Econ 520 Project

Sofia Lozano-Samper, Finn Owsley, Scott Masterson

May 2024

Contents

1 Introduction

In this investigation, we use ordinary least squares and regression adjustment to assess the factors that led to the fate of victims of the Titanic Disaster of 1912. We consider demographic information like age, passenger class, and gender to see if any had influencing factors into the survival rates of the passengers aboard. The passengers were divided into three classes: first class passengers were the wealthiest and had a multitude of accommodations aboard, while 3rd class were the poorest. We narrow our scope to the following research question:

How did being in first class affect one's chance of survival in the sinking of the Titanic?

We will answer this question using the previously mentioned inference methods. We expect that a higher-class status will translate into higher survival rates. Additionally, we expect younger passengers and women to be prioritized in emergency procedures at this time in history: "The practice of prioritising women and children gained widespread currency following the actions of soldiers during the sinking of the Royal Navy troopship HMS Birkenhead in 1852 after it struck rocks" [How02]. Also, men were often expected to die with honor in tragedies such as the Titanic:

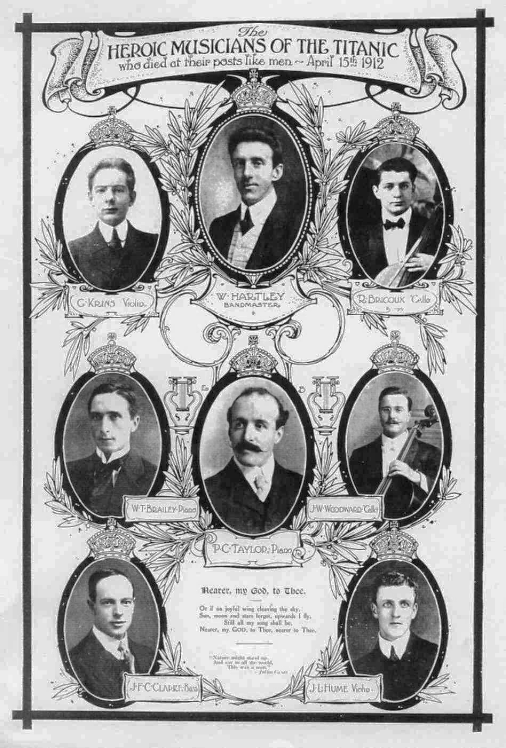
The
HEROIC MUSICIANS OF THE TITANIC
who died at their posts like men ~ April 15th 1912
G·KRINS Violin
W·HARTLEY
BANDMASTER
R·BRICOUX Cello
W·T·BRAILEY·Piano
J·W·WOODWARD·Cello
P·C·TAYLOR·Piano
Nearer, my God, to Thee.
Or if on joyful wing cleaving the sky,
Sun, moon and stars forgot, upwards I fly,
Still all my song shall be,
Nearer, my GOD, to Thee, nearer to Thee.
J·F·C·CLARKE·Bass
J·L·HUME Violin

Figure 1: Men who Died "Like Men" in the Tragedy were Hailed as Heroes

2 Literature Review

Previous work (Frey et. al, 2011) [FST11] explores the behavioral determinants of survival from Titanic. Through analysis of individual-level data, researchers can explore factors such as gender, social status, companionship, role, and nationality to understand their impact on survival rates, shedding light on the dynamics of social norms like "Women and children first!" and the influence of time duration on evacuation strategies, as evidenced by comparisons to other maritime disasters like the sinking of the Lusitania. This is furthered by research conducted by (Frey, Savage et. al, ND) [FST11], who found that women and children had greater odds of survival than men, while first-class passengers had a higher likelihood of being rescued than those in second or third class. British travelers faced a higher risk of perishing than individuals from other countries. In terms of methods, they use probit estimates to test their main hypotheses.

Further, the Titanic Dataset has been widely used for computer science classes, as well as ML exercises. A simplified version of the dataset is found in Kaggle. One of the most interesting uses of this dataset is Konos Papadopoulos' estimation of the effect of treatment (a passenger's assignment to a particular cabin) on passenger survival [P]. He uses propensity score matching (PSM) to create an artificial control group.

