Final Project

Summary

The RMS Titanic was a passenger ship carrying an estimated 2,224 people on its maiden voyage from Southampton to New York City. It sank on 15th April, 1912 after a collision with an iceberg. In this study we aim to quantify the effects of various factors that influenced the survival of a passenger. We find that the odds of death for women were 9.65 times higher compared to men when other factors were held constant. Secondly, probability of death decreased with age; with older people having a higher chance of surviving after accounting for other factors. Lastly, passengers who embarked in Cherbourg had the highest probability of death, followed by Queenstown and finally Southampton across all genders and social classes.

1 Introduction

In this analysis we aim to understand and quantify the effects of various factors that contributed to the survival of a passenger aboard the Titanic. The outcome variable is binary; indicating if the passenger survived the incident. This can be modelled using a generalized linear model with the logit link function, also referred to as logistic regression. To provide uncertainty estimates for our analysis, we use bayesian logistic regression. Additionally, we hypothesize that passengers who boarded the ship together are correlated with each other, implying a natural "grouping" in the dataset. To account for this correlation, we use a hierarchical variant of logistic regression.

2 Data

The dataset consists of 891 data points. Table 1 has a data dictionary that describes each feature. Certain features such as `PassengerId`, `Ticket`, `Fare`, `Name` and `Cabin` were excluded as they did not have a strong explanatory signal and were not required for a good first model. Rows with missing values were omitted (leaving us with 712 data points).

Feature	Type	Description
`PassengerId`	Numeric	
`Survived`	Factor (2 levels)	Target Variable
`Pclass`	Factor (3 levels)	
`Name`	String	
`Sex`	Factor (2 levels)	
`Age`	Numeric	
`SipSp`	Numeric	Number of siblings or spouses aboard
`Parch`	Numeric	Number of parents or children aboard
`Ticket`	String	
`Fare`	Numeric	
`Cabin`	String	
`Embarked`	Factor (3 levels)	Location where journey was started

Table 1: Each row corresponds to the characteristics of of one passenger and if they survived the incident

After visualization the dataset, we made the following observations:

- There were more survivors ($n = 424$) than people who died ($n = 288$). Amongst the people who died, there were a lot more females (68 %) than men. We can hypothesize that `Sex` could be an important explanatory feature.
- The survival rate was lowest for children (54.1 %), higher for young adults (60.5 %), even higher for adults (60.9 %) and highest for old people (80.9 %)[1]. This might be because older people were evacuated first.

[1] Ages were divided into four equal bins using the `cut` function

- Distributions of `Parch` and `SibSp` (indicating family size) are heavily left skewed with a median value of 0 for both and a 75 % quantile of 1.
- Age is approximately normally distributed with a mean of 29 with a slight left skew.

