Modeling Titanic Survival

Qiushi Yan

[a]Beijing, China

ARTICLE HISTORY
Compiled November 17, 2020

ABSTRACT
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 alleviate selection bias, disadvantages of the holdout sample approach in validation, and more. This analysis is fully reproducible with all source R code and text.

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.

After appropriate formatting and cleaning, the data underlying the whole analysis 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 data analysis to first identify key questions of investigation that facilitates exploration and interpretation, however vague at the beginning. Then we can approach the core problem, filtering out trivialities, with statistical expression by abstraction. For our purposes and building on many outstanding works of study (2004; 2009; 2018), we establish the following research questions.

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	
joined	port of embarkation	Belfast, Cherbourg, Queenstown, Southampton
class		1st, 2nd, 3rd or crew
nationality	motherland	from wiki passenger list
title	title	Extracted from name
spouse	# of spouse on board	
sibling	# of siblings on board	
parent	# of parents on board	
children	# of children on board	

Question 1: To which degree is the *women and children first* policy respected? The obvious fact is that significantly higher proportion of females (73.4%) and children (57.3%) rescued than that in males (20.5%) and adults (42.5%). The key question lies in identifying possible factors that might intervene this process. For example, a first class adult male may possess the socio-economic advantage or financial means to get lifeboat access unfairly from the deck crew. The policy can be disrupted in other ways: some Titanic subjects could behave more in line with the selfish *homo oeconomicus*, then people (especially male) in their prime with physical superiority would see higher probability of survival.

Question 2: What is the crew member's position in the rescue effort? On the one hand, the 908 crew of men and women are expected to be more experienced, skilled and better informed about the location of the lifeboats and the incoming danger. On the other hand, their obligation is to care for the safety of passengers first, and only abandon the ship after the task has been fulfilled. We want to know whether self interest dominates in the life-and-death situation and crew tend to look out for themselves.

Question 3: For those who traveled alone with no companions (spouse, sibling, parent, children) on the vessel, is their survival probability greater or less? People without companions could be in shortage of psychological and physical support during the sinking. While they might also be able to reach a life-saving decision faster without transaction cost and negotiation. Specifically, the effects of having parents or children on the ship has been widely studied. The theory of parental investment suggests 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. In statistical terms, gender-children interaction may exist.

Question 4: Did English passengers 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 that might give preference to compatriots.

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. With descriptive statistics, we examine data distribution character-

istics, 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 We fit a saturated model with each predictor allowed maximum complexity and nonlinear interactions. This is used to guide development of the final model, based on hypothesis testing and predictor importance ranking with bootstrap.
  - Multiple imputation: We use predictive mean matching to impute subject's age, resulting in 30 complete dataset.
  - Model fitting and validation. We fit the final pooled model. Bootstrap internal validation (the ".632" method) is used to study optimism-corrected index measuring discrimination and calibration.
- Interpratation and discussion. We describe the model with both graphical methods such as partial effect plots and statistical testing. This provides model-based explanation to address former research problems.
- 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::describle` 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.

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
```

---

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.[1]

Of special importance is the age variable, which has roughly 30% missingness. On the other hand, it has a nearly symmetric distribution with 80% known observations falling between 14 and 50. For further examination of patterns of missing data, we could fit a decision tree 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.

```
na_tree <- rpart(factor(is.na(age)) ~ .,
                 data = t %>% mutate(survived = as.factor(survived)) ,
                 minbucket = 50)
# figure 1
```

---

[1] 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.

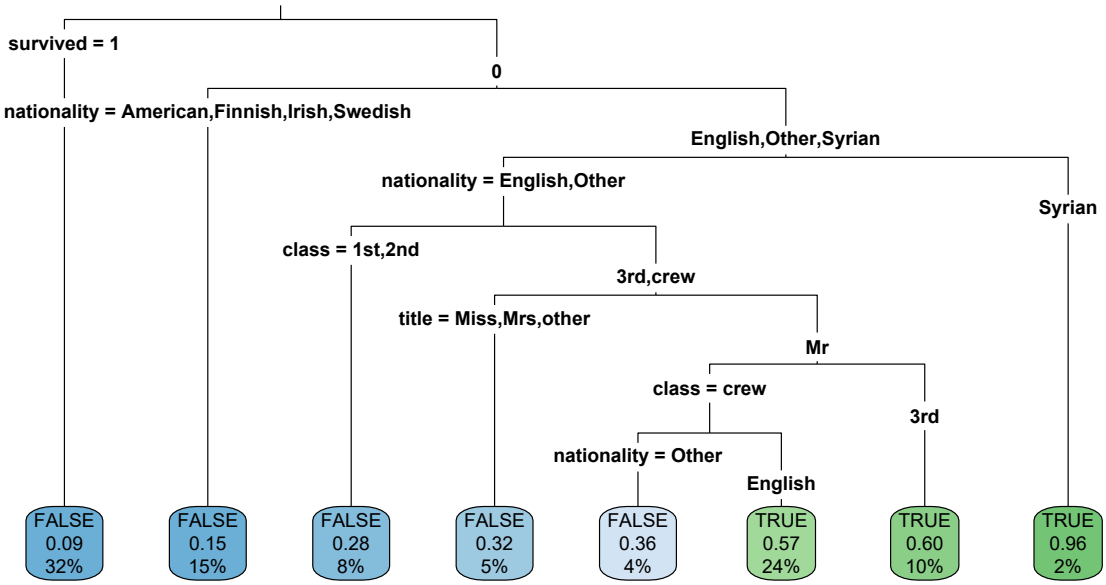

Figure 1. The decision tree for predicting `is.na(age)`, which finds strong patterns of missing related to class/department and gender (the Syrian node has very limited samples). Each node shows (top to bottom) the predicted class, the predicted probability of age being missing, the percentage of observations in the node.

```
rpart.plot::rpart.plot(na_tree, type = 3, cex = 0.6)
```

We see in figure 1 that survival status, gender and class are essential in determining age missingness. For a 3rd class male passenger who did not survive, age is missing with a probability of 60%. Interestingly, English male crew members are much more likely to have missing age than subjects of other nationality

Back to other variables in descriptive statistics. Distributions of subject's companion on Titanic are all too narrow, as illustrated in figure 2. There is too little variation to model them continuously. 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%.

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 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 the univariate method fails to detect.

The plot reveals appreciably strong effects of gender and cabin class on survival status. Age effects seem trivial except for the subjects with missing entry, but again, the downside of such kind of univariate display is that it force the audience to think linearly, and only after categorization. For the same reason effects of other variables cannot be determined.

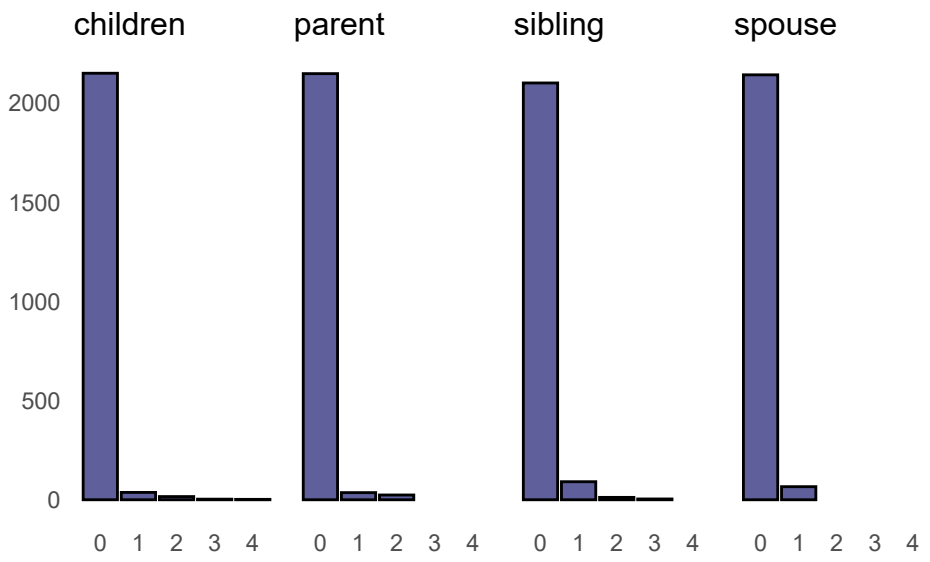

Figure 2. Few subjects have more than one companion in any of the 4 relations.

We finish with a redundancy analysis to study if data reduction of removing unnecessary predictors is possible. The checking algorithm expands continuous predictors into cubic splines and categorical predictors into dummy variables, it then uses OLS to predict each predictor with all remaining predictors. A predictor is deemed "redundant" if it can be predicted with an R^2 greater than 0.9.

