An Investigation on A Full Voter Turnout in the 2019 Canadian Federal Election

Cindy Gao

December 21, 2020

ABSTRACT

The Canadian Federal Elections are one the most important events that not only catches the attention of millions in Canada, but millions in the entire world. Canada, as a democracy, depends heavily on the opinions and votes of its citizens. However, there has never in history been a full voter turnout, with all turnouts roughly averaging only between 60%-75%. This investigation will achieve the impossible and simulate the results of a full voter turnout in the 2019 Canadian Federal Election between the top two competitors, The Liberal Party of Canada, and The Conservative Party of Canada. To accomplish this simulation, a multilevel logistic regression model with post-stratification will be built using a dataset from the Canadian Election Study (CES) and Statistics Canada. In the end, comparisons from the model results and the official 2019 results will be made to see if there are any differences in the outcome with a full voter turnout.

KEYWORDS

Multilevel Logistic Regression Model, Post-Stratification Estimate (Y^{PS}), First Past the Post, Area Under the Curve, The Liberal Party of Canada, The Conservative Party of Canada

INTRODUCTION

Voting serves as one of the most treasured basic right that allows Canadians to maintain a democratic society. Not only is voting a basic right, but it should be treated as a duty for Canadian citizens, as each citizen has a role in determining the outcome of the Canadian society. However, studies show that in both the 2015 and 2019 Canadian Federal Elections, only approximately 77% of eligible voters voted, therefore meaning roughly a quarter of eligible Canadians did not vote. This suggests that if there was a 100% voter turnout, there may have been a different outcome in the elected leader and/or a different outcome in the distribution of seats in the House of Commons. The goal in this study will be to investigate what would have happened in the 2019 elections if there was a 100% voters' turnout.

For this investigation, a dataset from the Canadian Election Study (CES), a large-scale survey conducted on voters yearly, and a dataset from Statistics Canada, a government-run production that generates statistics to help better understand Canada's economy, population, society and culture will be used. The CES will serve as the survey data, and the dataset from Statistics Canada will be used as the census data. Data from both the CES and Statistics Canada will be used to create a multiple linear regression model with post-stratification. The multiple linear regression model will be created with variables on gender, age, province, education, and voter's choice respectively from each of the two datasets. The multiple linear regression model can display and show which variables have the biggest effects on voter's choice and which have the smallest, as well as if there are any patterns leading to one's final voting choice.

Post-stratification is a great technique to use to combine both the survey data and the census to simulate a 100% voters' turnout. To perform post-stratification, different cells will be created based on different demographics from the census population and then applied to the survey estimate probabilities in each cell to estimates each response variable. Each cell's estimates will then be averaged together by weighing them respectively to their proportions in the census population, which is known as the post-stratification estimate, $\hat{Y}^{PS}$. By using post-stratification, the unattainable 100% voters' turnout can be attained.

METHODOLOGY

Data

For this investigation specifically, the 2019 Canadian Election Study (CES), a large-scale survey conducted on voters yearly, and the 2016 "Highest Level of Educational Attainment (general) by Sex and Selected Age groups" dataset from Statistics Canada, a government-run production that generates statistics to help better understand Canada's economy, population, society and culture, will be used. The 2019 CES will serve as the survey data, and the Statistics Canada dataset will serve as the census data.

To start off, both datasets were first cleaned by filtering out those who are not eligible to vote in the 2019 Canadian Federal Elections. Those who are eligible must satisfy the conditions of being a Canadian citizen over the age of 18. Then, from both the survey and census dataset, four predicator variables representing one's age, gender, education level, and province were chosen respectively to build further build the model. From the census dataset, they were, "Age", "Sex", "Education" and "Geographic.name", and in the survey dataset, they were "cps19_yob", "cps19_gender", "cps19_education" and "cps19_province". As predictor variables, they will represent the betas $(\beta_1, \beta_2 ... \beta_n)$ in the upcoming model for this investigation

These variables were carefully selected with the thought that they all play a role in one's decision. For example, the younger generation may have a different political than the older generation, and female voters opinions and decisions may differ from male voters, as female voters may value a party who implements and fights more for gender equality. These variables were also chosen so that one's personal level (level one), and group level's (level two) information could be obtained to help simulate a more accurate prediction of the final result. Along with that, the variable, "cps19_votechoice", was also selected to represent the response variable, which represents one's choice in the 2019 Canadian Federal Election.

