Determinants of Voting Behaviour in Recent Australian Elections

Joshua Myers

Abstract

The 2016 and 2019 Australian federal elections were both won by the Liberal-National Coalition (LNP).The Australian Election Study (AES) longitudinally surveyed voters after each election with questions in two main categories: attitudes towards election issues, such as the economy and education, and background information such as age, occupation and religion. The purpose of this study was to mine the AES survey data for insights into factors that shaped the choices of LNP voters in the 2016 and 2019 elections. Logistic regression was used to model the probability of voting LNP by fitting three models: one using only responses on election issues, another using background information only, and a third using responses from both categories. Models were fitted on a training sample, with a penalty applied to minimise over fitting, and robust standard errors used to account for repeated observations on respondents between the two elections. The model that used the election issues responses only was the best fitting, and was taken as the final model for further evaluation and interpretation. The model had high discriminative ability on the training sample with area under the receiver operating characteristic curve (AUC-ROC) of 0.94, which generalised well to the validation sample with AUC-ROC of 0.90. LNP voters were more likely to agree that the unions were too powerful, list the economy and superannuation as important issues, disagree that big business was too powerful, think the economy had improved compared to the previous year, and agree that all boats should be turned back. The model had high discriminative ability, and gave insight into the most important issues for voters who voted for the LNP in the 2016 and 2019 elections.

Introduction

The Liberal-National Coalition (LNP) was victorious in both the July 2016 and May 2019 federal elections. The LNP narrowly won in 2016, and again retained power in 2019 in defiance of the predictions of all the major polls (McAllister et al. 2019). The Australian Election Study (AES) longitudinally surveyed voters after the 2016 and 2019 elections, asking voters which party they voted for, attitudes toward election issues, such as the economy and climate change, and background information such as gender, religion and occupation (McAllister et al. 2019).

The purpose of this study was to mine the AES survey results to investigate factors that shaped the voting behaviour of LNP voters in the House of Representatives in the 2016 and 2019 elections.

Data

The AES was a nationally representative longitudinal survey of Australian voters in the 2016 and 2019 federal elections (McAllister et al. 2019). The data set is available online in both long and wide formats (McAllister et al. 2019). This analysis used the long format data set, which has a column for year, rather than a 2016 and 2019 variable for each question in the survey.

The long form data set contains 1628 observations from 814 participants, and 259 columns. A subset of 45 columns were selected for this analysis, including respondents' interest in politics, who they voted for in the

House of Representatives, attitudes toward election issues such as climate change, the economy and asylum seekers, and background variables such as education, occupation and gender (McAllister et al. 2019). An additional variable `age` was calculated by subtracting `year` (the year that the survey was completed) from `year_of_birth`. The `year_of_birth` variable was subsequently dropped from the data set.

A new variable `vote_lnp` was created that was "Yes" if the voter voted for the Liberal or National Party in the House of Representatives, otherwise "No". This variable served as the outcome for the analysis. A total of 697 respondents voted LNP out of the the 1628 rows in the data set. Table 1 shows each of the factor variables selected for the analysis, showing the variable name, number of missing observations, whether the factor was ordered, and the number of levels. Table 2 displays information on numeric variables showing the variable name, number missing, and distributional information. Categorical variables with less than 100 responses in a given category were grouped together with other categories. There were 25 variables for which some categories were grouped together in this way. For ordinal variables, infrequent levels were grouped with the closest category (e.g., for the variable `euthanasia` "strongly disagree" was grouped with "disagree"). For nominal variables, categories with fewer than 100 responses were grouped together as "Other".

Table 1: Summary of factor variables showing the name, number of missing values, whether it is ordered, and number of levels.

