GLM Coursework Assignment 2: Titanic

By: Saif Rehman

**Introduction**

The sinking of the titanic was one of the most

The dataset used is the “titanic.txt” dataset which contains data on 1313 passengers on the Titanic: their name, age, sex, passenger class, and whether they survived or not. This dataset was generated by the Biostatistics department for University of Vanderbilt and can be publicly accessed through their website or the statsci website.

The variables are:

|  |  |
| --- | --- |
| Name | Name of passenger |
| Age | Age of passenger |
| Sex | Sex of passenger |
| Pclass | Passenger Class (First, second, third) |
| Survived | Whether passenger survived or not (1 = survived, 0 = did not survive) |

In the subsequent analysis, I will treat Sex and PClass as factors and Age as continuous.

Upon some initial analysis and inspection of the titanic dataset, there are a considerable number of NA values present in the dataset. There are a total of 557 NA entries present in the dataset and upon some basic analysis, it was uncovered that all 557 of these NA entries are due to Age values not being present in the dataset for those 557 passengers. To do analysis on this data, each row which contains an NA value is completely removed from the dataset. So, 557 NA values are then removed from the dataset leading to a dataset for 756 passengers in total with no NA further values present in the dataset. So despite only Age being missing for those 557 passengers, the entire row for those passengers were removed. This can be taken for further consideration in future analysis which might include those rows and use them as a predictor value for improving the model or testing the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistics** | | | | | |
|  | | PClass | Age | Sex | Survived |
| N | Valid | 1313 | 756 | 1313 | 1313 |
| Missing | 0 | 557 | 0 | 0 |

The response variable of interest Survived, which is coded 1 for whether the passenger survived and 0 for if the passenger didn’t survive. Using this in a binomial logistic regression, I will model the probability of survival and whether survival of the passengers is associated with the explanatory variables listed above.

**The Statistical Model**

I fit a generalized linear model, choosing a binomial distribution and a logit link. This method is commonly known as logistic regression. It can be described mathematically as:

Where represents the probability of survival for passenger ἰ. are a set of *p* explanatory variables, and are a set of unknown regression coefficients which are to be estimated from the model. I assume that the observed response is binomially distributed with sample size 1 and with unknown probability

I chose this model because the response variable is binary. The logistic link - – allows me to make statements about the odds of survival from the generalized model.

The significance of terms in the model will be determined by looking at the differences of deviances (-2 log likelihood) and will be produced by the Anova function in R.

**Descriptive Analysis**

I use the summary() functions in R to provide a descriptive analysis of the dataset. The table below provides the mean and standard deviation for Age as it is a continuous variable while providing the category counts and percentages for Sex and PClass.

|  |  |  |  |
| --- | --- | --- | --- |
| **PClass** | | | |
|  | | Frequency | Percent |
|  | 1st | 226 | 29.9 |
| 2nd | 212 | 28.0 |
| 3rd | 318 | 42.1 |
| Total | 756 | 100.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Sex** | | | |
|  | | Frequency | Percent |
|  | female | 288 | 38.1 |
| male | 468 | 61.9 |
| Total | 756 | 100.0 |

|  |  |
| --- | --- |
| **Age** |  |
| Min | 0.17 |
| 1st Quartile | 21 |
| Median | 28 |
| Mean | 30.40 |
| 3rd Quartile | 39 |
| Max | 71 |

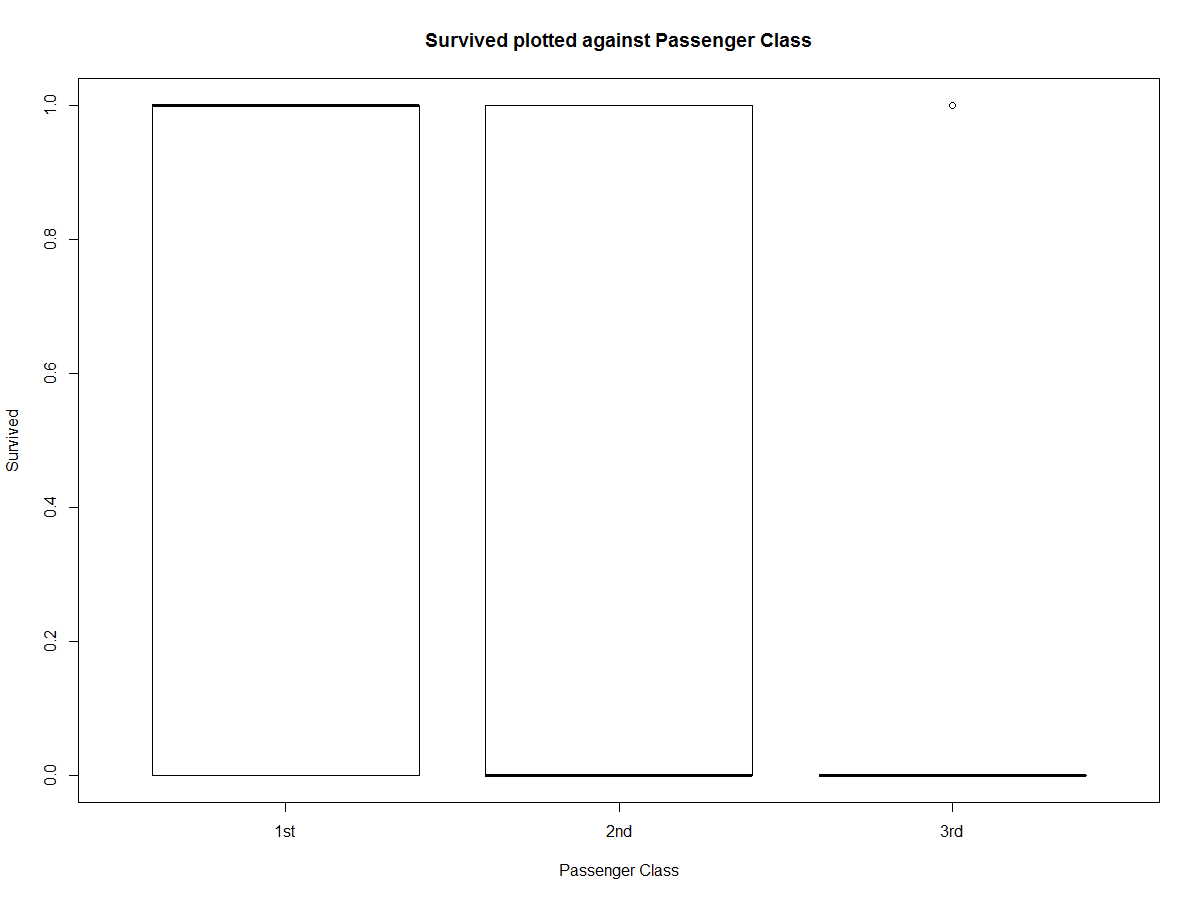
|  |  |  |  |
| --- | --- | --- | --- |
| **Survived** | | | |
|  | | Frequency | Percent |
|  | 0 | 443 | 58.6 |
| 1 | 313 | 41.4 |
| Total | 756 | 100.0 |

