Using Multilevel Regression with Post-stratification to Predict the Result of the Next Federal Election

STA304 - Assignment 2

Sizhang Lyu

November 24, 2022

Introduction

As federal election day approaches every once in five years, people form long queues outside the voting station regardless of the weather, and that is the magic of elections. People are enthusiastic about the federal election because it is one of the main ways everyone can participate in democracy. The result of the elections not only determines parliament's seating and the next government but, more importantly, everyone's future lifestyle. Forecasting of election results has been happening in statistical, economic, and political sciences for about half a decade, and its influences have spread beyond scholars to public media (Dowding, 2020). Epistemic gains from forecasting the elections are essential to politicians and the general public, so increasing attention gathered by election forecasting may not sound surprising (Campbell & Mann, 2016).

The Canadian political system consists of the governor, the Senate, and the house of commons representing the King of Canada (Canada, 2022). Three hundred thirty-eight seats are in the house of commons, which citizens elect its members, and thus form the government. One hundred five seats are in the Senate, which the governor general appoints its members on the prime minister's recommendation. The maximum time between federal general elections is five years, as set by the constitution acts in 1867 and 1982. The party's leader with the most elected house of commons representatives will generally form the government, and the governor general will invite the party's leader to become the prime minister. On the other hand, the party winning the second most number of seats is often considered an official opposition. Electoral, or ridings, are geographical divisions on which representation in the House of Commons is based, and each electoral district represents a seat in the house of commons, so there are in total of 338 electoral districts. Canada's electoral system is called the "single-member plurality" system, which means that in every electoral district, the nominee with the highest number of votes earns a position in the house of commons, and there can be any number of contenders running for each district. At the same time, each party can only have one nominee running for that district, and similarly, each nominee can only run for one electoral district (Canada, 2022).

Some terminologies may help readers better understand the context of the problem, and they are explained as follows. Electoral district or riding refers to the geographical unit of a seat in the house of commons represents. A political party is a political organization whose members share the same ideas that people can vote for in elections. A federal system is a political system in which two levels of government control the same territory. A parliamentary seat refers to a position as an elected member of the parliament. Finally, a runner-up party wins the second place of seats in an election.

When looking at the outcome of the 2019 election, tendencies for voting favor a particular party in various provinces, at each age level, different sex, et cetera, materializes. For example, almost all electoral districts in Alberta and Saskatchewan voted for the conservative, while in the Atlantic region, identical circumstances transpired for the liberal party (Official Voting Results, 2019). On the other hand, the intention to vote is significantly higher among the elderly than among the youth. Furthermore, seniors tend to vote for the conservative more than the youth. From the result of the previous election, we can form an initial hypothesis that the chance that the liberal party will win the next federal election still outstands. This report will predict the probability that an individual will vote for a particular party in the next Canadian federal election by examining features, such as income, gender, marital status, age, province, at an individual and provincial level using multi-level regression poststratification.

Data

The data employed in this study comes from a series of surveys compiled during the 2019 election period. The surveys were conducted using both English and French, and there were two phases in the gathering process. First, there were 4021 Canadian citizens interviewed during the election campaign (Stephenson, 2020). In addition, respondents were either mailed or called after the election based on their preferences and were asked to complete a post-election survey, of which 2889 have done so. The sample in the dataset represents the adult population of Canada. Interviewers typically make phone calls during the daytime and the evening to ensure the completeness of each survey, and interviewers may call back and schedule appointments with the respondents to enhance completeness. Post-election surveys may take more effort than the campaign stage as more proportion of the calls requires six or more attempts. Additionally, the design of careful attention to the number and timing of callbacks increases both the response rate and the representativeness of the sample. Overall, among the 95424 samples, the collection process achieved a response rate of 5.6%, a completion rate of 7.5%, and a participation rate (completion plus refusal) of 8.3% (Stephenson, 2020).

