Assuming 100% Voter Participation, Trudeau Would Win 48% of the Popular Vote in the 2019 Canadian Federal Election*

Xinyi Zhang

16 December 2020

Abstract

In the 2019 Canadian Federal Election, incumbent Prime Minister Justin Trudeau claimed a narrow victory over the Conversative Party amid 62% voter turnout. In this paper, we develop a multi-level regression model with post-stratification by training a multinomial logistic regression model using voter survey results (from the 2019 Canadian Election Study) and predicting the outcome of the popular vote using large-scale demographic data for the Canadian population (from the 2017 General Social Survey). Our model predicts that assuming 100% voter turnout, Trudeau would have won the popular vote in the 2019 Canadian Federal Election instead of his opponent Andrew Scheer. Our prediction speaks to the importance for each party to mobilize its voter base and our breakdown of votes by demographic groups provides political parties with information on how to effectively target voters during future elections.

Keywords: forecasting, 2019 Canadian Federal Election, multilevel regression with post-stratification, voter turnout

1 Introduction

In 2019, Trudeau defied election predictions and narrowly won enough seats to win reelection. This raises the question, what would the popular vote have been if everyone voted? By choosing 5 explanatory demographic variables closely associated with voting and political affiliation (age, gender, state, race, and education), we develop a multi-level regression model with post-stratification in order to predict the outcome of the popular vote.

In our analysis, we modeled the relationship between our selected demographic variables and a person's likelihood to vote for either Trudeau or Scheer. We analyzed the significance of our model, and the importance of using multilevel regression with post-stratification because the training data is not proportional to the Canadian population. We found that we were able to make predictions with approximately 45% accuracy and the strongest indicators of 2019 voting were education and gender.

This paper discusses the 2 datasets we used, how they were collected and key highlights of these datasets, followed by visualizations of the data. Next we explained the construction of our model and the positives and negatives of extrapolating information from a smaller voter survey to the Canadian population using a post-stratification dataset. Finally, we present our results and discuss how our results should inform future political strategies.

2 Data

To train our model to predict voting on an individual level for the 2019 Canadian Federal Election, we used data from the 2019 Canadian Election Study (CES). To make predictions on the outcome of the 2019

[*] Code and data supporting this analysis are available at: https://github.com/cindyzhang99/forecasting_canadian_election.

Canadian Federal Election through post-stratification, we used data from the 2017 General Social Survey (GSS).

In the following subsections (Individual-level Survey Dataset, Post-stratification Dataset), we will discuss how each dataset was collected and highlight their key features. Then, in the Data Visualization subsection, we'll graph the distribution of our variables of interest. We will use this data in the multilevel regression with post-stratification (MRP) technique that we will describe in the Model section.

2.1 Individual-level Survey Dataset

From September 13, 2019 to October 21, 2019, the Consortium on Electoral Democracy conducted online surveys on political views and voting intent prior to the 2019 Canadian Federal Election. Their target population was all Canadian citizens and permanent residents, aged 18 or older.

The online platform Qualtrics was used to conduct sampling. Qualtrics used several panels as the sampling frame for this survey. Sampling frames are lists of the individuals that will be selected for the survey sample, meaning that the members of panels aggregated by Qualtrics form a list of a subset of the target population. Then, a sample was selected from the frame using a purposive sampling method. This is a non-probability sampling method where the researcher decides which samples are most representative of the target population. According to the CES Codebook, demographic targets for province, gender (50% male and 50% female) and age (28% aged 18-34, 33% aged 35-54, and 39% aged 55 or older) were set. More specific information about the sampling method was not provided.

Of the 74,548 individuals contacted to take the survey, about 26% did not complete the questionnaire. Another 13% exceeded demographic quotas. Lastly, approximately 10% of responses were removed for speeding (spending less than 500 seconds completing the survey) or for "straight-lining" answers (selecting the same response for all questions), resulting in a final sample size of 37,822 respondents. To ensure results were representative of the Canadian population, survey responses were weighted using data from the 2016 Census. This ensures that the discrepancy between the target population and survey responses is minimized.

