Analysis of The 1912 Titanic Accident Dataset

Edwin Goh

October 2021

1 Abstract

This report analyses the passenger survival data from the 1912 Titanic ship accident using the *titanic* dataset given in the R's *Methods of Statistical Model Estimation* (*msme*) package. In this analysis, passengers are being classified into 3 different demographic variables and an attempt will be made in relating these variables to the likelihood of a passenger surviving the accident. A logistic regression, together with a random intercept, model will be used in the analysis. This analysis may provide insights to passenger behaviour and vulnerable sections of the ship, possibly highlighting potential loopholes in ship safety.

2 Introduction

The 1912 Titanic ship accident was one of the most catastrophic accidents in maritime history, with more than 1,500 deaths out of an estimated 2,224 people on board. Hence, the analysis of passenger survival data from this accident is vital in understanding factors that could have contributed to the likelihood of whether a passenger would have survived or otherwise. In addition, this analysis may also provide some insights into passenger behaviour during an emergency and highlight areas of ship safety that needs to be enhanced.

3 Data

The Titanic passenger survival dataset from R's msme package contains survival data of 1,316 passengers. The data includes 4 categorical variables, 3 of which are binary variables and the last being a factor variable with a total of 3 levels. Namely, the variables are:

1. survived: 0 = died; 1 = survived,
2. age: 0 = child; 1 = adult,
3. sex: 0 = female; 1 = male,
4. class: 1 = first class; 2 = second class; 3 = third class

For this analysis, the response variable is the *survived* binary variable while the other 3 variables are explanatory variables.

Before analysing and fitting the data to a statistical model, it would be beneficial to have a look at the distributions of all 4 variables which are shown on the following page:

