Modeling Titanic Survival

Qiushi Yan

[a]Beijing, China

ARTICLE HISTORY
Compiled November 4, 2020

ABSTRACT
This case study showcases the development of a binary logistic model to predict the possibility of survival in the loss of Titanic. I demonstrate the overall modeling process, including preprocessing, exploratory analysis, model fitting, adjustment, bootstrap 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. This analysis is fully reproducible with all source R code and text.

http://www.crema-research.ch/papers/2009-03.pdf

Who Survived Titanic? A Logistic Regression Analysis: https://sci-hub.do/https://journals.sagepub.com/doi/pdf/10.1177/084387140401600205

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

https://www.newscientist.com/article/dn22119-sinking-the-titanic-women-and-children-f

1. Introduction

The sinking of RMS Titanic brought to numerous machine learning competitions a quintessential dataset among others. After the “unsinkable” British passenger liner struck an iceberg in her maiden voyage on 15 April 1912 and was eventually wrecked, more than 1500 people perished. Decades of effort has been devoted to the study of the historic event, in which one major interest for statistical inquiries is to model and predict survival given a number of characteristics, since there was clear account that some people were allowed to get on the lifeboat first.

There are numerous variants of Titanic data existed on the web, with primary source based on 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 project is based on the most recent version as of October 2020, with following columns available (table 1). Source data and steps of data cleaning are elaborated in the data section in the appendix.

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

Table 1. Cleaned data with 2208 rows and 11 columns

Variable	Definition	Note
survived	Survival Status	0 = Lost, 1 = Saved
age	Age	In years, some infants had fractional values
gender	Gender	
class	Cabin class	1st, 2nd, 3rd or Crew
nationality	Motherland	from wiki passenger list
title	Title	Extracted from name
spouse	# of spouse on board	
sibling	Number of siblings on board	
parent	Number of parents on board	
children	Number of children on board	

After appropriate formatting and cleaning, the data at hand recorded the survival status 2208 Titanic travelers alongside his/her gender, age, companions on board, title, nationality, etc. There were 1496 victims and 712 survivors in total.

It is essential for every fruitful task of data analysis to first identify key questions of investigation that facilitates interpretation, however vague they are at the beginning. Then we can approach the core problem, filtering out trivialities, with statistical expression by abstraction. For our purposes, we could establish the following questions for which to quest

- To which degree is *Women and children first* policy respected? After the collision, the captain explicitly issued an order for women and children to be saved first.[1] Thus we should expect significantly higher proportion of females and children rescued than that in males and adults. If the opposite is true, that Titanic subjects behave more in line with the selfish *homo oeconomicus*, where everybody looked out for himself or herself and possibly even puts other people's lives in danger, then people in their prime with physical superiority would see higher probability of survival. This requires us to study gender and age effect.
- Did socio-economic advantages mean better chance of survival? If this is the case, passengers with higher financial means, i.e. who live in the first class are more likely to survive. Similarly, passengers from second class will have a higher change of survival than third class people. Cabin class's impact on survival status needs special notice here.
- For those who traveled alone with no companions (spouse, sibling, parent, children) on the vessel, is their survival possibility greater or less? On one hand, they are more likely to be in shortage of psychological and physical support. On the other hand, they would may be able to reach a life-saving decision faster without transaction cost and negotiation.
- Did English subjects receive any special care or given priority to aboard lifeboats? After all, Titanic was operated by British crew, and managed by British captain, masters and officers. Conversely, British nobility and elite
- Quantify interactions among various characteristics. Specifically, there are important interactions that need extra notice. For example, it has been widely studied in sociology and anthropology that human are sometimes driven by *procreation instinct* so that social norms would entail needs to protect females of

[1] Though there is no international maritime law enforcing this kind of chivalry.

reproductive age (Frey, Savage, and Torgler 2009).[2] Therefore, we could specify and study the interaction between age and gender. Another typical interaction is between offspring and gender. *Parental investment* suggest that women on average invest more in caring for their offspring than males. In times of a disaster, higher opportunity cost will alert females with offspring more than others, and make them seek more aggressively for changes to secure the children as well as themselves.

This case study has been greatly inspired by Dr. Frank Harrell's similar example in his *Regression Modeling Strategies* (2015, Chapter 12) book, here I attempt to propose my understanding and interpretation of model development that is as original as possible. To ensure reproducibility, all the analysis is done in R (R Core Team 2020) with code and text made public in this repo. A brief summary of each section is listed below

- Exploration. Use descriptive statistics to examine data distribution characteristics, data missing patterns and relative effects, followed by redundancy analysis to study dependencies among predictors. Finish with nonparametric loess regression exploring nonlinear trends.
- Model development. The key section in specifying, developing, validating and describing a binary logistic model, split into
  - Specification Prespecification of predictor complexity with a saturated main effect model.
  - Multiple imputation: Use predictive mean matching to impute subject's age, resulting in 30 complete datasets.
  - Model fitting, validation and calibration. Obtain pooled parameter estimates based on prespecified complexity and imputation results. Use bootstrap validation and calibration curve (the ".632" method) to study model performance and optimism.
  - Interpratation. Summarize the model with estimation and hypothesis testing, combined with graphical methods like partial effect plots and nomogram.
- Discussion. Model-based explanation to address some of our former questions.
- Conclusion. Conclusion and further study.

2. Exploration

2.1. *Descriptive statistics and data processing*

A graphical summary of of the data is given by the `Hmisc::describe` function. For numerical variables, a inline histogram is produced alongside summary measures such as the number of missing values and the mean. For discrete variables, we focus on the number of categories and their relative frequency.

```
# print a summary for the data
t %>%
  describe() %>%
  latex(file = "", size = "small", center = "none")
```

[2] The average peak reproductive period in females is between the ages of 16 and 35.

