Hypothesis and Regression Models Application Strategy

  Lisa Hall

University of Cumberlands

MSDS-530-A01 - First Bi-term

Dr. Afshin Zarenejad

February 23, 2025

**Abstract**

This paper explores how applying logistic regression can predict specific outcomes, using the survival rate of Titanic survivors as an example. Using binary logistic regression, we will study how independent variables like gender, age, and passenger class influence the dependent variable represented by survival. Using R, we will identify the influence of these independent variables on the dependent variable of survival. This demonstration illustrates how statistical models can influence decision-making.

First, we will explore the Titanic dataset, cleaning, transforming, and preparing it for analysis. We remove missing values, convert categorical variables into numerical variables, and normalize; this will improve model accuracy. Following the transformation, the data will be fitted using a generalized linear model function in R (glm()). We set survival as our dependent variable and then used passenger class, gender, and age as independent variables. The outcomes will predict how passenger class, gender, and age affect the survival outcome of passengers aboard the Titanic.

**Table of Contents**

1. Introduction
2. Background of the Problem Under Investigation
3. Problem Statement
4. Literature Review
5. Analysis of the Data Collected
6. Presentation of the Data
7. Methodology
8. Summary of the Findings
9. Conclusion and Future Work
10. List of References
11. Certificate of Authorship

**Hypothesis and Regression Models Application Strategy**

Binary logistic regression models estimate the relationship between a binary variable and other nonbinary independent variables (Penman, 2022). In practical application, evaluating the relationship between a binary variable, for example, presence or absence, and other attributes like gender, age, or income could help estimate the effect those have on the binary variable. In this study, we will examine how independent variables that are not binary help estimate a binary variable.

**Background of the Problem Under Investigation**

For this study, we will evaluate the preloaded Titanic dataset in R (R Core Team, 2023). Using the Titanic dataset, we will predict the probability of the binary outcome by using other variables. This dataset was chosen for its ease of availability and its clearly labeled passenger class, sex, and age, which all may have influenced the survival of those passengers on the Titanic. Since the variable “survived” is binary, using a binary logistic regression model to estimate the relationship between these variables is appropriate. This type of analysis underscores how logistic regression can be a significant tool in risk assessment.

**Problem Statement**

The historical sinking of the Titanic has provided an excellent context for researchers to use socioeconomic factors to model predictive outcomes. Due to these factors, it is one of the most studied maritime disasters in human history (Encyclopaedia Britannica, n.d.). It has captivated the minds of researchers since it is happening, thanks to the records and personal accounts available. The primary problem explored is:

* Did passenger class, age, and gender affect the odds of survival aboard the Titanic?

The statistical problem to be solved is how passenger class, sex, and age influence how likely passengers are to survive or not survive aboard the Titanic. What was the most significant variable that determined survival is also of interest. We can write the expression with p as the likelihood of surviving, β1 as passenger class, β2 as gender, β3 as age, as seen in equation 1:

… [1]

We can also form hypotheses as follows:

* H0: Passenger class, age, and sex do not affect the odds of survival.
* H1: Passenger class, age, and sex affect survival odds.

The Titanic dataset helps demonstrate how statistical modeling can use historical data to help form insights. The application has broad applications; this could be applied in various contexts, such as risk management and social sciences. The problems created by socioeconomic disparities are highlighted in this study but can be seen in many others. We hope to encourage further research into this critical study area by exploring this dataset.

**Literature Review**

Binary logistic regression can be applied widely to the context of historical datasets; the Titanic dataset is instrumental. According to Encyclopedia Britannica (n.d.), socioeconomic disparities influenced the survival rates of passengers, particularly age, gender, and passenger class. We can provide an example of statistical analysis and predictive modeling that considers these factors by utilizing the Titanic dataset. Understanding social structures offers insights into how they could affect survival rates in other contexts.

Fávero and Belfiore (2018) explain how logistic regression can provide easy-to-interpret results. Model accuracy is a key notation in their work; this is important when evaluating the suitability of logistic regression. Harris (2021) further explains how logistic regression estimates event probabilities. This is important to demonstrate predictive utility when applied to the Titanic dataset. This work underscores how important validating models is and how to create metrics for accuracy and precision.

The Titanic dataset is provided by The R Core Team (2023) for use in the R environment.

The dataset is so widely reviewed that it is essential to have it as part of the library. Variables included in the dataset which represent social and economic factors are age, passenger class and gender, when paired with survival status it is an excellent dataset to practice modeling. The dataset's structure allows for efficient application and producing meaningful results.

