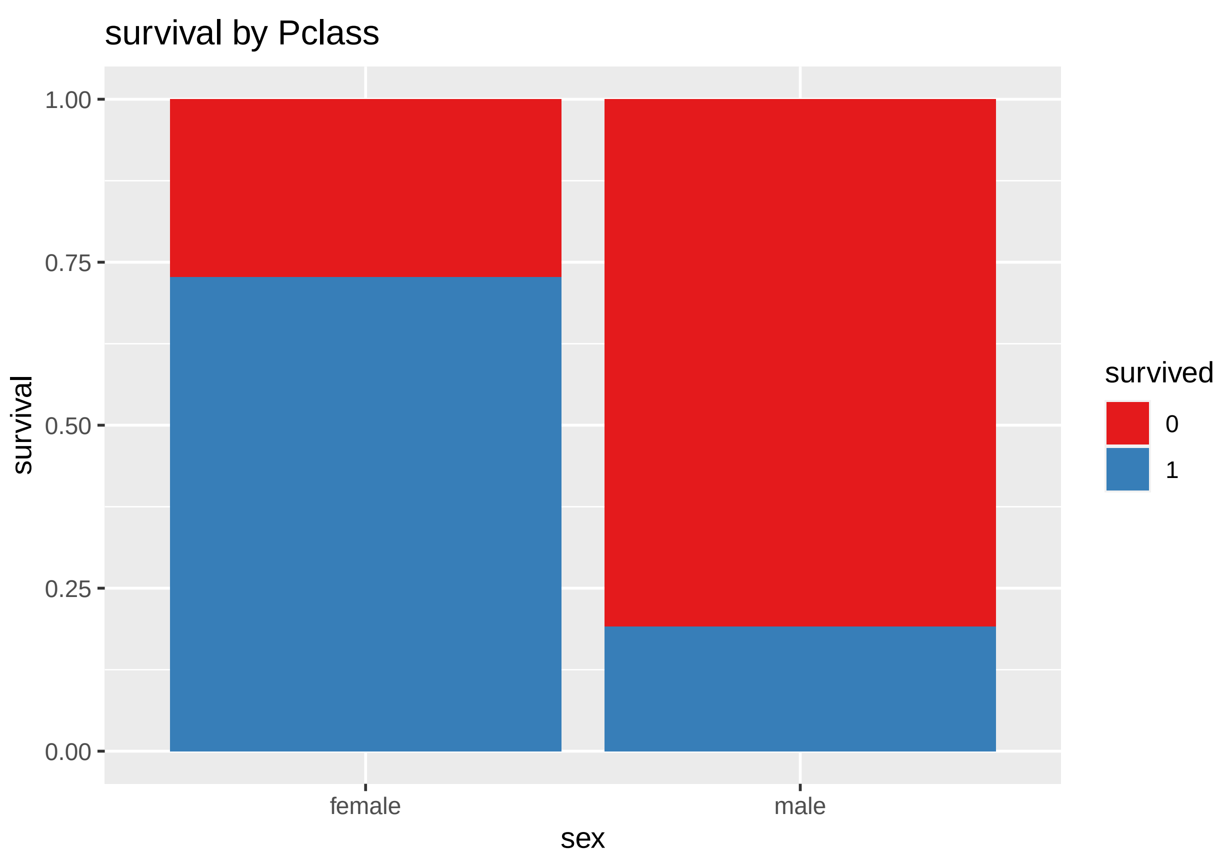
4). Now you are curious to know whether men or women, old or young, or people of different passenger classes have larger chances of survival. Build an appropriate model to answer this curiosity. Explain the choice of your model. Interpret results

```
model4.1 <- glm(survived ~ sex, data =titanic_dataset,family=binomial)
summary(model4.1)
```

```
##
## Call:
## glm(formula = survived ~ sex, family = binomial, data = titanic_dataset)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6124  -0.6511  -0.6511   0.7977   1.8196
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.9818     0.1040   9.437   <2e-16 ***
## sexmale      -2.4254     0.1360 -17.832   <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: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1368.1  on 1307  degrees of freedom
## AIC: 1372.1
##
## Number of Fisher Scoring iterations: 4
```

```
a1<-ggplot(titanic_dataset ,aes(sex,fill=survived))+
geom_bar(aes(fill=factor(survived)),position = "fill")+
  scale_fill_brewer(palette = "Set1")+
  ylab("survival")+
  ggtitle("survival by Pclass")

a1
```

