Statistics course project report

TITANIC

Logistic Regression
Poisson Regression
Bayesian Analysis

Students:
Marta Ortigas
Mònica Torner
Núria Mitjavila

Teacher Supervisor:
Jan Graffelman

Statistical Models and Stochastic Processes
First term, Second year, 2022 - 2023
Bioinformatics degree, ESCI - UPF

Contents

1. Introduction

This is the report of the final project for the Statistical Models and Stochastic Processes course. The main goal of this project is to analyze a data set with at least one statistical technique that we have treated during the course. The techniques that we could choose were maximum likelihood estimation, likelihood ratio tests, logistic regression, Poisson regression, mixed effects models, multiple testing and false discovery rate, Bayesian inference and Markov models, that are the different topics that we work in this course.

In our case, we decide to analyze the data we selected with the logistic regression model because it was one of the techniques that we found more interesting during the course and specially the exercises that we did with other data sets and because when we take a look at the data that we selected (this data will be explained in the following page, in the description section), it seems to fit in the logistic regression model. But we also use two more techniques, Poisson regression and Bayesian inference, to do a better analysis of our data set.

The structure of this report, is done with the goal to include all the necessary information, clearly distributed and easy to understand. First it has an introduction, that provides context and contains the statement of a goal of the project, and is followed by a description part, where explains the data set, including a descriptive analysis (numerical and graphical) of the data, and does an explicit reference to the origin of the data that we selected. Then, we have a new section, that came up before the results, and this is the techniques section, where it has a brief description of the techniques we used to analyze the data. In our case, it explains the different ways we analyze the data with logistic regression, Bayesian inference and Poisson regression.

The most important section of the report came after this introduction, description and techniques, and is the results section. In this part, we have explained the results of the analysis of the data, including the relevant statistics together with graphics, plots, and tables to understand the data set we are working with. The results are followed with the conclusions and discussion of the results, where we include a brief summary of the report's main points, if the main goal was reached and if we have found some problems during the data set analysis. Finally, in the last part of the report, you can found the bibliography with all the sources we used to do our project, and the appendix, where all the R scripts will be written. During the results, there will be some citations to the R scripts, because if not, there was not enough space in the results to put all the plots and tables. The R scripts, are also upload in a GitHub repository. [6].

The main hypothesis of our analysis is that our data fits the logistic regression and shows a relation with the variables Survive, Age, Class, and Sex. The hypothesis came from learning about the historical and social constructs of the beginning of the 20th century.

2. Description

For this project, we select the data set Titanic, that has the original data from Titanic competition but it has some changes that applied to be better suited for binary logistic regression. This data set has been taken from the Kaggle website [5], "an online community platform for data scientists and machine learning enthusiasts that allows the users to collaborate with the other users, find and publish datasets, use GPU integrated notebooks, and compete with other data scientists to solve different data science challenges". [8]

The overview of our data is detailed in the web page, and is explaining the following. The modifications that this data set has from the original data [7] from Titanic are:
- Merged the train and test data.
- Removed the "ticket" and "cabin" attributes.
- Moved the "Survived" attribute to the last column.
- Substituted values of "Sex" and "Embarked" attributes with binary and categorical values.
- Filled the missing values in "Age" and "Fare" attributes with the median of the data.

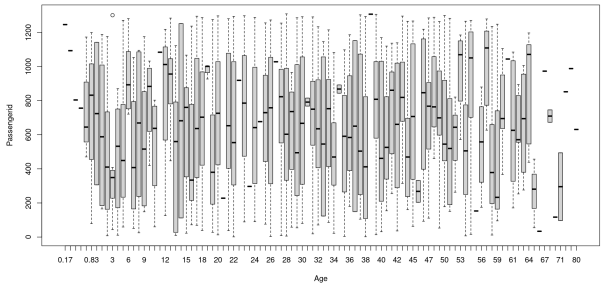
The different variables that are used in this data set and divided by columns are the following:
- Passengerid: the identifier for each one of the 1309 Titanic passengers.
- Age: is fractional if less than 1. If the age is estimated, is it in the form of xx.5
- Fare: indicates the fare for each one of the passengers.
- Sex: indicates the sex of the passenger: 0 = Male, 1 = Female.
- Sibsp: defines the family relations for sibling (brother, sister) and spouse (husband, wife).
- Pclass: ticket class, a proxy for socio-economic status: 1 = Upper, 2 = Middle, 3 = Lower.
- Embarked: the port of embarkation: 0 = Cherbourg, 1 = Queenstown, 2 = Southampton.
- Survived: indicates if the passenger survives or not: 0 = No, 1 = Yes.
- Parch: defines the family relations for parents (mothers, fathers) and child (daughters, sons). If the children was travelling only with a nanny, therefore the Parch = 0 for all of them.

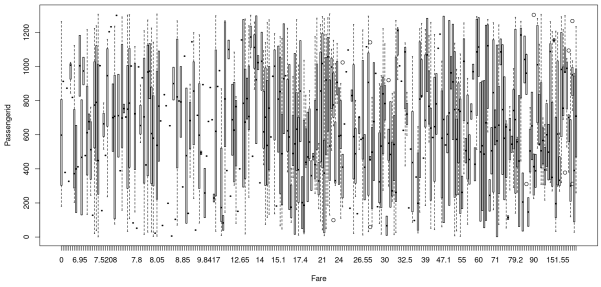
Passengerid	Age	Fare	Sex	sibsp	Parch	Pclass	Embarked	X2urvived
1	22	7.2500	0	1	0	3	2	0
2	38	71.2833	1	1	0	1	0	1
3	26	7.9250	1	0	0	3	2	1
4	35	53.1000	1	1	0	1	2	1
5	35	8.0500	0	0	0	3	2	0
6	28	8.4583	0	0	0	3	1	0

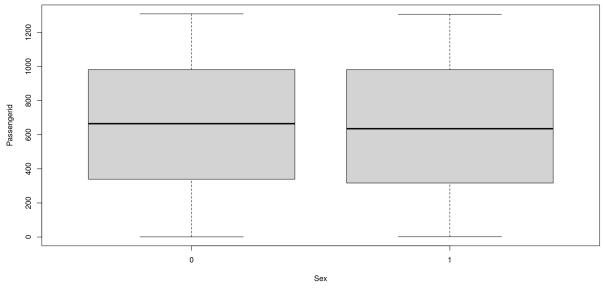
In the first overview of the data, we can observe that is a data frame of nine columns (passenger id, age, fare, sex, sibsp, parch, pclass, embarked and survived) and with 1309 rows, one for each passenger. If we do a global summary for the whole data, we can observe that we can have some important information for the titanic quantitative data like age and fare: [5]

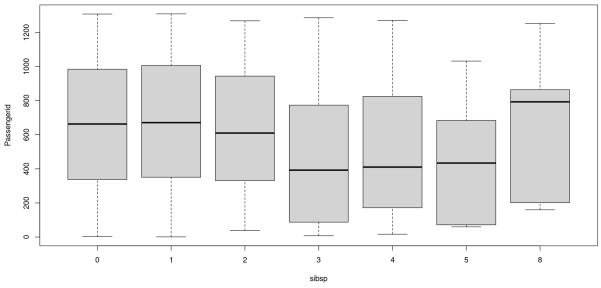
- Age: Min. 0.17, 1st Qu. 22.00, Median. 28.00, Mean. 29.50, 3rd Qu. 35.00 and Max. 80.00
- Fare: Min. 0.00, 1st Qu. 7.89, Median. 14.45, Mean. 33.28, 3rd Qu. 31.27 and Max. 512.33

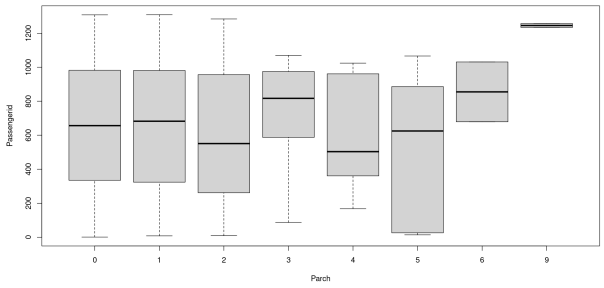
The best way to describe the data set graphically is doing a boxplot for each variable and the passengerID. The results of the boxplots that have been done in R are the following:

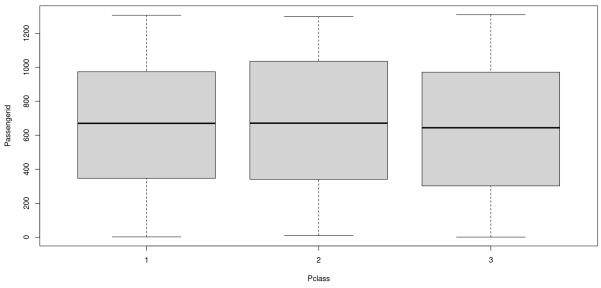


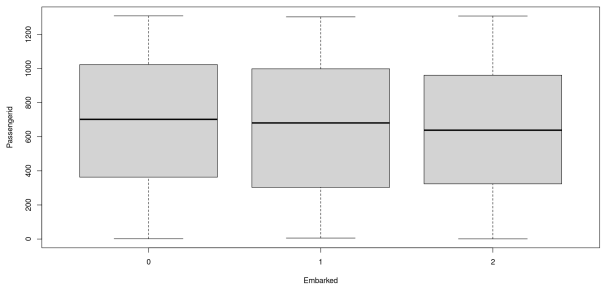


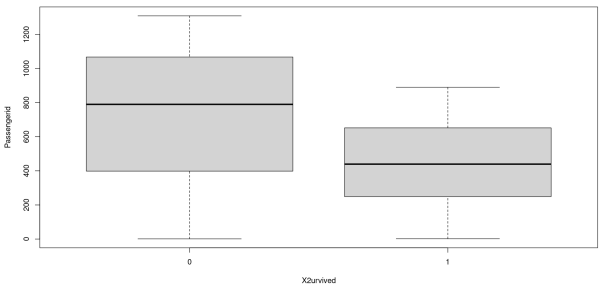












3. Techniques

For this project, to analyze the titanic data set, we decide to use three different techniques, Bayesian inference, Poisson Regression and Linear Regression. Although we know from looking at the Titanic dataset that this is fitting better in the logistic regression model, we decide to use these three different techniques so that we could explore the data from different points of views, with some models that fit better and other models that fit worst with it.

The first model that we use is Bayesian inference, as it is helpful to better understand our data and to formulate more accurate hypothesis. Also, computing the different probabilities and densities is useful to make better predictions and taking uncertainty into account. That is why we decided to compute the different probabilities and densities. Also, we used Bayesian Regression Models using 'Stan' (R package) to see if there was any issue with the data. [2] Lastly, in this part, the posterior densities in order to have a better view of our data and models.

The Poisson regression model, that is defined as "a generalized linear model form of regression analysis used to model count data and contingency tables" on Wikipedia [4]. In this case, to execute this model, we divided in five techniques to study properly our data. The first technique was to analyze the quantitative data, observing the relationship between them and the survived variables in scatter plots and barplots to observe how is this data distributed. The next technique, was doing different Poisson regression models comparing the most interesting variables and the survived variable, to observe if these variables were significant or not. And then, we used this models for the following techniques, finding the effect of the predictor on the average of the response and for looking to the overdispersion with the function dispersiontest from the AER library. Finally, we conclude this study after observing the deviance residuals, computing quasipoisson models for different variables and observing the results in plots.

And finally, the main model of this project is with the logistic regression technique. The logistic model is defined on Wikipedia as "a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model." [3]. To carry out this model, we first observed the data to see if our response variable was binary. Then, we did a logistic regression involving all the variables to select those with the lower P-values, and we compute other techniques to make sure that the explanatory variables selected were really important in our model. Once we knew this, we regressed our response variable in this explanatory ones individually to get the logit functions. Finally, from these functions we calculate π, which represents the proportion of 1's (survive/success) at any X to see which was the probability of surviving ("success") in the Titanic in each one of the cases. Also, we made some plots of the models which can be seen in the Appendix.

4. Results

4.1 Bayesian inference

We will use Bayesian inference as a mathematical tool to calculate probabilities for our hypothesis related to the Titanic dataset. The variables studied are Survival, Sex, Pclass and Fare, as these are the variables that we use in our hypothesis and the ones that will allow us to do better statistical analysis. The decision of not using the variable age came from observing that the data was not completely balanced in this aspect and had so many categories that hardened the analysis.