```
Redundancy Analysis

redun(formula = ~age + gender + class + nationality + title +
    spouse + sibling + parent + children, data = t)

n: 1497     p: 9    nk: 3

Number of NAs:   711
Frequencies of Missing Values Due to Each Variable
        age      gender       class nationality       title      spouse
        711           0           0           0           0           0
    sibling      parent    children
          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      gender       class nationality       title      spouse
      0.340       0.977       0.551       0.492       0.977       0.341
    sibling      parent    children
```

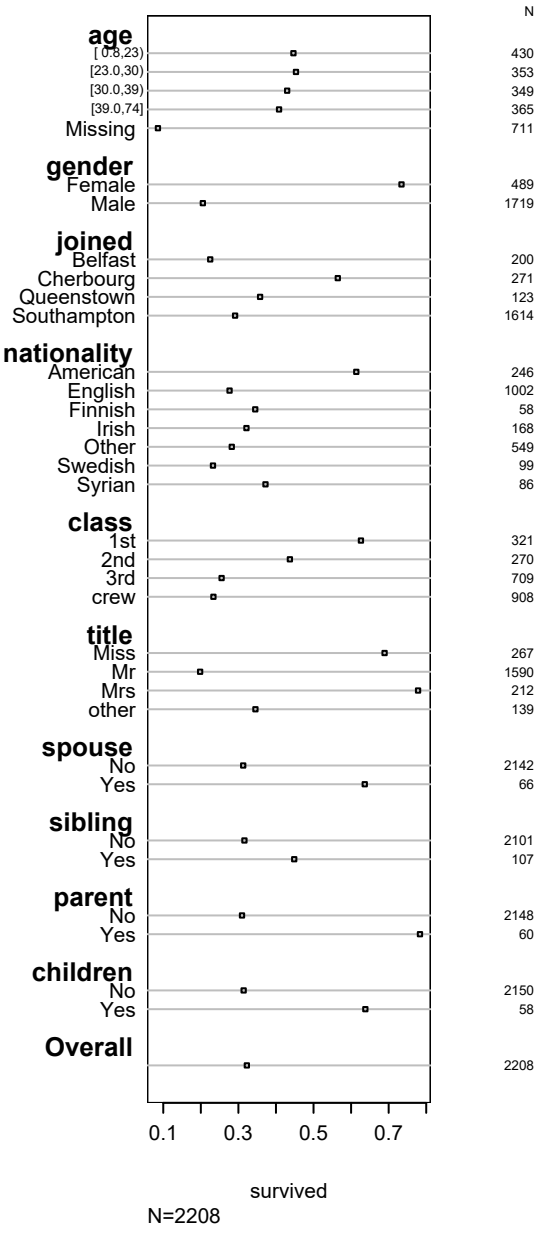

Figure 3. Univariate summary of relationship between survival and each predictor

```
      0.167       0.289       0.339

Rendundant variables:

title

Predicted from variables:

age gender class nationality spouse sibling parent children

  Variable Deleted   R^2 R^2 after later deletions
1            title 0.977
```

The redundancy analysis has reported `title` as redundant, with more than 97% of its variation explained by the rest of the predictors. It is not surprising that knowing a person' age could almost determine his/her title in the four categories. This means including `title` contains nearly no predictive value and could be readily deleted.

