Candian Election Forecast: A Real-world Application of Multilevel Regression Poststratification

STA304 - Fall 2023 -Assignment 2

GROUP 26: Leyuan (Tony) Chen, Xinyue (Eva) Wang, Xuening (Shirley) Wu

November 14, 2023

Introduction

Politics, though appearing far away, is arguably extremely close to individuals' daily lives since political changes can have considerable impacts on individuals' daily lives. For example, economic policies influence the amount of taxes that individuals pay, fiscal policies can significantly influence education and health care qualities, and policies towards other countries may even pose threats to people's safety and security as the result of national tensions and conflicts [9]. Despite being the key stakeholders, most individuals do not have direct access to political decision-making except for voting. Hence, voting becomes a very important way for citizens to express their political attitudes and for political leaders to understand the needs and requests of the public.

With its significance and uniqueness, voting has caught wide attention from researchers in various fields. Statisticians, for instance, are extending their interests in studying trends in fields like finance and biomedicals to trends in political votings. Specifically, statisticians are interested in identifying trends and relationships among individuals' voting choices and a series of auxiliary variables (e.g., age, gender, and area of residence) [10]. The identified relationships among the variables allow statistically informed inferences about the public opinions to be generated, providing valuable insights for politicians, and facilitating welfare-maximizing policies [10].

However, generating accurate predictions is extremely challenging, especially for complicated large-scale datasets like voting datasets. There are two main reasons for the problem: (1) Voting behaviors are influenced by too many factors, including income levels, education, sex and gender, race and ethnicity, and areas of residence. Not only does each factor have certain impacts on people's political attitudes and voting behaviors, but these factors also interact with each other and impact voting behaviors on aggregate. Imagine a person from Ontario who has a low income vs a person of the same income level but is from Quebec, it is likely that their voting choices will be completely different. (2) For many times, it is extremely hard to recruit a representative sample of the target population because of the complicated demographic compositions as well as the varied non-response rates. Such sample non-representativeness and non-response negatively impacts the accuracy of extrapolating predictions from samples to the entire population [8].

To tackle this problem, we plan to use a statistical tool that accounts for the two difficulties. Specifically, it allows us to (1) build models based on variations in individual voters and variations with respect to different groups (e.g., age groups) and (2) to adjust our model's predictions according to population demographics [8]. To prepare you for understanding further analyses, we will introduce the terminologies and materials that will be used in the following report:

- Statistical Tool: ** Multilevel Regression Poststratification (MRP): it combines Multilevel Regression and Poststratification to generate predictions. Specifically, "it poststratifies the population into a large number of cells based on combinations of various demographic attributes, use the sample to estimate the outcome of interest within each cell by fitting a multilevel regression model, and finally aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population." [5] ** Logistic Regression: it is an algorithm used for binary classification problems [6], which models the relationship between predictor variables and a binary response variable, e.g., vote or not.

- Materials/ Datasets ** Survey Dataset: Canadian Election Study (CES) 2021 Survey will be used as the survey dataset [7]. It collected survey participants' demographic variables and their voting preferences, which allows us to model voting behaviors based on various demographic attributes. ** Census Dataset: General Social Survey (GSS) will be used as the census dataset. It contains population demographics of the general Canadian population, which helps us to poststratify the population.
- Variables of Interest ** Predictor Variables: Age, Sex, Province, Family Income level. (These five predictor variables were selected based on data availability and plausibility). ** Response Variables: Votes. To accurately predict the winning political party, we will fit 3 models for votes for the 3 major political parties [11], Liberal, Conservative, New Democratic Party (NDP), respectively.

We choose Liberal party, Conservative party and NDP as our 3 response variables because these parties have the highest support rate in current years [11]. Thus, predicting these 3 responses can help us have a general prediction for the next election.

Through incorporating Logistic Regression with MRP, we aim to generate accurate election predictions. *We hypothesize that Age, Sex, Province, and Family Income level will be potent predictors in predicting the actual election results.*

We will conduct thorough analysis to support our hypotheses. In the following report, we will first introduce the datasets that we used for model-building and prediction-generating with data cleaning and preliminary analysis results presented. We will then showcase the methods used according to the data features. We will present our predictions in the results section. Finally, we will conclude the analysis with a brief discussion on our results.