Prior probabilities and Densities

Exploration of the data:
Prior probabilities:
(Here we are not using Survival-Fare as fare as many possible values)

Survival - Sex (row total, column total and table total) :

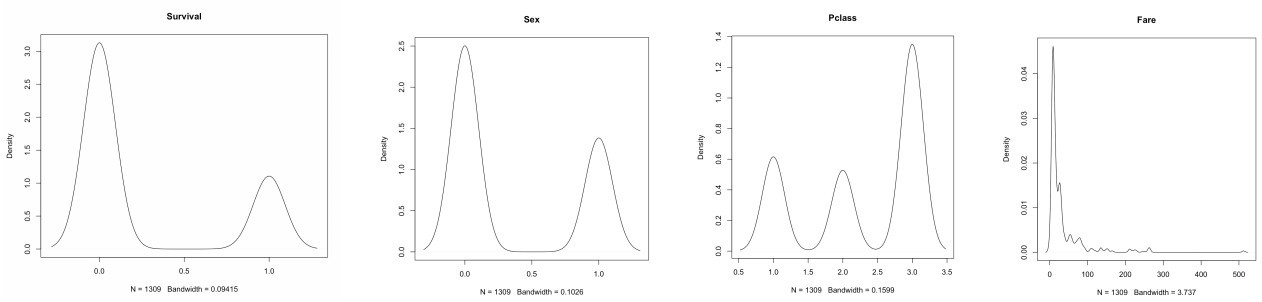
	0	1	Row Total
0	734	233	967
	19.874	35.952	
	0.759	0.241	0.739
	0.871	0.500	
	0.561	0.178	
1	109	233	342
	56.193	101.653	
	0.319	0.681	0.261
	0.129	0.500	
	0.083	0.178	
Column Total	843	466	1309
	0.644	0.356	

Survival - Pclass (row total, column total and table total):

	1	2	3	Row Total
0	187	190	590	967
	11.163	1.046	8.377	
	0.193	0.196	0.610	0.739
	0.579	0.686	0.832	
	0.143	0.145	0.451	
1	136	87	119	342
	31.564	2.957	23.686	
	0.398	0.254	0.348	0.261
	0.421	0.314	0.168	
	0.104	0.066	0.091	
Column Total	323	277	709	1309
	0.247	0.212	0.542	

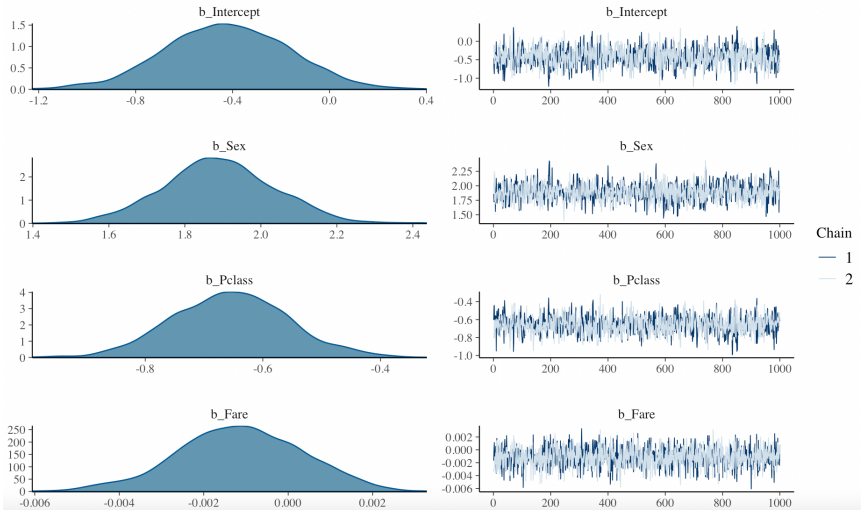
We observe that it is more likely to survive being a female than surviving being a male. Furthermore, the Pclass of each passenger also affected the probability of survival: those passengers that were in first-class, had more probabilities than those in second or third-class.

Densities:



With the densities, we can observe that survived and sex can only take 2 possible (values 0 and 1), while pclass has 3 possible values. Furthermore, fare has many possible values, and we observe a maximum close to 0.

Additionally, we compute the posterior densities and Markov-Chain-Montecarlo, with Bayesian Regression Models using 'Stan':

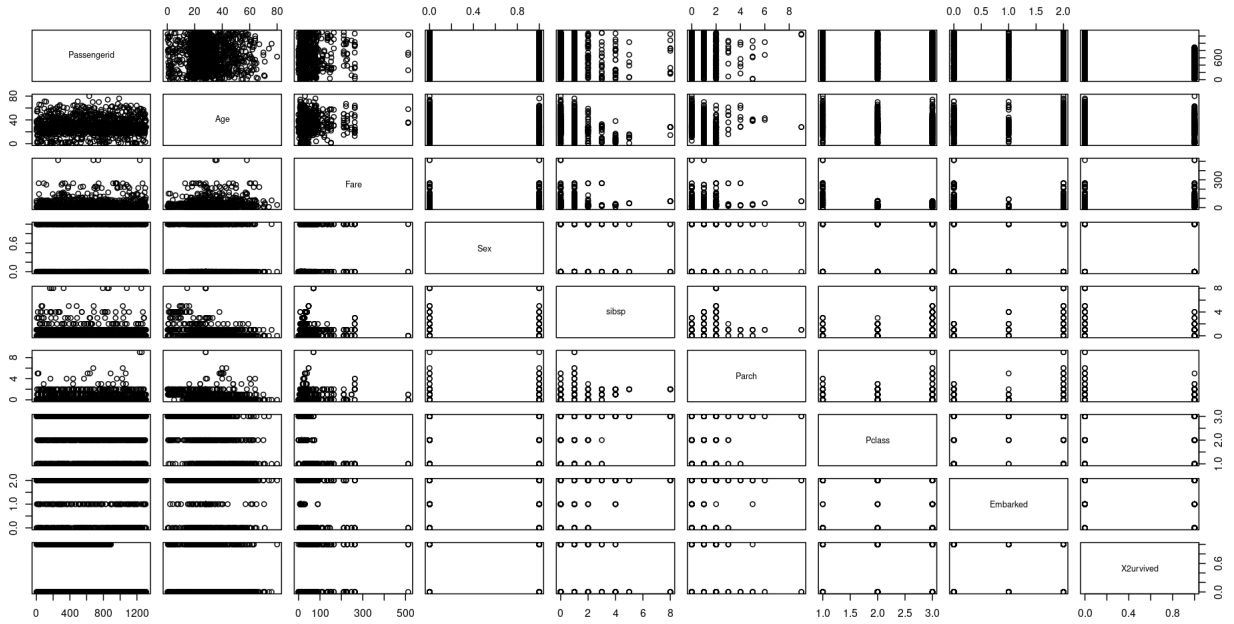



```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS
Intercept    -0.42      0.25    -0.94     0.07 1.00     1223
Sex           1.89      0.15     1.60     2.18 1.00     1106
Pclass       -0.66      0.10    -0.84    -0.46 1.00     1060
Fare         -0.00      0.00    -0.00     0.00 1.00     1897
          Tail_ESS
Intercept     1118
Sex           1031
Pclass        1085
Fare          1851
```

4.2 Poisson regression

Analysis of quantitative data

For starting our study about the Poisson regression, we take a first look at the distribution and show the relationship between the quantitative variables and the survive variable in a scatter plot. The variables that we will use in the Poisson regression will be the same that we used in the Bayesian analysis, age, class, sex, embarked and survive. To visualize if there is a relationship, we start doing scatter plots for each variable, using the function pairs() that returns a matrix of scatterplots. It is in these plots where we can observe that only the variables passengerid, age and fare are returning "interesting" scatterplots.



For this reason, our first "study" in the Poisson regression will be with the quantitative data, the variables Age and Fare, and show the relationship between them and the survived variables in a scatter plot. After doing a barplot for each variable we can observe that the age seems that it follows a more normal distribution (with a "strange" high pick for the people around 29 years old) and the fare can more or less follow a Poisson distribution if we pick the upper limits in ranges, but is not very clear. After this plots, we could show the relationship between survived and age and fair in scatter plots after putting the data in different ranges:

```
Age.cat
(-0.83,16]    (16,21]    (21,24]    (24,28]    (28,32]    (32,39]    (39,48]    (48,81]
       134        156        118        391        122        143        126        119

Fare.cat
  (-1,7.57] (7.57,7.85] (7.85,8.05] (8.05,10.5] (10.5,14.5] (14.5,21.6] (21.6,26.8] (26.8,41.6]
        131         144         146         108         129         127         131         134
```

Poisson regression models

After observing in the scatter plot that the age and the fair variables does not seem to follow a Poisson regression, the next step we did was to do different regression models comparing the number of surviving with sex, class and embarked, that are the variables that we are most interested in. And we also did a Poisson regression model with all predictors, taking out the ones that were less interesting, to have only the significant variables for our "final model".

From these models we get that all of them, sex, class and embarked were significant, being sex the most significant variable and embarked the less one. And when we did the final model, then embarked was not significant, as well as fare and parent-children relations. For each model, we also compute the regression equation and the ANOVA model, to take more conclusions about each variable. Here we can observe the screenshots of the model's results and the conclusions:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.04562    0.09578  -21.36   <2e-16 ***
Sex          1.35247    0.11604   11.65   <2e-16 ***


    Null deviance: 918.07  on 1308  degrees of freedom
Residual deviance: 768.95  on 1307  degrees of freedom
AIC: 1457
```

Significant evidence for association, for a P-value (lower than $\alpha = 0.05$), a positive estimator and a smaller residual deviance (difference larger than 3.84). The regression equation: survived $= -2.04562 + 1.35247 * sex$.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.37037    0.13154  -2.816  0.00487 **
Pclass      -0.45725    0.06149  -7.437 1.03e-13 ***


    Null deviance: 918.07  on 1308  degrees of freedom
Residual deviance: 863.31  on 1307  degrees of freedom
AIC: 1551.3
```

Significant evidence for association, for a P-value (lower than $\alpha = 0.05$), a positive estimator and a smaller residual deviance (difference larger than 3.84). The regression equation: survive $= -0.37037 - 0.45725 * class$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.09063    0.10049 -10.853   <2e-16 ***
Embarked    -0.17893    0.06199  -2.886   0.0039 **


    Null deviance: 915.65  on 1306  degrees of freedom
Residual deviance: 907.68  on 1305  degrees of freedom