Variable	Missing (n)	Ordered	Levels (n)
vote_lnp	29	FALSE	2
year	0	FALSE	2
interest_politics	4	TRUE	3
interest_election_campaign	8	TRUE	3
compulsory_vote	4	TRUE	4
lower_vote_age	5	TRUE	3
most_important_issue	58	FALSE	8
second_important_issue	99	FALSE	8
personal_finances_12_months	7	TRUE	4
economy_12_months	27	TRUE	4
unions_powerful	13	TRUE	5
banks_powerful	11	TRUE	4
death_penalty	10	TRUE	5
marijuana_legal	13	TRUE	5
harsher_penalties	9	TRUE	4
women_preferential	14	TRUE	4
turn_back_boats	16	TRUE	5
euthanasia	12	TRUE	4
indigenous_constitution	25	TRUE	4
first_priority	24	FALSE	4
second_priority	62	FALSE	4
immigrants_crime	11	TRUE	5
immigrants_good_economy	13	TRUE	4
immigrants_take_jobs	14	TRUE	4
global_warming	11	TRUE	4
qualification	37	FALSE	6
doing_last_week	35	FALSE	4
occupation	112	FALSE	6
who_work_for	130	FALSE	3
belong_union	66	FALSE	2
gender	15	FALSE	2
religion	49	FALSE	4
marital_status	19	FALSE	3
own_home	14	FALSE	4

Variable	Missing (n)	Ordered	Levels (n)
investment_properties	13	FALSE	2
self_managed_super	15	FALSE	2
social_class	16	FALSE	3
size_town	19	TRUE	4
annual_income	66	TRUE	4
own_shares	20	FALSE	2
state	1	FALSE	6

Table 2: Summary of numeric variables, showing the name, number missing, the mean, standard deviation (SD), minimum and maximum values.

Variable	n missing	Mean	SD	Minimum	Maximum
age_left_school	54	16.96	3.56	12	98
years_tertiary_study	135	3.65	2.91	0	21
age	22	57.34	15.33	18	95

Methods

The analysis was conducted in the `RStudio` programming environment using functions from base `R`, and the `tidyverse`, `mice`, `MASS`, `rms`, and `pROC` packages (R Core Team 2019; RStudio Team 2019; Wickham et al. 2019; van Buuren and Groothuis-Oudshoorn 2011; Venables and Ripley 2002; Harrell Jr 2019; Robin et al. 2011).

Missing data

The 29 observations with missing values for `vote_lnp` were removed from the data set, it is generally inadvisable to impute data for the outcome (Steyerberg 2019). Tables 1 and 2 show the number of missing values for all of the variables. Of the remaining 1599 observations after removing observations with missing `vote_lnp` data, the variable `years_tertiary_study` had the most missing observations, with 129 missing (0.08). The variable `who_work_for` had the next highest, with 127 (0.08), followed by `occupation` with 111 missing (0.07). The remaining variables ranged from 93 for `second_important_issue` missing values to for 0 for `year` and `id` (see Tables 1 and 2). The amount of missingness was considered acceptable (<10% for all variables), so missing values were imputed with multiple imputation using the `mice` function from the `mice` package. Numeric variables were imputed using predictive mean matching, binary factors with logistic regression, categorical factors with multinomial logistic regression and ordinal factor variables using ordinal logistic regression (van Buuren and Groothuis-Oudshoorn 2011).

Training and validation samples

The imputed data set was split into a training and a validation sample using 70 percent for training and 30 percent for validation using functions from the `dplyr` library (Wickham et al. 2019). The data set was first grouped by respondent `id`, and was split in such a way as that no `id`s in the training sample were included in the validation sample.

