Prediction for 2025 Federal Election

STA304 - Fall 2023 -Assignment 2

Group 5: Ji Hoon Kim, Geon Lim, Dongkeun Jang

2023-11-20

Introduction

In Canada, active groups of political parties play a crucial role in the parliamentary process. here are seven notable groups of political parties, namely the Liberals, Conservatives, New Democratic Party (NDP), Bloc Québécois, Green Party, Communist Party, and a category of 'Unknown' parties.

In the dynamic landscape of Canadian politics, accurately predicting electoral outcomes for future federal elections has become a key point. The upcoming election, tentatively in 2025, can be predicted statistically to analyze the complexities of voter behavior and demographic characteristics. While the measurement for the prediction is expected to be straightforward, there is no simple way of mapping the survey population in more detail that works for the entire population. To analyze our predictive model regarding the next election's results, each party with the highest number of votes lies with a deeper understanding of variables inside the survey population. This paper focuses on the three major parties in Canada's political landscape: 1) the Liberals, 2) the Conservatives, and 3) the New Democratic Party (NDP), especially chosen due to their significant representation in the current Members of Parliament, having the highest chance to win the next election. Members of Parliament or MP is a group of politicians who have been elected representing their particular area (Parliament of Canada, n.d.). Among 650 members, the proportion of each parties are different. For instance, the Liberal party holds 158 of these seats (Parliament of Canada, n.d.). This significant representation suggests a higher likelihood of that party winning in future elections due to their substantial presence in the current Parliament.

In the nature of the federal election, the challenge of resolving the percentage of people's votes to each party is likely to reflect the continuously evolving structures of society along with individual identities. This paper will further investigate not only sex and age but also other demographic variables like education, province, and income to capture the diversity of the Canadian electorates. This complexity is then addressed by showing which party will likely have the highest popularity, leading to the following future members of the House of Commons. In the past, federal elections had to be adjusted to provide a more accurate response rate. Hence, a variety of variables were taken into account. For instance, due to the significantly fewer female voters than men in 2021, adjustments were necessary to modify the weights of under-estimated and over-estimated populations (Dunham, 2021). Weighting techniques such as poststratification are employed to provide a precise prediction of the survey samples from national censuses. Poststratification is a technique of using known population totals to reweigh the sample to look more like the target population (Caetano et al., 2023). By using poststratification, there will be a high possibility of reducing the skewness of the data.

Consequently, this paper adapts statistical analysis to predict the 2025 federal election outcomes with a more precise result. This paper hypothesizes that the winning probability calculated from the logistic model will be a reliable source while predicting the next 2025 federal election. By comparing the probabilities for the parties, this paper deduce that the Liberal Party will be elected again in 2025. Using logistic regression and poststratification, three different winning probabilities will be calculated. Logistic regression is well suited for scenarios when the data outcome consists of binary numbers or categorical. The use of this regression model from the survey data will then be enhanced by the poststratification technique, forecasting the overall popular vote. By integrating the statistical analysis with a deeper understanding of the complex landscape

of Canadian society, this paper aims not only to identify the political party most likely to emerge victorious in the upcoming election but also to further enrich the knowledge of the consistently evolving Canadian politics and electoral dynamics.

Data

Survey Data is a survey done by the government of Canada and most of the questions are related to political parties. Participants completes required parts of the survey based on their location of residence and age. Questions include the main question "Who are you most likely to vote for?" as well as other questions regarding political parties such as "which parties deals with issues the most well?" The participation of this is completely voluntary.

Census Data is an observational data by government of Canada and it includes as much characteristics as possible about each Canadians. Although not all Canadians are included, this sample data is a great representation of the whole Canadian population.

Variable Description:

- age: The participant's age in years (numerical, grouped)
- sex: The participant's biological sex ("male" or "female")
- education: The participant's level of education ("Less than Bachelor" or "Bachelor or More")
- province: Name of the province the participant is from (character)
- income: The total salary of the participant after tax (numerical, grouped)

When cleaning the survey_data, we first decided to group the ages which matches the census data, that way we can make a prediction using post stratification later on. Participants aged between 18 and 25 inclusive will be grouped as "18-25" and so on. Here, if the participants were under 18, they were given NA values which later gets eliminated. Secondly, we decided to remove all other genders that aren't male or female and the detailed reasoning is described in subsequent paragraphs. Same as age, income values were grouped to match the census data. And we decided to group education into 2 categories, which were Less than Bachelor's and Bachelor's or higher. Finally, for province variable, no changes occur other than converting numbers into actual names of the provinces. On top of these cleaning, 3 new variables were created, one for each party which are indicators that have a value of 1 if the participant voted for them and 0 otherwise. Census data had same cleaning process for each variable and additionally 2 new variables were added which is the population size of each strata as well as the proportion in percentage.

The cleaned data of the survey along with the census data is shown in the appendix.

Table 1: Prediction of win rate soley based on survey data

Liberal	Conservative	NDP
0.2677019	0.2505866	0.1919255

To compare our final predictions using post-stratification and census data, this is the win rate of the Liberal, Conservative, NDP party solely based on the survey data. Throughout this report, we will see why this is inappropriate, however it is a good standard to compare our final results to.

Table 2: Summary statistics for important variables (Age, Income)

Measure	age_survey	income_number_survey	age_census	income_number_census
Mean	51.3051697	8.189343e+04	52.190452	6.138518e+04
Median	53.0000000	7.000000e+04	54.200000	3.749950e+04
Variance	295.8843828	5.587783e+09	314.952547	2.585439e+09
Skewness	-0.1372278	5.928472e+00	-0.194184	5.884558e-01

Here we calculated the important statistics for numerical variables which are age, and income in Canadian Dollars. First, we notice that the mean and median age of the survey data is slightly higher than the mean and median age of the census data, from this we can make an inference that there were more younger survey participants than older participants. This is reflected on the skewness of the data which is negative, meaning that the there are more people with higher age. Secondly, for income, prior to calculating the mean and median income number, we had to filter out certain observations in the survey data. This was due to some observations having abnormally high income such as 10^30 Canadian Dollars. Whether these values were incorrectly entered or it was entered as a joke, this would cause too much impact on the mean, therefore all data over 3 million dollars were removed as there weren't as many people with income over 3 million dollars. Additionally, if the participants decided to omit this question, the data was auto filled with a value of -99. Having this value in the calculation will obviously bring down the mean for incorrect reasons, they were removed as well. In the census data, there were no direct amount of income, instead it was a range of incomes. Although may not be appropriate, the mean of that range of income was used to compute this calculation. The mean income in the survey data was around 80 thousand with a median of 40, while the mean income in the census data was 70 thousand with a median of 37k. One possible explanation for such difference could be explained by the relationship with age. Since the participants in the survey data are older on average, and more work experiences implies higher salary, it is one of the possible causes. Nonetheless, it successfully shows that survey data is a poor representation of the canadian population and cannot be used to predict the election results.

