Reviewing 2019 Canadian Federal Election Results with MRP Model

Zhihuan Shao

Dec 20, 2020

Contents

1 Abstract

Federal Election is always one of the biggest and most important events in a country. As the 2019 Canadian Federal Election has ended for almost a year now, the fact that a large proportion of citizens were not able to votes in the 2019 election evokes the hypothesis of a potential completely different results should everyone had voted then. This report utilized Multilevel Regression Post-stratification technique, along with logistic modeling and random intercept modeling to predict the potential different election results based off 2019 Canadian Election Study data and General Social Survey (GSS) on Canadians at Work and Home in 2016 data. The report successfully predicts that even though every single citizen could have voted in 2019 Canadian Federal Election, the popular voting result was still less likely to change, although support rate for minor parties could increase, but not for major parties.

Keywords: Canadian Federal Election, Logistic model, Multilevel Regression Post-Stratification, Canadian Election Study, General Social Survey

2 Introduction

Voting and elections are the most basic elements of democracy. The 2019 Canadian federal election (formally the 43rd Canadian general election) was held on October 21, 2019, to elect members of the House of Commons to the 43rd Canadian Parliament. The writs of election for the 2019 election were issued by Governor General Julie Payette on September 11, 2019 [1]. The results were exciting and competition was fierce. Although once again, Liberal party has taken charge of the government, however, not without cost. The election results are demonstrate in Table 1 below:

Table 1: 2019 Canadian Election Results

Party Name	Party Leader	Seats Won	Popular Votes	Popular Percentage
Liberal	Justin Trudeau	157	6018728	33.12%
Conservative	Andrew Scheer	121	6239227	34.34%
Bloc Quebecois	Yves-Francois Blanchet	32	1387030	7.63%
New Democratic	Jagmeet Singh	24	2903722	15.98%
Green	Elizabeth May	3	1189607	6.55%
People's	Maxime Bernier	0	294092	1.62%

And the visualized results separated by provinces is shown as below:

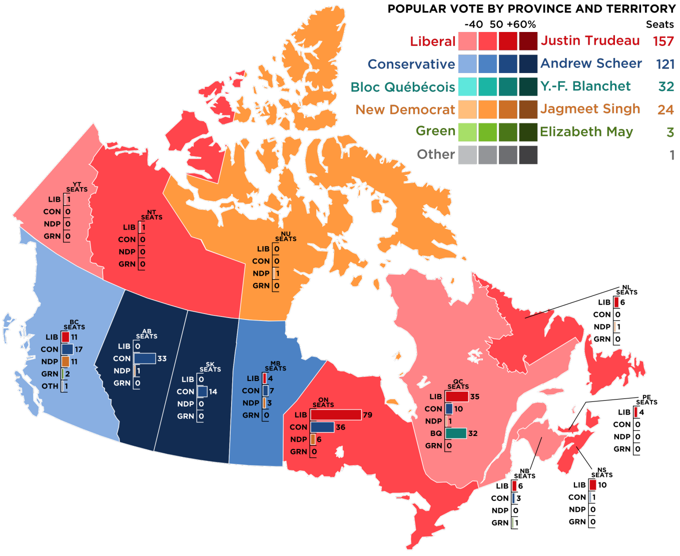
POPULAR VOTE BY PROVINCE AND TERRITORY
-40 50 +60%
Seats
Liberal Justin Trudeau 157
Conservative Andrew Scheer 121
Bloc Québécois Y.-F. Blanchet 32
New Democrat Jagmeet Singh 24
Green Elizabeth May 3
Other 1
YT SEATS LIB 1 CON 0 NDP 0 GRN 0
NT SEATS LIB 1 CON 0 NDP 0 GRN 0
NU SEATS LIB 0 CON 0 NDP 1 GRN 0
BC SEATS LIB 11 CON 17 NDP 11 GRN 2 OTH 1
AB SEATS LIB 0 CON 33 NDP 1 GRN 0
SK SEATS LIB 0 CON 14 NDP 0 GRN 0
MB SEATS LIB 4 CON 7 NDP 3 GRN 0
ON SEATS LIB 79 CON 36 NDP 6 GRN 0
QC SEATS LIB 35 CON 10 NDP 1 BQ 32 GRN 0
NL SEATS LIB 6 CON 0 NDP 1 GRN 0
PE SEATS LIB 4 CON 0 NDP 0 GRN 0
NB SEATS LIB 6 CON 3 NDP 0 GRN 1
NS SEATS LIB 10 CON 1 NDP 0 GRN 0

Figure 1: Election Results by Province and Territory, Wikipedia

Liberal party held 157 seats and won the election, however, it lost the competition of popular voting to the Conservative party. Compared to the overwhelming success of Liberal party in 2015, it was such a close one in the 2019 election. Political reasons are not the focus of this report, but the total number of voters seems far less than the number of Canadian citizens.

By roughly calculating the popular voting numbers, there were only approximately 16 millions of voters that had voted in 2019, compared to approximately 36 millions of citizens across Canada, according to 2016 census [2]. Although in reality, it is not possible for every citizen over 18 years old to vote for the election, due to restrictions in numerous aspects, the proportion of citizens that had voted in 2019 was still lower than expected. Nevertheless, is it possible that the election results could be different if EVERYONE had voted in 2019 Canadian Federal Election, especially the popular votes percentage for each party? What aspects can contribute to the outcome of whether a voter would vote for a certain party?

To study this problem, this report will review the 2019 Canadian Federal Election results using CES data and by assuming every qualified Canadian citizen could have voted, to use post-stratification technique based on General Social Survey data from 2016 to estimate whether the results might be different if ‘everyone’ had voted in 2019.