Logistic regression modelling

Logistic regression was used to model the probability of voting LNP in the training sample using the `lrm` function from the `rms` package (Harrell Jr 2019). Factor variables were prepared for one-hot encoding for modelling by setting `options(contrasts = c('contr.treatment', 'contr.treatment'))` in base `R`. The shrinkage coefficient (γ) was used to estimate the degree that a model was over fitting on the training set:

$$\gamma = \frac{\text{model } \chi^2 - df}{\text{model } \chi^2},$$

where model χ^2 is the statistical test for the model, and *df* the total degrees of freedom from all variables in the model (e.g., a factor variable with four levels *costs* three *df*). A penalty was applied to the model coefficients if γ was less than 0.95, as this indicated that results on the validation sample would likely be more than 5 percent worse than on the training set (Harrell Jr 2015). The shrinkage penalty was chosen by optimising corrected Akaike's information criterion (AIC) on the training sample (Hurvich and Tsai 1989; Harrell Jr 2015). AIC is a measure of model fit that penalises complexity (Harrell Jr 2015), and is useful for evaluating model fit with relatively small sample sizes. AIC was a useful criterion for model fit ror this analysis, since there were only 1127 observations from 571 participants in the training sample.

An assumption of logistic regression is that observations are independent, but this was not the case in the training sample, because respondents were surveyed in both the 2016 and 2019 elections. Violation of this assumption can result in biased standard errors and *p*-values. Therefore, the Huber-White robust standard errors method was used to account for this grouping in the data using the `robcov` function from the `rms` package (Huber 1967; White 1980; Harrell Jr 2019).

The data set had variables in two main categories, attitudes toward election issues, such as the economy and immigration, and background information such as age, occupation and religion. To evaluate the importance of information from these categories, three models were fit on the training sample. One model was fit using only the election issues variables, another using only the background information, and a third used all the variables (both issues and background). The `year` variable that the survey was completed was included in all models. The model with the lowest AIC on the training set was taken as the final model. This process also served as a variable reduction method, since variables from a category of questions were dropped if they did not contribute information to the model (Harrell Jr 2019).

Model evaluation and interpretation

The final model was evaluated on the training and test samples using accuracy, error, precision, recall, F1 score and area under the receiver operating characteristic curve (AUC-ROC). AUC-ROC was calculated using the `auc` and `roc` functions from the `pROC` package (Robin et al. 2011).

Analysis of variance (ANOVA) was used to infer important factors contributing to voters voting for the LNP. Variable importance was assessed using the `anova` function in the `rms` package (Harrell Jr 2019), which calculated χ^2 statistics for each variable in the model. The six most important variables in the model were investigated graphically for further insights into how these factors affected voting behaviour.

Results

Model fitting

Information for the three fitted models, including γ, penalty, and AIC is provided in Table 3. The γ was less than 0.95 for all three models, so a penalty was applied to each model to reduce over fitting. The penalty was chosen by optimising corrected AIC using the `pentrace` function in the `rms` package. The penalty for each model was applied using the `update` function and robust standard errors with the `robcov` function from the rms package (Harrell Jr 2019). The election issues model had the lowest AIC (892.7), and was therefore taken as the final model for further evaluation and interpretation.

Model evaluation

Table 4 shows the performance metrics for the final model on the training and validation samples. Overall, performance is very high on the training sample (AUC-ROC = 0.94), showing excellent discriminative ability that generalised well to the validation sample (AUC-ROC = 0.9). Other performance metrics also showed high accuracy and validated well, with performance on the validation sample between 3 and 8 percent lower than on the training sample.

Table 3: Model information for the three fitted models, including the shrinkage coefficient γ (gamma), penalty used and Akaike's information criterion (AIC).

Model	Gamma	Penalty	AIC
Full	0.87	6.5	920.80
Background	0.84	5.0	1349.46
Issues	0.91	4.2	892.70

Table 4: Performance metrics for the election issues model on the training and validation samples.

Metric	Train	Validation
Accuracy	0.87	0.83
Error	0.13	0.17
Precision	0.85	0.77
Recall	0.86	0.84
F1 score	0.85	0.80
AUC	0.94	0.90

Model interpretation

Figure 1 displays the ANOVA results for the election issues model, showing the importance of predictors determined by the amount of χ^2 contributed by a variable. The variable "unions too powerful" is clearly the most important predictor, followed by "most important issue", and then "big business too powerful". The variables "economy past year", "turn back all boats" and "second important issue" area also clearly separated from other variables in terms of the amount of χ^2 contributed to the model.

