Lost and Logistic: Navigating Titanic's Romance with Probability

Project by Arijit Dey as part of fulfilling the requirements for the Coursera course

August 2023

1 Introduction

The tragic sinking of the RMS Titanic on April 15, 1912, remains one of the most infamous maritime disasters in history. The luxury ocean liner, deemed 'unsinkable', struck an iceberg on its maiden voyage, leading to the loss of over 1,500 lives. Beyond its historical significance, the Titanic disaster has also captured the world's imagination through books, films, and documentaries. In this project, we embark on a data-driven journey to unravel the mysteries of survival aboard the Titanic using logistic linear regression.

2 Objective

The primary objective of this project is to employ statistical analysis to predict the probability of survival for Titanic passengers based on key factors such as sex, age, ticket fare, ticket class, and number of companions. By leveraging logistic linear regression, we aim to unveil patterns within the data that shed light on the various variables that influenced the passengers' chances of survival.

3 The Data

A wealth of data is accessible on the internet concerning the details of Titanic passengers, intended for the purpose of data analysis. We obtained the data from Kaggle. The data set comprises two distinct sets of data: one for training and the other for testing purposes, divided randomly. The training data set encompasses information about 891 passengers, while the testing data set comprises details of 418 passengers. Both data sets consist of 12 columns representing diverse passenger attributes. Refer to Table 1 for a breakdown of these details.

Variable	**Description**	**Variable**	**Description**
PassengerId	Unique ID for each passenger	Survived	Survived? 0 = No, 1 = Yes
Pclass	Ticket class: 1st, 2nd, or 3rd	Name	Name of the passenger
Sex	Sex: Male or Female	Age	Age in years
SibSp	Number of siblings/spouses aboard the Titanic	Parch	Number of parents/children aboard the Titanic
Ticket	Ticket number	Fare	Passenger fare (British pounds)
Cabin	Cabin number	Embarked	Port of Embarkation

Table 1: Column description of the data set obtained from Kaggle

3.1 Data Processing

We apply a statistical model to visualize how various factors influence the survival of passengers. As predictors, we select six columns from the data set: Passenger Class, Gender, Age, Sibling/Spouse Count, Parent/Children Count, and Passenger Fare. We specifically focused on these columns and filtered out rows that contained NA values. We transformed the entries in the 'Sex' column, replacing 'Male' or 'Female' with numeric values, where 1 represents 'Male', and 0 represents 'Female'. Additionally, we standardized and centered the values of the covariates.

3.2 Data Exploration

We conducted a preliminary data analysis to examine the correlations among various columns in the dataset. Figure 1 (left panel) displays the survival proportions during the tragic event of 1912, indicating a notably higher proportion of non-survivors. The right panel of Figure 1 presents correlation values and plots among the predictors derived from the dataset. The findings reveal that, except for the predictors 'Passenger Class' and 'Ticket Fare', correlations among other covariates are relatively modest.