Unfortunately, the non-probability sampling method employed by Qualtrics is a major weakness of the CES surveying methodology. Although the non-response rate is relatively low, this is clearly because the sampling frame consists of individuals who on survey panels who regularly take online questionnaires. Even so, several key features of the CES methodology should be highlighted as strengths. First off, the specificity of answer choices for many questions was very detailed. In fact, when matching CES variable levels to GSS variable levels, we often had to combine levels in the CES data because the GSS data was not as specific. Additionally, the CES recontacted some participants after the election to participate in a post-election survey. Although we will not be using the post-election data, there are a multitude of applications where a longitudinal survey about the Canadian Federal Election provides valuable insights. Lastly, as we briefly touched upon when discussing the non-response rate, the CES survey results met stringent data processing standards prior to being publicly released. Therefore, we can trust that the responses represent Canadian voter sentiments to a certain extent when working with the data.

Although the non-probability sampling method is a major weakness of the CES methodology, we will ensure that our prediction of the election popular vote will be representative of the Canadian population by conducting post-stratification using data from the GSS. Therefore, we can still make valid conclusions using the CES dataset even though the responses were drawn through non-probability sampling.

2.2 Post-stratification Dataset

From February 1, 2017 to November 30, 2017, Statistics Canada gathered data on the Canadian family unit by conducting voluntary telephone interviews for Cycle 31 of the General Social Survey. Their target population was all non-institutionalized individuals living in Canada, aged 15 or older.

Cross-sectional sampling was conducted in a two-stage design. The stratified simple random sampling method was used in the first stage. Here, the sampling frame consisted of telephone numbers from the

Census grouped as households using data from Statistic Canada's dwelling frame. Strata were formed at the census metropolitan area (CMA) level and at the province level (i.e., large CMAs formed their own stratum, smaller CMAs were grouped together, and the non-CMA regions of each province were grouped together), forming a total of 27 non-overlapping strata. Finally, households were sampled randomly from each stratum such that the number sampled units from each stratum corresponded to the population sizes of each stratum. To reiterate, the sampled population for this first stage was the chosen households from each stratum. The stratified simple random sampling method was also used in the second stage. Here, the sampling frame was a list of household members, aged 15 and older, from the households selected in the first stage. Then, one individual was randomly selected from each household, forming the sampled population. Approximately 43,000 individuals were contacted to participate in the survey.

Statistics Canada reported that the non-response rate was 52.4%. This presents problems for data analysis based on survey data if respondents differ significantly from non-respondents. To reduce the effects of non-response bias, survey estimates were adjusted based on the demographic characteristics of households that were non-responsive (by pulling their information from the 2016 Census). Another source of non-sampling error is imperfect coverage. For example, households without telephones are excluded from the sampling frame. Again, survey estimates were adjusted by weighing responses to represent all individuals in the target population. Lastly, another weakness of the survey methodology is the exclusion of the Canadian population residing in the Northwest Territories, Nunavut, and the Yukon Territory. As we found when trying to match variable levels in the CES dataset to the GSS dataset, we had to drop responses from the territories in the survey data because of the lack of information available for the territories in the post-stratification data. Due to this limitation in the GSS data coverage, our prediction of the popular vote of the 2019 Canadian Federal Election excludes the voters in the territories of Canada.

On the other hand, several key features of the GSS stand out as strengths of their surveying methodology. A major strength of the questionnaire is that it contains focused questions that comprehensively and extensively capture the subject of interest (the Canadian family). Extensive research and testing was conducted when designing the questionnaire. Upon reading through the questionnaire made available by Statistics Canada, the wording of each question is precise and clear, leaving little room for ambiguity. Additionally, another strength of the survey is that a vast majority of questions were objective (dates, events, counts) removing potential response biases that occur with subjective questions.

Overall, the GSS surveying method using two-stage simple random stratified sampling is effective in generating a sample that is geographically representative of the population living in the Canadian provinces. Despite limitations in sampling coverage, the demographic information available is much more detailed than in comparable post-stratification datasets (e.g., the 2016 Census).

2.3 Data Visualization

The full 2019 CES Online Survey and 2017 GSS datasets contain tens of thousands of observations for over 400 variables.

