Modeling Titanic Survival

Qiushi Yan[a]

[a]Beijing, China

ARTICLE HISTORY
Compiled October 17, 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, feature enginerring, 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 statistial decisions are explained. This analysis is also made fully reproducible with R code and text provided.

KEYWORDS
logistic regression; multiple imputation; model validation

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

- Do human beings behave more in line with the selfish homo oeconomicus, where everybody is out for himself or herself and possibly even puts other people's lives in danger

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 various machine learning competitions a quintessential dataset among others, in which one major interest is to predict possibility of survival given sex, age, class, etc. There are several variants of this data existed on the web, the one I will be using is accessed on Encyclopedia Titanica, namely `titanic3`, courtesy of Philip Hind, with the following variables (table 1)

This data frame recorded the survival status 1309 Titanic passengers[1] alongside his/her gender, age, family relations on board, ticket fare, etc. There were 809 victims

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

[1]The data does not involve crew members, and the total number of passengers is said to be 1317

Table 1. Data with 1309 passengers and 14 columns

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

and 500 survivors in total.

This case study has been greatly inspired by Dr. Frank Harrell's similar one in his *Regression Modeling Strategies* (2015, Chapter 12) book, here I attempt to propose my own idea 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) and RStudio with code and text made public in this repo.

- does socioeconomic advantage
- people in their prime
- Quantify predictive ability of each predictors, i.e. which predictor is most dominant in determining whether a passenger will survive
- Find Interactions between predictors. 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 reproductive age(Frey, Savage, and Torgler 2009)[The average peak reproductive period in females is between the ages of 16 and 35.]. 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.
- Whether the *Women and children first* policy is respected. After the collision, The captain explicitly issued an order for women and children to be saved first. [2]
- For those who traveled alone with no family relations on the vessal,

Here is a brief summary of the following sections

- Exploration, data preprocessing based on descriptive statistics and visualization, finish with a redundancy analysis

[2] Though there is no international maritime law that enforce this chivalry spirit.

2. Exploration

2.1. *Data processing and descriptive statistics*

Before any analysis, we'll start by some data munging. First exclude those variables that hardly bring any insight to the possibility of survival: `name`, `ticket`[3], `body`, `cabin`[4] and `home.dest`. The `boat` column is left out for another reason, because an non-missing entry in `boat` basically means survival and missing means death. For this reason "survive" and "get a life boat" is used interchangeably in the analysis. [5].

Next, for purposes of interpretation we will transform `fare` into today's US dollars with correction for inflation. According to discussion here, we make the transformation

$$\frac{\text{today's US dollar}}{\text{fare in 1912}} \approx \underbrace{5}_{\text{exchange rate then}} \times \underbrace{26}_{\text{inflation index from 1912 to 2020}}$$

At

Finally, a nice summary of all existing variables in the data is given by the `Hmisc::describe` function.

8 Variables 1309 Observations

---

pclass

n	missing	distinct
1309	0	3

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

---

survived

n	missing	distinct
1309	0	2

Value	0	1
Frequency	809	500
Proportion	0.618	0.382

---

sex

n	missing	distinct
1309	0	2

Value	female	male
Frequency	466	843
Proportion	0.356	0.644

---

age

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

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

---

[3]There may be some reasons to include ticket number since their prefix could represent placement of the cabins within the ship. However, the use of such a predictor would bring excessive degrees of freedom to the model. And their poor distribution would be declared redundant by redundant analysis (later) anyway.

[4]Because this was primarily an identification for class and most were missing.