Data

Data collection process:

The data collection process for this research project involved several key stages to ensure the integrity and coherence of the dataset used for analysis:

1. Survey Codebook Analysis: The initial step involved a detailed examination of the "Canadian Election Study 2021 Survey Codebook." This resource was utilized to identify specific questions and data variables believed to have a significant impact on the outcomes of the Canadian election.
2. Data Extraction from Census Data: Subsequently, an exploration of the "gss_clean.csv" file within the census data was carried out. The objective was to identify and extract data variables that corresponded to those found in the survey codebook. This step ensured that the data collected from the census data source was aligned with the variables of interest identified in the survey codebook, so that we can use Census Data to predict election using model fitted by survey data.
3. Data Alignment and Validation: The process was verifying that the data extracted from the census data aligned in meaning and context with the variables outlined in the survey codebook. This validation step aimed to eliminate discrepancies and confirm that the data collected accurately represented the intended information.

Cleaning process

This process involves cleaning both the “survey_data” and “census_data” datasets separately.

Two clean datasets were created:

“survey_data_clean” containing the clean data from “survey_data”.

“census_data_clean” containing the clean data from “census_data”.

For “survey_data_clean”:

1. Filter the “survey_data” to include only respondents who are 18 years of age or older and have citizenship status equal to 1, indicating the person has citizenship in Canada.
2. Create binary variables “vote_liberal”, “vote_Conservative”, and “vote_ndp” based on the values in the “cps21_votechoice” variable, indicating whether a respondent voted for the Liberal, Conservative, or NDP parties. (Yes = 1, No = 0)
3. Map the values in the “cps21_province” variable. To corresponding province to ensure consistency with the Census data, we dropped the people in Northwest Territories, Nunavut, and Yukon. And name it “province”.
4. Categorize respondents’ ages into predefined age ranges and name it “age”.
5. Map family income from “cps21_income_number” to income groups and name it “income_family” .
6. Map gender from “cps21_genderid” to “Male” and “Female” and remove those who respond as non-binary. Then name it “sex”.
7. Select specific variables of interest, including the three binary vote variables, “province”, “age”, “income_family”, and “sex”.
8. Remove rows with missing values.

For “census_data_clean”: 1. Filter the “census_data” to include only respondents who are 18 years of age or older and have citizenship status “By birth” or “By naturalization”, indicating the person has citizenship in Canada.

2. Categorize visitors’ ages into predefined age ranges. The age ranges is consistent with cleaned survey data.
3. Select specific variables “province”, “age”, “income_family”, and “sex”.
4. Remove rows with missing values.

Description of Important Variables:

vote_liberal: Binary variable indicating whether the respondent supports the Liberal Party (1 for support, 0 for not).

vote_Conservative: Binary variable indicating whether the respondent supports the Conservative Party (1 for support, 0 for not).