Model

To proceed with this investigation, the Frequentist approach was used, and a multilevel logistic regression model was built. A multilevel logistic regression model is a great fit for this investigation, as not only can it predict a valuable outcome based off of predictor variables, but it can also incorporate one's information on both the personal and group level. Along with that, a logistic level can assess and incorporate the data/predictor variables on both a numerical and categorical level. As stated in the Data section above, the model built for this investigation will include four predictor variables of one's age, sex, education and location (location). It will also include one response variable, which is the variable of interest, one's voting choice. The completed model has the following formula:

$$\log(\frac{Liber\hat{al}Party}{1 - Liber\hat{al}Party}) = 0.66346 + 0.06312 Age35to44 + ... + 0.12052 Age55to64$$

$$+ 0.32221 EducationCollegeLevelDegree + ... - 0.0466 EducationUniversityLevelDegree +$$

$$0.58145 GenderMale - 1.22381 ProvinceBritishColumbia + ... + 14.04089 ProvinceYukon$$

In this model, $\log(\frac{Liber\hat{al}Party}{1-Liber\hat{al}Party})$ is the predicted log odds of the Liberal Party winning. The log odds can later be used and transformed into the predicted probability of the Liberal Party's winning chances. $Liber\hat{al}Party$ represents the predicted probability of the Liberal Party winning the election in a full voter turnout. The estimated intercept, $\hat{\beta_0}$, is 0.66346, which means that while all other predictors in the model do not exist, or are zero, then the predicted log odds of the Liberal Party winning is 0.66346. Next, all predictor variables in this model, age, sex, education and location were divided into respective groups and were treated as dummy variables. The interpretations of these were that a unit's increase in each one of these predictor variables, would increase or decrease the predicted log odds of The Liberal Party's chances of winning by each predictor variables' corresponding estimated coefficients amount. For example, for every additional unit of increase in the Age35to44 predictor group, there will be a 0.06312 increase in the predicted log odds of The Liberal Party's chances of winning. This interpretation will hold and be consistent for all other predictors and their groups in this given model.

First Past the Post

Canada has a system called "First Past the Post", which is where each province has a certain number of seats that represents them in the House of Commons, with a total of 338 seats. Seats are then divided up in respect to each provinces' district's parliament representatives. Then, voters will vote for their own district's leader, and whichever leader receives a plurality of votes in his/her district will represent the district at the federal level and will also add one seat to the party he/her represents in the House of Commons. The party that secures the most seats in the House of Commons will then be able to form government and have their party's chosen leader as the Prime Minister of Canada.

The First Past the Post system was designed to help create a fairer voting outcome, by allocating provinces with more population or influence in the Canadian society with more seats. For example, Ontario has the greatest number of seats with 121 seats, which is approximately 36% of the seats in the House of Commons. This is because Ontario has the greatest amount of population in Canada and is the centre of Canada, and therefore it was allotted more seats to accurately portray its dominance. Another example would be Nunavut and Yukon, two of the lowest populated provinces, with each only having 1 seat in the House of Commons.

First Past the Post is also the reason as to why sometimes there is a majority government and why sometimes there is a minority government. In this investigation, because only two parties are being considered, then in order to form a government, the winning party must form a majority government and reach a minimum of 170 seats (more than half of 338 seats).

Including the First Past the Post system will simulate the real elections and result in a more accurate final prediction. To include First Past the Post in this investigation, first, each individual's voting choice will be predicted from the results of the above multilevel logistic regression model. Then, the predicted probability will find The Liberal Party's chances of winning in respect to each province and how many seats each province has. From thereon, the predicted amount of seats The Liberal Party and The Conservative Party would receive in each province will be predicted, and totalled together in the end to see who will receive the majority of the votes in the House of Commons, and which party will ultimately lead the Canadian society.

Post-Stratification

Post-stratification is another key technique used in this investigation to help simulate a full voter turnout. Not only can post-stratification reduce bias and error in the final prediction, but it can also allows the estimates and statistics found through the survey data to be applied at a constituency level with the census data. To perform post-stratification, cells will be created based demographics, and then the estimates found through the survey data, will be applied to each divided cell, and in the end will be combined by weighing them respectively to their proportions in the census. This value is known as the post-stratification estimate,$\hat{Y}^{PS}$, with its formal equation being:

$$\hat{Y}^{PS} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$$

where $\hat{y_j}$ is the estimate of each cell and where N_j is the population size of each j^{th} cell based off of demographics.