It can be seen that in terms of survival, 443 passengers (58.6%) did not survive while only 313 passengers (41.4%) survived the titanic crash. The majority of the passengers, 468 passengers in total (61.9%), are male and only 288 passengers are female (38.1%) so there is a greater proportion of males compared to females. The mean age of the passengers is 30.40 with the minimum being 0.17 and the maximum being 71 years old. For the different passenger classes, there are 226 passengers in the 1st class (29.9%), 212 passengers in 2nd class (28%), and 318 passengers in 3rd class (42.1%) so it does seem that 3rd class does have the most number of passengers while 1st and 2nd class have a smaller but roughly the same number of passengers.

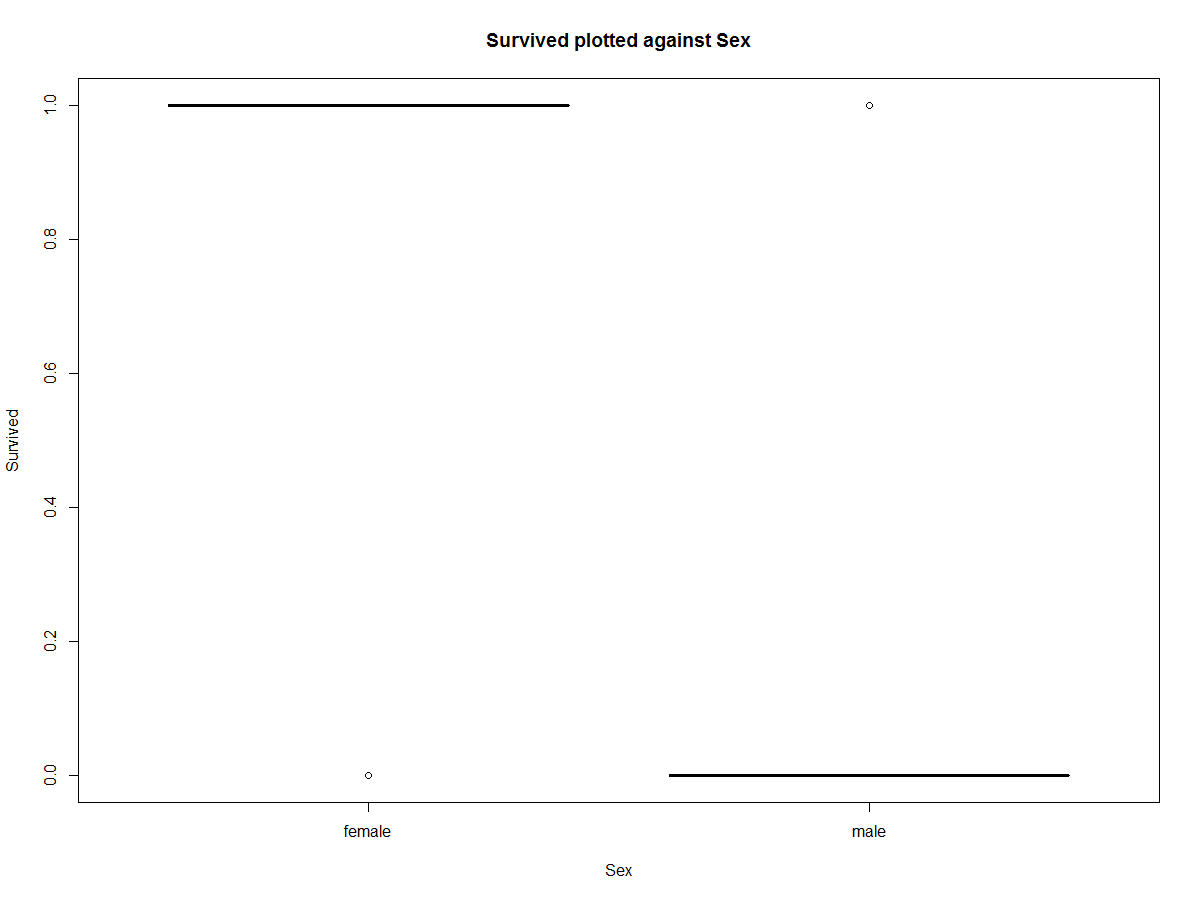
**Data Exploration**

I used a variety of graphs to explore different relationships between the data and to get a better visual representation of the dataset itself:

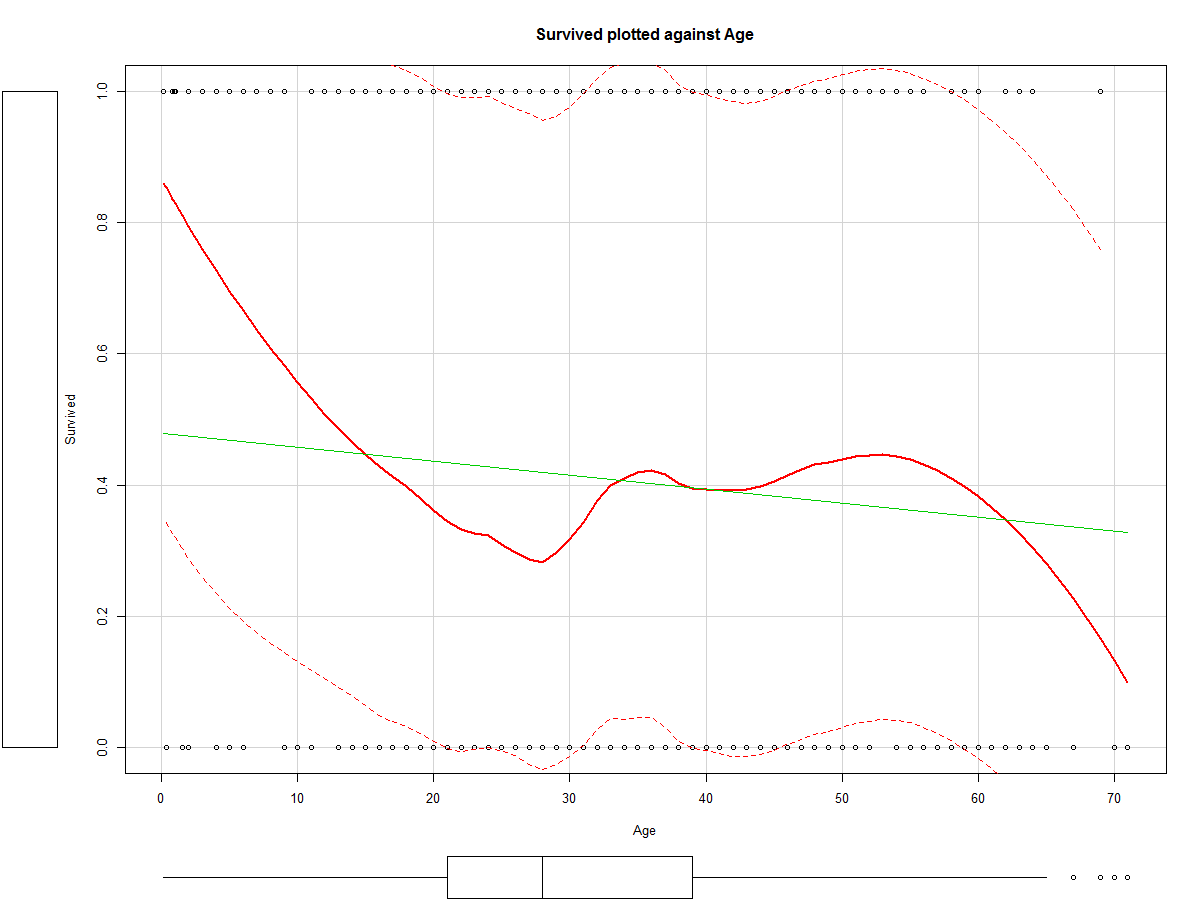
Plotting Survived against Passenger Class shows that on average, 1st class passengers survived while 2nd and 3rd class passengers on average didn’t survive. For 3rd class passengers, surviving is an outlier of the data.



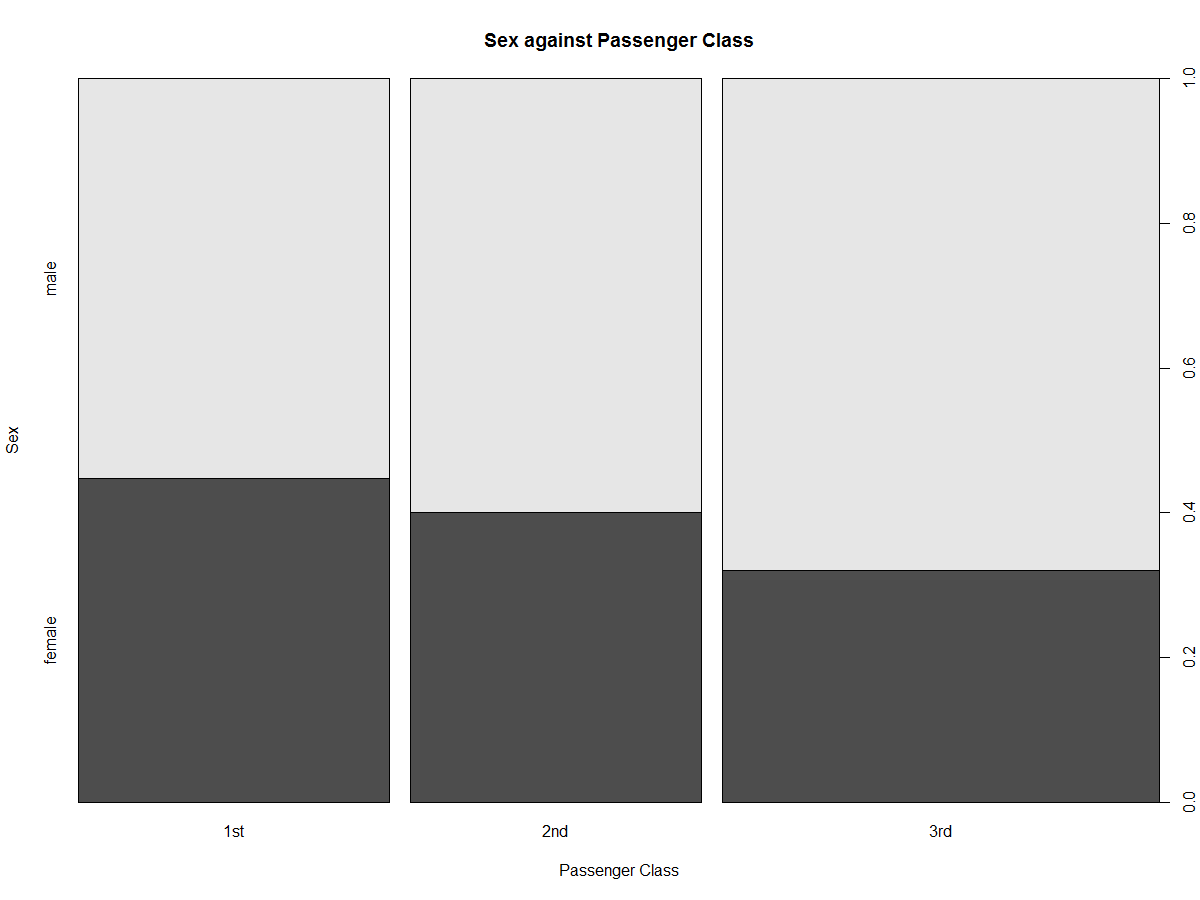
Plotting Survived against Sex shows a similar relationship as passenger class except females, on average, are surviving with not surviving as an outlier. On the other hand, males on average are not surviving with surviving as an outlier.