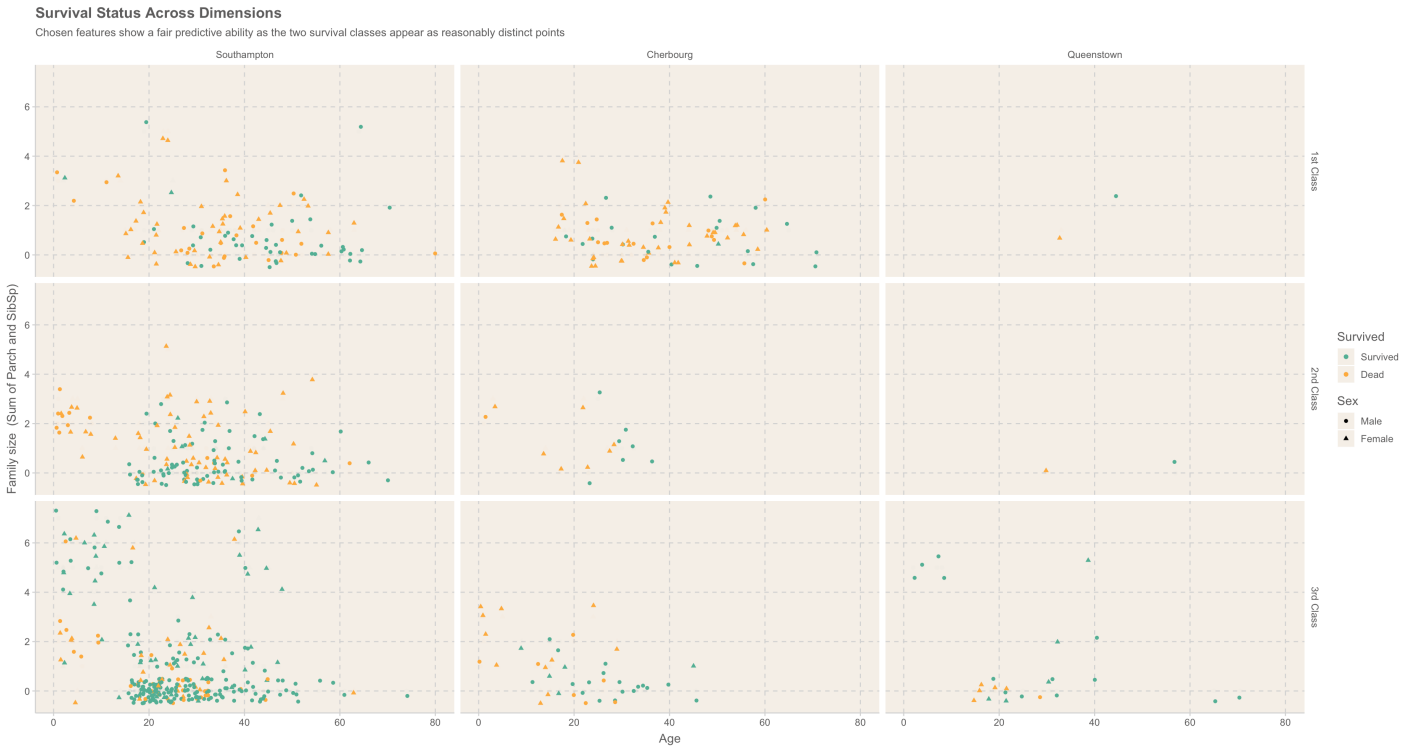


Figure 1: There is a high survival rate amongst passengers (particularly men) who embarked from Southhampton and from the third class

- We visualized all the dimensions in a single plot in figure 3. There are several observations to be made here. There seem to be a lot more people who embarked in Southampton compared to Cherbourg and Queenstown.
- Survival rate seems to differ significantly for different socio-economic classes (`Pclass`). It is highest for people in the third class, followed by second and first. This can be visualized be examining the different rows for different regions of embarkment (bottom row has the most relative green points).
- Specifically for people who embarked at Southampton and belonging to the third socio-economic class, a higher family size combined with a low age corresponds to a higher survival rate. Additionally, for the same group, a higher age with a small family size also corresponds with a higher survival rate. This also appears to be true for passengers who embarked at Cherbourg and Queenstown (for the third socio economic class).
- Overall, we do see that the classes (survived and dead) are fairly distinct after accounting for all the different factors.

3 Model

The dataset is a collection of ($n = 712$) pairs of points $\{x_i, y_i\}$ where $x_i \in \mathbb{R}^8$ and y_i is binary $\{0, 1\}$. We use the following models for our analyses.

3.1 Simple Logistic Regression

$$y_i \mid \phi_i \overset{\text{ind}}{\sim} \text{Bern}(\phi_i) \qquad i = \{1, .., n\}$$
$$\text{logit}(\phi_i) = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_d \cdot x_{di} \qquad \left(\text{where logit}(\phi) = \log \frac{\phi}{1-\phi}\right)$$
$$\beta_j \mid \mu, \tau \sim \mathcal{N}(\mu, \tau^2) \qquad j = \{1, ..., d\} \text{ and } \mu, \tau \text{ are hyper-parameters}$$