[5]More precisely, 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.

sibsp

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

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

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

---

parch

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

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

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

---

fare

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1308	1	281	1	4328	5019	939.2	983.8	1026.5	1879.0	4065.7	10146.6	17374.5

```
lowest :     0.000   412.204   521.625   650.000   810.875
highest: 29578.249 32177.704 34108.750 34190.000 66602.799
```

---

embarked

n	missing	distinct
1307	2	3

```
Value        Cherbourg  Queenstown Southampton
Frequency          270         123         914
Proportion       0.207       0.094       0.699
```

There are several interesting patterns to notice [6]

- age has roughly 20% missingness. On the other hand, the variable has a nice distribution with 80% known observations falling between 14
- Both sibsp and parch rarely have instances larger than 3, and can be readily categorized without losing too much information.
- The distribution of fare is heavily right skewed, caused by the large amount of third class passengers who make do with cheap cabins, as shown in figure 1. This may suggest a log transformation in the model.

- Approximately 20% age is missing. This calls for a necessary imputation method as we will see later, that survival can exhibit

Given this results, the last step in
Finally, redundant analysis

```
Redundancy Analysis

redun(formula = ~pclass + sex + age + cut2(sibsp, 0:3) + cut2(parch,
    0:3) + fare + embarked, data = titanic)

n: 1043      p: 7    nk: 3

Number of NAs:   266
```

[6] 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 family accompanies on the ship

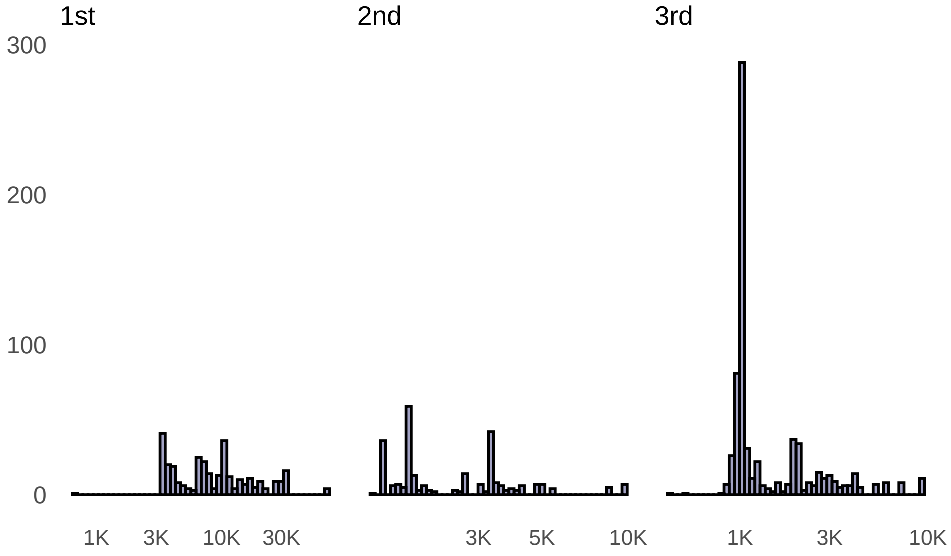


Figure 1. More than 75% of the third class passengers (700-plus in total) purchased tickets with price lower than \$2000, while the median fare for second and first class is \$3861. X axis is on log 10 scale

```
Frequencies of Missing Values Due to Each Variable
          pclass               sex               age cut2(sibsp, 0:3)
               0                 0               263                0
cut2(parch, 0:3)              fare          embarked
               0                 1                 2


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:

          pclass               sex               age cut2(sibsp, 0:3)
           0.769             0.116             0.308            0.409
cut2(parch, 0:3)              fare          embarked
           0.434             0.454             0.189

No redundant variables
```

2.2. *Data missing patterns*

There were 17 passengers whose `fare` is zero, all of whom males boarding in Southampton, the start of the voyage. It is suspected that some of them may be falsely included crew members, or this could be an error in data collection. I will treat these anomalies as missing entries for simplicity.