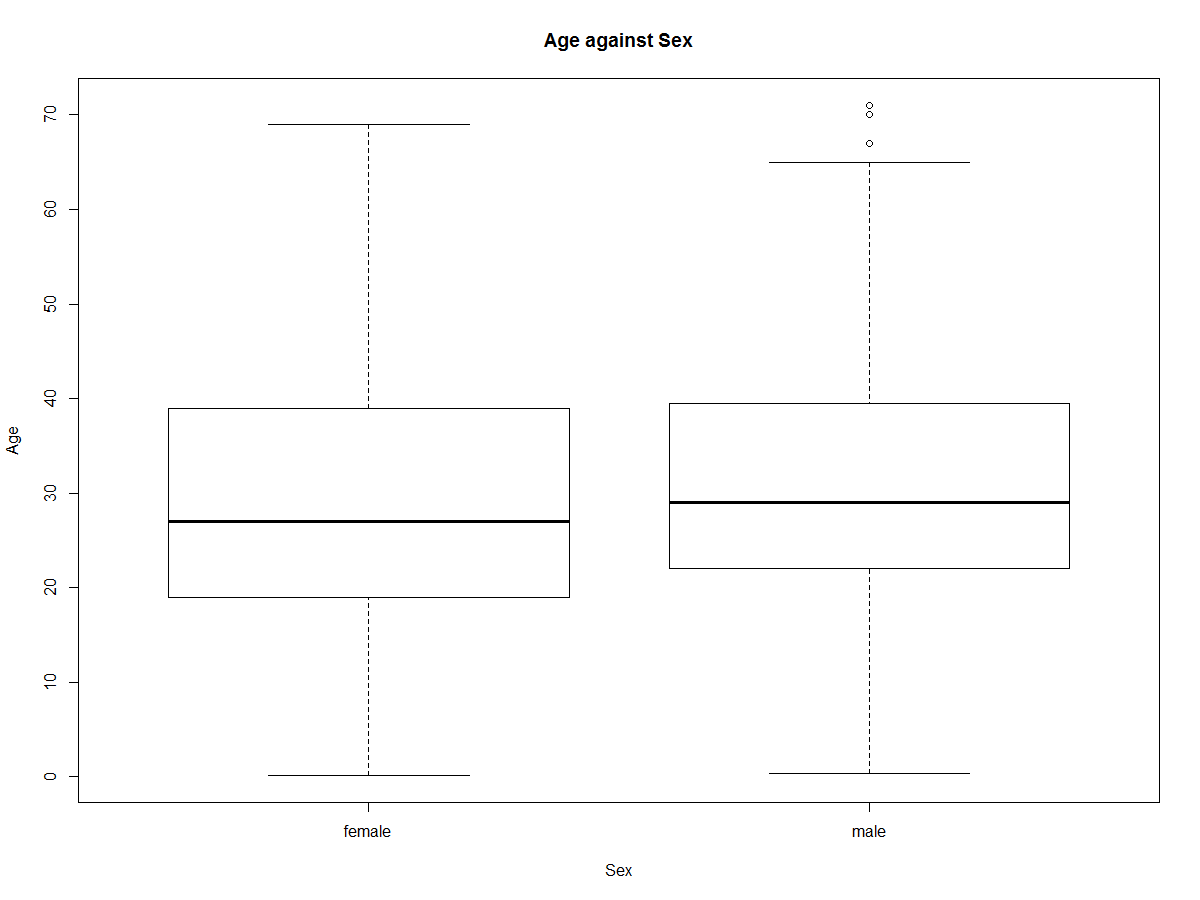
Doing a scatterplot for Survived plotted against Age shows that on average, inferring from the linear regression line, that younger passengers have the greatest chance of survival. So, the trend on average is that as age increases, the chance of survival decreases.



Plotting passenger class against sex shows that there are generally more males in each passenger class, with the first class having the most even distribution and the 3rd class having the least even distribution.



Plotting sex against age shows that there is a seemingly even distribution of age between the two genders. This is important because if there were predominantly more younger females, then that would affect conclusions stating that younger passengers have a greater chance of survival.



