Logistic Regression Analysis for Titanic Data

The data under investigation records the survival chances of 1313 passengers on the ill-fated liner. The aim of the investigation is to examine whether women and children had a higher survival rate than adult males, and to explore the part that passenger class played in one's likelihood of survival.

The initial analysis showed that of the 1313 passengers, 851 (65%) were male and 462 (35%) female. Of the total, only 450 (34%) survived. That said, women were more likely to survive than men. Of the 462 females, 308 (66%) survived. This compared favourably with the 142 (16.6%) of the men who escaped. This conceals steep discrepencies from one passenger class to another. The female survival rate was high in both 1st and 2nd class, at 97% and 88% respectively. This compared with relatively poor male survival rates of 32% in 1st class and 14% in 2nd class. It is in 3rd class where the variation in female and male survival rates evens out (though in an unfortunate way) with 37% of females surviving, compared to about 12% of the men. Therefore, female survival rates were quite high in 1st and 2nd class, but then declined considerably if the female passenger had the misfortune to be in 3rd class. The picture for male survival was somewhat more dismal, starting at under a third in 1st class before falling further in both 2nd and 3rd class. Interestingly, the difference in male survival rates between 2nd and 3rd class came to just 1 percentage point.

The age of the passengers was not always recorded, which made analysis more challenging. Of the 1313 passengers, age data was provided for 756 (58%) of them. Therefore, only a subset of the data could be examined. An initial decision had to be made about what constituted a 'child'. Initially, this was assumed to be a passenger under the age of 17. There were 83 children in this group and 63% of them survived – somewhat higher than the 34% of all passengers who survived. As before, there were strong class differences. Of the 10 children in 1st class, all except one (Helen Allison - a 2 year old female) survived. Similarly, of the 26 children in 2nd class, all except two (both male) escaped. As before, the picture for children in 3rd class was less than ideal with only 19 (12 female, 7 male) of the 47 children (40%) surviving. The decline in child survival rates from 2nd to 3rd (92% to 40%) was, broadly, as precipitious as that for the adults. For a child in 3rd class, the chances of survival were only marginally higher that of the female adults (40% vs 37%). When the defintion of a child was moved from under 17 to under 13, the general trend remained the same. The survival rate increased slightly to 67% (from 63%). Interestingly, the survival rate for children in 1st class fell marginally (from 90% to 80%), whilst rising to 100% in 2nd class, using this alternative definition of a child. In 3rd class, the survival rate grew marginally from 40% to 42%, with 14 of the 33 children surviving.

To undertake an initial logistic regression analysis, the categorical variables 'sex' and 'class' were coded as factors. The glm() command was then used to fit the model with the full dataset (1313 passengers), with 'survived' as the response variable (ie. whether or not one survived the shipwreck), and with 'Sex' and 'Class' as the predictor variables. The coefficients obtained when this was run are shown below:

```
Call:
glm(formula = Survived ~ PClass + Sex, family = "binomial", data = titanic2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0414  -0.6992  -0.3908   0.5156   2.2848

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9507     0.1718  11.356  < 2e-16 ***
PClass2nd    -0.7894     0.1954  -4.039 5.36e-05 ***
PClass3rd    -2.0391     0.1774 -11.496  < 2e-16 ***
Sexmale      -2.4454     0.1512 -16.177  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1 – Coefficient data from the logistic regression model on the 'Titanic' data

It is noticable that each of the (estimate) coefficients are negative, indicating that they each have a negative effect on one's chances of survival – they decrease one's chances of survival. For the 'Sex' variable, female is the reference value, and therefore the estimates for males are relative to those of the females. For the variable 'PClass', then 1st class is the reference value, meaning that 2nd and 3rd class estimates are defined in relation to those for 1st class. The magnitudes vary sharply, however. The coefficients for 'male' and for '3rd class' are both high (greater than 2), suggesting a stronger negative effect on survival. The coefficient for '2nd class' is more modest (though still negative). All of this concurs with our earlier examination. The Z values (estimate / std. error) are all negative and generally of high magnitude, pointing to their high significance in the model. The higher negative value for 'male' than for that of '3rd class' is interesting – again showing the prime significance of this variable. Finally, the p values are all close to zero, again showing that each of the variables has a significant impact on survival. This is also confirmed by the triple asterisks assigned to each of them by the software. A glance at the confidence intervals for the coefficients (CI = 95%), and the odds appears to concur with these findings. The values for the odds are all significantly less than one, again showing the strong negative effect of the 'male' and '3rd class' co-variates on the response variable.