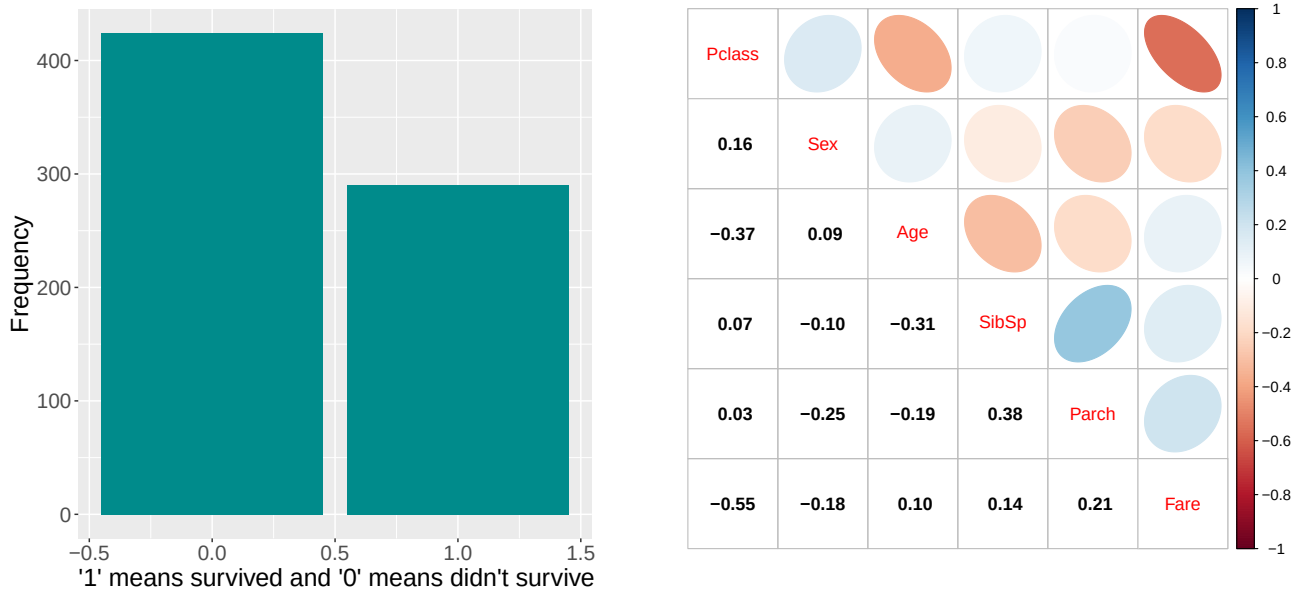

Figure 1: Bar chart depicting the proportion of survival (left panel), and a correlation plot illustrating the relationships among various covariates within the dataset (right panel).

4 Model Selection

It's important to note that our response variable is binary, taking the value 1 for survived passengers and 0 for non-survivors. Meanwhile, the predictor variables span the entire real number range. As a result, a Logistic Regression model is the most suitable choice for this scenario. Utilizing the available predictor variables, our objective is to construct a model for predicting the probability of passenger survival.

4.1 Logistic Linear Regression

Logistic regression is a powerful statistical technique used for predicting binary outcomes, making it particularly suitable for scenarios where we want to understand the relationship between a set of independent variables and the likelihood of an event occurring. In our case, we'll be using logistic regression to predict

the probability of survival for Titanic passengers based on their attributes, such as sex, age, ticket fare, ticket class, and number of companions.

The mathematical formulation of logistic regression involves transforming the linear combination of input features into a range between 0 and 1 using the logistic function. The logistic function models the probability of the binary outcome (in this project, survival or not) as a function of the linear combination of input features. Mathematically, the logistic function is defined as

$$P(Y=1) = \Big[1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P)\Big]^{-1} \tag{1}$$

where $P(Y=1)$ is the probability of survival (class 1); $\beta_0, \beta_1, \ldots . \beta_P$ are the coefficients associated with the intercept term and the input features $X_1, \ldots, X_P$, respectively.

4.2 Hierarchical structure and Prior specification

Our aim is to model the probability of a passenger's survival in relation to several predictors: the passenger's Class, Age, Sex, Sibling/Spouse count, and Parent/Children count, and Ticket fare. We introduce seven parameters, including one for the intercept term and one for each predictor value. When it comes to the prior specification, considering that the parameter values can encompass the entire real number line, we opt to use a Normal prior distribution with mean 0 and variance 100 for both the regression parameters and the intercept term. The hierarchical specification of the model is given as

$$Y_i \mid \phi_i \sim \text{Bernoulli}(\phi_i);\ i = 1(1)n$$

$$\text{Logit}(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \beta_0 + \sum_{p=1}^{6} \beta_p X_{ip};\ i = 1(1)n$$

$$\beta_i \sim \text{Normal}(0, 100);\ i = 0(1)6$$

5 Model Fitting

We employed the `rjags` package in R to fit the model. The model string was defined according to the details provided in 4.2. We conducted 1000 updates on the model to surpass the initial 'burn-in' phase. Subsequently, we collected samples from three chains, each comprising a length of 5000, for all seven parameters. To ensure result reproducibility, we performed all the aforementioned tasks with a seed value of 100 using the `set.seed()` function in R.