**Statistical Analysis**

I first fit a main effects model that involves all of the covariates, PClass, Sex, and Age, with Surivival as the response variable. The R code and the regression coefficients are as follows:

> fit1 <- glm(Survived ~ PClass + Age + Sex, data = titanic , family = binomial)

> summary(fit1)

Call:

glm(formula = Survived ~ PClass + Age + Sex, family = binomial,

data = titanic)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.7226 -0.7065 -0.3917 0.6495 2.5289

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.759662 0.397567 9.457 < 2e-16 \*\*\*

PClass2nd -1.291962 0.260076 -4.968 6.78e-07 \*\*\*

PClass3rd -2.521419 0.276657 -9.114 < 2e-16 \*\*\*

Age -0.039177 0.007616 -5.144 2.69e-07 \*\*\*

Sexmale -2.631357 0.201505 -13.058 < 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: 1025.57 on 755 degrees of freedom

Residual deviance: 695.14 on 751 degrees of freedom

AIC: 705.14

To test the significance of this model and the significance of all of the covariates, I used the Anova() command in library car :

> Anova(fit1)

Analysis of Deviance Table (Type II tests)

Response: Survived

LR Chisq Df Pr(>Chisq)

PClass 100.445 2 < 2.2e-16 \*\*\*

Age 28.454 1 9.595e-08 \*\*\*

Sex 214.776 1 < 2.2e-16 \*\*\*

---

The analysis of deviance table shows that all of the covariates are found to be significant. This means that passenger class, age, and sex have an effect of survival probability. If the covariates were found to be not significant using 5% as the threshold level, then I would perform backwards elimination until all of the covariates are found to be significant at the 5% level.

Testing the goodness of fit of this model, I used a goodness of fit test using a chi-squared distribution since this is binomial data with a sample size greater than one:

> 1-pchisq(695.14, 751)

[1] 0.9280469

As shown by the goodness of fit test, the probability is very high which means that the model fits very well. If the probability was less than 5%, or 0.05, then the model would not be a very good fit and it would need to be adjusted.

To interpret the regression coefficients, I exponentiate them:

> cbind(coef(fit1), exp(coef(fit1)))

[,1] [,2]

(Intercept) 3.75966210 42.93391621

PClass2nd -1.29196240 0.27473112

PClass3rd -2.52141915 0.08034550

Age -0.03917681 0.96158067

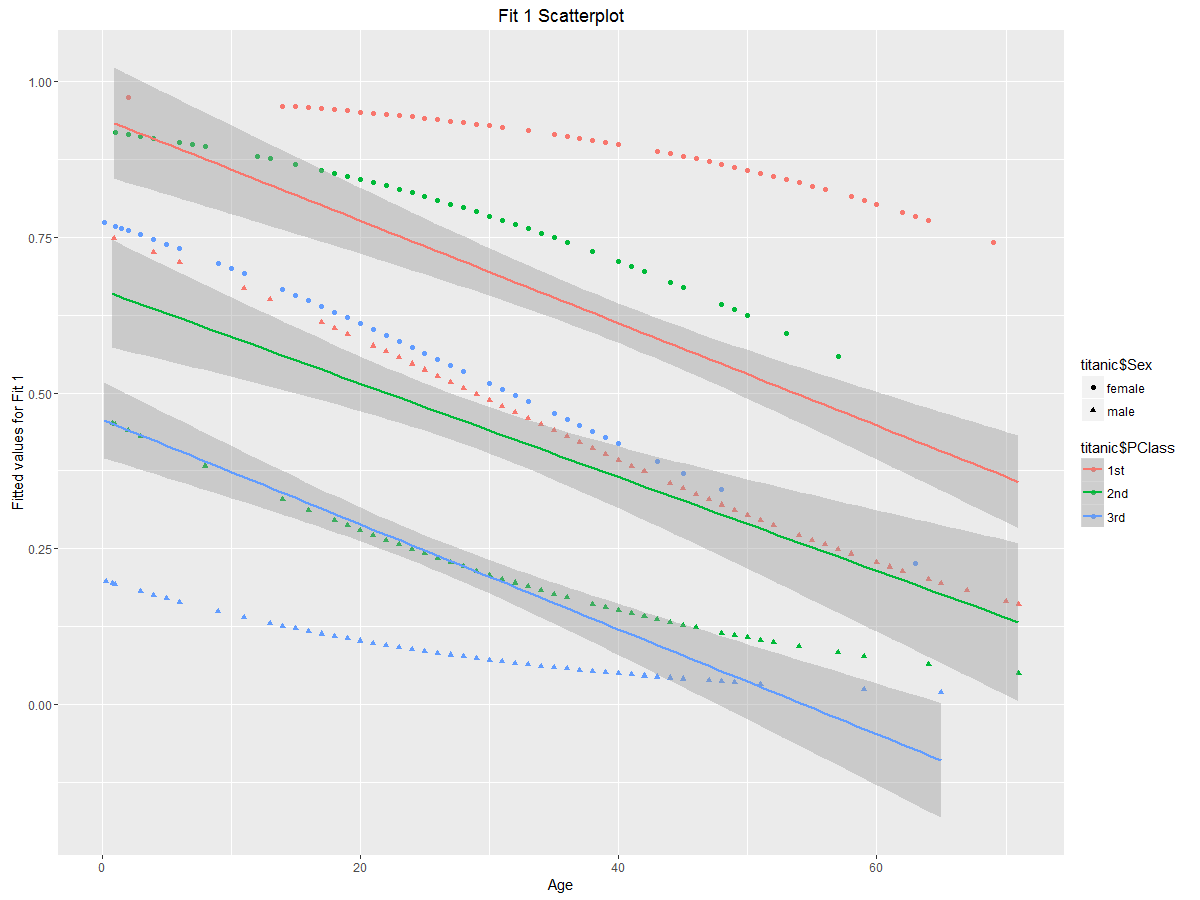
Sexmale -2.63135683 0.07198073

This shows the following effects:

Comparing the odds of survival with 1st class passengers and 2nd class passengers: 2nd class passengers have about a 73% decrease in odds of survival as compared with 1st class passengers (since the coefficient is 0.2747). Similarly, 3rd class passengers have a 92% decrease in odds of survival as compared with 1st class passengers.

For every year of increase in age, the odds of survival decrease by 4%.

Similarly, comparing female passengers with male passengers, male passengers have a 93% decrease in odds of survival as compared with female passengers.