```
t$title <- NULL
```

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 a loess curve exhibits a reasonable degree of nonmonotonicity, it will often pay to not assume linearity, e.g., modeling the predictor with polynomial transformation or with splines.

It was widely documented that gender interacts directly with age, in other words, age effects for men and women are most likely to be nonparallel in the logit scale. Another prominant joint effects, according to many follow up studies, happened between age and cabin class. Figure 4 displays loess estimates of survival probability given age under certain stratification. Not only in a powerful nonlinear fashion does age affect survival status (top left panel), we also observe it interact with other two factors in a nonlinear way. Despite that the loess estimation here would by no means serve as a strict inference-oriented model, and we only include 3 factors at this point, it inspires a treatment of these effects that is more thorough and cautious.

3. Model development

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 in nature, optimized with parameter tuning. We choose to develop a statistical model, a "simple" binary logistic regression, for the following reasons.

We prefer probabilistic predictions to classification with output label 0 and 1, for the emphasis placed upon the *tendency* of survival. And the value of the model consist not in a dichotomous prediction, but in what characteristics would increase or decrease the probability of survival. The notion has ruled out most of the machine learning models for classification, say, random forest, support vector machines and neural network,

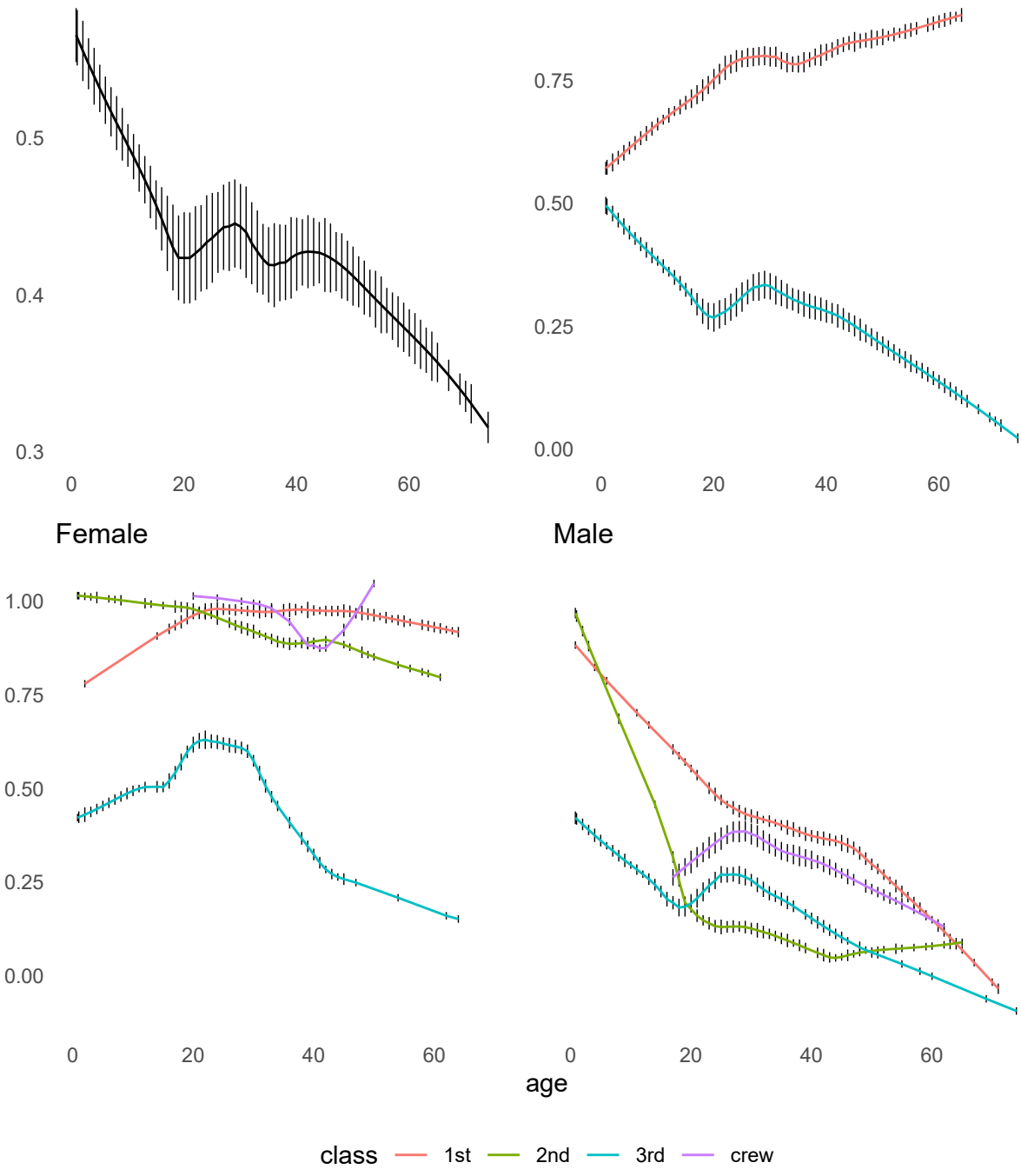

Figure 4. `loess` estimates of $P(\text{survived})$, with tick marks representing frequency counts within equal-width bins. Top left panel indicates nonlinear age effect without controlling other factors. Other plots give estimates under stratification by sex and class. The top right plot reflects women have a higher survival probability in general, and older females (red) and younger males (blue) are most likely to survive in their respective gender group. Two bottom plots depict survival pattern of subjects from different cabin class condition on sex.

which are not intrinsically probability oriented. Such classifiers can often only yield a forced choice.

Interpretability and inference matters. Many top data science competitions has reported moderately high signal to noise ratio (e.g., 90% prediction accuracy) that might tip the balance towards machine learning models, yet interpretability is harmed. Specifically, statistical models favours additivity and explicit specification. It follows that there are natural distinctions between main effects and interactions, linearity and non-linearity. And the inference procedure is well defined provided that the model is correctly specified. While in a multi-layer neural network, everything can interact 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 tree model instead of regression. While this case study uses a Titanic dataset that is about 1/3 larger than those only concerned with passengers, it is far less sufficient for a typical data-hungry machine learning model to validate well. The rationale is that a statistical model is a safter approach that can do without big data, 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, could we remove some term, 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 strategy as a starting point is also called prespecification of predictor complexity. It avoids creating phantom degrees of freedom when one has subjective judgment according to scatter diagrams or descriptive statistics on how to represent variables in a model. Commonly done, for example, is excluding a quadratic term simply because it is "non-significant", with p-value on the edge of 0.05. This approach is known to distort coefficient estimates, confidence intervals, p-value and calibration (too optimistic) of the final model, because it fails to accounts for sampling variability and suffers selection bias.(Grambsch and O'Brien 1991)[2] Therefore, it is essential to have extra caution, as demonstrated below using resampling, to do model simplification.

Prespecification of predictor complexity is done first by developing a saturated logistic model and then making necessary adjustments and improvements. In this model, we granted age effect maximal flexibility represented as natural splines with 5 knots, and 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. Since this is an initial model, observations with missing age are not used. The model equation is [3]

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

[2] confidence interval too narrow, p-value and standard errors too small and calibration too optimistic

[3] `rcs(x, n)` means "represent predictor x using natural splines with n knots". Knots are placed on evenly space percentiles by default.

Table 2. Hypothesis testing for the saturated model

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	59.76	24	<0.0001
All Interactions	34.91	20	0.0206
Nonlinear (Factor+Higher Order Factors)	45.20	18	0.0004
gender (Factor+Higher Order Factors)	183.28	9	<0.0001
All Interactions	50.76	8	<0.0001
class (Factor+Higher Order Factors)	72.11	18	<0.0001
All Interactions	54.92	15	<0.0001
parent (Factor+Higher Order Factors)	16.43	5	0.0057
All Interactions	8.23	4	0.0835
children (Factor+Higher Order Factors)	1.79	2	0.4080
All Interactions	0.76	1	0.3820
spouse	0.68	1	0.4099
sibling	0.41	1	0.5217
joined	5.60	3	0.1330
nationality	32.33	6	<0.0001
age × gender (Factor+Higher Order Factors)	8.58	4	0.0724
Nonlinear	8.28	3	0.0405
Nonlinear Interaction : f(A,B) vs. AB	8.28	3	0.0405
age × class (Factor+Higher Order Factors)	18.82	12	0.0929
Nonlinear	17.67	9	0.0392
Nonlinear Interaction : f(A,B) vs. AB	17.67	9	0.0392
gender × class (Factor+Higher Order Factors)	31.69	3	<0.0001
age × parent (Factor+Higher Order Factors)	8.23	4	0.0835
Nonlinear	1.47	3	0.6882
Nonlinear Interaction : f(A,B) vs. AB	1.47	3	0.6882
gender × children (Factor+Higher Order Factors)	0.76	1	0.3820
TOTAL NONLINEAR	45.20	18	0.0004
TOTAL INTERACTION	75.20	24	<0.0001
TOTAL NONLINEAR + INTERACTION	93.90	27	<0.0001
TOTAL	263.75	45	<0.0001

Table 2 selects specific hypothesis to test for general power, linearity and additivity assumptions of individual predictors, as well as their "chunky" version of global effects. We see dominant main effects of gender, age and cabin class, be it linear or nonlinear ($p < 0.0001$) More notably are the strong nonlinear interaction terms between the three. Of all four companion variables only parent manifests clear predictive ability. The impact of embarkation point is somewhat ambiguous ($p = 0.14$). As a graphical illustration, figure 5 plots "adjusted" partial χ^2 statistic of each predictor in the saturated model, with correction for degrees of freedom allocated to them.[4] This adjustment levels the playing field for comparison of predictive ability. The larger the adjusted χ^2, the more likely a variable would have a non-flat impact on survival status.

As mentioned before, the goal of the saturated model is guiding model complexity. More specifically, should we allocate more degrees of freedom to a certain term because some complex effects has been underrepresented? Or is there a term that is highly irrelevant thus could be deleted? The polynomial transformation on age and the resulting nonlinear interaction carry substantial statistical power, while further increasing knots or creating high-order interactions causes numerical problems. Therefore, it is positive advantage to us to keep them as is. There are also not sufficient reasons to collapse

[4]The correction is done by subtracting the d.f. from the partial χ^2 statistic, its expected value under the null hypothesis.

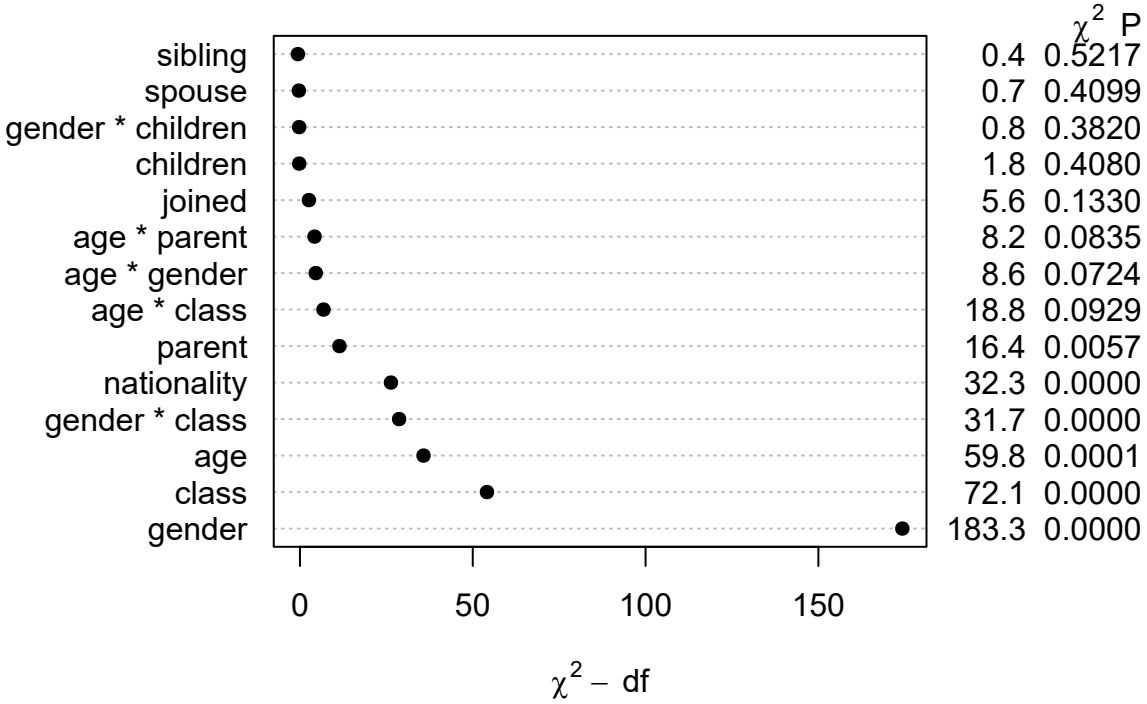

Figure 5. Ranking of predictive power in the saturated model based on adjusted χ^2

levels for nationality and port. Binary variables like spouse and sibling have extremely large p-values ($p > 0.5$), indicating relatively small predictive power. Still, great care should be taken when one attempts to conduct aggressive model simplification based on hypothesis testing and p-values. A reliable way is using bootstrap resampling. Figure 6 studies the importance of all terms including main effects and interaction over 500 bootstrap resamples. In each resample, we fit the saturated model, rank all 13 terms by the adjusted statistic $\chi^2 - \text{d.f.}$ in ascending order so that 13 is most important and 1 is least important. The height of a bar indicates the number of times a term is ranked at that position.

The importance ranking echoes previous findings that gender, age and classes are predominant factors. It also reveals great variability in terms of assessing predictive power. For example, we are only confident that `joined` is not one of the 5 most influential predictors. Nonetheless, rankings of the aforementioned "weak" variables, sibling and spouse, are highly concentrated at 1 to 3. In fact, if we perform backward selection in nearly 500 bootstrap resamples with AIC as stopping rule, none of the 3 binary variables entered the selected model more than 20 times. This results in the final decision to remove them in the final model.