3 Data

We sourced our data from Titanic Encyclopedia. We chose this source over others located on Kaggle since this data had 1311 observations, as opposed to around 900. Though the other datasets on Kaggle were in a more clean format, we wanted the dataset to be closer to a census of all passengers on board.

The data came in several large tables organized by passenger class. The following subsections overview cleaning operations we used in Google Sheets to get the data in a desirable format.

3.1 Merging the Data

The data came in three chunks organized by the classes on the Titanic: 1st, 2nd, 3rd. In order to get them into the same Google Sheet, we copy and pasted them manually. We then added a new column called “Class” where we stored values 1, 2, and 3 to indicate the class of the 1311 passengers in our dataset.

3.2 Creating New Variables: Gender

Using the “Name” column in our data-frame, we leveraged ChatGPT-4’s language recognition features to assign gender based off of the name. We thought of creating code for this ourselves, but since the passengers on the Titanic came from all over the world, many of the names would be difficult to analyze. We gave ChatGPT-4 the following prompt:

```
I am going to give you a large column of names
from around the year 1910.
I want you to give me back a new column of equal length
that I can copy and paste into Google sheets.
For each entry on the column I want you to write
Male or Female based on the gender of the corresponding name.
```

We are confident that ChatGPT-4 has a high degree of accuracy in this use case since many of the names included gendered prefixes like Mr. or Mrs.

and we manually validated the results for a small random subset of about 50 names. We found that subset to be 100% accurate.

3.3 Creating New Variables: Age Bins

From the original data, we had the age of all passengers. For causal inference, we needed to turn age into a binary variable with logical bins. After examining the following histogram of age and passenger count, we decided to make the bins 0-15, 16-30, 31-50, and 51+ to ensure each group had around the same amount of people:

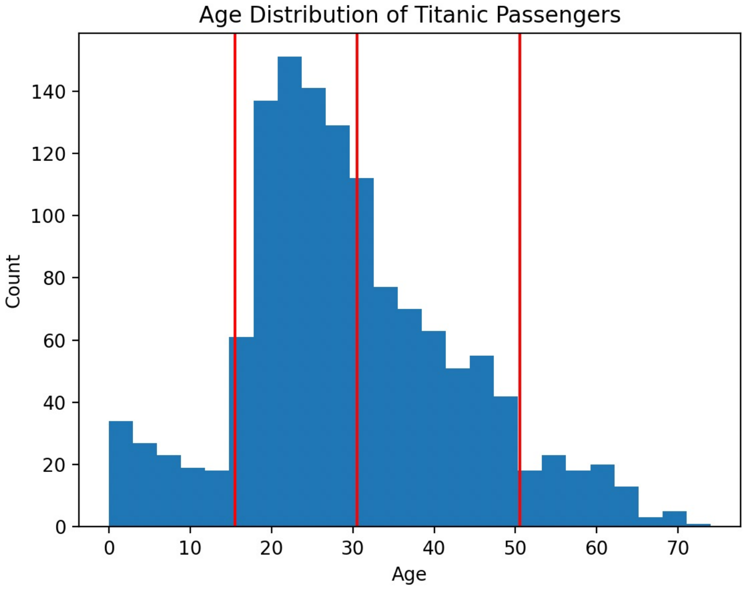

Figure 2: The Distribution of Age and Count is Skewed Left. The red lines indicate where we subdivided the data.

3.4 Creating New Variables: Survived

The dataset didn't explicitly say who survived the sinking of the Titanic; however, those who survived had their row highlighted blue in the data frame and those who died did not have any shading. In Google Sheets, we leveraged the feature of making a built-in function in Google Apps that checked if the row had blue shading or not. Using a function in javascript, we used color Hexcode matching that would display a "1" in the row adjacent to the blue-shaded row indicating if the person survived, and a "0" otherwise. We applied this function to all rows of the data-frame.