The practical applications of logistic regression are explored by Penman (2022). The Titanic dataset demonstrates how logistic regression can uncover valuable patterns, which can later be compared to historical outcomes to produce other insights. Real-world scenarios are helpful examples of handling categorical data and continuous predictors. When applying logistic regression to historical analysis such as this provides valuable insights into current applications of this approach.

**Analysis of Data Collected**

When evaluating the Titanic data set, we cannot identify variables imperative to our analysis. Survived is a binary variable, either yes for survived or no for did not survive. Pclass represents a categorical variable, the passenger class, with 1st class representing the more expensive tickets purchased and 3rd class representing the least expensive tickets purchased. Sex is categorical, and passengers are either male or female. Age is a continuous variable, representing the total years of age of the passengers in the dataset.

The dataset was first prepared by removing missing values. Then, the categorical and binary variables were transformed into numerical values. The continuous variable was normalized; this helped to improve model accuracy. The dataset was further split into training and testing subsets. Because of the structure of the Titanic dataset, it was exemplary to utilize a binary logistic regression model. The dataset includes 891 rows of passenger data.

**Presentation of Data**

An excellent tool for analyzing the Titanic dataset is logistic regression. Logistic regression is used when there is a qualitative way to estimate the probability of an event occurring (Fávero & Belfiore, 2018). There are important steps to logistic regression: data preparation, removing unnecessary data and normalizing data; model fitting, building models; model interpretation, determining what the model tells us; and validation, codifying the model's results. These steps ensure that the model is applied in a way that best predicts survival.

The first step of data preparation requires loading data from the Titanic dataset into R. This dataset includes independent variables like passenger class, gender, age, and fare, and our binary variable survived. To properly fit the logistic model, categorical classes such as passenger class must be converted into a numerical value. Missing data should be removed to avoid affecting the model, as logistic regression does not handle missing values. Normalization could also be considered for age and fare to improve the model. This will also involve appropriately loading libraries in R, like Titanic and dplyr, training the data, and changing the necessary values to numeric values. The following R code will be used:

> library(titanic)

> library(dplyr)

> data("titanic\_train")

> titanic\_data <- titanic\_train %>%

+ select(Survived, Pclass, Sex, Age) %>%

+ na.omit()

> titanic\_data$Survived <- factor(titanic\_data$Survived, labels = c("No", "Yes"))

> titanic\_data$Sex <- factor(titanic\_data$Sex)

**Methodology**

With the data prepared, the model-fitting step utilizes the generalized linear model function in R. The glm() function is particularly useful because it allows for selecting a regression family; in this case, the family will be binomial. Survival will be the dependent variable, where the formula will take the form “survived ~ pclass + sex + age + fare.” After this model is run, it will produce the beta coefficients for each predictor. Predictors indicate the effect of each variable on survival. This is crucial for interpreting the influence of the variables on survival probability. The following R code and output represent the modeling step:

> logit\_model <- glm(Survived ~ Pclass + Sex + Age, data = titanic\_data, family = binomial)

> summary(logit\_model)

Call:

glm(formula = Survived ~ Pclass + Sex + Age, family = binomial,

data = titanic\_data)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 5.056006 0.502128 10.069 < 2e-16 \*\*\*

Pclass -1.288545 0.139259 -9.253 < 2e-16 \*\*\*

Sexmale -2.522131 0.207283 -12.168 < 2e-16 \*\*\*

Age -0.036929 0.007628 -4.841 1.29e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 964.52 on 713 degrees of freedom

Residual deviance: 647.29 on 710 degrees of freedom

AIC: 655.29

Number of Fisher Scoring iterations: 5

In the model interpretation phase, each coefficient is assessed using confidence intervals and p-values; very small p-values are considered significant (< 0.05). We can review the previous summary function used to evaluate the coefficient output. First, we can note that each variable has a small p-value, indicating that the outputs are significant. We can determine from this that passenger class, age, and sex are all highly significant when predicting survival. For passenger class, the predictor is –1.289, which means that as passenger class increases, the odds of survival decrease. Second and third-class passengers have less chance of survival. The significant negative value for gender, -2.522, indicates that male passengers have less chance of survival than females. For age, the negative value, -0.037, indicates that the chances of survival decreased for each additional year. We can, therefore, discard the null hypothesis. Survival is impacted by passenger class, age, and sex.

Model validation provides a better understanding of model performance and answers how effective a model is in its estimation. The predict() function is used in R, which can generate probability scores to help determine survival. Using a confusion matrix, we can summarize the results and then evaluate further to get a numerical representation of how effective the model is. In practice, it could be valuable to understand if this model could be used in other applications, not just the estimations in this dataset. The following R code generates a confusion matrix and then uses that output to generate the model's accuracy. It shows the overall efficacy of the model to be 79%.

