2025 Canadian Federal Election Prediction

GROUP 199: Chenyang Pan, Yuanyuan Pan, Jingchi Wei, Zhuowen Dai

November 5, 2021

Introduction

Canada is a democratic country that usually holds Canadian federal election every 4 years. It is divided into 338 electoral districts, also known as ridings [7]. Canada's electoral system is a "first-past-the-post" system [7]. All Canadians over 18 years old get to vote to elect a member of parliament in their electoral area to represent them in the House of Commons [7]. The candidate who receives the most votes in a riding is the winner [7]. The party that has the most MPs forms the federal government and the leader of that party becomes the prime minister [8]. For example, in 2021 Canadian federal election, there are 338 seats in the House of Commons in total and there were 160 liberal members of parliament elected in the House of Commons [8]. Liberal party had the most MPs and thus, Justin Trudeau, as the leader of the liberal party, became prime minister in Canada [8]. Election is important because it actually is the hallmark of the democracy because citizens get to express their political choices. The party we are going to do an analysis on is the liberal party. The liberal party stands for individual rights, freedom of speech, press, and religion, democracy, and so on [9]. Since the liberal party is the oldest and longest political party in Canada, it has a huge influence on the country [9]. In addition, it has a very large chance to win in the next election. As a result, it is important and interesting to predict whether or not the liberal party will still be the federal government after next selection.

To study this topic, we will use two sets of data. Firstly, we will use 2019 Canadian election phone survey data. This set of data contains 4021 survey records/observations and it asked about which party they will most likely to choose in the next election. However, due to the limitation of this data set, there are some differences between the population from the survey data and general Canadian population. As a result, we use an an additional data set called General social survey data to better represent the general population. This data set contains 20602 survey records.

According to different researches, a lot of factors including gender, age, level of education can have significant influence on voting. As a result, we think that province, age, and level of education will influence whether or not liberal party will win in next election. Additionally, since the survey asked whether or not they are satisfied with Justin Trudeau, we will also use this response as a factor that could influence the final prediction. Our research question is whether or not liberal party will win in next election. In order to do this, we will use the factors "province", "age", "level of education", "will vote for liberal or not" from the 2 datasets mentioned above. ***According to the results from past elections, We predict that liberal party will have a huge chance of winning in next election.***

Terminology

Riding: also known as electoral district in Canada and is a geographical constituency, which elects members of parliament to the House of Commons in every election. There are 338 electoral districts in Canada [7].

First-past-the-post: the candidate who gets the most votes in each riding is the only winner in that riding [7].

Member of parliament: the representative of people who live in their constituency. He or she represents their electoral district [7].

House of Commons: it is a democratic body whose members are known as members of parliament. All MPs are elected by citizens in the federal election. The role of the House of Commons includes approving new laws and taxes, debating of the issues of the day, etc [7].

Liberal Party: one of the political parties of Canada. It stands for democracy, individual rights, etc [9].

Data

According to Statistics Canada, the Canadian General Social Survey census data uses random digital dialing (RDD) and computer-assisted telephone interview to collect demographic information from a random sample of Canadians. The interview is usually about 40 minutes. The completion could take up 6 to 12 months[4]. The election survey applies dual-mode, two-wave data collection with a rolling cross-section. It collects data with non-probability online survey and RDD internet survey according to a Canadian Election Study[5].

First, all the missing entries are removed in the census data. Only data with citizenship and age above 18 are kept since those non-citizens or those aged below 18 cannot vote. Then the variables mentioned above are kept in the data set.

Similarly, we keep only variables that we use for the study when cleaning survey data and they are age, gender, citizenship, the education level, province, satisfaction with the government under Justin Trudeau and the party that is most likely to win in the local riding. Vote choices are also kept. Then all the entries with missing values are removed.

By the way, almost all the answers to survey questions are represented with numbers. For example, question3 asking for gender in the survey provides five options, male, female, other, 'refused' and 'does not know', and they are 1, 2, 3, -8 and -9 in the data set. Because 'refused' and 'does not know' are useless and thus deleted. To predict whether the liberal party will win the election, a new variable named "liberal vote" is created. If the participants choose the liberal party, the value is "Yes", and all the other votes are "No". Then we replace these numbers with the options for better reading and understanding. Again, non-citizens and non-adults are deleted.

