Analysing Possible Results of the 2019 Canadian Federal Election if Everyone had Voted Based on the 2019 CES and 2017 GSS Datasets

Hongbo Zhou (1004832862)

Dec 21, 2020

Abstract

It is of great importance that citizens practice their voting privileges in federal elections. In this study, an estimation is obtained on the popular vote of the 2019 Canadian Federal Election in the case that everyone had voted to study the significance of turnout. I performed multilevel regression and poststratification using the *2019 Canadian Election Study - Online Collection* and the *2017 GSS* datasets. Poststratification is implemented on a logistics regression model with 5 predictor variables, leading to the result of an estimate of 35.12% of the popular votes going to the Liberal Party.

Keywords: Multilevel Regression with Poststratification, Logistic Regression, 2019 Canadian Federal Election, Popular Votes

Code and data supporting this analysis is available at: https://github.com/EleanorZhou/STA304-FinalProject

Introduction

Statistical analysis is widely used to make predictions nowadays, and one of its most popular usage in this area is to predict election results. Data collection of voters in the Canadian population regarding their demographics and political views is an important way of monitoring the social-economical changes and trends of Canadians and their current interests on political issues. Thus, it is quite essential to be able to predict the overall election results at a population level.

One popular way to make population level predictions is through the multilevel regression and poststratification (MRP) technique. Firstly introduced in 1997 (Gelman and T. Little) and subsequently expanded (Park, Gelman, and Bafumi), it has become broadly used in recent years, with one famous example of predicting the U.S. presidential election results using Xbox data (Wang et. al.). In this report, MRP will be used to see if every voter had voted, what could the 2019 Canadian Federal Election result be.

Two datasets will be used to apply MRP on predicting the 2019 Canadian Federal Election result, namely the *2019 Canadian Election Study - Online Collection* and the *2017 GSS* datasets. In the Methodology section, the data and the model will be described and specified. Results on the MRP analysis will be included in the Results section. Lastly, the conclusions, weakness and next steps will be provided in the Discussion section.

Methodology

Data

The datasets used in the study are: the *2019 Canadian Election Study - Online Collection* dataset and the *2017 GSS* dataset. More specifically, the *2019 Canadian Election Study - Online Collection* dataset is used for building a logistic regression model, and the *2017 GSS* dataset is used for poststratification purposes.

The *2019 Canadian Election Study - Online Collection* was conducted using online surveys during both the campaign period (CPS) and the post-election period (PES). Its population is the set of all adult Canadian citizens and permanent residents. The sampling frame of the study is the online members of the Canadian general population. 37822 samples were drawn for the CPS and 10337 of the respondents from the CPS were further involved in the PES as samples (Stephenson, Harell, Rubenson and Loewen). CPS respondents were found with targets stratified by region and balanced on gender and age within each region, and some were re-contacted after the election for the PES. The CES is thorough about data quality, as incomplete and duplicate responses, speeders, straightliners, and respondents whose postal code didn't match their province have all been removed from the data. However, some weaknesses are still present. For instance, a small portion of the remaining respondents that have been kept in the data are inattentive respondents and initial duplicate respondents, and although their responses might be useful, they pose certain potential problems to later analysis.

The *2017 General Social Survey on Family (GSS)* dataset targeted population of all persons 15 years of age and older in Canada, excluding residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions. From the sampling frame of the lists of telephone numbers in use available to Statistics Canada and the Address Register, data was collected via computer assisted telephone interviews on 20602 samples. Respondents were randomly selected from each household through simple random sample without replacement in each pre-processed stratum (Statistics Canada). There are both strengths and weaknesses present in the data. For example, the study implemented careful imputation and edition on invalid data, however, certain non-response was not permitted, which might cause a loss of information.

There are a total of 620 variables in the *2019 Canadian Election Study - Online Collection* dataset, and for this analysis, 5 are selected as independent variables and 1 as the response variable. Among the 5 independent variables, 4 are categorical and 1 is numerical, which cover different aspects of possible influence on the response. Some variables are imputed for the purpose of conciseness of the study, such as: year of birth is converted into age, and education is combined into two categories and so on.