```
t$sibling <- NULL
t$spouse <- NULL
```

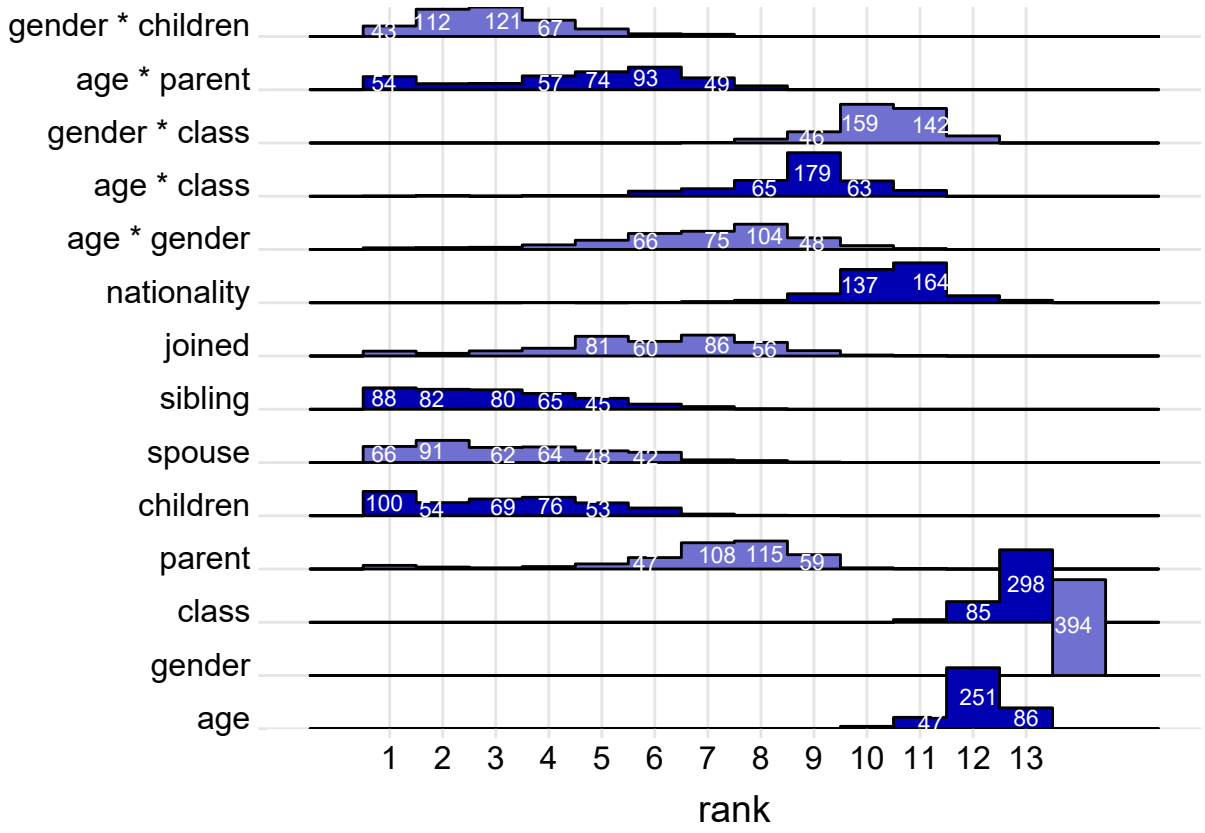

Figure 6. Distribution of importance ranking over 500 bootstrap resamples, 412 of which are actually fitted without numerical problems. Text indicates the number of times a term has a specific ranking, when the term ranked more than 40 times at that position. For example, gender ranks 13 (the most important) in all valid resamples.

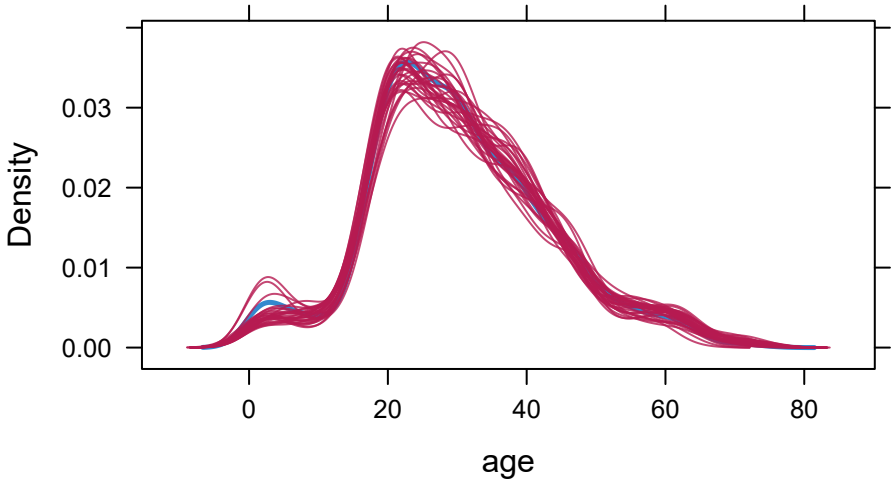

Figure 7. Density plot of observed and imputed data. In general, the imputed dataset mimic the age distribution seen in the observed data.

3.2. *Multiple imputation*