Graphically interpreting the fitted values plotted against age shows that for all of the passengers, the probability of survival decreases with age. The plot also shows that female passengers have a higher probability of survival as compared to the male passengers. Also, 1st class passengers have a higher probability of survival then 2nd class passengers, with 3rd class passengers having the lowest probability of survival. So in total, female 1st class passengers at age 0 theoretically have the greatest chance of survival and 3rd class male passengers at the maximum age have the lowest chance of survival. The three lines with the grayed-out confidence intervals show the linear regression lines for 1st class passengers (red), 2nd class passengers (green), and 3rd class passengers(blue) for both of the genders combined.

**Adding quadratic for age**

The previous model was monotonic in age; however, it might be a possibility that there may be a nonlinear effect of age so I fit a model that includes a quadratic in age:

> fit3 <- glm(Survived ~ PClass + Age + I(Age^2) + Sex, data = titanic, family = binomial)

> summary(fit3)

Call:

glm(formula = Survived ~ PClass + Age + I(Age^2) + Sex, family = binomial,

data = titanic)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8277 -0.7246 -0.3753 0.6191 2.5329

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 4.1178339 0.4744070 8.680 < 2e-16 \*\*\*

PClass2nd -1.2403997 0.2619086 -4.736 2.18e-06 \*\*\*

PClass3rd -2.4787835 0.2783369 -8.906 < 2e-16 \*\*\*

Age -0.0702583 0.0229622 -3.060 0.00222 \*\*

I(Age^2) 0.0004861 0.0003358 1.448 0.14774

Sexmale -2.6282307 0.2020141 -13.010 < 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: 1025.57 on 755 degrees of freedom

Residual deviance: 693.07 on 750 degrees of freedom

AIC: 705.07

In order to test the significance of this model, I have used the Anova command:

> Anova(fit3)

Analysis of Deviance Table (Type II tests)

Response: Survived

LR Chisq Df Pr(>Chisq)

PClass 95.488 2 < 2.2e-16 \*\*\*

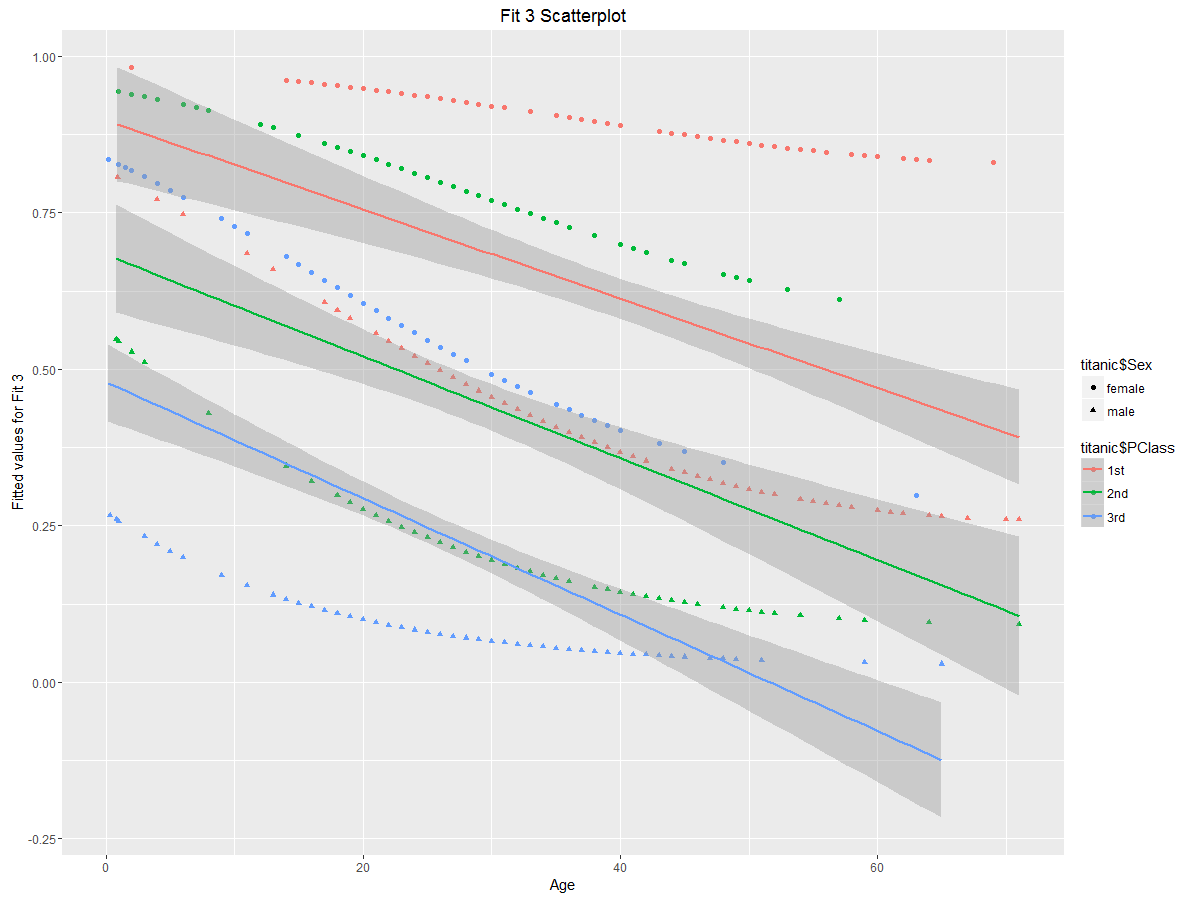
Age 9.454 1 0.002106 \*\*

I(Age^2) 2.075 1 0.149763

Sex 213.180 1 < 2.2e-16 \*\*\*

---

As seen through the analysis of deviance table for fit3, the quadratic term for Age is not significant at the 5% level so it is reasonable to drop the term.