Figure 2 shows how the probability of voting LNP changed for the six most important variables as the factor level changes with all other variables held at typical values. Typical values were the most common level for factor variables and the median for numeric variables. LNP voters were more likely to agree or strongly agree that unions are too powerful, more likely to list the economy as the most important issue, and more likely to either disagree or neither agree or disagree that big business is too powerful. They were more likely to think that the economy had improved over the past year, agree or strongly agree that all refugee asylum seeker boats should be turned back, and were more likely to list superannuation or the economy as the second most important election issue.

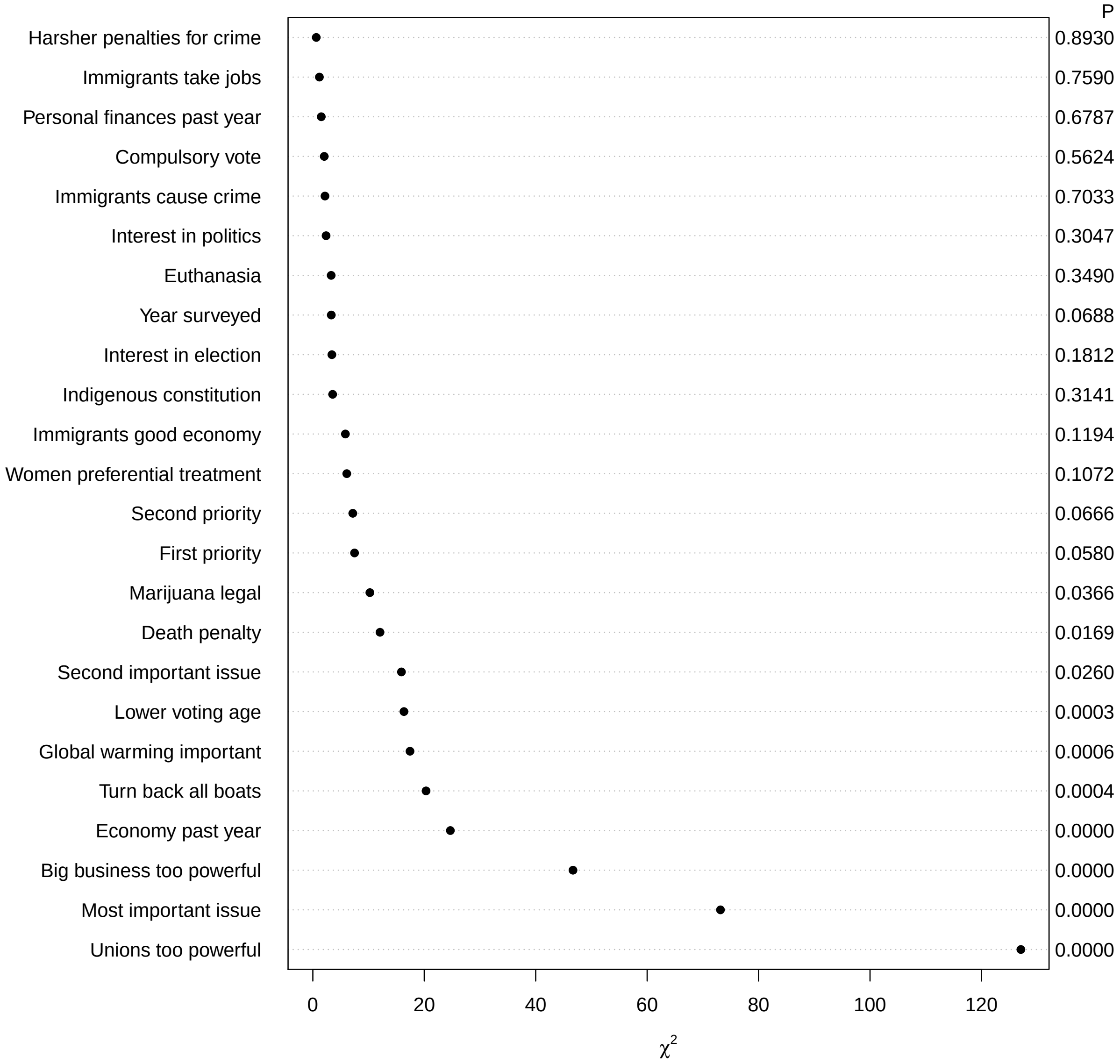

Figure 1: Importance of predictors in the election issues model in terms of amount of χ^2 that a variable contributed to the model (x-axis), p-values for each variable are shown on the right.