Here we use the data before excluding irrelevant columns

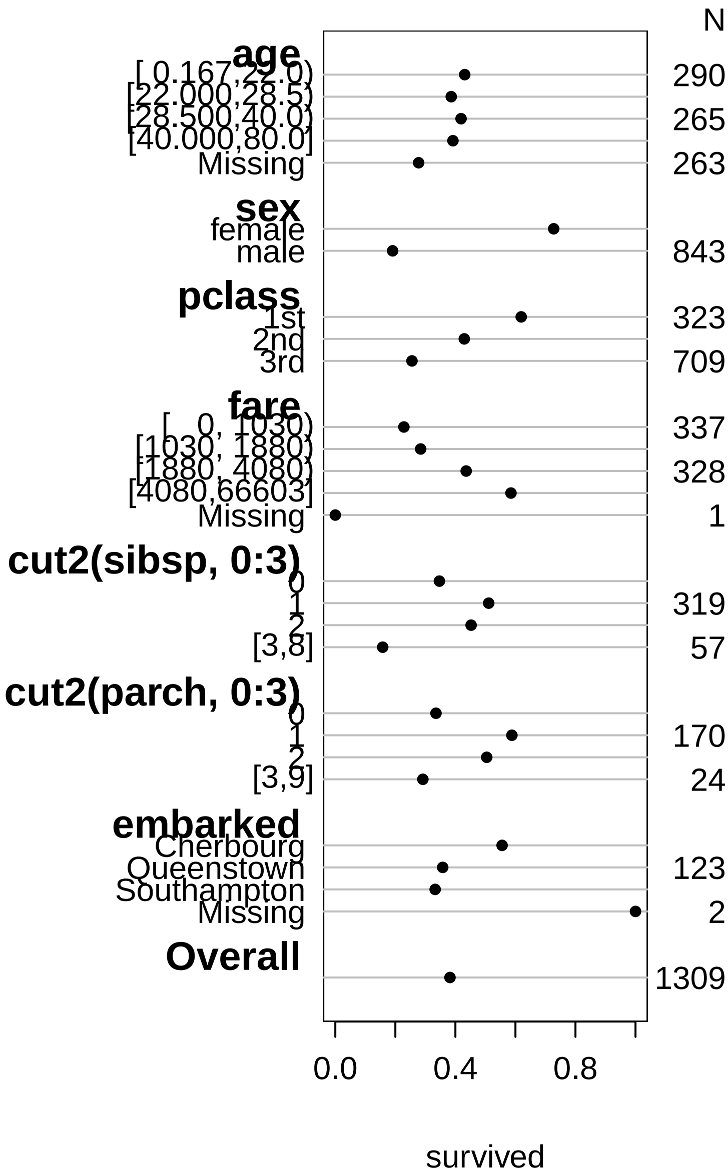