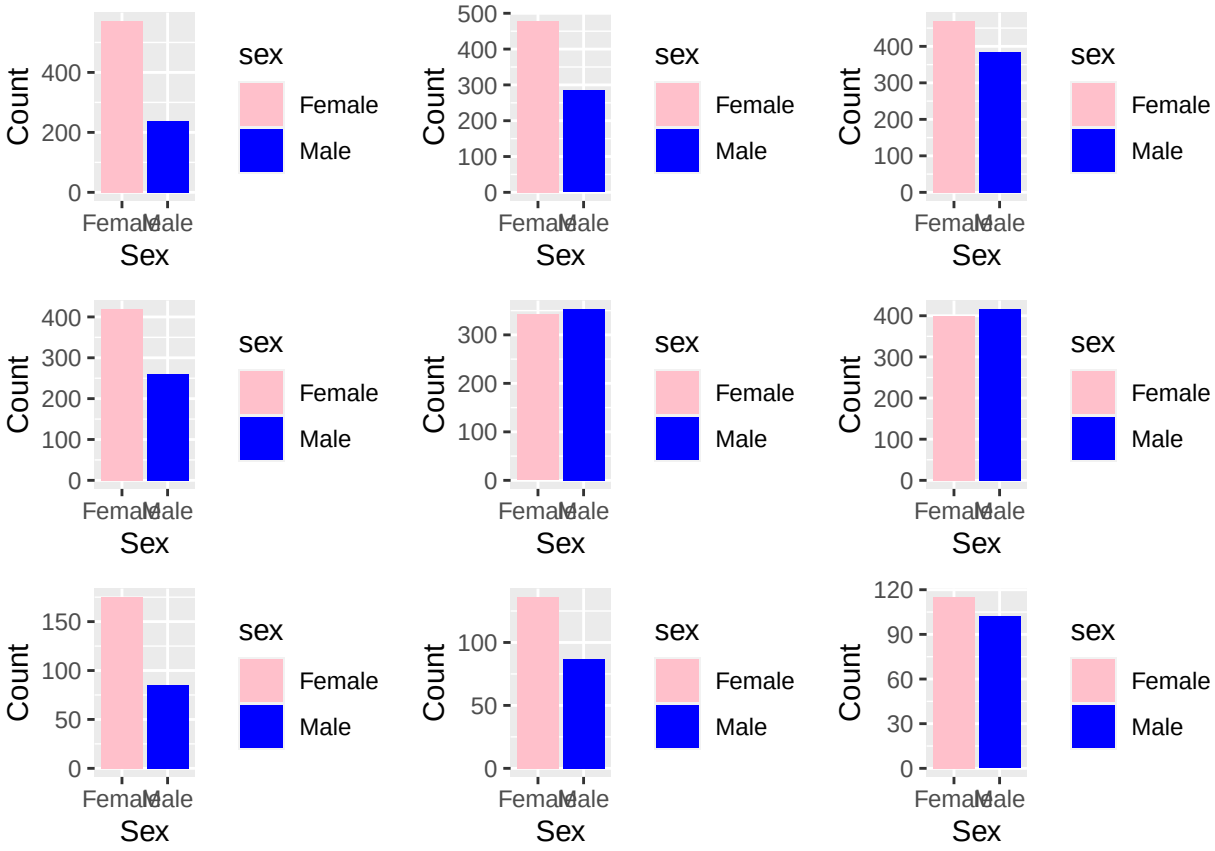
Table 3: Summary statistics of missing data based on census data

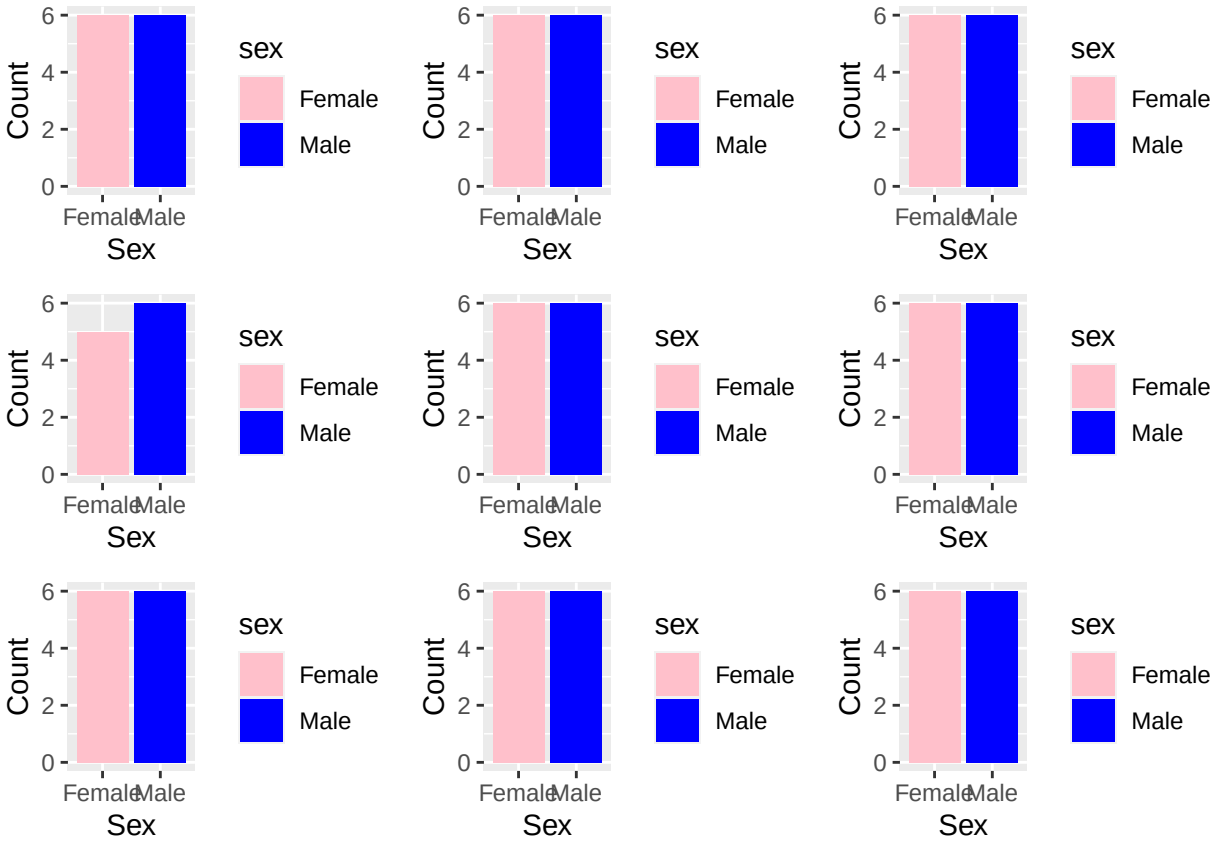
missing_age	missing_education	missing_sex	missing_income_respondent	missing_province
424	341	0	0	0

Table 4: Summary statistics of missing data based on survey data

missing_age	missing_education	missing_sex	missing_income_respondent	missing_province
0	0	124	642	0

Before fitting the model with the chosen predictors, during the cleaning process, some observations had to be removed. Other than omiting all NA values, as you cannot fit a model with NA values, there were some other cases which they had to be removed. For example, the sex variable was a binary variable in the census data "male" or "female". However, in the survey the participants had the option to choose "non-binary" or "other" which corresponds to values 3 and 4. Because it was inappropriate to assign random male or female values to these observations due to the sex ratio becoming different, we decided that it was best to omit them. As mentioned above, we removed negative values for income variable and we removed observations where their age was less than 18 to prevent affecting any results since they do not have voting rights yet.