11 Variables 2208 Observations

survived

```
       n  missing  distinct     Info    Sum     Mean      Gmd
    2208        0         2    0.655    712   0.3225   0.4372
```

age

```
       n  missing  distinct     Info   Mean      Gmd   .05   .10   .25   .50   .75   .90   .95
    1497      711        71    0.999  30.18    14.31     8    17    22    29    38    47    54

lowest :  0.8  1.0  2.0  3.0  4.0, highest: 67.0 69.0 70.0 71.0 74.0
```

gender

```
       n  missing  distinct
    2208        0         2

Value      Female   Male
Frequency     489   1719
Proportion  0.221  0.779
```

joined

```
       n  missing  distinct
    2208        0         4

Value          Belfast   Cherbourg  Queenstown Southampton
Frequency          200         271         123        1614
Proportion       0.091       0.123       0.056       0.731
```

nationality

```
       n  missing  distinct
    2208        0         7

lowest : American English  Finnish  Irish    Other    , highest: Finnish  Irish    Other    Swedish  Syrian

Value      American  English  Finnish    Irish    Other  Swedish    Syrian
Frequency       246     1002       58      168      549       99        86
Proportion    0.111    0.454    0.026    0.076    0.249    0.045     0.039
```

class

```
       n  missing  distinct
    2208        0         4

Value        1st   2nd   3rd  crew
Frequency    321   270   709   908
Proportion 0.145 0.122 0.321 0.411
```

title

```
       n  missing  distinct
    2208        0         4

Value       Miss    Mr   Mrs other
Frequency    267  1590   212   139
Proportion 0.121 0.720 0.096 0.063
```

spouse

```
       n  missing  distinct     Info    Sum      Mean       Gmd
    2208        0         2    0.087     66   0.02989   0.05802
```

sibling

```
       n  missing  distinct     Info      Mean      Gmd
    2208        0         4    0.138   0.05752   0.1103

Value          0     1     2     3
Frequency   2101    91    12     4
Proportion 0.952 0.041 0.005 0.002
```

parent

```
       n  missing  distinct     Info      Mean       Gmd
    2208        0         3    0.079   0.03804   0.07441

Value          0     1     2
Frequency   2148    36    24
Proportion 0.973 0.016 0.011
```

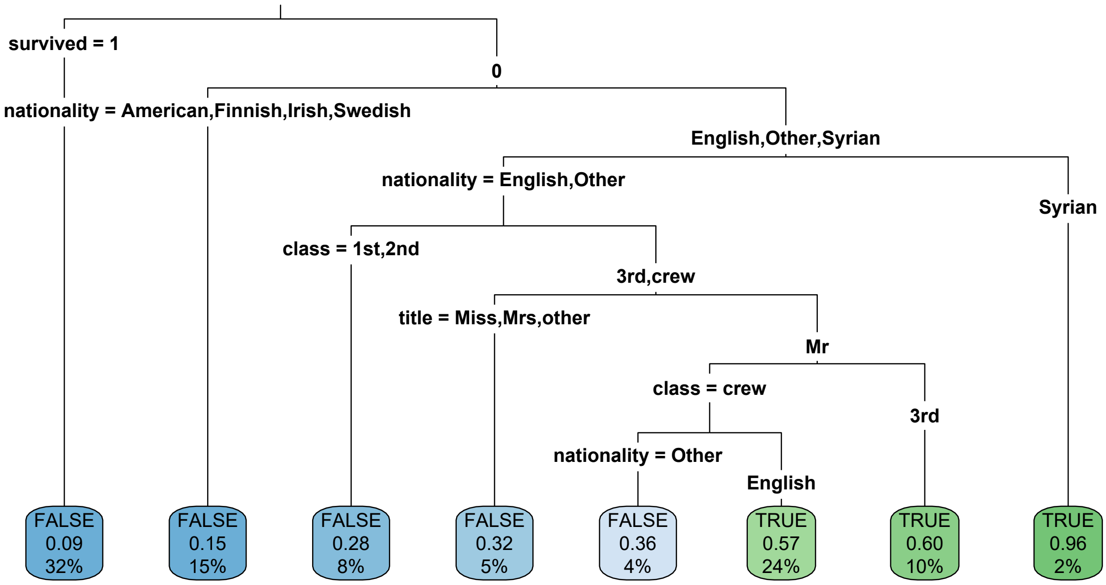


Figure 1. The decision tree for predicting `is.na(age)`, which finds strong patterns of missing related to class/department and gender (the Cherbourg node has very limited samples).

children

```
       n   missing   distinct    Info      Mean       Gmd
    2208         0          5   0.077   0.03895   0.07636

lowest : 0 1 2 3 4, highest: 0 1 2 3 4

Value          0     1     2     3     4
Frequency   2150    37    16     3     2
Proportion 0.974 0.017 0.007 0.001 0.001
```

There are several noteworthy patterns.[3]

Of special importance is the `age` variable, which has roughly 30% missingness. On the other hand, it has a nice distribution with 80% known observations falling between 14 and 50. For further examination of patterns of missing data, we could fit a decision tree (figure 1) to predict which type of subject tend to have missing ages. Generally, for some third class male passenger or crew, age is mostly to miss.

```r
na_tree <- rpart(factor(is.na(age)) ~ .,
                 data = t %>% mutate(survived = as.factor(survived)) ,
                 minbucket = 50)
# figure 1
rpart.plot::rpart.plot(na_tree, type = 3, cex = 0.6)
```

Back to other variables in descriptive statistics. Distributions of subject's companion on Titanic are all too narrow, as shown in figure 2. This motivates categorization since we will not lose too much information. Lastly, nearly half of the subjects are English. And if we focus on crew, the number rise to 85%.

[3] Though this may not be relevant to the model, it is still an surprising discovery that it wasn't until the late 19th century that the idea of women traveling alone gained ground. As a result, there were nearly twice as many males passengers as females on Titanic. In fact, only 40% female passengers have no companion on the ship.

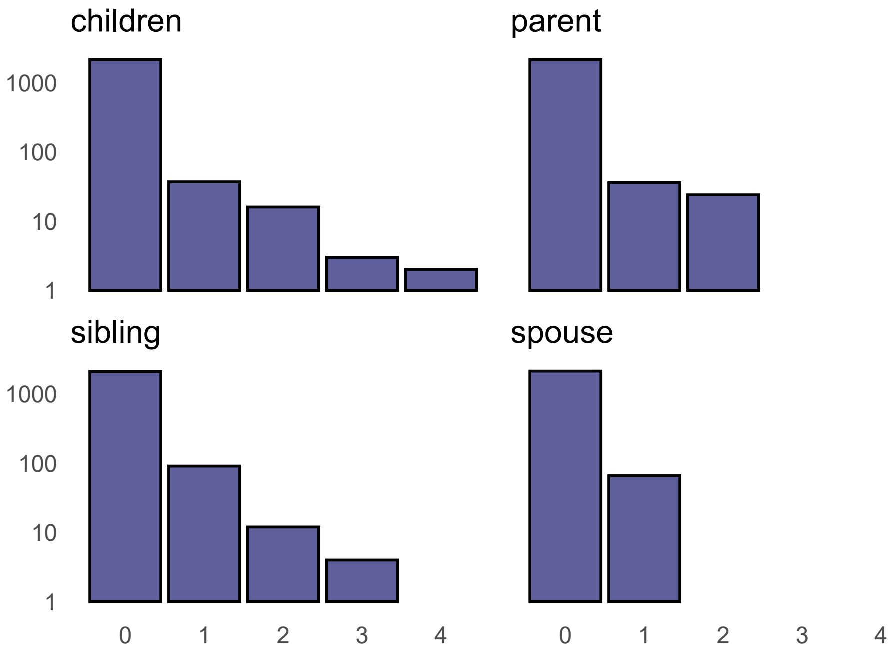


Figure 2. Few subjects have more than one companion in any of the 4 relations. Y axis on log scale.