In the interest of space, we will only discuss the variables in the two datasets that are relevant to our multinomial logistic regression model. We chose five demographic variables to serve as a independent factors that determine vote choice. These variables were present in both datasets and typically associated with political affiliation. They are:

- age (represented by the variables named `cps19_yob` in the CES dataset and `agec` in the GSS dataset): the age of the respondent at the time of the survey
- sex (`cps19_gender` in CES and `sex` in GSS): sex of the respondent
- province (`cps19_province` in CES and `prv` in GSS): province the respondent resides in
- education (`cps19_education` in CES and `ehg3_01b` in GSS): highest education level of the respondent
- religion importance (`cps19_rel_imp` in CES `rlr_110` in GSS): how important religion is to the respondent

To prepare the datasets, we selected responses from individuals who were 18 years of age or older. Additionally, we mapped the integer values of the `age` variable to 6 bins ("18 to 29", "30 to 41", "42 to 53", "54 to 65", "66 to 77", and "78 and above"). Comparing the two distributions, notice that individuals aged 78 and above are underrepresented in the CES dataset (Figure 1). This is likely because Canadians that are much older rarely serve as participants on online survey panels. This speaks to the importance of using post-stratification to adjust our predictions to be more representative of the Canadian population.

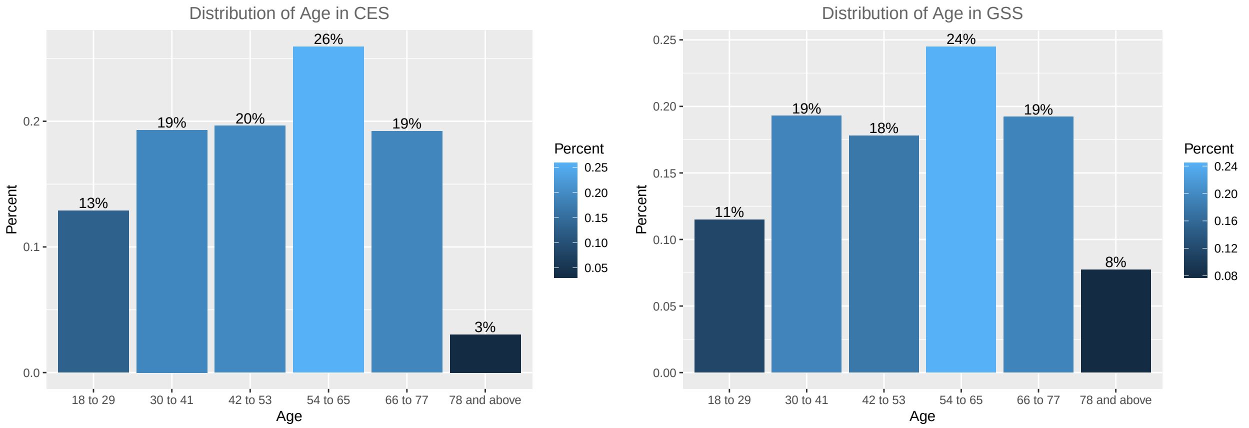

Figure 1: Distribution of Age in CES and GSS datasets.

In the CES dataset, we dropped responses of "Other" for the variable `sex` because the GSS dataset had only "Male" and "Female" levels. Comparing the distributions of sex in each survey, females are clearly overrepresented in the CES dataset (Figure 2).

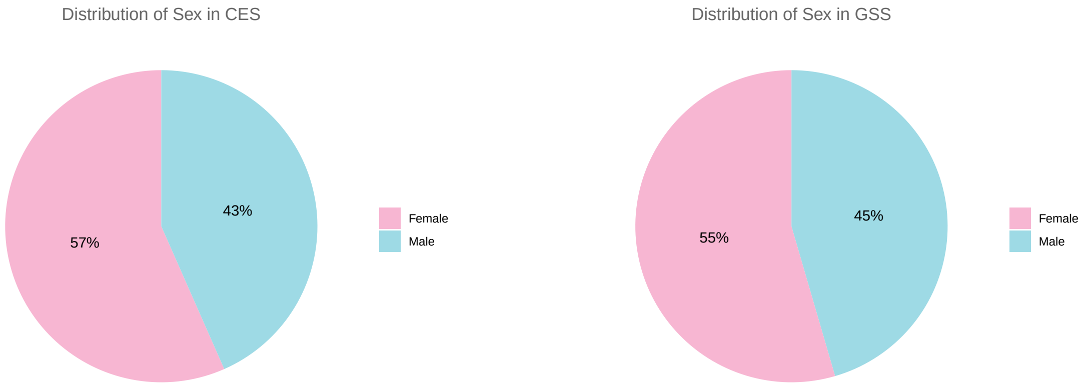

Figure 2: Distribution of Sex in CES and GSS datasets.

