PSTAT 100 Final Project

Natalie Samuels, Gabriel Ambrose, Tianhao Qin

2024-11-28

Introduction

Brief Overview of Data Set

One of history's most known tragedies was the Titanic disaster in April of 1912. After hitting a massive iceberg, the massive aquatic vessel sank in the middle of the Northern Atlantic Ocean. From this disaster we have collected the data of those who were aboard the ship at the time of the disaster.
In the Titanic data set, we have a structured collection of data that consists of the record of each passengers' information. From this data, we are able to get a more detailed perspective of the population aboard the titanic prior to it sinking, as well as those who survived the disaster. The data set includes each passenger's name and each passenger was assigned a Passenger ID from 1 to 891. From there, the data set includes two variables depicting whether the person survived and their class. The "Survived" variable is categorized by 0 and 1, 1 meaning they survived and 0 meaning they unfortunately did not survive. The "Pclass" variable is categorized by numbers 1-3, each representing the first, second and third class based on which ticket the passenger purchased. Then, each passenger's sex and age is recorded in the data set, the youngest being less than a year old and the oldest being 80 years old. The data set also includes information about the family status of each passenger including whether they boarded the ship with a sibling or spouse, as well as whether they had a parent or child aboard with them. More detailed information that the data set includes is the type of ticket purchased, the price, the passenger's cabin number and where they embarked from. Using this information, we have a collection of data that accounts for a large amount of the passengers aboard the Titanic.
Further research shows that the Titanic had 2240 aboard at the time of the sinking, 1317 when excluding the crew. With this information, we know that we do not have data of the complete population that was aboard the Titanic at the time of the sinking. However, we do have the data of a large sample, approximately 67.6% of the total number of passengers and 39.7% of the total population aboard the Titanic when it sank in 1912. With the sample data from the Titanic data set, we are able to analyze different components of this historic event and those who experienced it.

Research Questions and Hypotheses

In the Titanic disaster, approximately 1500 people did not survive. This extremely number raises many questions. Does the data show evidence that passenger survival was based on specific factors, or were the outcomes purely due to chance? Were there particular attributes or groups of passengers that influenced survivability at a higher rate than others? Were there reasons why more people from those groups or passengers with those attributes survived when compared to others? All of these questions can be analyzed and investigated using our data set.
For our analysis, we chose to use the variables in our data set to evaluate which impacted a passenger's survival. The variables we chose to assess were Age, Sex, and Class because we felt these could possibly have a significant impact on whether or not a passenger survived the sinking of the Titanic.

Research Question #1

On the Titanic, the evacuation protocol claimed that "women and children would be helped first" in the case of an emergency. Due to this, the common assumption would be that women and children had a higher survival rate due to the protocol insisting that they receive priority assistance. However, in a disaster as turbulent as the Titanic sinking, most of the passengers must have been frenzied and frantic meaning this protocol may have been impossible to instill in such chaos. This leads to the following question:
Given the evacuation protocol prioritizing women and children, was there a difference in survival rates between male and female children?

Hypothesis

Let μ_M be the mean survival rate for male minors and μ_F be the survival rate for female minors. Thus, our null hypothesis is that there is no difference in the two means, meaning male and female children were given equal access to the life boats. The alternative hypothesis is that the two means are not equal, and there is some difference in the two groups mean survival rates based on the gender of the children.

$$H_0 : \mu_M = \mu_F$$
$$H_A : \mu_M \neq \mu_F$$

Research Question #2

Although age and gender was prioritized in the evacuation protocol, there may have been an uneven distribution of resources. During this period in history, there was a strong divide between the classes as well as the resources available to each ones. For example, the third class cabins were located at the very bottom of the ship. This meant that they would have been in much more danger when the Titanic hit the iceberg as the ocean water would inevitably fill the bottom of the ship first when it began to sink. The first class cabins was located on the top deck, much closer to the lifeboats and the ship crew who could assist in evacuation. This poses the question:
Regardless of age and gender, which passenger class had the highest survival rate?

