# An MRP Model for the 2019 Canadian Election with Imputation of Missing Votes

Steven Tran

December 22 2020

The supporting code and data for this report are available at this Git repository: [https://github.com/st-tran/sta304-final-project](https://github.com/st-tran/sta304-final-project)](https://github.com/st-tran/sta304-final-project)

## Abstract

This report takes variables from survey data, including the respondents' province of residence, income, whether they were born in Canada, highest level of education attained, their gender, age, how many people live in their household, and use them as cells for poststratification of two logistic models that are fitted to predict the outcome of the 2019 Canadian Federal Election if everyone had voted. Voter turnout is an issue in many elections as competing parties for the majority vote are frequently close to each other.

## Keywords

MRP, multilevel, logistic, regression, poststratification, 2019, election, imputation, census, survey, Canadian

## Introduction

The Canadian Election Study (CES) is a regular study on "Canadians' political behaviour and attitudes" (Stephenson et al. 2020) that has been a rich source of data on political affiliations across demographics since 1965. However, many respondents choose not to answer questions about who they vote for or respond that they spoiled their vote[1] altogether, leading to missing data in statistical models to predict the popular vote. An additional issue common to surveys is that, due to sampling biases, the sampling frame is not representative of the target population of voting-age Canadians, so poststratification is key to adjust for this.

Logistic regression models are a natural fit for predicting binary outcomes because the predicted value, once transformed, is a valid probability. First-Past-the-Post, Canada's electoral system, tends to produce single-party majority governments, so a simple majority vote is employed. The goal of this report is to impute the missing data and include it in a multinomial logistical regression model with poststratification using Canadian census data to obtain the log-odd-ratios of winning the election for each party if everybody had voted.

In the Methodology section, variables are selected from the two datasets and organized in order to make a matching between cells. Two logistic regression models are fitted to predict the probability that the Liberals won as well as the probability that the Conservatives won (the two leading parties), and then poststratified using weights. The results, inferences, and conclusions are presented in the Results and Conclusion sections.

---

[1] A spoiled vote is a vote marked as invalid for various reasons, such as improperly filling out a ballot

# Methodology

## Data

In order to fit a model for the election outcome, the predictor variables below were selected from the CES survey and ordinal categorical variables were converted into levelled factors:

- `province`: nominal; describes the province where the respondent resides
    - Ontario, British Columbia, Alberta, Newfoundland and Labrador, Saskatchewan, Prince Edward Island, Quebec, Nova Scotia, New Brunswick, Manitoba
- `income`: ordinal; income of respondent
    - Less than $25,000 (baseline)
    - $25,000 to $49,999
    - $50,000 to $74,999
    - $75,000 to $99,999
    - $100,000 to $124,999
    - $125,000 and more
- `bornin_canada`: ordinal; describes whether the respondent was born in Canada
    - No (baseline)
    - Yes
- `education`: ordinal; highest level of education attained by the respondent
    - Less than high school (baseline)
    - High school
    - College
    - Bachelor's degree
    - Above bachelor's level
- `gender`: ordinal; gender of respondent[2]
    - A man (baseline)
    - A woman
- `age`: quantitative; age of respondent
- `household`: quantitative; describes how many people live in the respondent's household

The equivalent variables were selected from the GSS dataset used for later use in post-stratification to adjust the survey data to the population characteristics.

## Model

Next, two multilevel regression models were fitted to the CES data, using two more variables:

- `weight`: quantitative; CES-assigned weights for the respondent

|  | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| age | 18.0000000 | 43.0000000 | 57.0000000 | 54.9450877 | 67.000000 | 99 |
| household | 1.0000000 | 2.0000000 | 2.0000000 | 2.2987099 | 3.000000 | 6 |
| weight | 0.3451671 | 0.6150085 | 0.7836755 | 0.9538693 | 1.116848 | 5 |

- `votechoice`: nominal; the party which the respondent voted for in the 2019 Canadian Election

---

[2]TODO: note about gender

| Party | Votes |
|-------|-------|
| Liberal Party | 8949 |
| Conservative Party | 8713 |
| Don't know... | 4908 |
| ndp | 4328 |
| Green Party | 2456 |
| Bloc Quebecois | 1404 |
| People's Party | 605 |
| Another party | 201 |