Given this results, the final step in data munging is to dichotomize `spouse`, `parent`, `children` and `sibling` to denote if there is such relation. Thus we no longer have to deal with continuous predictors with poor distribution.

Univariate relationship between each independent variable and survival status is presented in figure 3. For each column, we can build a anova-type plot with no control over confounding variables, though it may still assist us in determining how to spend degrees of freedom. If a predictor's effect on the response is strong, it's more likely that we need to spend more parameters on it. However, if a variable's effect appears to be weak, it could either due to a truly flat relationship, or to nonlinearity and predictors among variables that univariate method cannot detect.

The plot shows appreciably strong effects of gender and cabin class on survival status. The effect of age seems trivial except for the missing subjects, but again, this figure exposes only linear relationship, and only after categorization. As we will see in the next section, age effect are much nonlinear and concentrated in the young subjects. The downside of this kind of univariate relationship is also exemplified in `title`, where "Miss". For the same reason effects of other variables cannot be determined.

We will finish with a redundancy analysis to study if any predictor can be readily explained by the rest of predictors, therefore does not much bring new information and may not enter the model. The checking algorithm involves

```
Redundancy Analysis

redun(formula = ~age + class + nationality + title, data = t)

n: 1497     p: 4    nk: 3
```

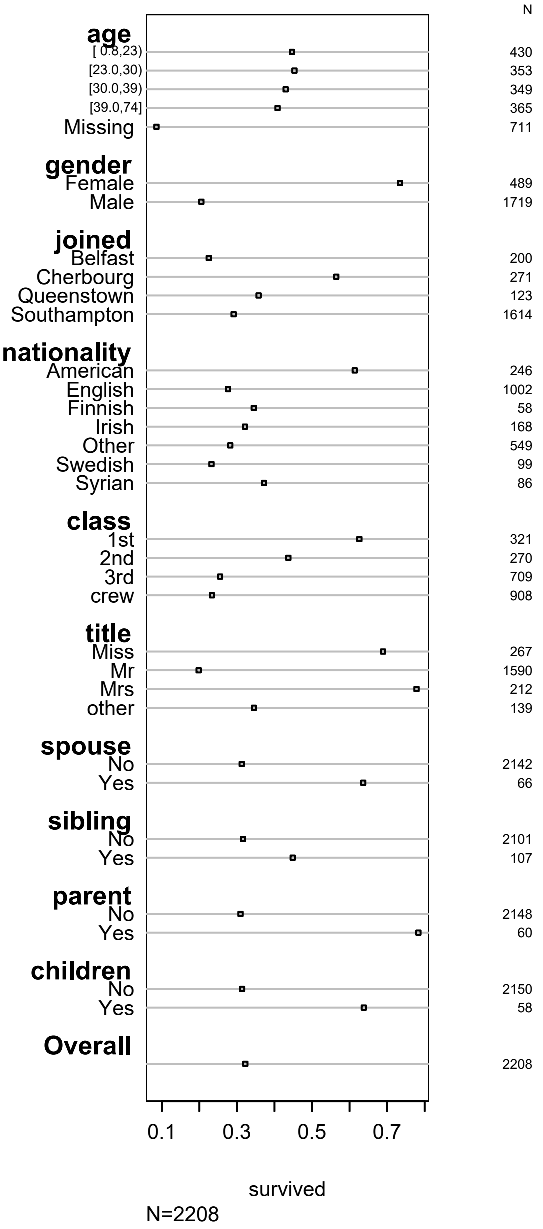


Figure 3. Summary of relationship between survival and each predictor

```
Number of NAs:   711
Frequencies of Missing Values Due to Each Variable
        age       class nationality       title
        711           0           0           0


Transformation of target variables forced to be linear

R-squared cutoff: 0.9   Type: ordinary

R^2 with which each variable can be predicted from all other variables:

        age       class nationality       title
      0.285       0.525       0.482       0.299

No redundant variables
```

2.2. *Loess regression for nonlinear pattern*

The loess method is a common nonparametric regression model to study nonlinear relationship. In the case of binary response, the fitted value at $x = x_0$ is the weighted proportion of positive cases near the neighborhood of x_0. If the trend of a loess curve shows nonmonotoncity, it is reasonable to include that nonlinearity relationship in the model, e.g., modeling the predictor with polynomial transformation or with splines.

Another important interaction, according to many follow up studies, is related to cabin class (for passenger) and department (for crew and staff).
4

3. Model devlopment

A typical modeling workflow begins with an choice of a statistical model or a machine learning model. A statistical model often stems from a hypothesized probabilistic data generating mechanism and assumes additivity, whereas machine learning models is algorithmatic, optimized through parameter tuning to achieve a higher performance score. We choose the "simple" binary logistic model for the following reasons.

We prefer probabilistic predictions to classification with output label 0 and 1, since we are placing emphasis upon the *tendency* of survival. And the value of our model consist not in a dichotomous prediction, but in what characteristics would increase or decrease the possibility of survival. The notion has ruled out most of the machine learning models for classification, say, random forest, support vector machines and neural network, which are not intrinsically probability oriented. Such classifiers can often only yield a forced choice.