```
# multiple imputation with predictive mean matching to generate 30 complete dataset
imp <- mice(t, method = "pmm", m = 30, printFlag = FALSE)
```

The last step before fitting the final model is imputing missing values for age. The goal of multiple imputation, in contrast to simple alternatives such as filling in conditional mean, is to provide an accurate estimate of the variance-covariance matrix that not only accounts for sampling variability, but also for the extra variance caused by missing values and finite number of imputations (Van Buuren 2018). Thus tests on individual parameters gain power and bias are reduced. The general idea is to generate multiple complete dataset, fit the model in parallel, and then obtain a pooled final estimate by averaging over all fitted models.

We use predictive mean matching with $m = 30$, since approximately 30% of age is missing. The nonparametric method selects a group of Titanic subjects from all complete cases that have predicted values closest to the predicted value for the subject with missing age. [5] One donor is randomly drawn from the candidates, and the observed age of the donor is taken to replace the missing value. We use the default "type 1 matching" and 5 donors (Van Buuren (2018), Section 3.4.2). Advantages of predictive mean matching in the Titanic age setting are manifold. Since imputations are based on values observed elsewhere, they are realistic (e.g., no negative age). For another, it is compatible with non-normaility which allows us to have fewer assumptions.

3.3. *Model fitting, validation and calibration*

We now fit the final logistic model across 30 complete dataset. Pooled estimates are obtained by averaging over all pieces. We also get an imputation-corrected variance–covariance matrix based on within– and between–imputation variances. The long table

[5] The predicted value is generated by fitting a linear main effect model conditional on all other variables.

of Individual estimates are in the appendix.

The model exhibits moderate discrimination power, e.g. the ability to separate perished and survived subjects (concordance probability = area under the ROC curve ≈ 0.81). Brier score, as a proper quadratic scoring rule that incorporate both aspects, is a promising 0.144.

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	2208	LR χ^2	787.09	R^2	0.419	C	0.814
0	1496	d.f.	43	g	1.789	D_{xy}	0.627
1	712	$\Pr(>\chi^2)$	<0.0001	g_r	5.993	γ	0.629
$\max \lvert \frac{\partial \log L}{\partial \beta} \rvert$	0.02			g_p	0.274	τ_a	0.274
				Brier	0.144		

Table 3. Hypothesis testing for the final model

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	55.79	24	0.0002
All Interactions	33.89	20	0.0269
Nonlinear (Factor+Higher Order Factors)	35.87	18	0.0073
gender (Factor+Higher Order Factors)	251.63	9	<0.0001
All Interactions	51.98	8	<0.0001
class (Factor+Higher Order Factors)	93.98	18	<0.0001
All Interactions	53.39	15	<0.0001
parent (Factor+Higher Order Factors)	13.87	5	0.0164
All Interactions	7.39	4	0.1168
children (Factor+Higher Order Factors)	5.48	2	0.0645
All Interactions	1.58	1	0.2092
nationality	4.70	6	0.5829
joined	7.54	3	0.0566
age × gender (Factor+Higher Order Factors)	6.79	4	0.1471
Nonlinear	6.22	3	0.1015
Nonlinear Interaction : f(A,B) vs. AB	6.22	3	0.1015
age × class (Factor+Higher Order Factors)	15.93	12	0.1944
Nonlinear	13.47	9	0.1426
Nonlinear Interaction : f(A,B) vs. AB	13.47	9	0.1426
gender × class (Factor+Higher Order Factors)	31.03	3	<0.0001
age × parent (Factor+Higher Order Factors)	7.39	4	0.1168
Nonlinear	1.46	3	0.6920
Nonlinear Interaction : f(A,B) vs. AB	1.46	3	0.6920
gender × children (Factor+Higher Order Factors)	1.58	1	0.2092
TOTAL NONLINEAR	35.87	18	0.0073
TOTAL INTERACTION	75.40	24	<0.0001
TOTAL NONLINEAR + INTERACTION	87.86	27	<0.0001
TOTAL	353.11	43	<0.0001

Table 3 again constructs meaningful hypothesis testing for the final model. The χ^2 statistic of age decreased by a minor amount, resulting from using patterns of association with survival status to impute missing age. Remaining predictors generally have larger χ^2 statistic and smaller p-value compared to the saturated model in table 2, due to larger sample size in model development.

Although there will not be a second Titanic, making prediction a lesser problem, validation can still be used for good purposes. It quantifies the degree of overfitting by presented unbiased, optimism-corrected measures. More accurately, we will be using

Table 4. Optimism-corrected metrics

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.6253	0.6416	0.5599	0.0413	0.5840	489
R^2	0.4190	0.4374	0.3077	0.0703	0.3487	489
Intercept	0.0000	0.0000	−0.2598	0.1642	−0.1642	489
Slope	1.0000	1.0000	0.6726	0.2069	0.7931	489
$E_{\max}$	0.0000	0.0000	0.0819	0.0819	0.0819	489
D	0.3560	0.3753	0.2526	0.0653	0.2907	489
U	−0.0009	−0.0009	Inf	$-Inf$	Inf	489
Q	0.3569	0.3762	$-Inf$	Inf	$-Inf$	489
B	0.1435	0.1407	0.1496	−0.0039	0.1474	489
g	1.7895	2.0282	1.3646	0.2685	1.5210	489
g_p	0.2742	0.2812	0.2164	0.0365	0.2376	489

bootstrap internal validation to study the "future" performance of the model. In an award-winning solution to this legendary dataset submitted by IBM Watson, a holdout test set was used to validate their model. The data-splitting approach is known to require a significantly larger sample size (> 20000) than resampling methods on average to work acceptably well (Harrell, Jr. 2020b). Moreover, when the model developed on training sample is validated, the researcher would recombine training and testing set to fit a full model. This model, however, is never validated.

