Modeling Titanic Survival*

Qiushi Yan

This case study showcases the development of a binary logistic model to predict the probability of survival in the loss of Titanic. I demonstrate the overall modeling process, including preprocessing, exploratory analysis, model fitting, adjustment, bootstrap internal validation and interpretation as well as other relevant techniques such as redundancy analysis and multiple imputation for missing data. The motivation and justification behind critical statistical decisions are explained, touching on key issues such as the choice of a statistical model or a machine learning model, using bootstrap to reduce selection bias, disadvantages of the holdout sample approach in validation, and more. The study finds that

- Females and kids are more likely to survive on average. But the *women and children first* policy was severly undermined in the first and third class.
- Crew members have the second highest survival probability, only smaller than first class passengers.
- Passengers who travelled with parents survive better, while other family relationship (sibling, children, spouse, etc.) is of lesser help.
- Nationality does not have a significant effect on survival.
- Passengers came aboard at Cherbourg have significantly higher survival likelihood than passengers boarded at other embarkation ports.

Introduction

The sinking of RMS Titanic has brought to numerous machine learning competitions a quintessential dataset. The unsinkable British passenger liner struck an iceberg on 15 April 1912 in her maiden voyage, and was eventually wrecked. More than 1500 people perished in the great loss. Decades of effort has been devoted to the study the tragic accident, in which one major interest for statistical inquiries is to model and predict the probability of survival given personal characteristics.

In recent years the web has witnessed the birth of numerous variants of Titanic data, with one primary source being Encyclopedia Titanica (1999), a site started in 1996 as an attempt to tell the story of every person that traveled the Titanic as a passenger or crew member. This case study grows from the most up-to-date version of the site's data as of October 2020, with the following columns available (Table 1). Source data and steps of data cleaning are elaborated in the data section in the appendix.

*This case study has been greatly inspired by Dr. Frank Harrell's similar example in his *Regression Modeling Strategies* (2015, chap. 12) book.