In this investigation, cells were created based off of one's gender, education and location. Location was classified by each province. This is an important factor to focus on because, Canada is very big in area, and thus Canada spreads over a great amount of land. Therefore, people of different places will have different wants and needs when considering which party is able to satisfy their needs and wants. Next, gender and education were also big factors to consider, as research show that gender and one's education level do indeed play a major factor on one's outlook on life, and thus will most likely influence one's voting choice as well. Therefore, in the post-stratification estimate for this model, $\hat{y}_j$ will be the estimate of the Liberal Party's chances of winning in each province based off of gender and education and N_j is each province's population size.

RESULTS

First, to evaluate the model's capabilities, an Area Under the Curve (AUC) plot was drawn. As it can be seen, the AUC curve is 0.7044, suggesting this is quite a fair model, as this means that this model is should be able to accurately predict 70.4% of the real results. This shows that this model does have strong predictive abilities as 70.04% is relatively high. At the same time, there evidently is also still room for improvements.

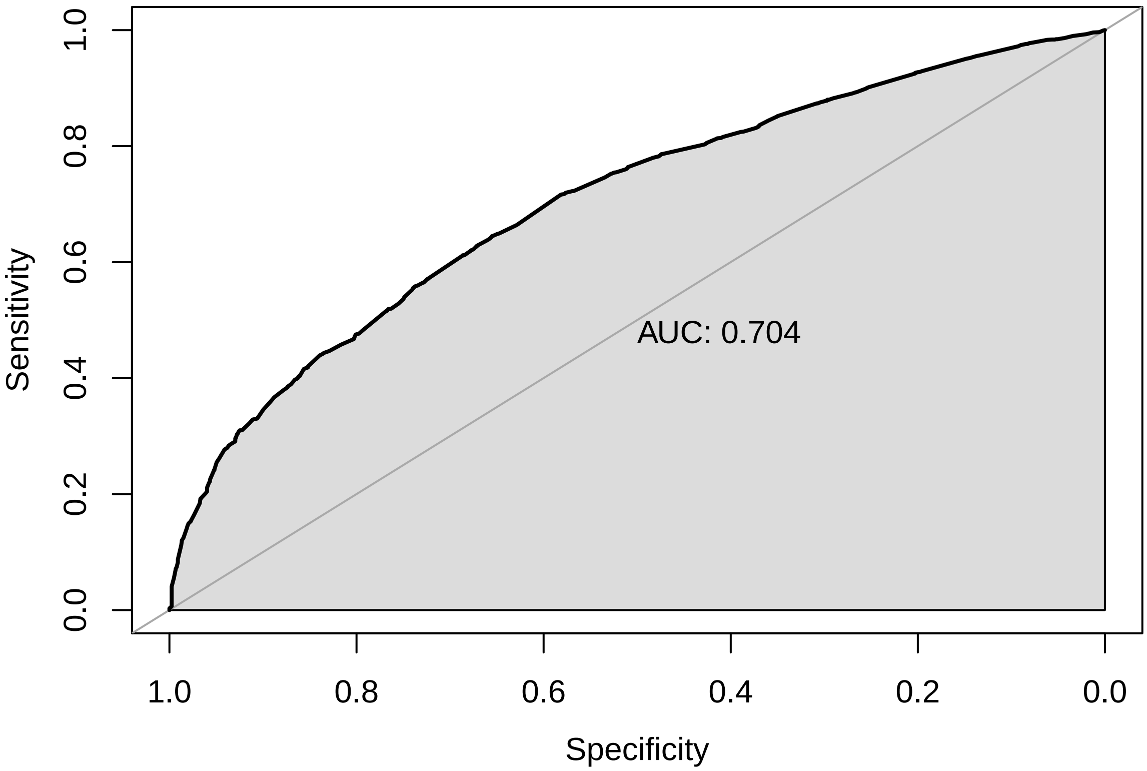

Table 1:

Estimates	Results
Post-stratification est. (Lib.)	0.5048968
Post-stratification est. (Con.)	0.4951032
Popular Votes (Lib.)	0.5064103
Popular Votes (Con)	0.4935897