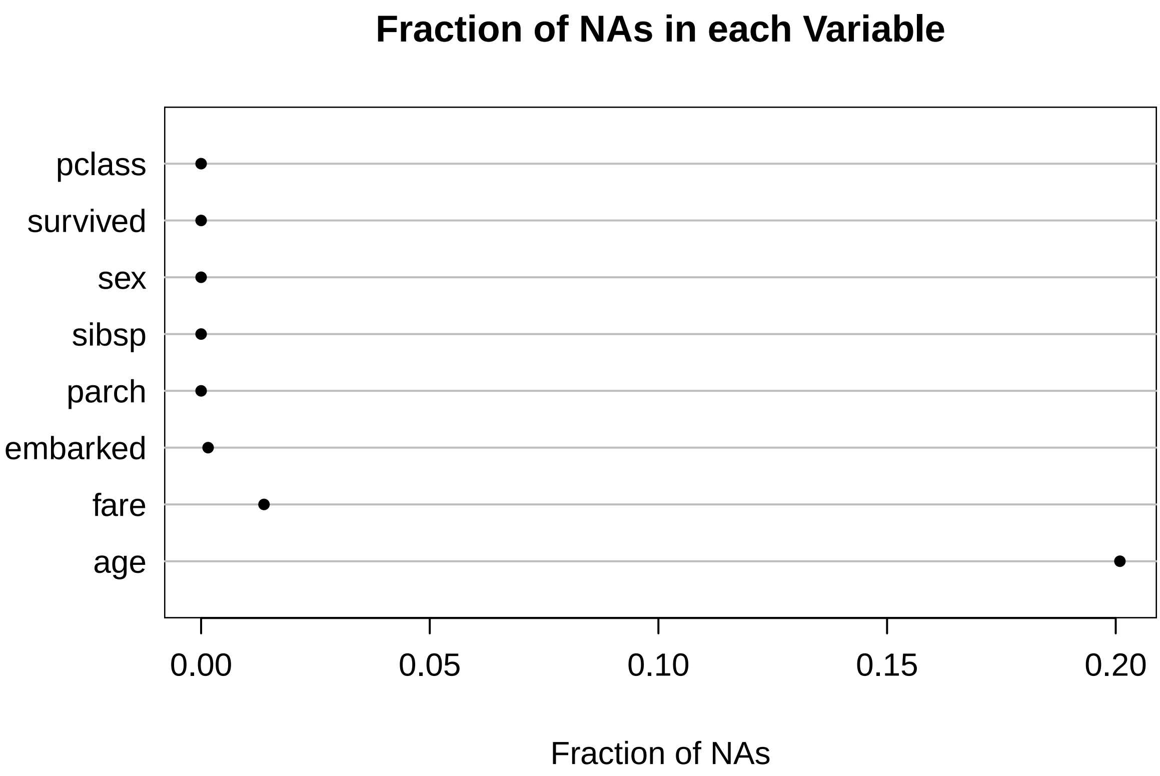
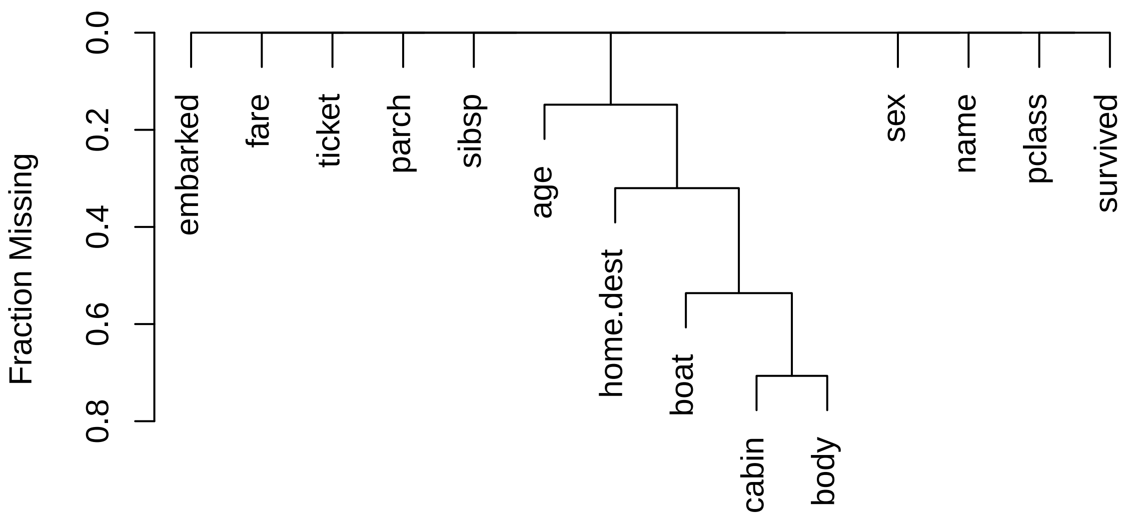