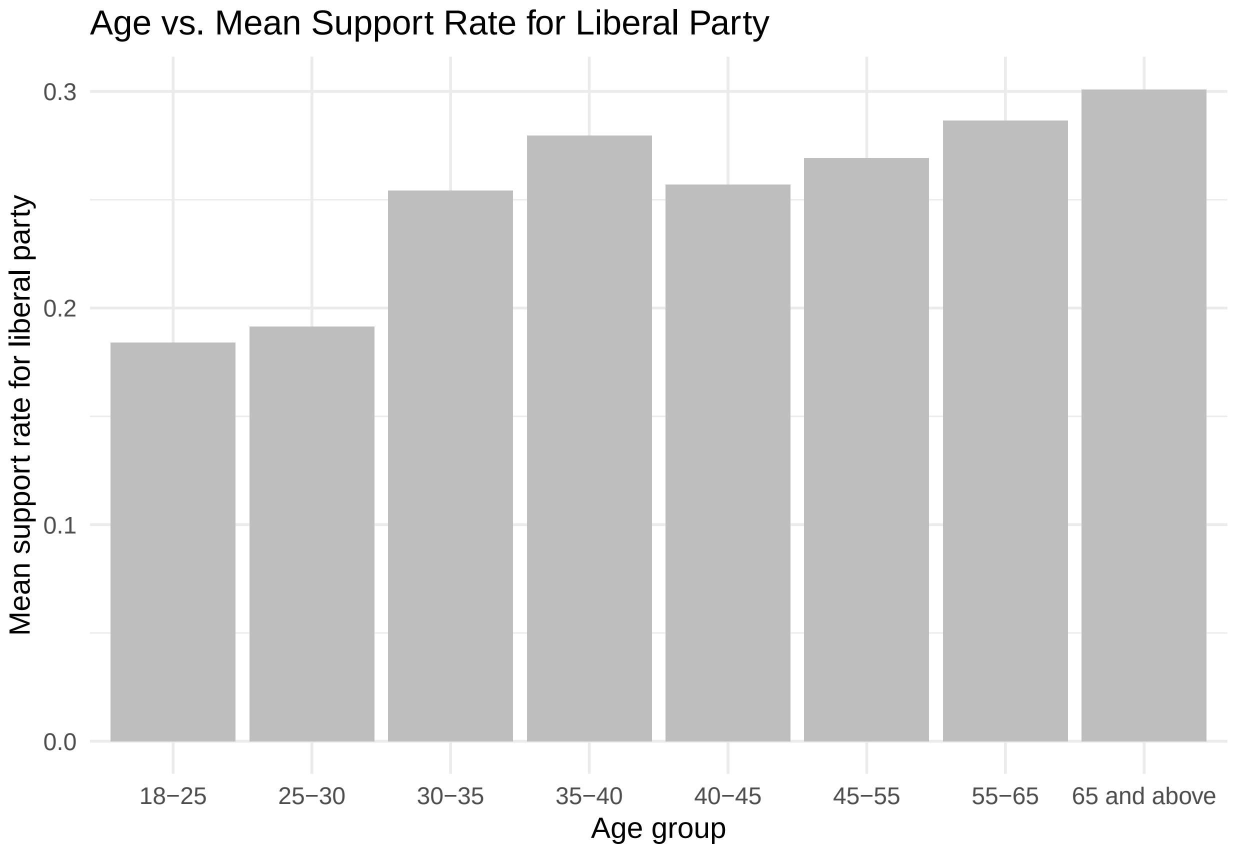
vote_ndp: Binary variable indicating whether the respondent supports the New Democratic Party (NDP) (1 for support, 0 for not).

province:This variable categorizes respondents based on their province of residence within Canada.

age: This variable groups respondents into age ranges to explore voting behavior by age groups. Age categories include "18-25", "25-30", "30-35", "35-40", "40-45", "45-55", "55-65", and "65 and above".

income_family: This variable classifies respondents based on their family income levels, providing insights into the relationship between income and voting behavior. Income categories include "Less than $25,000," "$25,000 to $49,999," "$50,000 to $74,999," "$75,000 to $99,999," "$100,000 to $124,999," and "$125,000 and more."

sex: This variable categorizes respondents by sex. Respondents are grouped into "Male" and "Female" two categories.