Count
Sex
Female
Male
sex
Female
Male



This is a two 3x3 plot one for each survey and census data that visualizes the sex ratio in each strata. We chose Ontario, Quebec, British Columbia and the age groups "26-35","36-45","46-55" because these strata contained the highest population. The first row of each plot grid represents the gender ratio of people in Ontario with age group "26-35","36-45","46-55" respectively. Second row is Quebec for the same age group and the third row is British Columbia also for the same age group. Regardless of the province, in the survey data, there is significantly more females than males in almost every strata. 7 out of 9 strata has more females than males. For British Columbia age group 26-35 there are more than double females than males. While in the census data, the ratio is precisely equal except for age group 26-35 in Quebec. From this, we once again realize making inferences about the election just based on the survey data will yield a very incorrect result assuming that gender plays a key role in their voting characteristics.

Methods

As mentioned in the introduction, the goal of our report is to predict the overall expected result of the vote for the next Canadian federal election using statistical models. To achieve it, we are going to use a logistic regression and a post-stratification to estimate the probability of winning rate for each party, focusing on Liberal party, Conservative party, and NDP.

There are a few missing data in the 'sex' and 'education' variables. In the sex variable, there are few observations of people who did not want to reveal their sex. On the other hand, for the education variable, there were also few NA observations. Even though these missing observations still affect our analysis, we decided to get rid of them from our data as their size was relatively small compared to the whole sample size. We believed that the impact of removing these observations on the analysis will not be huge. Additionally, we also decided to remove observations whose age is younger than 18. Since people under 18 are not allowed to vote, it was unreasonable to include them in our analysis.

Logistic regression model

We determined to use the logistic regression model to estimate the probability of voting for the Liberal party, Conservative party and NDP. There is a clear reason why we chose this model among several statistical models. This is because the outcome for the model would be 1 or 0, representing a binary response and we want to estimate the probability of winning rate for each party. Thus, the logistic regression model would be the best model for our analysis because it is usually used to show the relationship between a binary response variable and predictors and to estimate the probability of a event occurring using independent variables in the model. Therefore, this model will results in desired estimates.

As mentioned earlier, there will be three models for each party and each model has the same variables in common. There are 21 indicator variables in each model and they are all crucial for our analysis

Model for liberal party

$$log(\frac{P_{Liberal}}{1-P_{Liberal}}) = \beta_0 + \beta_1 x_{male} + \beta_2 x_{age26\sim35} + \beta_3 x_{age36\sim45} + \beta_4 x_{age46\sim55} + \beta_5 x_{age56\sim65} +$$

$$\beta_6 x_{age65+} + \beta_7 x_{underBachelor'sdegree} + \beta_8 x_{BritishColumbia} + \beta_9 x_{Manitoba} + \beta_{10} x_{NewBrunswick} +$$

$$\beta_{11} x_{NewfoundlandandLabrador} + \beta_{12} x_{NorthwestTerritories} + \beta_{13} x_{NovaScotia} + \beta_{14} x_{Nunavut} + \beta_{15} x_{Ontario} +$$

$$\beta_{16} x_{PrinceEdwardIsland} + \beta_{17} x_{Quebec} + \beta_{18} x_{Saskatchewan} + \beta_{19} x_{Yukon} + \beta_{20} x_{50k\sim under100k} + \beta_{21} x_{Under50k}$$

Model for conservative party

$$log(\frac{P_{Conservative}}{1-P_{Conservative}}) = \beta_0 + \beta_1 x_{male} + \beta_2 x_{age26\sim35} + \beta_3 x_{age36\sim45} + \beta_4 x_{age46\sim55} + \beta_5 x_{age56\sim65} +$$

$$\beta_6 x_{age65+} + \beta_7 x_{underBachelor'sdegree} + \beta_8 x_{BritishColumbia} + \beta_9 x_{Manitoba} + \beta_{10} x_{NewBrunswick} +$$

$$\beta_{11} x_{NewfoundlandandLabrador} + \beta_{12} x_{NorthwestTerritories} + \beta_{13} x_{NovaScotia} + \beta_{14} x_{Nunavut} + \beta_{15} x_{Ontario} +$$

$$\beta_{16} x_{PrinceEdwardIsland} + \beta_{17} x_{Quebec} + \beta_{18} x_{Saskatchewan} + \beta_{19} x_{Yukon} + \beta_{20} x_{50k\sim under100k} + \beta_{21} x_{Under50k}$$

Model for NDP

$$log(\frac{P_{NDP}}{1-P_{NDP}}) = \beta_0 + \beta_1 x_{male} + \beta_2 x_{age26\sim35} + \beta_3 x_{age36\sim45} + \beta_4 x_{age46\sim55} + \beta_5 x_{age56\sim65} +$$

$$\beta_6 x_{age65+} + \beta_7 x_{underBachelor'sdegree} + \beta_8 x_{BritishColumbia} + \beta_9 x_{Manitoba} + \beta_{10} x_{NewBrunswick} +$$

$$\beta_{11} x_{NewfoundlandandLabrador} + \beta_{12} x_{NorthwestTerritories} + \beta_{13} x_{NovaScotia} + \beta_{14} x_{Nunavut} + \beta_{15} x_{Ontario} +$$

$$\beta_{16} x_{PrinceEdwardIsland} + \beta_{17} x_{Quebec} + \beta_{18} x_{Saskatchewan} + \beta_{19} x_{Yukon} + \beta_{20} x_{50k\sim under100k} + \beta_{21} x_{Under50k}$$

Parameters/coefficients

$log(\frac{P_{liberal}}{1-P_{liberal}})$: The log odds for the liberal party.

$log(\frac{P_{conservative}}{1-P_{conservative}})$: The log odds for the conservative party.

$log(\frac{P_{NDP}}{1-P_{NDP}})$: The log odds for the NDP party.

$P_{liberal}$: The probability of voting for the liberal party.

$P_{conservative}$: The probability of voting for the conservative party.

P_{NDP}: The probability of voting for the NDP party.

β_0: The intercept of the model which means that every predictor is all zero. Moreover, it also represents the baseline.

