Assuming Everyone Voted, Trudeau Would Win 2019 Canadian Federal Election with 34% of Popular Vote*

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 and predicting the outcome of the popular vote using large-scale demographic data for the Canadian population. Our model predicts that assuming 100% voter turnout, Trudeau would have won the 2019 Canadian Federal Election with a larger margin. Our prediction speaks to the importance of encouraging voter turnout and our breakdown of votes by demographic groups provides political parties with information on how to target voters during future elections.

Keywords: forecasting, 2019 Canadian Federal Election, Justin Trudeau, Andrew Scheer, 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 election result 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 70% 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.

-vs riding -actual number of popular vote -general info about election -where dataset is from

*Code and data supporting this analysis are available at: https://github.com/cindyzhang99/sta304_ps5.

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 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, 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 20,602 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 outcome 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).

2.2 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 20,602 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 outcome 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

3 Model

4 Results

5 Discussion

6 Weaknesses

7 Next Steps

8 Appendix

8.1 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.

9 References

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection.

https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/DUS88V/HRZ21G&version=1.0

Rohan, A. & Caetano, S. (2020). 'GSS 2017 Cleaning Code'. MIT License.

https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501&lang=en&db=imdb&adm=8&dis=2

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.