Analyzing Survival Outcomes on the Titanic Using Iteratively Reweighted Least Squares (IRLS) and Simulation Techniques

Benson Cyril Nana Boakye
December 19, 2024

1 Introduction

The **Titanic disaster** remains one of the most tragic and well-known events in maritime history. On April 15, 1912, the RMS Titanic, a luxury ship famously deemed "unsinkable," struck an iceberg during its maiden voyage in the North Atlantic Ocean. Within hours, the ship sank, claiming the lives of 1,502 out of 2,224 passengers and crew aboard. While the sheer scale of the tragedy captured the world's attention, evidence suggests that **survival was not merely a matter of chance** but was influenced by various social, economic, and demographic factors.

This project aims to explore the survival outcomes from the Titanic disaster, focusing on how certain factors such as **gender**, **age**, **social class**, and **fare** impacted the likelihood of survival. While many of these factors may seem intuitive—such as women and children being prioritized for lifeboats—statistical analysis can reveal more intricate relationships and quantify the extent to which each factor influenced survival.

To conduct this analysis, we apply **Iteratively Reweighted Least Squares (IRLS)**, a method commonly used in **Generalized Linear Models**, to fit a **logistic regression model**. Logistic regression allows us to model the probability of survival (a binary outcome: survived or not) based on various predictors. Using **IRLS**, we estimate the coefficients for each predictor (e.g., gender, class, age, and fare) to understand how they influence the odds of survival.

Beyond basic logistic regression, we also employ **Monte Carlo simulations** to simulate different survival scenarios. For instance, we investigate how changes in **lifeboat capacity**, such as increasing the number of available lifeboats, might have altered survival rates. Additionally, we modify the **demographic distribution** of the passengers to explore "what-if" scenarios, such as a higher proportion of women or children onboard. These simulations help us estimate the effects of different conditions on survival, providing insights into how the distribution of resources (like lifeboats) could have impacted survival outcomes.

To ensure that our simulations are efficient and accurate, we incorporate **variance reduction techniques**. These techniques, including **stratified sampling**, **antithetic variates**, and **control variates**, improve the precision of our results by reducing the variability inherent in random sampling. Stratified sampling divides the population into subgroups (e.g., by class), ensuring that each subgroup is adequately represented. Antithetic variates involve generating pairs of complementary simulations, which help reduce the randomness in the results. Control variates adjust the outcomes based on known quantities, further stabilizing the estimates.By combining logistic regression with Monte Carlo simulations and variance reduction techniques, this analysis not only uncovers patterns in the Titanic survival data but also demonstrates the practical use of statistical methods in exploring historical events. For this analysis, we used the Titanic dataset, which is available on Kaggle.

2 Methodology

This study investigates Titanic survival outcomes by employing **Iteratively Reweighted Least Squares (IRLS)** for logistic regression and **Monte Carlo simulations**. The following sections describe the steps taken to implement the logistic regression model, conduct the simulations, and apply variance reduction techniques to ensure accurate and efficient results.

2.1 Logistic Regression via Iteratively Reweighted Least Squares (IRLS)

To model the survival probabilities of passengers, we employed **logistic regression**, a standard method for predicting binary outcomes. In this context, the binary outcome was survival (1) or non-survival (0). The **Iteratively Reweighted Least Squares (IRLS)** algorithm was used to fit the logistic regression model, as it efficiently handles the nonlinearities of logistic regression by transforming the problem into a series of weighted least squares problems.

2.1.1 Model Specification

The logistic regression model expresses the probability of survival as a function of a linear combination of passenger features. Mathematically, the probability $P(\text{Survived} = 1|\mathbf{X})$ is given by the logistic function:

$$P(\text{Survived} = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\mathbf{X}\boldsymbol{\beta})}}$$

- $\mathbf{X}$is the matrix of predictor variables (gender, age, class, fare) and $\boldsymbol{\beta}$ is the vector of regression coefficients.

2.1.2 Log-Likelihood Function

The log-likelihood function is used to estimate the model parameters by maximizing the likelihood of observing the given data. For binary outcomes $Y_i \in \{0, 1\}$, the likelihood function is: $f(Y_i|\boldsymbol{\beta}) = \mu_i^{Y_i}(1 - \mu_i)^{1-Y_i}$
The **log-likelihood** function is $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i \log(\mu_i) + (1 - Y_i) \log(1 - \mu_i)]$

2.1.3 Newton-Raphson Method for Optimization

To estimate the coefficients $\boldsymbol{\beta}$, we used the **Newton-Raphson method**, an iterative optimization technique that updates the coefficients based on the gradient and the Hessian matrix. The update rule is $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{H}^{-1}\nabla\ell(\boldsymbol{\beta})$

- $\nabla\ell(\boldsymbol{\beta})$ is the gradient and $\mathbf{H}$ is the Hessian matrix, the second derivative of the log-likelihood function.