3.2 Hierarchical Logistic Regression (Random Intercept Model)

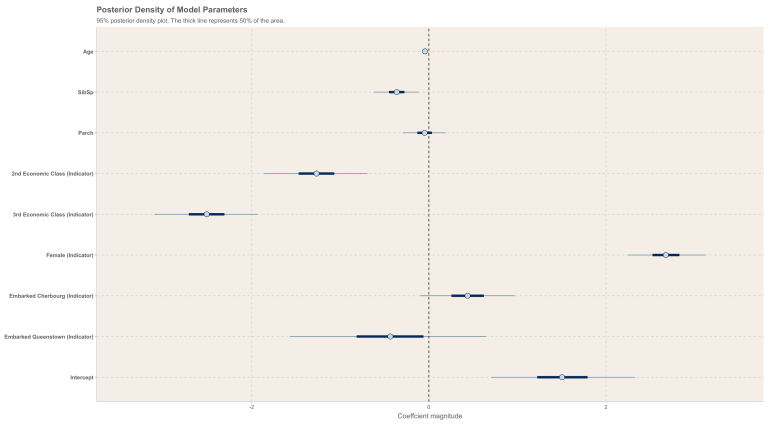
$$\begin{aligned} y_i \mid e_i, \phi_i &\sim \text{Bern}(\phi_i) && i = \{1, ..., n\} \quad e_i = \{1, 2, 3\} \\ \text{logit}(\phi_i) \mid \alpha_{e_i}, x_i, \beta &= \alpha_{e_i} + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + ... + \beta_d \cdot x_{di} && e_i = \{1, 2, 3\} \\ \alpha_{ei} \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau^2) && e_i = \{1, 2, 3\} \\ \beta_j \sim \mathcal{N}(a, b^2), \quad \mu \sim \mathcal{N}(c, d^2), \quad \tau &\sim \text{Gamma}(e, f) && j = \{1, ..., d\} \end{aligned}$$

4 Results

4.1 Convergence

We sampled 50,000 points from the posterior distribution of both models and assessed the *trace plots*, *effective sample size* and *gelman diagnostics* for convergence. The simple logistic regression model had a *penalized deviance* of 650.7 and the hierarchical model had a deviance of 650.6. Since both the models had a similar fit, we carried out further analysis using the simpler model. [2]

4.2 Inference



Coefficient	Posterior Mean	Posterior Std. Dev.
`Intercept`	1.50	0.417
`Age`	-0.04	0.008
`SibSp`	-0.364	0.129
`Parch`	-0.048	0.1233
`2nd Economic Class` (Indicator)	-1.27	0.298
`3rd Economic Class`	-2.511	0.296
`Female` (Indicator)	2.678	0.224
`Embarked Cherbourg` (Indicator)	0.436	0.272
`Embarked Queenstown` (Indicator)	-0.442	0.5625

Figure 2: The thin line represents a 95 % HPD posterior credible interval and the thick line is a 50 % HPD credible interval

If we denote the probability of death by ϕ, the fitted model (using posterior mean estimates) is given by

$$\log\left(\frac{\phi}{1-\phi}\right) = 1.50 - (0.04 \cdot \texttt{Age}) - (0.364 \cdot \texttt{SibSp}) - (0.048 \cdot \texttt{Parch}) - (1.27 \cdot \texttt{Economic.Class.2nd}) + (2.678 \cdot \texttt{Female}) - (2.511 \cdot \texttt{Economic.Class.3rd}) + (0.436 \cdot \texttt{Embarked.Cherbourg}) - (0.442 \cdot \texttt{Embarked.Queenstown})$$

[2] Occam's razor

- Here, the intercept of 1.50 represents the log-odds of death for the reference levels which are males of `Age, SibSp, Parch = 0` and the 1st class who embarked from Southampton. The odds of death are therefore, $e^{1.50} = 4.48$ but this does not have a meaningful interpretation as there are no people with `Age = 0` in the dataset.
- The coefficient of `Female` is high positive value away from zero, indicating a *higher* probability of death for females compared to males when all other variables are held constant. Specifically, the odds of death for females are $e^{2.678} = 9.65$ times *higher* compared to men after accounting for other variables.
- Meanwhile `Age` seems to have a smaller effect, but the negative sign indicates that as we *increase* age, the probability of death *decreases.* Specifically, with every unit increase in age, the odds of death *decrease* by a factor of $e^{-0.04} = 0.96 \approx 1$ when all other factors are held constant.
- The coefficients of the economic classes are also very significant. Specifically, the odds of death for passengers in the 2nd economic class are lower by a factor of $e^{-1.27} = 0.28$ compared to passengers in the reference (1st class) and for the 3rd economic class they are lower by a factor of $e^{-2.51} = 0.08$ with **all else equal.**
- The coefficient for `Parch` has a lot of probability mass over zero, indicating that it is potentially not statistically significant given our model.