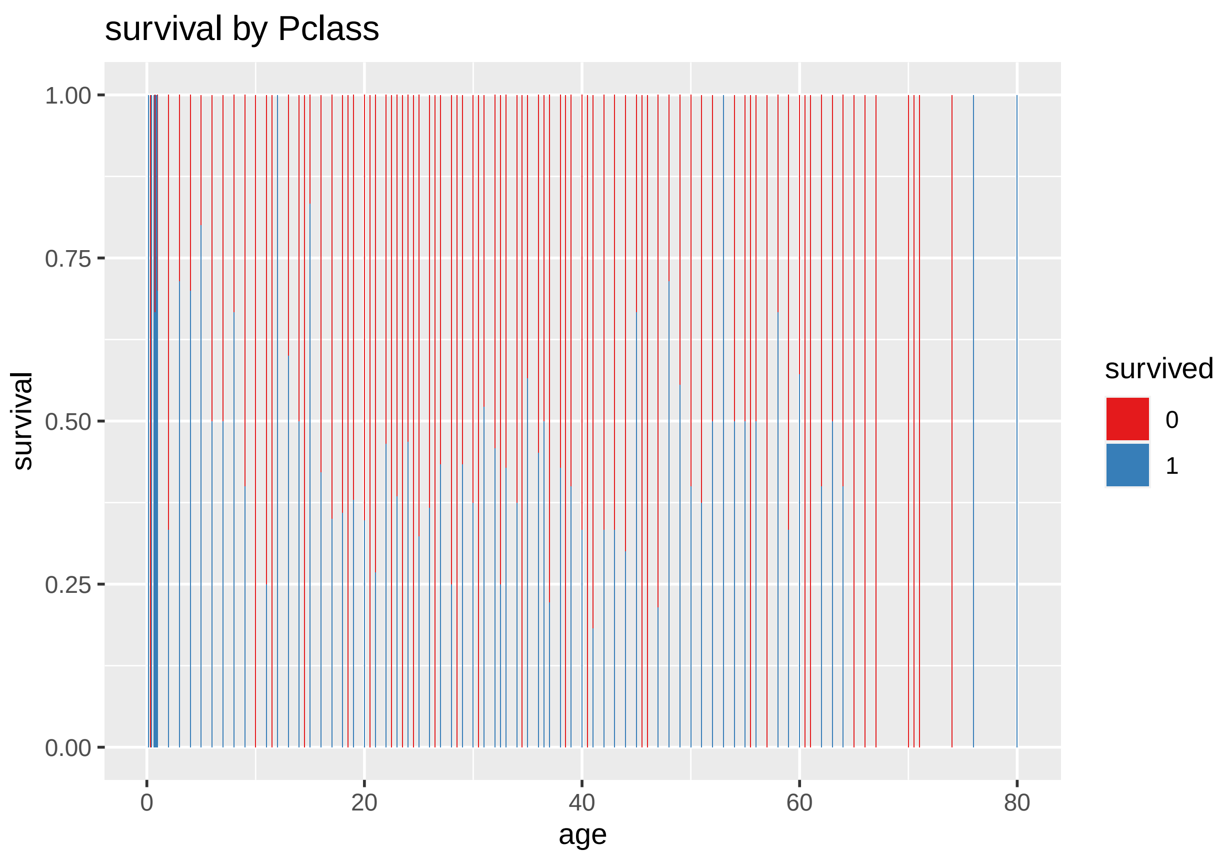


```r
#in the chart, it's clear that the survival of women is higher than men

model4.2 <- glm(survived ~ age, data =titanic_dataset,family=binomial)
summary(model4.2)
```

```
##
## Call:
## glm(formula = survived ~ age, family = binomial, data = titanic_dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1189  -1.0361  -0.9768   1.3187   1.5162
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.136531   0.144715  -0.943   0.3455
## age         -0.007899   0.004407  -1.792   0.0731 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance: 1411.4  on 1044  degrees of freedom
##   (263 observations deleted due to missingness)
## AIC: 1415.4
##
## Number of Fisher Scoring iterations: 4
```

```r
a2<-ggplot(titanic_dataset ,aes(age,fill=survived))+
geom_bar(aes(fill=factor(survived)),position = "fill")+
  scale_fill_brewer(palette = "Set1")+
  ylab("survival")+
  ggtitle("survival by Pclass")

a2
```

```
## Warning: Removed 263 rows containing non-finite values (stat_count).
```