β_1: The change in log odds when a voter is male,$x_{male} = 1$. If $x_{male} = 0$, then the voter is female.

$\beta_2 \sim \beta_6$: The change in log odds when a voter is within age group 26~35 ($x_{age26\sim35} = 1$), age group 36~45 ($x_{age36\sim45} = 1$), age group 46~55 ($x_{age46\sim55} = 1$),age group 56~65 ($x_{age56\sim65} = 1$), and age group 65+ ($x_{age65+} = 1$) respectively. If $x_{age26\sim35}$, $x_{age36\sim45}$, $x_{age46\sim55}$, $x_{age56\sim65}$ and x_{age65+} are all 0, then the voter is within age group between 18 and 25.

β_7: The change in log odds when the voter's degree is under bachelor, $x_{underBachelor'sdegree} = 1$. If $x_{underBachelor'sdegree} = 0$, then the voter's degree is bachelor or more.

$\beta_8 \sim \beta_{19}$: The change in log odds when a voter lives in British Columbia ($x_{BritishColumbia} = 1$), Manitoba ($x_{Manitoba} = 1$), New Brunswick ($x_{NewBrunswick} = 1$), Newfoundland and Labrador ($x_{NewfoundlandandLabrador} = 1$), Northwest Territories ($x_{NorthwestTerritories} = 1$), Nova Scotia ($x_{NovaScotia} = 1$), Nunavut ($x_{Nunavut} = 1$), Ontario ($x_{Ontario} = 1$), Prince Edward Island ($x_{PrinceEdwardIsland} = 1$), Quebec ($x_{Quebec} = 1$), Saskatchewan ($x_{Saskatchewan} = 1$) and Yukon ($x_{Yukon} = 1$) respectively. If all predictors corresponding to these coefficients are equal to zero, then the voter lives in Alberta.

β_{20},β_{21}: The change in log odds when the voter's income is between 50k dollars and under 100k dollars $x_{50k\sim under100k} = 1$, and under 50k dollars $x_{under50k} = 1$ respectively. If all predictors corresponding to these coefficients are zero then the voter's income is above 100k dollars.

After we constructed a model, we identified assumption errors to check our data contains any bias. As shown in the appendix, these 6 plots (3 residuals vs fitted values scatter plot and 3 normality plot) is used to check normality and constant variance which are some of the linear regression assumptions. In this case, normality seems to be violated as the points seriously deviate from the normal line in the qq plot for all 3 graphs. But constant variance assumptions seems to hold since there is no "fanning" pattern in the residual vs fitted scatter plot. Fanning pattern is when the distance between points become larger as predictors increase in value.

Post-Stratification

The second statistical method is the post-stratification. It is usually used when a simple random sample is not a proper representation of the population, decreasing the precision of estimates (Qian, 2010). Thus, this method alleviates this problem by reweighing each cell by its relative proportion in the population. In other words, it is possible for the survey dataset to have the different proportion compared to the population proportion since the sample size in each cell is random. Thus, in order to increase the accuracy of estimates and to look more like the target population, we reweigh the sample using population. Therefore, this method not only explains our data but also helps us estimate the desired results. One assumption that must be satisfied to use this method is that the population size and post strata sizes have to be known.

We used variables of sex, age, education, income and province to create cells.For example, the reason for choosing 'sex' variable to classify cells among other variables is because if there exists a specific party to offer more appealing suggestion to the certain sex, then it could influence the outcome. Therefore, it is likely to exist a prefer party depending on the sex. For the 'province' variable, each province have different taste to a particular party thus, we believed it can also influence the outcome of voter. Consequently, similar to the sex variables, depending on the political party's promises, 'age', 'education' and 'income' variables that we chose are treated as factors that have an ability to change the outcome of voter. While there are other variables in the data, we did not include them for the cell classification because they are not as significant nor influential as the five variables.

From the data section, we created 641 cells. We then use the logistic regression model we made to get estimated P for each cell. Afterward, following the formula below, we weight each cell by its relative proportion in the population. Then we will obtain $\hat{P}^{PS}$ for a specific party using the model below. Since we aim to obtain $\hat{P}^{PS}$ for three different parties, we have to repeat the same process twice with the corresponding model, the same set of cells and the different weight of each cell based on the population.

$$\hat{P}^{PS} = \frac{\Sigma N_j \hat{P}_j}{N}$$

$\hat{P}^{PS}$: The estimated response variable for the population.

$\hat{P}_j$ The estimated response variable for j^{th} cell.

N_j: The population size of the j^{th} cell.

N: The total population size.

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

Results

This paper successfully predicts the winning probability of the Liberal, the Conservative, and the NDP for the 2025 federal election. Due to high number of representatives for each of the parties in the House of Commons, we assume one of these three parties will be elected for 2025. Calculation based on survey data were made in the firsthand to analyze the winning probabilities for each party based on the survey data. To evaluate the accuracy of these calculations, this paper compares the census-based probabilities with those derived from survey data. The table below shows a probability based on the survey data we first calculated for the three major parties. The probabilities obtained are summarized in the table below:

Table 5: Winning Probability for 2021 Election based on Survey Data

Parameter	Value
The Liberal	26.8%
The Conservative	25.1%
The NDP	19.2%

The Liberal Party is shown to have the highest probability of winning the 2025 election, at 0.2677019 (26.8%). The Conservative Party follows with a 0.2505866 (25.1%) chance of winning, and the New Democratic Party (NDP) with 0.1919255 (19.2%) probability. The remaining 28.9% is collectively attributed to the other four political party groups. This data implies that, according to the survey, the Liberal Party is the most likely to win the upcoming 2025 election.

To provide a comprehensive analysis, voting patterns across different provinces was calculated to provide a comprehensive view of the overall voting patterns in 2021 election. The table below provide probabilities of people in each province voting for each of the three parties.

Table 6: Provincial Voting Probabilities for the Liberal, Conservative, and NDP Parties

provinces	Liberal	Cons	NDP
Alberta	0.2079066	0.3547287	0.2338397
British Columbia	0.2716292	0.2122038	0.2870964

provinces	Liberal	Cons	NDP
Manitoba	0.2576547	0.2550056	0.2532100
New Brunswick	0.3664075	0.1447612	0.1896889
Newfoundland and Labrador	0.4252715	0.1695597	0.2436812
Nova Scotia	0.3705627	0.1729236	0.2596885
Ontario	0.3315475	0.2426792	0.2056542
Prince Edward Island	0.2709795	0.2470439	0.0493697
Quebec	0.2435494	0.1225970	0.1204690
Saskatchewan	0.1465303	0.3669876	0.2748299

Based on the data presented in the table, it is observed that the percentage of votes for the Liberal, Conservative, and NDP parties varies across different provinces in Canada. From an initial overview, it seems the Liberal party holds a competitive position in many provinces, securing the highest proportion of votes. However, there are exceptions where the Conservative or the NDP party leads in some provinces. For instance, the Conservative party shows a strong lead in Alberta and Saskatchewan, while the NDP has a notable percentage in British Columbia. It is important to note that these are probabilities, and actual voting outcomes may vary. This data could be instrumental in understanding regional political patterns for 2025 elections.