<Here is a resource for grabbing the CES2019 data: https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-canadian-election-study>

Table of variables for Census Data

Variable Name	Description	Type
Age range	the range of the age	categorical
Sex	the sex of the person	categorical binary
Education	the education level of the people	categorical
Province	the current province stayed	categorical
Whether citizen	is the participant citizen	categorical

The above table describes the important variables from the census data; as we can see, all the variables are categorical.

Table of variables for Survey

Variable Name	Description	Type
Age range	the range of the age	categorical
Gender	the Gender of the person	categorical binary
Education	the education level of the people	categorical
Province	the current province stayed	categorical
Whether citizen	is the participant citizen	categorical
Satisfied with Justin	how well are people satisfied with Justin	categorical
Is Justin winning	do people think Justin will win	Categorical Binary
Vote liberal	will people vote liberal	Categorical Binary

The above table describes the important variables from the survey data; as we can see, same as census, all the variables are categorical.

We will use these variables to predict whether the liberal party will win the election. It is widely accepted that different genders think in different ways. As age increases, people tend to think about more aspects of voting and more comprehensively. So, age affects decision on voting for the liberal party or not. citizenship, education level, and province can also differentiate people in the vote decision. Moreover, participants who are not so satisfied with government performance under Justin Trudeau will probably vote for other parties. The herd mentality may encourage participants to vote for the party that is most likely to win in participants' local riding.