Table 1:

```
## # A tibble: 6 x 6
##     age sex    province        education        marital_status      vote_liberal
##   <dbl> <chr>  <chr>           <chr>            <chr>                  <dbl+lbl>
## 1    50 Female Ontario         Other degree     Living common-l~ 0
## 2    72 Female Ontario         Bachelor or hig~ Separated        0
## 3    28 Female Ontario         Bachelor or hig~ Never Married    1 [Liberal Par~
## 4    59 Male   British Columb~ Bachelor or hig~ Married          0
## 5    63 Male   Ontario         Bachelor or hig~ Married          0
## 6    29 Female Ontario         Bachelor or hig~ Living common-l~ 1 [Liberal Par~
```

Table 1 gives an overview of the survey data (*2019 Canadian Election Study - Online Collection*) for modeling purposes.

As for the *2017 GSS* dataset, similar variables are selected and edited in order to perform poststratification. After grouping, each demographic cell is counted for convenience to obtaining the final prediction with the model of survey data.

Table 2:

```
## # A tibble: 6 x 6
## # Groups:   age, sex, province, education [6]
##     age sex    province         education    marital_status     n
##   <dbl> <chr>  <chr>            <chr>        <chr>          <int>
## 1    15 Female Manitoba         Other degree Never Married      1
## 2    15 Female Ontario          Other degree Never Married      1
## 3    15 Female Saskatchewan     Other degree Never Married      1
## 4    15 Male   British Columbia Other degree Never Married      1
## 5    15 Male   Ontario          Other degree Never Married      1
## 6    15 Male   Quebec           Other degree Never Married      2
```

Table 2 gives an overview of the census data (*2017 GSS*) for poststratification.

Model

Model Specifics Firstly, a regression model will be constructed using the RStudio software based on the *2019 Canadian Election Study - Online Collection* survey data.

Since the response variable, the proportion of voters who voted for the Liberal Party, is binary, I will use a logistic regression model for modeling. 5 predictor variables are selected based on my assumption of influential factors of voting choice, including 1 numerical variable age and 4 categorical variables sex, province, education and marital status. I use age as a numerical variable since it is difficult to divide by age groups that are homogeneous within each group, education is grouped into bachelor degree or higher and other degree for simplicity and gender is imputed as sex in order to match the census data.

I choose to include age because the time a person was born can result in different ways of thinking. Variables sex, province, education and marital status are also involved because a vote can depend on how one likes the way a candidate interacts with these traits and certain policies proposed by the Party regarding them.

The logistic regression model I am using is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sexMale} + \beta_{3i} x_{provinceI} + \beta_4 x_{educationOther} + \beta_{5j} x_{maritalstatusJ}$$

$\hat{p}$ represents the estimated percentage of voters who voted for the Liberal Party. β_0 is the measure of the model intercept and describes the probability of voting for the Liberal Party when the voter is a divorced female with a Bachelor degree or higher from Alberta of age at 0. β_i, where $1 \leq i \leq 5$, stands for the slope of the model.

When age increases by one unit, voting for the Liberal Party measured in log odds is expected to increase by β_1. The log odds of being a male Trudeau supporter is β_2 times that of female voters. Other degree holders have a β_4 time of log odds of voting for the Liberal Party than of Bachelor or higher degree holders.

The β_{3i}'s represent the comparisons of the log odds supporting the Liberal Party of different provinces, including ON, BC, MB, NL, SK, QC, NS, NB and PE, with Alberta voters. Similar representation goes with the β_{5i}'s, which are the compared multiples of the log odds voting for Trudeau of being separated, never married, married, living common-law and widowed with the divorced. These two β's are abbreviated with subscripts in order to being concise on the expression.