As an improved alternative, we choose Efron's 0.632 method for bootstrap internal validation. In each of the 494 bootstrap resamples, a logistic model is developed and evaluated on observations omitted from bootstrap samples. Per-bootstrap optimism is then the apparent index of accuracy subtracting that in the test sample formed by omitted observations. An weighted average $\hat{\varepsilon}_0$ over all 494 bootstrap resamples is computed to estimate the true optimism, while the bias-corrected estimate of predictive accuracy is calculated as $0.632(\text{apparent accuracy} - \hat{\varepsilon}_0)$. Table 4 displays the results. It validates two general aspects of model accuracy, discrimination and calibration. Calibration is the ability to make unbiased estimates of survival status, while discrimination is the a measure in how separated predictions are for survivors and victims.

The output does not deviate much from the apparent index. The validated area under the ROC curve as well as the concordance probability is now 0.79, and pseudo R^2 0.35. This indicates certain shortage of discrimination ability in comparison to some published models on the older dataset. Slope = 0.8 signals small amount of overfitting, with coefficients shrinking by 20% on new data.[6]. Brier score is now near 0.15.

Discrimination index are often associated with rank correlations between predictions and response, thus may not be sensitive enough to evaluate their closeness. Figure 8 aims to gauge the concordance between predicted values and observed data, or in other words, calibration. The actual probability is estimated with loess regression, and the bias correction is computed in a similar way as in table 4. The 45 degree line indicates the ideal scenario in which prediction perfectly matches observation. The model is well-calibrated by and large, with slight departure from the straight line in central region where there were few observations. The mean squared error is 0.00018, and the 0.9 quantile of absolute error is 0.023. Similarly, the unreliability index $E_{max} = 0.54$ measures maximum error in predicted probabilities. All these metrics reflects strong calibration.

[6] $D_{xy} = 2(\text{concordance proabability - } 0.5)$

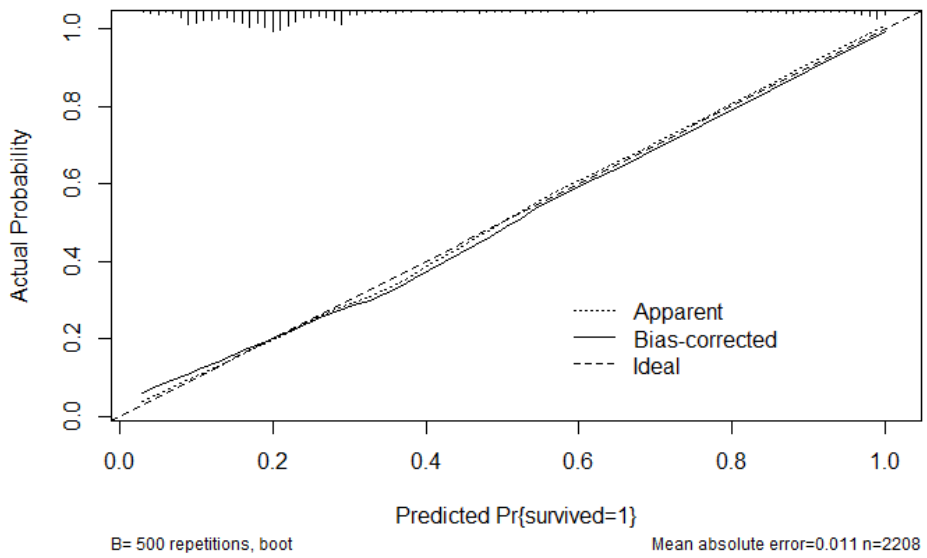

Figure 8. Bias-corrected calibration curve from internal bootstrap validation. Predictions output by the final model are plotted against estimated actual probabilities with loess. Bars on the top indicates number of observations in the same prediction bin. The bias corrected line is close to the ideal 45 degree line

To sum up, the model presents reasonable discrimination power and satisfactory quality of fitting. Excellent performance in calibration lifts overall metrics such as Brier score.

4. Interpretation and discussion

With non-monotonic relations and interactions involved, interpretations built upon parameter estimates or interquartile-range odds ratios are rarely informative. In this case, the model can be effectively described by partial effects plots, where we plot each characteristics against $\hat{P}$(survived) while holding other predictors constant at mean or median for continuous variables and mode for categorical variables. When interactions are involved, the survival probability are estimated separately for various levels of the interacting factors. For categorical predictors, level-specific effects are computed comparing to a reference category.

Figure 9 is an example partial effect plot for age under stratification of sex and cabin class.

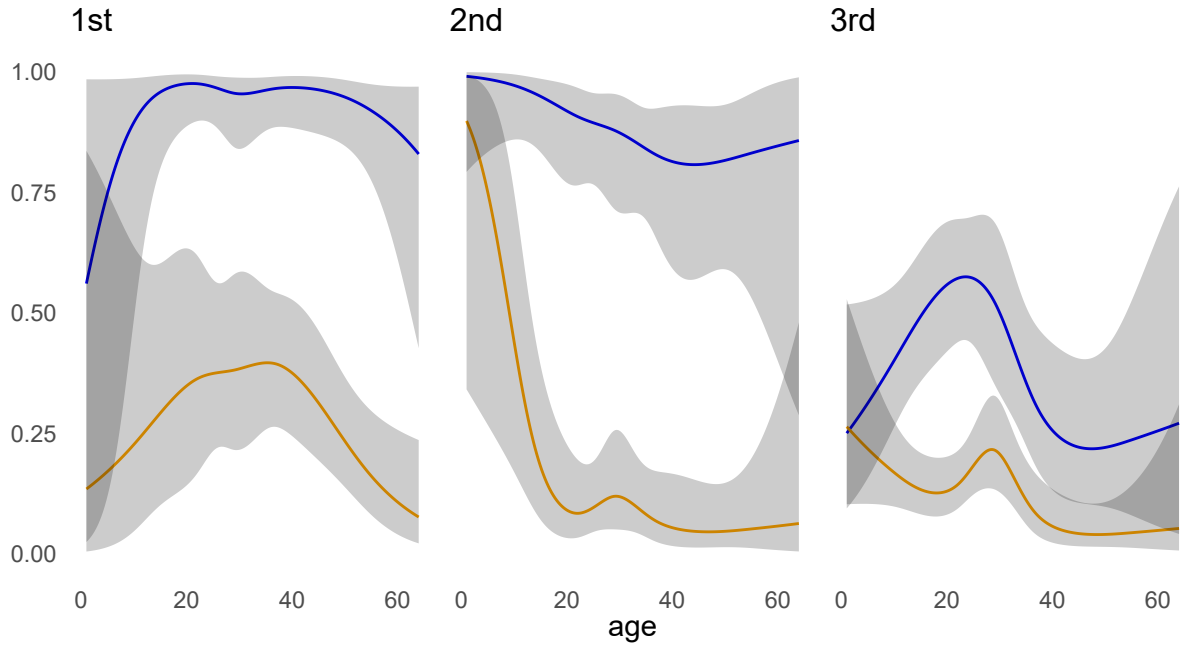

Figure 9. Age effects on survival status for **female** and **male** , stratified by cabin class. The default control setting is parent = 0, children = 0, nationality = English, and port = Southampton. Shaded region shows 95% confidence interval.

Question 1: Study the effects of the *women and children first policy*, with focus on gender, age and classes.

Women and children are evacuated first on average, as mentioned before using total proportion. The model helps us bring into considerations possibly intervening factors that could distort the established order of lifeboat access. There was recollection that at the beginning of the sinking, the Second Officer on the Titanic suggested to Captain Smith, "Hadn't we better get the women and children into the boats, sir?", to which the captain responded: "put the women and children in and lower away". Fifth Officer Harold Lowe in the American Inquiry as reinforced this policy as one regardless of classes

> It was simply the first woman, whether first class, second class, third class, or sixty-seventh class. It was all the same; women and children were first.

As we think not only in total numbers and the average, particularly after taking a closer look at numbers under stratification, another story presents itself. To begin with, not all classes of women and children receive the same level of help during the sinking, whether intended or unintended. We observe that first class passengers have the highest predicted survival possibility in general, and the third class being the lowest. The chasm between classes is most manifest in adult women. A 30-year-old first class women is likely to survive with a likelihood between 84.0% and 98.8%), and had she been in the third class, 33.2% and 66.9%. For children, knowing a 12-year-old comes from the first class increases the upper bound of predicted survival probability by nearly 30%. The most decisive explanation 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 were located at the quarters down the stern, from which the designer deliberately made it difficult to reach the upper decks.