AIC: 1591.7
```

Significant evidence for association, for a P-value (lower than $\alpha = 0.05$), a positive estimator and a smaller residual deviance (difference larger than 3.84). The regression equation: survived $= -1.09 - 0.18 * embarked$.

Effect of the predictor

The next step after finding that the three models were significant, specially in the case of sex, we decide to interpret this models by quantifying the effect of the predictor on the average of the response giving a 95% confidence interval for the parameter representing that effect. To find these predictors, we first extract the standard errors and coefficients from the summary of the model, then, we get the estimator of the model and then, from the estimator and the standard error, we calculate the confidence intervals with the function qnorm(0.975). These intervals, can also be exponenciated to take more conclusions. The results for the 3 models are:

- Sex: Confidence interval (1.125030, 1.579914) and interval exponenciated (3.080310, 4.854536)

- Pclass: Confidence interval (-0.57776, -0.33674) and interval exponenciated (0.56115, 0.71409)

- Embarked: Confidence interval (-0.3004, -0.0574) and interval exponenciated (0.7405, 0.9441)

We can observe that without exponenciating the intervals, only the variable sex has a positive confidence interval, whereas the variables pclass and embarked are located between minus one and zero. When we exponenciate these intervals, we get that for the sex variable it gets an interval with a big difference between it and the other two models are now located around one, but one is not included on the interval. We can also observe that the confidence interval is more narrow, in this two intervals, because there is a difference of 0.2 approximately.

Overdispersion and T-test

Another interesting point when we are doing a test is looking if there is any indication that overdispersion is a problem for our models. To observe if there is any kind of overdispersion, this can be estimated as the division of the residual deviance from the model and the degrees of freedom. But there is a package in R named Applied Econometrics with R, AER, that has a function named dispersiontest that tests the null hypothesis of equidispersion in Poisson GLMs against the alternative of overdispersion and/or underdispersion. Knowing that if the value is larger than one, indicates that we might have overdispersion and that this might happen because of data heterogeneity (fluctuating covariates, variables are not constant) or correlation between observations (variables are not really independent). The results are the following:

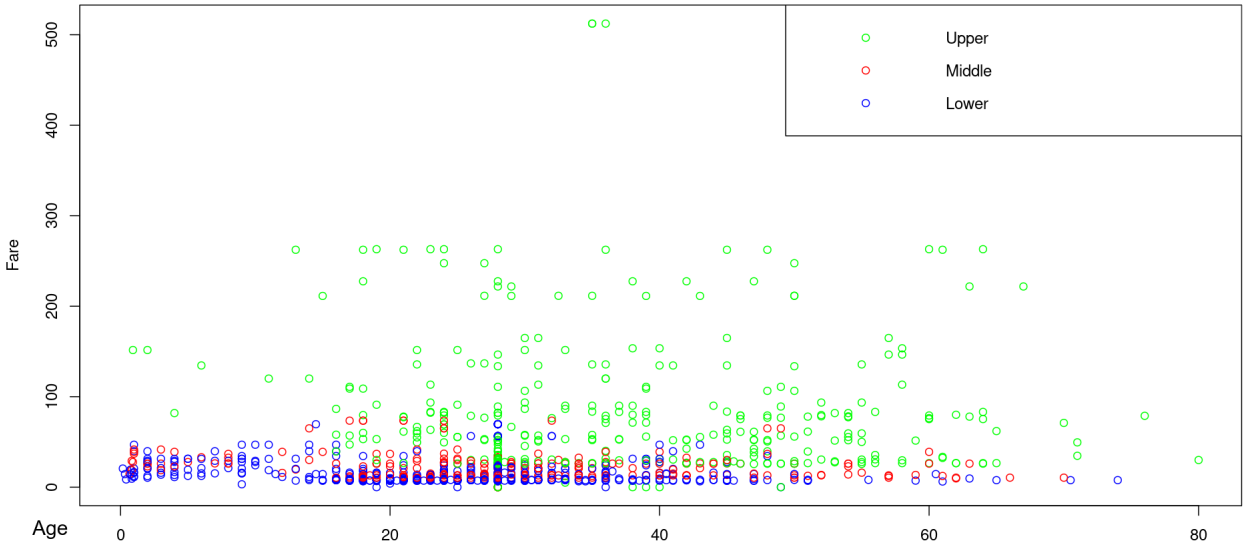
- First model: z = -11.48, p-value = 1, dispersion: 0.7387319 – there is no overdispersion

- Second model: z = -11, p-value = 1, dispersion: 0.7387319 – there is no overdispersion

- Third model: z = -10.75, p-value = 1, dispersion: 0.7398623 – there is no overdispersion

Deviance residuals

Finally, we decide to calculate the deviance residuals according to the different variables, sex, pclass, embarked and survived and plot these as a function of the predicted values, using a different color for each variable classification. To observe the results, we first do a barplot differentiating the variables and then a general plot with fare and age as the axis variables and the selected variables with different colors, explained in the legend. This is one of the plots:



Conclusions

After doing this study, we can conclude that for our dataset, the Poisson regression model is not fitting quite properly. The main problem is that most of our data is binary, except for the age and the fair. Despite that the models gave us significant P-values and that this models did not have overdispersion, the way our data is distributed, shows us that Poisson regression might not be the best model to test our data.

If we observe to our data, from eight variables, two of them are not binary and the other six are binary. Taking into account that the Poisson regression model is a statistical method for the modeling of count data, where the response is assumed to follow a Poisson distribution, and that the Logistic regression model is a particular case of a generalized linear model, for binary response variables with a Bernoulli distribution, we can conclude that maybe the Poisson regression model is not the good one for our data. It is for that reason that the next model that we decided to analyze is the Logistic regression model, that knowing that is specific for binary response variables, it might be the right model to analyze our data.

4.3 Logistic regression

Analysis of the data

To start our Logistic regression study, first, we will see if our data fit a logistic regression model. Then, we will use logistic regression to find out which explanatory variables are more significant in the survival of the Titanic.
For this part of the project, we will need the packages pscl and caret.

To get started, we observe the data:

Age <dbl>	Fare <dbl>	Sex <int>	sibsp <int>	Parch <int>	Pclass <int>	Embarked <int>	X2urvived <int>
22	7.2500	0	1	0	3	2	0
38	71.2833	1	1	0	1	0	1
26	7.9250	1	0	0	3	2	1
35	53.1000	1	1	0	1	2	1
35	8.0500	0	0	0	3	2	0
28	8.4583	0	0	0	3	1	0

6 rows | 3-10 of 9 columns

As we can see, our response variable, survive the Titanic (X2urvived), is binary, which is perfect when doing a logistic model.

Now, we do a general logistic regression including all our explanatory variables to see which of those variables explain better our response variable (survival on the Titanic).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.2061733  0.4770709   6.721 1.81e-11 ***
Passengerid -0.0030985  0.0002478 -12.506  < 2e-16 ***
Age         -0.0343026  0.0068425  -5.013 5.35e-07 ***
Fare        -0.0002592  0.0017834  -0.145  0.88445
Sex          2.2640321  0.1750012  12.937  < 2e-16 ***
sibsp       -0.3132645  0.1010252  -3.101  0.00193 **
Parch        0.0122141  0.1032653   0.118  0.90585
Pclass      -0.9859739  0.1229424  -8.020 1.06e-15 ***
Embarked    -0.1162352  0.0983299  -1.182  0.23717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1498.3  on 1306  degrees of freedom
Residual deviance: 1001.4  on 1298  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 1019.4
```

As we can see, the explanatory variables with a better significance level (P-value lower than $\alpha = 0.05$) are Age, Sex and Pclass (we do not count Passengerid because it is not relevant).

Also, we get that:
$ln(\pi/(1-\pi)) = 3.2061733 - 0.0343026 * age + 2.2640321 * sex - 0.9859739 * Pclass$

To continue, and just to get more information about our explanatory variables, we observe the importance of each predictor in the model using the function varImp() from the package caret. [1]

Results of our variables of interest: Age = 5.0131849, Sex = 12.9372385, Pclass = 8.0198030.

Higher values indicate more importance, so the results obtained match up very well with the p-values from our model. Being Age, Sex and Pclass good predictors. This results can be observed in the Appendix in the Logistic Regression section.

Next, we are going to calculate McFadden's R^2 to assess how well our model fits the data. The value ranges from 0 to just under 1, with higher values indicating better model fit.
McFadden's $R^2 = 0.3316875$
This means that the logistic model fits our data pretty well.

Logistic regression models

Secondly, we will regress our response variable in these explanatory ones (Age, Sex and Pclass).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.746635   0.156381  -4.774  1.8e-06 ***
Age         -0.010059   0.004986  -2.017   0.0437 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1503.7  on 1308  degrees of freedom
Residual deviance: 1499.6  on 1307  degrees of freedom
AIC: 1503.6
```

Age is a significant predictor (P-value lower than $\alpha = 0.05$).
Logistic regression equation: $ln(\pi/(1-\pi)) = -0.746635 - 0.010059 * age$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9072     0.1026  -18.58   <2e-16 ***
Sex           1.9072     0.1383   13.79   <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: 1503.7  on 1308  degrees of freedom
Residual deviance: 1295.2  on 1307  degrees of freedom
AIC: 1299.2
```

Sex is a significant predictor (P-value lower than $\alpha = 0.05$).
Logistic regression equation: $ln(\pi/(1-\pi)) = -1.9072 + 1.9072 * sex$