Following that, the post-stratification estimate, $\hat{Y}^{PS}$ for this study is 50.49%.(Table 1) This means that after performing a thorough study based on each voter's gender, education level and location, the combined weighted result is that approximately 50.49% of Canadians will vote for the Liberal Party, and approximately 49.51% (Table 1) will vote for the Conservative Party. From the post-stratification estimate, it can be seen that the results are pretty evenly distributed amongst these two parties. This is not unexpected, as in real life, the votes between The Liberal Party (33.12%) and The Conservative Party (34.34%), only differed by approximately 220,000 votes which is a difference of approximately 1.2%. These statistics and the post-stratification estimate are only based on the weighted popular votes.

Next, to simulate a more realistic prediction, the final model factored in the First Past the Post system. This system, as described above, tries to factor the weights of each province through seats to make the elections fairer. With each province allotted a certain number of seats, the party who attains the greatest number of seats in the House of Commons compared to the rest of the parties will become the leader. Since this investigation only includes The Liberal and The Conversative parties, in order to win, the parties much attain more than 50% of the 338 seats in the House of Commons. In the end, the model will predict both the overall popular votes, and the overall seat distribution amongst both parties in the House of Commons. With all this data factored in, the model predicted that The Liberal Party will only secure about 49.36% (Table 1) of the votes with only 146 seats (43.19%) in the House of Commons (Table 2). The Conservative Party on the other hand is predicted to win approximately 50.64% (Table 1)of the votes in Canada and receive 192 seats (56.80%) in the House of Commons (Table 2), and thus will become the ultimate predicted winner and form government.

Table 2:

Provinces	Predicated Liberal Seats	Predicted Conservative Seats
Alberta	34	0
British Columbia	24	18
Manitoba	11	3
New Brunswick	3	7
Newfoundland and Labrador	1	6
Northwest Territories	0	1
Nova Scotia	0	11
Nunavut	1	0
Ontario	48	73
Prince Edward Island	1	3
Quebec	8	70
Saskatchewan	14	0
Yukon	1	0
TOTAL	146	192

DISCUSSION

Summary

In this investigation, a successful multilevel logistic regression model using the 2019 CES and a 2016 Canadian educational-based census to simulate a full voter turnout in the 2019 Canadian Federal Election was built. Using only predictors representing one's age, sex education level and location status, the model was able to successfully apply the survey data outcomes on a constituency level and predict a winner between the two leading parties, The Liberal Party of Canada, and The Conservative Party of Canada, in the case off a full voter turnout in the 2019 Canadian Federal Elections. Many different aspects were also factored into the final results, such as a post-stratification estimate, and the First Past the Post system. In the end, in all cases, it seems that given what was factored into the logistic regression model, the Conservative Party would win not only win the popular vote but would also win majority seats in the House of Commons, and in the end become the ultimate leader of the country.

Conclusion

According to the overall popular vote, and the seat distribution in the House of Commons, the Conservative Party of Canada would ultimately win the 2019 Canadian Federal Election in the case of a full voter turnout. The post-stratification estimate, the overall popular vote, and the seat distribution in the House of Commons for the Conservative Party respectively are, 49.51%, 50.64% and 192 seats (56.80%). It can be seen that both the post-stratification estimates, and the overall popular vote is very split amongst both parties, and in the end, it was ultimately the amount of seats that distinguished a clear winner. This suggests that parties should ultimately focus on important provinces and gain their votes, as seats are what ultimately determine the winner in the end.

In comparison to the actual outcome of the 2019 Canadian Federal Election, it can be seen that there is indeed a big difference in the predicted full voter turnout outcome and the actual outcome. In the real outcomes of the 2019 Canadian Federal Elections, The Liberal Party won by 157 seats (46.44%) and a popular vote of 33.12%, whereas The Conservative Party only acquired 121 seat (35.80%) and a popular vote of 34.34%. This is completely different that the predicted model's outcome, as it predicted that the Conservative Party would win with 192 seats (56.80%) and 50.64% of the popular votes. It can also be seen here in the actual results, much like the predicted model's results, that there is a less than 1% difference in the overall popular votes, but the seats distribution percentage differed by a lot. This once again suggests that it may be more important for parties to focus on certain provinces when campaigning, as the end results heavily depend on the number of seats won.

