The 2019 Canadian Federal Election and its Results: Forecasting with Post Stratificiation

Yaqi Feng 1003925443

December 21, 2020

github: https://github.com/LindaFeng0830/STA304FinalProject

Abstract

This study aims to use multilevel regression with post-stratification to predict the 2019 Canadian federal election results using *the 2019 Canadian Election Survey* ("CES") and *the 2016 Census Education Highlight Tables* ("Census"). The prediction is that the Liberal Party wins the general election with 82,520 votes, becoming a minority government of Canada with Justin Trudeau serving as the Prime Minister. The results can be taken and compared to the actual results in hopes of reproducing a similar model and predictions for the upcoming Canadian federal election, or any country's election that adopts a similar system.

Keyword

Multilevel Regression, *Post-stratification*, *Electoral Results Prediction*

Introduction

Canada's electoral system is based on a parliamentary system of government, modeled on that of the United Kingdom, commonly referred to as a "single-member plurality" or "first-past-the-post" system. In this system, any political party, with one candidate, who wins the highest number of votes wins the election and the right to serve as Prime Minister. An absolute majority is not required, unlike the system in the United States, for a candidate to be elected (The Electoral System of Canada, 2020).

The goal of this report is to predict the results of the 2019 Canadian Federal Election under said system, by using data from the CES and the 2016 Census data. Specifically, the model will be multilevel logistic regression, and post-stratification will be applied afterward.

The sections that follow will elaborate on the data selection basis, methodology and models used, results, and concluding with a discussion on the results, model limitations, and possible improvements. Analysts or organizations, who are interested in electoral outcomes, can apply the model and result in a similar electoral system in hopes of predicting the outcome.

Data

The two datasets used in this study are: the *the 2019 Canadian Election Survey* ("CES") and *the 2016 Census Education Highlight Tables* ("Census"). The former is loaded in directly from R and the latter is retrieved from Statistic Canada. Before cleaning, the CES contains responses from 37,822 respondents, and the Census contains information 1,512 entries [1]. Both datasets are reduced so that no reoccurring variables and NAs. Generally speaking, the **target population** of the CES is every Canadian citizen who has registered to vote and will be 18 or older on election day. The **frame population** is a portion of the target population to which the online CES survey delimit, identify, and subsequently allow access to (Fricker, 2015). The **sample population** is hence the 37,822 respondents who completed the CES survey online [2].

Both the CES and the Census are detailed in information collecting and some variables may represent the same information under different titles. After a close inspection and careful selection, four variables, represented in both the CES and the Census, are selected to become the predictor variable of interest. They