3.5 Summary Statistics

There were 1311 total observations in our dataset. 65% of the passengers were male. 35% of the passengers were female. 25% were 1st class. 9% were in the age group <15. 50% were in the age group 16-30. 33% were in the age group of 31-50. 8% were in the age group of 51+. 38% survived, while 62% perished.

4 Empirical Methods Overview

In this section, we will discuss the statistical methods that we employed in order to answer our fundamental research question from the introduction section (1). To answer this question, we will employ causal inference methods, seeking to tie the dependent variable of survival status to the treatment status of being in first class, as opposed to second or third class. We will also use the other variables as controls.

In order to run our desired causal inference methods and generate valid results, it was necessary for us to operate under a number of assumptions. First is the assumption of unconfoundedness, which requires us to make the treatment status be as if random. To do so, we must control for possible confounding variables, include age and gender.

Additionally, we assume overlap, which simply states that there is a minimum requirement in terms of number of individual of both treatment statuses within each subgroup. This is satisfied in our dataset.

Lastly, we have the assumption of homogeneity, which states that the effect of each variable is not dependent on the others. We believe this to be our weakest assumption, and therefore we will employ causal inference methods which account for heterogeneity.

The basis of our analysis was ordinary least square multiple linear regression, for which we used the following equation:

$$Y_i = \alpha + \beta D_i + \gamma_1 X_{1i} + \gamma_2 X_{2i} \gamma_3 X_{3i} + \gamma_4 X_{4i} + \epsilon_i$$

This ties the treatment status (first class status) Di to the survival status Y_i while controlling for variables X_1 through X_4. α represents the intercept, which is equivalent to the outcome for a theoretical control group. The control variables account for gender and age. We also employed the regression adjustment method in order to account for heterogeneity. In this method, we separate the data into treated and untreated subsets and run a separate regression on each, which are then applied to the data as a whole, and the outputs are compared to get a final ATE, based on the according to the following equation:

$$\widehat{ATE}_{ra} = \hat{\alpha}_{D=1} - \hat{\alpha}_{D=0} + \bar{X}\hat{\beta}_{D=1} - \bar{X}\hat{\beta}_{D=0}$$

To calculate the standard error of this ATE, we employed a bootstrap method, with 10,000 simulations featuring random sampling, calculating sample ATEs based on each.

We also examined heterogeneity, specifically between the age and gender variables, using the interaction variables method. We first combined the young adult and middle age categories into one category, 16-50. Then we added the interaction variables Male*(16-50) and Male*Old. These dummy variables will be equal to 1 only if the individual is both male and between the ages of 16-50, or male and older than 50, respectively. The rationale behind this decision was that gender would only be an important factor if the individual in question was an adult, since children of both genders were prioritized.

5 Research Findings

5.1 Ordinary Least Squares

Below, we show our results for OLS.

Table 1: OLS Regression Results

Variable	Coefficient	Standard Error	t-Statistic	P-value	95% C.I
Const	0.7654	0.038	20.284	0.000	(0.691, 0.839)
Male	-0.5031	0.023	-21.850	0.000	(-0.548, -0.458)
Young Adult	-0.1296	0.039	-3.333	0.001	(-0.206, -0.053)
Middle Age	-0.1439	0.042	-3.463	0.001	(-0.225, -0.062)
Old	-0.2694	0.056	-4.848	0.000	(-0.378, -0.160)
First Class	0.2941	0.028	10.639	0.000	(0.240, 0.348)

When looking at the linear regression results, the intercept coefficient is positive. This implies that holding all other variables constant at their reference categories (children, females, and any other class except first class), the estimated survival probability for such an individual is 0.7654. Essentially, this coefficient represents the baseline survival probability for individuals who are children, female, and not in first class.

Secondly, the coefficient estimate of 0.2941 for a first-class status indicates that holding all other variables constant, being a first-class passenger is associated with a 0.2941 increase in the probability of survival. The small standard error of 0.028 suggests high precision in this estimate. With a t-value of 10.639, which is higher than the critical value of 1.96 at the 95% confidence level, the coefficient is highly significant. Additionally, this is reflected in the fact that the 95% confidence interval is (-0.548 -0.458), which doesn't include zero.