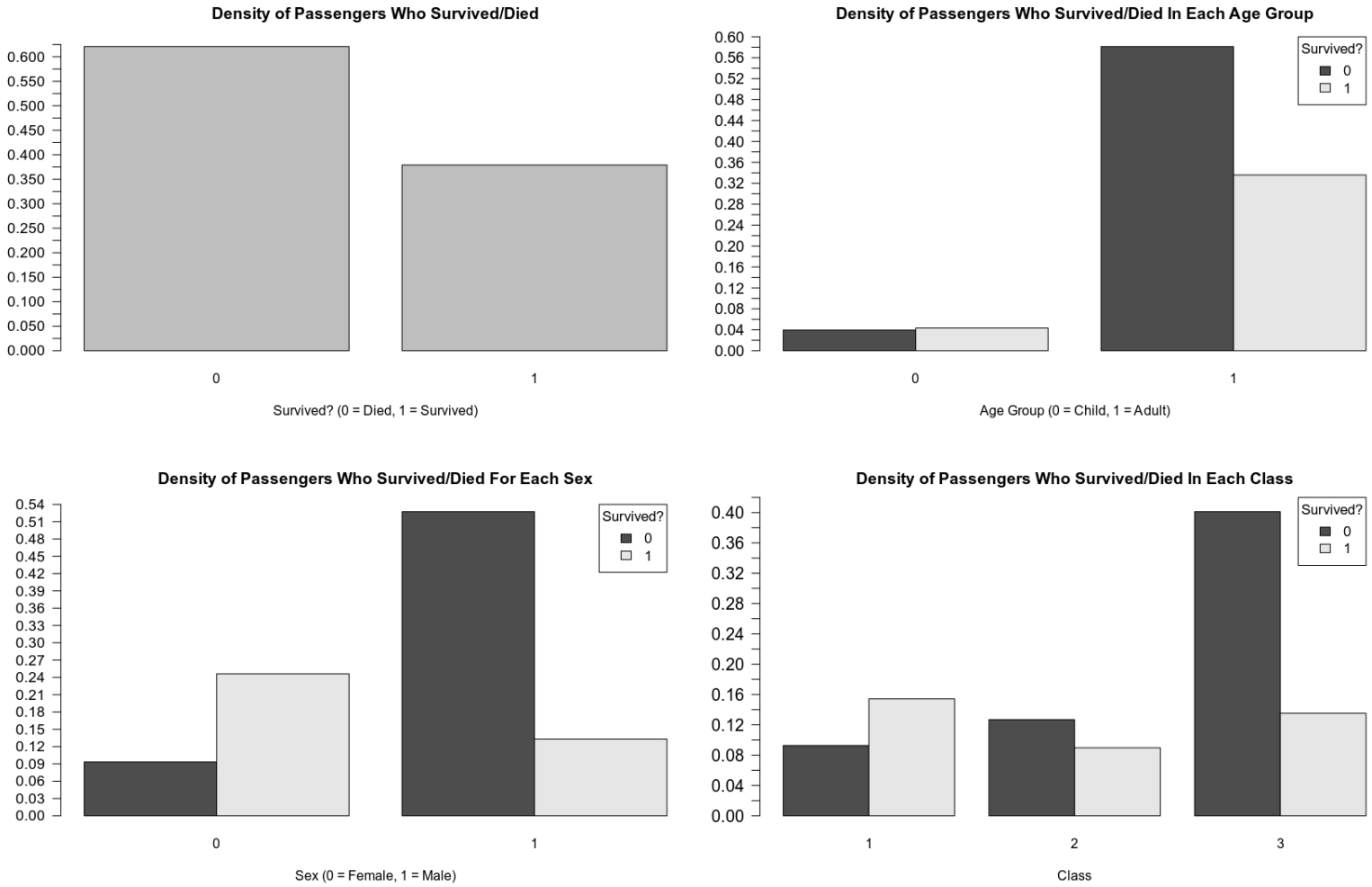

Figure 1: Distributions of All 4 variables: First Bar Plot - Response Variable, Subsequent Bar Plots - Explanatory Variables Classified In Terms of Both Values of The Response Variable

A general trend can be observed for each of the above bar plots. In the first plot, it can be seen that there are more passengers who did not survive. The second plot clearly shows that there are significantly less children aboard, and the proportion of child survivors is marginally more than those who did not while there are more adults who did not survive. From the third plot, it can be observed that the total number of male passengers is almost twice of the female population on board. However, there is larger proportion of female survivors compared to that of their male counterparts. In the final plot, there is an increasing proportion of passengers who did not survive through the classes, where third class passengers contributed the largest proportion of fatalities.

4 Model

Since the response variable is in a binary format, a logistic regression model, whose samples are drawn from a Bernoulli probability distribution, would be the most appropriate in this analysis. In addition, a random intercept model was used since an intercept in this model would imply that the passenger is a girl.

4.1 Fitted Model

$$
\begin{aligned}
Y_i|\phi_i &\overset{\text{ind}}{\sim} \text{Bern}(\phi_i)\,, \text{where } i = 1, 2, ..., 1316 \\
\text{logit}(\phi_i) &= \alpha_{\text{class}_i} + (\beta_1 \times \text{age}_i) + (\beta_2 \times \text{sex}_i)\,, \text{where class}_i = \{1, 2, 3\} \\
\alpha_j &\overset{\text{iid}}{\sim} \text{N}\left(\mu, \frac{1}{\tau^2}\right), \text{where } j = 1, 2, 3 \\
\mu &\sim \text{N}\left(0, 10^6\right) \\
\tau^2 &\sim \Gamma(1, 25) \\
\beta_k &\overset{\text{iid}}{\sim} \text{N}\left(0, 10^4\right), \text{where } k = 1, 2
\end{aligned} \tag{1}
$$

4.2 Alternative Model

$$
\begin{aligned}
Y_i|\phi_i &\overset{\text{ind}}{\sim} \text{Bern}(\phi_i)\,, \text{where } i = 1, 2, ..., 1316 \\
\text{logit}(\phi_i) &= \alpha_{\text{class}_i} + (\beta_1 \times \text{age}_i) + (\beta_2 \times \text{sex}_i)\,, \text{where class}_i = \{1, 2, 3\} \\
\alpha_j &\overset{\text{iid}}{\sim} \text{N}\left(0, 10^6\right), \text{where } j = 1, 2, 3 \\
\beta_k &\overset{\text{iid}}{\sim} \text{N}\left(0, 10^4\right), \text{where } k = 1, 2
\end{aligned} \tag{2}
$$

4.3 Model Evaluation

The simulation for each model was initialised with 3 Markov Chains and an initial burn-in of 1,500 iterations was conducted before Monte Carlo samples were collected. A total of 10,000 samples per chain was collected for each simulation.