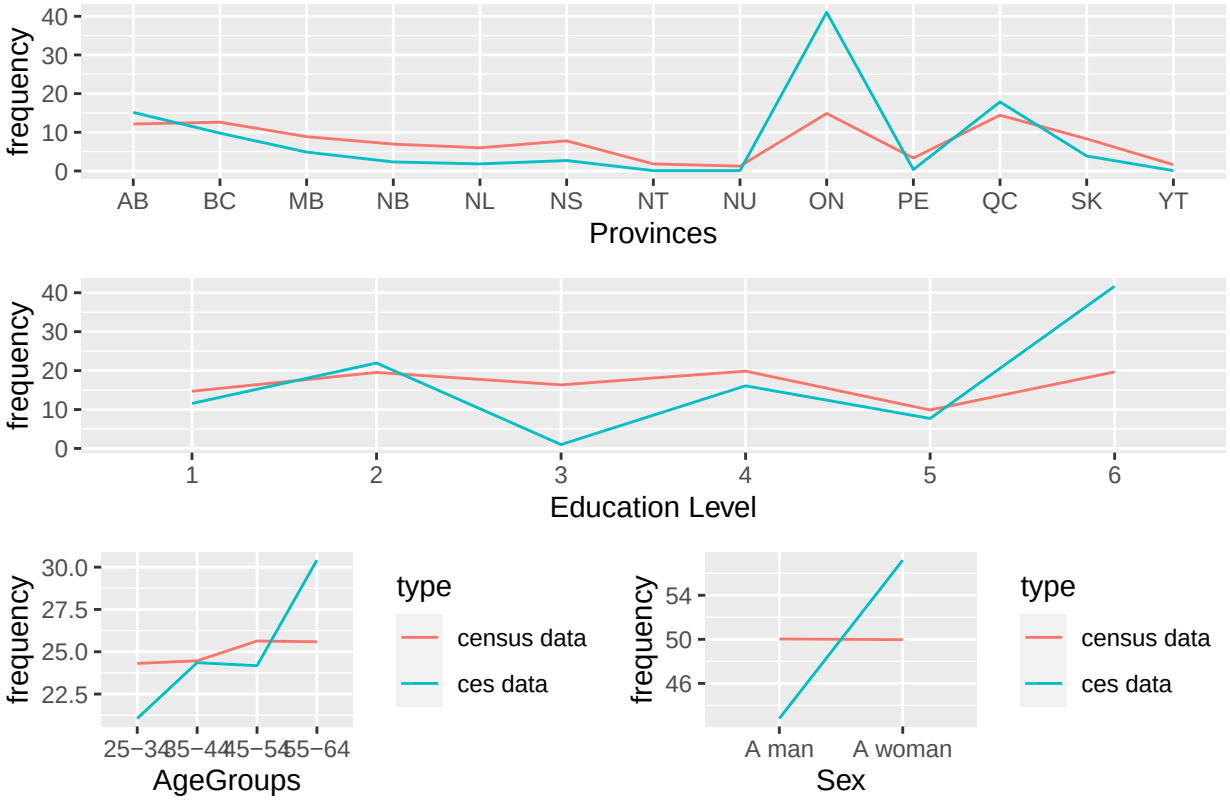
[1] The Census contains visibly less observation here because it is a fraction, dedicated to education and basic demographic information, of the complete 2016 Canadian Census. This snapshot of Census is chosen rather than choosing the complete Census because it better captures for variables that are represented in CES.

[2] The target population of the Census consists of: Canadian citizens (by birth or by naturalization); landed immigrants (permanent residents); and non-permanent residents and their family members living with them in Canada. The frame population would be those who have received the web access codes and capable of completing the census. The sample population is the proportion of the frame population, reportedly 98.4%, who responded to the census.

are sex, age, geographic location, and education level. Note that the choice "Other (e.g. Trans, non-binary, two-spirit, gender-queer)" available under Sex in the CES is filtered out because it is not contained in the Census data. For similar reasons, in the Census, "All ages, 15-plus" is also filtered out because only citizens older than 18 are eligible to vote, and the dataset does not provide further detailed grouping for the group labeled "15-plus". In the end, the cleaned CES and Census data contain 12,359 and 624 observations, respectively.

Below is a comparison graph of the CES and Census data. [3] [4]

Figure 1: CES data vs. Census data



Model and Methodology

All of the following model-building processes and analyses are done in R studio.

A multilevel logistic regression model is built to predict the electoral results. The CES data is carefully portioned into 8 cells, each cell representing an interaction variable of sex and age, which are the two variables deemed significant to the voting choice. Sex is also treated as a random intercept variable to capture the impacts of different sexes on the model results.

A multilevel logistic regression is suitable for predicting vote choices because it accounts for the clustering of subjects within clusters of higher-level units when estimating the effect of subject and cluster characteristics

[3] Provinces abbreviation by Canada Post. Alberta "AB"; British Columbia "BC"; Manitoba "MB"; New Brunswick "NB"; Newfoundland and Labrador "NL"; Nova Scotia "NS"; Northwest Territories "NT"; Nunavut "NU"; Ontario "ON"; Prince Edward Island "PE"; Quebec "QC"; Saskatchewan "SK"; Yukon "YT".