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.37750    0.16950   2.227   0.0259 *
Pclass      -0.64652    0.07478  -8.645   <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: 1503.7  on 1308  degrees of freedom
Residual deviance: 1427.4  on 1307  degrees of freedom
AIC: 1431.4
```

Pclass is a significant predictor (P-value lower than $\alpha = 0.05$).
Logistic regression equation: $ln(\pi/(1-\pi)) = 0.37750 - 0.64652 * Pclass$

Calculating probabilities

To finish, we will use the logit function that we get in each one of the models before to determine π = proportion of 1's (survive/success) at any X.

SURVIVAL & AGE
Logistic regression equation: $ln(\pi/(1-\pi)) = -0.746635 - 0.010059 * age$
How the age can vary very widely, we are not going to compute the probability of surviving in each one of the possible ages.

SURVIVAL & GENDER
Logistic regression equation: $ln(\pi/(1-\pi)) = -1.9072 + 1.9072 * sex$
Remember that in our data women = 1, men = 0

For women:
$ln(\pi/(1-\pi)) = -1.9072+1.9072*1 = 0 \rightarrow e^{ln(\pi/(1-\pi))} = e^0 \rightarrow \pi/(1-\pi) = e^0 \rightarrow \pi/(1-\pi) = 1 \rightarrow \pi = 1(1-\pi) \rightarrow \pi = 1-\pi \rightarrow 2\pi = 1 \rightarrow \pi = 1/2 \rightarrow \pi = 0.5$

The probability of surviving the Titanic being a woman was 50%.

For men:
$ln(\pi/(1-\pi)) = -1.9072+1.9072*0 = -1.9072 \rightarrow e^{ln(\pi/(1-\pi))} = e^{-1.9072} \rightarrow \pi/(1-\pi) = e^{-1.9072} \rightarrow \pi/(1-\pi) = 0.1484956 \rightarrow \pi = 0.1484956(1-\pi) \rightarrow \pi = 0.1484956 - 0.1484956\pi \rightarrow 1.148496\pi = 0.1484956 \rightarrow \pi = 0.1484956/1.148496 \rightarrow \pi = 0.1292957$

The probability of surviving the Titanic being a man was 12.93%. Let's say 13%.

SURVIVAL & CLASS
Logistic regression equation: $ln(\pi/(1-\pi)) = 0.37750 - 0.64652 * Pclass$
Remember that in our data Upper = 1, Middle = 2 and Lower = 3

For Lower class:
$ln(\pi/(1-\pi)) = 0.37750 - 0.64652 * 3 = -1.56206 \rightarrow e^{ln(\pi/(1-\pi))} = e^{-1.56206} \rightarrow \pi/(1-\pi) = e^{-1.56206} \rightarrow \pi/(1-\pi) = 0.2097036 \rightarrow \pi = 0.2097036(1-\pi) \rightarrow \pi = 0.2097036 - 0.2097036\pi \rightarrow 1.209704\pi = 0.2097036 \rightarrow \pi = 0.2097036/1.209704 \rightarrow \pi = 0.1733512$

The probability of surviving the Titanic, traveling in lower class was 17.34%.

For Middle class:
$ln(\pi/(1-\pi)) = 0.37750 - 0.64652 * 2 = -0.91554 \rightarrow e^{ln(\pi/(1-\pi))} = e^{-0.91554} \rightarrow \pi/(1-\pi) = e^{-0.91554} \rightarrow \pi/(1-\pi) = 0.4003004 \rightarrow \pi = 0.4003004(1-\pi) \rightarrow \pi = 0.4003004 - 0.4003004\pi \rightarrow 1.4003\pi = 0.4003004 \rightarrow \pi = 0.4003004/1.4003 \rightarrow \pi = 0.2858676$

The probability of surviving the Titanic, traveling in middle class was 28.59%.

For Upper class:
$ln(\pi/(1-\pi)) = 0.37750 - 0.64652 * 1 = -0.26902 \rightarrow e^{ln(\pi/(1-\pi))} = e^{-0.26902} \rightarrow \pi/(1-\pi) = e^{-0.26902} \rightarrow \pi/(1-\pi) = 0.764128 \rightarrow \pi = 0.764128(1-\pi) \rightarrow \pi = 0.764128 - 0.764128\pi \rightarrow 1.764128\pi = 0.764128 \rightarrow \pi = 0.764128/1.764128 \rightarrow \pi = 0.4331477$

The probability of surviving the Titanic, traveling in upper class was 43.31%.

Conclusions

After doing this study on our data in the logistic regression technique, we can conclude that this model fits quite well our data thanks to the binary format of our response variable and the results that we have obtained. As the Poisson regression model, we also get significant P-values for different variables but in this case, the data was fitting better to this model. In general, we can observe that women have higher probability of surviving rather than men and that the probability of surviving is higher if you increase the class in which you were traveling.

5. Conclusions

After analyzing the data set with the different techniques, we can arrive to the conclusion that our main hypothesis was true, i.e., accepting the alternative hypothesis. This hypothesis was that our analysis was fitting the logistic regression, showing a relationship between different variables, that the probability to survive in the Titanic was related with the age, the class, and the sex. This conclusion was obtained after observing the results of the different techniques:

Bayesian analysis allowed a flexible approach to understand our dataset and give a starting point for the other techniques used, furthermore, it allowed the group to compute probabilities and densities taking into account uncertainty. When computing the probabilities, we observed that the probability of surviving being a woman was higher than being a man. Also, the class where passengers travelled did affect its probability of surviving: those who did in first class survived in a greater percentage than does who were in second or third. However, the most remarkable difference is between gender.

For the Poisson regression analysis, we could observe that most of the tests gave us significant P-values for the selected variables, but the way our data is distributed, shows us that the Poisson regression might not be the best model to test our data. As we could observe from the beginning, most of our variables are binary and the Poisson regression is a statistical method for the modeling of count data, where the response is assumed to follow a Poisson distribution. For this reason, we can conclude from this section that the Logistic regression model is the technique that fits better our data because it is also a particular case of a generalized linear model, but for binary response variables with a Bernoulli distribution.

After doing the Logistic regression, we can arrive to the conclusion that this is the model where the data fits very well and this happens because of the binary form that our response variable has. Also, we can support our alternative hypothesis explained before, because we have seen that the explanatory variables that we have studied (age, sex, and class) are the variables that are most related with the probabilities of surviving in the Titanic.

Finally, from our observations and computations, we can say that there was an evident relation between age, sex, and class when it comes to survival in the Titanic's tragedy; showing that the group with more probabilities of surviving were women travelling in first-class. Furthermore, it is also possible to conclude that women, in general, had more chances of surviving. An interesting aspect of this dataset is observing the quantity of passengers that were travelling without travelling (Fare = 0). That makes us consider that it is possible that the dataset lacks information, or even that there is a general lack of information about those passengers.

Bibliography

[1] The caret Package. Max kuhn. https://topepo.github.io/caret/index.html
.

[2] Crosstable(). Bayesian analysis. https://medium.com/epidence-en-espa%C3%B1ol/obtener-proporciones-en-r-no-deber%C3%ADa-ser-un-dolor-de-cabeza-24a2b26ab745
.

[3] Free encyclopedia. Logistic regression. https://en.wikipedia.org/wiki/Logistic_regression
.

[4] Free encyclopedia. Poisson regression. https://en.wikipedia.org/wiki/Poisson_regression
.

[5] Khashayar Baghizadeh Hosseini. Titanic data set suited for binary logistic regression, 2017. https://www.kaggle.com/datasets/heptapod/titanic
.

[6] Mònica Torner Marta Ortigas and Núria Mitjavila. The scripts done for the analysis project. https://github.com/NuriaMitjavila/StatisticProject
.

[7] Machine Learning from Disaster Titanic Competitions. Original data of our data set. https://www.kaggle.com/competitions/titanic/rules
.

[8] Çağlar Uslu. What is kaggle? https://www.datacamp.com/blog/what-is-kaggle
.

Appendix, R Scripts

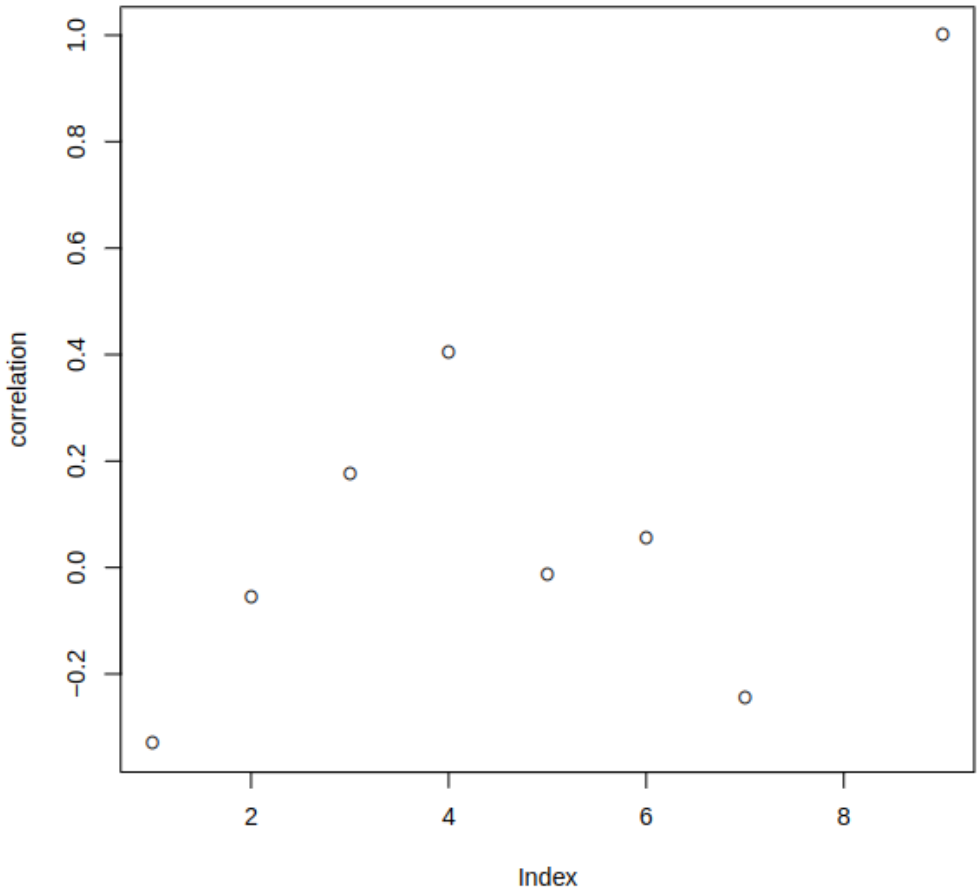
First Analysis

```
# We take a first look to our data
titanic <- read.csv("titanic.csv")
class(titanic)

## [1] "data.frame"

dim(titanic)

## [1] 1309    9

head(titanic)

##   Passengerid Age    Fare Sex sibsp Parch Pclass Embarked X2urvived
## 1           1  22  7.2500   0     1     0      3        2         0
## 2           2  38 71.2833   1     1     0      1        0         1
## 3           3  26  7.9250   1     0     0      3        2         1
## 4           4  35 53.1000   1     1     0      1        2         1
## 5           5  35  8.0500   0     0     0      3        2         0
## 6           6  28  8.4583   0     0     0      3        1         0

summary(titanic$Age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.17   22.00   28.00   29.50   35.00   80.00

summary(titanic$Fare)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   7.896  14.454  33.281  31.275 512.329

# boxplot(Passengerid~Age)
# boxplot(Passengerid~Fare)
# boxplot(Passengerid~Sex)
# boxplot(Passengerid~sibsp)
# boxplot(Passengerid~Parch)
# boxplot(Passengerid~Pclass)
# boxplot(Passengerid~Embarked)
# boxplot(Passengerid~X2urvived)
```

Bayesian Analysis

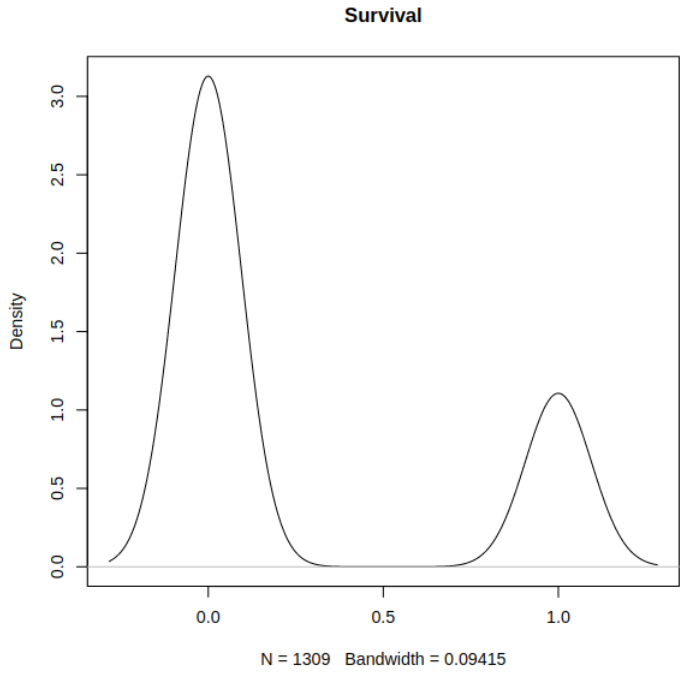
```
table(titanic$Pclass)

##
##   1   2   3
## 323 277 709

table(titanic$Sex)

##
##   0   1
## 843 466

table(titanic$Embarked)

##
##   0   1   2
## 270 123 914

table(titanic$X2urvived)

##
##   0   1
## 967 342

# We take a first look to our data
correlation = cor(titanic, titanic$X2urvived)
correlation

##                    [,1]
## Passengerid -0.33149303
## Age         -0.05586165
## Fare         0.17378625
## Sex          0.40402004
## sibsp       -0.01437545
## Parch        0.05490813
## Pclass      -0.24468559
## Embarked             NA
## X2urvived    1.00000000
```

```
# plot(correlation)
```

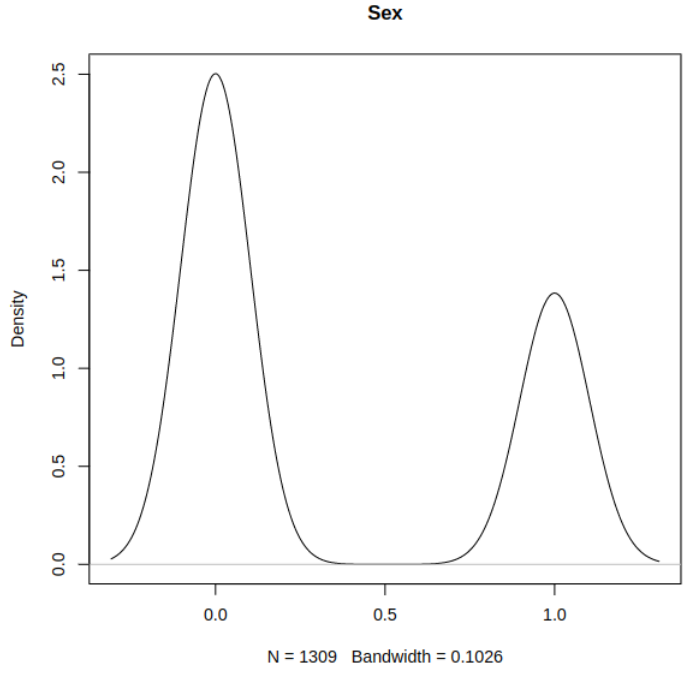



```
variables <- c("X2urvived","Pclass", "Sex", "Fare")

newdataset = titanic[variables]

head(newdataset)