The gradient $\nabla\ell(\boldsymbol{\beta})$ is computed as: $\nabla\ell(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu})$ Where:

- $\mathbf{y}$ is the vector of observed survival outcomes and $\boldsymbol{\mu}$ is the vector of predicted survival probabilities.

$\mathbf{H} = -\mathbf{X}^T\mathbf{W}\mathbf{X}$. Where $\mathbf{W}$ is the diagonal weight matrix with elements $w_i = \mu_i(1 - \mu_i)$, representing the variance of the predicted probabilities.

2.1.4 Reformulation into IRLS

The **IRLS algorithm** is derived from the Newton-Raphson method. Substituting the gradient and Hessian into the update rule, the coefficients are updated iteratively as follows: $\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{y} - \boldsymbol{\mu})$
Where $\mathbf{W}$ is the weight matrix, and $\mathbf{y} - \boldsymbol{\mu}$ is the vector of residuals. This iterative update process continues until the coefficients converge, which is determined when the change in $\boldsymbol{\beta}$ falls below a predefined threshold (e.g., 10^{-6}).

2.1.5 Working Response

To simplify the iterative procedure, a **working response** $\mathbf{z}$ is introduced: $\mathbf{z} = \eta + \frac{\mathbf{y}-\boldsymbol{\mu}}{\boldsymbol{\mu}(1-\boldsymbol{\mu})}$
Where $\eta = \mathbf{X}\boldsymbol{\beta}$ is the linear predictor. Substituting $\mathbf{z}$ into the update rule results in the final IRLS update:

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}$$

2.2 Monte Carlo Simulation

In addition to the logistic regression model, **Monte Carlo simulations** were employed to simulate different survival outcomes and estimate the impact of various factors, such as changes in lifeboat capacity or demographic distributions.

2.2.1 Generating Synthetic Passenger Data

Synthetic data for **10,000 passengers** was generated using the following parameters:

- **Gender**: 50% male, 50% female.
- **Class**: Randomly assigned to 1st, 2nd, or 3rd class with proportions 20%, 30%, and 50
- **Age**: Normally distributed with a mean of 30 years and a standard deviation of 10 years.
- **Fare**: Exponentially distributed with a mean of 50.

These attributes were generated using `numpy` random functions to simulate real-world distributions for each variable. For each simulated passenger, the **predicted survival probability** was computed using the logistic regression model obtained from the **IRLS** algorithm. The predicted survival probability μ_i is given by: $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, Where $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$ is the linear predictor. Using the predicted probabilities μ_i, survival outcomes for each passenger were simulated based on a **binomial distribution**: $Y_i = \text{Binomial}(1, \mu_i)$, Where $Y_i = 1$ indicates survival, and $Y_i = 0$ indicates death.

2.2.2 Calculating Baseline Survival Probability

The **baseline survival probability** was calculated as the proportion of survivors out of the total simulated passengers:

$$\text{Baseline Survival Probability} = \frac{\text{Number of survivors}}{\text{Total number of simulations}}$$

2.2.3 Impact of Increased Lifeboat Capacity

To evaluate the effect of increased lifeboat capacity on survival outcomes, a hypothetical scenario was simulated in which 95% of passengers were assumed to have access to lifeboats, representing an improved resource allocation compared to historical conditions. Survival probabilities predicted by the logistic regression model were adjusted by capping them at 95%, ensuring that no passenger's survival likelihood exceeded this threshold. This was achieved using the following formula:

$$\text{Adjusted Survival Probability} = \min(\mu_i, 0.95)$$

These adjusted probabilities were then used to simulate survival outcomes for 10,000 passengers through a binomial sampling process. The overall survival probability under this enhanced lifeboat scenario was computed as the proportion of passengers who survived.

2.3 Variance Reduction Techniques

To improve the precision and efficiency of the Monte Carlo simulations, **variance reduction techniques** were applied. These methods help reduce the variability of the simulation estimates, leading to more accurate results with fewer simulations.

2.3.1 Stratified Sampling

In **stratified sampling**, the population was divided into subgroups based on passenger class (1st, 2nd, and 3rd class). The survival outcomes were simulated for each class separately, and the overall survival probability was computed as the **weighted average** of the class-specific probabilities.