The Liberal Party and Conservative Party competed for the majority vote, so the `votechoice` variable was converted into two variables `voted_conservative` and `voted_liberal`, both of which have binary outcomes. The two models are given by (where 'party' is one of the aforementioned parties):

$$\ln\left(\frac{P_{\text{party}}}{1 - P_{\text{party}}}\right) = \beta_0 + \beta_{\text{age}}x_{\text{age}} + \beta_{\text{household}}x_{\text{household}}$$

$$+ \sum_{\substack{i \in \text{gender}, \\ i \neq \text{'male'}}} \beta_i x_i + \sum_{\substack{i \in \text{education}, \\ i \neq \text{'Less than high school'}}} \beta_i x_i + \sum_{\substack{i \in \text{bornin\_canada}, \\ i \neq \text{'No'}}} \beta_i x_i$$

$$+ \sum_{\substack{i \in \text{income}, \\ i \neq \text{'Less than \$25,000'}}} \beta_i x_i + \sum_{\substack{i \in \text{province}, \\ i \neq \text{'Ontario'}}} \beta_i x_i$$

Here, there is an intercept term $\beta_0$ which represents the log-odds of voting for a particular party when all other variables are at baseline (the respondent's age is 0 and nobody lives with them, they are male, haven't finished high school, weren't born in Canada, make less than \$25,000, and live in Ontario). In this context, it isn't meaningful due to such individuals being ineligible to vote in Canada. The quantitative variables `age` and `household` have coefficient terms $\beta_{\text{age}}$ and $\beta_{\text{household}}$ which represent the average change in log odds for a unit increase in each with all other variables held constant, respectively. The remaining variables, which are expressed as sums, use one-hot coding that represents the change in log odds compared to the baseline if the categorical variable of concern moves to the 'next' level; refer to the Data section for the order of these factors.

Associated with the model are the Wald test statistics and corresponding $P$-values for each of the individual regression coefficients under the null hypothesis that the regression coefficient is equal to zero.

## Results

The coefficients of the fitted models can be seen in Table [...]. Across them, many of the predictor variables are similar in terms of statistical significance – that is, a variable that's statistically significant in the model that predicts the Liberal Party's log-odds of winning the election is statistically significant in the model that predicts the Conservative Party's log-odds of winning.

| Variable | Liberal Model | | | Conservative Model | | |
|---|---|---|---|---|---|---|
| | OR | Beta (SE) | P | OR | Beta (SE) | P |
| Intercept | - | -2.34 (0.32) | <0.001 *** | - | 0.04 (0.31) | 0.90 |
| bornin_canada | | | | | | |
| — No (ref) | - | - | - | - | - | - |
| — Yes | 0.86 | -0.15 (0.10) | 0.12 | 0.91 | -0.10 (0.10) | 0.33 |
| province | | | | | | |
| — Alberta (ref) | - | - | - | - | - | - |
| — British Columbia | 2.24 | 0.81 (0.14) | <0.001 *** | 0.31 | -1.17 (0.12) | <0.001 *** |
| — Manitoba | 2.11 | 0.75 (0.19) | <0.001 *** | 0.45 | -0.80 (0.17) | <0.001 *** |
| — New Brunswick | 3.95 | 1.37 (0.23) | <0.001 *** | 0.24 | -1.41 (0.23) | <0.001 *** |
| — Newfoundland and Labrador | 3.71 | 1.31 (0.24) | <0.001 *** | 0.25 | -1.38 (0.25) | <0.001 *** |
| — Nova Scotia | 4.38 | 1.48 (0.20) | <0.001 *** | 0.15 | -1.88 (0.25) | <0.001 *** |
| — Ontario | 3.28 | 1.19 (0.12) | <0.001 *** | 0.30 | -1.21 (0.10) | <0.001 *** |
| — Prince Edward Island | 3.03 | 1.11 (0.49) | 0.02 * | 0.21 | -1.54 (0.82) | 0.06 |
| — Quebec | 2.78 | 1.02 (0.13) | <0.001 *** | 0.10 | -2.30 (0.14) | <0.001 *** |
| — Saskatchewan | 0.79 | -0.24 (0.22) | 0.27 | 0.90 | -0.11 (0.17) | 0.53 |
| education | | | | | | |
| — Less than high school (ref) | - | - | - | - | - | - |
| — High school | 1.40 | 0.33 (0.20) | 0.09 | 0.86 | -0.15 (0.18) | 0.42 |
| — College | 1.31 | 0.27 (0.20) | 0.18 | 0.87 | -0.14 (0.19) | 0.45 |
| — Bachelor's degree | 2.21 | 0.79 (0.20) | <0.001 *** | 0.58 | -0.54 (0.19) | 0.004 ** |
| — Above bachelor's level | 1.91 | 0.65 (0.21) | 0.002 ** | 0.51 | -0.68 (0.20) | <0.001 *** |
| household | 1.00 | -0.00 (0.03) | 0.96 | 1.05 | 0.05 (0.03) | 0.12 |
| age | 1.00 | 0.00 (0.00) | 0.16 | 1.01 | 0.01 (0.00) | <0.001 *** |
| income | | | | | | |
| — Less than $25,000 (ref) | - | - | - | - | - | - |
| — $25,000 to $49,999 | 1.17 | 0.16 (0.15) | 0.28 | 0.95 | -0.05 (0.16) | 0.75 |
| — $50,000 to $74,999 | 1.14 | 0.13 (0.14) | 0.36 | 1.28 | 0.25 (0.15) | 0.09 |
| — $75,000 to $99,999 | 1.05 | 0.04 (0.14) | 0.75 | 1.28 | 0.24 (0.15) | 0.11 |
| — $100,000 to $124,999 | 1.22 | 0.20 (0.15) | 0.19 | 1.40 | 0.34 (0.16) | 0.03 * |
| — $125,000 and more | 1.20 | 0.18 (0.14) | 0.20 | 1.56 | 0.45 (0.15) | 0.003 ** |
| gender | | | | | | |
| — A man (ref) | - | - | - | - | - | - |
| — A woman | 1.22 | 0.20 (0.07) | 0.003 ** | 0.68 | -0.38 (0.07) | <0.001 *** |