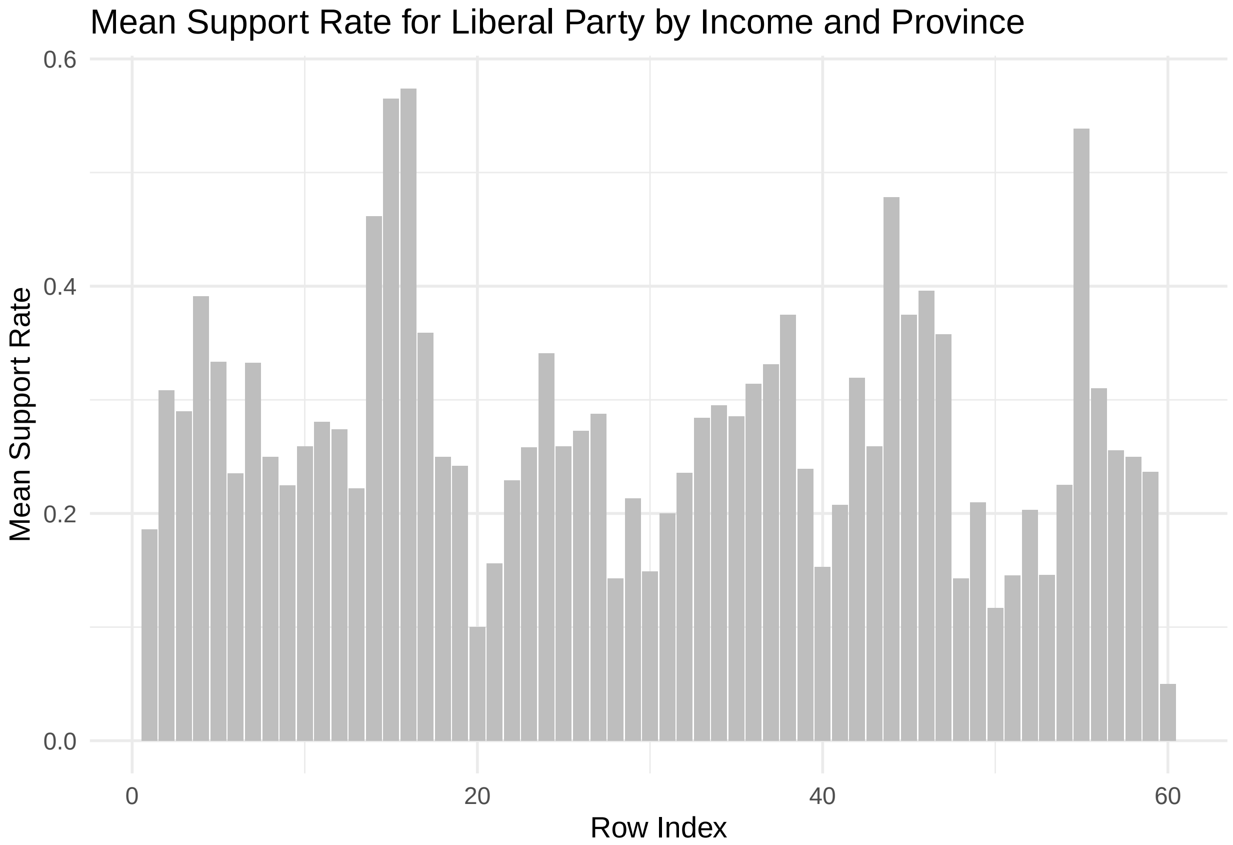


The graph titled "Age vs. Mean Support Rate for Liberal Party" offers a detailed perspective on the relationship between age groups and their respective mean support rates for the Liberal Party from survey data. From age "18-25" to age "35-40", there is a growing level of support, from approximately 17% to 28%. Respondents aged "40-45" demonstrated a support rate of around 25%, showing a slight decrease in support compared to the previous age category. In the age "45-55" , support for the Liberal Party rebounded, with approximately 26% of respondents favoring this political choice. Then the level of support is growing to 30% at age "65 and more". Overall, different age groups have different mean support rates for Liberal Party.

Table 1: Summary of Support for Liberal Party by Gender

sex	Counts	Number of Voting liberal party	Support rate
Female	7912	2079	0.2627654
Male	6545	1793	0.2739496

The table "Summary of Support for Liberal Party by Gender" offers a comprehensive overview of support for the Liberal Party in a Canadian election, categorized by sex. This table illustrates that there is no significant difference in support for the Liberal Party between men and women, respectively 27.4% and 26.3%.



The graph titled "Mean Support Rate for Liberal Party by Income and Province" visually presents key insights regarding the level of support for the Liberal Party within the different combinations of income and province. The x-axis of the graph is labeled as "Row Index," each row index indicating one combination of income and province. The y-axis is labeled as "Mean Support Rate," representing the average level of support for the Liberal Party in each gourp.