[4] Education level by dataset order. 1 for "Apprenticeship or trades certificate or diploma (2016 counts"; 2 for "College, CEGEP or other non-university certificate or diploma (2016 counts)"; 3 for "No certificate, diploma or degree (2016 counts)"; 4 for "Secondary (high) school diploma or equivalency certificate (2016 counts)"; 5 for "University certificate or diploma below bachelor level (2016 counts)"; 6 for "University certificate, diploma or degree at bachelor level or above (2016 counts)".

on subject outcomes (Austin & Merlo, 2017). The logistic regression allows for accurate estimation when the dependent variable is binary, a choice between conservative or liberal in this case. In simpler terms, this model is appropriate to use here as its results can be used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

The regression model can be expressed in mathematical terms as follows:

$$Pr(Y_i \in Conservative, Liberal) = logit^{-1}(a_m + a_{j[i]}^{Education} + a_{j[i]}^{Province})$$

Where a_n is the random intercept variable indicating a particular sex group's impact on a respondent's voting choice, and n is that sex group. Education and Province are the predictor variables because a respondent's education level and geographic location are assumed to have a considerable impact on his or her voting choice. For example, individuals living in a province, where there is a historical pattern of voting for liberal, may be impacted by the environment and more likely to vote for liberal. However, if he or she has completed a certain level of education, that experience may lead them to choose otherwise instead. The expressions $a_{j[i]}^{Education}$ and $a_{j[i]}^{Province}$ are the coefficients for level 1 categorical variable, education and province, and the footnote j[i] indicates the cell that the ith respondent belongs. All variables used in this model can be found in both the CES and the Census.

The glmer() function, found in R package lme4 is used to formulate an approximate marginal maximum likelihood estimate to transform the datasets into estimates of the 2019 Canadian Federal Election results. In the preceding section titled "Discussion", a graph depicting the Area Under the ROC Curve ("AUC") will be presented and the model accuracy, as well as performance, will be discussed thoroughly.

Post-stratification

Post-stratification corrects for non-sampling error and allows for fewer variable estimates. That is, given a population composed of distinct groups, strata, or clusters that differ with regard to the quantity relevant to model estimation, and if the size of these strata can be observed, then post-stratification can obtain a more accurate estimate of the quantity of interest by correcting for any imbalance in the representation of the strata in the sample (Reilly, Gelman, & Katz, 2001).

To apply the post-stratification technique, using the Census data, it is assumed that there is no significant change in the Canadian population between 2016 and 2019. The post-stratified proportion of voters, using the Census data, voting for either the Liberal Party, represented by Justin Trudeau, or the Conservative Party, represented by Andrew Scheer, can be expressed as follows:

$$\widehat{y}^{PS} = \frac{\sum N_j \widehat{y_j}}{\sum N_j}$$

This expression aims to estimate the weighted average of the proportion of voters choosing Liberal or Conservative, where N represents each person's assigned weight and $\widehat{y_j}$ represents the voting proportion estimator for the parties. Notice that in this model, the election results are calculated as if there is only a binary choice between Liberal and Conservative. However, Canada, by adopting the "first-past-the-post" electoral system, actually allows for multiple candidates to be elected at once. In the "Weaknesses" section that follows, this point will be discussed in detail.

The post-stratification process consists of two steps. To begin with, a logical function is used to determine whether a particular party has over half (0.5) of the support in a specific demographic group [5]. If so, the number of votes that belong to this specific demographic group is awarded to that party.

[5]The "demographic group" here refers to the groups represented in the Census, comprised of Age, Sex, Province, and Education criteria. For example, one group could be a man aged 25-34 from Ontario with no certificate, diploma, or degree.

Table 1: Table 1: Election Result

Liberal Party	Conservative Party
82520	71894

Results

The regression results after post-stratification indicate that the liberal party will receive 82,520 votes, winning the conservative party by 10,626 votes. The conservative party will receive 71,894 votes. In practical terms, the liberal party will win the federal election and becomes the minority government of Canada. And, its party leader, Justin Trudeau, gets the right to serve as the Prime Minister. The conservative party loses by a slight disadvantage and becomes the Official Opposition instead.

The Area under the ROC Curve ("AUC"), depicted in Figure 1 below, is 0.69. AUC tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting values at their actual value. In this study, the higher the AUC, the better the model is at distinguishing between voters who choose to vote for liberal and for conservative. An AUC value of 0.69 indicates that this model is relatively effective in doing so.