Hypothesis

Let β_{1st} be the estimated coefficient representing the odds of survival for first class, β_{2nd} be the coefficient for second class, and β_{3rd} be the same for third class. Our null hypothesis states that these coefficients are all equal and there is no effect (or an identical effect) on the survival rate depending on the passenger class. Our alternative hypothesis states that at least one of these coefficients will be greater or less than the others, with {i, j, k $\in$ (1st, 2nd, 3rd)}, indicating that one of our passenger classes has a positive or negative effect on the predicted survival of a passenger based on their class.

$$H_0 : \beta_{1st} = \beta_{2nd} = \beta_{3rd}$$
$$H_A : \beta_i > \beta_j \geq \beta_k$$

Preliminary Data Analysis and Cleaning

We first examine our Titanic data set by checking for any missing values that may cause uncertainty in our future model building and statistical analysis. While checking the data, we find that there are missing values in the variables age and cabin, and the percentage of observations which had a missing value in either of those variables was 8.1%. Moving forward we use a data set that has these missing observations cleaned out for any analysis that involves the age of the passengers, and fortunately we will at no point be considering the cabin variable, so the missing values for that variable will not impact any of our analysis. No cleaning was necessary for the pclass variable.

Analysis: Research Question #1

Assumption Checking for Minor Age Groups (Male/Female)

Levene's Test

statistic	p.value	df	df.residual
0.0325096	0.8571802	1	137

The output of Levene's Test gives us **p-value = 0.8572**, which is much greater than $\alpha = 0.05$. This means we fail to reject the null hypothesis of equal variances. Thus, the assumption of homogeneity of variances is satisfied for the male and female minor groups.

Shapiro-Wilk Test

Male p-value	Female p-value
1.17003102449359e-12	1.70288846910683e-12

The Shapiro-Wilk test results for the two groups indicate that neither the survival rates of male minors nor female minors follow a normal distribution, since both p-values are significantly less than $\alpha = 0.05$ and we reject the null hypothesis of following a normal distribution for both groups.

We have found that our data satisfies the assumption of homoscedasticity that is often required for hypothesis testing, but the distributions of our groups do not follow a normal distribution. In order to obtain accurate results as we move forward with testing our hypothesis, we should use a non-parametric method. The following method allows us to accurately test our hypothesis despite our model not satisfying all the necessary assumptions we needed for other hypothesis testing methods.

Hypothesis Testing

Wilcoxon Rank-Sum Test

We chose to use the Wilcoxon Rank-Sum Test as our non-parametric method of testing our hypothesis. This test is applicable in this situation because there is no assumption of normality, and our data consists of independent observations since each individual passenger forms a separate entry and one passenger's survival

does not directly influence another passengers survival. Therefore, we satisfy the necessary assumptions to use the Wilcoxon Rank-Sum Test to receive the following results:

The Wilcoxon Rank-Sum Test results are:

- **W = 3231**
- **p-value = 7.111e-05**

Interpretation:

1. **p-value**:
    - The p-value is very small (**7.111e-05**) and much less than the significance level of $\alpha = 0.05$.
    - This indicates a statistically significant difference in survival rates between male and female minors.
2. **Hypothesis Testing Results**:
    - The test supports the rejection of our null hypothesis, which states that the mean survival rates for male and female minors aboard the Titanic were the same. This means that we have found statistically significant evidence that one group has a higher chance of surviving than the other. We demonstrate this difference in means with the following visualizations, showing that the female minors on board the Titanic had a higher mean survival rate.

Figure 1: Survival Proportions by Gender

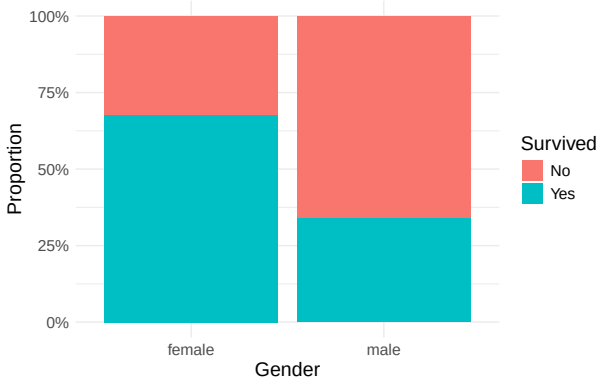
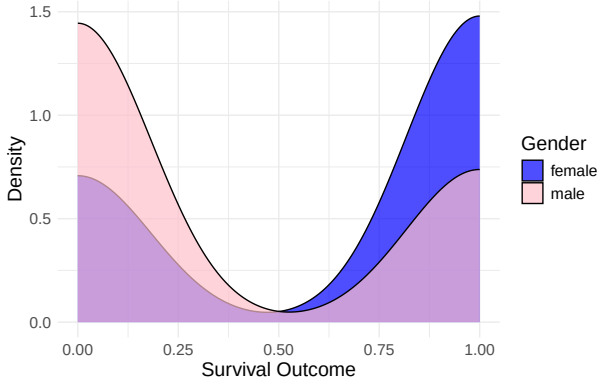


Figure 2: Density Plot of Survival Rates by Gender



The stacked bar plot in **Figure 1** simply shows the proportion of each group, female/male minors, that survived the sinking of the Titanic. A much higher percentage of the female group is shown to have survived than the male group.