In the end, through this investigation, it can be seen that there may have been a big difference in the outcome of the 2019 Canadian Federal Elections if there was indeed a full voter turnout. This result and finding strongly highlights the saying "every vote matter", as through this investigation and the actual results of the 2019 Canadian Federal Election, it can be seen that the competition between the two leading parties is extremely slim. Therefore, any small changes in voting numbers can heavily influence the ultimate outcome of the elections.

Weaknesses

As it can be seen, the Canadian Federal Elections are very intricate and involves many complex factors that all may potentially influence it. To successfully simulate a full voter turnout, many factors and inputs had to be simplified. Due to this, there definitely exist many aspects that might have been missed in this investigation.

One major drawback is that in the 2019 CES and the 2016 Highest Level of Educational Attainment (general) by Sex and Selected Age groups" census, there were only four predictor variables were chosen to create the multilevel regression model. This is because they were the only four variables that related to this study, and that matched up between both the survey and census data. However, his may be a major weakness because four variables is most likely not quite enough to get a detailed picture of each individual, which is what is needed to perform a MRP.

Along with that, while cleaning both the survey and census dataset, there were many missing or unnecessary values. For example, in the census data, there were many people who answered their province as "Canada", which is not very helpful, and thus were deleted from this investigation so that it would not clutter the rest of the data. However, at the same time, those missing values could have been relatively important, as it can be seen that any amount make a difference in the final outcome. Another important missing data example would be that there were no individual between the age of the age of 18 and 24 studied, as there simply were no classifications of this age group in the census dataset. This may have had a heavy impact on the final results, as nowadays, the younger generation play a big role in the election outcome, as their opinions are very valuable and may differ a lot from the older generations.

Following that, in this investigation, only two outcomes, The Liberal Party and The Conservative Party were considered. In reality, there are eight parties that ran in the 2019 Canadian Federal Elections. The decision of only assessing these two parties is due to the fact that in the actual 2019 elections, approximately 71% of those votes went to these two parties, thus meaning that they are the leading competitors, and that the final decision was truly only split between them. Due to this, a lot of information from the other parties are missing.

Lastly, in this investigation the census used is from the year 2016. Although, most of the data in this investigation were adjusted to the 2019 information, it is still important to keep in mind that, there most likely are many now eligible voters that were not captured by the 2016 dataset. Also, between 2016 and 2019, many things could have happened to all Canadian citizens. Therefore, this may be a weakness as the census data is not up-to-date with the year of the elections being studied.

Next Steps

The next steps to better improve for this investigation should start with fixing the weaknesses. Finding a stronger and more up-to-date census would solve many of the weaknesses above. A stronger and more up-to-date census would allow there to be more predictor variables and more information to work with, and thus will result in a more accurate final result. Next would be to work with models using more complex functions so that all parties in the actual election can be incorporated. By doing this, final results will include information and estimates on not only the Liberal and Conservative Party, but all other parties such as The New Democratic Party (NDP) and The Green Party. Lastly, nowadays, modern technology is very developed, and therefore perhaps stronger and more complex models can be created to study a case as intricate and complex as the Canadian Federal Election for a more accurate result.

REFERENCES

"A History of Vote in Canada." Elections Canada, www.elections.ca/ content.aspx?section=res&dir=his&document=intro&lang=e.

Government of Canada, Statistics Canada. "Education Highlight Tables, 2016 Census." Government of Canada, Statistics Canada, 27 Nov. 2017, www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm.

Government of Canada, Statistics Canada. "Reasons for Not Voting in the Federal Election, October 21, 2019." The Daily - , 26 Feb. 2020, www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm.

Publishing, Custom and Community. "Your Vote Is Your Voice: Use It! Why It's Important to Exercise Your Right to Vote as a Canadian: The Chronicle Herald." Business-Voice | More | The Chronicle Herald, 7 Oct. 2019, www.thechronicleherald.ca/more/business-voice/your-vote-is-your-voice-use- it-why-its-important-to-exercise-your-right-to-vote-as-a-canadian-361063/.

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

"Welcome to the 2019 Canadian Election Study." Canadian Election Study, www.ces-eec.ca/.

"2019 Canadian Federal Election." Wikipedia, Wikimedia Foundation, 17 Dec. 2020, en.wikipedia.org/wiki/2019_Canadian_federal_election.

APPENDIX

https://github.com/ccxin0/STA304-Final-Project-CG