Figure 2: ROC Curve

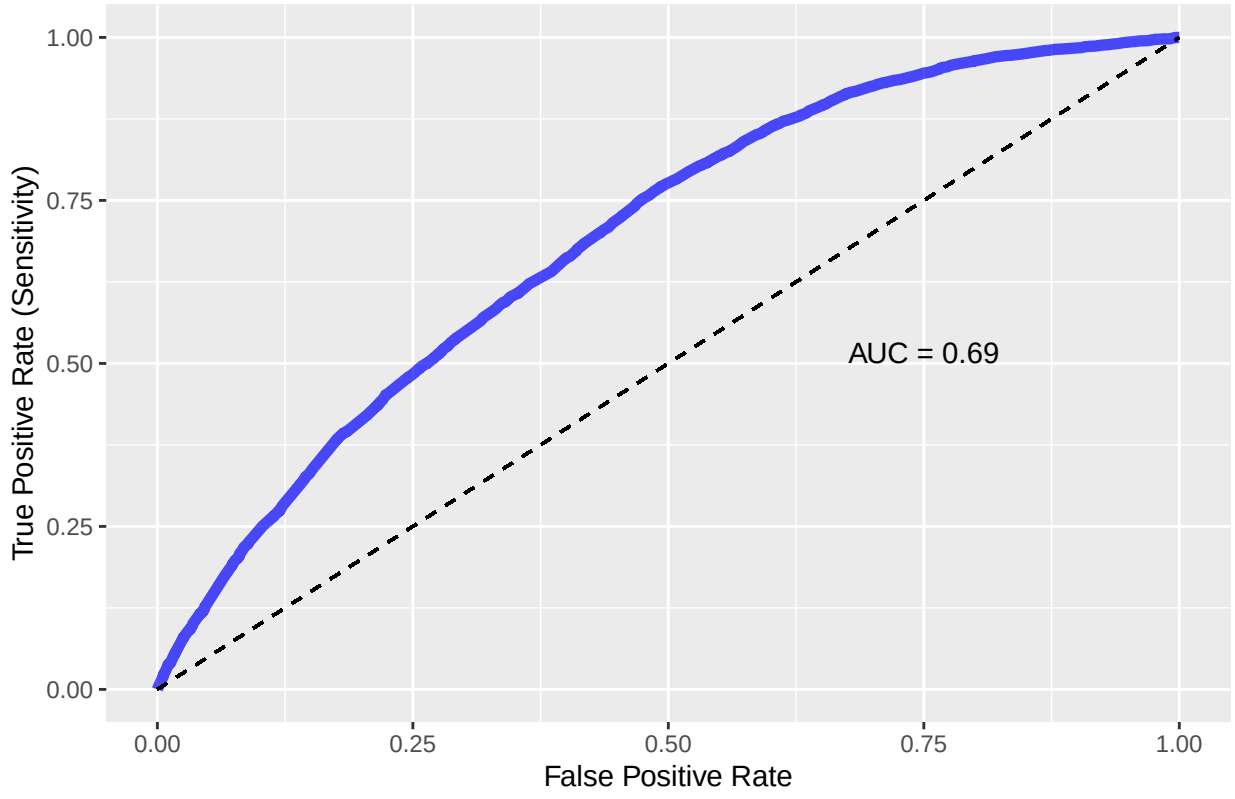


Table 2: Table 2: Votes for Liberal or Conservative by cell

cell	Liberal Party	Conservative Party
A man 25 to 34	9928	8960
A man 35 to 44	9367	9550
A man 45 to 54	6963	12748
A man 55 to 64	8535	11201
A woman 25 to 34	12576	6073
A woman 35 to 44	10883	7978
A woman 45 to 54	11916	7957
A woman 55 to 64	12352	7427

Discussion

In this report, the CES and Census data have been used to predict the 2019 Canadian federal election results. The CES, conducted in 2019, was done through an online survey and thereby is subject to sampling biases like the volunteer bias. Therefore, the Census, conducted in 2016, is used to apply the post-stratification technique onto the multilevel logistic regression model built using the CES data.

The CES data was carefully portioned into 8 cells, as described in the previous sections, each representing an interaction variable of sex and age. The table below summarizes the predicted number of votes awarded to each party in each cell:

In conclusion, based on the estimation, the liberal party wins the federal election at a slight advantage, 82,520 over 71,894 votes, and becomes a minority government of Canada. Its party leader, Justin Trudeau, subsequently wins the right to serve as Prime Minister. The model has an AUC of 0.69, suggesting that the model performs fairly well on distinguishing the proportion of voters voting for the liberal party or the conservative party.