While the first class is used to exemplify the uneven treatments of different classes, the policy was severely undermined within the third class after removing class effects. But let us first take the second class as a reference group, where the prediction roughly matches what should happen if children and women received the asserted assistance on the part of the authorities. The likelihood of survival for both gender starts approximately at the same higher region, then the probability for male plunged as he ages, while female only see a minor decrease after leaving the "girl" region. In the observations children are saved 100% (24 out of 24) and 88.7% of the women are rescued (94 out of 106).

The third class deviates from the reference group in two mains ways. Firstly, people in their prime have higher predicted survival probability. The survival curve displays an inverted U-shape, where people aged between 20 and 30 are mostly likely to survive, for both men and women. Secondly, the gender gap in survival probability narrowed drastically for third class. For example, being an 28-year-old man in the third class increases survival probability by 10% compared to the second class on average, whereas for a female age-mate the third class made her 33% more unlikely to survive. In addition, third class children had a slight 6% advantage in survival rates over adults, whereas in the other 3 groups, this disparity rose to 56%.

In summary, women and children first was only true on average, it was a not a class-blind effort, and third class women and children were faced with an extra harsh situation.

Question 2:
Crew members have a survivial pattern very skin to that of first class passengers, in both absolute value

Question 3:
Having parents or children around could significantly increases the chance

Question 4: Did British passengers gained survival advantage for some reason (e.g., activated national tie or simply because they can understand the instructions)?

Nationality is highly insignificant in the final model ($p = 0.58$). For the same class, people from different nations have almost identical survival pattern. (figure 10. One explanation, according to the titanica forum, is that the crew's instructions in English tended to be along the lines of "Wait down here for further orders", and a lack of understanding might . Also many of the immigrants in third 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.

Our last discovery concerns embarkation point. Passengers boarded at Cherbourg totaled 271, and had a surprisingly higher marginal survival rate of over 50%. The predicted survival probability is about 10% higher at Cherbourg than the average. The other two non-reference locations (Southampton and Queenstown) did not observe

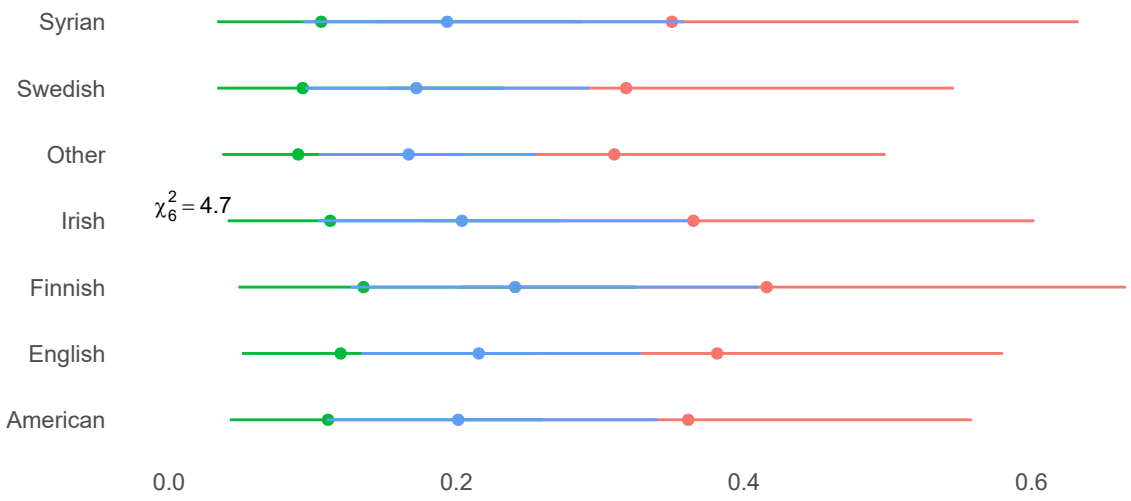

Figure 10. Partial effects of nationality on predicted survival probability.

significant difference ($p = 0.9$ and 0.75).

I have much confusion about this peculiar role of embarkation. One relevant point I found lately is that the Cherbourg list of passengers vanished long ago, according to a famous Titanic tract *Who Sailed on Titanic*, by Debbie Beavis. Getting the number of passengers on that port is difficult, not to mention the exact list of who went on which ship. Though it is still hard to imagine why Cherbourg could almost shoulder the entire χ^2 statistic for `joined`, even with the presence of some amount of incorrectness.

5. Conclusion

It may also also be informative to split crew members further into groups of different responsibilities, e.g., officers, deck and engine crews, victualing, restaurant staff. These working groups differs in comparative information advantage and authority, which affects survival rate.

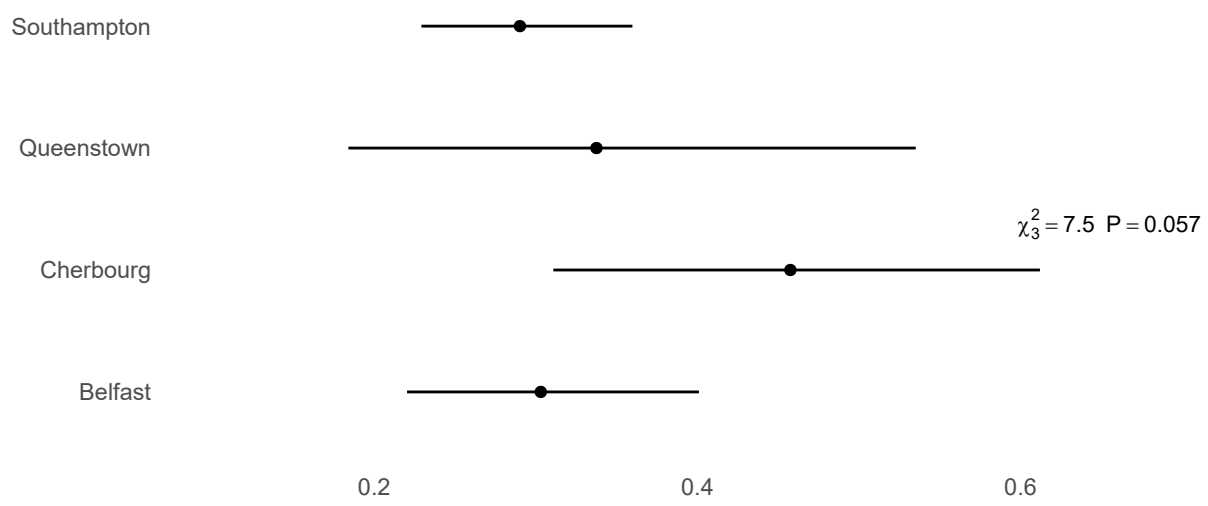

Figure 11. Partial effects of port on predicted survival probability

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 variables. Additional columns 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.
- Improved accuracy. `titanic3` was an effort to study Titanic in the 20th century, lastly updated and improved by Thomas Cason in 1999. During the recent two decades the data has been constantly revised, many errors corrected, many missing ages filled in, and new variables created. Now it reflects our our most up-to-date understanding of the event, in the digital form, as of 21 October 2020.

The data cleaning process involves converting 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.[7] This is a predictor that has been widely reported to have good predictive ability in many submissions. However, as we see in the redundancy analysis at the end of Section 2.1, it should not even be accepted in the tentative, saturated model.

Passengers are classified according to their cabin class. Others on the vessel fall into one of crew and staff members. Crew includes victualling crew[8], 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[9], fare, and cabin number.[10]

Appendix B. Model formula and paramter estimates

$$\text{Prob}\{\text{survived} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}$$