The density plot in **Figure 2** shows that the distributions of survival outcomes differ significantly between male and female minors, the density being higher around 0 for males and higher around 1 for females, indicating more female minors survived than male survivors.

Analysis: Research Question #2

Assumption Checking for Class Groups (1st/2nd/3rd)

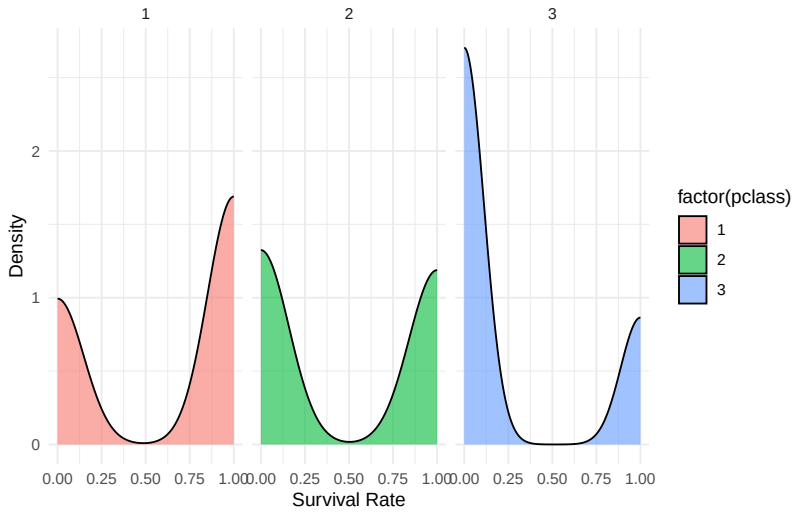
Our first step in approaching this research question was to group our passengers according to their class. Then we check those groups (1st, 2nd, and 3rd class) to see what assumptions we can use for the hypothesis testing we intend to do. We use the Shapiro-Wilk Normality test on each of the passenger classes, and then ran the Levene's Test to check the variance of each non-normally distributed class group.

Table 3: Hypothesis Testing Results

Test	Results
Shapiro Wilk: Class 1	7.715e-22
Shapiro Wilk: Class 2	1.234e-19
Shapiro Wilk: Class 3	4.518e-34
Levene's Test: Variance Between Classes	1.233e-08

Using the Shapiro-Wilk Test for each class, we obtain a significant p value that is less than 0.05 which indicates that the data follows a significantly non-normal distribution, and so we know that we cannot make any assumptions of normality moving forward. We then ran the Levene's Test to check the variance of each non-normally distributed class group, and find that **p-value = 1.233e-08**, telling us that the variances are not the same between the different classes, and as such we cannot use any assumption of homogeneous variances in the testing we will choose to do. These conclusions regarding normality and homogeneity of variances are supported by the following visualization of each classes distributions as well.