```
> BETA
                betaLB       beta     betaUB
(Intercept)  1.613686  1.9506533  2.2876205
PClass2nd   -1.172811 -0.7894247 -0.4060388
PClass3rd   -2.387082 -2.0391172 -1.6911526
Sexmale     -2.741946 -2.4453990 -2.1488521
```

Figure 2 – Confidence Intervals for coefficients

```
> exp(BETA)
                betaLB       beta     betaUB
(Intercept) 5.02128549 7.03328058 9.8514685
PClass2nd   0.30949584 0.45410596 0.6662843
PClass3rd   0.09189748 0.13014356 0.1843070
Sexmale     0.06444482 0.08669154 0.1166180
```

Figure 3 – Confidence Intervals for Odds

To investigate the effect of low age on ones' chances of survival, the subset of the data with the age details (756 passengers) was isolated. The age data was re-coded so that passengers whose age was 16 or less were coded with a '1' whilst the older passengers were coded with a '0'. The 'age' variable was then coded as a factor. The glm() command was used to fit the model, again with 'Survived' as the response variable, and 'Age' as the predictor variable. The coefficients obtained (**figure 4**) show a strong positive effect on survival. In otherwords, being aged 16 or under meant that one was significantly more likely to survive. When this model was adjusted to include 'Class' and 'Sex' (**figure 5**), the results were broadly the same, with age again having a positive effect on survival, and class and sex having a significant negative impact on survival (>2 for 3rd class and males).

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.45650    0.07911  -5.770 7.91e-09 ***
Age1         0.97376    0.24031   4.052 5.07e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4 – Coefficient data – Age only

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.1033     0.2204   9.542  < 2e-16 ***
Age1          1.3765     0.3017   4.562 5.07e-06 ***
PClass2nd    -0.9074     0.2372  -3.825 0.000131 ***
PClass3rd    -2.0702     0.2423  -8.544  < 2e-16 ***
Sexmale      -2.6117     0.2003 -13.040  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5 – Coefficient data – Age, Class, Sex

Returning to the original model of the full dataset with 1313 passengers, it is interesting to note how survival is impacted by the interaction between passenger class and sex. The relationship between being in 3rd class and being male is especially significant.

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.7006     0.3443   7.843 4.41e-15 ***
PClass2nd          -0.7223     0.4540  -1.591    0.112
PClass3rd          -3.2014     0.3724  -8.598  < 2e-16 ***
Sexmale            -3.4106     0.3793  -8.992  < 2e-16 ***
PClass2nd:Sexmale  -0.3461     0.5274  -0.656    0.512
PClass3rd:Sexmale   1.8827     0.4283   4.396 1.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6 – Interactions between Class and Sex

In order to assess the goodness of fit of the original model (**figure 1**), the Pearson residuals were plotted. At the top left of **figure 7**, there are a number of observations where the difference between the observed and the predicted value is quite large, and where the residuals are greater than 3. In other words, these observations have not been modelled very well. However, the difference between these observations and the others is not very large. As we are dealing with a large number of observations, the existence of these is to be expected.

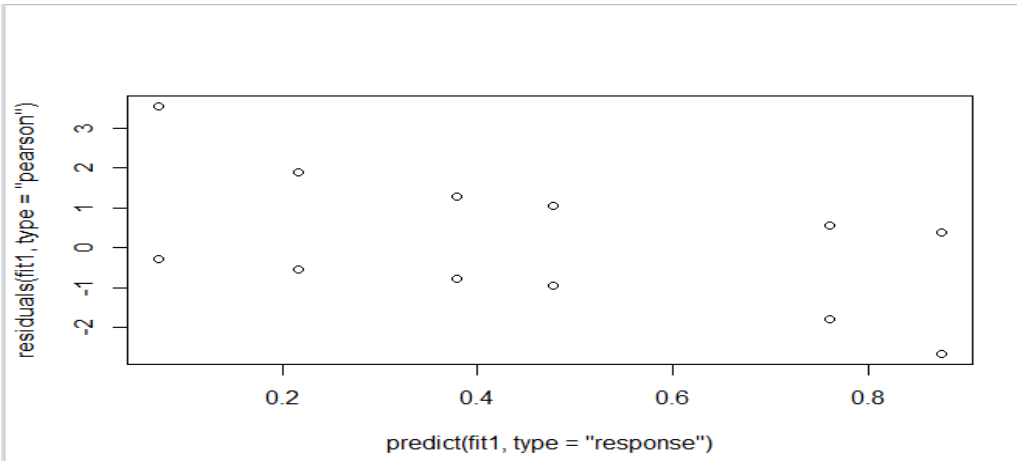

Figure 7 – Pearson Residuals for the original glm model

Finally, the predictive performance of the original model (from **figure 1**) can be examined used the ROC curve. This plots the false positive rate against the true positive rate. The curve lies entirely above the 45 degree line (simulating a coin toss approach to predicting survival). The area under the curve (**AUC**) comes to 0.8256. Both of these suggest a good degree of effectiveness for predicting the survival of the passengers.

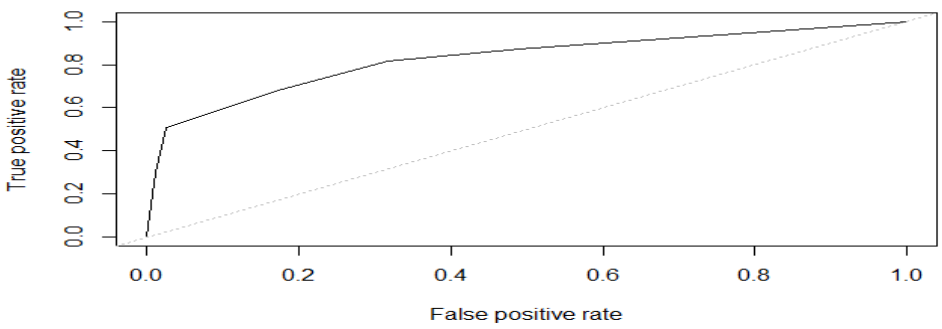

Figure 8 – ROC curve for the original glm model