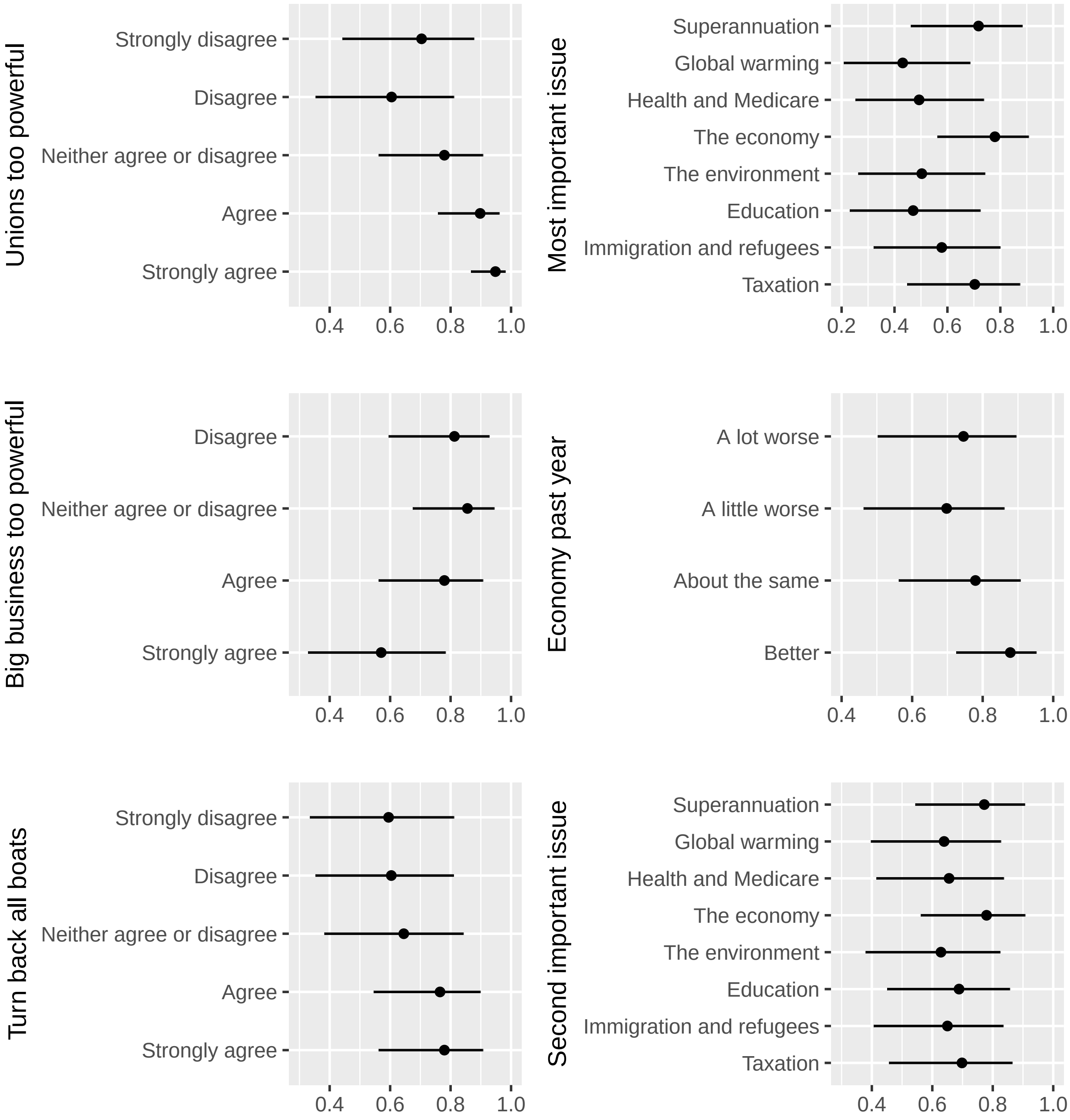

Figure 2: The six most important variables in the logistic regression model, showing how the probability (x-axis) of voting Liberal changes as the factor level changes with all other variables held at typical values (the most common level for factor variables and the median for numeric values). The points are the prediction, and the lines extending to either side depict the 95% confidence intervals.

Discussion

The election issues model indicated that the issues of union power, management of the economy, boat turn-backs and superannuation played a central role in the outcomes of the 2016 and 2019 elections. The issues model was better fitting (lower AIC) than both the full model (all variables), and the background variables model, indicating that knowing where a voter stands on election issues is more informative than knowing their background and demographic information. Furthermore, including background details did not contribute substantial information once attitudes toward election issues were known. This indicates that knowing voters attitudes toward election issues is more important than background or demographic information when predicting voting behaviour.

Choosing the the best fitting model from a small pool of candidate models served as a data reduction mechanism by excluding 20 variables that did not contribute additional information to the model (Harrell Jr 2015). This process helped to reduce overfitting and made the model easier to interpret (James et al. 2013). The year that the survey was completed was not statistically significant in the final model ($p = 0.06$), indicating that voting behaviour was fairly stable between elections. Indeed, in the data set, of the 1599 participants in the data set, only 104 changed their vote between the 2016 and 2019 elections.

An important limitation to this study is that it is not designed to predict future LNP voting behaviour. Whilst the model provided insights into the 2016 and 2019 elections, issues that are important to voters will change over time, and the relative importance may change over time, even for long standing issues such as the economy.

Conclusion