Some cleaning was needed in order for the data to fit our problem. For the GSS data (data from the most recent census), all the data is cleaned, so we only need to filter out the required variables, which are the age, sex, province, and family income range of the respondent. On the other hand, for the CES data (data from the 2019 federal election), the cleaning procedures are as follows. First, since the categorization of income range differs between the GSS and CES datasets, we need to categorize the numerical income data manually into categories of Less than $25,000, $25,000 to $49,999, $50,000 to $74,999, $75, 000 to $99, 999, $100,000 to $124, 999, $125,000 and more, which is consistent with the GSS data. Second, as the data representing each province is in number, we need to adjust them accordingly, with numbers 1 to 12 corresponding to provinces of Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia, Northwest Territories, Yukon respectively. Likewise, we need to adjust 1 to male, 2 to female, and 3 to others for variable gender. Then we need to filter out observations that refused to answer the questions or were unable to provide an answer by omitting all observations with values -8 and -9. Also, we need to filter out respondents who were not Canadian citizen, did not vote in the elections, or was under the age of 18 so that the remaining respondents were eligible. Then, we need to calculate each respondent's age by subtracting 2019 from their birth year. Finally, we need to select our desired variables: age, income, the party voted, and province.

Some critical variables will be used in further analysis. For example, party_voted is a categorical variable indicating the party the respondent voted for in the federal elections. Age is a numerical variable denoting the age of the respondent. Province is a categorical variable representing the respondent's province of residence. Gender is a categorical variable illustrating the gender of the respondent in three classes: male, female, and others. Finally, income_range is a categorical variable describing the respondent's household income, with six categories ranging from below 25000 dollars to above 125000 dollars. The same variables can be found in both the GSS and the CES datasets, while the variable party_voted cannot be found in the GSS dataset.

Both table 1 and table 2 are generated using the `vtable` package in `R`. Table 1 is a summary of the GSS data. The table shows that the average age of all the participants in the census is 52.19 years old, there are more female than male participants, and families earning between 25000 and 50000 and families earning more than 125000 occupy the most proportion.

On the other hand, table 2 is a summary table for the CES data. The table shows that the average age of all the participants in the census is 52.356 years old, over 30 percent of people voted for the liberal or the conservative party, there are more males than females in the sample, and families earning below 25000 and families earning more than 125000 occupy the most proportion.

Table 1: Summary Table of the GSS Data

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
age	20602	52.19	17.747	15	37.3	66.775	80
sex	20602						
... Female	11203	54.4%					
... Male	9399	45.6%					
income_family	20602						
... $100,000 to $ 124,999	2158	10.5%					
... $125,000 and more	4707	22.8%					
... $25,000 to $49,999	4345	21.1%					
... $50,000 to $74,999	3696	17.9%					
... $75,000 to $99,999	2921	14.2%					
... Less than $25,000	2775	13.5%					

Table 2: Summary Table of the CES Data

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
party_voted	2461						
... Bloc Québécois	139	5.6%					
... Conservative Party	769	31.2%					
... Green Party	263	10.7%					
... Liberal Party	788	32%					
... NDP	460	18.7%					
... People's Party	42	1.7%					
gender	2461						
... Female	1045	42.5%					
... Male	1415	57.5%					
... Other	1	0%					
age	2461	52.356	16.745	18	39	65	95
income_range	2461						
... $100,000 to $ 124,999	260	10.6%					
... $125,000 and more	588	23.9%					
... $25,000 to $49,999	269	10.9%					
... $50,000 to $74,999	364	14.8%					
... $75,000 to $99,999	288	11.7%					
... Less than $25,000	692	28.1%					