This report will include four major parts:

1. Data Introduction and Exploring, in which original CES survey data will be visualized to explore several characteristics in preparation of further analysis;
2. Model Specifications, in which mathematical and statistical models that have been used in this report will be specified, explained, and evaluated;
3. Results, in which estimated results from models and post-stratification technique will be demonstrated and compared to the actual election results, as well as discoveries from model exploration;
4. Discussion, in which the results will be evaluated, along with the weaknesses of this analysis and potential future directions to make this analysis better

3 Data Introduction and Exploring

Throughout its long history, the CES has been a rich source of data on Canadians’ political behaviour and attitudes, measuring preferences on key political issues such as free trade with the US, social spending and Quebec’s place in Canada; political actors, such as parties, party leaders and the government; and social concerns, such as women’s place in the home, support for immigration, and attitudes toward gays and lesbians; as well as political preferences and engagement. These data provide an unparalleled snapshot and record of Canadian society and political life [3]. The multifaceted data from CES dataset makes a perfect survey sample to investigate which aspects or characteristics of voters would affect the decision of voting for a certain party.

General Social Survey (GSS) on Canadians at Work and Home in 2016, which is a sample survey with cross-sectional design and conducted from August 2nd to December 23rd 2016. The target population includes all non-institutionalized persons 15 years of age and older, living in the 10 provinces of Canada. This survey aimed at taking a comprehensive look at the way Canadians live by incorporating the realms of work, home, leisure, and overall well-being, and thus knowing more about the lifestyle behaviors of Canadians that impact their health and well-being both in the workplace and at home [4]. The multi-diversity and detailed information of each individual involved in this survey makes it a good source for data post-stratification. Both data are observational data, meaning the data were collected based on random selection upon observations rather than a fully experimental data collection and introducing a systemetic intervention to study any effects.

The 2019 CES data have quite a lot of predictors, a few of which are considered important to this analysis are: province of residence, sex, age, education background, income, vote choice, and satisfaction towards government. The full data demonstration is attached to the Appendix page. The voting choices of survey participants are demonstrated below:

Based on 2019 CES data, Figure 2a demonstrates that most voters preferred Liberal Party, as expected, however, Conservative Party is chasing right behind. New Democratic is the third, followed by Green and Bloc Quebecois party. People’s Party received the least number of supporters. The results are a little bit different from what we have seen in 2019 Canadian Federal Election data, that Conservative party should lead in popular voting and Liberal Party should follow behind, although Liberal won eventually. Green Party should have less supporters than Bloc Quebecois according to 2019 election results. Could this imply that the actual outcome could be different?

Besides, the data also present some other interesting aspects. For example, the satisfaction of government:

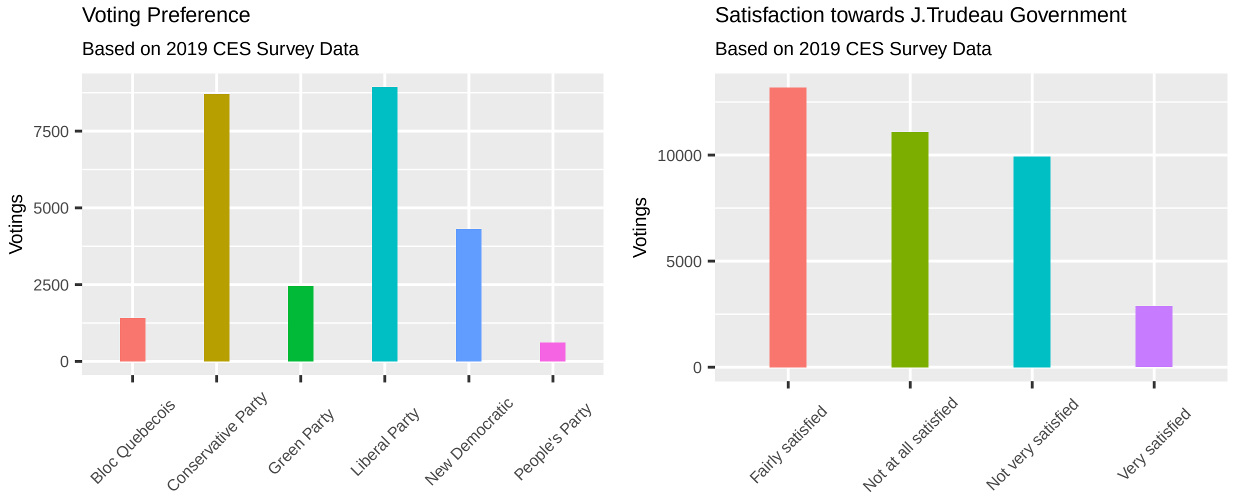

Figure 2a

Figure 2b

Figure 2: Voting Choices and Govenment Satisfaction Index

The fact that a large numbers of voters were “Not At All Satisfied” with Trudeau government might be the reason why Liberal party lost popular voting to Conservative party in 2019 election. Only very few portion of voters thought “Very satisfied”. Without no further details, it is hard to specify which aspect the majority of voters thought unsatisfied with Trudeau government. To further look into potential correlation of satisfaction level with other aspects:

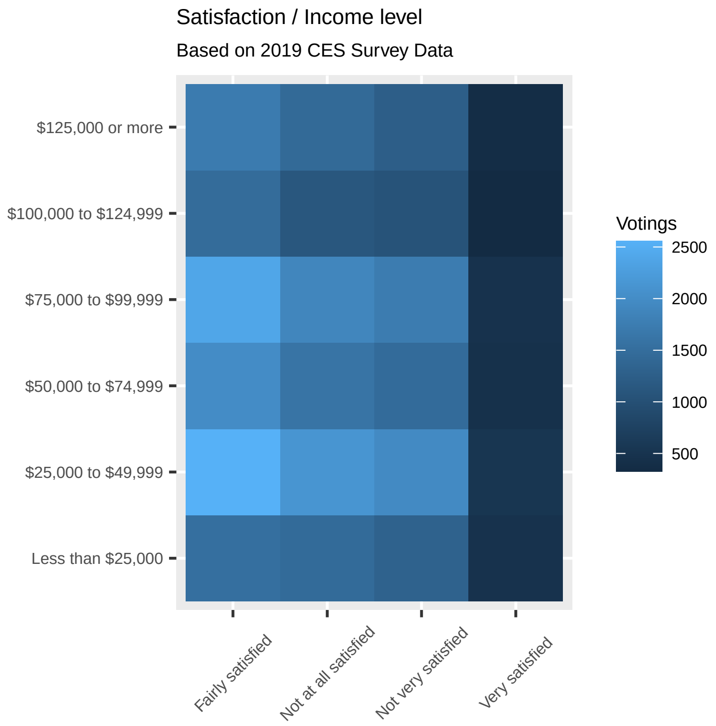

Figure 3: Government Satisfaction vs Income Level

Mapping satisfaction with income level, Figure 4 shows a weak correlation between two indicators, that voters with incomes lower than 100,000 annually were more unsatisfied with the government, especially with incomes ranging from 75,000 to 100,000, potentially related to the tax policies. Voters with higher incomes are more likely to be satisfied with Trudeau government.

4 Model

In order to look into potential different outcome if "everyone" had voted in 2019 election, this report utilized Rstudio software to implement Multilevel Regression Post-stratification with logistic model and random interception model for calculating and estimating the voting preference of different parties for voters.

4.1 Study Design

The report will follow the study design in Figure 4:

To estimate the voting preference of individuals based on certain criteria, logistic model with random interception model will be used on CES survey data. Logistic model has the advantage of estimating binary outcomes, which is either 1 (yes) or 0 (no). Transforming each voter's preference for parties into "whether will vote for a certain party or not" can greatly take advantage of logistic model to estimate voting preference. In this way, by using logistic model, it allows for estimating the outcome on a large participants basis and calculating possibilities of each individual's voting preference for each Party (mainly focused on the six largest parties, smaller parties will be labeled as "others"). The consideration of using random interception arose from the facts that parties might have a preference for regions, for example, participants from Quebec might have a stronger preference for Bloc Quebecois. In attempt to minimize the influences of regional differences, a random interception model will be applied in addition to logistic model.

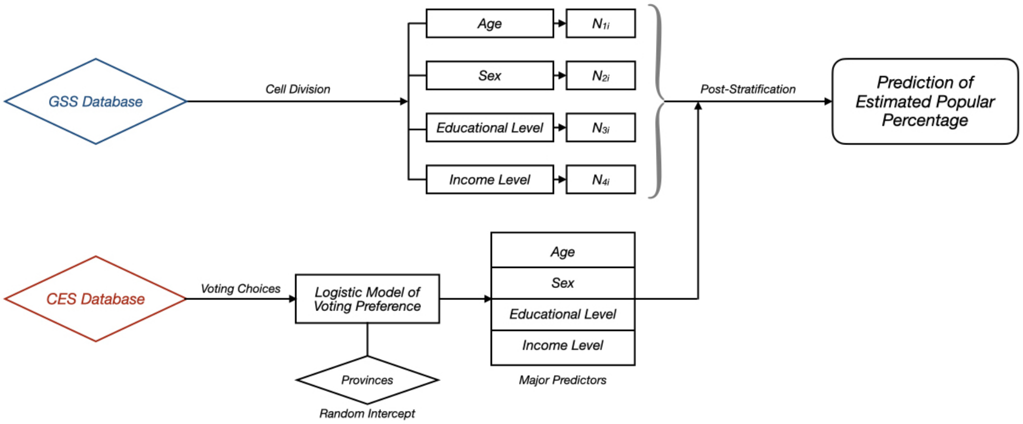

Figure 4: Study Design

By applying the model estimated by logistic models to post-stratification data, it is possible to estimate if 'everyone' had voted, how the results could be different.

4.2 Model Specifics

The logistic model will be focused on four large aspects of voters: age, sex, educational level, and income level, with a random interception model of provinces. The rationality are described as below:

1. **Age:** Age will be a comparable larger aspect compared to other aspects, due to the nature that younger voters will tend to favor more modernized and innovative policies, while elder voters will tend to favor conservative policies. The ages will be divided into 7 categories, ranging from 0 (18-25 years old) to 6 (older than 75 years old).
2. **Sex:** Sex could also be an interesting aspect, that voters of different sexes might have a preference for a certain party will the corresponding leader. The image and appearance of a party leader could sometimes also affect the affinity of voters to a party.
3. **Educational Level:** The educational level is considered a high-weighted aspect. Although the understanding of society and policy is not necessarily based off higher education backgrounds, the potential benefit from higher educations could still make implicit differences when deciding to vote for any parties. For example, higher education in research areas might have a preference for parties with scientific benefits in corresponding areas.
4. **Income Level:** The income level is considered a heavily-weighted aspect, because the tax policies will heavily impacted the income of voters. Thus, whether the policies will impact higher income or lower income groups would tend to affect whether the corresponding voters would vote for the party or not.