##   X2urvived Pclass Sex    Fare
## 1         0      3   0  7.2500
## 2         1      1   1 71.2833
## 3         1      3   1  7.9250
## 4         1      1   1 53.1000
## 5         0      3   0  8.0500
## 6         0      3   0  8.4583

densitysurvived <- density(newdataset$X2urvived)
```

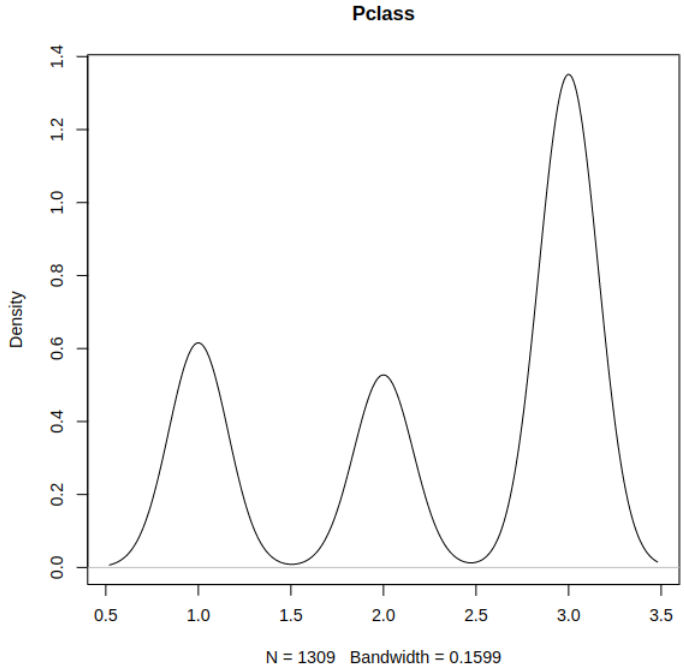
```
# plot(densitysurvived, main='Survival')
```



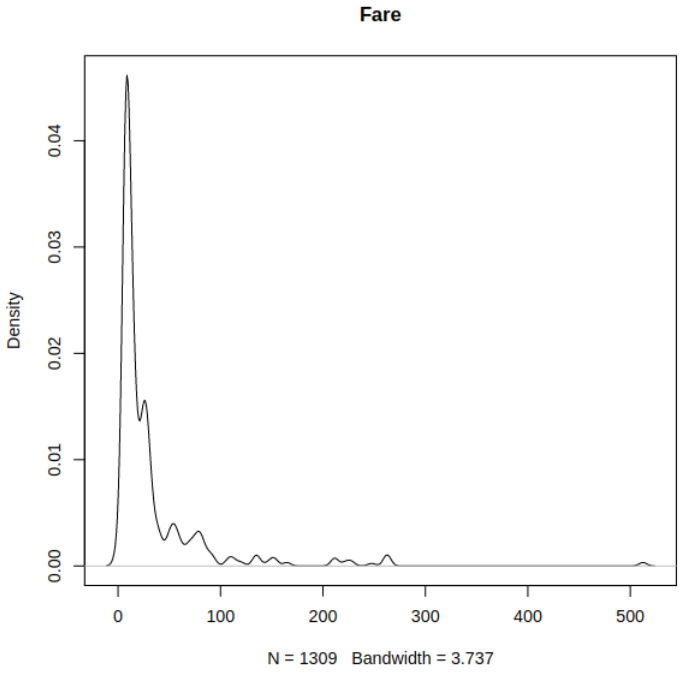

```
densitysex <- density(newdataset$Sex)
# plot(densitysex, main='Sex' )
```



```r
densitypclass <- density(newdataset$Pclass)
# plot(densitypclass, main='Pclass' )
```




```r
densityfare <- density(newdataset$Fare)
# plot(densityfare, main='Fare' )
```



```r
# library(brms)
library(gmodels)
# require(brms)

tabsex <- table(newdataset$X2urvived,newdataset$Sex)
CrossTable(tabsex) #to have row total, column total and table total
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  1309
##
##
##              |
##              |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       734 |       233 |       967 |
##              |    19.874 |    35.952 |           |
##              |     0.759 |     0.241 |     0.739 |
##              |     0.871 |     0.500 |           |
##              |     0.561 |     0.178 |           |
## -------------|-----------|-----------|-----------|
##            1 |       109 |       233 |       342 |
##              |    56.193 |   101.653 |           |
##              |     0.319 |     0.681 |     0.261 |
##              |     0.129 |     0.500 |           |
##              |     0.083 |     0.178 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       843 |       466 |      1309 |
##              |     0.644 |     0.356 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(tabsex, prop.chisq=FALSE, prop.r = FALSE, prop.t = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1309
##
##
##              |
##              |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |       734 |       233 |       967 |
##              |     0.871 |     0.500 |           |
## -------------|-----------|-----------|-----------|
##            1 |       109 |       233 |       342 |
##              |     0.129 |     0.500 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       843 |       466 |      1309 |
##              |     0.644 |     0.356 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
tabpclass <- table(newdataset$X2urvived,newdataset$Pclass)
CrossTable(tabpclass)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
```

```
##
## Total Observations in Table:  1309
##
##
##              |
##              |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            0 |       187 |       190 |       590 |       967 |
##              |    11.163 |     1.046 |     8.377 |           |
##              |     0.193 |     0.196 |     0.610 |     0.739 |
##              |     0.579 |     0.686 |     0.832 |           |
##              |     0.143 |     0.145 |     0.451 |           |
## -------------|-----------|-----------|-----------|-----------|
##            1 |       136 |        87 |       119 |       342 |
##              |    31.564 |     2.957 |    23.686 |           |
##              |     0.398 |     0.254 |     0.348 |     0.261 |
##              |     0.421 |     0.314 |     0.168 |           |
##              |     0.104 |     0.066 |     0.091 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       323 |       277 |       709 |      1309 |
##              |     0.247 |     0.212 |     0.542 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
CrossTable(tabpclass, prop.chisq=FALSE, prop.r = FALSE, prop.t = FALSE)
```

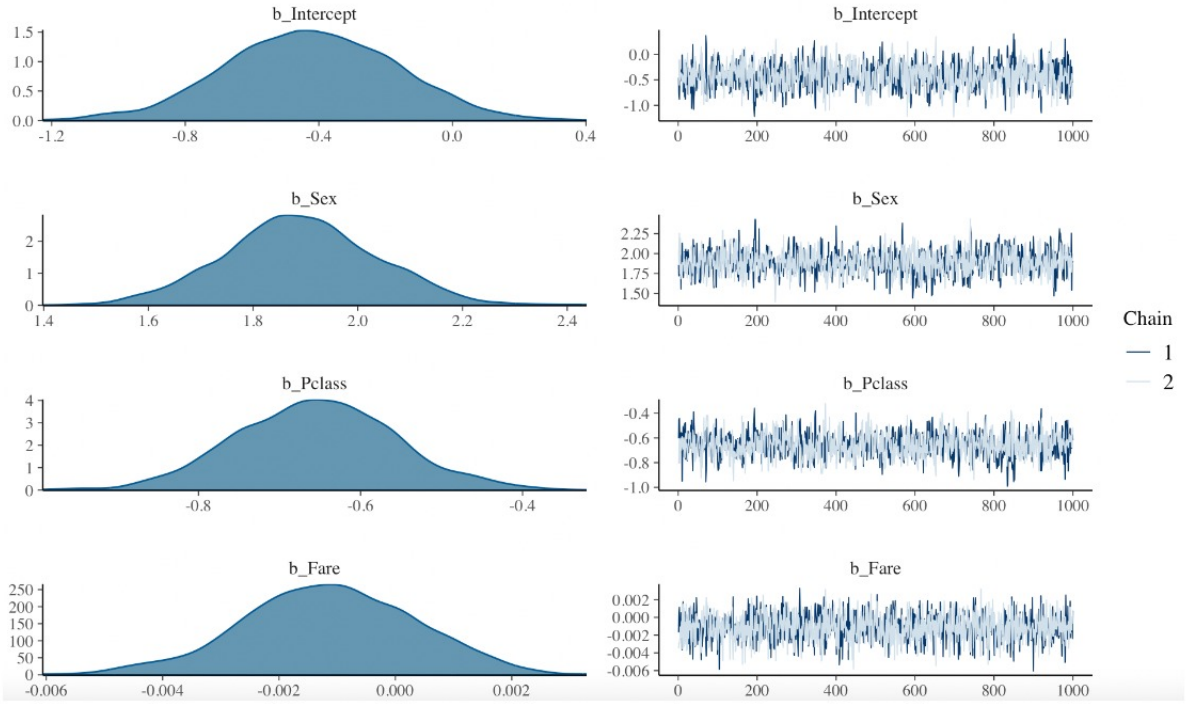
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1309
##
##
##              |
##              |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##            0 |       187 |       190 |       590 |       967 |
```

```
##               |     0.579 |     0.686 |     0.832 |           |
## -------------|-----------|-----------|-----------|-----------|
##            1 |       136 |        87 |       119 |       342 |
##              |     0.421 |     0.314 |     0.168 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |       323 |       277 |       709 |      1309 |
##              |     0.247 |     0.212 |     0.542 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
tabfare <- table(newdataset$X2urvived,newdataset$Fare)
# CrossTable(tabfare)
CrossTable(tabfare, prop.chisq=FALSE, prop.r = FALSE, prop.t = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1309
##
##
##              |
##              |         0 |    3.1708 |    4.0125 |         5 |    6.2375 |    6.4375 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##            0 |        16 |         1 |         1 |         1 |         1 |         3 |
##              |     0.941 |     1.000 |     1.000 |     1.000 |     1.000 |     1.000 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##            1 |         1 |         0 |         0 |         0 |         0 |         0 |
##              |     0.059 |     0.000 |     0.000 |     0.000 |     0.000 |     0.000 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
## Column Total |        17 |         1 |         1 |         1 |         1 |         3 |
##              |     0.013 |     0.001 |     0.001 |     0.001 |     0.001 |     0.002 |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##
##
```

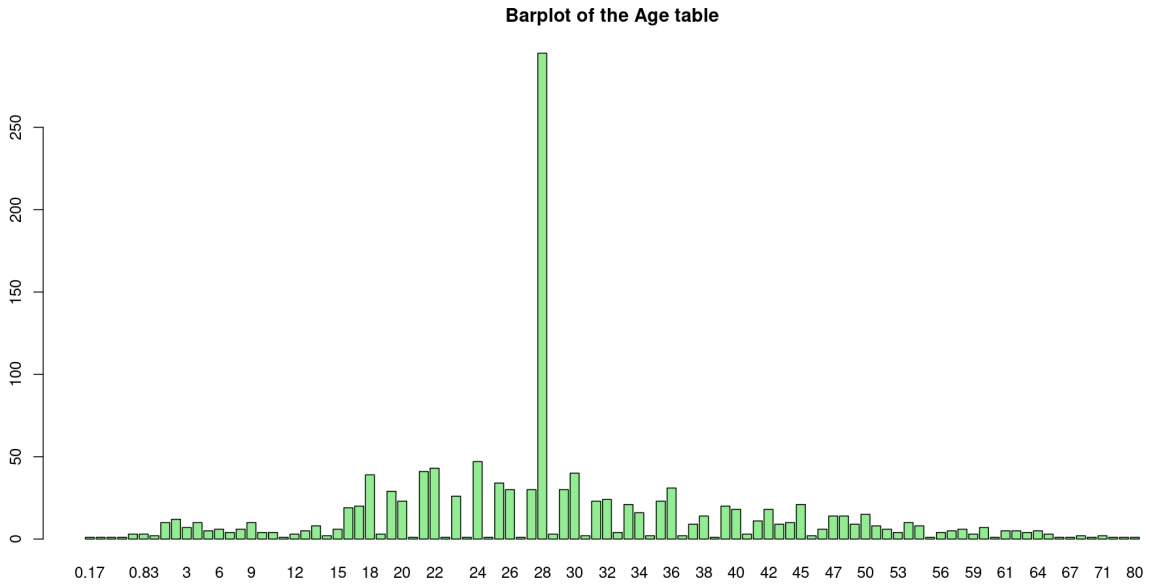
```
# fited <- brm(X2urvived ~ Sex + Pclass + Fare, family = bernoulli(),
#     data=newdataset, chains = 2, iter = 2000, refresh = 0)
# plot(fited)
```