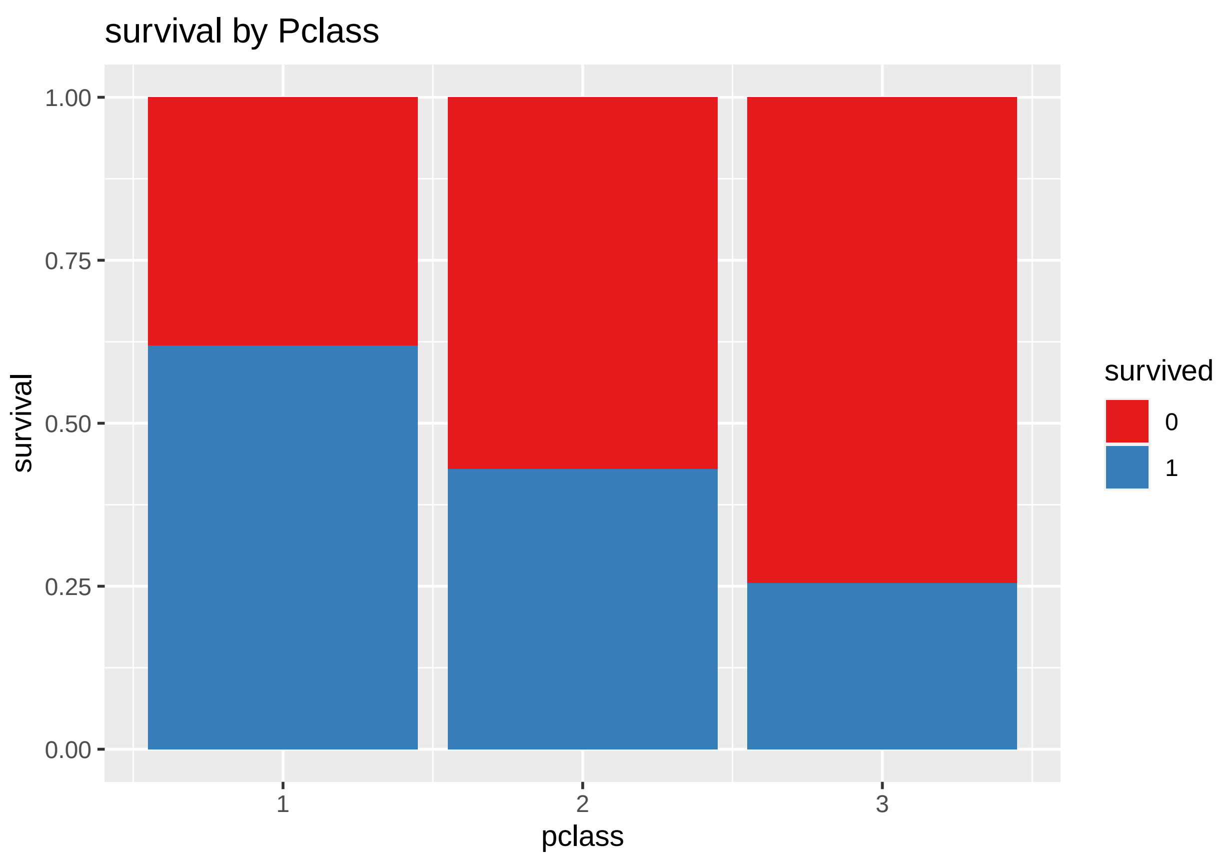



```r
#in the chart, it's clear that the survival of the oldest and the youngest
#is higher

model4.3 <- glm(survived ~ pclass, data =titanic_dataset,family=binomial)
summary(model4.3)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = titanic_dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3909  -0.7683  -0.7683   0.9780   1.6518
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.26802    0.16792   7.551 4.31e-14 ***
## pclass      -0.77900    0.07096 -10.978  < 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: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1307  degrees of freedom
## AIC: 1617.3
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
a3<-ggplot(titanic_dataset ,aes(pclass,fill=survived))+
  geom_bar(aes(fill=factor(survived)),position = "fill")+
  scale_fill_brewer(palette = "Set1")+
  ylab("survival")+
  ggtitle("survival by Pclass")

a3
```




```
#in the chart, it's clear that the survival of lower class is higher than others
```