Figure 3: Density Plot of Survival Rates by Passenger Class



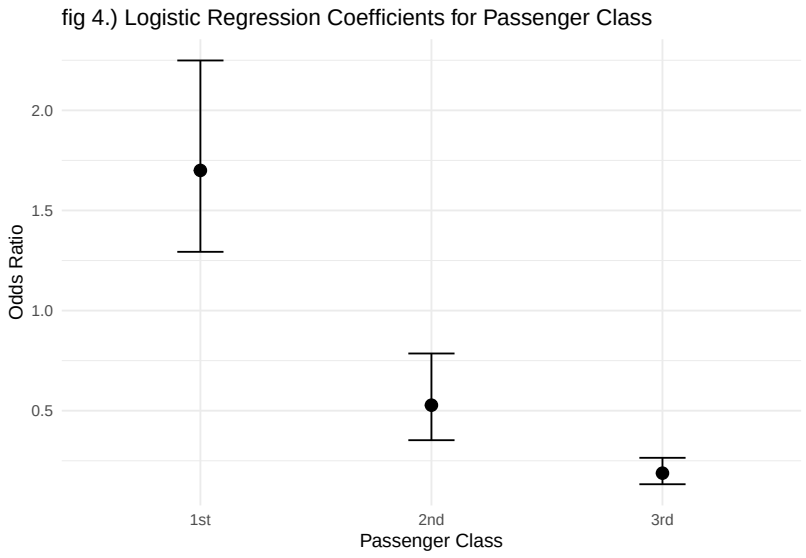
We see in **Figure 3** that the distribution for each class is clearly non-normal, and the difference in skew between each distribution confirms the non-homogeneity of the variances between the groups. With no satisfied assumptions of normality or homogeneous variances, we then consider non-parametric methods of analysis, but find that the difference in distribution between the groups and the binary nature of our dependent variable will lead to issues for several of the standard non-parametric methods such as the Kruskall-Wallis test. Due to the lack of assumptions that we can make about the data here for the sake of testing, the binary aspect of the survival variable, and the fact our predictor variable pclass is categorical, we decide to move forward using logistic regression to analyze the survival rates of the three classes.

Logistic Regression Modeling

We now create our logistic regression model to evaluate the relationship between passenger class and survival. Below are the key results of that model :

1. **Intercept (1st Class as Reference)**:
    - **Estimate**: 0.5306
    - **Odds Ratio**: $e^{0.5306} = 1.700$
    - **Interpretation**: Passengers in the 1st class have 1.7 times higher odds of survival compared to the baseline odds.
    - **P-value**: 0.0001659
2. **2nd Class** ($\beta = -0.6394$):
    - **Odds Ratio**: $e^{-0.6394} = 0.528$
    - **Interpretation**: Passengers in 2nd class have approximately 47.2% lower odds of survival compared to 1st class passengers.
    - **P-value**: 0.0017306
3. **3rd Class** ($\beta = -1.6704$):
    - **Odds Ratio**: $e^{-1.6704} = 0.188$
    - **Interpretation**: Passengers in 3rd class have approximately 81.2% lower odds of survival compared to 1st class passengers.
    - **P-value**: $<$ 2e-16

In **Figure 4** we visualize the odds ratio of survival for each passenger class, as well as 95% certainty confidence intervals for those odds ratios.



Interpretation of Logistic Model Results

- The p-values for all coefficients are all $<$ **0.002**, which at a significance level of $\alpha = 0.05$ is strong evidence that we should reject the null hypothesis that all of the passenger classes have the same survival rate, and accept the alternative hypothesis that there is a significant difference in the odds of a passenger's survival depending on their class.

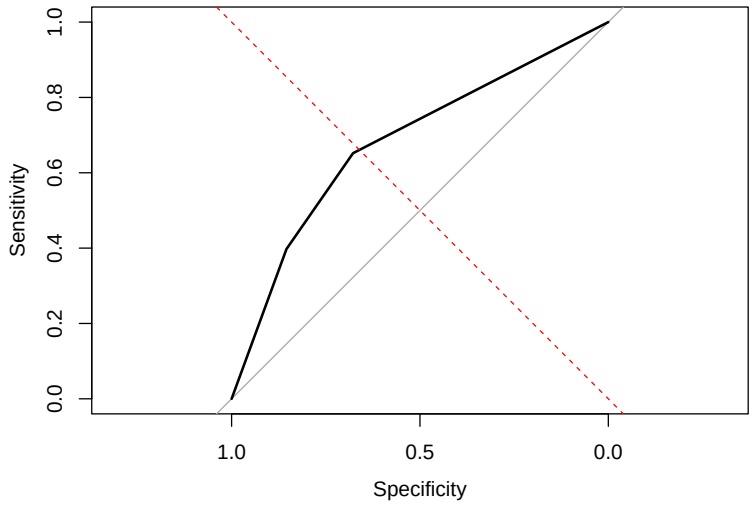
Model Evaluation

McFadden's Psuedo-R^2

We find that evaluating McFadden's Pseudo-R^2 for our logistic regression model gives us a value of approximately **0.087**, which indicates a low explanatory power for the model. This conclusion logically makes sense, as passenger class alone may not fully explain the passenger's survival outcome.