Graphical Analysis

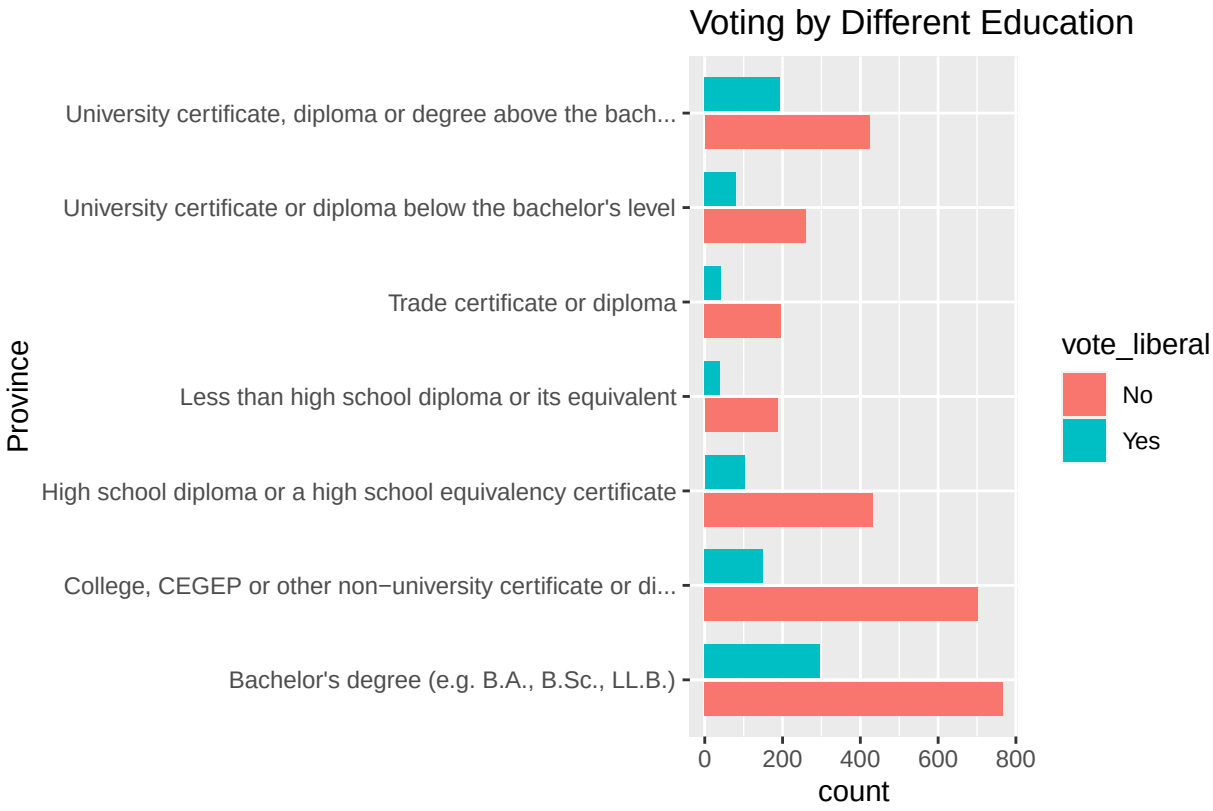

Figure 1

Figure1 shows that among liberal party voters, education level of bachelor's degrees is the majority. Next come those with diplomas above the bachelor's degree, followed by those with non-university certificates. Participants with certificates below the bachelor's including high school or less than high school diploma, and trade certificate take up the minority liberal party voters. There are many choices besides the liberal party so, votes for the liberal are not minority though in each education level, liberal party voters are less than half those who vote for other parties. So, education level does affect decisions on voting. It is better to do post-stratification

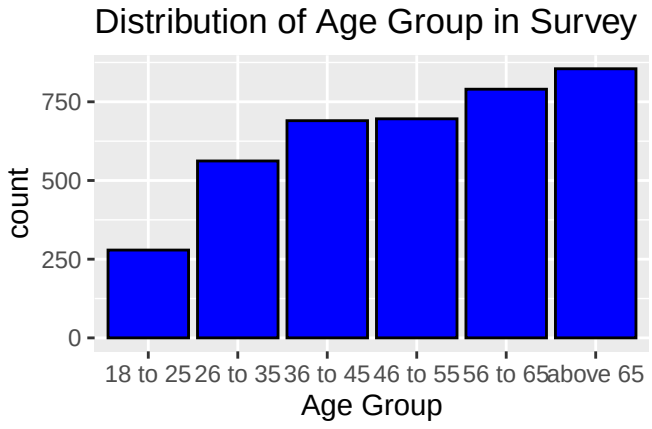

Figure 2

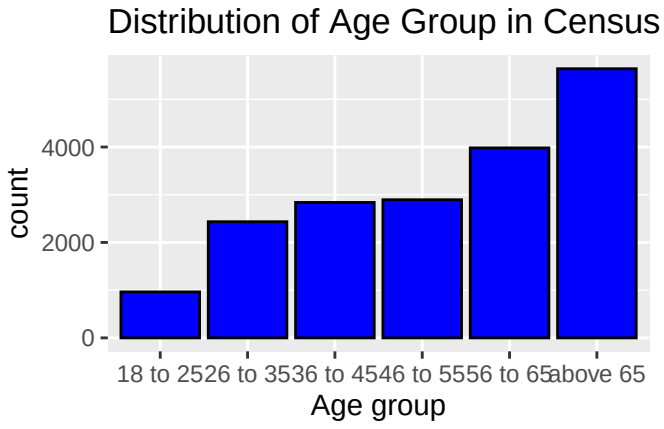

Figure 3

According to figure2 and figure3, most of the participants of the survey are distributed in the middle of age

groups. While in the census data, most of the people are aged from 56 and over. Moreover, the number of people in different age groups in census data is extremely large and it can reach up to more than 5000 of age above 65 while the number of people in each age group of the survey study is all below 800. This indicates that the actual population distribution differences in age groups are much larger than it looks. So, the age distributions are quite different between survey data and census data.

The survey data shows that more than a thousand participants completed non-university certificates or university diploma of bachelor's degree. It is the same for census data but those with high school certificate are much higher. In survey study, a few of the participants received only high school education or degree above the bachelor's, which is the same for census data. In both data, few of the participants received certificate below the bachelor's or failed high school. The education conditions are not the same between the survey data and the census data.