Comparing the previous graph of Fit 1 and this graph, Fit 3, the main trends for gender and passenger class remain exactly the same. Female passengers still have the greatest odds of survival, with odds decreasing with increasing age. Furthermore, 1st class passengers of the respective gender have higher odds of survival, then the 2nd class passengers, then the 3rd class passengers with the lowest odds of survival. So, a male with the maximum age in 3rd class will have the lowest odds of survival. The only thing that marginally changes is the slope of each line. However, the change is not that significant as we can see from the Anova test.

**Testing for interactions**

After checking that that Age^2 is not significant, it makes sense to check if there are any interaction terms in the model so I have modeled three two-way interactions between the covariates in the model:

> fit6 <- glm(Survived ~ PClass + Age + Sex + PClass:Age + PClass:Sex + Age:Sex, data = titanic, family = binomial)

> summary(fit6)

Call:

glm(formula = Survived ~ PClass + Age + Sex + PClass:Age + PClass:Sex +

Age:Sex, family = binomial, data = titanic)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.5882 -0.6394 -0.3678 0.4206 2.7230

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.463599 0.845664 2.913 0.003577 \*\*

PClass2nd 1.117752 1.018111 1.098 0.272262

PClass3rd -2.807199 0.833634 -3.367 0.000759 \*\*\*

Age 0.013492 0.020250 0.666 0.505246

Sexmale -0.946499 0.833357 -1.136 0.256055

PClass2nd:Age -0.065120 0.024761 -2.630 0.008539 \*\*

PClass3rd:Age -0.007055 0.020288 -0.348 0.728047

PClass2nd:Sexmale -1.410385 0.729056 -1.935 0.053047 .

PClass3rd:Sexmale 1.033269 0.625510 1.652 0.098558 .

Age:Sexmale -0.068052 0.018607 -3.657 0.000255 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1025.57 on 755 degrees of freedom

Residual deviance: 641.37 on 746 degrees of freedom

AIC: 661.37

Number of Fisher Scoring iterations: 5

I use the Anova test to check the significance of each of the interaction terms in the model:

> Anova(fit6)

Analysis of Deviance Table (Type II tests)

Response: Survived

LR Chisq Df Pr(>Chisq)

PClass 103.479 2 < 2.2e-16 \*\*\*

Age 32.998 1 9.223e-09 \*\*\*

Sex 219.110 1 < 2.2e-16 \*\*\*

PClass:Age 8.581 2 0.0136958 \*

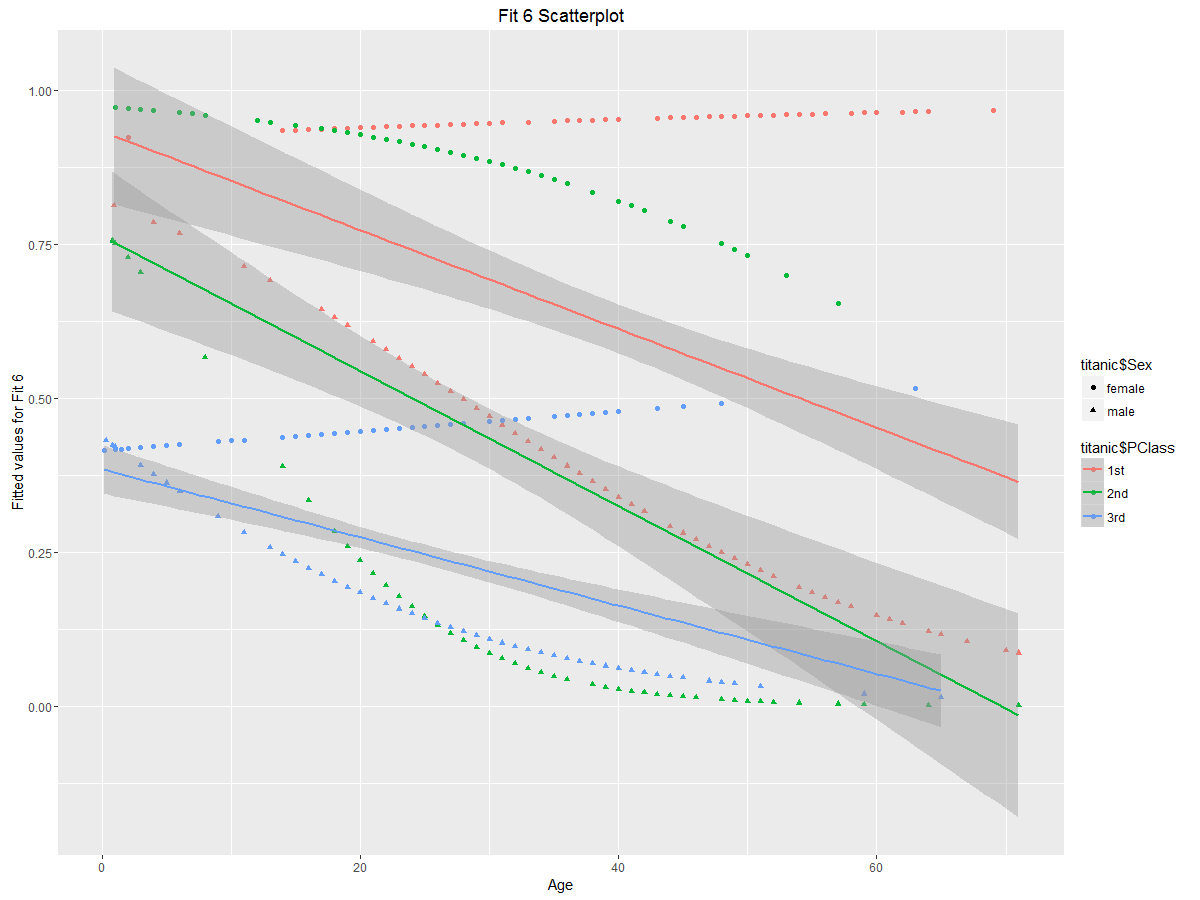
PClass:Sex 21.080 2 2.646e-05 \*\*\*

Age:Sex 13.671 1 0.0002178 \*\*\*

---

From the results of the Anova test, it is shown that all three interaction terms are deemed to be significant by using a threshold at the 5% level so there is no need to perform backwards elimination.