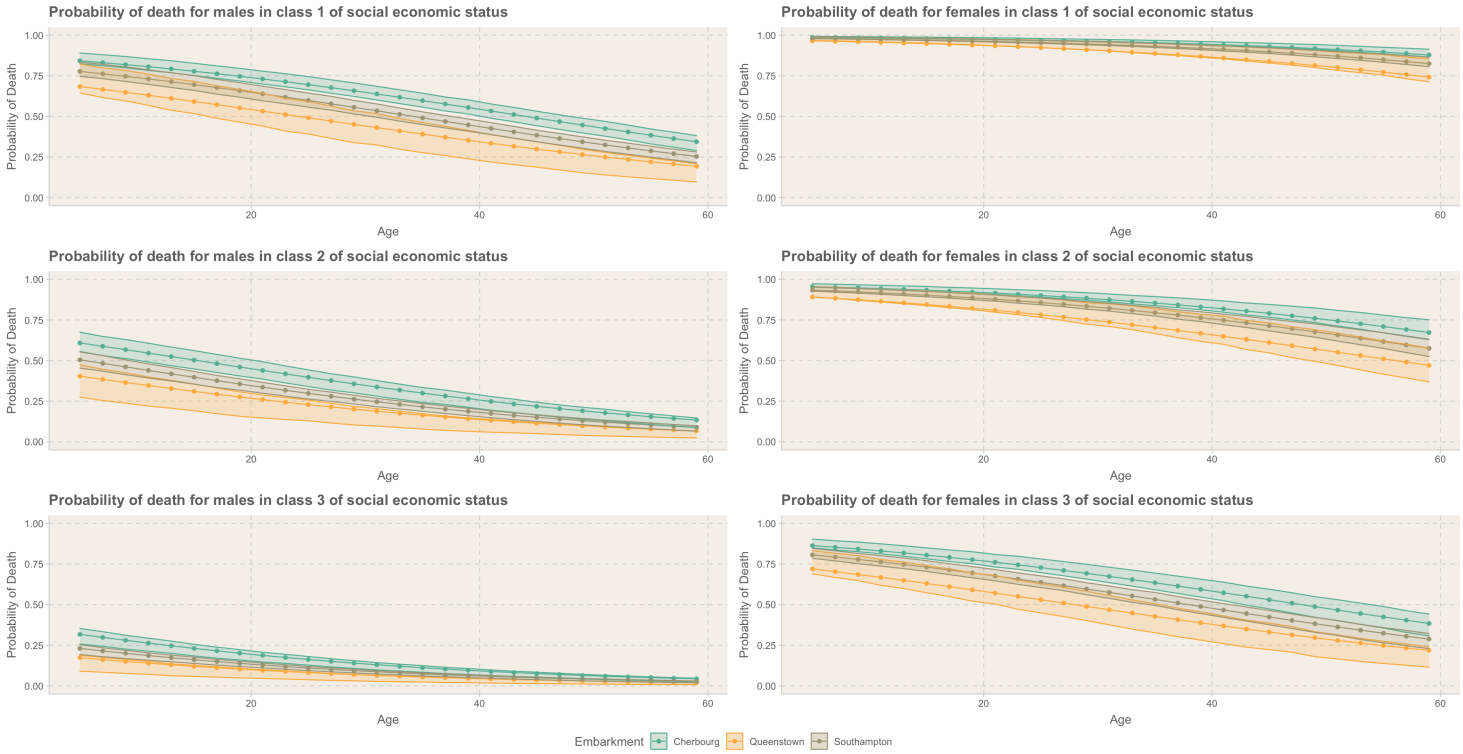


Figure 3: We visually represent $P(\text{death})$ for a fixed value of `SibSp`$= 0$, `Parch` $= 0$ [3], varying `Age`, `Sex` and social class. The shaded area represents a 50% credible interval [4] over the posterior predictive distribution. (i) Males have a lower probability of death across all age groups as compared to females. Furthermore, (ii) the probability of death is highest for people who embarked in Cherbourg followed by Queenstown and finally Southampton. Finally, (iii) probability of death reduces as age increases

5 Conclusions

Thus, we have studied various factors that account for the survival of passengers aboard the Titanic. However, several improvments are possible. Our models do not have any interaction or polynomial terms and could benefit from them. We could also try using alternative priors and carry out a prior sensitivity analysis. Additionally, we could try out alternative hierarchical models, perhaps based on factors other than the port of embarkment for a better fit.

[3] Median values

[4] The choice of 50 % was arbitrary and mainly for figure aesthetics. A higher value of 95 % results in wider overlapping intervals