**A Logistic Analysis to Assess Survival Chance among Passengers based on Class and Gender on Titanic**

**Abstract**

Due to the “lady first” principle during the rescue of Titanic, we observe high survival chance for female passengers on Titanic. High survival chance also exists among first-class passengers because they were the priority during the rescue mission. However, if we compare third-class female and first-class men, which factor, gender or class, would be more crucial on survival chance? The purpose of this study is to examine the significance of class and gender effect on survival chance on Titanic and to answer the question above. We find that both effects are statistically significant, and survival chance for third-class female are higher than first-class men but might not be significantly. The effect of class might be more obvious among female group than that of male group.

**Introduction**

The well-known movie Titanic in 1997 directed by James Cameron depicts the disaster happened during the first, and also the last, trip of the gorgeous and gigantic Cruise Titanic. The movie leaves audience a romantic impression between two main characters and the heroic “lady-first” principle when facing the upcoming death. However, the inequality between social classes during the last hours of Titanic might disgrace the above impressions. There is higher survival rate of passengers from first and second class compared to the third-class passengers, highlighting the demonstration of social class affects mortality rate (Pearson, 2008). Therefore, a question was proposed: what about the survival rate of female passengers of the third class? Is the survival rate of the third-class female significantly larger than that of the first-class male? Is “lady-first” still a golden principle when confronting social class difference?

It is well-known that the chance of surviving the sinking distributes unequally among passengers (Hall, 1986). The purpose of this research is to examine details of the difference of survival rate among sex and passenger classes. Through the statistical modeling techniques, we aim to answer the key question: is social class have stronger effect than sex on the survival rate of titanic passengers? The research would not make any definite claim about the Titanic incident.

**Methods**

**Details of Data**

The present study is based on the 1313 observations of passengers on the Titanic comprising of 462 females and 851 males. Among them, 322 passengers owned first-class tickets (143 female and 179 male), 280 passengers owned second-class tickets (107 females and 173 males) and 711 passengers were third-class (212 females and 499 males). Ages of passengers range from 0.17 to 75 years old. Age, Sex, PClass are factors. Age has 75 levels and 557 missing values. Sex has two levels indicating female and male, coding “1” as female and “2” as male. PClass contains three levels with “1st” indicating first-class, “2nd” indicating second-class and “3rd” indicating third class.

The variable Survival is an integer variable with values 1 indicating survived passengers and 0 indicating death. There are 450 passengers survived in the data, which comprises around 30% of all passengers.

The data is obtained from the website of Encyclopedia Titanica (Hinde, Philip, 1998).

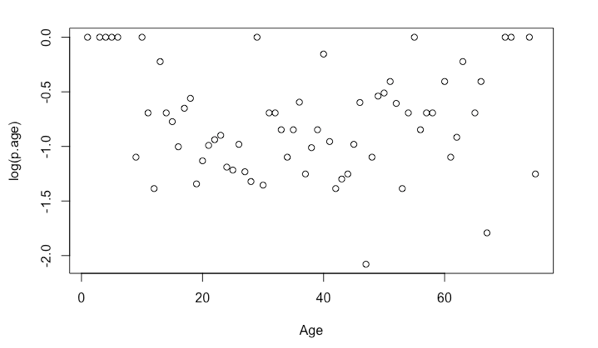
**Statistical Techniques:**

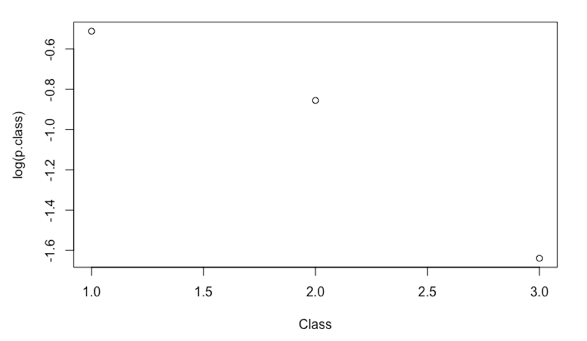
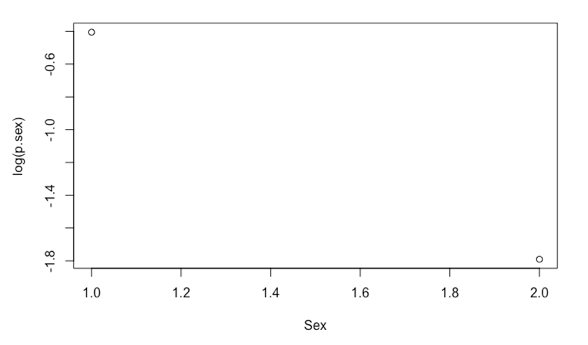
Due to the presence of binary data of variable Survived, we use logistic regression to analyze the survival rate among different characteristics of passengers. The present study utilizes the linear link between the log of odds of survival probabilities (logit function) and variables Age, Sex, and PClass in order to test the significance of these variables.

