Modeling Titanic Survival

Qiushi Yan[a]

[a]Beijing, China

ARTICLE HISTORY
Compiled October 12, 2020

ABSTRACT
This short analysis showcases the development of a binary logistic model with spline transformations on predictors, to predict the possibility of survival in the loss of Titanic. It starts with exploratory analysis with descriptive statistics and visualization and then proceeds to modeling. I demonstrate the overall process of model fitting, adjustment, validation and intepretation as well as other relevant techniques such as multiple imputation for missing data. This analysis is fully reproducible with R code and text provided in supplemental materials.

KEYWORDS
logistic regression; multiple imputation; model validation

http://www.crema-research.ch/papers/2009-03.pdf
https://www.insider.com/titanic-secrets-facts-2018-4#at-the-memorial-of-frederick-fleet-one-of-the-lookouts-a-prankster-left-a-pair-of-bino
http://rpubs.com/edwardcooper/titanic1
https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic/report
https://www.kaggle.com/startupsci/titanic-data-science-solutions/comments

1. Introduction

The sinking of RMS Titanic brought to various machine learning competitions a quintessential dataset among others, in which one major interest is to predict possibility of survival given sex, age, class, etc. There are several variants of this data existed on the web, the one I use here comes by courtesy of Encyclopedia Titanica founded by Thomas Cason, namely `titanic3` with following variables available (table 1):

The raw data contains 1309[1] rows and 14 variables, with each row corresponding to the survival status of one passenger, alongside with his/her gender, age, family relations on board, ticket fare, etc. In the data there are 809 victims and 500 survivors in total.

Inspired by Dr. Frank Harrell's similar case study on the same topic in his *Regression Modeling Strategies* (2015) book, here I attempt to propose my own idea and

CONTACT Qiushi Yan. Email: qiushi.yann@gmail.com, website: https://qiushi.rbind.io

[1] Approximately 60% of all Titanic's passengers and crew, which is 2208

Table 1. Data Dictionary

.	Variable	Definition	Note
1	pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
2	survival	Survival Status	0 = No, 1 = Yes
3	name	Name	
4	sex	Sex	
5	age	Age	In years, some infants had fractional values
6	sibsp	Number of Siblings/Spouses Aboard	
7	parch	Number of Parents/Children Aboard	
8	ticket	Ticket Number	
9	fare	Passenger Fare	in Pre-1970 British Pounds
10	cabin	Cabin	
11	embarked	Port of Embarkation	Cherbourg, Queenstown or Southampton
12	boat	Lifeboat	
13	body	Body Identification Number	
14	home.dest	Home/Destination	

interpretation of model development that is as original as possible. To ensure reproducibility, all the analysis is done in R (R Core Team 2020) and RStudio with code and text provided in supplemental materials.

- quantify predictive ability of each predictors, i.e. which predictor is most dominant in determine whether a passenger will survive
- interactions between predictors
- whether the *Women and children first* policy is respected

2. Exploration

Before any analysis, let's first exclude those variables that bring little insight to prediction: `name`, `embarked`, `body`, `cabin`[2], `home.dest`. Then, a nice summary of all existing variables in the data is given by the `Hmisc::describe` function

```
cols <- setdiff(names(titanic_raw),
                c("name", "embarked", "body", "cabin", "home.dest"))
titanic_excluded <- titanic_raw[, cols]

latex(describe(titanic_excluded), file = "",
      size = "small", center = "none")
```

titanic_excluded
9 Variables 1309 Observations

pclass

n	missing	distinct
1309	0	3

```
Value        1st   2nd   3rd
Frequency    323   277   709
Proportion 0.247 0.212 0.542
```

[2]because this is primarily an identification for class

survived

n	missing	distinct	Info	Sum	Mean	Gmd
1309	0	2	0.708	500	0.382	0.4725

---

sex

n	missing	distinct
1309	0	2

```
Value      female   male
Frequency     466    843
Proportion  0.356  0.644
```

---

age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1046	263	98	0.999	29.88	16.06	5	14	21	28	39	50	57

```
lowest :  0.1667  0.3333  0.4167  0.6667  0.7500, highest: 70.5000 71.0000 74.0000 76.0000 80.0000
```