With the rationality, the logistic models will be fitted onto all voters and their potential choices for parties based off age, sex, educational level, and income level, with a random interception model on provinces. The formula is shown below:

$$\log\left(\frac{\hat{p_i}}{1-\hat{p_i}}\right) = \hat{\beta_0} + \hat{\beta_1}X_{Age} + \hat{\beta_2}X_{Male} + \hat{\beta_3}X_{Education} + \hat{\beta_4}X_{Income} + \hat{\epsilon}_{Province}$$

In which:

i. $\hat{p_i}$ represents the voters estimated probability of voting for a party called party i.

ii. $\hat{\beta_0}$ represents the intercept of the model, and is the logistic estimator of probability of voting for a voter that is a female between 18-25 years old with minimal educational level and lower annul income less than \$25000. Additionally, $\hat{\beta_1}$, $\hat{\beta_2}$, $\hat{\beta_3}$, $\hat{\beta_4}$ represent the factors that voters with different sex, educational level, and income levels can contribute to different logistic estimators, which represent the probability of voting the certain candidate respectively. Positive coefficients indicate the predictors have positive impacts on whether voters are more willing to vote, for example, a positive $\hat{\beta_4}$ indicate the higher the income is, the higher the willingness of the voter to vote for this party, and vice versa. Also, the larger the absolute value is, the heavier the predictor can impact the preference for a certain party. Need to mention that the range for X_{Age} is from 0-6, range for $X_{Education}$ is from 0-7, and range for X_{Income} is from 0-5.

iii. $\hat{\epsilon}_{Province}$ indicates the random interception model predictor that used to correct the impact of party preferences in difference provinces.

iv. For each individual party, separate models will be applied, thus the coefficients will also vary, meaning $\hat{\beta_1}$, $\hat{\beta_2}$, $\hat{\beta_3}$, $\hat{\beta_4}$ and $\hat{\epsilon}_{Province}$ will be different in different models

Using logistic models to fit survey data for all parties:

1. Logistic model of popular voting - Liberal Party:

$$\log\left(\frac{\hat{p_1}}{1-\hat{p_1}}\right) = -1.689 + 0.043X_{Age} - 0.009X_{Male} + 0.121X_{Education} + 0.027X_{Income} + \hat{\epsilon_1}_{Province}$$

Education levels seem to have the highest impact on voters' preference for Liberal party, whereas age and income has less impact. Female seem to have a slight stronger favor for Liberal party.

2. Logistic model of popular voting - Conservative Party:

$$\log\left(\frac{\hat{p_2}}{1-\hat{p_2}}\right) = -1.396 + 0.109X_{Age} + 0.444X_{Male} - 0.113X_{Education} + 0.133X_{Income} + \hat{\epsilon_2}_{Province}$$

The conservative party has much higher preference for male voters, and a quite high preference for elder voters and voters with higher income. However, educational level seem to have negative impacts, that voters with lower educational levels have higher preference for the Conservative party.

3. Logistic model of popular voting - Bloc Quebecois:

$$\log\left(\frac{\hat{p_3}}{1-\hat{p_3}}\right) = -17.750 + 0.239X_{Age} + 0.230X_{Male} - 0.081X_{Education} + 0.012X_{Income} + \hat{\epsilon_3}_{Province}$$

Interestingly, the intercept is extremely low for this logistic model estimating voter preference for Bloc Quebecois, meaning a female voter between 18-25 years old with minimal educational level and lower annul income has almost no tendency for Bloc Quebecois.

4. Logistic model of popular voting - New Democratic Party:

$$\log\left(\frac{\hat{p_4}}{1-\hat{p_4}}\right) = -0.809 - 0.270X_{Age} - 0.356X_{Male} + 0.024X_{Education} - 0.126X_{Income} + \hat{\epsilon_4}_{Province}$$

With only educational levels be positively affect the potency of voters voting for NDP, other predictors indicate that voters with younger ages and lower incomes will have more preference for NDP, potentially related to the party's policies.

5. Logistic model of popular voting - Green Party:

$$\log\left(\frac{\hat{p_5}}{1-\hat{p_5}}\right) = -1.898 - 0.136X_{Age} - 0.088X_{Male} + 0.042X_{Education} - 0.104X_{Income} + \hat{\epsilon_5}_{Province}$$

Similarly, voters with younger ages and lower incomes have a higher preference for the Green Party. However, voters with higher educational levels have a higher preference for the Green party rather than NDP.

6. Logistic model of popular voting - People's Party:

$$\log\left(\frac{\hat{p_6}}{1-\hat{p_6}}\right) = -2.890 - 0.223X_{Age} + 0.601X_{Male} - 0.100X_{Education} - 0.119X_{Income} + \hat{\epsilon_6}_{Province}$$

As for People's party, female voters seem to favor more than male voters. Besides, younger ages, lower educational level and lower incomes seem to favor People's party.

7. Logistic model of popular voting - Other Parties:

$$\log\left(\frac{\hat{p_7}}{1-\hat{p_7}}\right) = -5.343 + 0.086X_{Age} + 0.327X_{Male} + 0.018X_{Education} - 0.089X_{Income} + \hat{\epsilon_7}_{Province}$$

As for other parties, due to the fact that voter base is much smaller than that of other larger parties, it does not really affect the decision of most voters with the same predictors. However, male voters do seem have a higher preference.