Following the result from table 2 portraying that the Liberal party holds the majority of the votes, we have refined the survey data to more accurately reflect the broader population demographics applying poststratification. The hypothesis is that the Liberal Party stands a high chance of winning in the 2025 elections, provided that the winning probabilities generated from the logistic model hold true as a reliable source of indicator. Displayed beneath is the table outlining the winning probabilities we have calculated representing the overall population:

Table 7: Winning Probability for 2025 Election

Party	Probability
The Liberal	28.6%
The Conservative	21.5%
The NDP	20.5%

By comparing these two sets of probabilities identified in table 1 and 3, these two tables are useful while assessing the accuracy and consistency of our predictive model. The two tables mentioned offer a similar conclusion regarding the winning probabilities of various political parties in the upcoming 2025 Canadian election. According to table 3, the Liberal Party leads with the highest probability of victory at 0.2858506 (28.6%). This is followed by the Conservative Party at 0.2148609 (21.5%) and the New Democratic Party (NDP) at 0.2046199 (20.5%). Combined, these probabilities account for 70.6% of the total.

The remaining 29.4% is attributed to the other parties, including the Bloc Québécois, Green Party, Communist Party, and those categorized as 'Unknown'. It's important to note that the statistical analysis was specifically conducted for the three major parties: the Liberals, Conservatives, and NDP. This focus was based on the assumption that one of these parties is most likely to win, considering they are the major groups among the total of seven party categories. This approach was deemed appropriate for our analysis, given their significant representation and influence in Canadian politics.

Consequently, table 1 and 3, illustrates our prediction accepts our hypothesis based on statistical analysis. Both suggest that the Liberal Party has the highest estimated probability of winning. These results indicates a similarity in the data analysis, as both tables reflect a similar conclusion regarding the leading position of the Liberal Party in terms of winning probability. Analyzing the previous 2021 election outcomes, we observed that the Liberal party secured a victory with a total of 160 seats in Parliament while the Conservative party finished with the total of 119 seats ("Federal election 2021 live results," n.d.). This historical data aligns

with our current findings, which also indicate the Liberal party as having the highest probability of winning. The consistency of the Liberals' past performance with our calculated results reasonably increase the result's credibility and strengthens the plausibility of our analysis.

Conclusions

In conclusion, our initial hypothesis was that the Liberal Party is likely to win the 2025 election. To test this hypothesis, we employed logistic regression modeling, an appropriate choice given that our indicator variable of "voting for a specific party" was binary in nature. Our analysis involved segmenting the target population into five strata: age, sex, education, province, and income. This stratification resulted in the creation of 641 distinct cells, enabling us to calculate individual estimate values for each cell. Our initial step involved deriving the winning probabilities for the political parties from the survey responses, a critical measure to guarantee the precision and dependability of our subsequent analysis. The results indicated that the Liberal party held a 26.8% chance of winning.

Following this, we determined the likelihood of each party securing votes across the various provinces. Observing the Liberals' leading position, we progressed to poststratification, applying weights to these estimates in order to represent the probability of the overall population. After calculating the winning probabilities for each of the Liberal, the Conservative, and the NDP, this process led us to conclude that the Liberal Party holds the highest chance of winning the 2025 election, with a probability of 28.6%. Therefore, comparing the probability derived from the survey data in the firsthand, the estimates we derived representing the census data highlight the Liberal party have the highest chance of winning. The comparison of these two sets of probabilities was essential to confirm the integrity of our conclusions.

Despite that our result follows our hypothesis and derived a result that the Liberal Party is likely to win, there are still limitations regarding our process of logistic regression and poststratification. Since the data this paper adapts to predict the next federal election is based on the past data gathered from 2021, the data might not represent accurately. Due to reliance of historical data, there are chances of unexpected events which might disturb the result.

During the process of deriving our result, there were limitations regarding the process of deriving the results. We had to calculate out the results by using the data provided by the Canadian Election survey in 2021. In the observed data, there were data that were missing due to personal reasons or systematic reason. There were some respondents who did not revealed their gender along with naturally appearing NA values in the variables we selected for the process. Therefore, due to these issues, we had to carry out our analysis by deleting these data. Even though, those missing data were taking a small proportion, reconstructing a model using imputation is likely to be appropriate while constructing a valid prediction. Additionally, as shown in the QQ plot attached in the appendix, we can see normality was violated. Due to severe deviation from the ab-line, box-cox transformation can be adapted to mitigate this violence. One method such as maximum likelihood estimation can be used to provide a more reasonable data, thus giving a more credible result.

Furthermore, while discussing about our overall procedure of analysis process, our calculation only considered 5 variables among total of 80 variables. As the paper did not take into account of other variables, the predicted power might be reduced leading to a change in the overall winning probability for the three parties. These unmeasured factors can be an important factor while examining the outcome of an election. These disturbances might chance the value for the estimations or the actual result for the future 2025 election, also might result a different party to win the election.

Taking these limitations into account, for future analysis, we might use multilevel modelling. In this process, level 1 will be the individual voters, while level 2 can be defined by provinces. This structure allows us to analyze how individual-level factors interact with provincial-level characteristics to influence voting behavior, making more accurate predictions. Since each of the provinces have similarities leading to dependency of data within the province, adopting multilevel modelling can be useful while examining how individual and provincial identities might interact shaping unique voting activities.

Nevertheless, this paper successfully addressed the 2025 election winner providing a comprehensive method of the process of logistic regression and postratification along with comparison with the prediction gathered

by the sample data, concisely creating a conclusion that the Liberal party have the highest chance of winning in 2025.

Bibliography

1. Allaire, J.J., et al. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)

2. Caetano, S.J., & Somerset, E. (2023). *Multilevel Regression and Postratification*. [PowerPoint slides]. University of Toronto.

3. Dunham, J. (2021). *Highest percentage ever of female and gender-diverse candidates running in this election*. CTVNews. https://www.ctvnews.ca/politics/federal-election-2021/highest-percentage-ever-of-female-and-gender-diverse-candidates-running-in-this-election-1.5570913.

