Preregistration

Sex and social class determined survival on the Titanic

Lina Aragon-Baquero[1], Kristen Bill[2], Leah D'Aloisio[3], Jack Goldman[4], Briar Hunter[5,6]

[1] University of Waterloo

[2] Wilfrid Laurier University

[3] University of British Columbia Okanagan

[4] University of Toronto

[5] Laurentian University

[6] Toronto Zoo

06. October 2021

Study Information

Title	Sex and social class determined survival on the Titanic
Description	It is well known that survivorship upon the sinking of the Titanic was heavily influenced by sex and class of the individual. Previous studies reveal the role of sex and class was a strong predictor of survival due to policies and social/physical barriers put in place prioritizing first class and women/children Bruno S. Frey, Savage, & Torgler (2009) Hall (1986). This study seeks to replicate previous findings while also statistically testing the degree to which these variables influenced survival.

Therefore, using an alternative dataset, we will run a mixed effects model to test the effects of class and sex on survival on the Titanic.

Hypotheses Similar to other studies done on Titanic survivorship data, we predict class and sex will have a strong impact on survivorship, with women/children and higher class passengers (i.e. first class) being more likely to survive than men or third class passengers.

Design Plan

Study type This is a retrospective observational study using a dataset based on known outcomes from the historical event of the sinking of the Titanic.

Blinding Not applicable.

Study design This study will utilize retrospective data from a one-time natural event, therefore data collection is constrained by information gathered/documented from passengers and survivors. This information was then compiled into the dataset that will be used for this study. On account of this being a one-time event with a defined number of people, our design is a cross-sectional observation study.

To understand the driving factors of survival aboard the titanic, we will use a mixed effect model approach in which sex and class will be fixed factors and age a random factor: Survival ~ Class + Sex + (1|Age).

Randomization Not applicable.

Sampling Plan

Existing data As of the date of this submission, we have accessed and analyzed some of the data relevant to the research plan, including preliminary analysis of variables, calculation of descriptive statistics, and observation of data distributions.

Explanation of existing data	The data we will use is a dataset retrieved from https://github.com/paulhendricks/titanic. The data provides information on the fate of passengers of the Titanic including variables such as class, age, sex, ticket fare, and nationality. This dataset is an array resulting from cross-tabulating 2201 observations. Prior to conducting our own analyses, we will not read any papers performing statistical analyses and thus remain unaware of any summary statistics on the data. We have only read prior articles using a different dataset to discuss effects of social class and sex on survival. These papers also gathered information based on interviews of the survivors following the event.
Data collection procedures	This study will use the publicly available titanic dataset found in R (https://github.com/paulhendricks/titanic). The principal source for the data in this package was built on previously existing data in which two authors re-visited the death count to best represent the historical event in this dataset Dawson (1995) Simonoff (1997). These contributions then led to the creation of this simulated data, which is formatted in a machine learning context for training purposes. The study population we will use is not reproducible, as it relied on the specific circumstances of the Titanic sinking, which we hope never happens again. Comparative analyses, however, can be performed on similar events such as the sinking of the Lusitania B. S. Frey, Savage, & Torgler (2010), but the exact conditions are unlikely to be repeated.
Sample size	The sample size of our study will be based on available passenger data. The total sample size in the dataset is 1309 subjects. Since 263 passengers are missing information about their age, we will be using a sample size of 1046.
Sample size rationale	Since these data are obtained from an historical event, we are constrained to the information made available to the public. The sample size in this dataset is not true to the total number of passengers and crew on board due to lack of records obtained prior to boarding the Titanic. Furthermore, not all information is provided for each passenger in the dataset, as certain details were only able to be obtained from the survivors. Since we are not adding observations from this historical event, we will not be doing a power analysis. After removing missing variables, we will use 1046 passengers from the dataset.

Stopping rule Not applicable

Variables

Manipulated variables Not applicable.

Measured variables

The measured variables will be:

- Passenger class on the boat = (1 = 1st; 2 = 2nd; 3 = 3rd)
- Survival = (0 = No; 1 = Yes)
- Sex (those of uknown sex were added to male category)
- Age

Other variables provided in the dataset, but will not be included in this confirmatory analysis are:

- Number of family members = number of siblings/spouses and parents/children on board
- Fare = Passenger Fare in British pounds
- Boat number = Which lifeboat a passenger was on
- Home/Destination = Destination of travels

Indices To understand the summary statistics of the Titanic data, means will be extracted from the raw data. No other indices will be used for these data. Estimated marginal means will be used for multiple comparisons of simple logistic regressions.