Additionally, the adjusted R-squared value of 0.341 suggests that the model explains approximately 34.1% of the variance in survival status based on the chosen variables. In summary, these regression results strongly suggest that being in first class is associated with a large and statistically significant increase in the probability of survival on the Titanic, even after adjusting for other covariates included in the model.

Furthermore, when looking at the rest of the covariates, it's evident that women were more likely to survive than their male counterparts, indicated by a coefficient of -0.5031 for the "Male" group: being a male reduced survival chance by 50.31%. Regarding age, all groups ("Young Adult", "Middle Age" and "Old") had negative coefficients, indicating that compared to children (the reference group), all other groups had a lower likelihood of surviving. However, the older you are, the greater this negative effect is (-0.1296 for young adults, vs -0.1439 for middle age, vs. -0.26940 for old). For each one of the covariates, the results are statistically significant at the 95% confidence level (the p-values associated with each coefficient are less than 0.05).

5.2 Regression Adjustment

The results of the regression adjustment analysis provide an estimated Average Treatment Effect (ATE) of approximately 0.299 with a standard error of about 0.041. This is similar to our initial ATE estimate of 0.2941 using the simple OLS method. Since the confidence interval (0.2186, 0.3794) does not include zero, we can conclude that the estimated ATE of 0.299 is statistically significant at the 95 percent confidence level.

5.3 Interaction Variables

Table 2: Interaction Variables Regression Results

Variable	Coefficient	Std. Error	t-Statistic	P-value	95% C.I.
const	0.5340	0.051	10.393	0.000	(0.433, 0.635)
Male	-0.0704	0.070	-1.003	0.316	(-0.208, 0.067)
16-50	0.1264	0.056	2.271	0.023	(0.017, 0.236)
Old	0.0807	0.087	0.930	0.353	(-0.090, 0.251)
16-50 x Male	-0.4737	0.075	-6.357	0.000	(-0.620, -0.327)
Old x Male	-0.5954	0.108	-5.502	0.000	(-0.808, -0.383)
First Class	0.2798	0.026	10.703	0.000	(0.228, 0.331)

From the heterogeneity analysis, the age groups were changed to have clearer interaction terms show that being adult (16-50 years old) and male has a negative effect on survival. However, the negative effect is higher for old males. Additionally, the coefficient for the intercept decreased, indicating a shift in the baseline survival probability when accounting for heterogeneity effects.

The regression with interaction terms reveals that being a male and an adult has a significant negative effect on survival, but the effect of being a male is no longer statistically significant because the confidence interval includes zero.

6 Conclusion and Project Extension

This study utilized ordinary least squares and regression adjustment to analyze survival factors during the Titanic disaster, particularly focusing on passenger class. The results confirmed with a high degree of confidence that first-class passengers had significantly higher survival rates, underscoring the historical bias towards wealthier individuals during emergencies. The findings also supported the “women and children first” policy, as gender and age significantly influenced survival chances. Our use of interaction variables further confirmed that gender played a role in the survival of adults, but not children.

We could expand upon this investigation by investigating the effect of a larger number of variables, and determining whether that had an effect on survival outcomes. If used as control variables, these additional variables could improve the accuracy of our ATE calculations. Possible variables to add include:

- Destination
- Place of origin
- Cabin location on ship
- If a passenger made it onto a lifeboat or not
- What side of the ship the lifeboat the passenger boarded was on

References

[How02] W. W. Howells. *The Steamboat Disaster*. Accessed May 2024. 1902. URL: `https://archive.org/details/steamboatdisaste01howl/page/150/mode/2up`.

[FST11] Bruno S. Frey, David A. Savage, and Benno Torgler. “Behavior under Extreme Conditions: The Titanic Disaster”. In: *Journal of Economic Perspectives* 25.1 (2011), pp. 209–222.

[P] Konstantinos P. *Propensity Score Matching Analysis*. GitHub repository. Accessed May 2024. URL: `https://github.com/konosp/propensity-score-matching/blob/main/propensity_score_matching_v2.ipynb`.