4. Grolemund, G. (2014, July 16). *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)

5. n.a. (n.d.). *Federal election 2021 live results*. CBCnews. https://newsinteractives.cbc.ca/elections/federal/2021/results/

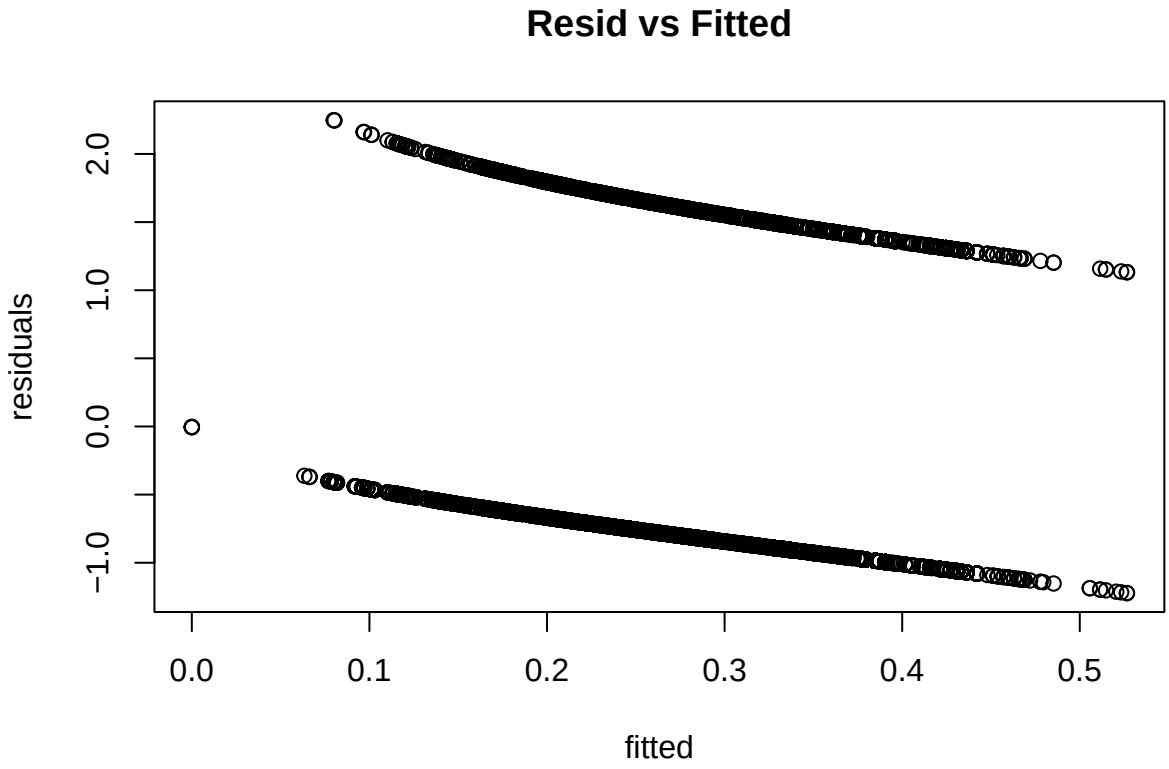
6. OpenAI. (2023). *ChatGPT (September 13 version) [Large language model]*. https://chat.openai.com/chat (Last Accessed: September 13, 2023).

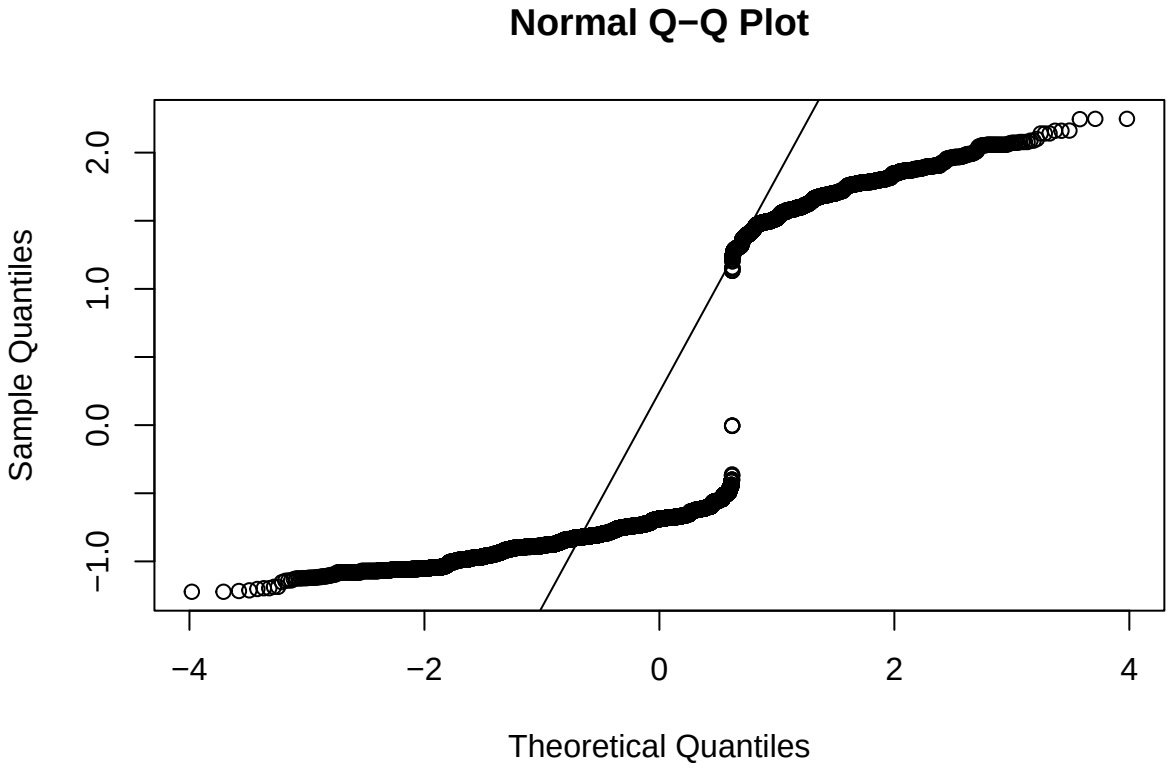
7. Parliament of Canada. (n.d.). *Senate of Canada*. SenCanada. https://sencanada.ca/en.