Let the logistic regression is defined by the following:

log(Prob[Survived] / (1- Prob[Survived])) = b0 + b1\*Age + b2\*Sex + b3\*PClass

For determining the function form of these independent variables, plots of log odds of survival rate against each variable are used in figure 1. The log odds ratio was calculated for each level variables. When fitting the model, we re-level the variable Sex and PClass so that the baseline in models describes the log odds of survival rate of third-class female.



Figure

First problem existed in the data is that Age is coded as a factor with 75 levels. With such many levels, the statistical significances are compromised, and the model will have more than 75 variable which makes the modeling process redundant. Two methods are presented. The first one is to re-group the age factor to 5 new levels based on human life cycle (“0-12”, “13-18”, “19-29”, “30-59” and “60-”); the second method is to convert Age to integers. Two models are fitted based on above methods and being compared by using AIC. The Analysis of Deviance Chi-Square Test will be used to test if predictors significantly improve the prediction power from the null model.

The second problem of the data is that variable age has 557 missing values. With nearly 42% of data missing. There is potential information of Age’s distribution is missing. when modeling process simply delete observations with missing data. The study uses multiple imputation to impute missing values by randomly drawing from the conditional distribution of the target variable given other variables. Multiple imputation applies bootstrapping techniques while considering the target variable’s correlations with other variables. To properly account for missing values’ variabilities, the study set the repeating time of bootstrapping to be 100 multiplies the percentage missing values of the total number of the variable, which is 40 in this study (Harrell, 2015). To check if the imputed values align with actual available values in Age, we plot both distributions of imputed values and non-missing values together in figure 2 and the plot suggests that their distributions appears similar patterns.

Means of imputed values are incorporated into the original data set afterwards, and two models based on the two methods: one is modeling with factor age and another one is modeling with age as integers, are fitted with the new dataset. Model comparison is conducted also by comparing AIC. The Analysis of Deviance Chi-Squared Test will be used to check if the model is significantly improved from null model.

Model Assessment is conducted by evaluating the pattern in deviance residual against fitted values plots and by checking estimated dispersion through Pearson residuals. The pattern of residuals is detected by running LOWESS smoother. Dispersion of the model is estimated by Pearson residuals divided by the model’s degree of freedom. The study compares the dispersion calculated with 1 and reports inequality.