Interpretability and inference matters. Although some top data science competitioners has reported moderately high signal to noise ratio (e.x., 90% prediction accuracy) that might tip the balance towards machine learning models, interpretability is harmed. Specifically, statistical models favour additivity have explicit specification. As a result, there are natural distinctions between main effects and interactions, linearity and nonlinearity. And the inference procedure is well defined provided that the model is correctly specified. While in a multilayer neural network, everything can in-

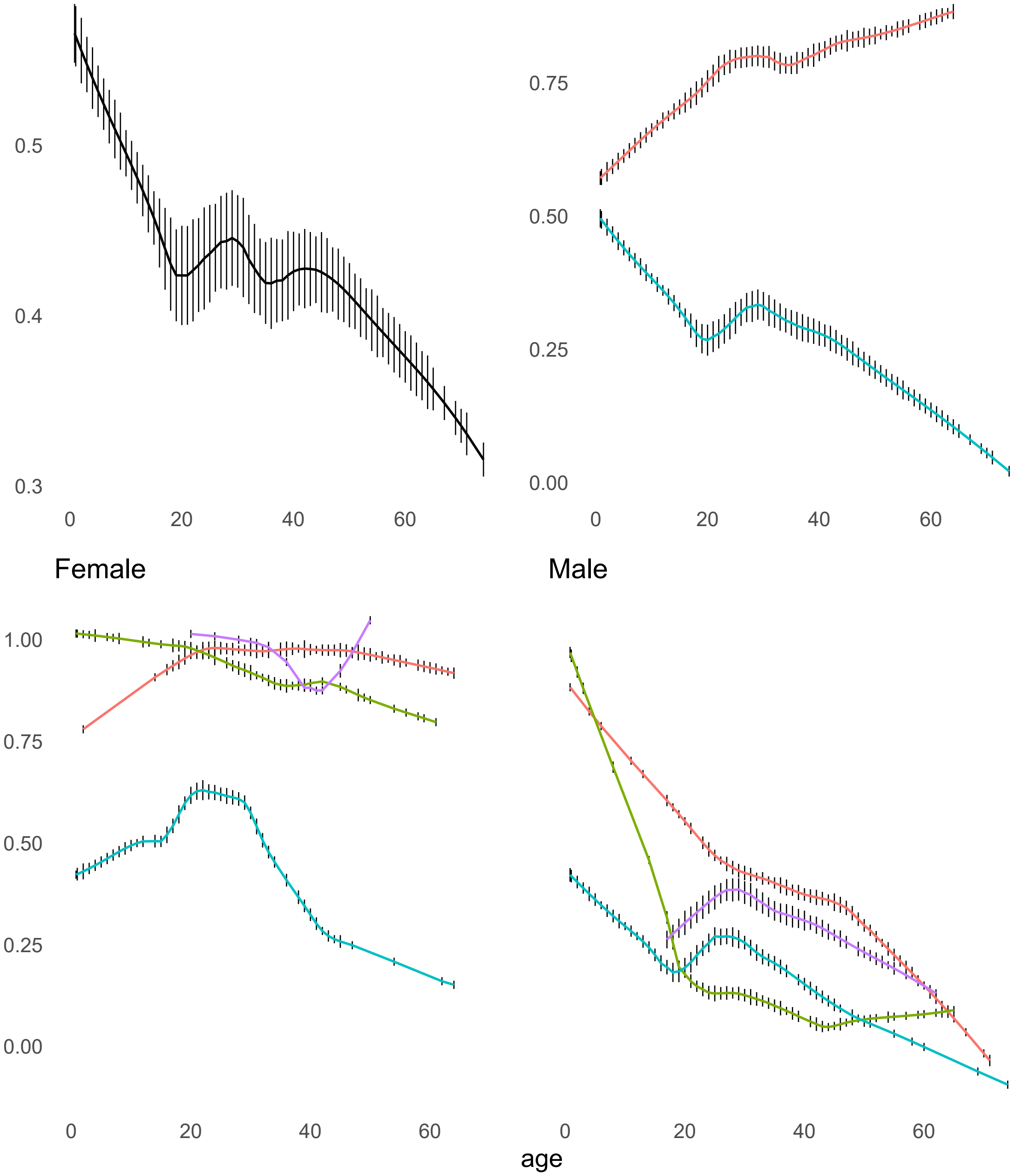


Figure 4. `loess` estimates of $P(\text{survived})$, with tick marks representing frequency counts within equal-width bins. Top left panel shows the nonlinear relationship between age and survival status without controlling confounding variables. Other plots give estimates under stratification by sex and class.

teract with one another and it could be daunting to isolate effects and conduct former inference.

Machine learning models are data hungry and sometimes create the need for big data (van der Ploeg, Austin, and Steyerberg 2014). To guard against overfitting, the analyst has to have a sample size that is 10 times larger at least if he chooses a decision tree instead of regression models. The rationale is that a statistical model is a safer approach as Dr. Harrell commented

> If n is too small to do something simple, it is too small to do something complex

3.1. *Specification*

We start by fitting a relatively large model, to decide how model complexity should be properly represented. This includes deciding the number of knots for continuous predictors and the number of categories of categorical predictors, where should we place interaction, etc. The large model also gives an overall sense of the predictive ability of each subject characteristics on survival status.

This is done by developing a saturated logistic model, with maximum flexible non-linear age effect represented as natural splines with 5 knots, and with all categorical predictors retain their original categories without pooling. Two way interactions have been specified between age and gender, age and class, and age and parent. We will not create three-way or higher interactions to avoid singularity. Since this is an initial model, observations with missing age are not used. The model equation is

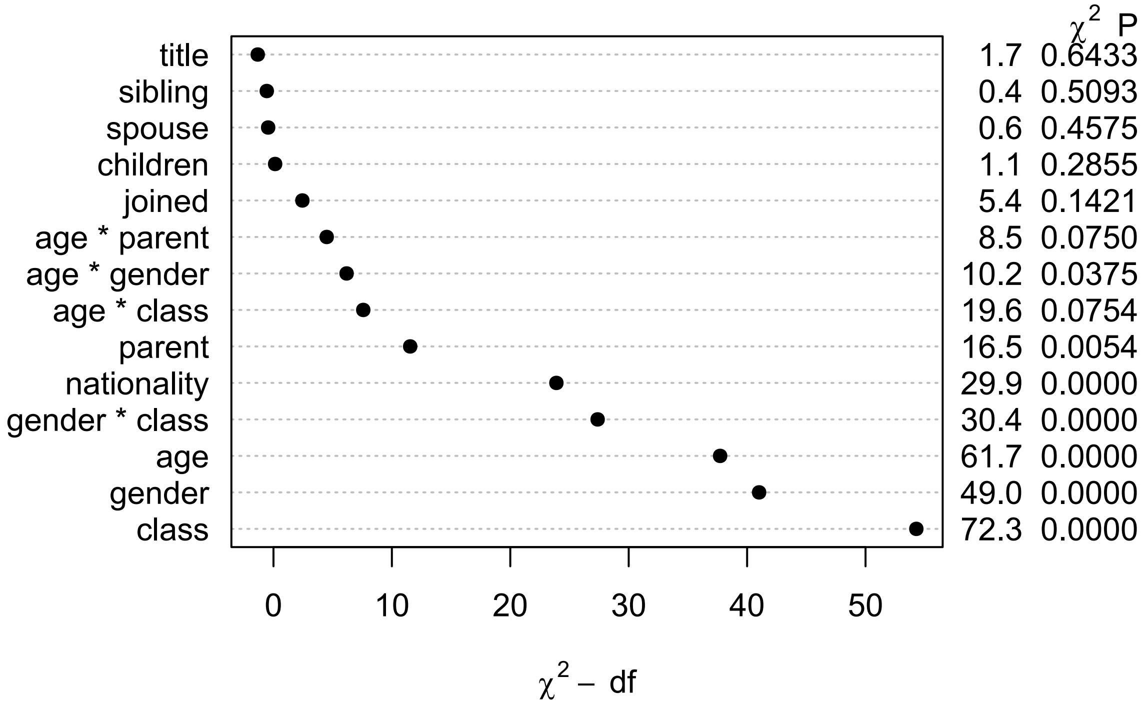
```
survived ~ (rcs(age, 5) + gender + class)^2 + (rcs(age, 5) * parent) +
            joined + spouse + sibling +
            children + nationality + title
```