Some variables that are highly statistically significant at standard thresholds of 0.001, 0.01, and 0.05 are the province in which the respondent resides, the highest level of education attained, and their gender. These all make sense in reality and they reaffirm some intuition:

- Certain provinces have specific facts of living such as common industries, laws, and recreational activities which may influence voter opinions on parties that are for or against policies that govern them. For example, Alberta's oil reserves are among the largest in the world, so the topic is likely to be divisive in that province. The same can be said about the topic of fishing in the Atlantic provinces or farming in Saskatchewan.
- People who are highly educated may be more willing to support Liberal policies on education, which tend to give more to students who are in need of financial aid. Those who haven't reached the same levels of postsecondary education may hold the opinion that taxpayers shouldn't foot the bill, viewing programs such as Ontario's OSAP as extravagant unnecessities.

Some variables that are notably insignificant are `income` in the Liberal model and `bornin_canada` in both models. The interpretations of these may be that, among those who voted for the Liberal Party, income didn't significantly impact their decision, and likewise for whether or not they were born in Canada.

By poststratifying the models across the defined cells in order to predict the log-odd voting outcomes, the following values were computed:

- The Conservative Party would earn 0.312087467270546 of the votes
- The Liberal Party would earn 0.340578739518302 of the votes

# Discussion

## Summary & Conclusions

The results of this report show that the Liberal Party would be predicted to be the winners if everybody voted in the 2019 Canadian Federal Election, with 2.85% more votes. In reality, the Liberals won a minority government with 157 seats and 33.12% of the popular vote compared to the Conservative Party's 121 seats and 34.34% of the popular vote. By adjusting the weights within each cell across a number of predictors, the models were more representative of the Canadian population.

It was found that a person's province of residence, highest level of education attained, and their gender is very important in determining who they would vote for. Income and age are less important for those who vote for Liberals than Conservatives, and other factors, such as whether they were born in Canada and the number of people in a household, are not statistically significant.

With these facts in mind, there is evidence that the decreased voter turnout may have negatively affected the Liberal Party's chance at achieving a majority government. Under the two fitted models, the majority vote swung in favour of the Liberals.

## Weaknesses & Next Steps

# References

Stephenson, Laura B, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2020. "2019 Canadian Election Study." *Harvard Dataverse.* https://doi.org/10.7910/DVN/DUS88V.