This analysis mined the AES survey data set for insights into determinants of LNP voting behaviour in the House of Representatives the 2016 and 2019 federal elections. The election issues model was better fitting than both the background information model and the model using all variables, indicating that voter attitudes toward issues was more informative than background information, and adding background information did not contribute additional information once attitudes toward issues was known. The election issues model had excellent performance on the training sample that generalised well to the validation sample, and gave insight into important issues for LNP voters. LNP voters were more likely agree that unions were too powerful, list the economy and superannuation as important issues, disagree that big business was too powerful, think that the economy had improved in the previous year, and agree that all refugee asylum boats should be turned back.

Reference List

Harrell Jr, Frank E. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* New York: Springer.

Harrell Jr, Frank E. 2019. *rms: Regression Modeling Strategies.* https://CRAN.R-project.org/package=rms.

Huber, Peter. 1967. “The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions.” *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 59–82.

Hurvich, CM, and CL Tsai. 1989. “Regression and Time Series Model Selection in Small Samples.” *Biometrika* 76: 297–307.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Vol. 112. New York: Springer.

McAllister, Ian, Clive Bean, Rachel Gibson, Toni Makkai, Jill Sheppard, and Sarah Cameron. 2019. “Australian Election Study, 2019.” ADA Dataverse. https://doi.org/10.26193/KMAMMW.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for R and S+ to Analyze and Compare Roc Curves." *BMC Bioinformatics* 12: 77.

RStudio Team. 2019. *RStudio: Integrated Development Environment for R.* Boston, MA: RStudio, Inc. http://www.rstudio.com/.

Steyerberg, Ewout W. 2019. *Clinical Prediction Models.* New York: Springer.

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/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48: 817–38.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.