The graph's content is derived from a dataset that has been summarized and grouped by both income family and province. For each combination of income and province, the mean support rate for the Liberal Party is calculated. This allows for the identification of variations in support rate based on the income levels and provinces of respondents.

This graph is valuable because it provides a clear visual representation of how income and province impact the average level of support for the Liberal Party. It allows for the identification of regions or income brackets where the party enjoys stronger support. Thus, multilevel regression may be a good choice in this case.

Table 2: Proportion of Respondents by Province in Survey and Census Data

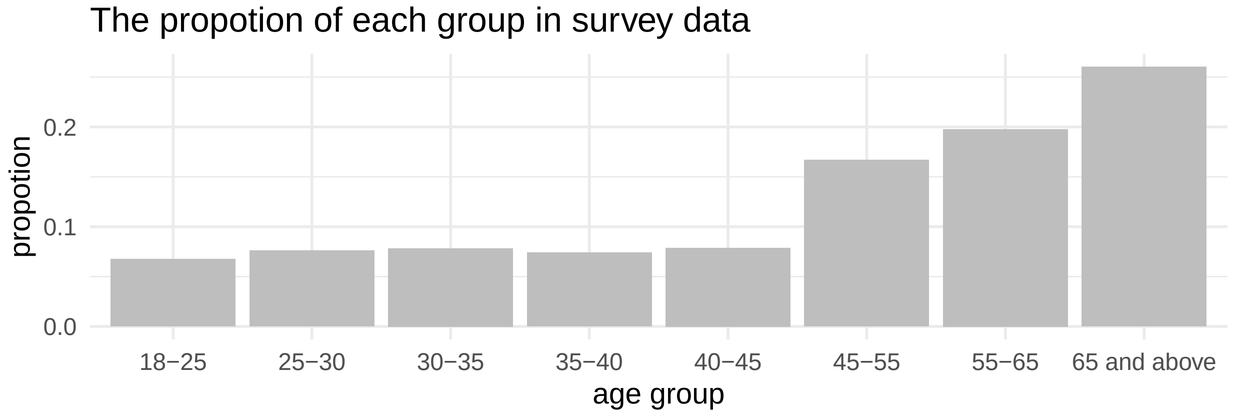
province	propotion in survey data	propotion in census data
Alberta	0.1186277	0.0824829
British Columbia	0.1081829	0.1198317
Manitoba	0.0379747	0.0567596
New Brunswick	0.0187452	0.0664913
Newfoundland and Labrador	0.0094764	0.0552867
Nova Scotia	0.0246939	0.0716465
Ontario	0.3569897	0.2698054
Prince Edward Island	0.0029052	0.0347712

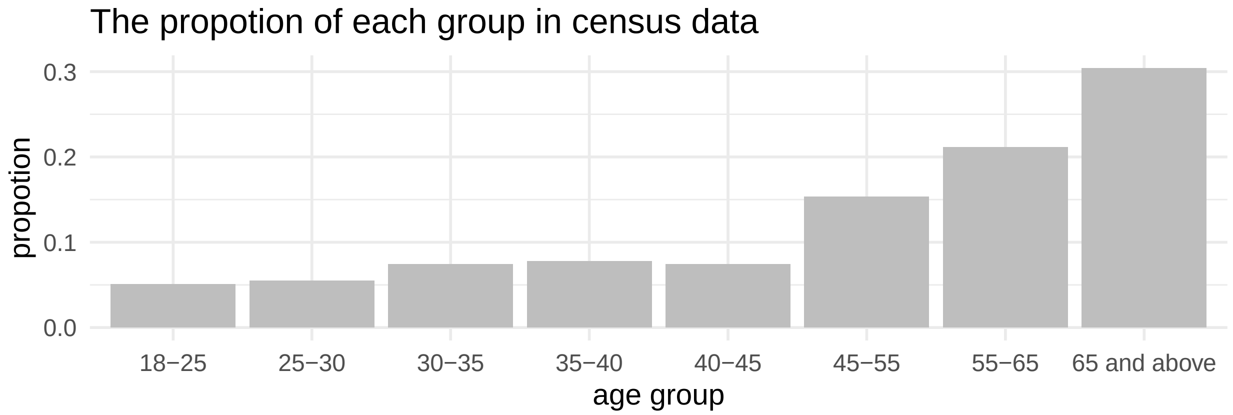
province	propotion in survey data	propotion in census data
Quebec	0.3012382	0.1875855
Saskatchewan	0.0211662	0.0553393

The table titled ‘Proportion of Respondents by Province in Survey and Census Data’ offers a comparative view of the distribution of respondents across different provinces based on data collected from both survey and census sources.