---

sibsp

n	missing	distinct	Info	Mean	Gmd
1309	0	7	0.67	0.4989	0.777

```
lowest : 0 1 2 3 4, highest: 2 3 4 5 8

Value         0     1     2     3     4     5     8
Frequency   891   319    42    20    22     6     9
Proportion 0.681 0.244 0.032 0.015 0.017 0.005 0.007
```

---

parch

n	missing	distinct	Info	Mean	Gmd
1309	0	8	0.549	0.385	0.6375

```
lowest : 0 1 2 3 4, highest: 3 4 5 6 9

Value         0     1     2     3     4     5     6     9
Frequency  1002   170   113     8     6     6     2     2
Proportion 0.765 0.130 0.086 0.006 0.005 0.005 0.002 0.002
```

---

ticket

n	missing	distinct
1309	0	929

```
lowest : 110152      110413      110465      110469      110489
highest: W./C. 6608  W./C. 6609  W.E.P. 5734 W/C 14208   WE/P 5735
```

---

fare

```
      n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
   1308        1      281        1     33.3    38.61    7.225    7.567    7.896   14.454
    .75      .90      .95
 31.275   78.051  133.650

lowest :   0.0000   3.1708   4.0125   5.0000   6.2375, highest: 227.5250 247.5208 262.3750 263.0000 512.3292
```

