Change in the 2019 Canadian Federal Election Result if Everyone had Voted

Rubing Mai

December 21, 2020

Change in the 2019 Canadian Federal Election Result if Everyone had Voted

Rubing Mai

December 21, 2020

Abstract

The goal of this research is to identify how the 2019 Canadian Federal Election would have been different if "everyone" had voted. 2019 Canadian Election Study(2019 CES) data provided by Harvard Dataverse was used as survey data to establish different multi-level logistic regression model for the Liberal Party and Conservative Party. General Social Survey Cycle 31: Families(GSS 31) provided by CHASS Data Center was used as the census data to do post stratification. Variables selected from both data sets were renamed to "age", "gender" and "province". The results show that the Liberal Party will won the popular vote instead of the Conservative Party, which is different from the real 2019 Canadian Federal Election results.

Keywords

2019 Canadian federal election, Multi-level Logistic Regression Model, Post Stratification

Introduction

In 2019 Canadian Federal Election, liberals which is led by Justin Trudeau won 157 seats to form a minority government. However, liberals got less popular vote than conservatives, 33.1% and 34.4% respectively.(Federal election 2019 live results, n.d.) Elections Canada's report stated that only 67% of eligible Canadian voters cast a ballot in 2019 Federal Election. The top reason why Canadians did not vote in 2019 was being uninterested in politics. The other reasons given by non-voters were too busy, out of town, an illness, disability and so on.(Aiello, 2020) It is obvious that a large proportion of Canadians did not vote in 2019 Federal Election. This suggests that the result would change with full voters' turnout. The goal in this study is to identify how the 2019 Canadian Federal Election would have been different if "everyone" had voted.

Two data sets will be used to complete this analysis. The survey data, 2019 Canadian Election Study(2019 CES) data, is provided by Harvard Dataverse. 2019 CES online survey data contains two parts, Campaign Period Survey(CPS) data and Post-Election Survey(PES) data. The data set, General Social Survey Cycle 31: Families(GSS 31) which is provided by CHASS Data Center will be used as the census data. In order to complete this study, we will add people who have a preferred party but ultimately did not vote or not eligible to vote into our analysis. Their first preferred party will be labeled as their final voting decision.

Their favorite party will be obtained from CPS data. For those who voted during the 2019 Canadian election, their final vote will be obtained from PES data.

Then we are firstly using logistic regression model and then employing a post-stratification technique. Survey data will be used in model section to produce the logistic regression model. A total of two logistic multi-level regression model will be built for the Liberal Party and Conservative Party. Stepwise regression with AIC is a commonly-used method to find a parsimonious model. The census data helps us post stratify. Firstly, cells will be created based off different gender, age, province. These variables are chosen because they are significant, and they provide the model with smaller AIC. Then using the logistic regression models to estimate the proportion of voters in each bin of cells. Next, weight each proportion estimate within each bin of cells by the corresponding population size of that bin, then add these values and divide by the total population size. More details about data, Logistic multi-level regression model and post-stratification calculation will be described in the methodology section. The results of this study will be shown in the results section. More information about summary, conclusion, weakness and next steps will be talk about in discussion section. The external references used in this study will be shown in the references section.

Methodology

Data

The 2019 Canadian Election Study data provided by Harvard Dataverse is used as the survey data for this research. The data was collected from online surveys in two periods, namely the Campaign Period Survey (CPS) and the Post-Election Survey (PES). The target population is Canadian citizens and permanent residents, aged 18 or older. For CPS, it collected 37,822 samples of the Canadian general population through Qualtrics online platform. The targets were stratified by region and with roughly equal distribution of gender and age within each region. 2019 CES data aimed for 50% men and 50% women. Also it aimed to have 28% of respondents aged 18-34, 33% aged 35-54, 39% aged 55 and higher, 80% French, 20% English within Quebec, 10% French within the Atlantic region, and 10% French nationally. The regions involved in the data were Atlantic, Quebec, Ontario Prairies, and British Columbia. For PES, it re-contacted 10,340 respondents from the CPS after the election for a follow-up survey using Qualtrics online platform. Respondents responses from PES were matched to CPS by using respondent's panel IDs. If the respondent did not answer some question or part of the question, then their response was recorded as missing. The general advantages of online surveys are low cost, overall convenience and flexibility. However, due to false or fake responses, the reliability of the data will be reduced. In addition, certain groups of people cannot access the Internet in response to online surveys. Therefore, the 2019 CES data removed those who were not the target population, those who did not complete the core survey, those who had duplicate information with previous respondents, those who answered too quickly, and those whose postal codes did not match the province.

The variables selected in this study are "cps19_gender", "cps19_age", "cps19_province", "cps19_votechoice", "cps19_votechoice_pr", "pes19_votechoice2019".