Also, by taking a look at the random intercept predictor (province), it allows for an analysis of potential province-party preference:

Table 2: Province Preference for Parties

	Liberal Party	Conservative Party	Bloc Quebecois	New Democratic	Green Party	People's Party
Alberta	-0.736	1.202	-0.023	-0.271	-0.714	0.008
British Columbia	-0.071	-0.018	-0.023	0.406	0.509	-0.019
Manitoba	-0.074	0.427	-0.009	0.178	-0.118	-0.042
New Brunswick	0.261	-0.119	-0.005	-0.602	0.777	0.036
Newfoundland and Labrador	0.519	-0.370	-0.003	0.349	-0.943	-0.022
Nova Scotia	0.414	-0.608	-0.006	-0.013	0.370	0.010
Ontario	0.277	-0.005	-0.069	0.197	-0.064	0.059
Prince Edward Island	0.278	-0.634	-0.001	-0.275	1.009	-0.012
Quebec	0.092	-0.720	15.810	-0.425	-0.230	-0.036
Saskatchewan	-0.947	0.870	-0.007	0.489	-0.535	0.021

Perhaps a heatmap could be more informative and provide visualized comparasion of province preference differences:

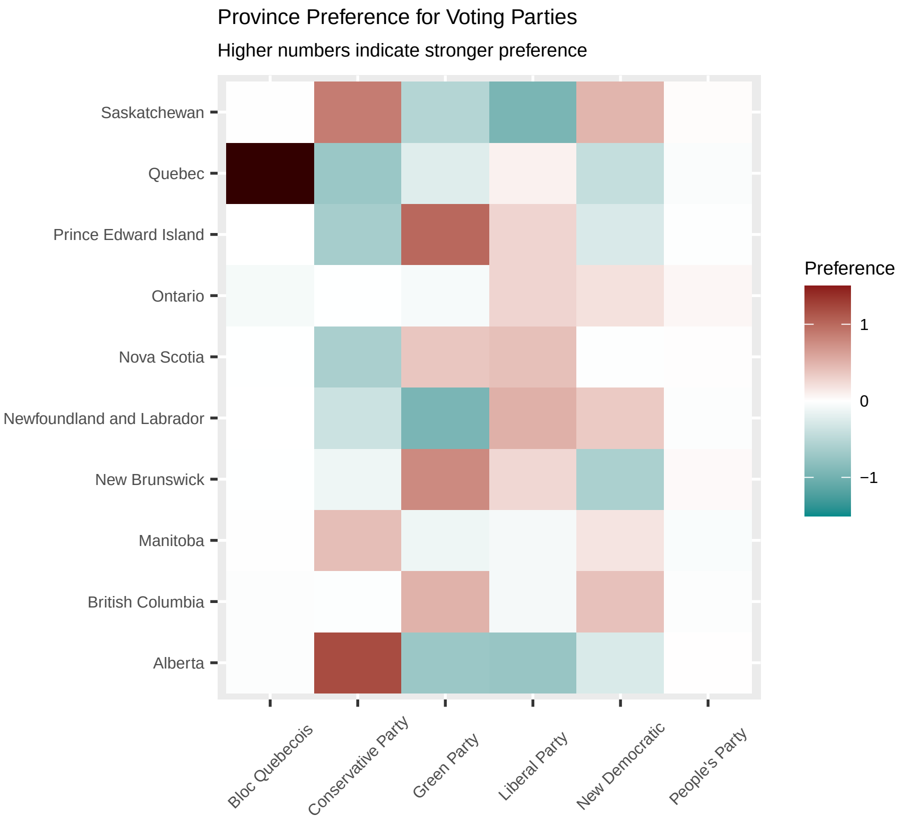

Figure 5: Provincial Voting Preference

The warmer the color is, indicating the higher preference for a certain party of voters from a certain province, and the cooler the color is, the lower the preference is. As expected, voters from Quebec have an extremely higher preference for Bloc Quebcois, and voters from Saskatchawan and Alberta have a strong preference for Conservative Party. Voters from Prince Edward Island and New Brunswick also have a stronge preference for the Green Party, and they really dislak the Liberal Party to some extent. However, voters from Newfounder and Labrador citizens seem to disagree. Other than those findings, no significant high or low preference for other parties based off geological differences.

With the models indicated above, it now allows for an estimation of popular votes for each parties with post-stratification technique and data from GSS in 2016.

4.3 Post-Stratification

To estimate the voting preference of voters in a large scale, this report will also apply post-stratification technique, using data from General Social Survey (GSS) on Canadians at Work and Home in 2016. This dataset has a very detailed information regarding various aspects of participants, thus it allows for a more accurate estimation of preference for a certain party.

The post-stratification technique will need to divide the initial GSS dataset into 3686 cells based off province, age, sex, educational level, and income level, with a counted number of participants inside each cell. The following formula shows how post-stratification technique works:

First, calculating the logistic estimator $\hat{y_{ij}}$ of cell j based on the logistic model of party i and each individual predictors (age, ses, education, income and province), and the estimated probability or proportion of voters in the cell j that are willing to voter for the party $\hat{p_{ij}}$ could also be calculated:

$$\hat{y_{ij}} = \log\left(\frac{\hat{p_{ij}}}{1-\hat{p_{ij}}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{Age_{ij}} + \hat{\beta}_2 X_{Male_{ij}} + \hat{\beta}_3 X_{Education_{ij}} + \hat{\beta}_4 X_{Income_{ij}} + \hat{\epsilon}_{Province_{ij}}$$

$$\Downarrow$$

$$\hat{p_{ij}} = \frac{e^{\hat{y_{ij}}}}{e^{\hat{y_{ij}}}+1}$$

Then, it allows for the proportion of voters in all cells to be merged together and form the finalized estimated proportion of all voters that are willing to vote for party i

$$\hat{p_i}^{ps} = \frac{\sum N_j \hat{p_{ij}}}{\sum N_j}$$

In which $\hat{y_{ij}}$ indicate the logistic estimator of the voting preference for a certain party based off the corresponding logistic model in a certain cell j divided by province, age, sex, educational level, and income level. N_j indicates the number of participants in cell j, and $\hat{p_{ij}}$ indicates the proportion of voters in favor of party i within cell j By using this formula, we can estimate the at what percentage of voters $\hat{p_i}^{ps}$ in a large scale that are willing to vote for a specific party, which also represent the estimated popular voting rates.