---

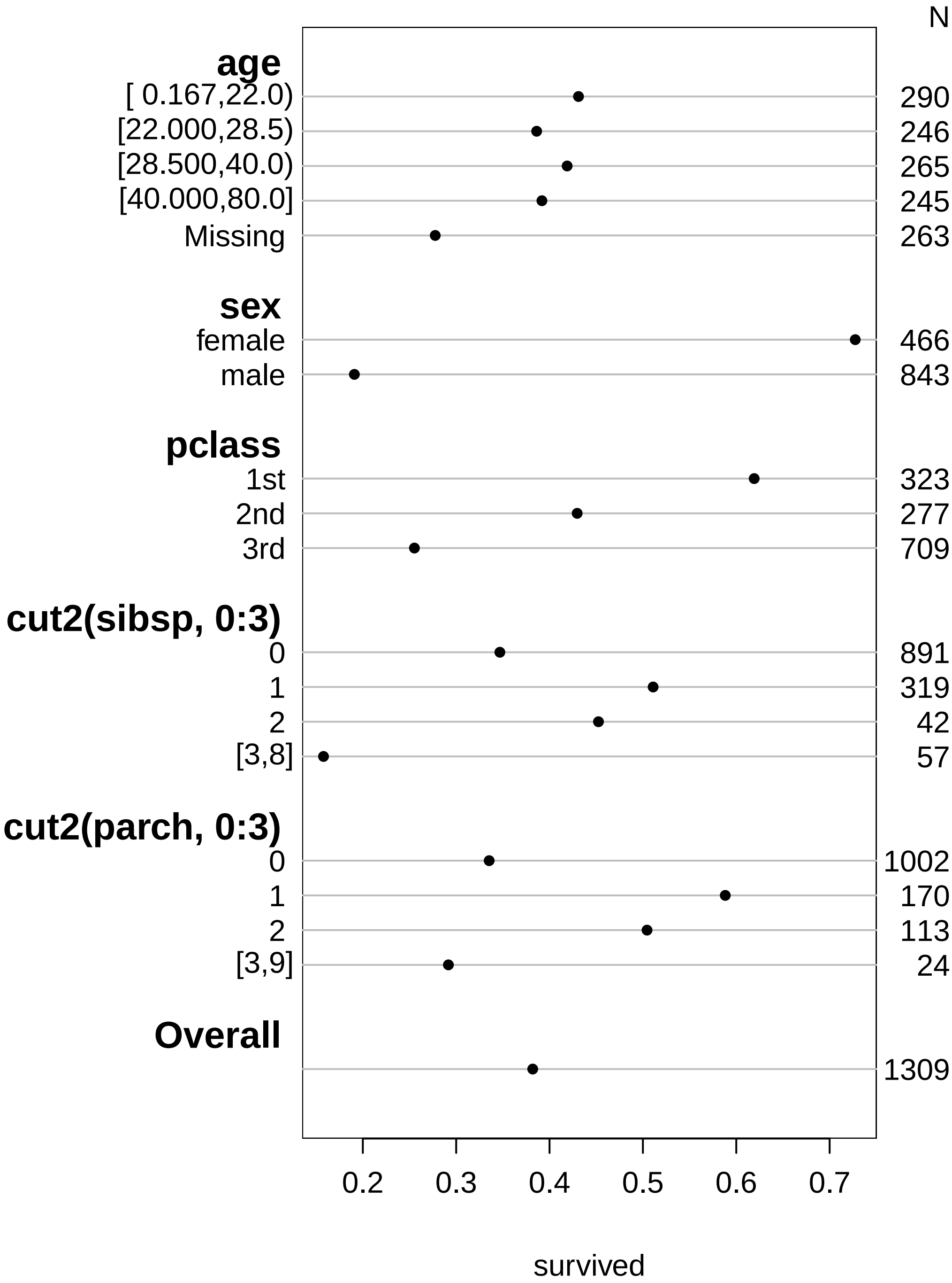
boat

n	missing	distinct
486	823	27

```
lowest : 1   10  11  12  13 , highest: A   B   C   C D D
```

There are several interesting patterns to notice

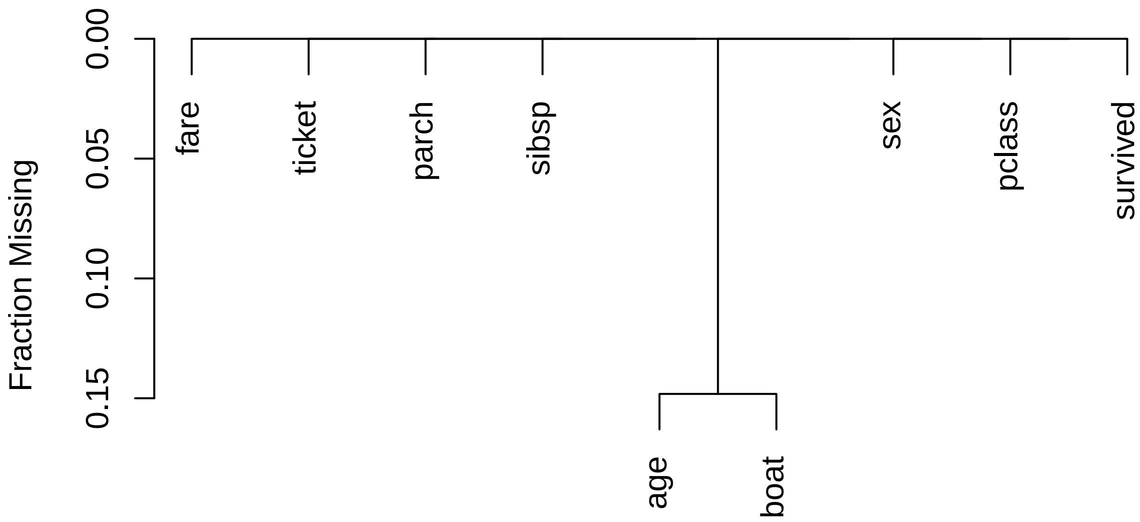
- there were nearly twice as many man as women.
- the problem of missing data:



male nearly twice as women: women are not allowed to travel alone

2.1. *Data missing patterns*

```
plot(naclus(titanic_excluded))
```



3. Modeling

3.1. *Initial model*

3.2. *Multiple imputatiin*

3.3. *Validation*

In the award-winning solution to this legendary dataset presented by IBM Watson, they used a holdout sample to validate their model. https://www.fharrell.com/post/split-val/

3.4. *The final model*

4. Discussion

A more detailed explanation of some of these measures is presented in the appendix.

5. Conclusion

Appendix A. Assessment of binary logistic model

This will be Appendix A.

References

Harrell Jr, Frank E. 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.