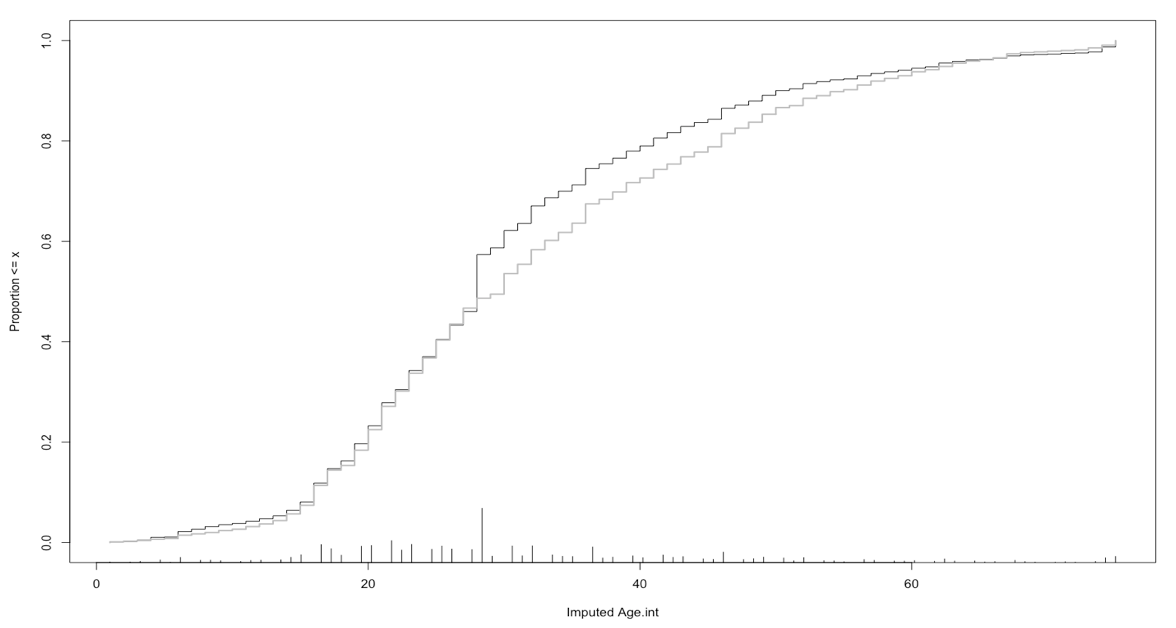


Figure Distributions of imputed and actual ages. The black curve is for imputed values and the grey one indicates actual values

**Results**

For models fitted without imputation missing values, m.logit, the model with 5 levels factor Age, is preferred because it has less AIC 718.22 compared to 727.99 of the model with Age as integers, and the p-value of the Analysis of Deviance Chi-Squared test shows that predictors in m.logit significantly reduces the deviance from null model. The summary table of the m.logit is given in table 1. All variables have at least 95% statistical significance.

The intercept indicates female, third-class passengers of age 0-12 has survival odds, 4.82. The survival odds will decrease by 76.5% if female, third-class passengers are “Age13-18”, decrease by 75.5% for “Age19-29”, decrease by 80% for age30-59 and 87.7% for age of more than 60. Male, third class passengers of age 0-12 has 92.8% less of survival chance. Female, first class passengers of age 0-12 could increase the survival chance by 180.6% while the increase could be 67.74% for female second class passengers for age 0-12.

For models fitted with imputation missing values, newfit.factor, the model with variable Age as factors, is preferred because its less AIC, 1195.2, compared to 1198 of the model with Age as integers, and the p-value of the Analysis of Deviance Chi-Squared test shows that predictors in newfit.factor significantly reduces the deviance from null model. The summary table of the newfit.factor is given in table 2. According to table 2, all coefficients except two of Age 13-18 and Age 19-29, has at least 95% statistical significance.

The intercept indicates female, third-class passengers of age 0-12 has survival odds, 2.53. The survival odds will decrease by 49.3% if female, third-class passengers are “Age13-18”, but the value is not statistically significant. Age 19-29 could decrease 58.9% chance of survival, but the value also does not have statistical significance. There is 76.3% less chance of survival for age 30-59 and 77.7% less chance of survival of ages more than 60. Male, third class passengers of age 0-12 has 91.2% less of survival chance. Female, first class passengers of age 0-12 could increase the survival chance by 413.5% while the increase could be 152.04% for female second class passengers of age 0-12.

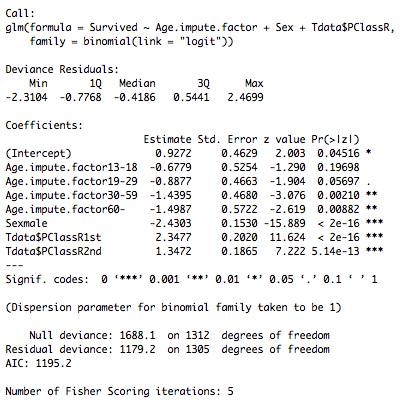
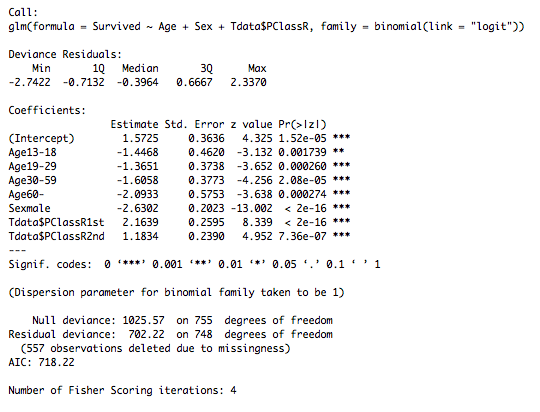
****