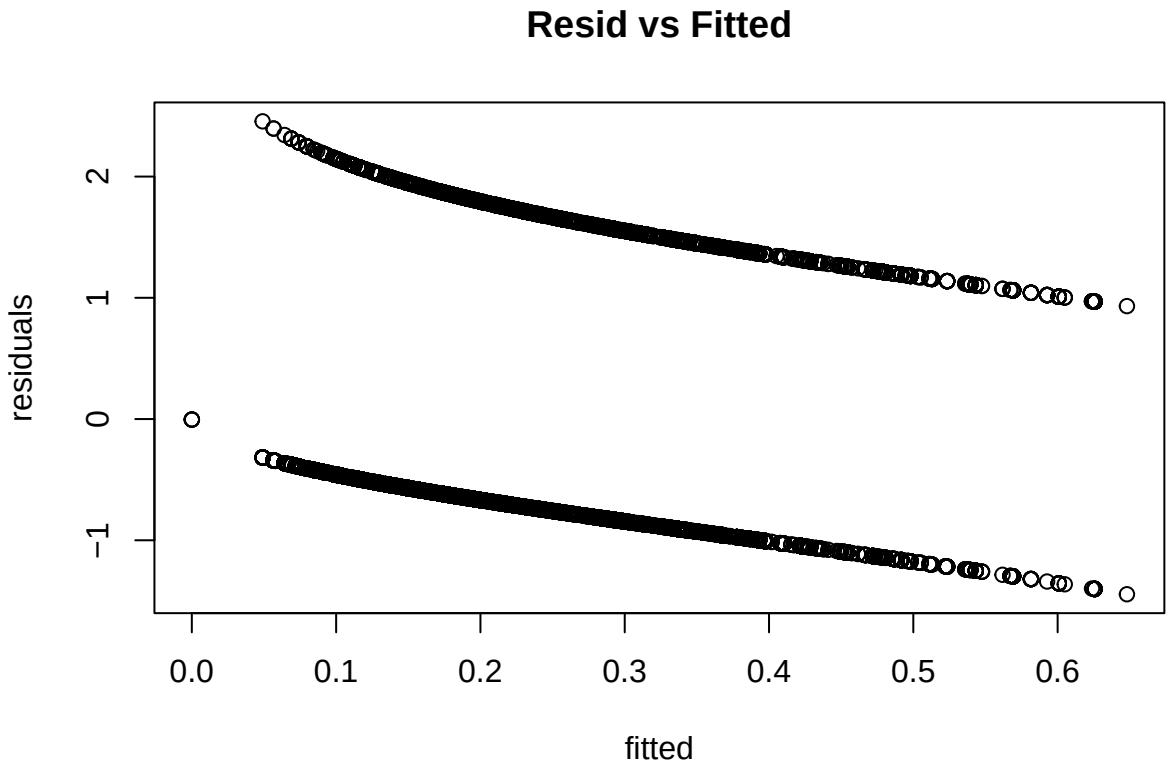
8. Qian, J. (2010). Sampling. *International Encyclopedia of Education (Third Edition)*, 390-395. https://doi.org/10.1016/B978-0-08-044894-7.01361-0.

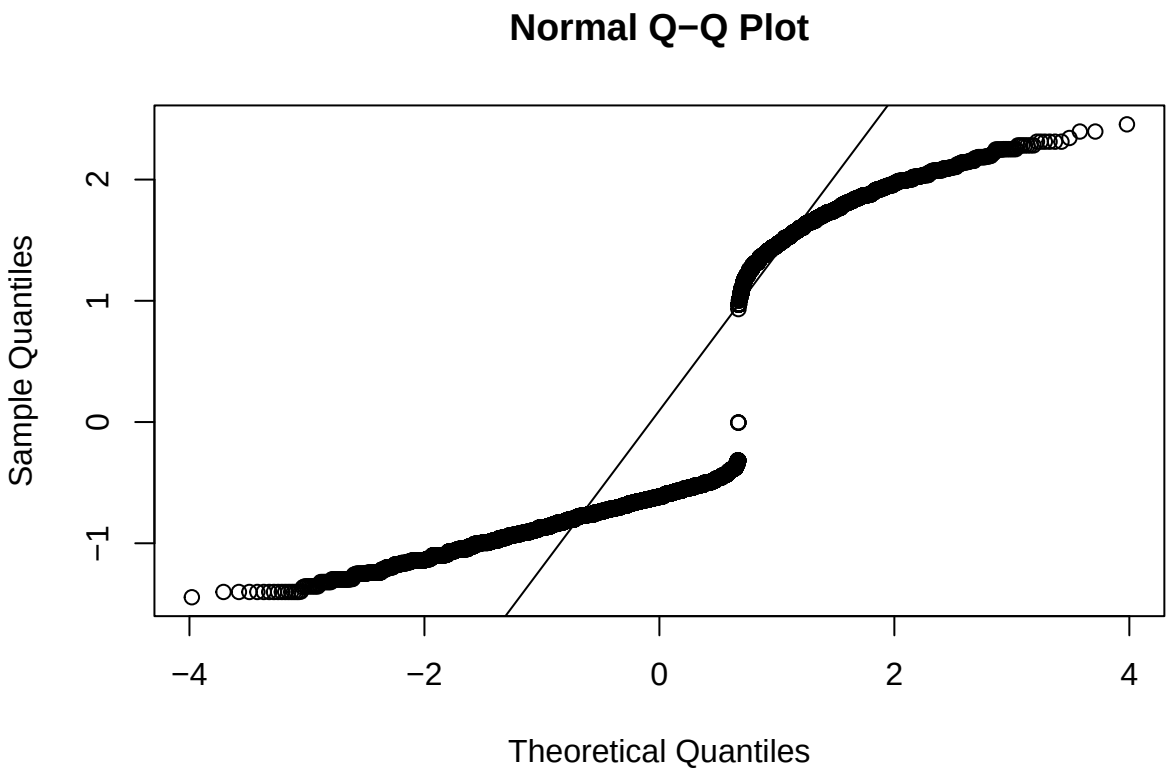
9. RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA. http://www.rstudio.com/.