As for the province distribution, it is the same for both survey data and Canadian census data that there are large quantities of participants from Quebec and Ontario, and people from the rest of the provinces are almost equally distributed. The quantity of participants from British Columbia are bit smaller than that of Ontario and Quebec in the survey but in census data, there are about 2500 fewer British Columbians than Ontarians and Quebec

Generally building a multi-level model to predict the results of the election is a good choice since the demographic characters of participants differs from each other. The second level is age group, and the individual level is based on respondents' province and education. Because the survey data cannot reflect the real characteristics of age groups and education level of the population, it is better to do post-stratification on these variables.

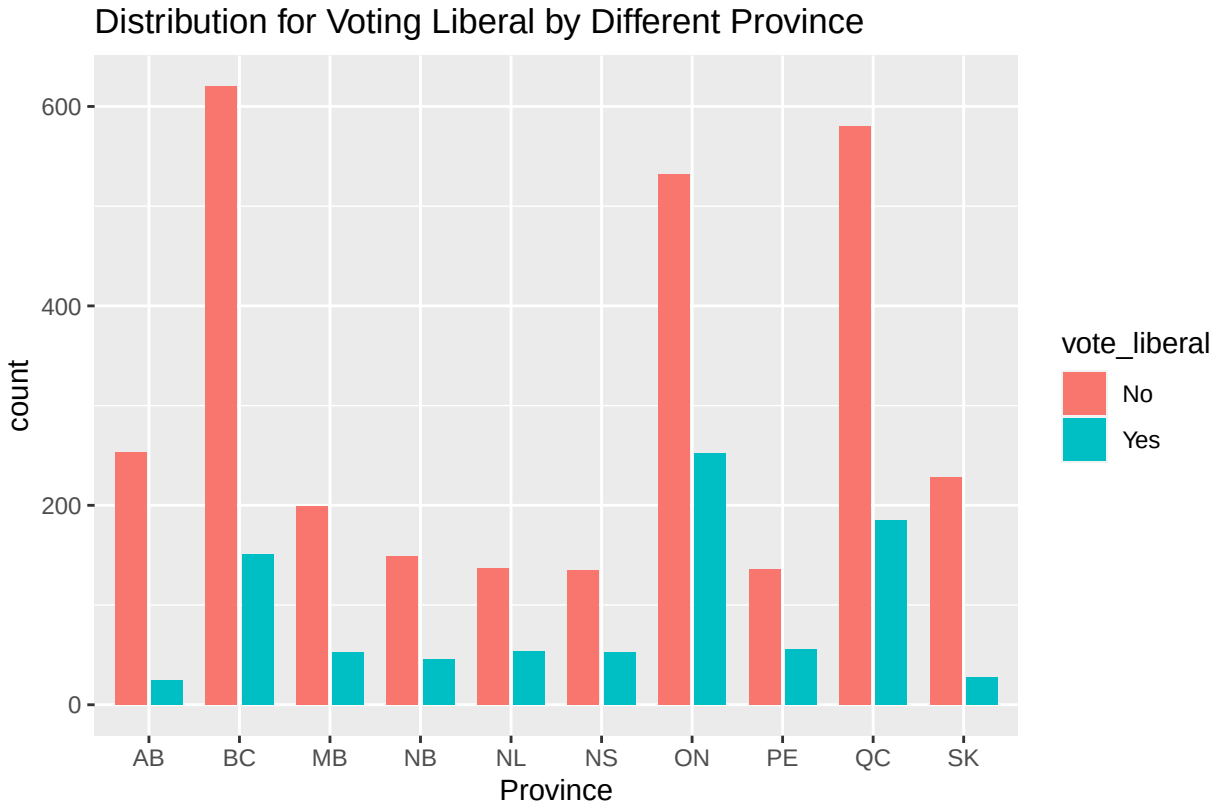

Figure 4

Distribution of Voting Liberal by Different Province bar plot(figure4) shows that Ontario votes the most for the liberal party compared to other provinces. Next is Quebec. Then about one-fifth of British Columbia are liberal. In Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island,

liberal voters take up less than one-third of the total participants. In Alberta and Saskatchewan, only about ten percent of the survey participants voted or will vote for the liberal party. Though for each province, the proportion of liberal party voters are much smaller than that of those who vote other parties, they are not minority because non-liberal party voters are distributed in other five major parties. The survey study demonstrates the quantity of non-liberal supporters is about three times of those who vote liberal. And there are more males than females while in the census data, the situation is just the opposite, but the difference is small.