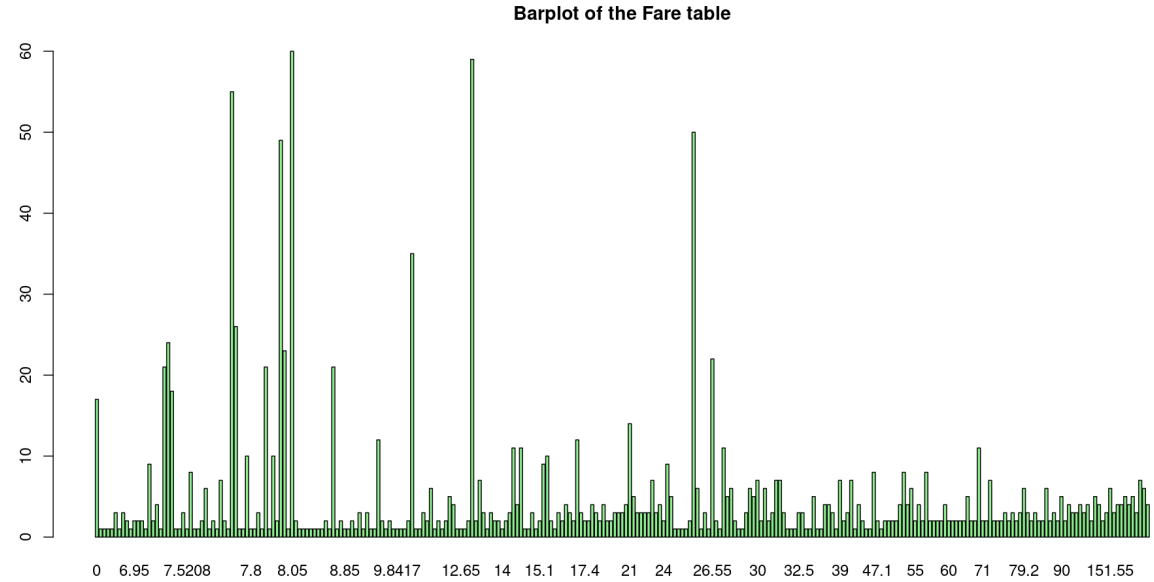

```
# Family: bernoulli
#   Links: mu = logit
#   Formula: X2urvived ~ Sex + Pclass + Fare
#   Data: newdataset (Number of observations: 1309)
#   Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
#          total post-warmup draws = 2000

# Population-Level Effects:
#           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
# Intercept    -0.42      0.25    -0.94     0.07 1.00     1133     1169
# Sex           1.90      0.14     1.63     2.18 1.00     1081     1265
# Pclass       -0.66      0.10    -0.85    -0.46 1.00     1170     1255
# Fare         -0.00      0.00    -0.00     0.00 1.00     1877     1727

# Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
# and Tail_ESS are effective sample size measures, and Rhat is the potential
# scale reduction factor on split chains (at convergence, Rhat = 1).
```

Poisson Regression

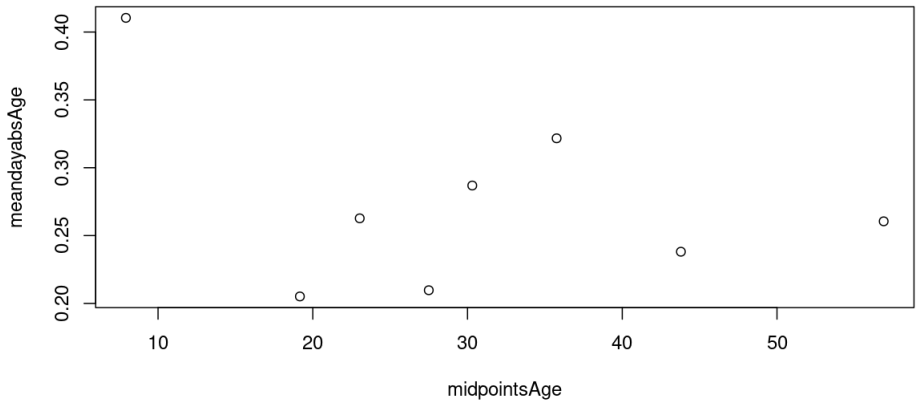
```
attach(titanic)
pairs(titanic)
barplot(table(Age), main="Barplot of the Age table", col="lightgreen")
```



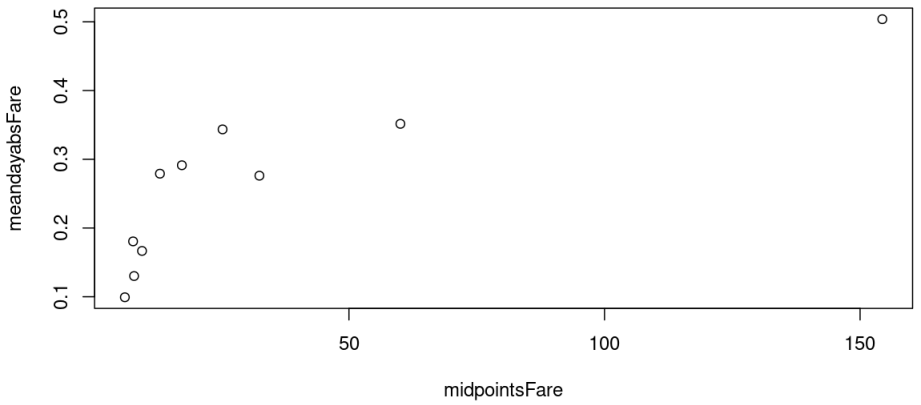

```
barplot(table(Fare), main="Barplot of the Fare table", col="lightgreen")
```



```r
# Show the relationship between survived and age in a scatter plot.
quAge <- unique(quantile(Age,probs=seq(0.1,0.9,by=0.1)))
miAge <- min(Age)
maAge <- max(Age)
breakpoints <- c(miAge-1, quAge, maAge+1)
Age.cat <- cut(Age, breakpoints)
midpointsAge <- tapply(Age, Age.cat, mean)
meandayabsAge <- tapply(X2urvived, Age.cat, mean)
plot(midpointsAge, meandayabsAge)
```




```r
# Show the relationship between survived and fare in a scatter plot.
quFare <- quantile(Fare,probs=seq(0.1,0.9,by=0.1))
miFare <- min(Fare)
maFare <- max(Fare)
breakpoints <- c(miFare-1, quFare, maFare+1)
Fare.cat <- cut(Fare, breakpoints)
midpointsFare <- tapply(Fare, Fare.cat, mean)
meandayabsFare <- tapply(X2urvived, Fare.cat, mean)
plot(midpointsFare, meandayabsFare)
```



```
# Poisson regression of the number of survived on sex
model1 <- glm(X2urvived~Sex,family=poisson(link="log"))
summary(model1)
eq1 <- -2.04562 + 1.35247*Sex
anova(model1)

# Poisson regression of the number of survived on class
model2 <- glm(X2urvived~Pclass,family=poisson(link="log"))
summary(model2)
eq2 <- -0.37037 - 0.45725*Pclass
anova(model2)

# Poisson regression of the number of survived on embarked
model3 <- glm(X2urvived~Embarked,family=poisson(link="log"))
summary(model3)
eq3 <- -1.09063 - 0.17893*Embarked
anova(model3)

# Poisson regression with all predictors.
detach(titanic)
final.model <- glm(X2urvived~., family=poisson(link="log"), data=titanic)
summary(final.model) # all the variables
final.model <- glm(X2urvived~., family=quasipoisson(link="log"), data=titanic)
summary(final.model) # overdispersion tried to be solved with quasipoisson
final.model <- glm(X2urvived~.-Parch-Fare-Embarked,
                   family=quasipoisson(link="log"), data=titanic)
summary(final.model) # we left out the none significant variables
attach(titanic)
eq4 <- 0.450957 - 0.0016257 * Passengerid - 0.0155141 * Age + 1.2324323 *
        Sex - 0.1558248 * sibsp - 0.48229 * Pclass
anova(final.model)
'
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4509570  0.2060773   2.188  0.02883 *
Passengerid -0.0016257  0.0001286 -12.637  < 2e-16 ***
Age         -0.0155141  0.0036310  -4.273 2.07e-05 ***
Sex          1.2324323  0.0941085  13.096  < 2e-16 ***
sibsp       -0.1558248  0.0534777  -2.914  0.00363 **
Pclass      -0.4822900  0.0529614  -9.106  < 2e-16 ***
```

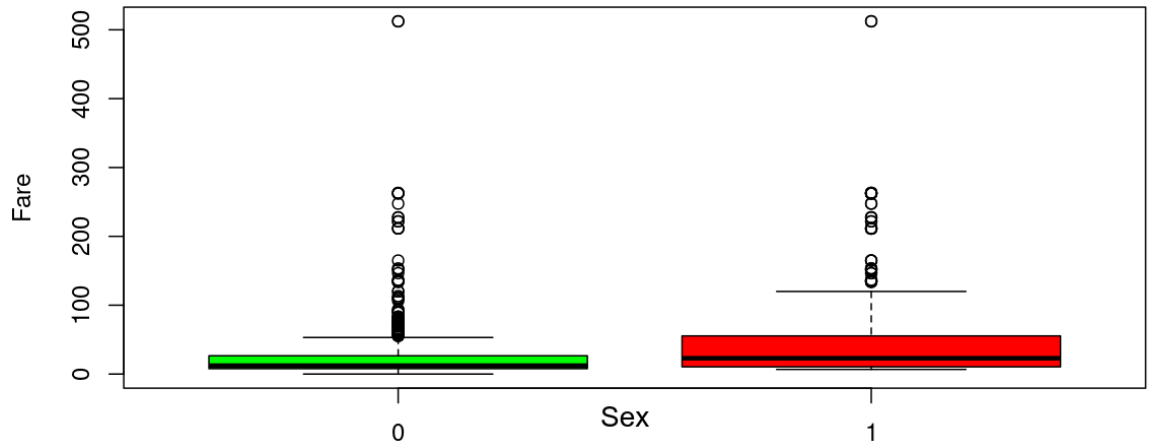
This are the confidence intervals obtained from the previous models, computed in this page:

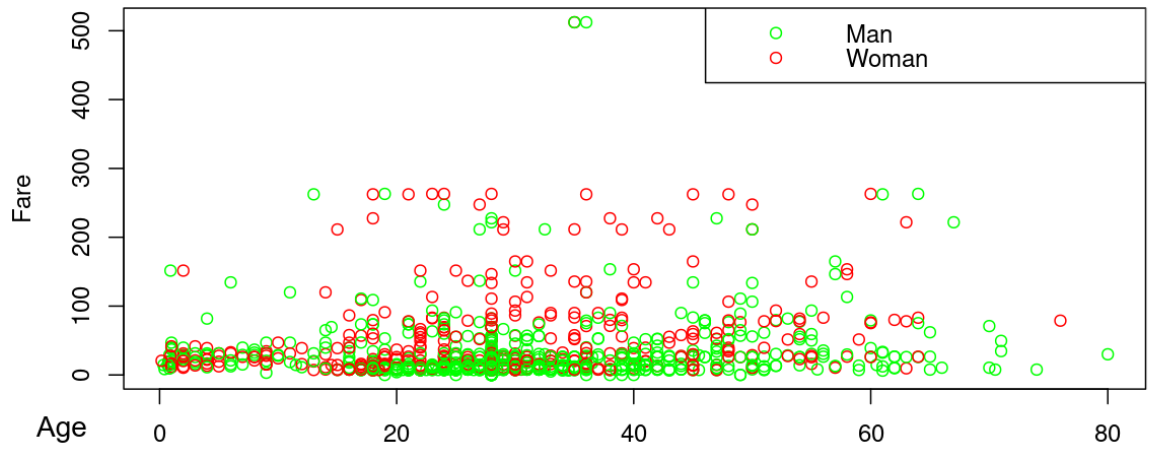
```
      Sex       Sex    Pclass    Pclass  Embarked  Embarked Passengerid Passengerid
3.080310 4.854536 0.5611516 0.7140920 0.7405033 0.9441839   0.9981239   0.9986274
```

```
# Interpret the first model by quantifying the effect of the predictor
M1 <- summary(model1)$coefficients # standard errors and coefficients
se1 <- M1[,2] # get the standard error of the first model
b1 <- coef(model1) # get the estimate of the first model
llb1 <- b1 - qnorm(0.975)*se1 # confidence intervals of our model
ulb1 <- b1 + qnorm(0.975)*se1 # confidence intervals of our model
c(llb1[2], ulb1[2])
c(exp(llb1[2]), exp(ulb1[2])) # Value 0 or 1 inside the interval?

# Interpret the second model by quantifying the effect of the predictor
M2 <- summary(model2)$coefficients # standard errors and coefficients
se2 <- M2[,2] # get the standard error of the second model
b2 <- coef(model2) # get the estimate of the second model
llb2 <- b2 - qnorm(0.975)*se2 # confidence intervals of our model
ulb2 <- b2 + qnorm(0.975)*se2 # confidence intervals of our model
c(llb2[2], ulb2[2])
c(exp(llb2[2]), exp(ulb2[2])) # Value 0 or 1 inside the interval?

# Interpret the third model by quantifying the effect of the predictor
M3 <- summary(model3)$coefficients # standard errors and coefficients
se3 <- M3[,2] # get the standard error of the third model
b3 <- coef(model3) # get the estimate of the third model
llb3 <- b3 - qnorm(0.975)*se3 # confidence intervals of our model
ulb3 <- b3 + qnorm(0.975)*se3 # confidence intervals of our model
c(llb3[2], ulb3[2])
c(exp(llb3[2]), exp(ulb3[2])) # Value 0 or 1 inside the interval?

# Quantify the effect of the different predictors on the response, 95% CI
Mfinal <- summary(final.model)$coefficients # standard error and coefficient
sefinal <- Mfinal[,2] # get the standard error of the final model
bfinal <- coef(final.model) # get the estimate of the final model
llbfinal <- bfinal - qnorm(0.975)*sefinal # confidence intervals of our model
ulbfinal <- bfinal + qnorm(0.975)*sefinal # confidence intervals of our model
c(llbfinal[2], ulbfinal[2])
c(exp(llbfinal[2]), exp(ulbfinal[2])) # Value 0 or 1 inside the interval?
'
```

