1 Introduction

We consider data on the survival of passengers of the RMS Titanic. The data contains many variables about the passengers of the Titanic, some of which we will use and some of which we will not. We are going to perform logistic regression in order to ascertain the effect that certain variables have on the survival of a passenger. We have 1309 individual passenger records to analyse.

We are particularly interested in the affect of socio-economic status on survival probabilities, as well whether the adage of 'women and children first' holds.

2 Statistical Analysis

2.1 Introduction to our data

The survival variable is binary (they either survived or died). This is the reason for using simple logistic regression. The passenger class is a factor variable with three levels corresponding to 1st, 2nd, and 3rd class. The sex is a factor variable with two levels corresponding to male and female. The age is a continuous variable (with rounding applied). We have data on the sum number of siblings and spouses aboard of a passenger, as well as of the number of parents and children. These are highly correlated, so we will only use one. In our case we will use the siblings and spouses. We have the ticket number, which is useless for our analysis. We have the fare paid by the passenger. This is continuous and has properties that warrant a discussion later. We have the cabin number, which we will use to ascertain whether a passenger had a cabin (this is highly correlated with passenger class). We have the port of embarkation, which is one of Cherbourg, Queenstown, and Southampton. We have the lifeboat number, which is useless, as well as the body number. These are both useless as they are a posteriori of the survival. We also have the home/destination of the passenger. This is also useless.

This time around our data has no missingness that is not appropriate. Moreover none of the variables that we are analysing have missingness (the NA in cabin number means that the passenger did not have a cabin).

2.2 Initial Analysis

We first need to conduct an analysis of the independence of the variables we think we may include in our model. Performing a chi squared test we find that most variables of interest for the model are highly interdependent. However the type of model that we are fitting is somewhat resilient to this. What is important is that the variables are not co-linear. Most of our variables cannot be co-linear as they are factor variables. The only ones of concern are the sum number of siblings and spouses and the sum number of parents and children. We calculate the variance inflation factors for these variables. They come out as 1.24 and 1.23 respectively. The standard threshold for multicollinearity is 3, so we are very safe.

Looking at the fare paid by the passenger it is clear that in our data the fare paid is that of the ticket purchase, not per person. This means that those travelling in large groups will seem to pay a very large fare, whilst in reality no paying quite so high a price. Couple this with the fact that the fare is inextricably linked with both the passenger class and whether the passenger received a cabin and we are safe to remove this factor from our model.

We will also note that there are some mistakes in the data; some ages are recorded incorrectly. As there are thousands of records we hope that the correct values will swamp the effect of the incorrect values. No attempt will be made to fix the data, as there is no reasonable way to do so in the time-frame given for this analysis.

The data that we have been provided has been subject to imputation. This was particularly heavy in the age variable, with 263 values being imputed. As we have only the one dataset we cannot bring to bear many of the tools that are available to deal with the uncertainty brought about by missing data imputation. We must simply keep in mind that the statistics for age will be artificially more certain.

2.3 Model Design

We are going to use a logistic regression with the response variable being the logit probability of survival. This means that the response variable is of the form $y = \ln(P/1-P)$, where P is the probability of survival. We do this as we are modelling a categorical variable, so we want a response that is close to categorical. We can simply round the response to get an indicator as to whether the passenger would likely survive or perish.

We create the model with the predictors being age, sex, class, port of embarkation, whether or not the passenger had a cabin, and the sum number of siblings and spouses the passenger had aboard (hereon referred to as sibsp).

The reason for excluding the sum number of parents and children of the passenger is that it is extremely closely interlinked with the age and sibsp. It does not add much to the model when included as the corresponding coefficient is not statistically significantly different from 0.

We could include a parsed title of the passenger as a factor variable, however the overwhelming majority of the passengers go by the standard three (Mr., Mrs., Miss.). This means that this would be highly susceptible to outliers, as well as being completely dependent on sex.

We do not include fare as it is completely dependent on class and the number of people travelling in a certain party. An example would be the wealthy Sir. Duff Gordon's fare being recorded as 56.93 pounds, and the 3rd class passenger Mr. Frederick Sage having his fare recorded as 69.55 pounds. This is due to his travelling in a party of 11. This means that this should be accounted for by passenger class and sibsp.

We make the cabin into a binary variable. The idea for this is that people with a cabin are likely to have easier access to the decks to evacuate. We note that not all 1st class passengers had cabins, and not all cabins were 1st class. This makes it quite the interesting predictor.

3 Discussion of Results

Graphs showing the predicted survival by age with variable class and sex are given in Figure 1. We see that women had a far greater chance of survival than men of the same class. In fact 3rd class women had a better survival chance than 1st class men. This, coupled with the decreasing survival probabilities with age, points to 'women and children first' very much being a thing.

More explicitly we have that the coefficients for the model are found in Table 1. Note that these effect the logit of the probability (as given above), so we need to invert this in order to get the predicted probability of survival. The predicted probability of survival is given by $P_{surv} = \exp(y)/(1+\exp(y))$, where y is our response variable.

3.1 Effect of Age

4 Conclusions

A Plots and Figures

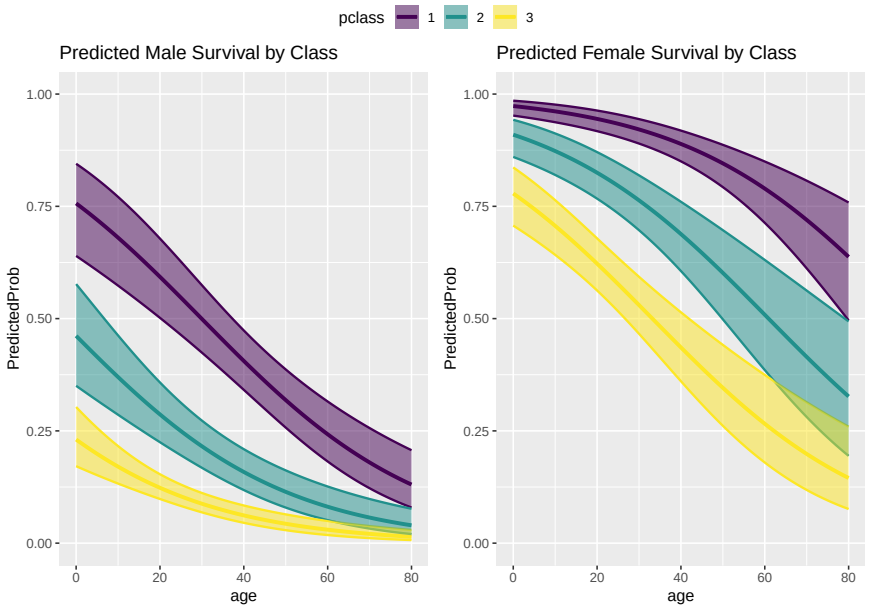

Figure 1: Predicted survival probabilities with 95% prediction intervals

Figure 2: Visualisation of data relating to gas consumption

Figure 3: Model Q-Q plots for the transformed and non-transformed models

Figure 4: Model diagnostic plots for our full model

Figure 5: Model diagnostic plots for our sub-models

B Tables

	term	estimate	std.error
1	(Intercept)	3.79	0.41
2	pclass2	-0.61	0.29
3	pclass3	-1.61	0.30
4	sexmale	-2.59	0.16
5	age	-0.05	0.01
6	sibsp	-0.37	0.09
7	embarkedQ	-0.50	0.30
8	embarkedS	-0.56	0.19
9	cabinY	0.88	0.26
10	parch	-0.04	0.09

Table 1: Regression coefficients for the logistic fit