And the results are shown in Table 3

5 Results

Using logistic model and post-stratification technique, this report is able to calculate and estimate the percentage of voters that favor each party, showing below (only showing finalized results here, all peripheral parameters or estimators will be included in Appendix):

Table 3: Popular Votes for All Parties

	Popular Votes in Percentage
Liberal Party	27.06
Conservative Party	27.20
Bloc Quebecois	3.55
New Democratic	14.57
Green Party	9.17
People's Party	2.38
Other Parties	0.70

From the results, this report successfully estimates that **27.06%** of voters would vote for the Liberal Party, however, the Conservative Party still takes the lead in popular votes, with **27.20%**, slightly higher than that of the Liberal Party. New Democratic party has a much higher than expected support rate according to this model, at **14.57%**, followed by the Green party, Bloc Quebecois, and the People's party. Only **0.70%** of voters would consider other parties.

Nevertheless, the result is not finalized. Noting that the overall percentage of voters number is not at 100%, but only at 84.63%. This is due to the fact that from CES survey data, there was still quite a large number of participants that were not willing to vote or had not decided which party to vote. Statistics shown as below:

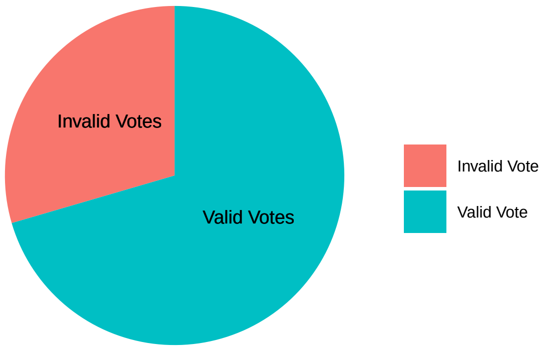

Figure 6: Valid and Invalid Votes

Considering the purpose of this study is to look into potential popular voting differences if every single citizen had voted, it is necessary to make an assumption that all citizens would vote. Thus, to adjust the results, assume all votes that are used to estimate the results are the total number of votes, and the adjusted result is shown below:

Table 4: Popular Votes for All Parties - With Adjustment

	Popular Votes in Percentage	Adjusted Value
Liberal Party	27.06	31.97
Conservative Party	27.20	32.14
Bloc Quebecois	3.55	4.19
New Democratic	14.57	17.22
Green Party	9.17	10.83
People's Party	2.38	2.81
Other Parties	0.70	0.83

The adjusted result indicates that there would be **31.97%** of citizens willing to vote for the Liberal Party, while **32.14%** of voters are in favor of the Conservative Party. The Conservative party still has a slight advantage over the Liberal Party. More voters seem to favor smaller parties, especially the New Democratic Party and the Green Party, at **17.22%** and **10.83%** respectively, followed by Bloc Quebecois and the People's Party. This result is based off our post-stratification analysis of the proportion of voters in favor of each party modeled by a Multilevel Regression Post-stratification model with logistic modeling and random intercept modeling, and the results are adjusted upon the assumption that all citizens would vote and all votes are valid votes.

Comparing the result to the actual 2019 election results:

Comparing to the actual 2019 results, the differences are not as huge as expected, especially for the two largest parties. Conservative party still leads a slight bit compared to Liberal party, however, also still lost electoral seats. The advantages of Conservative party is diminished as per the prediction of this model. Nevertheless, more voters seem to favor smaller parties according to this model, except for Bloc Quebecois. For Green party, the popular votes increase around 4%, which could be considered huge; the New Democratic party and the People's party could have slight more supporters, around **1.3%** percent both. However, the only party that seem to lose in popular votes is Bloc Quebecois, dropped around **3.5%**.

Table 5: 2019 Canadian Election Results vs Predicted Results

Party Name	Party Leader	Seats Won	Popular Votes	Popular Percentage	Popular Percentage - Predicted
Liberal	Justin Trudeau	157	6018728	33.12%	31.97%
Conservative	Andrew Scheer	121	6239227	34.34%	32.14%
Bloc Quebecois	Yves-Francois Blanchet	32	1387030	7.63%	4.19%
New Democratic	Jagmeet Singh	24	2903722	15.98%	17.22%
Green	Elizabeth May	3	1189607	6.55%	10.83%
People's	Maxime Bernier	0	294092	1.62%	2.81%

6 Discussion

This report utilizes MRP technique with logistic model and random intercept model based of CES and GSS data to estimate if every single citizen had voted in 2019, would the election results be any different? The predicted result is quite interesting, showing that Conservative party could still have more supporter ratio compared to the Liberal party, at **32.14%** to **31.97%**. It appears that this model predicts, if every citizen could have voted, the popular votes advantage for the Conservative party could be diminished from over 1% to less than **0.2%**. For smaller parties, they might acquire a stronger support from voters, that the Green party could have **4%** more support ratio, and both the New Democratic party and the People's party could earn around **1%** more support ratio. However, as for Bloc Quebecois, things might not end as smooth as expected, because the model predicts that it could lose around **3.5%** in popular votes.

In conclusion, this model predicts that if all citizens had voted in 2019, the result was still less likely to change on a large scale. Two largest parties, Liberal and Conservative would still lead in popular votes, and the fact that electoral seats won't change significantly even though everyone could have voted implies that the result would still be exactly the same, that the Liberal party will still run the government, the only difference would be still leading a minority government.

There could be a few reasons that contribute to the result differences:

1. For large parties, the differences between number of popular votes would further decrease as more citizen could have voted, since most citizens that did not vote might due to geological restrictions, technical restrictions, and many other minor restrictions. From the province and voting preference correlation plot above (Figure 6), voters from provinces that have large populations (e.g. Ontario, British Columbia, Alberta) seem to have a stronger preference for large parties rather than smaller parties, and it is foreseeable that a large portion of voters should come from those large provinces. This could induce bias towards large parties. As the model suggests, if voters could overcome any restrictions to vote, the advantage of this bias could be weakened, thus the advantage between large parties could be further diminished, as the model predicted.