Weaknesses

In the data cleaning process, the size of the Census data and the CES data is significantly reduced. This is due to a few key filtering decisions. Firstly, only respondents who indicated that they are "certain to vote" or "likely to vote" are kept in the CES dataset to maximize the accuracy of the voting choice results. However, this decision may be subject to sampling bias since the responses may not reflect the respondent's actual level of willingness to vote, thereby hindering the accuracy of the prediction results. Secondly, the category "Other (e.g. Trans, non-binary, two-spirit, gender-queer)" is removed from the CES data because it is not represented in the Census data, which only contains a binary choice of a male or a female. The removal of an entire category may be subject to an under-coverage bias, suggesting that some members of the population, in this case, the people who identify their sex as "Others", are inadequately represented in the sample. The underrepresentation of an entire demographic group will negatively impact the model's accuracy, causing deviations from the actual result, especially when both the random intercept variable and the cell involves the variable "Sex" [6].

Another weakness of this model, and possibly the most visible one, is that this model assumes that the Canadian electoral outcome is binary. As discussed in the model section, the multilevel logistic regression model can appropriately describe the data and explain the relationship between one dependent binary variable and numerous independent variables. However, in reality, the Canadian electoral system allows for multiple candidates, and a candidate does not have to win an absolute majority to declare an electoral victory. In fact, in the actual 2019 election, none of the parties won over 50% of the votes. Therefore, even if this model did predict correctly that the liberal party wins the election, the predictions on the exact number

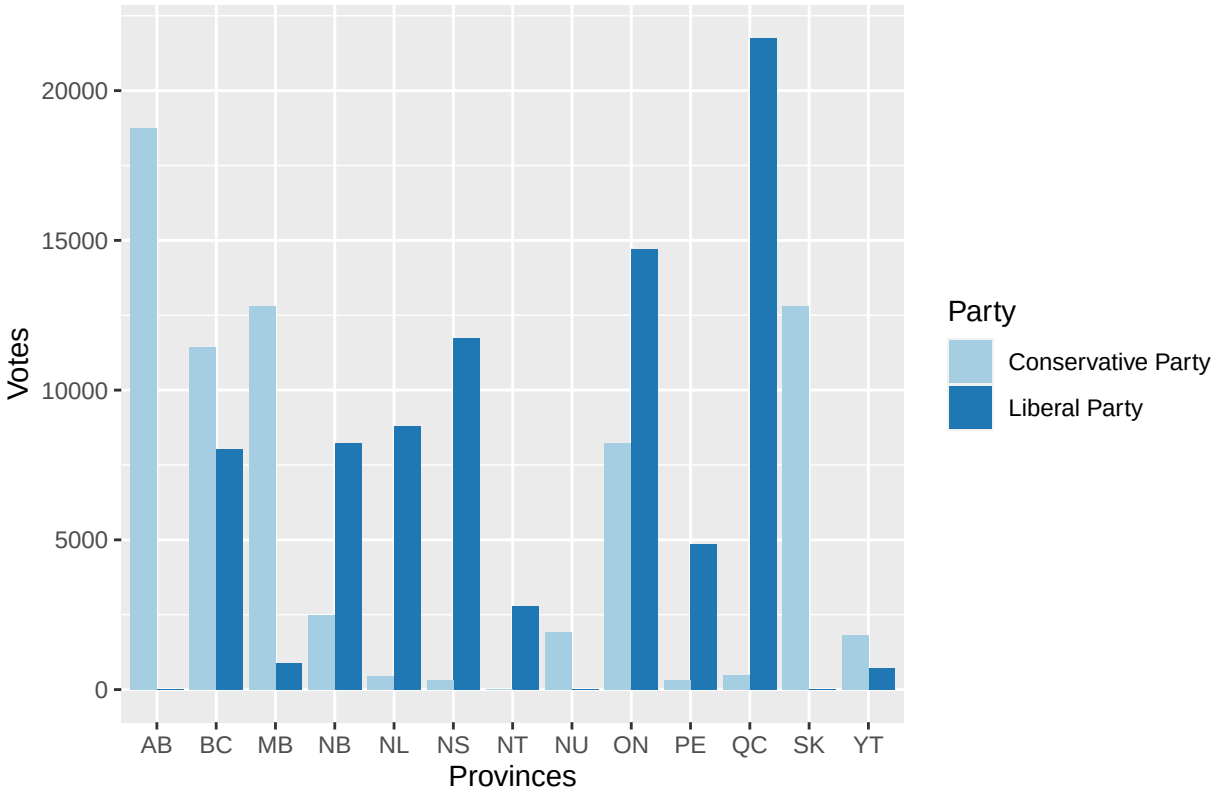
[6]The age group "All ages, 15-plus" is removed for similar reasons as removing the "Other" group in Sex.