All analysis for this report was programmed using `R version 4.0.2`.

Methods

In the purpose of predicting the result of Canada Election, the methodology we will use to achieve our goal is Multilevel Regression Post-stratification (MRP). By using MRP, we are able to make use of various respondent types, which can improve the accuracy of our prediction. There are two stages in MRP; the first stage is to estimate a survey response by using a multilevel model, which is capable of handling clustered data and can have its parameters more than one level. The second stage is poststartification, it best minimizes the differences between survey samples and the actual population and weights the estimates for each respondent type in percentages of each type.

Model Specifics

The model we will be using to model the proportion of voters who will vote for liberal party is Multilevel Logistic Regression Model with random intercept. There are two levels in this model; the individual level, which is level one, will be based on respondents' education and their current living province; the group level, which is level two, is the age group.
The random intercept model is:

$$log(\frac{p_{ij}}{1-p_{ij}}) = Y_{ij} = \alpha + \beta_1 x_{ij} + a_j + \epsilon_{ij}$$

where:

$$\epsilon_{ij} \sim N(0, \sigma_e^2)$$

and

$$a_j \sim N(0, \sigma_a^2)$$

Where Y is the response variable, which is whether or not each respondent will vote liberal, and p represents the probability of voting liberal of each respondent, which is also our parameter of interest. In this model, the intercept, which used to be represented as β_0, of the regression line, is $\alpha + a_j$, where a_j is the impact of the group on our response variable Y, and α is the baseline which is also intercept mean. $\beta 1$ is the fixed slope in our model. Unlike $\beta 1$, a_j is not fixed and will be changing in our model since it follows normal distribution, and its variation will lead to the change of intercept. At last, ϵ is the error term which follows the normal distribution.

The reason we choose this model is because our response variable vote_liberal is binary, which has only two possible observations - 'Yes' or 'No', and based on the bar plot of age range in graphical summary session, the level two variable age range has different tendency for different data point. As we have mentioned before, age is a big factor that can significantly impact the voting results, and as a person getting older, he/she has more tendency to vote due to caring more about the politic issues. Therefore, there is a linear relationship between age and the voting results, and thus our assumption of linearity won't be violated.
Moreover, the assumption of variance is probable to be violated since the our chosen variables have sub-factors that can have various outcomes.
Finally, since the a_j within our model is following normal distribution, $a_j \sim N(0, \sigma_a^2)$, and all the slopes are fixed, whereas the intercept is changing based on the variation of a_j, and the error also follows the normal distribution, thus the residuals are likely to be normally distributed. However, deviation is highly possible to exist yet the overall pattern won't be affected.
As a whole, our model is appropriately chosen and ready to be used in later analysis.

Post-Stratification

Post-Stratification is a method to rearrange or accommodate the weight of data collected from the survey, which is the CES 2019 phone survey in this study experiment, based on the attributes from the data of target population, which is the GSS census data.The reason we do a Post-Stratification is that, the choices of some factors in the data of survey may be biased and may differ from the population, for example, in above data section, we had two bar plots comparison based on the age of people in the data from survey and the data from the census, what has been found was that in the census data, there is a significant difference between the people above 65 than others. However, this significant difference has not been shown in data from the survey. In such case, Post-stratification would be a good way to accommodate this issue.[10]

In order to estimate the proportion of voters for voting liberal party, we first count number of people in every combination of the age range, education and province from our census data, then we would calculate an estimate of proportion of voting liberal for each combination by using the model we have established earlier with the survey data. Lastly for each combination, we use the number of this combination to multiply with the proportion of voting liberal for this combination, and sum up this calculation for every combination and divided by the summation the number of every combination. Eventually, the predicted proportion of voters for voting liberal party is acquired.[10]

The Post Stratification equation

$$\hat{y}^{PS} = (\sum_{i=1}^{n} N_i \hat{y}_i) / \sum_{i=1}^{n} N_i$$

In this equation of Post-stratification, $\hat{y}^{PS}$, the parameter of interest, represents the proportion of voters for voting liberal party we want to predict, N_i represents the number of one single type of combination, $\hat{y}_i$ represents one estimate of the proportion of voting liberal based on that combination.