Problem 1: Part c - Predictions with a categorical output Now let's try to do some predictions with the Titanic data. Our goal is to predict the survival of passengers by considering only the socioeconomic status of the passenger.

1). After loading the data, split your data into a *training* and *test* set based on an 80-20 split. In other words, 80% of the observations will be in the training set and 20% will be in the test set. Remember to set the random seed.

```
library(simEd)
```

```
## Loading required package: rstream

##
## Attaching package: 'simEd'

## The following objects are masked from 'package:base':
##
```

```
##     sample, set.seed
```

```
set.seed(1)

train_row <- sample(1309,1309*0.8)
train <- titanic_dataset[train_row, ]
test <- titanic_dataset[-train_row, ]
```

2). Fit the model described above (that is in Problem 1 (c), that only takes into account socio-economic status).

```
model5 <- glm(survived ~ pclass, data =train,family=binomial)
summary(model5)
```

```
##
## Call:
## glm(formula = survived ~ pclass, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4283  -0.7645  -0.7645   0.9457   1.6570
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.39950    0.18821   7.436 1.04e-13 ***
## pclass      -0.82671    0.07959 -10.387  < 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: 1399.2  on 1046  degrees of freedom
## Residual deviance: 1283.6  on 1045  degrees of freedom
## AIC: 1287.6
##
## Number of Fisher Scoring iterations: 4
```

```
#according to the model, since the p-value of the pclass is less than 0.05
#so it shows a regression between pclass and survived
```

3). Predict the survival of passengers for each observation in your test set using the model fit that you just fitted. Save these predictions as yhat.

```
library(broom)
library(dplyr)

titanic_dataset_new <- augment(model5,newdata=test,type.predict = 'response')%>%
  mutate(survive_predict=round(.fitted))
```

4). Use a threshold of 0.5 to classify predictions. What is the number of false positives on the test data? Interpret this in your own words. *Hint: You need to show confusion matrix*

```
install.packages('SDMTools', repos = "http://cran.us.r-project.org")
```

```
## Warning: package 'SDMTools' is not available (for R version 3.6.2)
```

```
library(SDMTools)
```

```
confMatrixNew<-confusion.matrix(titanic_dataset_new$survived,titanic_dataset_new$survive_predict,thresh
confMatrixNew
```

```
##     obs
## pred   0  1
##    0 140 62
##    1  29 31
## attr(,"class")
## [1] "confusion.matrix"
```

```
#number of false = 29+62=91
#number of positive =140+31=171
```

5). Pick a different threshold to classify predictions and interpret your results again. Did you have a rationale when picking a different threshold? Did you see any change? Reflect on your results.

```
titanic_dataset_new$survive_predict<-
  if(titanic_dataset_new$.fitted > 0.4){
    titanic_dataset_new$survive_predict=1
  }else{
          titanic_dataset_new$survive_predict=0
  }
```

```
## Warning in if (titanic_dataset_new$.fitted > 0.4) {:     > 1
##
```

```
confMatrixNew2<-confusion.matrix(titanic_dataset_new$survived,titanic_dataset_new$survive_predict)
confMatrixNew2
```

```
##     obs
## pred   0  1
##    0   0  0
##    1 169 93
## attr(,"class")
## [1] "confusion.matrix"
```

```
#number of false =0+169=169
#number of positive =0+93=93

#when the threshold is decreased, the number of false increase.
#As a result, the accuracy of the matrix decreases.
```

Problem 2: Customer Churn data In this problem, you will work with the churn dataset. Documentation of the dataset can be found here: https://www.rdocumentation.org/packages/bayesQR/versions/2.3/topics/Churn

The dataset is random sample from all active customers (at the end of June 2006) of a European financial services company. The data captures the churn behavior of the customers in the period from July 1st until December 31th 2006. Here a churned customer is defined as someone who closed all his/her bank accounts with the company.