To interpret this graphically:



Plotting the fitted values against Age does change the shape of the graph drastically. The graph does show basically, female passengers in the 1st class do mainly survive with their odds of survival increasing with age. Also, 2nd class female passengers have the greatest chance of survival when they are youngest, with their chance of survival decreasing with age. Similarly, males in the 1st class have a fairly higher chance of survival as compared with the previous graphs, however, their chance of survival decreases greatly with age (with the biggest degree negative slope). All males in general have the same trend of their odds of survival decreasing with age, with 2nd and 3rd class male passengers having the least odds of survival. One difference in this graph is that 3rd class female passengers and 1st class female passengers have increasing odds of survival with age while in the previous graphs it is predicted that the higher the age, the lower the odds of survival. However, the overall trend as seen by the three linear regression lines for each respective passenger class remains relatively the same as it still decreases with age.

> cbind(coef(fit6), exp(coef(fit6)))

[,1] [,2]

(Intercept) 2.46359906 11.74701374

PClass2nd 1.11775231 3.05797310

PClass3rd -2.80719940 0.06037384

Age 0.01349163 1.01358305

Sexmale -0.94649853 0.38809756

PClass2nd:Age -0.06511982 0.93695519

PClass3rd:Age -0.00705466 0.99297017

PClass2nd:Sexmale -1.41038489 0.24404933

PClass3rd:Sexmale 1.03326894 2.81023733

Age:Sexmale -0.06805151 0.93421235

Interpreting the coefficients for this model becomes a bit tricky with all of the interaction terms. For example, a female in the 2nd class passenger class has triple the odds of surviving as compared to a first-class passenger class however, as age increases for that passenger, then the odds decreases by around 6% per age. The graphical interpretation makes the model much easier to understand.

Testing the goodness of fit of this model, I used a goodness of fit test using a chi-squared distribution since this is binomial data with a sample size greater than one:

> 1-pchisq(641.37, 746)

[1] 0.9976684

As shown by the goodness of fit test, the probability is very high which means that the model fits very well. If the probability was less than 5%, or 0.05, then the model would not be a very good fit and it would need to be adjusted. It fits even better than the other model, fit1, as that only gave a probability of 92.8%.

**Testing accuracy of the models**

As a test of the models, I have recreated the titanic data with all of the NA values removed and implemented them in a new data frame called “test”. Then I used the model, fit1, and used the predict function, rounding the answers to obtain binomial responses and stored in the column “model”. Then the model responses in the test was compared to the actual survival of the passengers given in the titanic data set and seeing the percentage of correct responses.

> test <- read.table("titanic.txt", header = TRUE)

> test <- na.omit(test)

> test$model <- round(predict(fit1, titanic, type = "response"))

> nrow(test[test$model == test$Survived,])

[1] 594

> nrow(test[test$model == test$Survived,])/nrow(test)

[1] 0.7857143

This shows that the model predicted 594 entries for the survived and not survived passengers correctly using the fit1 model. This divided by the total number of rows in the titanic data set (with NA values removed, total of 756 rows) gives the percentage that the model is correct for the titanic data set which is 78.57%. This means model that the fit1 accurately predicted the survival of the passengers 78.57% of the time. Although this is a fairly okay percentage, it could be further improved.

Now using the same technique using the final model, fit6:

> test <- read.table("titanic.txt", header = TRUE)

> test <- na.omit(test)

> test$model <- round(predict(fit6, titanic, type = "response"))

> nrow(test[test$model == test$Survived,])

[1] 616

> nrow(test[test$model == test$Survived,])/nrow(test)

[1] 0.8148148

Using the fit6 model, it predicted 616 rows corrected which is a slight improvement from the previous model. The fit6 model predicts the correct survival of the passengers 81.48% of the time which is a slight improvement from the fit1 model, but there is still room for improvement. That could mean that there are some passengers which were given preference that are not modeled (such as some high priority males that were given preference and survived even though males generally didn’t survive according to the model), that there may be other covariates affecting the model that are not included in the model, or that could also mean that the model needs to be further improved.

**Conclusion**

From the various models, it can be shown that the crew of the titanic did give priority to women and children in the lifeboats, thereby saving a greater proportion of them. This is especially graphically clear in the first model (fit1), as when using the graphical representation of the fitted values for fit1, females in general always had higher odds of surviving as compared with their male counterparts in their respective passenger classes. Furthermore, the graphical interpretation for fit1 did show that as age increased, the odds of survival did decrease for all passengers, males and females.

Furthermore, there is also evidence to suggest that the crew also gave priority to passengers in the higher passenger classes thereby saving a greater proportion of them. This is also visually evident in the graphical representation of the fitted values for fit1 as the 1st class passenger class for females and males had greater odds of survival as compared with the 2nd class passenger’s counterparts which further compares with 3rd class.

When adding a quadratic term to age did not make a significant impact to the model except just changing the slope a bit. The quadratic term was found to be not significant so it was then removed.

However, when adding interaction terms to the model, it did show that some newer relationships. The graphical interpretation of the fitted values for fit6 showed that first class females did have the overall highest odds of survival. 2nd class females also had high odds of survival, however, it did drop with age as compared with the 1st female passengers which didn’t vary with age. 1st class and 2nd class male passengers that were babies did have high odds of survival, with the odds decreasing with age while 2nd and 3rd class male passengers after the age of 25 basically had the lowest odds of survival.

Overall, there was evidence to conclude that female passengers, babies, and passengers of the higher class were given priority by the crew which thereby increased their odds of survival. So females on average had higher odds of survival. Odds of survival for all passengers, on average, did decrease with increases in age. Finally, on average, passengers of higher class had higher odds of survival as compared with lower class.