Table 2

anova plot

Table 2. Hypothesis testing for the saturated model

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	61.72	24	<0.0001
All Interactions	37.11	20	0.0114
Nonlinear (Factor+Higher Order Factors)	45.56	18	0.0003
gender (Factor+Higher Order Factors)	49.01	8	<0.0001
All Interactions	48.84	7	<0.0001
class (Factor+Higher Order Factors)	72.29	18	<0.0001
All Interactions	54.24	15	<0.0001
parent (Factor+Higher Order Factors)	16.55	5	0.0054
All Interactions	8.50	4	0.0750
joined	5.44	3	0.1421
spouse	0.55	1	0.4575
sibling	0.44	1	0.5093
children	1.14	1	0.2855
nationality	29.89	6	<0.0001
title	1.67	3	0.6433
age × gender (Factor+Higher Order Factors)	10.18	4	0.0375
Nonlinear	9.47	3	0.0237
Nonlinear Interaction : f(A,B) vs. AB	9.47	3	0.0237
age × class (Factor+Higher Order Factors)	19.58	12	0.0754
Nonlinear	18.43	9	0.0305
Nonlinear Interaction : f(A,B) vs. AB	18.43	9	0.0305
gender × class (Factor+Higher Order Factors)	30.38	3	<0.0001
age × parent (Factor+Higher Order Factors)	8.50	4	0.0750
Nonlinear	1.47	3	0.6895
Nonlinear Interaction : f(A,B) vs. AB	1.47	3	0.6895
TOTAL NONLINEAR	45.56	18	0.0003
TOTAL INTERACTION	74.12	23	<0.0001
TOTAL NONLINEAR + INTERACTION	89.64	26	<0.0001
TOTAL	264.28	47	<0.0001



Model simplification based on hypothesis testing in table 2 must be done with extreme care. If we delete a predictor based on a p-value of 0.08, it leads to severe problems phantom degrees of freedom that distort coefficient estimates, confidence intervals, p-value and calibration of the final model. A more robust way is to use backward selection on bootstrap resamples.

Table 3. d.f. budget in the saturated model. Row 1: main effects. Row 2: interactions.

age	gender	class
4	1	1
4	4	4

3.2. *Multiple imputation*

The pooled estimates are obtained by averaging over m fitted model based on one piece of multiple imputation. The variance-covariance matrix T is calculated using Rudin's rule

$$T = \frac{1}{m}\sum_{i=1}^{m} U_i + (1 + \frac{1}{m})B$$

where U_i is the estimated complete-data variance-covariance matrix in each impu-

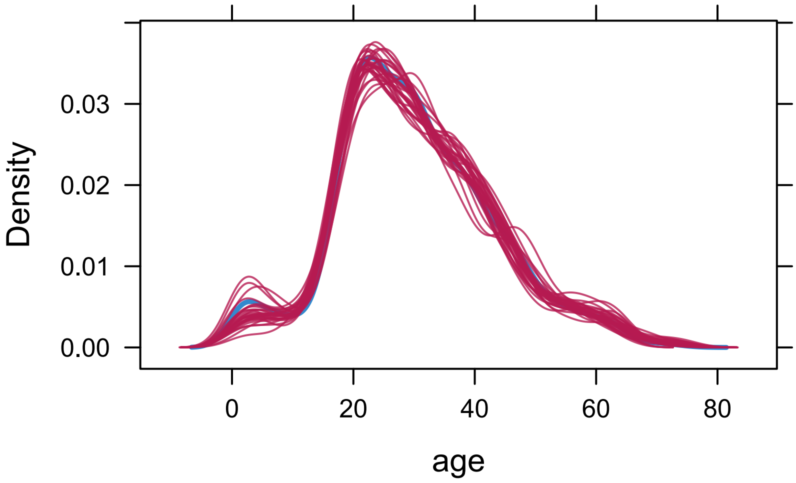


Figure 5. Density plot of observed and imputed data. In general, the imputed dataset mimic the age distribution seen in the observed data.

tation, and B the estimated variance-covariance matrix between the m complete-data estimates. Here we see the one major advantage of multiple imputation over single imputation is that not only does its variance estimates accounts for sampling variability, but also for the extra variance caused by missing values and finite number of imputations.

There are some simple workarounds

- complete-case analysis: That is, we delete all incomplete observations. Needless to say this will translate into a major harm on sample size since over 60% of `boat` are missing, not to mention other columns. Even if we remove `boat` and then delete rows with missing `age` we still lose over 1/5 of data. Moreover, figures in 2 have shed light on the relatively strong influence of `age` on survival. Also, the deletion of incomplete observations assumes date are missing completely at random (MCAR). When it's not the case, this could severely bias estimates of coefficients (Van Buuren 2018)
- single imputation:
- multiple imputation

https://www.encyclopedia-titanica.org/community/threads/passengers-who-spoke-other-languages.20103/

Since the crew's instructions (in English) tended to be along the lines of "Wait down here for further orders" a lack of understanding might well have saved many lives. Also many of the immigrants in 3rd class were traveling in family or neighborhood groups which included at least one English-speaker (often an established immigrant returning to the US from a visit back home) who could act as their spokesperson.

