Anqi Xia (U84859358)

Research of Survive rate of passengers in Titanic

1. Questions.

RMS Titanic was a British passenger liner operated by the White Star Line that sank in the North Atlantic Ocean on 15 April 1912.by researching the dataset I found the survive rate is only 38.38%. So, I wonder what kind of variables have an influence on the survive rate of a passenger. Here are some questions I want to research. First, I wonder is there a difference between the survive rate between males and females. Second, is there any other variables also influence the survive rate a lot? I also want to create a model to predict weather a passenger can survive if he is in Titanic.

2. Dataset description

Here is the definition of each variables in the dataset.

Variable	Definition	Key
survival	Survival	0 = No 1 = Yes
pclass	Ticket class	1 = 1st 2 = 2nd 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	ticket price	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg Q = Queenstown S = Southampton

I will use the passenger class, sex, age, number of siblings or spouses, number of parents or children to finish my research. Because all these variables may have some influence on the survive rate if a passenger. There are some reasons why I didn't use the other variables. For ticket number, I don't think it has some value for the analysis. For ticket price, I notice the price of 1st class tickets is highest. And the price of 2nd class is lower, and 3rd class have the cheapest ticket. So, I think the price and passenger are two dependent variables, so we just need to consider one of them. and for cabin, there are too many

missing values in the data set, so I have to abandon this variable, finally, for the port of embarkation, I also think it doesn't have too much influence on the survive rate.

```
> summary(data)
  PassengerId       Survived          Pclass                                      Name         Sex          Age            SibSp
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Abbing, Mr. Anthony                  :  1   female:314   Min.   : 0.42   Min.   :0.000
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Abbott, Mr. Rossmore Edward          :  1   male  :577   1st Qu.:20.12   1st Qu.:0.000
 Median :446.0   Median :0.0000   Median :3.000   Abbott, Mrs. Stanton (Rosa Hunt)     :  1                Median :28.00   Median :0.000
 Mean   :446.0   Mean   :0.3838   Mean   :2.309   Abelson, Mr. Samuel                  :  1                Mean   :29.70   Mean   :0.523
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000   Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00   3rd Qu.:1.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000   Adahl, Mr. Mauritz Nils Martin       :  1                Max.   :80.00   Max.   :8.000
                                                  (Other)                              :885                NA's   :177
     Parch             Ticket         Fare               Cabin     Embarked
 Min.   :0.0000   1601     :  7   Min.   :  0.00             :687    :  2
 1st Qu.:0.0000   347082   :  7   1st Qu.:  7.91   B96 B98   :  4   C:168
 Median :0.0000   CA. 2343:  7   Median : 14.45   C23 C25 C27:  4   Q: 77
 Mean   :0.3816   3101295  :  6   Mean   : 32.20   G6        :  4   S:644
 3rd Qu.:0.0000   347088   :  6   3rd Qu.: 31.00   C22 C26   :  3
 Max.   :6.0000   CA 2144  :  6   Max.   :512.33   D         :  3
                  (Other)  :852                    (Other)   :186
```

From the summary above we can find there are 177 missing values in age and 687 empty cells in cabin. So I decide to use average value to replace the missing value in age. And abandond the variable cabin.

Link of the dataset: https://github.com/awesomedata/awesome-public-datasets/tree/master/Datasets

3. Statistical methods

First, because I want to know is there a difference between the survive rate of males and females so I want to do a two-sample test for the proportion. Second, I want to build a multiple logistic regression model to help me to predict if a customer can survive in the boat. And use c-statistic score to check the model.

4. Result.

PART A.

	female	male	sum
not survived	81	468	549
survived	233	109	342
sum	314	577	891
proportion	0.742038217	0.188908146	0.383838384

PART B.

Risk difference = p1 -p2 = 0.742 – 0.189 = 65.3%

1. Set up the hypotheses and select the alpha level

H0 : p1 = p2 (the proportion of survive rate is same across males and females)

H1 : p1 ≠ p2 (the proportion of survive rate is not same across males and females)

α = 0.05

2. Select the appropriate test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

3. State the decision rule

Determine the appropriate critical value from the standard normal distribution associated with a right-hand tail probability of $\alpha/2 = 0.05/2 = 0.025$.

Decision Rule: Reject H0 if $|z| \geq 1.960$, Otherwise, do not reject H0

4. Compute the test statistic and the associated p-value

P1=74.2%

P2=18.9%

Z=(0.742-0.189)/sqrt(0.383*(1-0.383)*(1/314+1/577))= 16.2

5. Conclusion

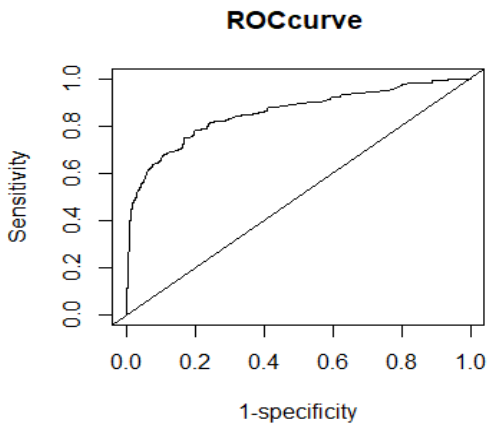
Reject H0 since $16.2 \geq 1.960$.

We have significant evidence at the $\alpha = 0.05$ level that $p1 \neq p2$. We reject the null hypothesis that the proportion of survive rate is same across males and females. The survive rate of females is 65.3% higher than the survive rate of males

PART C.

```
> m<-glm(data$Survived~data$Sex+data$Pclass+data$Age+data$SibSp+data$Parch, family =binomial)
> summary(m)

Call:
glm(formula = data$Survived ~ data$Sex + data$Pclass + data$Age +
    data$SibSp + data$Parch, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6536  -0.6147  -0.4224   0.6133   2.4324

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.241404   0.483589  10.839  < 2e-16 ***
data$Sex1   -2.768190   0.198718 -13.930  < 2e-16 ***
data$Pclass -1.172846   0.119687  -9.799  < 2e-16 ***
data$Age    -0.040103   0.007778  -5.156 2.52e-07 ***
data$SibSp  -0.334326   0.108557  -3.080  0.00207 **
data$Parch  -0.081624   0.114688  -0.712  0.47665
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  790.33  on 885  degrees of freedom
AIC: 802.33

Number of Fisher Scoring iterations: 5
```

I also build a multiple logistic regression model that can predict weather a passenger can survive in the event. And the c-statistic score is 0.8541.



5. Conclusion and limitations.

After I analyze the dataset, I found we have evidence to show the survive rate of females is higher than males. And other two important factors are passenger class and age. Because the coefficients of these two variables are negative so it seems the first-class passengers may have a higher survive rate and young people may also have a higher survive rate.

I think the research also have some limits.

First, in the model the class of passenger is an important variable, but we can not make sure all passengers were staying in their own room when the boat was sinking. Maybe many passengers are in the restaurant or on the deck of the ship. So, it may influence the relationship between survive rate and passenger class.

Second, 'sibsp' and 'parch' were considered as two factors that are not important. But I think there might be some condition was not considered in this variable. For example, 'silsp' is number of siblings / spouses aboard the Titanic. If a female aboard the ship with her boyfriend, the condition will not be collected that will influence the accuracy of this variable. And another example is 'parch' is the number of parents / children aboard the Titanic. If a baby board the ship with a nanny. The condition also will not be collected. All of these conditions will influence the accuracy of the dataset.

Third, because Titanic's right side hit the iceberg and start to sink first, so the people in that part might have a very low survive rate, but the reason seems was not considered in the dataset.