2.3.2 Antithetic Variates

Antithetic variates were generated for each simulated passenger. For a passenger with a predicted survival probability p, an **antithetic sample** was generated with survival probability $1 - p$. The survival outcomes for each pair were averaged to reduce the randomness in the simulation.

2.3.3 Control Variates

Control variates were applied by using a known baseline survival rate (e.g., from historical Titanic data) as a control value. The difference between the simulated survival probabilities and the baseline survival rate was calculated, and the simulation outcomes were adjusted accordingly to reduce variability.

3 Results

3.1 Logistic Regression Model (IRLS)

The **IRLS algorithm** was applied to fit a **logistic regression model**, predicting the probability of survival based on several key passenger characteristics, including **gender**, **age**, **class**, and **fare**. The coefficients estimated by the model are as follows:

$$\boldsymbol{\beta} = [2.048, 2.607, -1.152, -0.033, 0.00059]$$

3.1.1 Interpretation of Coefficients

- **Intercept ($\boldsymbol{\beta_0} = 2.048$)**: The log-odds of survival when all other variables are set to zero (e.g., male, third-class, age 0, fare 0). The positive intercept indicates a higher probability of survival when other factors are not considered.

- **Gender ($\beta_1 = 2.607$)**: Female passengers had a higher log-odds of survival compared to male passengers. The coefficient suggests that being female increases the odds of survival significantly, consistent with historical reports of women and children being prioritized for lifeboats.
- **Class ($\beta_2 = -1.152$)**: Passengers in third class (the reference category) had lower odds of survival compared to passengers in first and second class. The negative coefficient indicates that lower-class passengers had significantly lower survival chances, reflecting the prioritization of wealthier passengers.
- **Age ($\beta_3 = -0.033$)**: Each additional year of age decreased the log-odds of survival, though the effect is small. Younger passengers were slightly more likely to survive, reflecting the general trend of prioritizing women and children.
- **Fare ($\beta_4 = 0.00059$)**: Higher-paying passengers (those who paid higher fares) had a slightly higher chance of survival. This suggests that wealthier passengers were more likely to have access to lifeboats and better accommodations.

The logistic regression model converged in **6 iterations** using the IRLS algorithm, with an accuracy of **80.02%**. This indicates that the model successfully predicted survival outcomes for 80% of the passengers in the simulation.

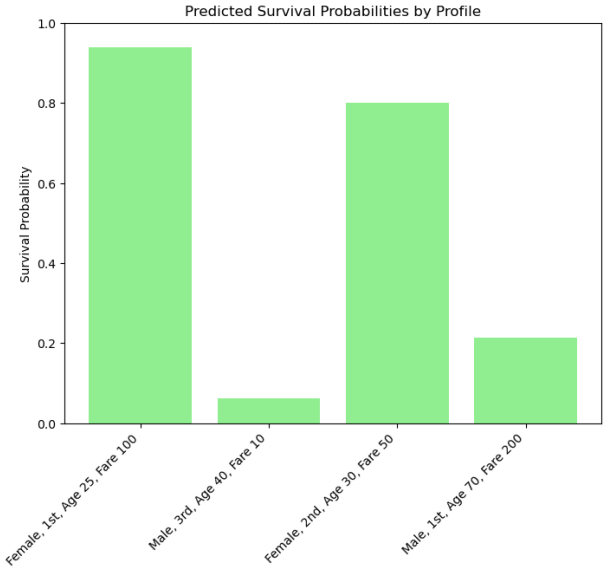

Figure 1: Predicted survival probabilities by profile

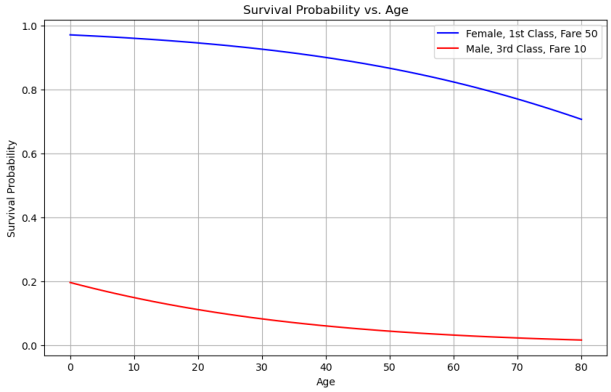

Figure 2: Survival Probability vs Age

From Figure 1, I simulated and predicted the survival probabilities of different passenger profiles on the Titanic. These profiles were converted into a DataFrame and an intercept term is added. Using the logistic regression coefficients (denoted as beta), calculates the survival probabilities for each profile by applying the logistic function. The results, visualized through a bar chart, reveal distinct survival trends. A female passenger in 1st class, aged 25, with a fare of 100 exhibited the highest survival probability, reflecting the prioritization of women and higher-class passengers during the evacuation. In contrast, a male passenger in 3rd class, aged 40, with a fare of 10 demonstrated the lowest survival probability, highlighting the compounded disadvantages associated with being male and in a lower socioeconomic class. Further comparisons illustrate that a female passenger in 2nd class, aged 30, paying a fare of 50, has a relatively high probability of survival, albeit lower than her 1st-class counterpart. Similarly, a male passenger in 1st class, aged 70, paying a fare of 200, achieves a moderate survival probability where the benefits of class and fare are tempered by the adverse impact of advanced age and gender.