In the CES dataset, we dropped responses of "Northwest Territories", "Nunavut", and "Yukon" for the variable `province` because the GSS dataset lacked levels for the territories. Comparing the distributions of provinces in each survey, all of the other provinces are underrepresented while Ontario is overrepresented (by 50%) (Figure 3). Clearly, this has significant implications for our prediction of the popular vote because if we were just to use the results of the CES dataset, the opinions of Canadians residing in Ontario would be significantly overrepresented. Luckily, we are conducting post-stratification using the GSS dataset, which better represents the demographics of the Canadian population.

We simplified the levels for the variable `education` in both datasets, highlighting only the most important categories ("High school or less", "Some college or trade school", "Undergraduate degree", "Some graduate school or more"). Individuals who pursued any level of post-secondary education are overrepresented and

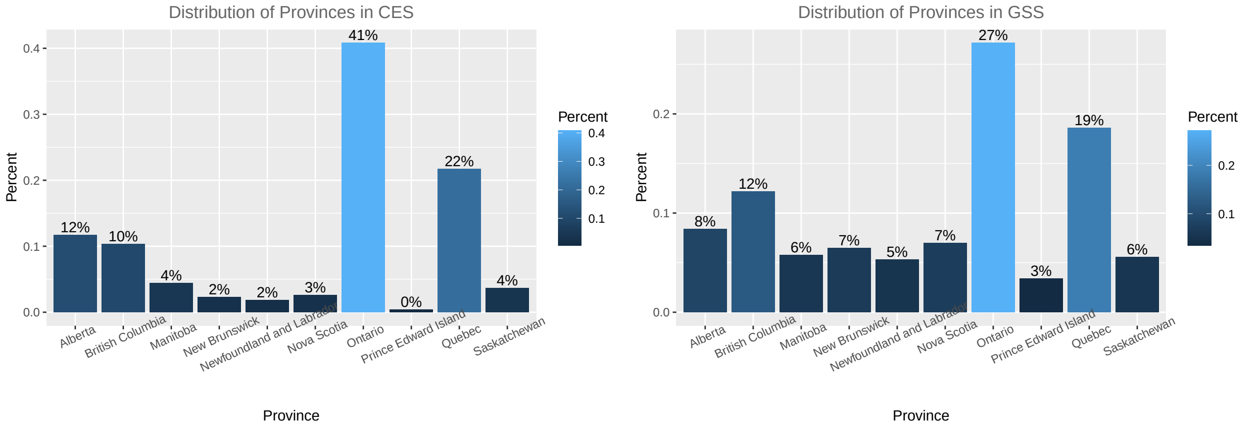

Figure 3: Distribution of Provinces in CES and GSS datasets.

individuals who have at most a high school diploma are underrepresented in the CES dataset (Figure 4). Again, this has significant implications for our prediction of the popular vote if we were to use only the CES dataset but our post-stratification technique using the GSS dataset will mitigate this discrepancy between survey respondents and the demographics of the Canadian population.

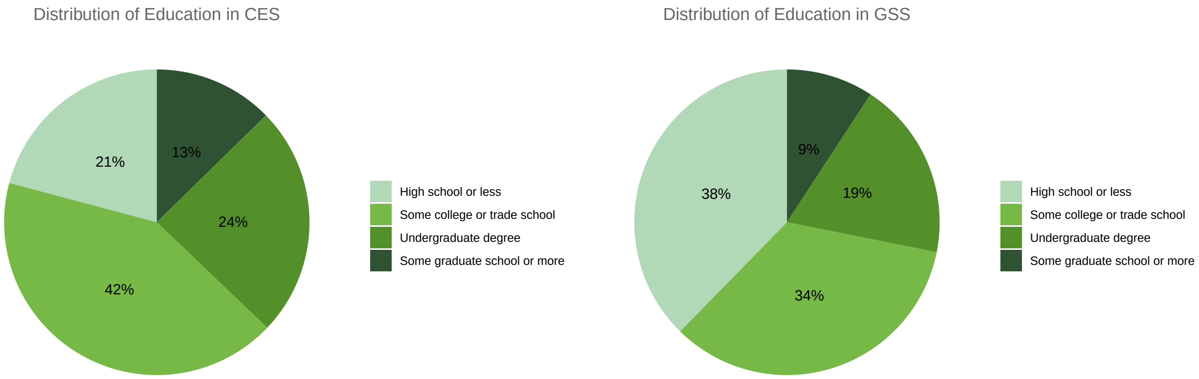

Figure 4: Distribution of Education in CES and GSS datasets.