5.1 Convergence diagnostics

We have employed standard MCMC convergence diagnostics on our simulated samples. The trace plots of the six regression parameters and the intercept term exhibit satisfactory mixing, indicating favorable convergence. Additionally, we utilized the Gelman-Rubin statistic for convergence testing, yielding PSRF values of 1.01, 1.00, 1.01, 1.00, 1.00, 1.00, and 1.00 for the six regression parameters and the intercept term, respectively, further confirming robust convergence. Considering autocorrelation, even though the lag-1 autocorrelation is approximately 0.5, it diminishes to zero after 3-4 lags. The effective sample sizes are calculated as 3545.801, 7246.173, 4797.413, 6205.071, 5945.725, 4233.709, and 8404.084 for the quantities of interest.

5.2 Model Prediction Analysis

To assess the model's quality of fit, we conduct a predictive analysis using the fitted model. The test data used is sourced from the designated test data set within the Kaggle data set. For the 418 passengers

identified within the test data, we compute the survival probabilities based on the model's predictions. The obtained fitted results exhibit an accuracy of 91% when applied to the test data. This indicates a reasonable level of confidence in the validity of the fitted model.

6 Model Summary

With the confidence of good fit, we now present the model summaries and the estimates of the regression parameters along with the confidence intervals to quantify the estimation uncertainty. Table 2 provides the posterior summaries of the model parameters obtained form MCMC samples.

Parameter	**Mean**	**SD**	**Naive SE**	**Time Series SE**	**95% CI**
β_0	-0.51245	0.1015	0.0008285	0.001107	(-0.7116,-0.3155)
β_1	-1.04578	0.1402	0.0011450	0.002353	(-1.3190,-0.7695)
β_2	-1.28492	0.1065	0.0008696	0.001252	(-1.4968,-1.0791)
β_3	-0.64705	0.1193	0.0009742	0.001723	(-0.8877,-0.4155)
β_4	-0.35915	0.1198	0.0009783	0.001521	(-0.5954,-0.1286)
β_5	-0.05923	0.1064	0.0008687	0.001381	(-0.2735,0.1466)
β_6	0.13559	0.1416	0.0011562	0.002176	(-0.1238,0.4327)

Table 2: Posterior summary of the model parameters; β_0: parameter for the intercept term, β_1 parameter for the Class of passenger, β_2: parameter for Sex, β_3 parameter for Age, β_4: parameter for Number of siblings/spouse, β_5: parameter for Number of children/parents, β_6: parameter for Price of ticker.

7 Predicting Jack and Rose's Survival Probabilitie

With the model parameters now estimated, we can delve into discussing the survival probabilities of Jack, potentially addressing the debate surrounding who might have lived had certain events transpired differently. At the time of the tragedy, Jack Dawson was a 20-year-old passenger. He embarked on the ship without any family or spouse, purchasing his third-class ticket for 7.5 British pounds. On that fateful day, Rose Bukater was 19 years old, accompanied by her fiancé (considered as a spouse) and her mother. She obtained her first-class ticket at a cost of 512 British pounds.

Utilizing these details as covariates and applying the estimated parameter values, we find that for Jack's given covariate values, his survival probability was merely 0.12. In contrast, considering Rose's covariate values, her probability of survival was high at 0.98. Thus, although our emotions may have been stirred as Rose found refuge on the boat while Jack faced the waters, the unforgiving nature of statistical probability attests that events unfolded in alignment with the dictates of probabilistic laws.

8 Disclaimer

Commenting on a passenger's survival in this manner lacks ethical and logical soundness. Logistic regression stands as a potent tool in the realm of regression, adept at modeling binary outcomes alongside covariates. Our presentation of this project serves as an application showcasing the capabilities of logistic regression. However, we must exercise caution and refrain from treating the subject unseriously, as it pertains to the poignant reflection of a common sentiment: 'Jack might have survived'.