Fig 1: Relation Between Gender and Votes

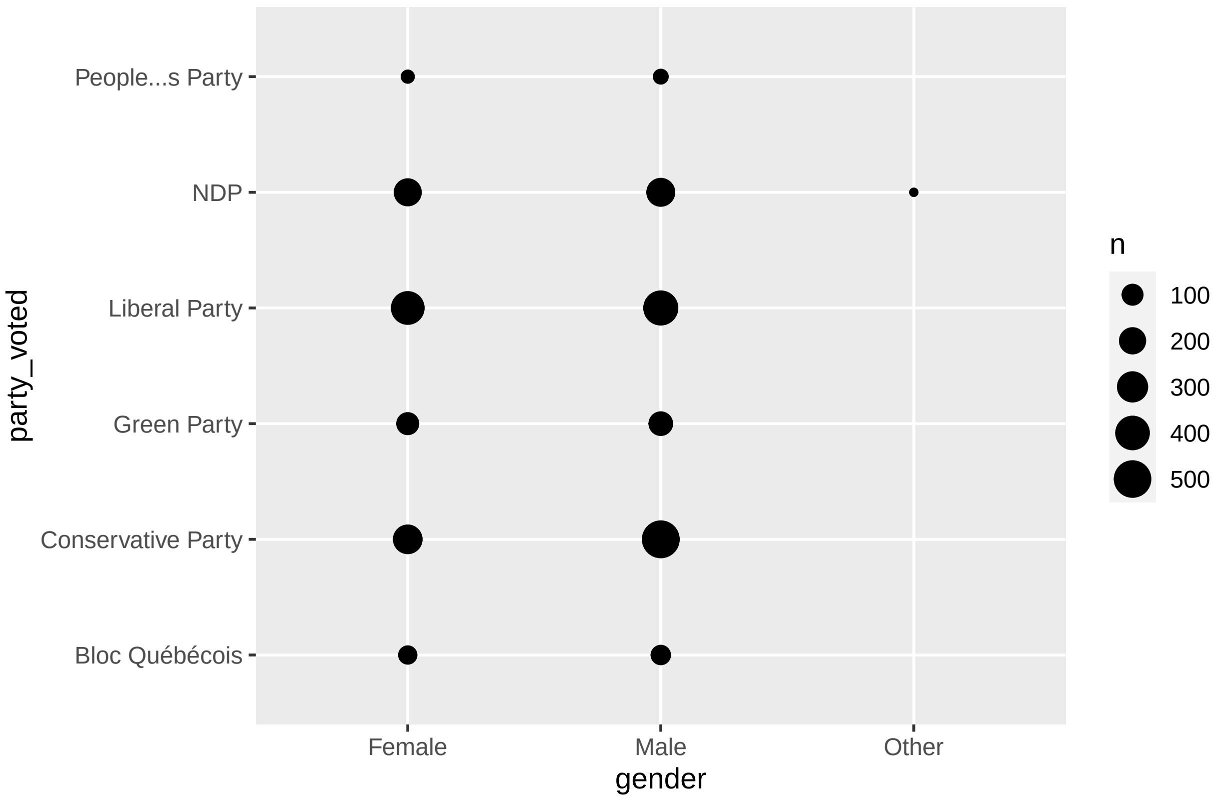


Fig 2: Relation