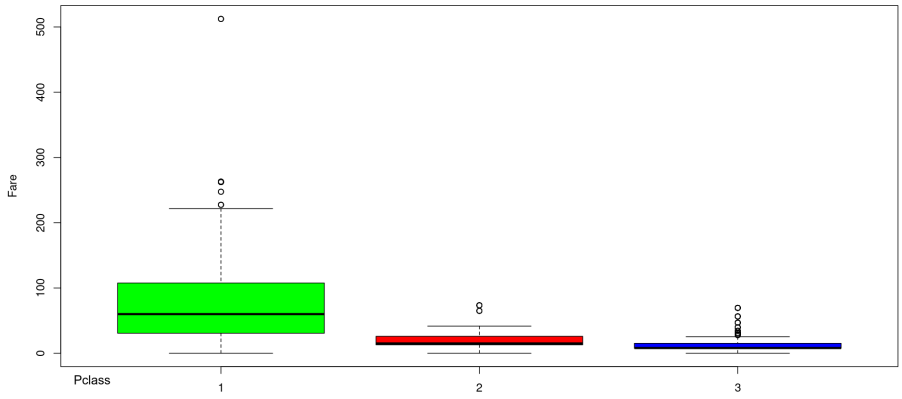
```r
# Is there any indication that overdispersion is a problem for you model?

library(AER)
## Loading required package: car
## Loading required package: carData
## Loading required package: lmtest
## Loading required package: zoo
## Attaching package: 'zoo'
## This objects are masked from 'package:base': as.Date, as.Date.numeric
## Loading required package: sandwich
## Loading required package: survival

dispersiontest(model1)
##
## Overdispersion test
##
## data: model1
## z = -11.48, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion 0.7387319

dispersiontest(model2)
##
## Overdispersion test
##
## data: model2
## z = -11, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion 0.7387319

dispersiontest(model3)
##
## Overdispersion test
##
## data: model3
## z = -10.75, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion 0.7398623
```

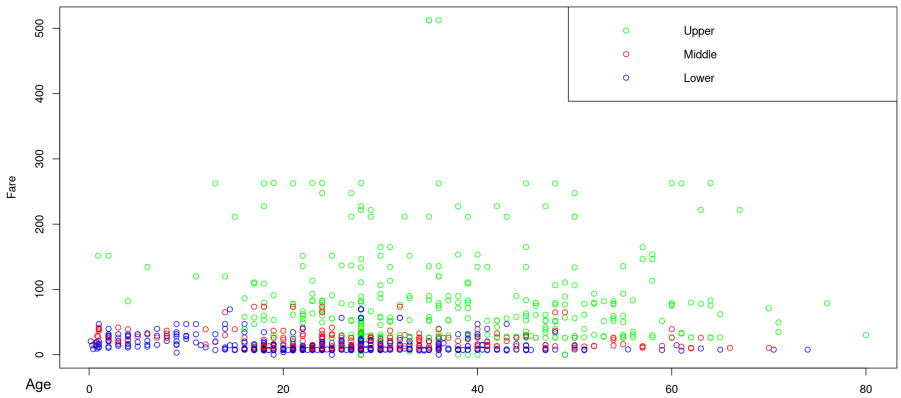
```r
# Deviance residuals according to the first model and sex
Man <- titanic[titanic$Sex==0,]
Woman <- titanic[titanic$Sex==1,]
Man <- Man[,-1]
Woman <- Woman[,-1]
model1.1 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Man)
summary(model1.1) # performance a little bit significant
model1.2 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Woman)
summary(model1.2) # performance a little bit significant
boxplot(Fare~Sex,data=titanic, col=c("green","red"))
n <- nrow(titanic)
color <- rep(NA, n)
color[titanic$Sex==0] <- "green"
color[titanic$Sex==1] <- "red"
plot(Age, Fare, col=color,main="Divided by sex")
legend("topright", c("Man", "Woman"), col=c("green", "red"), pch=1)
```



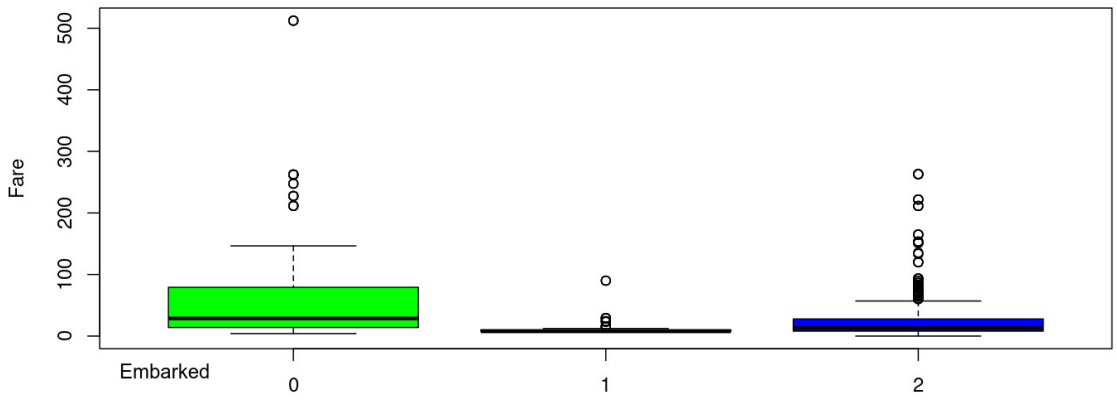


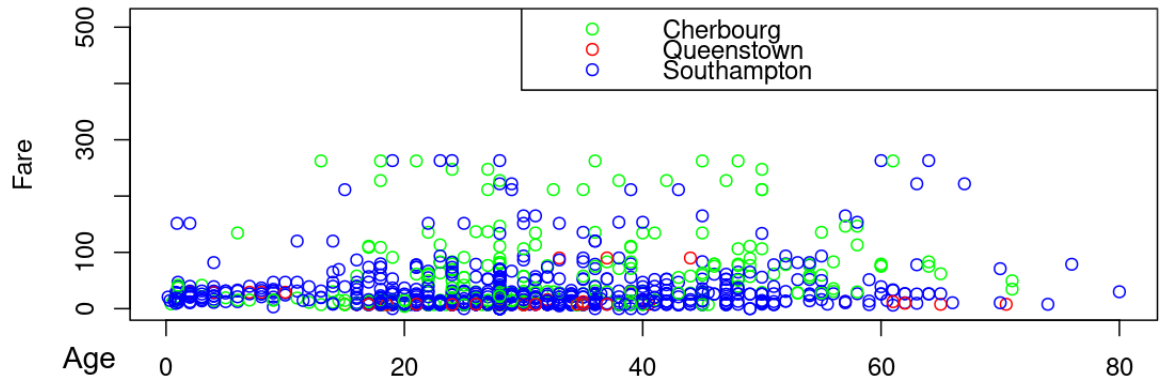
```r
# Deviance residuals according to the first model and class
Upper <- titanic[titanic$Pclass==1,]
Middle <- titanic[titanic$Pclass==2,]
Lower <- titanic[titanic$Pclass==3,]
Upper <- Upper[,-1]
Middle <- Middle[,-1]
Lower <- Lower[,-1]
model2.1 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Upper)
summary(model2.1) # performance a little bit significant
model2.2 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Middle)
summary(model2.2) # performance a little bit significant
model2.3 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Lower)
summary(model2.3) # performance a little bit significant
boxplot(Fare~Pclass,data=titanic, col=c("green","red","blue"))
n <- nrow(titanic)
color <- rep(NA,n)
color[titanic$Pclass==1] <- "green"
color[titanic$Pclass==2] <- "red"
color[titanic$Pclass==3] <- "blue"
plot(Age, Fare, col=color,main="Divided by class")
legend("topright",c("Upper","Middle","Lower"),col=c("green","red","blue"))
```



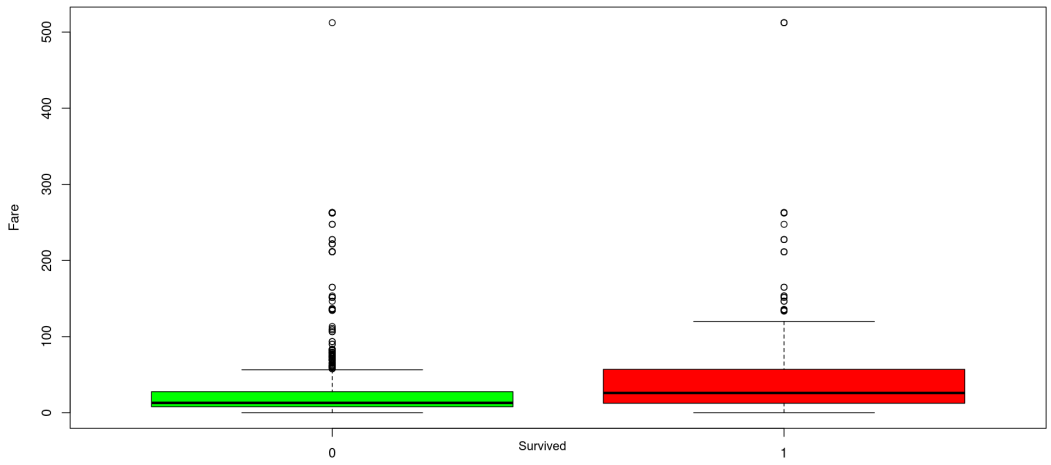


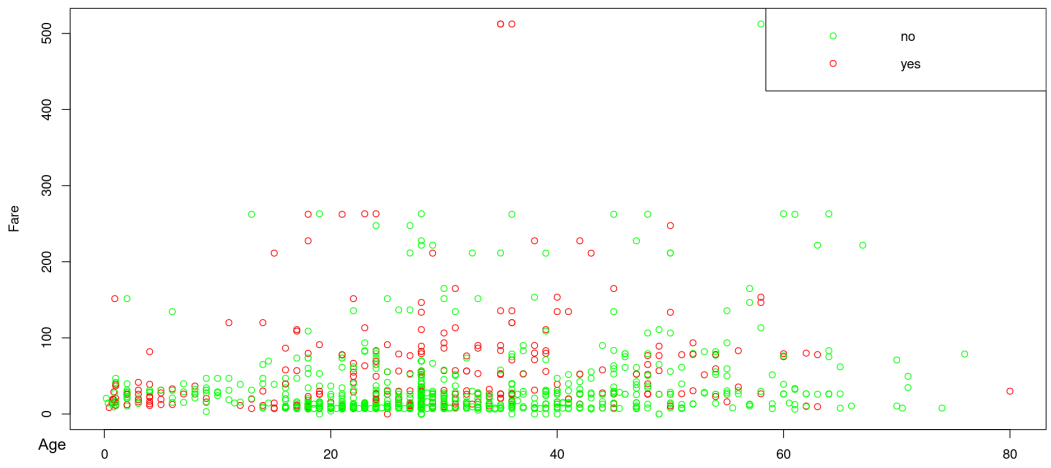
```r
# Deviance residuals according to the first model and embarked
Cherbourg <- titanic[titanic$Embarked==0,]
Queenstown <- titanic[titanic$Embarked==1,]
Southampton <- titanic[titanic$Embarked==2,]
Cherbourg <- Cherbourg[,-1]
Queenstown <- Queenstown[,-1]
Southampton <- Southampton[,-1]
model3.1 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Cherbourg)
summary(model3.1) # performance a little bit significant
model3.2 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Queenstown)
summary(model3.2) # performance a little bit significant
model3.3 <- glm(Fare~Age, family=quasipoisson(link="log"), data=Southampton)
summary(model3.3) # performance a little bit significant
boxplot(Fare~Embarked,data=titanic, col=c("green","red","blue"))
n <- nrow(titanic)
color <- rep(NA,n)
color[titanic$Embarked==0] <- "green"
color[titanic$Embarked==1] <- "red"
color[titanic$Embarked==2] <- "blue"
plot(Age, Fare, col=color,main="Divided by emarked")
legend("topright",c("Cherbourg","Queenstown","Southampton"),col=c("green","red","blue"),p
```





```
# Deviance residuals according to the first model and embarked
no <- titanic[titanic$X2urvived==0,]
yes <- titanic[titanic$X2urvived==1,]
no <- no[,-1]
yes <- yes[,-1]
model4.1 <- glm(Fare~Age, family=quasipoisson(link="log"), data=no)
summary(model4.1) # performance a little bit significant
model4.2 <- glm(Fare~Age, family=quasipoisson(link="log"), data=yes)
summary(model4.2) # performance a little bit significant
boxplot(Fare~X2urvived,data=titanic, col=c("green","red"))
n <- nrow(titanic)
color <- rep(NA,n)
color[titanic$X2urvived==0] <- "green"
color[titanic$X2urvived==1] <- "red"
plot(Age, Fare, col=color,main="Divided by emarked")
legend("topright",c("no","yes"),col=c("green","red"),pch=1)
'
```