All analysis for this report was programmed using `R version 4.0.2`.

Results

The final model of our Multilevel Regression with random intercept is:

$$log(\frac{p_{ij}}{1-p_{ij}}) = Y_{ij} = \alpha + \beta_{j[i]}^{education} + \beta_{j[i]}^{province} + \beta_{[i]}^{age} + a_j + \epsilon_{ij}$$

where all the slope are fixed, yet the intercept is vary as a_j changing. Below is a table displaying the different intercept based on each age range:

Age Range	Intercept
18 to 25	-2.014483
26 to 35	-2.080883
36 to 45	-2.113921
46 to 55	-2.139692
56 to 65	-2.035327
above 65	-1.859879

By examining above table, the intercept is gradually increasing as the age range getting older, yet the overall intercept is negative. Below is another table displaying the slope, which is fixed for every age range, of the model:

Predictors	Slope (coefficient)
educationCollege, CEGEP or other non-university…	-0.5650457
educationHigh school diploma or a high school equivalency…	-0.4571145
educationLess than high school diploma its equivalent…	-0.6129801
educationTrade certificate or diploma	-0.5870221
educationUniversity certificate or diploma below…	-0.2147769
educationUniversity certificate, diploma or degree above…	0.1336748
provinceBC	0.8317763
provinceMB	0.9488381
provinceNB	1.073056
provinceNL	1.418033
provinceNS	1.354195
provinceON	1.475085
provincePE	1.384192
provinceQC	1.132071
provinceSK	0.22702

Below is another table displaying the summary of the multilevel logistic regression model with random intercept:

Variable	Estimate Std.	Error	p-value
(Intercept)	-2.0415	0.2248	< 2e-16
educationCollege, CEGEP or other non-university…	-0.5650	0.1152	9.32e-07
educationHigh school diploma or a high school equivalency…	-0.4571	0.1325	0.000559
educationLess than high school diploma or its equivalent	-0.6130	0.1943	0.001607
educationTrade certificate or diploma	-0.5870	0.1857	0.001572
educationUniversity certificate or diploma below…	-0.2148	0.1499	0.152029
educationUniversity certificate, diploma or degree above…	0.1337	0.1125	0.234604
provinceBC	0.8318	0.2298	0.000294
provinceMB	0.9488	0.2619	0.000292
provinceNB	1.0731	0.2707	7.35e-05
provinceNL	1.4180	0.2661	9.88e-08
provinceNS	1.3542	0.2668	3.85e-07
provinceON	1.4751	0.2245	5.03e-11
provincePE	1.3842	0.2647	1.71e-07
provinceQC	1.1321	0.2275	6.51e-07
provinceSK	0.2270	0.2913	0.435712

By examining above model summary table, we can clearly see that most of the p-value of the variables are very

significant, yet there are three exceptions: educationUniversity certificate or diploma below the bachelor's level, educationUniversity certificate, diploma or degree above the bachelor's level, and provinceSK all have a p-value higher than 0.1, and provinceSK even have a p-value around 0.4. This shows that these three factors won't make much impact on the voting for liberal. The p-value for educationLess than high school diploma or its equivalent and educationTrade certificate or diploma are all less than 0.002, which is quite significant even though they are little bit less significant when comparing to other significant factors. Nevertheless, although we do have insignificant factors exist within our model, the model overall is very significant.
This result is reasonable because University students or people who have higher education level can be busy at academic work or occupation, and they might won't pay much attention on the election. Since NDP has been in power in Saskatchewan for 47 years, this province has less impact on which party is going to win, thus it won't be significant since our parameter of interest is whether liberal is going to win.
As a whole, our model is appropriate based on this reasonable result, and the overall factors of our model are significant.