Between Age and Votes

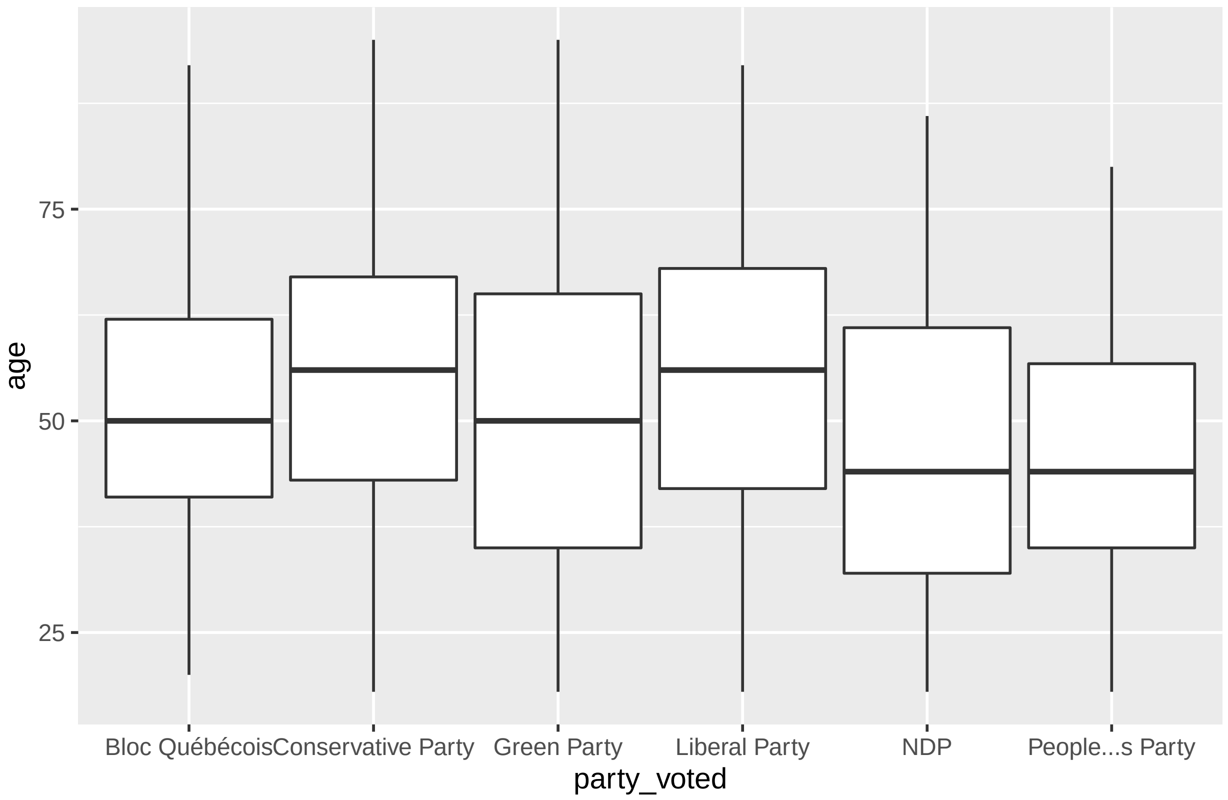
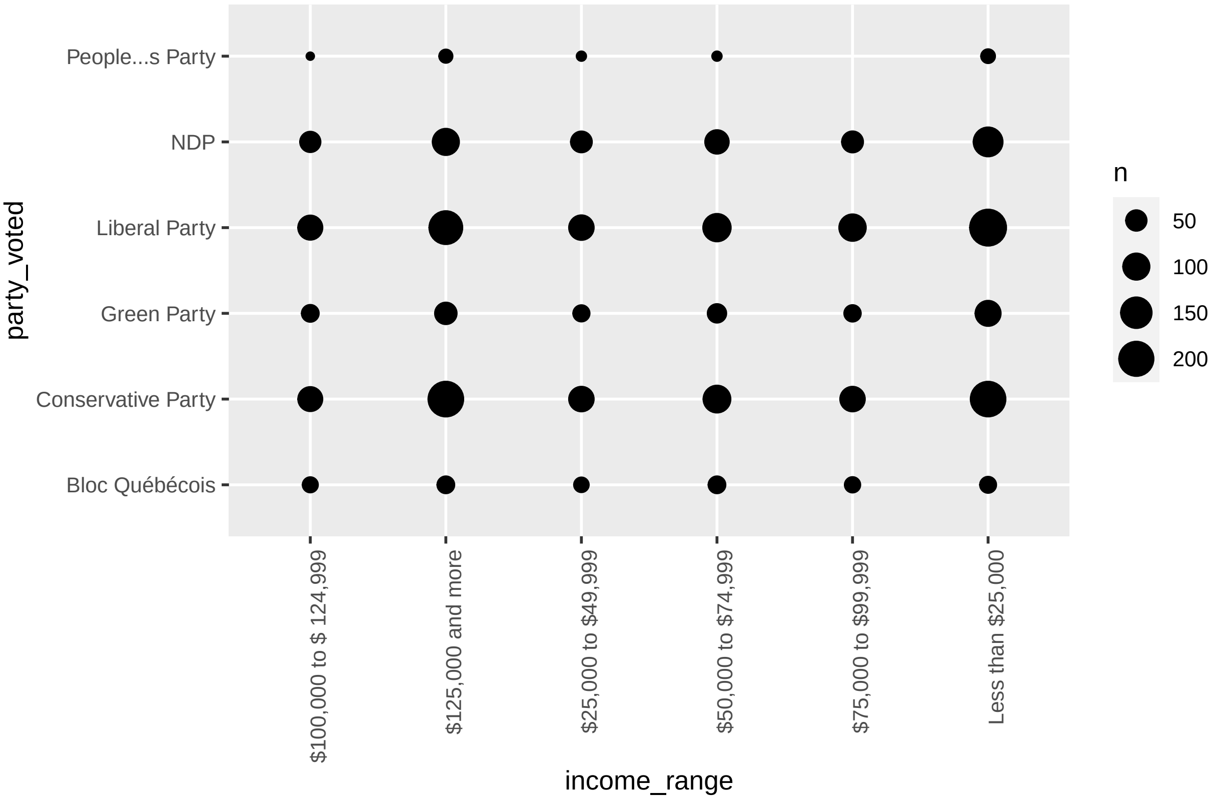