Logistic Regression

```r
# To get started, we observe the data:
head(titanic)
# Passengerid Age    Fare Sex sibsp Parch Pclass Embarked X2urvived
# 1           1  22  7.2500   0     1     0      3        2         0
# 2           2  38 71.2833   1     1     0      1        0         1
# 3           3  26  7.9250   1     0     0      3        2         1
# 4           4  35 53.1000   1     1     0      1        2         1
# 5           5  35  8.0500   0     0     0      3        2         0
# 6           6  28  8.4583   0     0     0      3        1         0

general_logisitc <- glm(X2urvived ~ ., data = titanic, family = "binomial")
summary(general_logisitc)
#
# Call:
# glm(formula = X2urvived ~ ., family = "binomial", data = titanic)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -2.4428  -0.6011  -0.3050   0.3766   2.7279
#
# Coefficients:
#               Estimate Std. Error z value Pr(>|z|)
# (Intercept)  3.2061733  0.4770709   6.721 1.81e-11 ***
# Passengerid -0.0030985  0.0002478 -12.506  < 2e-16 ***
# Age         -0.0343026  0.0068425  -5.013 5.35e-07 ***
# Fare        -0.0002592  0.0017834  -0.145  0.88445
# Sex          2.2640321  0.1750012  12.937  < 2e-16 ***
# sibsp       -0.3132645  0.1010252  -3.101  0.00193 **
# Parch        0.0122141  0.1032653   0.118  0.90585
# Pclass      -0.9859739  0.1229424  -8.020 1.06e-15 ***
# Embarked    -0.1162352  0.0983299  -1.182  0.23717
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
# Null deviance: 1498.3  on 1306  degrees of freedom
# Residual deviance: 1001.4  on 1298  degrees of freedom
# (2 observations deleted due to missingness)
# AIC: 1019.4
# Number of Fisher Scoring iterations: 5
```

To continue, we will see the importance of each predictor in the model using the function varImp() from the package caret.

```
# The importance of each predictor in the model using the function varImp().
caret::varImp(general_logisitc)
```

	Overall <dbl>
Passengerid	12.5055240
Age	5.0131849
Fare	0.1453253
Sex	12.9372385
sibsp	3.1008550
Parch	0.1182793
Pclass	8.0198030
Embarked	1.1820937

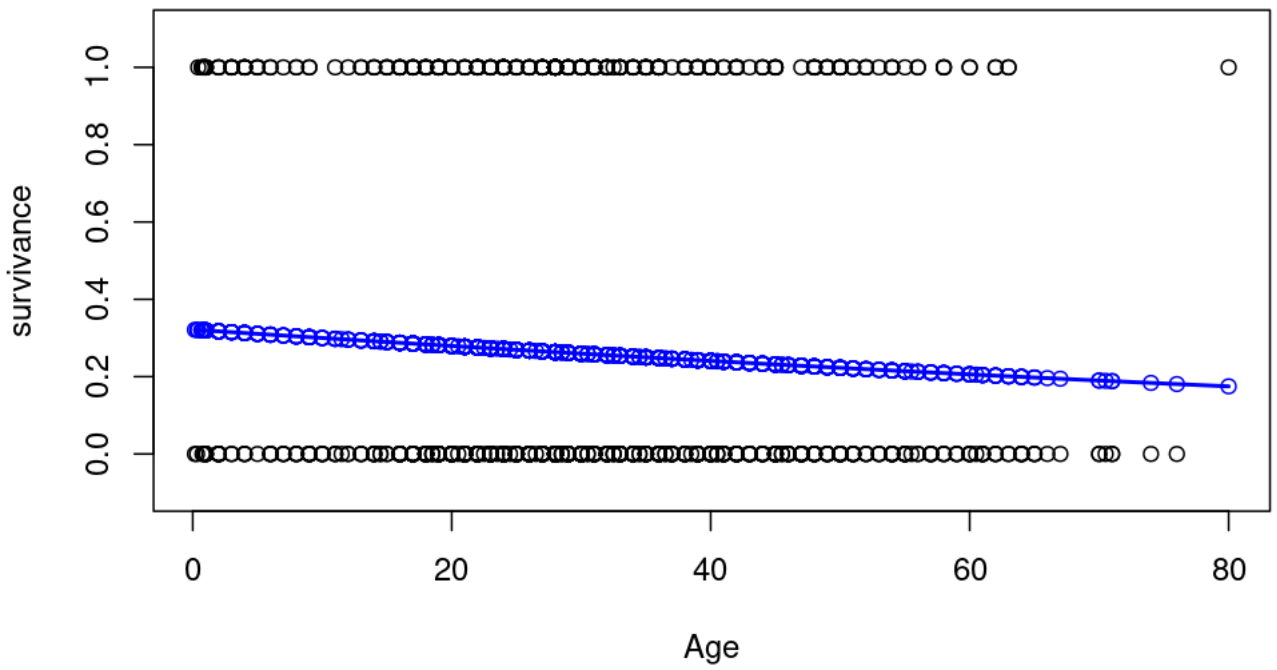
8 rows

Higher values indicate more importance, so the results obtained match up very well with the p values from our model. Being Age, Sex and Pclass good predictors. Next, we are going to calculate McFadden's R^2 to assess how well our model fits the data. The value ranges from 0 to just under 1, with higher values indicating better model fit.

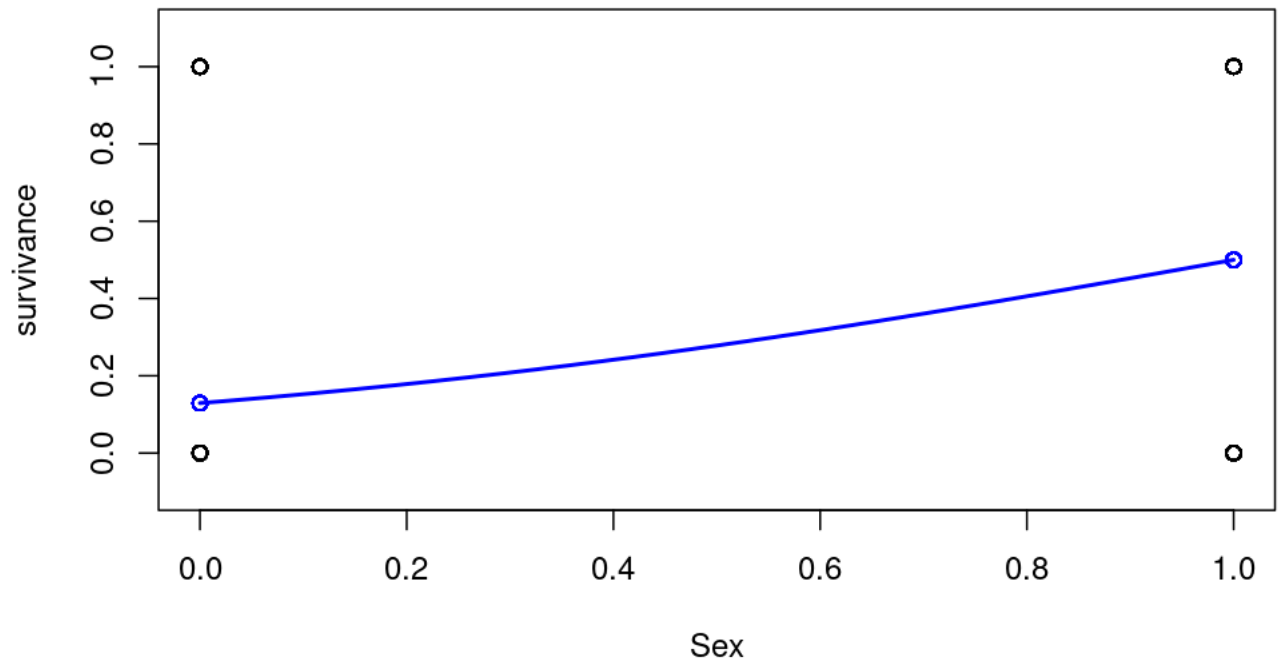
```
# Calculate McFadden's R^2 to assess how well our model fits the data.
pscl::pR2(general_logisitc)["McFadden"]
#
# fitting null model for pseudo-r2
#  McFadden
# 0.3316875
#
```

Our McFadden's R^2 is equal 0.3316875 which means that the logistic model fits pretty well our data.

```r
# SURVIVAL & AGE
logisitc_1 <- glm(X2urvived ~ Age, data = titanic, family = "binomial")
summary(logisitc_1)
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept) -0.746635   0.156381  -4.774  1.8e-06 ***
# Age         -0.010059   0.004986  -2.017   0.0437 *
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# (Dispersion parameter for binomial family taken to be 1)
# Null deviance: 1503.7  on 1308  degrees of freedom
# Residual deviance: 1499.6  on 1307  degrees of freedom
# AIC: 1503.6
# Number of Fisher Scoring iterations: 4


# Now let's plot the fitted logistic model:
# In blue we will represent the predicted values according to the model.
plot(titanic$Age, titanic$X2urvived, xlab = "age", ylab = "survive")
curve(predict(logisitc_1,data.frame(Age=x),type = "resp"),add=TRUE,col="blue")
points(titanic$Age,fitted(logisitc_1),pch=1,col="blue")
```



```r
# SURVIVAL & SEX
logisitc_2 <- glm(X2urvived ~ Sex, data = titanic, family = "binomial")
summary(logisitc_2)
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)  -1.9072     0.1026  -18.58   <2e-16 ***
# Sex           1.9072     0.1383   13.79   <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: 1503.7  on 1308  degrees of freedom
# Residual deviance: 1295.2  on 1307  degrees of freedom
# AIC: 1299.2
# Number of Fisher Scoring iterations: 4


# Now let's plot the fitted logistic model:
# In blue we will represent the predicted values according to the model.
plot(titanic$Sex, titanic$X2urvived, xlab = "Sex", ylab = "survive")
curve(predict(logisitc_2,data.frame(Sex=x),type = "resp"),add=TRUE,col="blue")
points(titanic$Sex,fitted(logisitc_2),pch=1,col="blue")
```



```r
# SURVIVAL & CLASS
logisitc_3 <- glm(X2urvived ~ Pclass, data = titanic, family = "binomial")
summary(logisitc_3)
#
# Coefficients:
#              Estimate Std. Error z value Pr(>|z|)
# (Intercept)  0.37750    0.16950   2.227   0.0259 *
# Pclass      -0.64652    0.07478  -8.645   <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: 1503.7  on 1308 degrees of freedom
# Residual deviance: 1427.4  on 1307 degrees of freedom
# AIC: 1431.4
# Number of Fisher Scoring iterations: 4


# Now let's plot the fitted logistic model:
# In blue we will represent the predicted values according to the model.
plot(titanic$Pclass, titanic$X2urvived, xlab = "Pclass", ylab = "survive")
curve(predict(logisitc_3,data.frame(Pclass=x),type = "resp"),add=TRUE,col="blue")
points(titanic$Pclass,fitted(logisitc_3),pch=1,col="blue")
```