ROC Curve and AUC Evaluation

Figure 5: ROC Curve for Logistic Regression



The ROC Curve for our logistic regression model demonstrates the trade-off between it's ability to give us accurately predicted outcomes of survival and it's tendency to tell us a passenger would have survived when they really died. The AUC (Area Under the Curve) is used to check the degree to which our model gives us a true positive vs. a false positive, and we find that **AUC = 0.681**. This tells us that the model performs better than random guessing but still leaves significant room for improvement.

Conclusion

Research Question #1 Results

Our first research question sought to investigate the differences in survival rates between female minors and male minors. Our null hypothesis claimed there was no difference in the average survival rate of male and female minors while our alternative hypothesis claimed there were significant difference based on the gender of Titanic passengers under the age of 18.

We began by checking assumptions using a Levene's Test and Shapiro Wilk Test. While the Levene's test failed to reject our hypothesis of equal variances, satisfying the assumption of homogeneity of variances, our Shapiro-Wilk Test did not satisfy the assumption of normality. This meant that our two groups did not have normal distributions, threatening the validity of our hypothesis test. In order to counteract this, we chose to conduct a Wilcoxon Rank-Sum Test, a non-parametric method that did not require normality as an assumption.

From the Wilcoxon Rank-Sum Test, we obtained a significant p-value of **7.111e-05**, which allows us to reject our null hypothesis at the 0.05 significance level. In regards to our hypothesis, this test tells us that there are significant differences between the avergae survival rate of passengers under the age of 18 based on

gender. Utilizing visualizations, we were able to create plots that depict the survival rates of minors based on gender. In **Figure 1** and **Figure 2**, there are significant differences in the average survival rate of male minors and female minors, which correlate with the results from our hypothesis test. Based on the visual plots, we can see that the average rate of survival among female minors was much higher than male minors (**Figure 1**).
In regards to our data set, the results of this hypothesis test tell us that gender did in fact have a significant impact on the survival rate of passengers under 18 that were aboard the Titanic at the time of the disaster. Despite children being prioritized upon evacuation when the Titanic was sinking, the gender of the minor did effect their chances of survival.
This difference in group distributions found in **Figure 1** and **Figure 2** could slightly affect the validity of the Wilcoxon Rank-Sum Test. However, the test is robust to moderate differences in distributions, and as such we can remain confident in our conclusion that there was a difference in the mean survival rates for male and female minors on board the Titanic, with female minors having a higher rate of survival.

Research Question #2 Results

Our second research question aims to analyze whether the predictor variable, a passenger's class, was increasingly impactful on their chances of survival when the Titanic sunk. A passenger's class was determined by which ticket they bought and determined their living quarters, access to different areas of the ship and more, as stated in the introduction. Using hypothesis testing, we were able to analyze whether a passenger's class significantly increased their chances of survival.
We begun by testing which assumptions were satisfied based on the data. Under further inspection using the Shapiro-Wilk Test and Levene's Test, we concluded that both assumptions of normality and homogeneous variances were not satisfied. Based on these conclusions, we opted to use logistic regression to test our hypothesis.
From the logistic regression model, we obtained results for the survival rate of each class while utilizing the first class as a reference variable. Our results shows that passengers in the first class were 1.7 times more likely to survive than other classes. The second class were 47.2% less likely to survive the Titanic sinking when compared to first class passengers and the third class was 81.2% less likely to survive compared to first class passengers. The p-value for each class was below the significance level of $\alpha = 0.05$ meaning that we could reject the null hypothesis and conclude that a passenger's class did have a significant impact on their chances of survival.
Although our results proved to be significant, after evaluating our model we found that the overall model did not explain a large amount of the variance in the data. As previously stated, this was expected as a person's class may have an impact on their chances of survivability but it is not enough information to predict their rate of survival as a whole. However, it does provide more information that could assist in predicting the survival of a passenger rather than using random guessing.
In conclusion, by using logistic regression to model the relationship between the passenger classes and the survival of thepassenger's aboard the Titanic, we found that there is a statistically significant difference in the odds of a passenger surviving based on which passenger class they are in. We did find that our model has low explanatory power and only has a moderately acceptable predictive accuracy, but still performs better at predicting survival than a random model. Taking this into consideration, our model was able to confirm that the different passenger classes had different rates of survival we would want to consider adding more predictors to the model in an attempt to increase it's explanatory power if we wanted to do more analysis of the passengers survival rates.