1). Read and inspect the data. *Hint: the file is an fst fast-storage format file. Check your regression lab to figure out how you can read this file*

```
data("Churn")

churn_data<-read_fst(
  '/Users/leechenhsin/Desktop/study@USA/07_UW_School/IMT573/churn.fst',
  columns = NULL,
```

```
  from = 1,
  to = NULL,
  as.data.table = FALSE,
  old_format = FALSE
)
```

2). Describe the data and variables that are part of the churn dataset.

```
#churn : churn (yes/no)
#gender : gender of the customer (male = 1)
#Social_Class_Score : social class of the customer
#lor : length of relationship with the customer
#recency : number of days since last purchase
#time_since_first_purchase : the standardization of time since first purchase
#time_since_last_purchase : the standardization of time since last purchase
```

3). Considering this data in context, what is the response variable of interest?

```
#churn is the response variable of interest
```

4). Our goal is to determine customer churn. Which variables do you think are the most important ones to describe customer churn? How should those be related to the churn? Interpret your results.

```
model1 <- glm(churn ~ gender, data =Churn,family=binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = churn ~ gender, family = binomial, data = Churn)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21977  -1.13479   0.00042   1.13563   1.22063
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.09909    0.14089   0.703    0.482
## gender      -0.20019    0.20026  -1.000    0.317
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.52  on 399  degrees of freedom
## Residual deviance: 553.52  on 398  degrees of freedom
## AIC: 557.52
##
## Number of Fisher Scoring iterations: 3
```

```
#since its p-value is  0.317, which is larger than 0.05,
#so it means that the gender does not have a strong regression with churn

model2 <- glm(churn ~ Social_Class_Score, data =Churn,family=binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = churn ~ Social_Class_Score, family = binomial,
##     data = Churn)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21117  -1.17670   0.00729   1.17709   1.21405
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.001438   0.100127   0.014    0.989
## Social_Class_Score 0.027575   0.092429   0.298    0.765
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.52  on 399  degrees of freedom
## Residual deviance: 554.43  on 398  degrees of freedom
## AIC: 558.43
##
## Number of Fisher Scoring iterations: 3
```

```
#since its p-value is 0.765, which is larger than 0.05,
#so it means that the Social_Class_Score does not have a strong regression
#with churn

model3 <- glm(churn ~ lor, data =Churn,family=binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = churn ~ lor, family = binomial, data = Churn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3646  -1.1488   0.1584   1.1166   1.6143
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.01518    0.10156  -0.150  0.88115
## lor         -0.35479    0.11095  -3.198  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.52  on 399  degrees of freedom
## Residual deviance: 543.73  on 398  degrees of freedom
## AIC: 547.73
##
## Number of Fisher Scoring iterations: 4
```

```
#since its p-value is 0.00139, which is smaller than 0.05,
#so it means that the lor,length of relationship with the customer
#has a strong regression with churn

model4 <- glm(churn ~ recency, data =Churn,family=binomial)
summary(model4)
```

```
##
## Call:
## glm(formula = churn ~ recency, family = binomial, data = Churn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8716  -1.1146  -0.2007   1.1996   1.2937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.03502    0.10152  -0.345  0.73013
## recency      0.26921    0.09812   2.744  0.00607 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 554.52  on 399  degrees of freedom
## Residual deviance: 546.40  on 398  degrees of freedom
## AIC: 550.4
##
## Number of Fisher Scoring iterations: 4
```

```
#since its p-value is 0.00607, which is smaller than 0.05,
#so it means that the recency has a strong regression with churn


#lor,length of relationship and recency are the most important ones to
#describe customer churn
```