Figure 2. Summary of relationship between survival and each predictor

```r
naplot(naclus(titanic), "na per var")
```




```r
plot(naclus(titanic_raw))
```



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

For demonstration purposes I will fit two decision trees to predict

2.3. *Loess regression for nonlinear pattern*

The loess is a common nonparametric regression method to study nonlinear relationship. In the case of binary response, the fitted value at $x = x_0$ is the proportion of positive cases near the neighborhood of x_0 [7]. If the trend of a loess curve shows non-monotoncity, it is reasonbale to include that nonlienarity relationship in the model, e.g., modeling the predicotr with polynomial transformation or with splines.

figure 2

[7] with varying weights according to their distance to x_0

2.4. *Multiple imputation*

3. Modeling

the choice of model. In this setting, it is obvious that we would prefer probabilistic predictions to classification with output label 0 and 1, since we are placing emphasis upon the *tendency* of survival. And the true value of our model consist not in the decision on who will survive, but in what characteristics would increase or decrease the possibility of survival. This idea has ruled out most of the black box machine learning models for classification, say, random forest, support vector machines and neural network. Not only are they not intrinsically probability oriented, it is hard to interpret main effects and interactions as everything seems to be interacted with one another.

3.1. *Saturated model*

First and foremost,

The limiting sample size for binary outcome would be the number of minority class, in our case 500. Using the 15:1 rule, that will give us some confidence spending roughly 33 parameters or degrees of freedom.

This plot is useful in identifying weather the relationship between survival status and any predictor is flat [8].

likely shrinkage

3.2. *Validation*

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.

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

3.3. *The final model*

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.

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

[8] A misuse of this plot would be checking nonlinearity. Even with spline transformation and large corrected χ^2 there is no guarantee for nonlinearity.

5. Conclusion

Appendix A. Measures used in valiation

This will be Appendix A.

Appendix B. Original 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] mice_3.11.0     rms_6.0-1       SparseM_1.78    Hmisc_4.4-1
[5] Formula_1.2-4   survival_3.1-12 lattice_0.20-41 ggplot2_3.3.2
[9] dplyr_1.0.2

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.5          mvtnorm_1.1-1       tidyr_1.1.2
 [4] png_0.1-7           zoo_1.8-8           assertthat_0.2.1
 [7] digest_0.6.25       R6_2.4.1            backports_1.1.10
[10] MatrixModels_0.4-1  evaluate_0.14       pillar_1.4.6
[13] rlang_0.4.8         multcomp_1.4-14     rstudioapi_0.11
[16] data.table_1.13.0   rticles_0.16.1      rpart_4.1-15
[19] Matrix_1.2-18       checkmate_2.0.0     rmarkdown_2.4
[22] labeling_0.3        splines_4.0.2       readr_1.4.0
[25] stringr_1.4.0       foreign_0.8-80      htmlwidgets_1.5.2
[28] munsell_0.5.0       broom_0.7.1         compiler_4.0.2
[31] xfun_0.18           pkgconfig_2.0.3     base64enc_0.1-3
[34] htmltools_0.5.0     nnet_7.3-14         tidyselect_1.1.0
[37] tibble_3.0.4        gridExtra_2.3       htmlTable_2.1.0
[40] bookdown_0.21       codetools_0.2-16    matrixStats_0.57.0
[43] fansi_0.4.1         crayon_1.3.4        conquer_1.0.2
```

```
[46] withr_2.3.0         MASS_7.3-51.6      grid_4.0.2
[49] nlme_3.1-148        polspline_1.1.19   gtable_0.3.0
[52] lifecycle_0.2.0     magrittr_1.5       scales_1.1.1
[55] cli_2.1.0           stringi_1.5.3      farver_2.0.3
[58] latticeExtra_0.6-29 ellipsis_0.3.1     generics_0.0.2
[61] vctrs_0.3.4         sandwich_3.0-0     TH.data_1.0-10
[64] RColorBrewer_1.1-2  tools_4.0.2        glue_1.4.2
[67] purrr_0.3.4         hms_0.5.3          jpeg_0.1-8.1
[70] yaml_2.2.1          colorspace_1.4-1   cluster_2.1.0
[73] knitr_1.30          quantreg_5.73
```

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 2.4, 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.

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

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

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.

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.

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

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.

Wickham, Hadley, and Jim Hester. 2020. *readr: Read Rectangular Text Data*. R package version 1.4.0, https://CRAN.R-project.org/package=readr.