Analysis Plan

To understand how social class and gender (independent) drive survivorship (dependent) on the Titanic, we will use general mixed effect models (GLMMs) with age

as a random effect in the model (R package: lme4). We will test that accounting for age as a random effect improves the model. Interactions between gender and social class will also be explored to assess whether the interaction of the two variables are driving the outcome. Assumptions of the GLMMs will be confirmed.

Statistical models

To understand how social class and gender drive survivorship on the Titanic, we will use generalized mixed effect models with age as a random effect in the model (R package: lme4). Interactions between gender and social class will also be explored to assess whether the interaction of the two variables are driving the outcome. Multicollinearity will also be assessed using variable inflation factor (VIF) (R package: car). Chi-squared test for comparisons of models will be used to test the differences among models. Estimated marginal means (R package: emmeans) with Bonferroni adjustments will be used for multiple comparisons for simple logistic multiple regression models. Model prediction assessment will be done using confusion matrices.

Transformations

No transformations will be used for these data unless transformations are needed to meet assumptions. There are several variables that are binary and will be coded. The dependent variable survived will be coded as "yes" and "no." Gender, which is a predictor variable, will be coded as "male" or "female."

Inference criteria

We will use the standard $p<.05$ criteria from the Chi-squared analysis to determine the most parsimonious glm model, which would suggest that the models are significantly different from those expected if the null hypothesis were correct. AIC and BIC values will also be compared between models to further assess which model is the most parsimonious model. The post-hoc estimated marginal means analysis, adjusted using Bonferroni.

Data exclusion

Passengers who did not have their age documented in the dataset will be excluded from the study. Outliers will be included in the analysis. No other checks will be performed to determine the eligibility for inclusion of data.

Missing data Passengers who did not have their age in the dataset will be automatically excluded from the analysis.

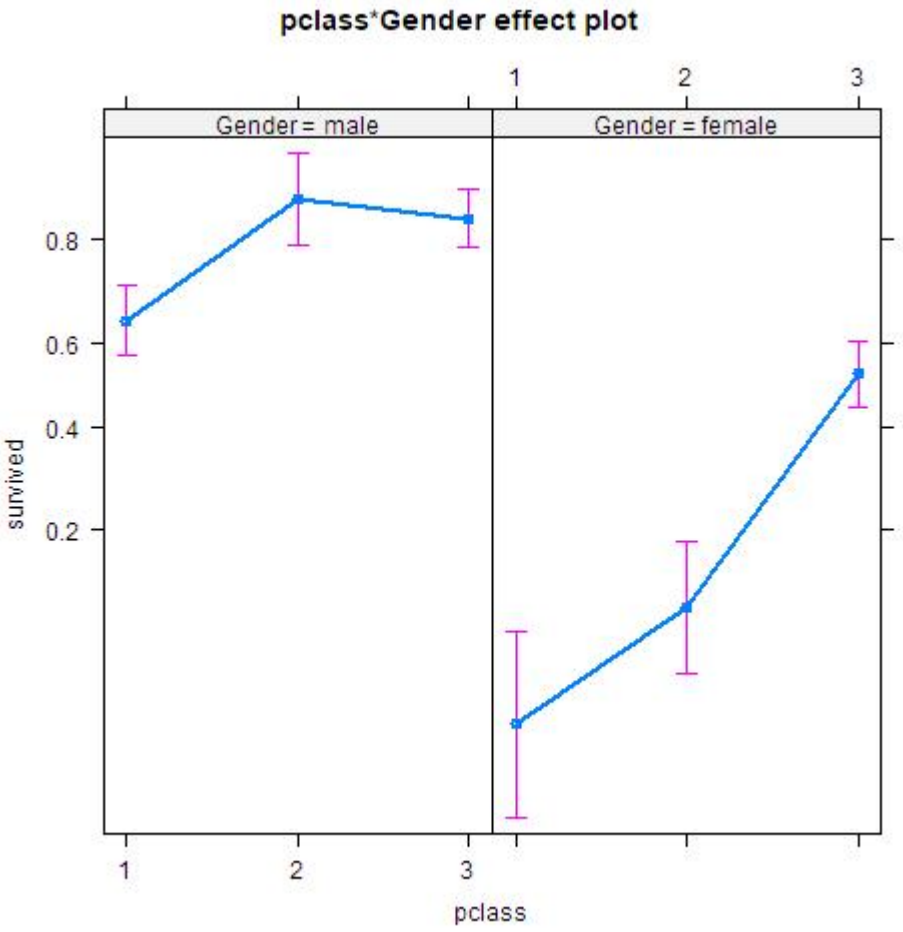

Figure 1: Interactive effects for most parsimonious Titanic model predicting survivability using gender and class as interactive factors.