2. For smaller parties, with more voters that could potentially overcome any restrictions and vote, the increase in popular votes is also reasonable. The fact that geological and technical restrictions could potentially delay the vote activities in smaller areas that might have a higher preference for smaller parties could result in the large gap between smaller parties and larger parties.

3. As for Bloc Quebecois, it is interesting to see a decrease in popular votes, however, it is also not unpredictable. Figure 6 also suggests an extremely higher preference for Quebec citizens to vote for Bloc Quebecois, this could induce a bias as well, since Quebec is also one of the larger provinces and could have a higher overall vote rate compared to other smaller provinces. Similarly, with more voters supporting other parties from smaller provinces being considered as valid votes, the province bias could be weakened, and this is especially worse for a party that has an extremely higher province preference, such as Bloc Quebecois. The total number of votes might still be higher, but the ratio will be lower that those of other parties.

6.1 Weaknesses

There are a few weaknesses of this study:

1. The first major weakness is the timeliness of both data.

CES data: The CES dataset is an excellent dataset for studying the election, however, there could still be a timeliness problem with it. The CES survey was conducted around May.2019 [3], a few months before the election, and it is not difficult to predict that there could be a portion of people not actually voted for the exact same party as they mentioned when taking the survey. The nearer the election is, the fiercer the competition would be, and throughout those last months, situations could had changed drastically.

GSS 2016 data: The GSS data is also an excellent data for post-stratification, but the fact that it was collect in 2016 makes it quite outdated in comparison to the CES data. Although GSS 2017 data and census data were both initially considered as the alternative dataset for post-stratification, the fact that GSS 2017 dataset lacks a lot of basic predictors (e.g. income level) and census 2016 dataset was collected in a completely difference framework makes both not as good as GSS 2016 dataset. But the outdated information still could induce increase errors into this model and the predicted results

2. The second major weakness is the reliability of post-stratification data.

Another big problem with GSS 2016 dataset is the smaller survey population pool. CES 2019 has 37,822 online survey participants and 4,021 phone survey participants, while GSS 2016 only has 19,609 participants. The post-stratification dataset being smaller than the model dataset makes the post-stratification data less reliable in a large scale, especially it is used for estimating results on a full population basis. While doing the data cleaning, it is not hard to notice that a lot of cells divided based off the predictors have only 1 subject. The bias towards smaller cells could induce errors while estimating the results for parties with less supporters, for example, the People's party, and result in a larger than actual number.

3. The third major weakness is the bias of survey itself.

According to the description of both CES and GSS survey collection [1] [4], CES data were collected mainly through online survey (survey over phone had only limited questions and almost no useful predictors, which was abandoned in this analysis) and GSS data were collected main through landline or telephone. Thus the difference between data collection methods induced an bias of target population: the prediction models are established based off online survey with participants that have higher chance of accessing internet and modern communication methods, while the post-stratification estimations are calculated based off landline and phone survey, which may have limited access to internet and modern communication methods. Overall, GSS data should be less biased regarding technical restrictions, and the bias of models induced by CES data could potentially affect the predicted results.

6.2 Next Steps

The next steps of this analysis could mainly include better post-stratification data and modeling.

For post-stratification data, census is still considered the best source for estimating results on a full population basis. The last census was 2016, which implies that the next census is in the near future (2021 census). Using the newsest census data, with complicated data cleaning method to make the data usable as post-stratification data, the estimation of results should be much more accurate than the current one.

As for modeling, a few more techniques could be implemented, for example, using Bayesian MRP model, or using random intercept model for all categorical predictors instead of transforming them into numerical predictors, could both produce better results.

7 References

1. 2019 Canadian federal election, Wikipedia, https://en.wikipedia.org/wiki/2019_Canadian_federal_election

2. 2016 Census of Population – Data products, Statistics Canada, https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm

3. 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

4. General Social Survey (GSS), Cycle 30, 2016 : Canadians at Work and Home

5. Rohan Alexander and Sam Caetano, October 2020, gss_cleaning.r, License from MIT

6. 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.

7. Population and Dwelling Count Highlight Tables, 2016 Census, Statistics Canada, https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Comprehensive.cfm

8 Appendix

8.0.1 CES survey data demonstrating

```
## # A tibble: 20 x 10
##       ID province age_group age_level sex     edu income_level income
##    <dbl> <chr>    <fct>         <dbl> <chr> <dbl> <fct>         <dbl>
##  1     1 Quebec   25 to 34~         1 Fema~     7 <NA>             NA
##  2     2 Quebec   15 to 24~         0 Fema~     7 <NA>             NA
##  3     3 Ontario  15 to 24~         0 Fema~     5 <NA>             NA
##  4     4 Ontario  15 to 24~         0 Male      5 <NA>             NA
##  5     5 Ontario  15 to 24~         0 Fema~     2 $50,000 to ~      2
##  6     6 Ontario  15 to 24~         0 Fema~     5 <NA>             NA
##  7     7 Ontario  15 to 24~         0 Fema~     2 Less than $~      0
##  8     8 Ontario  15 to 24~         0 Fema~     5 $100,000 to~      4
##  9     9 Ontario  15 to 24~         0 Fema~     5 <NA>             NA
## 10    10 Ontario  15 to 24~         0 Male      2 <NA>             NA
## 11    11 Ontario  15 to 24~         0 Fema~     2 <NA>             NA
## 12    12 Ontario  15 to 24~         0 Fema~     5 <NA>             NA
## 13    13 Ontario  15 to 24~         0 Male      2 $125,000 or~      5
## 14    14 Ontario  15 to 24~         0 Fema~     5 $25,000 to ~      1
## 15    15 Ontario  15 to 24~         0 Male      5 Less than $~      0
## 16    16 Ontario  15 to 24~         0 Fema~     2 <NA>             NA
## 17    17 Ontario  15 to 24~         0 Fema~     2 $75,000 to ~      3
## 18    18 Ontario  15 to 24~         0 Fema~     2 Less than $~      0
## 19    19 Ontario  15 to 24~         0 Fema~     5 $25,000 to ~      1
## 20    20 Ontario  15 to 24~         0 Fema~     5 Less than $~      0
## # ... with 2 more variables: cps19_votechoice <chr>, cps19_fed_gov_sat <chr>
```