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

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

[9] 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.

[10] Although some study used this attribute to find cabin locations, its large amount of missingness could be a major source of complexity.

$$
\begin{aligned}
X\hat{\beta} = \\
&0.008881433 \\
&+0.2114074\text{age} - 0.000573145(\text{age}-10)_+^3 + 0.00246216(\text{age}-22)_+^3 \\
&-0.003007257(\text{age}-29)_+^3 + 0.001265583(\text{age}-37)_+^3 \\
&-0.0001473412(\text{age}-54.65)_+^3 \\
&-1.965745[\text{Male}] \\
&+4.801849[\text{2nd}] - 1.207512[\text{3rd}] - 2.311638[\text{crew}] \\
&+3.482035[1] \\
&+1.722792[1] \\
&+0.08636608[\text{English}] + 0.230042[\text{Finnish}] + 0.0155819[\text{Irish}] - 0.2309403[\text{Other}] \\
&-0.1927661[\text{Swedish}] - 0.04923775[\text{Syrian}] \\
&+0.6624393[\text{Cherbourg}] + 0.1582776[\text{Queenstown}] - 0.06188784[\text{Southampton}] \\
&+[\text{Male}][-0.1407618\text{age} + 0.0004672268(\text{age}-10)_+^3 - 0.002094754(\text{age}-22)_+^3 \\
&+0.002485191(\text{age}-29)_+^3 - 0.0009185903(\text{age}-37)_+^3 \\
&+0.00006092609(\text{age}-54.65)_+^3] \\
&+[\text{2nd}][-0.3392543\text{age} + 0.0007260332(\text{age}-10)_+^3 - 0.003023828(\text{age}-22)_+^3 \\
&+0.003693232(\text{age}-29)_+^3 - 0.001610244(\text{age}-37)_+^3 + 0.0002148067(\text{age}-54.65)_+^3] \\
&+[\text{3rd}][-0.1339377\text{age} + 0.0004332387(\text{age}-10)_+^3 - 0.002523439(\text{age}-22)_+^3 \\
&+0.003882666(\text{age}-29)_+^3 - 0.002070494(\text{age}-37)_+^3 + 0.0002780277(\text{age}-54.65)_+^3] \\
&+[\text{crew}][0.05161241\text{age} + 0.00001196427(\text{age}-10)_+^3 - 0.0004874508(\text{age}-22)_+^3 \\
&+0.001074983(\text{age}-29)_+^3 - 0.0006907784(\text{age}-37)_+^3 \\
&+0.00009128236(\text{age}-54.65)_+^3] \\
&+[\text{Male}][-0.4206718\ [\text{2nd}] + 2.177372\ [\text{3rd}] + 0.4894051\ [\text{crew}]] \\
&+[1][-0.1193207\text{age} - 0.0004343827(\text{age}-10)_+^3 + 0.005184355(\text{age}-22)_+^3 \\
&-0.01314738(\text{age}-29)_+^3 + 0.01061509(\text{age}-37)_+^3 - 0.00221768(\text{age}-54.65)_+^3] \\
&-1.206593\ [\text{Male}]\ \times\ [1]
\end{aligned}
$$

and $[c] = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise

Parameter estimates, standard error, Wald statistic and p-value

	$\hat{\beta}$	S.E.	Wald Z	$\Pr(>\|Z\|)$
Intercept	0.0089	2.1148	0.00	0.9966
age	0.2114	0.1251	1.69	0.0910
age'	-1.1426	0.7426	-1.54	0.1239
age''	4.9086	3.5265	1.39	0.1639
age'''	-5.9953	4.7267	-1.27	0.2047
gender=Male	-1.9657	1.0625	-1.85	0.0643
class=2nd	4.8018	2.4761	1.94	0.0525
class=3rd	-1.2075	1.9539	-0.62	0.5366
class=crew	-2.3116	2.9322	-0.79	0.4305
parent=1	3.4820	1.2035	2.89	0.0038
children=1	1.7228	0.8076	2.13	0.0329
nationality=English	0.0864	0.2762	0.31	0.7545

	$\hat{\beta}$	S.E.	Wald Z	$\Pr(>\|Z\|)$
nationality=Finnish	0.2300	0.4309	0.53	0.5934
nationality=Irish	0.0156	0.3898	0.04	0.9681
nationality=Other	-0.2309	0.2674	-0.86	0.3877
nationality=Swedish	-0.1928	0.3815	-0.51	0.6134
nationality=Syrian	-0.0492	0.4463	-0.11	0.9122
joined=Cherbourg	0.6624	0.3345	1.98	0.0477
joined=Queenstown	0.1583	0.4123	0.38	0.7011
joined=Southampton	-0.0619	0.1998	-0.31	0.7568
age × gender=Male	-0.1408	0.0552	-2.55	0.0108
age' × gender=Male	0.9315	0.4138	2.25	0.0244
age" × gender=Male	-4.1761	2.2651	-1.84	0.0652
age"' × gender=Male	4.9545	3.3787	1.47	0.1425
age × class=2nd	-0.3393	0.1538	-2.21	0.0274
age' × class=2nd	1.4474	0.9459	1.53	0.1260
age" × class=2nd	-6.0284	4.4551	-1.35	0.1760
age"' × class=2nd	7.3629	5.9182	1.24	0.2135
age × class=3rd	-0.1339	0.1158	-1.16	0.2476
age' × class=3rd	0.8637	0.7003	1.23	0.2174
age" × class=3rd	-5.0308	3.3216	-1.51	0.1299
age"' × class=3rd	7.7406	4.4306	1.75	0.0806
age × class=crew	0.0516	0.1684	0.31	0.7592
age' × class=crew	0.0239	0.8542	0.03	0.9777
age" × class=crew	-0.9718	3.6543	-0.27	0.7903
age"' × class=crew	2.1431	4.4525	0.48	0.6303
gender=Male × class=2nd	-0.4207	0.7009	-0.60	0.5484
gender=Male × class=3rd	2.1774	0.6237	3.49	0.0005
gender=Male × class=crew	0.4894	0.8663	0.56	0.5721
age × parent=1	-0.1193	0.1091	-1.09	0.2742
age' × parent=1	-0.8660	1.3351	-0.65	0.5166
age" × parent=1	10.3356	10.8394	0.95	0.3403
age"' × parent=1	-26.2109	29.2993	-0.89	0.3710
gender=Male × children=1	-1.2066	0.9608	-1.26	0.2092

Appendix C. Computing environment

```
sessionInfo()
```

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)

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] ggridges_0.5.2  scales_1.1.1    rpart_4.1-15    patchwork_1.0.1
 [5] mice_3.11.0     rms_6.0-1       SparseM_1.78    Hmisc_4.4-1
 [9] Formula_1.2-4   survival_3.1-12 lattice_0.20-41 ggplot2_3.3.2
[13] 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] plyr_1.8.6          htmltools_0.5.0     Matrix_1.2-18
[19] conquer_1.0.2       pkgconfig_2.0.3     broom_0.7.2
[22] bookdown_0.21       purrr_0.3.4         mvtnorm_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." .

Gleicher, David, and Lonnie K Stevans. 2004. "Who survived Titanic? A logistic regression analysis." *International Journal of Maritime History* 16 (2): 61–94.

Grambsch, Patricia M, and Peter C O'Brien. 1991. "The effects of transformations and preliminary tests for non-linearity in regression." *Statistics in Medicine* 10 (5): 697–709.

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

Harrell, Jr., Frank E. 2020b. "Split-Sample Model Validation." Https://www.fharrell.com/post/split-val/.

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.