Luckily, the levels for religion importance perfectly overlapped in both datasets. However, the distribution of each level in both datasets indicates that individuals with strongly held religious opinions (religion is “Very important” or “Not at all important”) are underrepresented while individuals with middle of the road opinions (“Somewhat important”, “Not very important”) are overrepresented in the CES dataset (Figure 5). Once again, we will be using the GSS dataset for post-stratification to ensure our prediction is more representative of the Canadian population than the demographics of the respondents to the CES is.

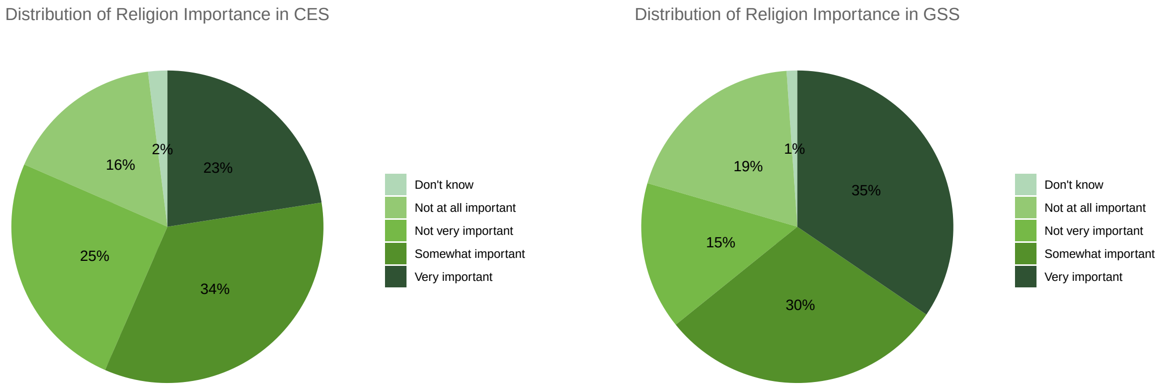

Figure 5: Distribution of Religion Importance in CES and GSS datasets.

3 Model

The purpose of our model is to predict the popular vote of the 2019 Canadian Federal Election assuming 100% voter turnout. We found when exploring and graphing the data in the in the Data Visualization section, the CES has underrepresented or overrepresented certain demographics compared to their actual distributions in the population. Namely, older voters are underrepresented, women are overrepresented, voters from Ontario are overrepresented, voters with some form of post-secondary education are overrepresented, and so forth. Therefore, if we were to predict the outcome of the popular vote based on the demographic proportions found in the CES dataset, we would find that the prediction is biased because the survey sample is not representative of the Canadian population. Hence, we must use multilevel regression with post-stratification (MRP) to base our estimate of the popular vote on demographic distributions are are closer to the actual population distributions. This method can be broken down into two parts: multinomial regression modeling and post-stratification.

First, we want to train a model on the CES dataset to predict vote choice. Our dependent variable `vote` is a categorical variable with six possible values ("New Democratic Party", "Green Party", "Liberal Party", "Conservative Party", "People's Party", and "Bloc Québécois"), which describes how a respondent would vote in the 2019 Canadian Federal Election. The independent variables in our model are the demographic characteristics of the respondent (age, gender, province, education, and religion importance). (As previously established, these were chosen based on their association with determining political affiliation and their persistence in both datasets).

Since we are representing the vote as a classification problem (with more than 2 possible values), multinomial logistic regression is the most suitable choice to model the relationship between vote choice and our independent demographic variables. Given an input of values for the demographic variables, the model will output six probabilities, one for each possible vote choice. The vote choice corresponding to the largest probability is our model's prediction for the given demographic values.