Exploratory analyses (optional)

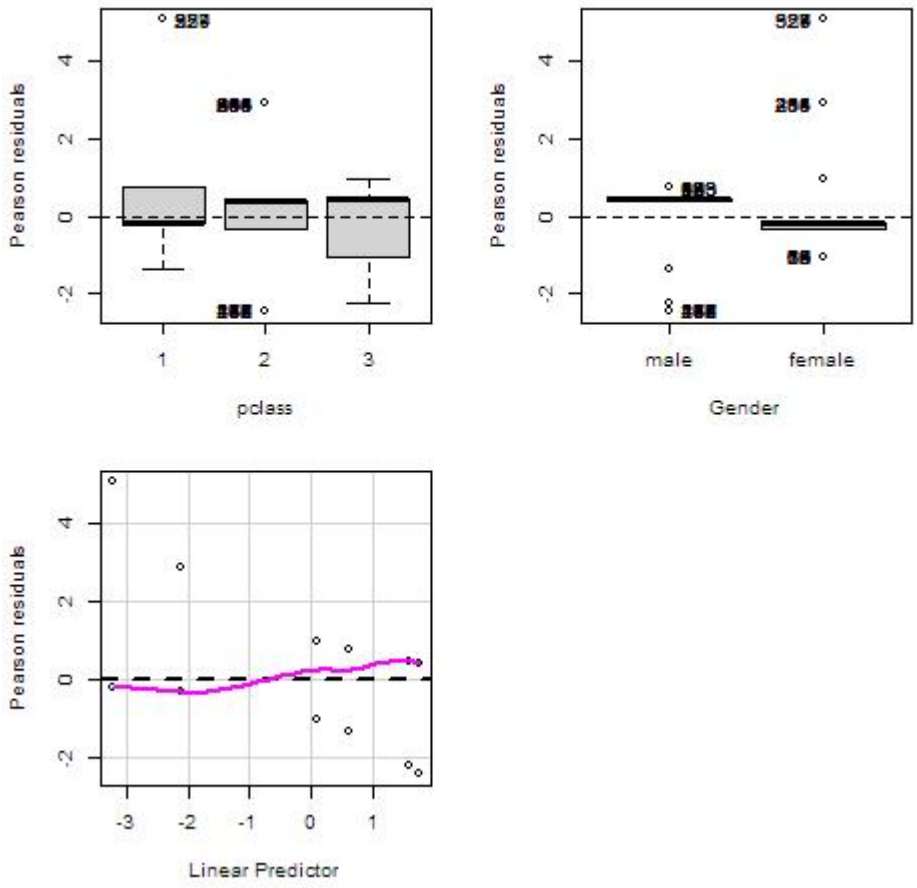

Figure 2: Residual plots for the Best Titanic model. These figures visualize the residuals to see if there are any differences in the variability of residuals as the value for each predictor variable increases.

References

Dawson, R. J. MacG. (1995). The "Unusual Episode" Data Revisited. *Journal of Statistics Education*, *3*(3), 7. doi:10.1080/10691898.1995.11910499

Frey, Bruno S., Savage, D. A., & Torgler, B. (2009). CESifo Working Paper no. 2551, 32.

Frey, B. S., Savage, D. A., & Torgler, B. (2010). Interaction of natural survival instincts and internalized social norms exploring the Titanic and Lusitania disasters. *Proceedings of the National Academy of Sciences*, *107*(11), 4862–4865. doi:10.1073/pnas.0911303107

Hall, W. (1986). Social class and survival on the S.S. Titanic. *Social Science & Medicine*, *22*(6), 687–690. doi:10.1016/0277-9536(86)90041-9

Simonoff, J. S. (1997). The "Unusual Episode" and a Second Statistics Course. *Journal of Statistics Education*, *5*(1), 4. doi:10.1080/10691898.1997.11910524