Table 2 Summary of newfit.factor

Table 1 Summary table of model, m.logit

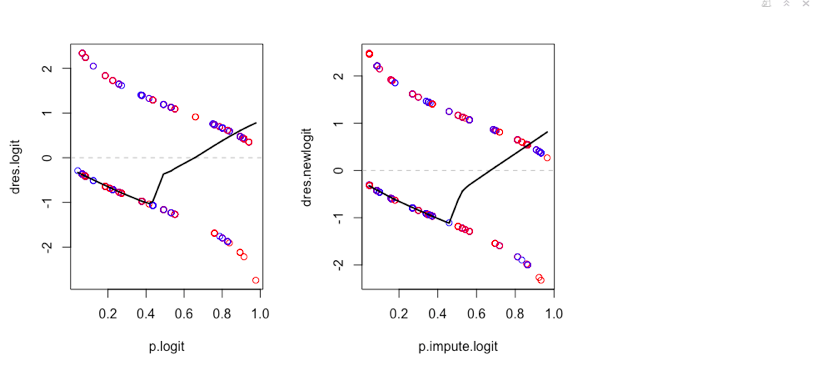
**Model Validation:**

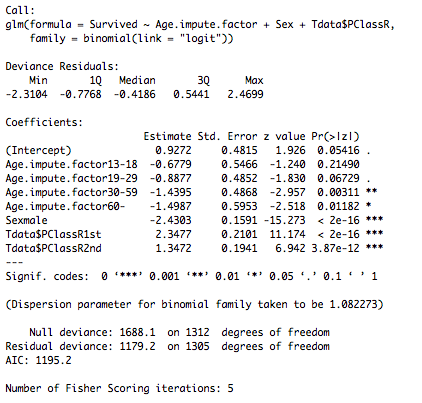
For model validations, plots of deviance residuals against fitted values of both models are drawn in Figure 3. The LOWESS smoother from both models show a curved pattern, which suggests that there exists a missing higher order term in the model. However, since the both models are fitted with only factors, it is inappropriate to include any higher order term.

Estimated dispersions of both models are calculated by the formula defined below, N means the total observation number, P is the number of predictors:

sum (Pearson-Residuals ^2) / (N – P – 1)

The estimated dispersion of m.logit is 1.106653 and for newfit.factor is 1.082273. They are both greater than, but at the same time, very close to one. The over-dispersion seems not to be significant. However, when re-summarize the model while taking into account of the estimated dispersion, the intercept term of newfit.factor becomes no longer significant at 95% confidence level. It suggests that the survival odds of female, third class passengers of age 0 – 12 is no longer significantly different than 1.



Figure Deviance Residuals against Fitted Values Plots. m.logit is the left one and newfit.factor is the right one

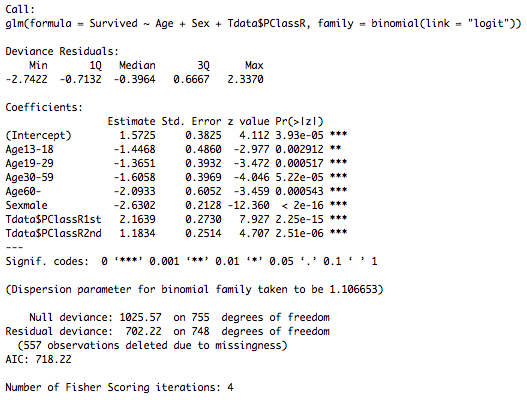


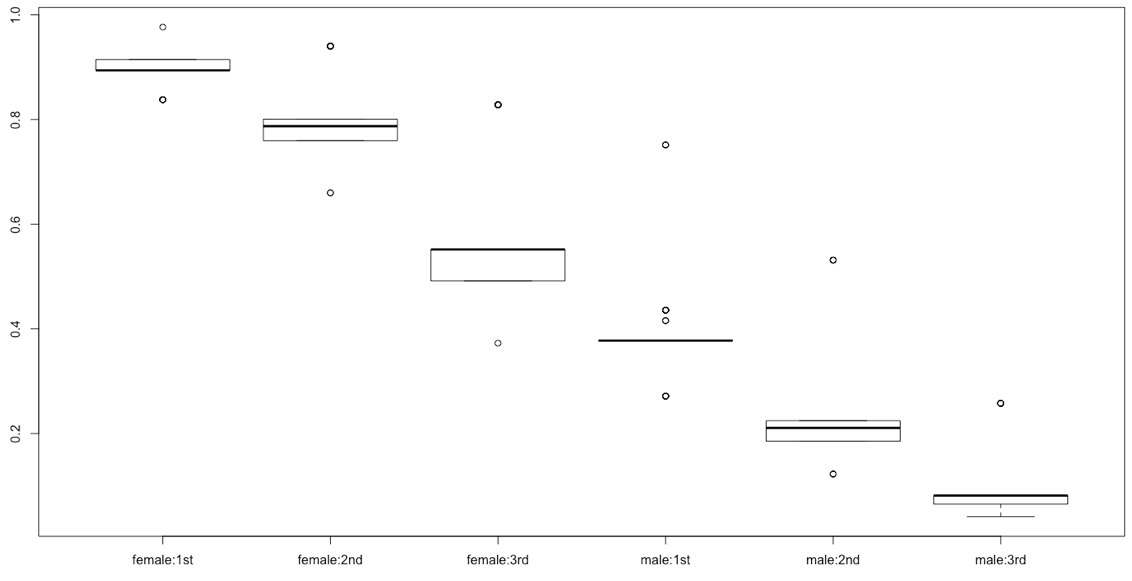
Table 3 Summary table of m.logit (left) and newfit.factor (right) after taking into account of estimated dispersion.