Fig 3: Relation Between Income and Votes



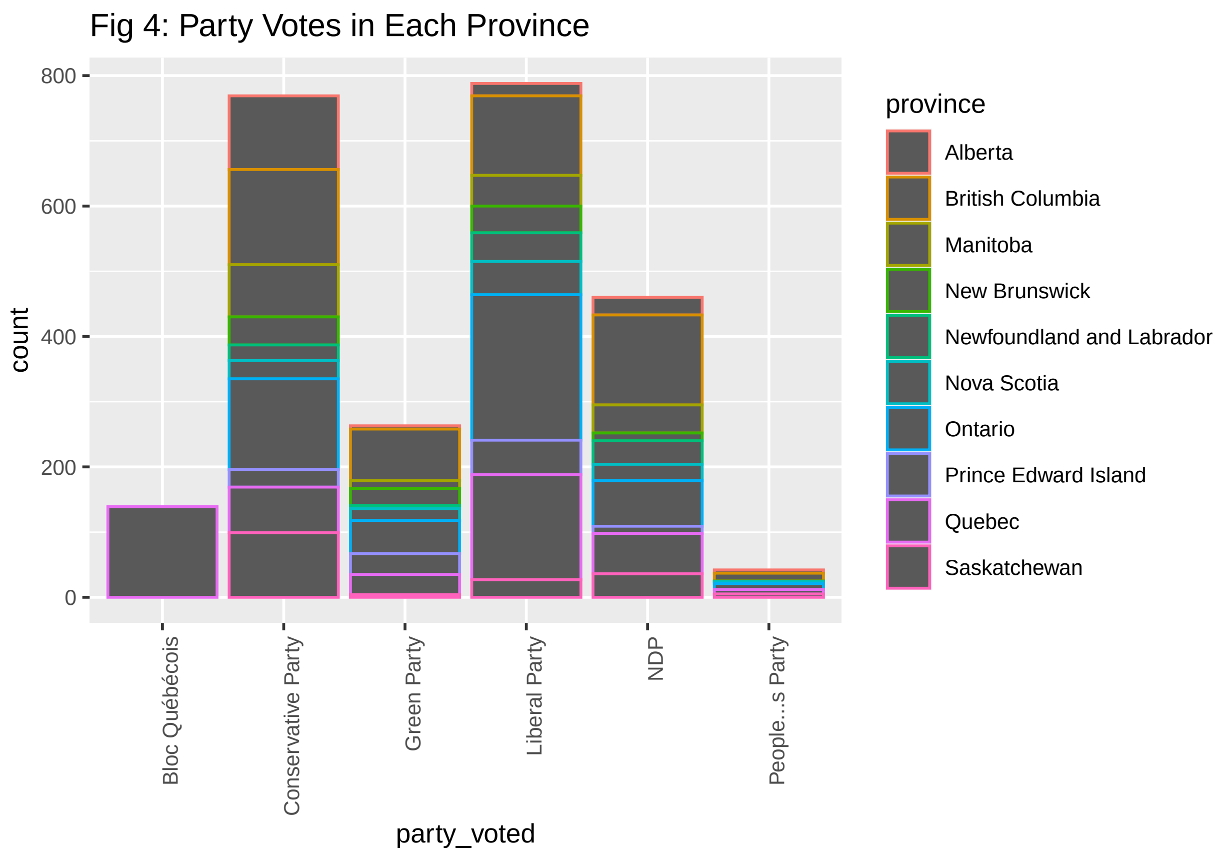


Figure 1 illustrates the relationship between gender and votes. The count plot shows that the distribution of votes for males and females is approximately identical, with most respondents voting for the liberal party, the conservative party, and the NDP. However, it is worth noticing that people who identify themselves as other than male or female voted in favor of the NDP. Nevertheless, as the sample size for people who identify themselves as other than male or female is too tiny, it could hardly influence the overall tendency. Therefore, the plot could imply no regressional relationship between gender and the party the respondent voted for in the election.