8.0.2 GSS data demonstrating

```
## # A tibble: 20 x 9
##    caseid age_group sex   income_level province education_level   edu income
##     <dbl> <fct>     <chr> <chr>        <chr>    <chr>           <dbl>  <dbl>
##  1      1 55 to 64~ Fema~ $25,000 to ~ Alberta  High school di~     2      1
##  2      2 55 to 64~ Fema~ $75,000 to ~ Ontario  Bachelor’s deg~     6      3
##  3      3 65 to 74~ Fema~ $25,000 to ~ Manitoba Less than high~     1      1
##  4      4 65 to 74~ Fema~ Less than $~ Quebec   University cer~     5      0
##  5      5 25 to 34~ Male  Less than $~ British~ Bachelor’s deg~     6      0
##  6      6 35 to 44~ Male  $100,000 to~ Ontario  University cer~     7      4
##  7      7 45 to 54~ Fema~ $75,000 to ~ Saskatc~ High school di~     2      3
##  8      8 35 to 44~ Male  $100,000 to~ Ontario  University cer~     7      4
##  9      9 15 to 24~ Fema~ Less than $~ Ontario  University cer~     5      0
## 10     10 45 to 54~ Fema~ $25,000 to ~ Ontario  College/CEGEP/~     4      1
## 11     11 75 years~ Fema~ $25,000 to ~ Nova Sc~ High school di~     2      1
## 12     12 25 to 34~ Male  $25,000 to ~ New Bru~ High school di~     2      1
## 13     13 15 to 24~ Fema~ Less than $~ Alberta  Less than high~     1      0
## 14     14 45 to 54~ Male  $50,000 to ~ Nova Sc~ High school di~     2      2
## 15     15 25 to 34~ Male  $75,000 to ~ Nova Sc~ University cer~     5      3
## 16     16 75 years~ Fema~ $25,000 to ~ Nova Sc~ College/CEGEP/~     4      1
## 17     17 65 to 74~ Fema~ Less than $~ Ontario  Trade certific~     3      0
## 18     18 55 to 64~ Fema~ Less than $~ New Bru~ Bachelor’s deg~     6      0
## 19     19 45 to 54~ Fema~ $100,000 to~ Quebec   University cer~     7      4
## 20     20 25 to 34~ Male  Less than $~ Manitoba Less than high~     1      0
## # ... with 1 more variable: age_level <dbl>
```

8.0.3 Logistic model results and coefficients

Table 6: Coefficients for Logistic Model - Liberal Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6895	0.1572	-10.7466	0.0000
age_level	0.0434	0.0084	5.1873	0.0000
sexMale	-0.0089	0.0280	-0.3173	0.7511
edu	0.1208	0.0081	14.9472	0.0000
income	0.0267	0.0089	3.0051	0.0027

Table 7: Coefficients for Logistic Model - Conservative Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3955	0.2064	-6.7598	0
age_level	0.1088	0.0088	12.3691	0
sexMale	0.4436	0.0290	15.3023	0
edu	-0.1133	0.0084	-13.5562	0
income	0.1333	0.0093	14.2823	0

Table 8: Coefficients for Logistic Model - Bloc Quebecois

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-17.7503	3.8876	-4.5658	0.0000
age_level	0.2390	0.0210	11.3925	0.0000
sexMale	0.2307	0.0651	3.5468	0.0004
edu	-0.0809	0.0184	-4.3977	0.0000
income	0.0122	0.0220	0.5530	0.5803

Table 9: Coefficients for Logistic Model - New Democratic Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8090	0.1354	-5.9741	0.000
age_level	-0.2704	0.0111	-24.3620	0.000
sexMale	-0.3558	0.0382	-9.3089	0.000
edu	0.0237	0.0105	2.2565	0.024
income	-0.1262	0.0118	-10.7329	0.000

Table 10: Coefficients for Logistic Model - Green Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8980	0.2145	-8.8478	0.0000
age_level	-0.1365	0.0136	-10.0380	0.0000
sexMale	-0.0882	0.0467	-1.8901	0.0588
edu	0.0422	0.0132	3.2075	0.0013
income	-0.1040	0.0148	-7.0119	0.0000

Table 11: Coefficients for Logistic Model - People's Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8900	0.1320	-21.9014	0e+00
age_level	-0.2231	0.0261	-8.5392	0e+00
sexMale	0.6007	0.0875	6.8646	0e+00
edu	-0.1002	0.0248	-4.0414	1e-04
income	-0.1194	0.0288	-4.1430	0e+00

Table 12: Coefficients for Logistic Model - Other Parties

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.3426	0.2582	-20.6885	0.0000
age_level	0.0855	0.0465	1.8385	0.0660
sexMale	0.3270	0.1549	2.1112	0.0348
edu	0.0178	0.0443	0.4030	0.6869
income	-0.0885	0.0505	-1.7523	0.0797