Table 3: Table 3: Votes for Liberal or Conservative by Province

Province	Liberal Party	Conservative Party
Alberta	0	18761
British Columbia	8031	11437
Manitoba	875	12821
New Brunswick	8245	2481
Newfoundland and Labrador	8803	466
Northwest Territories	2804	0
Nova Scotia	11722	314
Nunavut	0	1937
Ontario	14712	8245
Prince Edward Island	4856	312
Quebec	21770	492
Saskatchewan	0	12822
Yukon	702	1806

of votes in total and in each province are likely flawed. This weakness can also be easily observed as the model indicates some provinces contribute zero votes to the parties.

Figure 2: Votes for Liberal or Conservative by Province



Next Steps

Going forward, the model results should be compared closely to the actual 2019 federal election results to evaluate the significance of each predictor variable. The model efficiency and accuracy should also be analyzed but compared in a sophisticated way so that the weaknesses can be somewhat neglected. For example, perhaps the proportion of voters voting for liberal or conservative should be compared and analyzed rather than the actual number output from the model. Moreover, remodeling can be done using a different approach. For example, a multinomial regression is perhaps more suitable for the Canadian electoral system. Such next steps to take should grant the model presented in this report more predictor variables, or even more model to work with, thereby improving both the model efficiency and accuracy.

References

1. Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

2. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, '2019 Canadian Election Study - Online Survey', https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1 Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science.

3. 2016 Education Census

4. Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES Datasets a Little Easier.. R package version 0.1.0.

5. Joseph Larmarange (2020). labelled: Manipulating Labelled Data. R package version 2.7.0. https://CRAN.R-project.org/package=labelled

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

7. Millo G (2017). "Robust Standard Error Estimators for Panel Models: A Unifying Approach." *Journal of Statistical Software*, *82*(3), 1-27. doi: 10.18637/jss.v082.i03 (URL: https://doi.org/10.18637/jss.v082.i03).

8. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

9. David B. Dahl, David Scott, Charles Roosen, Arni Magnusson and Jonathan Swinton (2019). xtable: Export Tables to LaTeX or HTML. R package version 1.8-4. https://CRAN.R-project.org/package=xtable

10. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

11. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

12. Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

13. Kay M (2020). *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi: 10.5281/zenodo.1308151 (URL: https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1, <URL: http://mjskay.github.io/tidybayes/>.

14. Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

15. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005). "ROCR: visualizing classifier performance in R." *Bioinformatics*, *21*(20), 7881. <URL: http://rocr.bioinf.mpi-sb.mpg.de>.

16. Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

17. Xavier Robin, 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, p. 77. DOI: 10.1186/1471-2105-12-77 http://www.biomedcentral.com/1471-2105/12/77/

18. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

19. Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

20. Austin, P. C., & Merlo, J. (2017). Intermediate and advanced topics in multilevel logistic regression analysis. Wiley Statistics in Medicine.

21. Fricker, R. D. (2015). Sampling Methods for Online Surveys.

22. Reilly, C., Gelman, A., & Katz, J. (2001). Poststratification Without Population Level Information on the Poststratifying Variable, With Application to Political Polling. Journal of the American Statistical Association.

23. The Electoral System of Canada. (2020, December 9). Retrieved from Elections Canada: https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e

24. What is Logistic Regression? (n.d.). Retrieved from Statistics Solutions: https://www.statisticssolutions.com/what-is-logistic-regression/