Result for Post-stratification

$$\hat{y}^{PS} = 0.2249677$$

The predicted proportion of voting liberal acquired from the Post-stratification calculation is 0.2249677. From this result, we predict that there are 22.49677% of Canadian citizens will vote for liberal. According to the last Canada election 2015, Liberal party won the election with 54.4% of seats and Liberal party won again in 2019 with 46.4% of seats[6]. However, the percentage of seats for Liberal parties in 2019 had dropped about 8% from 2015. Also according to the Canadian election results by party provided by Simon Fraser University, in 2015 and 2019, the liberal party won in the election and its percentages of popular vote are all above 30%, which indicates that the probability of voting for the liberal should be at least 30 for the liberal to win the election[6]. Whereas, our predicted proportion of voting liberal acquired from the Post-stratification is 0.2249677. The past election trend might provide information for us that the seat percentage of Liberal parties is in a dropping state which means that our prediction may be reasonable and supported by the trend of the election. Therefore, we conclude that Liberal party would not be successfully elected in 2025.

Conclusions

In this study, we chose factors "province", "age", and "level of education" as independent variables that can influence Canadian citizens' voting. We predicted that liberal party will win in next election. In specific, we selected four factors from the survey data, which are "province", "age", "level of education", and "vote for liberals or not". We used a multilevel logistic regression model to fit these variables. In this model, we set "vote for liberal party or not" as the binary dependent variable, "age range" as the level 2 independent variable, and "province" and "level of education" as the level 1 independent variables. In this model, we had random intercepts. Since the data from the CES 2019 phone survey does not have a large enough sample size, the results can be biased and cannot be generalized to the general population in Canada. As a result, we used post-stratification to accommodate the weight of data collected from the survey based on the data collected in the GSS census data. **The main results found that the multilevel logistic regression model is significant as a whole with only three sub-factors being insignificant. This result confirmed that factors "age", "level of education", and "province" can influence the voting. In this case, we found that the probability of liberal party winning in next election is 0.225. This means that liberal party will have a low chance winning in next election, which rejected our hypothesis that liberal party will win in next election.**

There are a few limitations in this study. First off, since the census was conducted in 2007, the data is out of date. Additionally, there are many other factors that are not included in the data that can also influence the voting, but we only chose three factors. Thus, our model is limited to only three factors. Last but not least, some of the assumptions made by the multilevel logistic regression model was not satisfied, such has the normality assumption.

For future studies, researchers should first find an updated census as their data. Additionally, their models could include more factors that might possibly influence the voting. Lastly, they could also try other models that can fully explain the dependent variable.

In conclusion, this study found that the factors "age", "level of education, and"province" can influence the voting and the probability estimated is 0.225, which rejected our hypothesis that liberal party has a huge chance winning in next election. The model we built was appropriate and was used properly in this study.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Statistics Canada. (2013) **General Social Survey: An Overview, 2013** Statistics Canada. (https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm)

5. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John. (2020) **2019 Canadian Election Study - Online Survey** Canadian Election Study. (https://doi.org/10.7910/DVN/DUS88V), Harvard Dataverse, V1.

6. Heard, A. (n.d.). *Canadian election results by party 1867 to 2019. Canadian Election Results*: 1867-2019. Retrieved November 5, 2021, from [https://www.sfu.ca/~aheard/elections/1867-present.html].

7. Canada, E. (2021, October 27). *Home.* – Elections Canada. Retrieved November 6, 2021, from [https://www.elections.ca/content2.aspx?section=faq&document=fedelect&lang=e].

8. Wikimedia Foundation. (2021, November 5). *2021 Canadian federal election*. Wikipedia. Retrieved November 6, 2021, from [https://en.wikipedia.org/wiki/2021_Canadian_federal_election].

9. Wikimedia Foundation. (2021, October 26). *Liberalism*. Wikipedia. Retrieved November 6, 2021, from [https://en.wikipedia.org/wiki/Liberalism].

10. Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). *Forecasting elections with non-representative polls*. International Journal of Forecasting, 31(3), 980-991.