"cps19_votechoice": The party that the respondent is likely to vote.

"cps19_votechoice_pr": The party that the non-Canadian citizens may vote.

"pes19_votechoice2019": The party that the respondent voted.

Independent Variables:

"cps19_province"(renamed to "province"): Province or region of residence of the respondent.

"cps19_age"(renamed to "age"): Age of respondent at time of the survey.

"cps19_gender"(renamed to "gender"): Gender of respondent.

For this research, we created a new variable named "vote_choice" to combine people who voted in the 2019 Canadian election with people who have a preferred party but ultimately did not vote or were not eligible to vote. For respondent who voted in the election, their choice obtained from "pes19_votechoice2019" were directly put into "vote_choice". For those who did not vote, their preferred party obtained from "cps19_votechoice" and "cps19_votechoice_pr" were labeled as their final voting decision, and then put into "vote_choice". Then we creating two dummy variables "vote_Justin_Trudeau" and "vote_Andrew_Scheer". "vote_Justin_Trudeau" = 1 if the respondent voted for the Liberal Party and 0, if not. "vote_Andrew_Scheer" = 1 if the respondent voted for the Conservative Party and 0, if not. These two variables will be response variables in the logistic regression models described in the model section.

The census data is General Social Survey Cycle 31: Families(GSS 31), provided by CHASS Data Center. The survey was conducted through telephone interviews and used stratified random sampling method. A total of 20,602 samples were collected. The population was divided into strata according to the different provinces they lived in, and then a simple random sampling method (without replacement) was used to randomly select respondents from households. The target population of this survey is all individuals aged 15 and over in Canada excludes people in Yukon, Northwest Territories, and Nunavut. The data frame of the survey was obtained from the lists of available telephone numbers that provided by Statistics Canada and The Address Register. Respondent who refused to be interviewed and who did not get the call were re-contacted. If the respondent was still unable to contact, the weight was adjusted.

Variables chosen from census data should be corresponding to survey data. Here, variables "agedc", "sex", "prv" were selected from census data.

"agedc"(renamed to "age"): Age of respondent at time of the survey interview.

"sex"(renamed to "gender"): Sex of respondent.

"prv"(renamed to "province"): Province of residence of the respondent.

These three variables were renamed to the same name with the variables selected from survey data to do post stratification. Then the categories of each variable were reorganized to the same with survey data. It is worth noting that there are only variable "sex" in census data, and no variables about gender. This means that some people of other genders(e.g. Trans, non-binary, two-spirit, gender-queer) were not involved in the GSS 31 survey. When doing post stratification, in order to unify the categories of each variable, people of other genders were removed from the survey data. This involves some ethical issues. Also, people in Yukon, Northwest Territories, and Nunavut were not in the GSS 31 survey, and so they were removed from survey data.

Model

Model Specifics

Since we want to know the proportion of people vote for the Liberal Party and Conservative Party, the response variable should be binary. Thus, different logistic regression model(frequentist) will be used to model the proportion of voters who will vote for liberals and conservatives. We will use backward AIC to find a parsimonious model. Backward AIC starts with all the potential predictors in the model, then removes the predictor with the largest p-value each time to give a smaller AIC. In addition, all the model process will be done by using RStudio.

Logistic regression model for the Liberal Party:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{province}$$

For the first model, we are interested in how variables, such as age(x_{age}) and province($x_{province}$) influence the the proportion of voters who will vote for the Liberal Party(p). For example, holding other variables constant, for every additional unit increase in age we expect the log odds of voting for the Liberal Party to increase by 0.0028 on average. $x_{province}$ is a categorical variable with 10 levels. So we have 9 dummy variables appeared in our model. Province Alberta is our reference group. So we could say that under other variables constant, we expect the log odds of voting for the Liberal Party of voters in British Columbia to increase by 0.7235 on average. In addition, β_0 is the intercept of the model. Here, it means that under other variables constant, we expect the the log odds of voting for the Liberal Party of voters in Alberta to decrease by 1.8143 on average. β_1 and β_2 are coefficients of predictor age and province respectively.

Logistic regression model for the Conservative Party:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{province}$$

For the second model, we are interested in how variables, such as age(x_{age}) gender(x_{gender}) and province($x_{province}$) influence the the proportion of voters who will vote for the Conservative Party(p). For example, holding other variables constant, for every additional unit increase in age we expect the log odds of voting for the Conservative Party to increase by 0.0124 on average. x_{gender} is a dummy variable. Male is our reference group. So we could say that under other variables constant, we expect the log odds of voting for the Conservative Party of male voters to increase by 0.4395 on average. $x_{province}$ is a categorical variable with 10 levels. So we have 9 dummy variables appeared in our model. Province Alberta is our reference group. We could say that under other variables constant, we expect the log odds of voting for the Conservative Party of voters in British Columbia to decrease by 1.26 on average. β_0 is the intercept of the model. β_1 and β_2 are coefficients of predictor age and province respectively.