Model convergence and diagnostics are not especially required for this MRP analysis since the model is not Bayesian and the data is not too large. However, the p-values would still be noted as a representation of the performance of the model. A multilevel regression model might also be appropriate for this analysis since there are different individual and group levels in the population. However, multilevel regression models tend to be difficult to converge and it is also hard to determine the choice between random slope or intercept, therefore I still stick with the logistic regression model.

Post-Stratification In order to estimate the proportion of voters who would have voted for the Liberal Party in the population, poststratification would be performed using the *2017 GSS* dataset. Poststratification partitions data into demographic cells and estimate response variables for each cell. Then these cell-level estimates are aggregated up to a population level, leading to a final result. This is done by weighting each cell by its relative proportion to the population, as described by $\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$. In this analysis, cells are split based on age, sex, province, education and marital status. Using the model described in the previous sub-section I will estimate the proportion of voters in each bin. Then I would weigh each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size to calculate the final estimation on the population.

Results

Table 3:

```
## # A tibble: 18 x 5
##    term                              estimate std.error statistic  p.value
##    <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)                       -1.50      0.155      -9.70  2.98e-22
##  2 age                                0.00700   0.00173     4.04  5.29e- 5
##  3 sexMale                           -0.257     0.0495     -5.20  2.01e- 7
##  4 provinceBritish Columbia           0.790     0.104       7.57  3.60e-14
##  5 provinceManitoba                   0.752     0.138       5.47  4.54e- 8
##  6 provinceNew Brunswick              1.13      0.168       6.71  1.96e-11
##  7 provinceNewfoundland and Labrador  1.26      0.182       6.94  4.04e-12
##  8 provinceNova Scotia                1.47      0.149       9.88  5.23e-23
##  9 provinceOntario                    1.20      0.0893     13.4   7.17e-41
## 10 provincePrince Edward Island       1.25      0.333       3.74  1.85e- 4
## 11 provinceQuebec                     1.05      0.101      10.5   1.26e-25
## 12 provinceSaskatchewan              -0.155     0.172      -0.900 3.68e- 1
## 13 educationOther degree             -0.499     0.0487    -10.2   1.21e-24
## 14 marital_statusLiving common-law   -0.0698    0.110      -0.633 5.27e- 1
## 15 marital_statusMarried             -0.0475    0.0919     -0.517 6.05e- 1
## 16 marital_statusNever Married        0.218     0.106       2.06  3.93e- 2
## 17 marital_statusSeparated           -0.130     0.166      -0.782 4.34e- 1
## 18 marital_statusWidowed             -0.0311    0.137      -0.226 8.21e- 1
```

From the Table 3 summary of the logistic regression model, we find that the individual p-values of age, sex, province, education and marital status estimates are all lower than the significance level at 0.05. Therefore, all the predictor variables - age, sex, province, education and marital status - have a statistically significant impact on voting for the Liberal Party. More specifically, for one unit increase in age, we expect an increase of 0.0070 in the log odds of voting for the Liberal Party. In terms of sex, the log odds of voting for the Liberal Party in males is -0.2571 times that of the female population. As for education, the log odds for supporting Justin Trudeau in presidential elections in other degree holders is -0.4988 times that of amongst holders of Bachelor or higher degree.The similar results can be interpreted for the provinces and marital status.

With cells split by age, sex, province, education and marital status, I performed poststratification analysis on the voter proportions who are in support of the Liberal Party modeled by the logistic model discussed above. The estimate is that 35.12% of voters in the population would vote for Justin Trudeau, who represents the Liberal Party.

Discussion

Summary

This analysis aims to estimate the popular votes outcome of the 2019 Canadian Federal Election. A logistic regression model was built based on the *2019 Canadian Election Study - Online Collection* dataset (Stephenson, Harell, Rubenson and Loewen). Then the post-stratification method was applied on the *2017 GSS* dataset (Statistics Canada). It is estimated that the Liberal Party, represented by Justin Trudeau, would collect 35.12% of the votes of the entire voter population.