Figure 2 is a series of boxplots for the age of the voters of each party. The plot shows that the median age of the voters is the highest for the liberal and conservative parties, while the NDP and the people's party have the lowest median age. Trends for the first and the third quartiles are the same as the median, while all parties have similar minimums for the age of their voters.

Figure 3, like figure 1, is a count plot for each income level and the party they voted for in the election. From the plot, we can see that people in the lower income level tend to vote for the NDP, green party, and conservative party more, while people with higher incomes tend to vote for the liberal more. We can also see that the number of voters forms a "valley" regarding their incomes, with the median income level having the least number of voters.

Figure 4 is a bar plot counting the votes each party got in each province. We can see from the plot that voters in the prairie provinces have a massive trend for the conservative party, while Ontario and the Atlantic region, on the other hand, tend to favor the liberal party. However, for the green party and the NDP, British Colombia contributed a significant portion of its votes, while all the votes for Bloc Québécois came solely from Quebec.

Methods

The problem of this study is to predict the probability that an individual will vote for a specific party based on age, income, province of residence, and gender. This means the data was sampled via a non-probability method, and the problem will include characteristics at two levels: individual and group levels, making ordinary logistical regression inappropriate. Furthermore, including predictors from multiple levels would violate the assumption that all the predictors must be strictly independent. Therefore, a multilevel regression with a poststratification model would be the most suitable, with age, gender, and household income at the individual level (level 1) and the province of residence at the group level (level 2).

Model Specifics

Below is a formula for the model, both level 1 and level 2 are included:

level 1:

$$log(\frac{p\hat{_{h_0}}}{1-p\hat{_{h_0}}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{income} + \beta_3 x_{gender} + \epsilon$$

level 2:

$$\beta_0 = r_0 + r_0 W + u_0$$

Where $log(\frac{p\hat{_{h_0}}}{1-p\hat{_{h_0}}})$ represents the chance each party will be voted by the certain respondent; β_0 represents first level intercept represented by the second level (province); β_1 is the coefficient regarding the variable age; β_2 is the coefficient regarding the variable household income; β_3 is the coefficient regarding the variable gender; ϵ is the error term; r_0 represents the intercept for the provinces, each term in the vector represents the intercept for each province; W represents the weight for the provinces, each term in the vector represents the weight for each province; u_0 is the error at level 2.

Post-Stratification

Post-stratification will be applied to help determine in each province which party will obtain the most amount of votes. Grouping by province most closely mimics Canada's electoral district system. We will apply the census data to conduct poststratification, and the procedures are as follows. First, we must group the census data by age, province, and income range. Second, we will use census data to estimate the probability that an individual will vote for a specific party using the model. Then, we can calculate the product of the population total and the sample mean for each group and divide it by the population total. Finally, we sum up all the values calculated from the previous step and obtain the post-stratified estimator. The formula is written as follows:

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

Where $\hat{y}^{PS}$ is the post-stratified estimator; N_j is the number of observations within each group; $\hat{y_j}$ is the estimated sample mean for each cell; N is the population total.

We will use the methodologies explained above to obtain the results.

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

Results

The below table at the bottom of the section represents the estimated intercept for each province, the estimated coefficient for income level, and the estimated coefficient for gender. Models containing all of age, income level, province, and gender were also fitted. However, the model had too large an eigenvalue for the coefficients so that the result is likely to be statistically insignificant. Models with 2 variable combinations other than income level and gender turns out to fail a t-test for its coefficients at level 1. Therefore, a simpler model featuring provinces at level two, income range and gender at level 1 was adopted. Categorical variables were adopted using corresponding integers for computational convenience. For example, integer one represents

Table 3: Post-stratified Predictors for Each Province

province	predict
1	0.7687321
2	0.8743136
3	0.8376670
4	0.8728951
5	0.9529093
6	0.7419976
7	0.7975574
8	0.8199681
9	0.8216925
10	0.9444740

voting for the liberal party in the response variable, since the coefficients at level one is fixed, provinces featuring a lower intercept could imply favor for voting the liberal party. As in the table, provinces 1 and 6, corresponding to Newfoundland and Ontario, had the lowest value of intercept, so we can infer that the liberal party is the most likely to win the election in Newfoundland and Ontario. Similarly, on the other hand, as provinces 5 and 10, corresponding to Quebec and British Colombia, had the highest intercept, we can infer that the liberal party is the least likely to win the election in Quebec and British Colombia.