We fully derive the mathematical representation of the multinomial logistic regression model in the Appendix. In the interest of space, here we will just provide the equations we derived for the probability of each outcome:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta_1}\cdot\boldsymbol{X_i}}}{1 + \sum_{k\in\{1,2,3,5,6\}} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta_2}\cdot\boldsymbol{X_i}}}{1 + \sum_{k\in\{1,2,3,5,6\}} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 3) = \frac{e^{\boldsymbol{\beta_3}\cdot\boldsymbol{X_i}}}{1 + \sum_{k\in\{1,2,3,5,6\}} e^{\boldsymbol{\beta_k}\cdot\boldsymbol{X_i}}}$$

$$\Pr(Y_i = 4) = \frac{1}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 5) = \frac{e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 6) = \frac{e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

We will explain each variable that appears in the equations above (again, please see the Appendix for a more natural definition of each variable as we derive these equations). Y_i represents the outcome of the response variable for the ith observation. Since we have a total of 6 possible outcomes ("New Democratic Party", "Green Party", "Liberal Party", "Conservative Party", "People's Party", and "Bloc Québécois"), we represent each of them as 1, 2, 3, 4, 5, 6 in these equations in the interest of space.

$\boldsymbol{\beta_k}$ is the row vector with elements that are the coefficients for the explanatory variables for the kth outcome. More explicitly, $\boldsymbol{\beta_k} = [\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}]$ where $\beta_{0,k}$ is the intercept for the kth outcome, $\beta_{1,k}$ is the coefficient for the first explanatory variable (age) for the kth outcome, $\beta_{2,k}$ is the coefficient for the second explanatory variable (sex) for the kth outcome, $\beta_{3,k}$ is the coefficient for the third explanatory variable (province) for the kth outcome, $\beta_{4,k}$ is the coefficient for the fourth explanatory variable (education) for the kth outcome, and $\beta_{5,k}$ is the coefficient for the fifth explanatory variable (religion importance) for the kth outcome.

$\boldsymbol{x_i}$ is the row vector of values for the explanatory variables for the ith observation. Specifically, $\boldsymbol{x_i} = [1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}]$ where $x_{1,i}$ is the value of the first explanatory variable (age) for the ith observation, $x_{2,i}$ is the value of the second explanatory variable (sex) for the ith observation, and so on.

Therefore, the first equation represents the probability that the vote choice of the ith observation is "New Democratic Party" given values for age, sex, province, education, and religion importance The other 5 equations can be interpreted in a similar manner.

We implemented the multinomial logistic regression model using `multinom()` from the nnet library written in R (R Core Team (2020)). At first the model failed to converge because the maximum number of iterations would be reached before convergence. Consequently, we increased the maximum number of iterations from the default value of 100 to 200 and we observed the model converging after approximately 120 iterations.

Finally, we used post-stratification to accurately predict the popular vote because of the discrepancy between the demographic distribution of the CES respondents and the demographics of the Canadian population. We conducted post-stratification by calculating the counts for all combinations of explanatory variables in the GSS.

4 Results

In Figure 6, we see that the distribution of votes in the CES is in favor of the Conservative party, followed closely by the Liberal Party. However, as we've noticed when observing the distribution of our chosen demographic factors, the CES dataset does not appear to be representative of the Canadian population. Therefore, it is imperative that we use post-stratification when predicting the popular vote of the 2019 Canadian Federal Election to mitigate the effects of sampling bias.

In Figure 7, we see that the distribution of votes in post-stratification is led by the Liberal Party with 48% of the vote followed by Conservative Party with 44% of the vote.