**Prediction**

The survival odds of female, third-class passengers predicted by the m.logit is exp(1.5725 -1.4468 – 1.3651 – 1.6058 - 2.0933), 0.007165338. The survival odds of male, first-class passengers predicted by m.logit is exp(1.5725 -1.4468 – 1.3651 – 1.6058 - 2.0933 - 2.6302 + 2.1639), 0.004494953. Therefore, the survival chance of female, third-class passengers in Titanic is 0.007165338 / (1 + 0.007165338) = 0.711% and the survival chance of male first-class passengers in Titanic is 0.004494953 / (1 + 0.004494953) = 0.447%. The difference of the survival chance between female third-class passengers and male first-class passengers is 0.27%, which is 38% of the survival probability of female third-class passengers.

The survival odds of female, third-class passengers predicted by the newfit.factor is exp(0.9272 - 0.6779 - 0.8877 - 1.4395 - 1.4987), 0.02797064. The survival odds of male, first-class passengers predicted by m.logit is exp(0.9272 - 0.6779 - 0.8877 - 1.4395 - 1.4987-2.4303 + 2.3477), 0.02575311. Therefore, the survival chance of female, third-class passengers in Titanic is 0.02797064 / (1 + 0.02797064) = 2.72% and the survival chance of male first-class passengers in Titanic is 0.02575311 / (1 + 0.02575311) = 2.51%. The difference of the survival chance between female third-class passengers and male first-class passengers is 0.21%, which is 7.72% of the survival probability of female third-class passengers.

In figure 4, predicted log odds of survival chance by both models are plotted based on combinations of sexes and classes. We see that the log odds of survival chance exist obvious difference between sexes no matter what class the individual is. However, first