Post-Stratification

To estimate the proportion of voters who will vote for the Liberal Party and Conservative Party, I will do a post-stratification analysis. For the Liberal Party, I create cells based off different age and province. For the Conservative Party, I create cells based off different age, gender and province. Different variables were chosen from census data to corresponding to the variables used in different logistic regression models respectively. Then using the logistic regression models for the Liberal Party and Conservative Party in the previous subsection to estimate the proportion of voters in each bin of cells. Next, I will weight each proportion estimate within each bin of cells by the corresponding population size of that bin, then add these values and divide by the total population size. The final results of the proportion of voters who support the Liberal Party and the Conservative Party are 30.65% and 27.86% respectively.

Results

To complete our study about the proportion of voters who will vote for the Liberal Party and Conservative Party, we firstly built two different logistic regression models for the Liberal Party and Conservative Party, and then we usd post stratification technique to get the results. There are summary tables for these two models.

Table 1 – The Liberal Party Model Summary

	Est.	S.E.	z val.	p
(Intercept)	-1.8143	0.0570	-31.8310	0.0000
age	0.0028	0.0008	3.6895	0.0002
provinceBritish Columbia	0.7235	0.0569	12.7217	0.0000
provinceManitoba	0.6404	0.0735	8.7161	0.0000
provinceNew Brunswick	0.9593	0.0892	10.7498	0.0000
provinceNewfoundland and Labrador	1.3075	0.0965	13.5555	0.0000
provinceNova Scotia	1.2363	0.0811	15.2400	0.0000
provinceOntario	1.0702	0.0474	22.5598	0.0000
provincePrince Edward Island	1.0131	0.1795	5.6430	0.0000
provinceQuebec	0.8510	0.0505	16.8443	0.0000
provinceSaskatchewan	-0.1947	0.0960	-2.0283	0.0425

Standard errors: MLE

According to the first table, we could see that under other variables constant, for every additional unit increase in age we expect the log odds of voting for the Liberal Party to increase by 0.0028 on average.

Under other variables constant, we expect the log odds of voting for the Liberal Party of voters in British Columbia to increase by 0.7235 on average. It is equivalent to say that holding other variables constant, the proportion of voters in British Columbia who will vote for the Liberal Party is 25.15% higher on average than voters not in British Columbia. This is calculated by isolating p using formula:

$$p = \frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}}$$

Same for other dummy variables. Under other variables constant, we expect the log odds of voting for the Liberal Party of voters in Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island and Quebec to increase by 0.6404, 0.9593, 1.3075, 1.2363, 1.0702, 1.0131 and 0.8510 on average respectively. Then, the log odds of voting for the Liberal Party of voters in Saskatchewan will decrease by 0.1947 on average with other variables constant.

Table 2 – The Conservative Party Model Summary

	Est.	S.E.	z val.	p
(Intercept)	-0.5788	0.0504	-11.4825	0.0000
age	0.0124	0.0008	15.1327	0.0000
genderMale	0.4395	0.0262	16.7509	0.0000
provinceBritish Columbia	-1.2600	0.0494	-25.4863	0.0000
provinceManitoba	-0.8369	0.0638	-13.1134	0.0000
provinceNew Brunswick	-1.4026	0.0926	-15.1440	0.0000
provinceNewfoundland and Labrador	-1.6075	0.1109	-14.4890	0.0000
provinceNova Scotia	-1.7624	0.0937	-18.8120	0.0000
provinceOntario	-1.2454	0.0382	-32.6053	0.0000
provincePrince Edward Island	-2.0932	0.2442	-8.5721	0.0000
provinceQuebec	-2.0593	0.0468	-43.9918	0.0000
provinceSaskatchewan	-0.2803	0.0672	-4.1736	0.0000

Standard errors: MLE

According to the second table, we could say that holding other variables constant, for every additional unit increase in age we expect the log odds of voting for the Conservative Party to increase by 0.0124 on average. Also under other variables constant, we expect the log odds of voting for the Conservative Party of male voters to increase by 0.4395 on average. It is equivalent to say that holding other variables constant, the proportion of male voters who will vote for the Conservative Party is 46.52% higher on average than female voters. This is calculated by isolating p using formula

$$p = \frac{e^{\beta_0+\beta_2}}{1+e^{\beta_0+\beta_2}}$$

Then under other variables constant, we expect the log odds of voting for the Conservative Party of voters in British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec and Saskatchewan to decrease by 1.26, 0.8369, 1.4026, 1.6075, 1.7624, 1.2454, 2.0932, 2.0593 and 0.2803 on average respectively. Similarly, we could say that holding other variables constant, the proportion of voters in British Columbia who will vote for the Conservative Party is 13.72% higher on average than voters not in British Columbia. This is calculated by isolating p using formula:

$$p = \frac{e^{\beta_0+\beta_3}}{1+e^{\beta_0+\beta_3}}$$

Then, We estimate that the proportion of voters in favor of voting for the Liberal Party or Conservative Party to be 30.65% and 27.86% respectively. This is based off my post-stratification analysis of the proportion of voters in favor of the Liberal Party or Conservative Party modeled by logistic regression models, which accounted for age, province, and possibly gender. There are also 41.49% of population vote for other parties such as Bloc Québécois, New Democrat, Green, Independent, People's Party and so on. From our results, the Liberal Party which is led by Justin Trudeau will win the popular vote if "everyone" had voted. This is different with the real 2019 Canadian election results that the Conservative Party won the popular vote. Thus, the popular vote results in 2019 Canadian Federal Election would have been different if "everyone" had voted.

Discussion

Summary & Conclusions

In 2019 Canadian election, liberals got less popular vote than conservatives, 33.1% and 34.4% respectively. The goal in this study is to identify how the 2019 Canadian Federal Election would have been different if "everyone" had voted. So, in our study, we assume all the individuals in Canada could participate in voting. To do that, we added people who have a preferred party but ultimately did not vote or not eligible to vote into our analysis. Their first preferred party was labeled as their final voting decision. Their favorite party was obtained from CPS data. For those who voted during the 2019 Canadian election, their final vote was obtained from PES data. Then we created a new variable "vote_choice" to keep the voting results. Then we built different multi-level logistic regression model for the Liberal Party and Conservative Party. Finally, we got the overall forecast results using post stratification method.

Two data sets are used to complete the estimation of the proportion of voters who will vote for the Liberal Party and Conservative Party. The survey data, 2019 Canadian Election Study data is provided by Harvard Dataverse. The census data, General Social Survey Cycle 31: Families(GSS 31) is provided by CHASS Data Center.

Based off the estimated proportion of voters in favor of voting for the Liberal Party and Conservative Party being 30.65% and 27.86% respectively, we got the result that the popular vote results in 2019 Canadian Federal Election would have been different if "everyone" had voted. Then the Liberal Party which is led by Justin Trudeau will win the popular vote if "everyone" had voted.

Weakness & Next Steps

The general advantages of online surveys(survey data) are low cost, overall convenience and flexibility. However, due to false or fake responses, the reliability of the data will be reduced. In addition, certain groups of people cannot access the Internet in response to online surveys. There is ethic problem about census data. Only "sex" variable is included in data, no "gender" variable. So census data excludes people from other genders(e.g. Trans, non-binary, two-spirit, gender-queer). Then in post-stratification section, in order to unify the categories of each variable, people of other genders were removed from the survey data. Also, people in Yukon, Northwest Territories, and Nunavut were not in census data, and so they were removed from survey data. This will reduce our sample size and accuracy of results. Finally, there is only small amounts in each bin of cells(age, gender, province). This makes the prediction less accurate.

Since we only built logistic regression models for the Liberal Party and Conservative Party, the sum of the voting ratios of the two parties is not 100%. We may try to build more logistic regression models of other parties to check if there is a 100% turnout. The logistic regression models we obtained from backward AIC have lowest p-value and AIC. All the predictors are significant. But we could check if there is multicollinearity between variables. If there is a problem of multicollinearity, Box-Cox transformation and weight least squares can be considered.

References

1. Aiello, R. 2020, February 26. 'Not interested in politics' top reason Canadians didn't vote in 2019, StatCan says, https://www.ctvnews.ca/politics/not-interested-in-politics-top-reason-canadians-didn-t-vote-in-2019-statcan-says-1.4828540

2. Alexander, R., & Caetano, S. 2020, October 7. Gss_cleaning.R., https://q.utoronto.ca/courses/184060/files/9422740/download?download_frd=1
3. Federal election 2019 live results. n.d.. https://newsinteractives.cbc.ca/elections/federal/2019/results/
4. 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
5. Laine Ruus. 2017. General social survey on Family (cycle 31), 2017, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31
6. Long JA. 2020. *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1.0, https://cran.r-project.org/package=jtools.
7. RStudio Team. 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, http://www.rstudio.com/
8. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John. 2020. 2019 Canadian Election Study - Online Survey, Harvard Dataverse, V1, https://doi.org/10.7910/DVN/DUS88V
9. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, 2019 Canadian Election Study - Online Survey Technical Report and Codebook.pdf, 2019 Canadian Election Study - Online Survey, Harvard Dataverse, V1, https://doi.org/10.7910/DVN/DUS88V/HRZ21G
10. Wickham et al.. 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Appendix

https://github.com/RubingMai/STA304-Final-Project