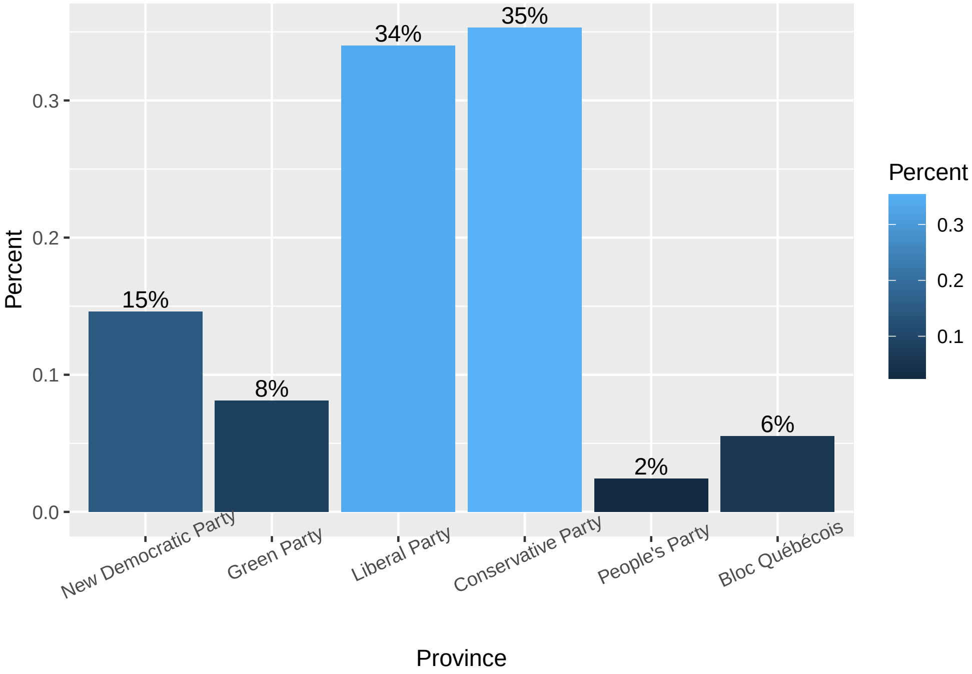

Figure 6: Distribution of Votes in CES.

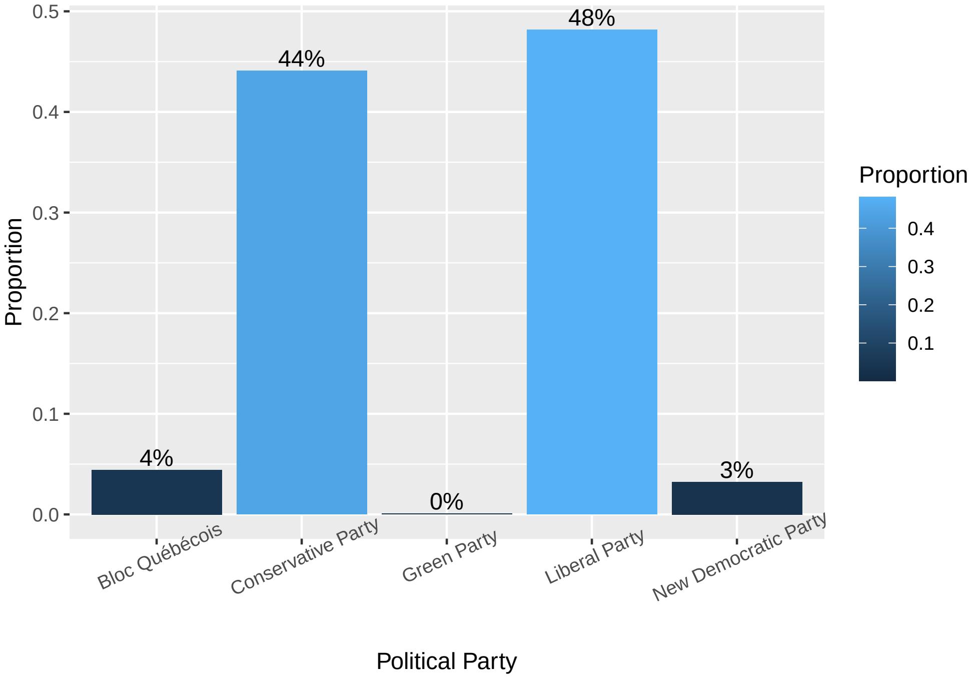

Figure 7: Distribution of Votes in Post-Stratification.

5 Discussion

5.1 Weaknesses

5.2 Next Steps

6 Appendix

6.1 Post-stratification dataset collection cost

Statistics Canada did not disclose the true cost of conducting the survey but we can make some speculations based on the available information about their field work methodology. Surveying was conducted using Computer Assisted Telephone Interviewing (CATI) wherein interviewers read aloud the computerized questionnaire and immediately record the respondent's answers. Although this allows for a reduction in costs compared to traditional in-person surveying, labor costs still include time spent computerizing the survey, training interviewers, and having interviewers administer the questionnaire. Other labor costs include designing the questionnaire and surveying methodology as well as conducting quality control (data consistency was checked by the CATI system during surveying and unresolved inconsistencies were handled afterwards by support staff). Non-labor costs likely included paying for equipment, phone service, offices, and so forth. Again, although we don't have exact costs, we can conclude that the time and costs associated with conducting the GSS is a clear reason why it is only administered once a year.