Conclusions

Based on the poststratification output, it is estimated that 35.12% of the voting population will support the Liberal Party, and 64.88% of the votes will go to the other Parties, including the Conservative Party, the NDP, Bloc Québécois, Green Party, People's Party and other possible Parties, concluding that the Liberal Party would possibly have lost the popular vote in the 2019 Canadian Federal Election.

This study output serves as a reflection on the general voters' opinions and therefore brings broad impacts to multiple parties. For Canadian voters who failed to practice their voting privileges for the 2019 Election, this study might shed light on the importance of popular votes and therefore assist them in their voting decisions possibly on the next election. Moreover, all the competing parties could analyze and respectively modify their publicity strategies in the coming years, aiming to win more votes in each province for the next election.

Weakness

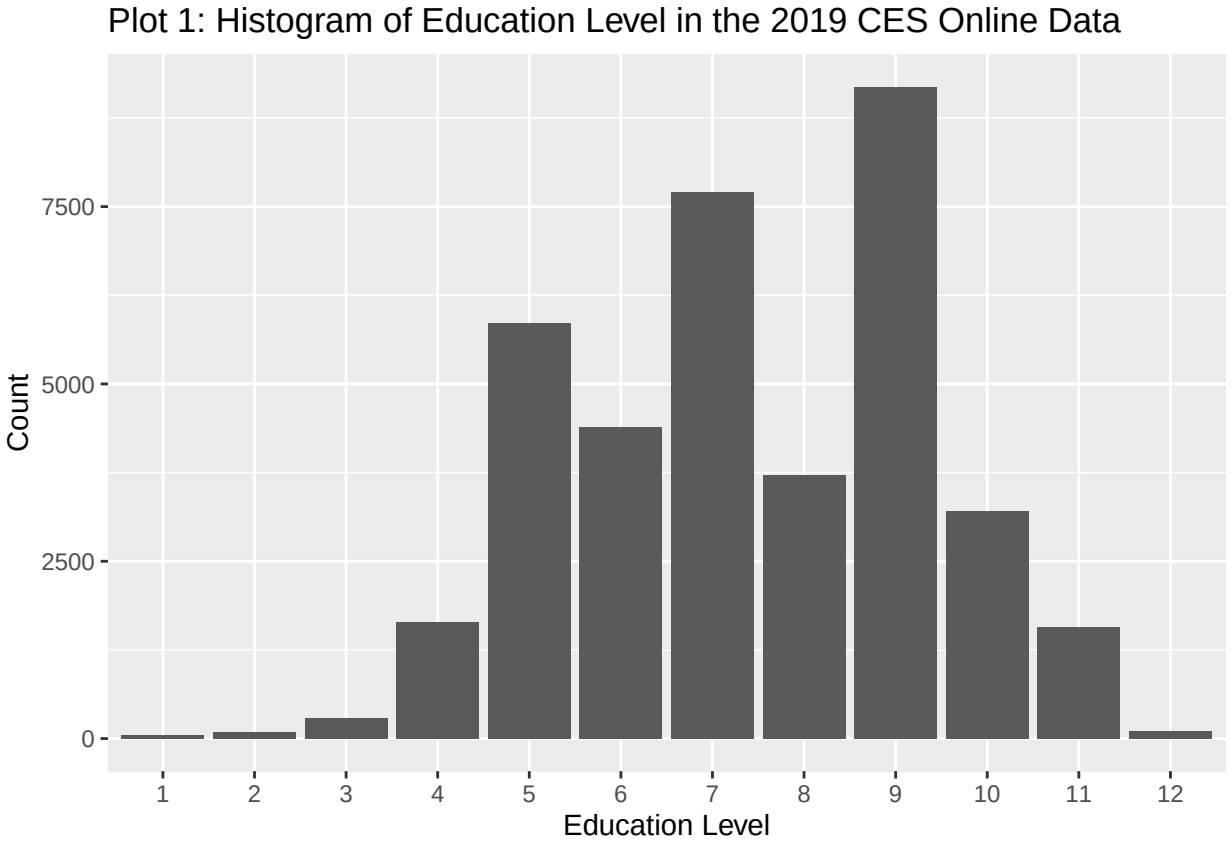
Possible biases are included due to the formatting of the response variable, vote_liberal. Amongst the sample population, I dichotomize voters in support of the Liberal Party as 1's whilst those who are in favor of other parties as 0's. Answers including "I spoiled my vote" and "Don't know / Prefer not to answer" are excluded from this data analysis, which could potentially lead to data misrepresentations. However, these biases are minor and are beyond my reach to correct for, so for this study's purpose I believe it is appropriate to use the binary response variable, although it might pose certain bias.