Table 3 illustrates the result of post stratified estimator in each province. As we can see from the table, all of the predicted values are close to one. This could imply that the liberal party could even extend its dominance in the next federal elections.

```
## $province
##    (Intercept) income_range       gender
## 1    0.7936089  -0.00400478 -0.003515468
## 2    0.8923099  -0.00400478 -0.003515468
## 3    0.8604931  -0.00400478 -0.003515468
## 4    0.8893939  -0.00400478 -0.003515468
## 5    0.9732240  -0.00400478 -0.003515468
## 6    0.7646606  -0.00400478 -0.003515468
## 7    0.8156335  -0.00400478 -0.003515468
## 8    0.8370367  -0.00400478 -0.003515468
## 9    0.8410896  -0.00400478 -0.003515468
## 10   0.9627438  -0.00400478 -0.003515468
##
## attr(,"class")
## [1] "coef.mer"
```

Conclusions

In the introduction sectoin, we hypothesized that that the chance that the liberal party will win the next federal election still outstands. During the analysis, multilevel regression is applied to model the survey data, and the result was fitted to the census data to calculate the overall post-stratification estimator. The result of the analysis conforms with our initial hypothesis that liberal party is likely to dominate the next federal election as the calculated estimators for all the provinces leaned towards the liberal party.

As mentioned in the introduction section, forecastings of election results has been happening in all areas of acedemia, and its influences have spread beyond scholarly fields. The influences of such studies will continue to widespread in the future. However, there still exists some weaknesses in the study. For example, due to the covid-19 pandemic, people's ethusiasm for participating the federal election has decreased dramatically, and this will mean more difficulty in the data collection process. As a result, the accuracy and amou nt of the sample is likely to reduce as probability sampling could be hard to employ. Another issue is the lack of data for each electoral district. The only readily available sources of data is the provincial data, which are large regions in the country, and this could further reduce the accuracy of the model. Moreover, the model accuracy could also be dramatically increased if we have more data from previous elections. In the end, given the census and surcey data, I was able to give a rough prediction for the next federal election. I encourage futher students and scholars to dig deeper into this topic and elaborate upon previous studies.

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. Dowding, K. (2020). Why forecast? the value of forecasting to political science. PS: Political Science & Politics, 54(1), 104–106. https://doi.org/10.1017/s104909652000133x

5. Campbell, J. E., & Mann, T. E. (2016, July 28). Forecasting the presidential election: What can we learn from the models? Brookings. Retrieved November 24, 2022, from https://www.brookings.edu/articles/forecasting-the-presidential-election-what-can-we-learn-from-the-models/

6. Canada, E. (2022). The Electoral System of Canada. – Elections Canada. Retrieved November 24, 2022, from https://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=e

7. Forty-third general election 2019. Official Voting Results. (2019). Retrieved November 24, 2022, from https://www.elections.ca/res/rep/off/ovr2019app/51/table9E.html

8. Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, (2020), "2019 Canadian Election Study - Phone Survey Technical Report.pdf", 2019 Canadian Election Study (CES) - Phone Survey, https://doi.org/10.7910/DVN/8RHLG1/1PBGR3, Harvard Dataverse, V2

9. Xie, Yihui Christophe Dervieux. "10.1 The Function Knitr::Kable() |R Markdown Cookbook." Kntr::Kable, (2020), bookdown.org/yihui/rmarkdown-cookbook/kable.html.

10. "Access the CES Datasets a Little Easier." Accessed 24 Nov 2022. CES Data, hodgettsp.github.io/cesR.

11. "LaTeX -A Document Preparation System." Latex, Accessed 24 Nov 2022., www.latex-project.org.