From Figure 2, I analyzed how survival probabilities change with age for two passenger profiles: a female in 1st class paying $50 do and a male in 3rd class paying $10. The plot shows that the female's survival probability starts high for younger ages and decreases slightly with age but remains significant. In contrast, the male's survival probability starts much lower and drops sharply with age, reaching nearly zero for older passengers.
This highlights how gender, class, and fare strongly influenced survival, with the female 1st-class passenger consistently having a much higher chance of survival than the male 3rd-class passenger.

3.2 Monte Carlo Simulation: Baseline Survival Probability

The baseline survival probability, calculated as the proportion of simulated passengers who survived, is: 44.89%
This suggests that approximately **44.89%** of the simulated passengers survived, based on the given demographic distribution and survival predictions from the logistic regression model.

3.3 Impact of Increased Lifeboat Capacity

The adjusted survival probability, calculated after increasing lifeboat capacity, was 45.26%. This represents a modest improvement in survival rates from **44.89%** to **45.26%**. The result suggests that increasing the number of available lifeboats to 95% would have slightly improved survival chances for the passengers, though the effect is limited.

3.4 Impact of Variance Reduction Techniques

The application of variance reduction techniques significantly improved the precision of the Monte Carlo simulation results. The baseline survival probability, without any variance reduction, was 44.89%. Stratified sampling increased the survival probability to 45.96%, reflecting a more precise estimate that accounted for class-based differences in survival. Antithetic variates provided the most substantial improvement, raising the survival probability to 50.09%, indicating a significant reduction in variability and greater stability in the estimate. Control variates led to a moderate improvement, with the survival probability adjusted to 45.01%, offering a more stable result compared to the baseline but with a smaller effect than antithetic variates. Overall, these techniques enhanced the accuracy of the survival estimates, with antithetic variates proving to be the most effective in reducing variability, followed by stratified sampling and control variates.

4 Conclusion

This study applied **Iteratively Reweighted Least Squares (IRLS)** for logistic regression and **Monte Carlo simulations** to analyze the survival outcomes of passengers aboard the RMS Titanic. The findings underscore the significant influence of **gender** and **class** on survival chances, with females and higher-class passengers exhibiting a markedly higher likelihood of survival. Simulating an increased lifeboat capacity scenario revealed a modest improvement in survival probability, highlighting the potential benefits of enhanced resource allocation during crises, though it was secondary to the impact of socioeconomic factors. The use of **variance reduction techniques**—including **stratified sampling**, **antithetic variates**, and **control variates**—substantially enhanced the precision of the Monte Carlo simulations. **Antithetic variates** proved to be the most effective, resulting in the most reliable survival estimates, while **stratified sampling** and **control variates** also contributed to improved accuracy. While the findings offer valuable insights, certain limitations must be acknowledged. First, the analysis assumes linear relationships between the predictors and the log-odds of survival, which may oversimplify the complexity of real-world interactions. Additionally, the synthetic data generated for the Monte Carlo simulations, though based on historical distributions, may not fully capture the actual demographic and situational nuances of the Titanic's passengers. Future research could explore more complex models or incorporate additional variables to improve the robustness of the findings. In conclusion, this study demonstrates the power of statistical methods to extract meaningful insights from historical data, offering a clearer understanding of the dynamics that influenced survival during the Titanic disaster. The application of advanced modeling and simulation techniques not only sheds light on past events but also provides valuable lessons for future decision-making in crisis management and resource allocation.

References

[1] Cox, D. R., 1972. *Regression Models and Life-Tables.* Journal of the Royal Statistical Society: Series B (Methodological), 34(2), pp.187-220.

[2] Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., and Rubin, D. B., 2013. *Bayesian Data Analysis* (3rd ed.). CRC Press.

[3] Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[4] Ross, S. M., 2010. *Simulation* (4th ed.). Academic Press.

[5] Kaggle, 2024. *Titanic: Machine Learning from Disaster.* [online] Available at: <https://www.kaggle.com/c/titanic>.