$\chi^2 - \text{df}$ is the "adjusted"

Table 4. Wald Statistics for `survived`

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	54.40	24	0.0004
All Interactions	33.62	20	0.0288
Nonlinear (Factor+Higher Order Factors)	36.19	18	0.0067
gender (Factor+Higher Order Factors)	251.76	8	<0.0001
All Interactions	50.24	7	<0.0001
class (Factor+Higher Order Factors)	100.57	18	<0.0001
All Interactions	50.32	15	<0.0001
parent (Factor+Higher Order Factors)	15.12	5	0.0099
All Interactions	8.77	4	0.0672
spouse	1.67	1	0.1957
children	1.77	1	0.1839
nationality	9.27	6	0.1587
age × gender (Factor+Higher Order Factors)	6.67	4	0.1544
Nonlinear	5.38	3	0.1458
Nonlinear Interaction : f(A,B) vs. AB	5.38	3	0.1458
age × class (Factor+Higher Order Factors)	14.92	12	0.2460
Nonlinear	12.80	9	0.1720
Nonlinear Interaction : f(A,B) vs. AB	12.80	9	0.1720
gender × class (Factor+Higher Order Factors)	31.09	3	<0.0001
age × parent (Factor+Higher Order Factors)	8.77	4	0.0672
Nonlinear	2.32	3	0.5096
Nonlinear Interaction : f(A,B) vs. AB	2.32	3	0.5096
TOTAL NONLINEAR	36.19	18	0.0067
TOTAL INTERACTION	73.33	23	<0.0001
TOTAL NONLINEAR + INTERACTION	83.83	26	<0.0001
TOTAL	342.62	40	<0.0001

3.3. *Model fitting, validation and calibration*

There will not be another Titanic, and any model on Titanic will not be used for prediction. Therefore, the goal of model validation is primarily to provide quantify the degree of overfitting with various bias-corrected measures.

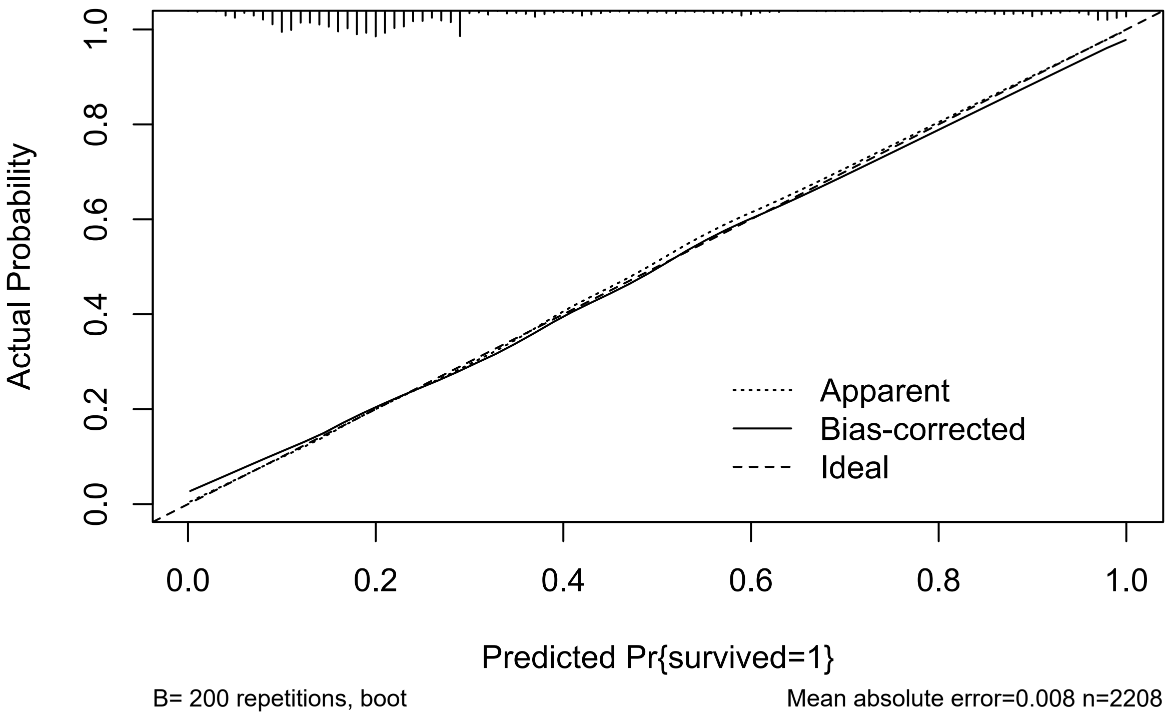
The van Houwelingen–Le Cessie heuristic shrinkage estimate

$$\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2}$$

where p is the total degrees of freedom and χ^2 the global likelihood ratio statistic for all predictors.

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/

Index	Original	Training	Test	Optimism	Corrected	n
	Sample	Sample	Sample		Index	
D_{xy}	0.6282	0.6428	0.6134	0.0295	0.5988	196
R^2	0.4161	0.4375	0.3859	0.0516	0.3644	196
Intercept	0.0000	0.0000	−0.1302	0.1302	−0.1302	196
Slope	1.0000	1.0000	0.8478	0.1522	0.8478	196
$E_{\max}$	0.0000	0.0000	0.0603	0.0603	0.0603	196
D	0.3530	0.3750	0.3233	0.0517	0.3014	196
U	−0.0009	−0.0009	0.0114	−0.0124	0.0114	196
Q	0.3539	0.3759	0.3119	0.0640	0.2899	196
B	0.1441	0.1404	0.1468	−0.0064	0.1505	196
g	1.7949	2.0421	1.7438	0.2983	1.4966	196
g_p	0.2733	0.2807	0.2595	0.0212	0.2521	196



```
n=2208   Mean absolute error=0.008   Mean squared error=0.00009
0.9 Quantile of absolute error=0.016
```

As a integral component of model validation, calibration aims to gauge the concordance between predicted values and observed data.

3.4. *Interpretation*

influence

`which.influence`

4. Discussion

The most decisive explanation for such effect is that first-class passengers had better access to information about the imminent danger and were aware that the lifeboats were located close to the first class cabins. Thus, their marginal effort costs to survive were lower. In contrast, most third-class passengers had no idea where the lifeboats were located (safety drills for all passengers were introduced after the Titanic disaster), and they did not know how to reach the upper decks where the lifeboats were stowed.

Wyn Craig Wade: there was a class culture on Titanic akin to the notion of a ”culture of poverty

> Undoubtedly, the worst barriers were the ones within the steerage passengers themselves.