The datasets themselves also bring conceivable bias to the study. In the *2019 Canadian Election Study - Online Collection* dataset, inattentive data, initial duplicate data, data of who completed the core survey but not the modules are all included in the final dataset according to the removal criteria (Stephenson, Harell, Rubenson and Loewen), leading to 3917 samples of lower data quality in the CPS and 2027 of medium-quality in the PES, which is a considerable amount especially for the PES, with a near 20% of the PES data not being of high quality. As for the *2017 GSS* dataset, using the dataset itself would introduce bias into the study, since the *2017 GSS* dataset is a survey dataset. Although the selection of samples are delicately designed through stratified sampling and SRS, there are still chances that it might fail to represent the entire population as a census. In addition, in the *2017 GSS* dataset, for respondents who answered the survey phone call but were unwilling to answer certain questions, the institution either imputed the value based on data from other sources or left as "NA/Not willing" to participate in the dataset, which would cause information loss and mispresentation of the data.

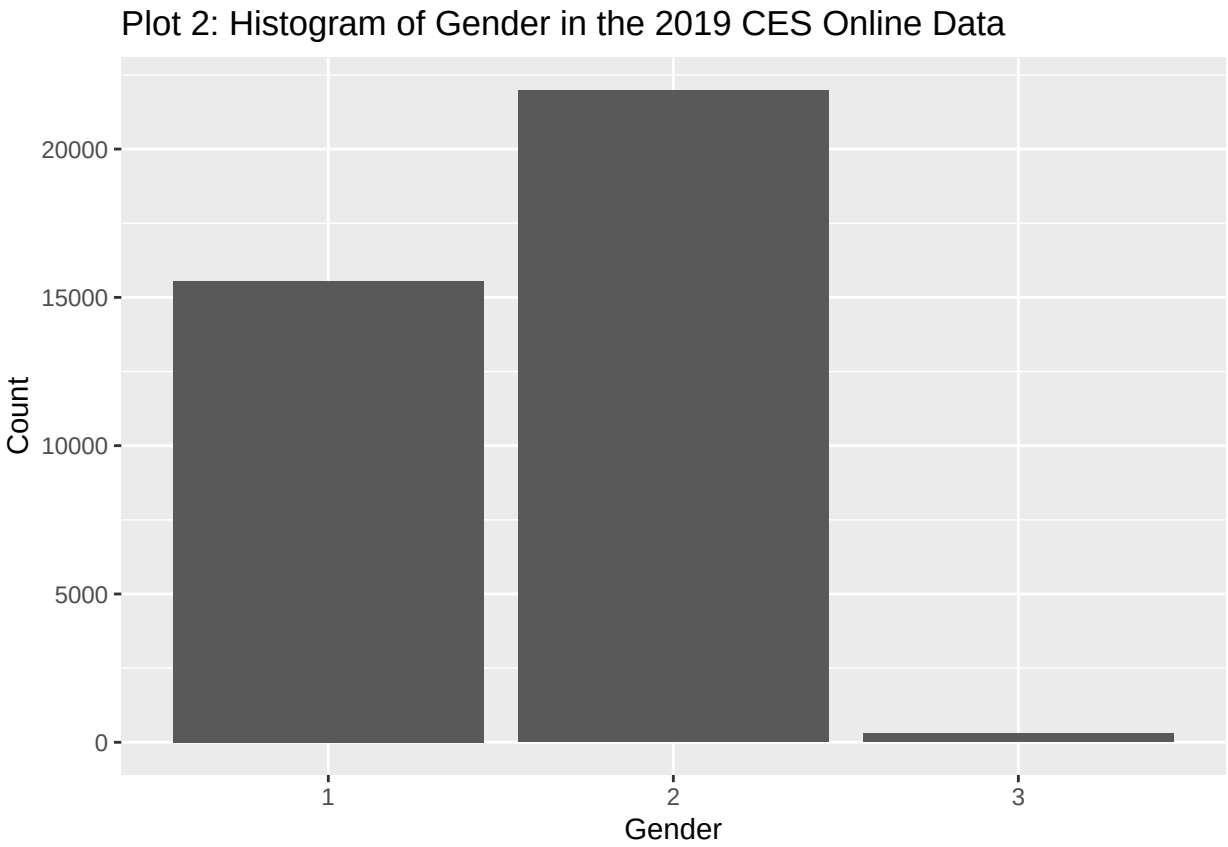
For model simplicity, the initial stages of choosing and imputing variables also pose weaknesses as to the overall model expressivity.