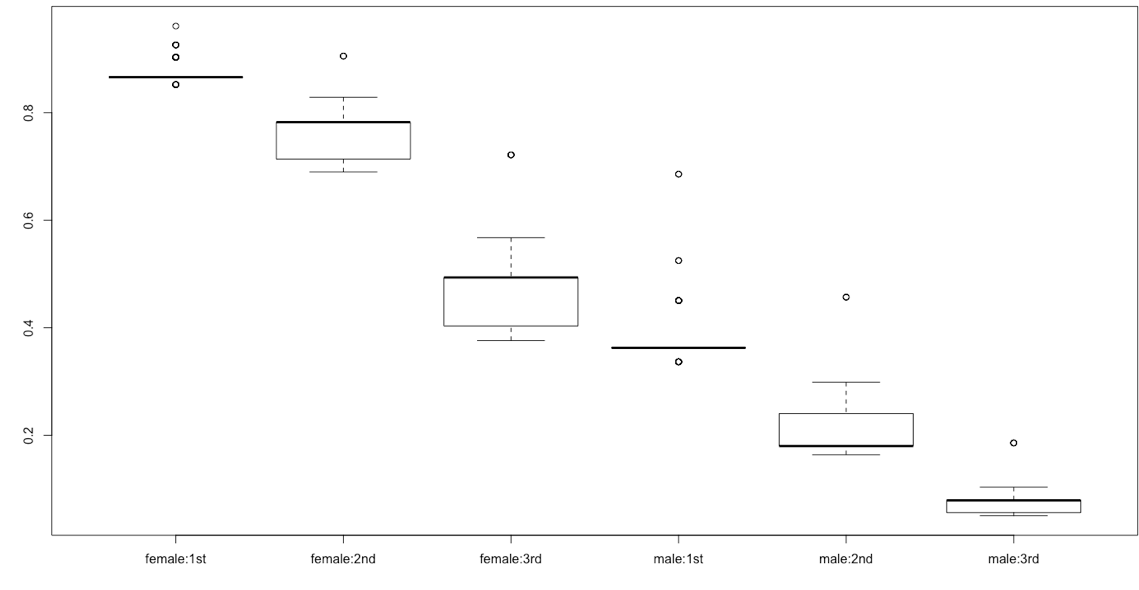


Figure Prediction plot of sex:class. m.logit above the newfit.factor one

class female has greater log odds of survival chance than that of third-class female in both models’ predictions. The difference of log odds of survival chance among classes are more obvious in female group than in the male group.

**Discussion**

Survival chances of female third-class passengers and male first-class passengers predicted by two models are very different. The study suspects the source of difference is from the incorporation of imputation values to the newfit.factor model. When fitting m.logit, R-Studio will adopt case-wise deletion for generalized linear modeling, which means, only observations with no missing value in any variable could be used for analysis. The case-wise deletion might diminish the statistic power of prediction due to the possibility of non-randomness for the reason missing (Roth, P. L, 1994). In Titanic dataset, the proportion of missing value in Age variable is nearly 42%, which greatly reduces the sample size and prediction power. With the limit of knowledge about the reason of missing values, it is hard to judge the randomness.

Multiple imputation is a technique that is designed to minimize the bias when dealing with missing values. However, the process requires the proper specification of imputation model, or bias may be created, and this issue is mostly noticeable for factor variables (Horton and Kleinman, 2007). This study adopts aregImpute function of HMisc package in R for the multiple imputation process and have no idea what model or method the function uses to impute the missing value. It is impossible for this study to identify the potential bias created by multiple imputation.

With potential possibility of bias from both models, this study has limited knowledge to test which model is more significant or valid for the prediction. Both results are reported. First-class male has 38% less survival chance than third class female based on prediction from m.logit (case-wise deletion), but only 7.72% less survival chance than third class female based on prediction from newfit.factor (multiple imputation). The difference in the survival chance between two groups are more obvious and probably more significant in the prediction of the multiple imputation model. It is difficult to tell which value is more reliable.

For model validation, the curved pattern in the deviance residual against fitted values plots might also suggests inappropriate link function. This needs a further investigation.

**Conclusion**

Sex and Class are both significant factors contributing to the survival chance on Titanic. Based on m.logit, if the sex is male, the odds of survival will be exp (-2.6302) = 7.21% of female’s survival odds. If the individual is from first class, the survival odds will be exp (2.1639) = 871% of survival odds for passengers from the third class, and for second class, the survival odds will be 326.55% of the third-class passengers’ survival odds.

Based on newfit.factor, if the sex is male, the odds of survival will be exp (-2.6302) = 8.798% of female’s survival odds. If the individual is from first class, the survival odds will be exp (2.1639) = 956.59% of survival odds for passengers from the third class, and for second class, the survival odds will be 368.95% of the third-class passengers’ survival odds.

The class difference in terms of survival chance is more obvious in female group than male group. Third-class female therefore, has less chance to survive since they are third-class, but their gender would eliminate this disadvantage. When comparing difference between survival chances of third-class female and that of first-class male, the male group has, 38% or 7.72%, less survival chance compared to third-class female based on two models. Currently, this study could not tell which value is more accurate and therefore, could not draw any conclusion from results.

**Reference**

[Pearson, Jay](https://www.psc.isr.umich.edu/people/profile/634/Jay_Pearson). 2008. "Can't Buy Me Whiteness: New Lessons from the Titanic on Race, Ethnicity, and Health." *Du Bois Review*, 5(1): 27-47.

Hall W. Social class and survival on the S.S. Titanic. Soc Sci Med 1986;22:687–90.

Hinde, Philip (1998). Encyclopedia Titanica. <http://www.encyclopedia-titanica.org/titanic_passenger_list/>

Harrell, Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. 2nd ed. Vol. 3.8. Series 692. Cham: Springer, 2015.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537-560. [http://dx.doi.org/10.1111/j.1744-6570.1994.tb01736.x](http://psycnet.apa.org/doi/10.1111/j.1744-6570.1994.tb01736.x)

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 79–90. http://doi.org/10.1198/000313007X172556