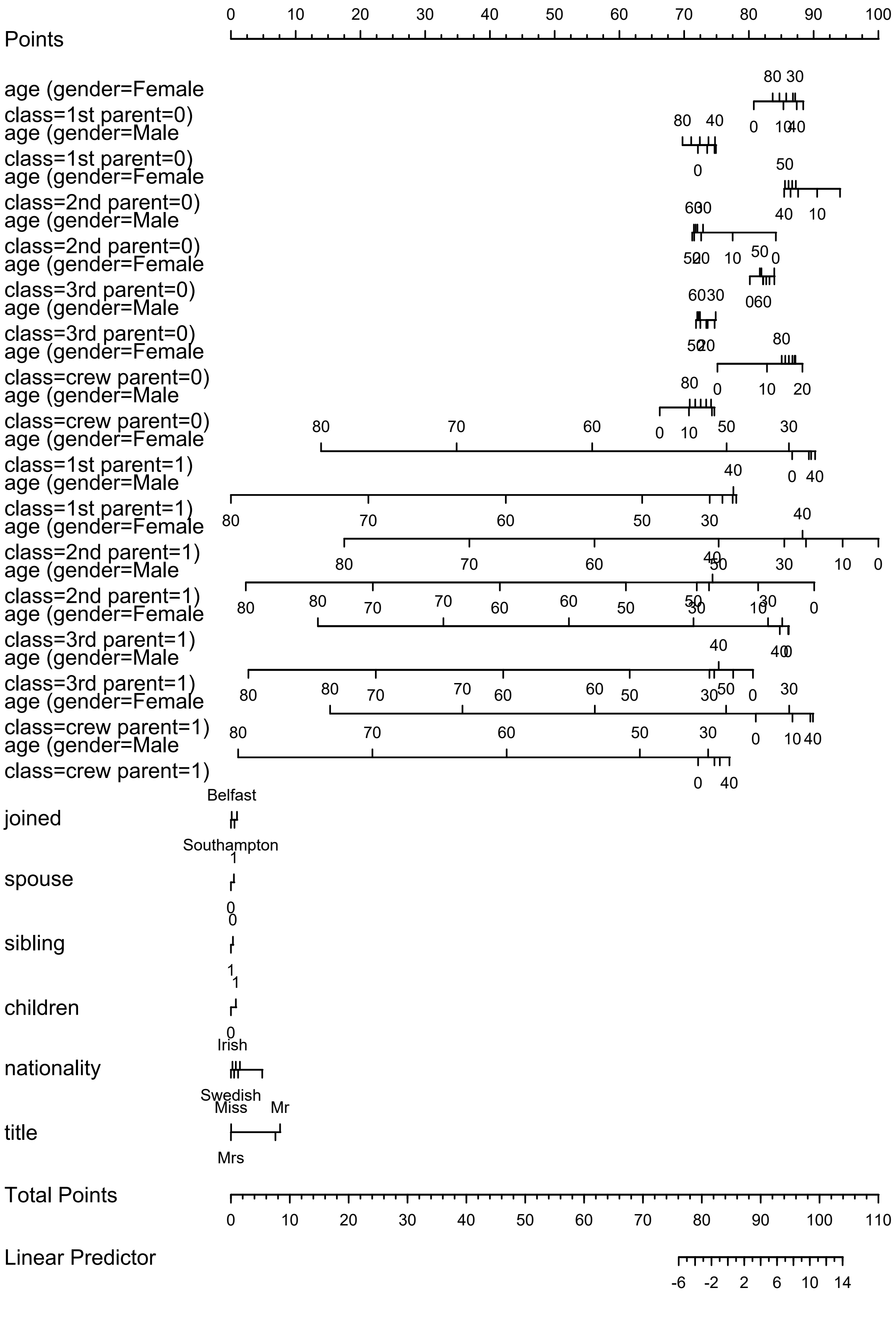


Figure 6. nomogram

> Years of conditioning as third-class citizens led a great many of them to give up hope as soon as the crisis became evident ... Barriers to steerage? Yes, but of a kind less indictable to the White Star Line than to the whole of civilization.

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

Women and children first only for higher class passengers. If you are a third class female

5. Conclusion

Appendix A. Data

A variety of other versions and forms of Titanic data sources have been collected due to public's constant interests in the tragedy as well as modern efforts trying to unveil the mystery. A comprehensive overview of several data variants is given by Symanzik, Friendly, and Onder (2018). Data in this case study is accessed on Encyclopedia Titanica, a leading archive on titanic facts. In contrast to the the famous titanic dataset (known as `titanic3`) distributed by kaggle for introductory level machine learning practices, the case study uses a more up-to-date and complete dataset in the following ways

- **Larger sample size**. Our data includes crew and staff members alongside passengers, while titanic3 only incorporate passenger information. We do not use a separate test set approach for validation either. As a result, the sample size is about 2.5 times larger.
- **More columns**. Additional variables such as role on the ship, nationality and occupation are added. A major difference is made by separating the travel companion data into four distinct columns: number of parents, children, sibling and spouses that each passenger traveled with. These were combined into two columns before.
- **More accurate**. `titanic3` was an effort to study Titanic in the 20th century, lastly updated and improved by Thomas Cason in 1999. The data has been constantly revised, many errors corrected, many missing ages filled in, and new variables created. Now it reflects the state of the data as of 21 October 2020.

The data cleaning process involves using appropriate data types, creating new features, adjusting levels for categorical variable and excluding irrelevant columns. Code can be found at clean.R.

`title` is extracted through each person's name with regular expressions and then collapsed into 4 levels.[4]

Passengers are classified according to their cabin class. Others on the vessel fall into one of crew and staff members. Crew includes victualling crew[5], engineering crew, deck crew and officers, substitute crew and guarantee group. Staff members include restaurant staff and orchestra.

Rare nationality (lower than 50 people) is collapsed.

Age information is presented as non-missing on the surface yet there is an indicator column representing when a person's age is only approximate and cannot be fully determined from current facts. These inaccurate age have been assigned NA. There were also ten subjects whose four companion variables were all explicitly missing. For simplicity, the mode `0` is filled in. Therefore, the problem of missing data is reduced to univariate missing of `age`.

Variables we do not utilize in this project includes name, date of birth and death, lifeboat number[6], fare, and cabin number.[7]

[4] For example, the title for passenger "Abbing, Mr Anthony" is "Mr".

[5] crew in charge of food, housekeeping, laundry, room service, etc.

[6] There were 9 recorded passengers who got on the lifeboat yet died before reaching Carpathia, another RMS which spearheaded the rescue of Titanic survivors. There were also 13 passengers who survived with no boat information documented, and this is most likely due to data quality issues after looking up on Encyclopedia Titanica. Even with these exceptions, whether a passenger got on a lifeboat yields perfect prediction on his/her survival. If one fits a logistic regression model on survival based on whether `boat` is missing, the apparent accuracy will be nearly 1. In this sense `boat` is more the result of survival, rather than a cause.