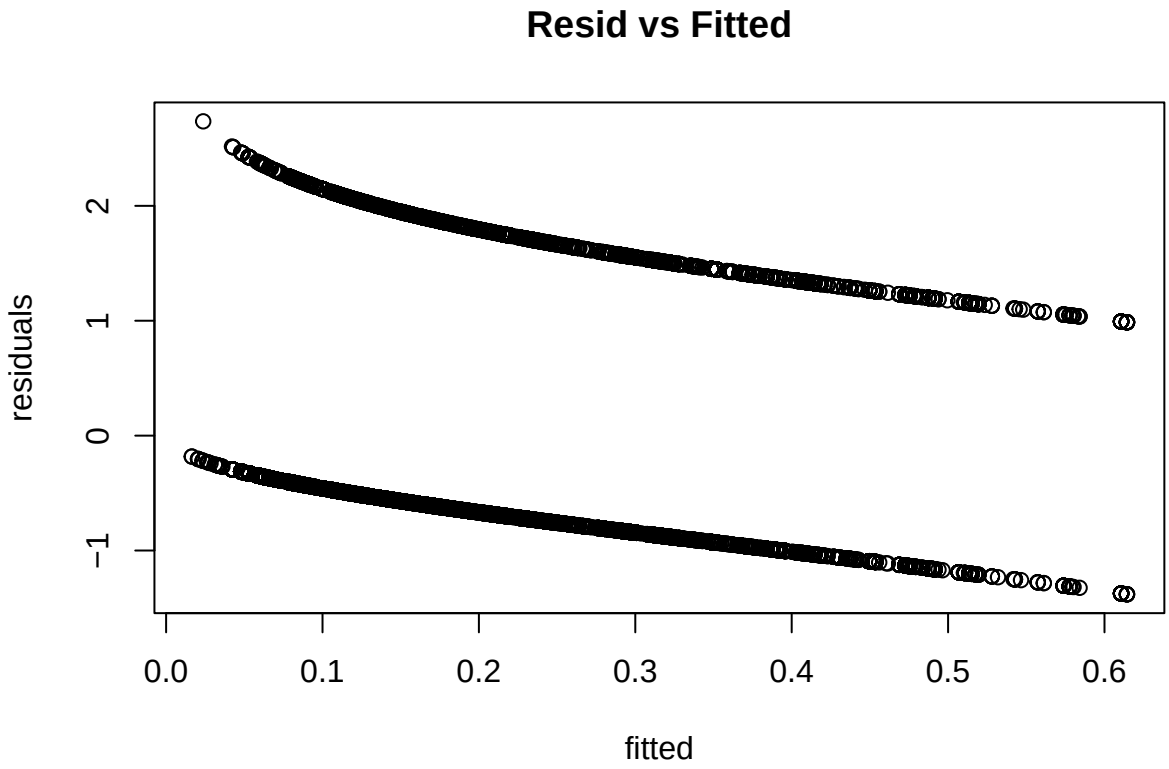
Appendix

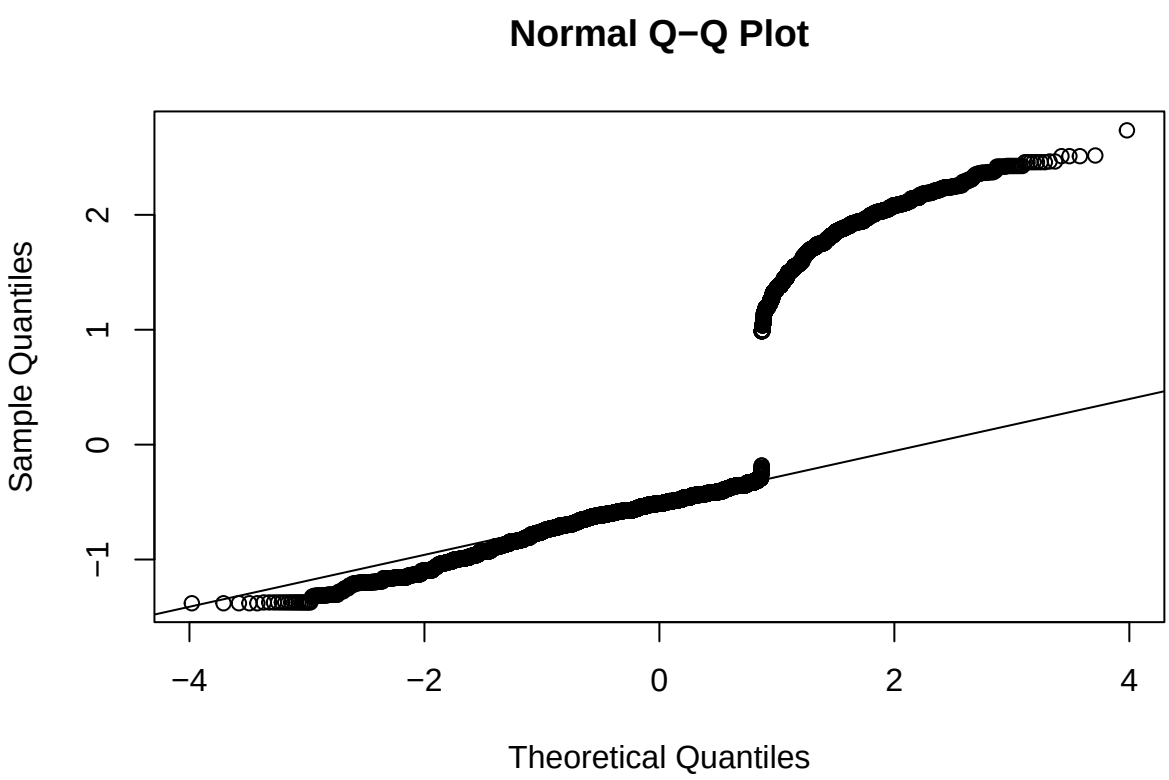

Resid vs Fitted
residuals
2.0
1.0
0.0
−1.0
0.0
0.1
0.2
0.3
0.4
0.5
fitted


Normal Q−Q Plot
Sample Quantiles
2.0
1.0
0.0
−1.0
−4
−2
0
2
4
Theoretical Quantiles


Resid vs Fitted
residuals
2
1
0
−1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
fitted


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−4
−2
0
2
4
−1
0
1
2


Resid vs Fitted
residuals
2
1
0
−1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
fitted



Glimpse of Cleaned Survey Data

```
## Rows: 14,490
## Columns: 8
## $ age          <chr> "18-25", "26-35", "36-45", "56-65", "46-55", "65+", "46-5~
## $ sex          <chr> "Female", "Female", "Female", "Female", "Female", "Male",~
## $ education    <chr> "Bachelor's degree or more", "Under Bachelor's degree", "~
## $ province     <chr> "British Columbia", "British Columbia", "Quebec", "Quebec~
## $ income       <chr> "$100k and above", "$100k and above", "$50k ~ Under $100k~
## $ vote_liberal <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, ~
## $ vote_cons    <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, ~
## $ vote_ndp     <dbl> 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, ~
```

Glimpse of Cleaned Census Data

```
## Rows: 641
## Columns: 10
## Groups: age, sex, education, province [239]
## $ age                          <chr> "18-25", "18-25", "18-25", "18-25", "18-25~
## $ sex                          <chr> "Female", "Female", "Female", "Female", "F~
## $ education                    <chr> "Bachelor's degree or more", "Bachelor's d~
## $ province                     <chr> "Alberta", "British Columbia", "British Co~
## $ income                       <chr> "Under $50k", "$50k ~ Under $100k", "Under~
## $ number                       <int> 8, 2, 8, 4, 3, 5, 3, 6, 29, 4, 8, 1, 3, 1,~
## $ cell_prop_of_division_total <dbl> 4.286556e-04, 1.071639e-04, 4.286556e-04, ~
## $ estimate                     <dbl> 0.10116067, 0.16276580, 0.13696128, 0.1263~
```

```
## $ estimate2                  <dbl> 0.22556966, 0.15164449, 0.11839512, 0.1543~
## $ estimate3                  <dbl> 0.5134925, 0.5738260, 0.6106148, 0.5573214~
```

Generative AI Statement

While working with Rstudio to create and display multiple graphs, I encountered a technical difficulties when attempting to knit my document into a PDF format. The process was dsirupted by an error message stating that "Error 'contrib.url()': !trying to use CRAN without setting a mirror backtrace. 1. utils::install.packages("gridExtra")".

Therefore, to seek a solution for this error, I used ChatGPT in the Data section of my project for assistance. After seeing the issue, the solution provided was straightforward. I needed to specify a CRAN mirror in my R environment settings. By incorporating the line options(repo = c(CRAN = "https://cran.rstudio.com/")) into my R script, I directed the package installation to utilize the CRAN repository in RStudio, thus resolving the package installation issue. This allowed the grid.arrange function to operate thoroughly, and I was subsequently able to knit the document into the PDF output successfully.