For the education level of the respondents, I made a judgment call to combine all categories into two – Bachelor degree or higher, and other degree. This was due to concerns including the level of difficulty in handling the inputs, as this variable has too many categories, which would make the model difficult to interpret thus decreases its usability. However, in the original *2019 Canadian Election Study - Online Collection* dataset there were 12 different categories and contains a much more variety of information. As it

can be notices from Plot 1, except for factor levels of 9, 10 and 11 which stand for Bachelor degree or higher, a large amount of data is simply categorized as other degree, which decreases the expressivity of the model.



On the other hand, this study came across the mismatch of sex and gender between the survey and census data. Although removing respondents who do not identify as either male or female from the sample does fix the problem, this method suggests that the responses of non-binary individuals are not counted for when making population generalizations and might even be a form of discrimination (Kennedy, Khanna, Simpson and Gelman). However, this is still the most convenient way and as in Plot 2, the number of "Other" responses are relatively minor compared to male and female. Thus, this study would continue with this method aside from its flaws.



Next Steps

The 2019 Canadian federal election popular vote result is already available in details. Then the actual election data output could be compared with the estimated result from this study. The difference in general popularity vote of the candidates could then be further studied. For instance, variables from the PES can be alternatively used to generate another model to compare with the initial one for analyzing which variables are more statistically significant. Additionally, another survey could be conducted for collecting voters' opinions on the 2019 Election result, which would allow me to perform a post-hoc analysis. With these modifications, I would be able to improve the estimation ability of the current model and provide a more accurate output in future election studies.

References

1. 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

2. 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. LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

3. "General Social Survey, Cycle 31 : Families." Statistics Canada, Minister Responsible for Statistics Canada.

4. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. International Journal of Forecasting, 31(3), 980-991. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/forecasting-with-nonrepresentative-polls.pdf

5. "Multilevel regression with poststratification." *Wikipedia*, https://en.wikipedia.org/wiki/Multilevel_regression_with_poststratification#History. Accessed 8th December 2020.

6. Abbreviations and codes for provinces and territories. "2011 Census of Population." Statistics Canada. https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/prov/tbl/tbl8-eng.htm

7. "2019 Canadian federal election." *Wikipedia*, https://en.wikipedia.org/wiki/2019_Canadian_federal_election. Accessed 20th December 2020.

8. Kennedy, L., Khanna, K., Simpson, D., and Gelman A. (2020). Using sex and gender in survey adjustment. https://arxiv.org/abs/2009.14401v1

9. Alexander, R., and Caetano, S. "gss_cleaning.R". 7 October 2020.