Analysing the trace plots and other convergence diagnostics of both simulations, it can be concluded that the chains in both simulations have converged to their respective stationary distributions. The Deviance Information Criterion (DIC) were almost identical for both models, with a slight difference in the penalties for the complexity of both models.

The mean residuals for both models were in orders of magnitude of -4 to -5, *i.e.* very close to 0. The mean of the squared residuals for both models were approximately 0.155.

Despite both models being almost identical in terms of their convergence and residual analyses, the model with its random intercept α_j drawn from a normal distribution with variable mean μ and variance $1/\tau^2$ was chosen. The variable mean and variance adds another level of randomness to the model, which is ideal, without adding significant penalty to its model complexity.

5 Results

Using the fitted model, predictions can be made to determine whether a passenger from a particular population will survive the accident or otherwise. Before such predictions can be made, a threshold for the probability of survival needs to be set, above which the passenger in question would be classified as a survivor. From the safety perspective, it would be more desirable for the model to predict the passenger to be a fatality when he/she is in fact a survivor. For this reason, the threshold was chosen to 0.6, where the classification accuracy was calculated to be approximately 0.787. With this threshold, it can then be determined whether a passenger in a particular class and population

would have survived. The mean probability of survival for a passenger in each class, gender and age group is shown in a table below:

Class	Girl	Boy	Woman	Man
First	0.954 (1)	0.881 (1)	0.663 (0.844)	0.410 (0)
Second	0.883 (1)	0.730 (0.999)	0.420 (0.004)	0.203 (0)
Third	0.782 (0.999)	0.560 (0.108)	0.255 (0)	0.106 (0)

Table 1: Mean Probability of Survival Across Classes, Genders and Age Groups.
The Probability of Survival > 0.6 Are Given Within The Brackets

From the table above, it can be seen that children, being the most vulnerable group of passengers, have the highest chances of survival other than a boy in third class, who is predicted to have less likely survive. This indicates that this vulnerable population was well protected, probably, by their parents or guardians. It is also observed that women, being the second most vulnerable population, have a higher probability of survival, especially those in first class, compared to men. A possible reason for the significantly lower probability of men surviving could be the fact that men are more willing to sacrifice their lives for their loved ones.

6 Conclusion

As discussed in the previous section, it would be preferred for the fitted model to predict more fatalities than the actual number of fatalities. Hence, a threshold of 0.6 was chosen for the minimum probability of survival above which a passenger would be classified as a survivor. Using this threshold, it can be postulated whether a particular population in a particular class would be able to survive the accident. It is believed that vulnerable populations were much more protected as a result of the sacrifices of their loved ones. It can be seen in Figure 1 that significantly more fatalities are observed in third-class as compared to the other 2 classes. A similar story is also painted in Table 1 where a passenger in third class is predicted to have a lower probability of survival. This would imply that safety measures need to be enhanced to better protect passengers in third class. Despite children having the highest predicted probabilities of survival, it is important not to neglect this population in the consideration of safety enhancements.

Despite its relatively decent performance, the fitted model was based on a few assumptions that might not be entirely true. It was assumed that the probability of a passenger's survival is independent to that of another passenger. However, as discussed previously, since it is believed that quite a significant proportion of passengers could be related to other passengers, the previous assumption might no longer be valid. In addition, the Titanic passenger survival dataset from R‘s msme package is imbalanced in terms of the distributions of the 3 explanatory variables. This is especially so for the age variable, where there are significantly more adults than children. Similarly, though to a smaller extent, there are almost twice as many male passengers as there are female passengers aboard. The imbalanced dataset have caused the fitted model to be skewed towards adult and male passengers. In other words, the posterior estimates of the β's parameters tend to be biased towards this particular population of passengers. Furthermore, as a result of the uneven distribution of passengers over all 3 classes and their different locations on the Titanic, the posterior samples of the intercept for each of the 3 classes cannot be considered as independent and identical (iid) to one another.