6.2 Derivation of Multinomial Logistic Regression Model

The multinomial logistic regression model consists of several binary logistic regression models. Like binary logistic regression, the multinomial logistic regression predicts the probability that the ith observation has outcome k using the following function:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + ... + \beta_{M,k}x_{M,i}$$

where $\beta_{m,k}$ is the coefficient for the mth explanatory variable and the kth outcome while $x_{m,i}$ is the value of the mth explanatory variable for the ith observation. In our case, we have $M = 5$ (age, sex, province, education, and religion importance) explanatory variables so the function as applicable to our model is:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \beta_{3,k}x_{3,i} + \beta_{4,k}x_{4,i} + \beta_{5,k}x_{5,i}$$

Note that we can represent $\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}$ and $1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}$ as row vectors $\boldsymbol{\beta_k}$ and $\boldsymbol{x_i}$, respectively. Then, the function can be simplified as follows:

$$f(k, i) = \boldsymbol{\beta_k} \cdot \boldsymbol{x_i}$$

where we take the dot product of the two row vectors we just defined.

As previously mentioned, the multinomial logistic regression model is a series of binary logistic regressions where the probability of each outcome of the response variable (vote choice for the 2019 Canadian Federal Election) is regressed against a chosen pivot outcome. Let Y_i represent the outcome of the response variable for the ith observation. We have a total of 5 possible outcomes ("New Democratic Party", "Green Party", "Liberal Party", "Conservative Party", "People's Party", and "Bloc Québécois"), represented as 1, 2, 3, 4, 5, and 6 respectively. Let's choose the "Conservative Party" (or 4) as the pivot. In mathematical notation, this is:

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = 4)} = \boldsymbol{\beta_1} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = 4)} = \boldsymbol{\beta_2} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 3)}{\Pr(Y_i = 4)} = \boldsymbol{\beta_3} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 5)}{\Pr(Y_i = 4)} = \boldsymbol{\beta_5} \cdot \boldsymbol{X_i}$$

$$\ln \frac{\Pr(Y_i = 6)}{\Pr(Y_i = 4)} = \boldsymbol{\beta_6} \cdot \boldsymbol{X_i}$$

Then, we solve for the probabilities by exponentiating both sides:

$$\Pr(Y_i = 1) = \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 2) = \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 3) = \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 5) = \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}}$$

$$\Pr(Y_i = 6) = \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}$$

The probability of the pivot outcome can be calculated because we know that the probability of all outcomes must sum to 1:

$$\Pr(Y_i = 4) = 1 - \big(\Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}\big)$$

$$1 = \Pr(Y_i = 4) + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}} + \Pr(Y_i = 4) \cdot e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}$$

$$1 = \Pr(Y_i = 4) \left(1 + e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}} + e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}} + e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}} + e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}} + + e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}\right)$$

$$1 = \Pr(Y_i = 4) \left(1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}\right)$$

$$\Pr(Y_i = 4) = \frac{1}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

Having the expression for $\Pr(Y_i = 4)$, we can represent the probabilities of the other outcomes as follows:

$$\Pr(Y_i = 1) = \frac{e^{\boldsymbol{\beta_1} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 2) = \frac{e^{\boldsymbol{\beta_2} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 3) = \frac{e^{\boldsymbol{\beta_3} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 4) = \frac{1}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 5) = \frac{e^{\boldsymbol{\beta_5} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

$$\Pr(Y_i = 6) = \frac{e^{\boldsymbol{\beta_6} \cdot \boldsymbol{X_i}}}{1 + \sum_{k \in \{1,2,3,5,6\}} e^{\boldsymbol{\beta_k} \cdot \boldsymbol{X_i}}}$$

References

"Factors Associated with Voting." 2015. *Statistics Canada: Canada's National Statistical Agency / Statistique Canada : Organisme Statistique National Du Canada*. https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.

Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://github.com/rstudio/bookdown.