[7] While some study used this attribute to find cabin locations, its large amount of missingness could be a

Appendix B. Model formula

The formula for our binary logistic model

Appendix C. Criterion used in model validation

Somer's D_{xy} index is a calibration measure, which is the rank correlation between predicted and actual response. It has a close relationship with the C index

$$D_{xy} = 2(c - 0.5)$$

Appendix D. Computing environment

```r
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18362)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252
system code page: 936

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] rpart_4.1-15    patchwork_1.0.1 mice_3.11.0     rms_6.0-1
 [5] SparseM_1.78    Hmisc_4.4-1     Formula_1.2-4   survival_3.1-12
 [9] lattice_0.20-41 ggplot2_3.3.2   dplyr_1.0.2

loaded via a namespace (and not attached):
 [1] tidyr_1.1.2         splines_4.0.2       assertthat_0.2.1
 [4] latticeExtra_0.6-29 ymisc_0.0.0.9000    yaml_2.2.1
 [7] pillar_1.4.6        backports_1.2.0     quantreg_5.75
[10] glue_1.4.2          digest_0.6.27       RColorBrewer_1.1-2
[13] checkmate_2.0.0     colorspace_1.4-1    sandwich_3.0-0
[16] htmltools_0.5.0     Matrix_1.2-18       conquer_1.0.2
[19] pkgconfig_2.0.3     broom_0.7.2         bookdown_0.21
```

major source of complexity.

```
[22] purrr_0.3.4         mvtnorm_1.1-1       scales_1.1.1
[25] jpeg_0.1-8.1        MatrixModels_0.4-1  htmlTable_2.1.0
[28] tibble_3.0.4        rticles_0.17        farver_2.0.3
[31] generics_0.1.0      ellipsis_0.3.1      TH.data_1.0-10
[34] withr_2.3.0         nnet_7.3-14         cli_2.1.0
[37] magrittr_1.5        crayon_1.3.4        polspline_1.1.19
[40] evaluate_0.14       fansi_0.4.1         nlme_3.1-148
[43] MASS_7.3-51.6       foreign_0.8-80      tools_4.0.2
[46] data.table_1.13.2   hms_0.5.3           lifecycle_0.2.0
[49] matrixStats_0.57.0  multcomp_1.4-14     stringr_1.4.0
[52] rpart.plot_3.0.9    munsell_0.5.0       cluster_2.1.0
[55] compiler_4.0.2      rlang_0.4.8         grid_4.0.2
[58] rstudioapi_0.11     htmlwidgets_1.5.2   labeling_0.4.2
[61] base64enc_0.1-3     rmarkdown_2.5       gtable_0.3.0
[64] codetools_0.2-16    R6_2.5.0            gridExtra_2.3
[67] zoo_1.8-8           knitr_1.30          readr_1.4.0
[70] stringi_1.5.3       Rcpp_1.0.5          vctrs_0.3.4
[73] png_0.1-7           tidyselect_1.1.0    xfun_0.19
```

References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020a. *rmarkdown: Dynamic Documents for R*. R package version 2.5, https://github.com/rstudio/rmarkdown.

Allaire, JJ, Yihui Xie, R Foundation, Hadley Wickham, Journal of Statistical Software, Ramnath Vaidyanathan, Association for Computing Machinery, et al. 2020b. *rticles: Article Formats for R Markdown*. R package version 0.16.1, https://github.com/rstudio/rticles.

Frey, Bruno S, David A Savage, and Benno Torgler. 2009. “Surviving the Titanic disaster: economic, natural and social determinants.” .

Harrell, Frank E, Jr. 2020. *Hmisc: Harrell Miscellaneous*. R package version 4.4-1, https://CRAN.R-project.org/package=Hmisc.

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

Harrell, Jr., Frank E. 2020. *rms: Regression Modeling Strategies*. R package version 6.0-1, https://CRAN.R-project.org/package=rms.

Hind, Philip. 1999. https://www.encyclopedia-titanica.org/.

Koenker, Roger, and Pin Ng. 2019. *SparseM: Sparse Linear Algebra*. R package version 1.78, http://www.econ.uiuc.edu/~roger/research/sparse/sparse.html.

Milborrow, Stephen. 2020. *rpart.plot: Plot rpart Models: An Enhanced Version of plot.rpart*. R package version 3.0.9, http://www.milbo.org/rpart-plot/index.html.

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

Roecker, Ellen B. 1991. “Prediction error and its estimation for subset-selected models.” *Technometrics* 33 (4): 459–468.

Sarkar, Deepayan. 2008. *Lattice: Multivariate Data Visualization with R*. New York: Springer. ISBN 978-0-387-75968-5, http://lmdvr.r-forge.r-project.org.

Sarkar, Deepayan. 2020. *lattice: Trellis Graphics for R*. R package version 0.20-41, http://lattice.r-forge.r-project.org/.

Symanzik, Jürgen, Michael Friendly, and Ortac Onder. 2018. “The Unsinkable Titanic Data.” .

Terry M. Therneau, and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the*

Cox Model. New York: Springer.

Therneau, Terry, and Beth Atkinson. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15, https://CRAN.R-project.org/package=rpart.

Therneau, Terry M. 2020. *survival: Survival Analysis*. R package version 3.1-12, https://github.com/therneau/survival.

Van Buuren, Stef. 2018. *Flexible imputation of missing data*. CRC press.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://www.jstatsoft.org/v45/i03/.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2020. *mice: Multivariate Imputation by Chained Equations*. R package version 3.11.0, https://CRAN.R-project.org/package=mice.

van der Ploeg, Tjeerd, Peter C Austin, and Ewout W Steyerberg. 2014. "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." *BMC medical research methodology* 14 (1): 137.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020a. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. R package version 3.3.2, https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2020b. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.2, https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman and Hall/CRC. ISBN 978-1466561595, http://www.crcpress.com/product/isbn/9781466561595.

Xie, Yihui. 2015. *Dynamic Documents with R and knitr*. 2nd ed. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1498716963, https://yihui.org/knitr/.

Xie, Yihui. 2016. *bookdown: Authoring Books and Technical Documents with R Markdown*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 978-1138700109, https://github.com/rstudio/bookdown.

Xie, Yihui. 2020a. *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21, https://github.com/rstudio/bookdown.

Xie, Yihui. 2020b. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30, https://yihui.org/knitr/.

Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9781138359338, https://bookdown.org/yihui/rmarkdown.

Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9780367563837, https://bookdown.org/yihui/rmarkdown-cookbook.

Zeileis, Achim, and Yves Croissant. 2010. "Extended Model Formulas in R: Multiple Parts and Multiple Responses." *Journal of Statistical Software* 34 (1): 1–13.

Zeileis, Achim, and Yves Croissant. 2020. *Formula: Extended Model Formulas*. R package version 1.2-4, https://CRAN.R-project.org/package=Formula.