> titanic\_data$predicted\_prob <- predict(logit\_model, type = "response")

> titanic\_data$predicted\_class <- ifelse(titanic\_data$predicted\_prob > 0.5, "Yes", "No")

> table(titanic\_data$Survived, titanic\_data$predicted\_class)

No Yes

No 356 68

Yes 83 207

> conf\_matrix <- table(titanic\_data$Survived, titanic\_data$predicted\_class)

> accuracy <- sum(diag(conf\_matrix)) / sum(conf\_matrix)

> print(paste("Model Accuracy:", round(accuracy, 4)))

[1] "Model Accuracy: 0.7885"

**Summary of the Findings**

The findings lend an understanding of how the independent variables described in this study affect the dependent variable of survival. Reviewing the results, we can determine that passenger class had the most significant effect on survival rate when the passenger had a third-class ticket. Historical accounts of survival and how passengers were compartmentalized on the ship align with the results. The first-class passengers were roomed closer to the ship's deck, making evacuation easier. The logistic regression model confirmed that passengers with more expensive tickets had significantly higher survival odds.

Gender played a role in survival; male passengers had a lower probability of survival than female passengers. The logistic regression model demonstrated that the male gender had a negative coefficient, demonstrating lower survival odds. Historical accounts align with this as well; since women and children were called to evacuate first, male passengers were largely left to fend for themselves. This underscores how gender can affect survival in certain situations, especially historically, when women and children were often prioritized in emergencies above the safety of men.

The younger passengers aboard the Titanic, the more likely their chance of survival is. Older passengers had a higher mortality rate; this aligns with passenger accounts that women and children were prioritized during evacuation. During emergencies, it could also indicate that older individuals were not as able to access lifeboats. The regression model's negative coefficient exemplified this: being older decreased the likelihood of passengers surviving.

This study highlights the factors of age, gender, and passenger class and their effect on survival. The logistic regression model produced had an accuracy of 78.85%; this indicates a high accuracy in the model. The model showed it could predict who did and did not survive well. These findings demonstrate how logistic regression can be utilized in survival analysis, especially considering the influence of disparities. Insights gained from this study have practical applications and can help professionals who assess risk mitigate potential barriers to survival.

**Conclusion**

When applied similarly, models can draw predictions from other historical instances with datasets that could be valuable examples of future work. The efficacy of binary logistic regression models in predictive analytics is codified and exemplifies the diversity of how the model could be applied. Exploring how these models could be applied could allow for further refinement and more nuanced conclusions based on the data provided. Understanding complex datasets that represent a defined result and their influences could help address how to mitigate potential adverse outcomes best.

The effectiveness of binary logistic regression was displayed by achieving an accuracy of 78.85%. The results confirm that independent variables that could describe social and economic factors, like age, gender, and passenger class influence survival probabilities. The historical narratives surrounding the sinking of the Titanic align with the probabilities drawn from the model. Continued exploration of historical incidents such as these could provide valuable context for current problems and help scientists consider what could be valuable in that research.

**References**

Encyclopaedia Britannica. (n.d.). *Titanic*. Britannica. <https://www.britannica.com/topic/Titanic>

Fávero, L. P., & Belfiore, P. (2018). Data Science for Business and Decision Making (1st ed.). Academic Press. https://doi.org/10.1016/C2016-0-01101-4

Harris J. K. (2021). Primer on binary logistic regression. *Family medicine and community health*, *9*(Suppl 1), e001290. https://doi.org/10.1136/fmch-2021-001290

Penman, P. (2022, February 10). *Binary Logistic Regression An Introduction*. Retrieved February 9, 2025, from https://www.datascienceinstitute.net/blog/binary-logistic-regression-an-introduction

R Core Team. (2023). *Titanic dataset*. R: A language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>



University of the Cumberlands

School of Computing and Information Sciences

# Certification of Authorship

**Submitted to**: Dr. Afshin Zarenejad

**Course**: MSDS-530-A01

**Student’s Name**: Lisa Hall

**Date of Submission**: 02/23/24

**Purpose and Title of Submission**: Hypothesis and Regression Models Application Strategy

**Certification of Authorship**: I certify that I am the author of this document and that any assistance I received in its preparation is fully acknowledged and disclosed. I have also cited all sources from which I obtained data, ideas, or words copied directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by (me or my group #) for this purpose.

**Student’s Signature**: Lisa Hall