This table plays a crucial role in the analysis because by examining the proportion of respondents by province in both survey and census data, which can help us assess the accuracy and completeness of the survey dataset, and potential need for adjusting differences in non-probability sampling and non-response bias.

From the table we can see significant difference of shares between survey data and census data. Survey data may not be representative for population therefore multilevel regression with post-stratification may be helpful.





The graph above offers a visual representation of the distribution of respondents across different age groups for both survey dataset and census datset. The x-axis of the graph categorizes respondents into distinct age groups and the y-axis represents the proportion of respondents within each age group in the survey and census dataset.

It help us visualize the distribution of participants across different age categories between the survey and census dataset, which is crucial for identifying potential biases in the survey data so that reminding us adjust for differences in non-probability sampling and non-response bias.

However, from the graph, there is no significant difference in the age distribution between two datasets. Thus, age can be considered as a dependent variable instead of mutilevel effect in our model.

Methods

To predict which parties will win the election, we fit a logistic regression with random-effect on intercept using the glmer function from the "lme4" package and logit as link function.

This specific model is designed to analyze the relationship between the dependent binary variable (vote the party or not) and several independent variables, including age, sex, and a random effect related to the combination of province and family income.

The 5 binary response variables denote whether an individual voted for the that corresponding party (1 for yes, 0 for no).

Thus, we will have 5 models with same structure but different responses. Each of them can estimate the proportion of voting its corresponding party.

We will apply post-stratification for the estimates from 5 models to adjust for the issues of non-response or non-probability sampling. After that, we can get 5 $\hat{y}^{PS}$: the overall weighted estimated proportion of voting for each party. The party with the highest proportion will be predicted to win the election.

Model Specifics

We fitted a logistic mixed model to predict the proportion of voting with age and sex (vote ~ age + sex). The model included a random effect related to the combination of province and family income as random effects on intercept.

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}} + \beta_8 SEX_{\text{ij}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

p represents the proportion of voting for the specific party. β_{0j} represents a Random Intercept that affected by the combination of province and income_family. I_{ij} is the indicator that if the observation is in the particular age group (Yes = 1, No = 0). SEX_{ij} is a categorical variable: male or female. β_1 to β_8 are the coefficients of these variables.

Since we want to fit a logistic regression with random-effect on intercept, we will have the following assumptions:

The response variable is binary: We are very certain that the survey data satisfies this assumption.

Independence of Observations: We believe that each observation in the dataset should be independent of the others.

No muticollinearity issue and No outilers issues: Since all of our variables are categorical, we believe there are no such issues.

Appropriate Random Effects Structure: The specification of random effects, in this case, (1|province:income_family), should accurately reflect the structure of the data. We need to ensure that it is justified and that it accounts for any correlations or nesting in the data

After we build the initial model, we will consider how to choose a better model or drop the variable(s) that does not significantly impact the response. We will use AIC and BIC as critera to compare different models and choose the one that best balances goodness of fit. If AIC and BIC are lower, after a change for the model, it typically indicates an improvement in the model's fit to the data.

Post-Stratification

Post-stratification is a technique used in survey analysis to improve the accuracy of estimates by adjusting for differences in non-probability sampling and non-response bias. It involves dividing the population into strata based on certain characteristics (variables) and then weighting the observations in each stratum to account for different estimates and sum them up to get a overall average.

In our logistic regression with random-effect on intercept for predicting voting proportion, we can apply post-stratification.

In order to estimate the proportion of voting one party, we group our cleaned census data by the variables we use as predictors in our regression models. For example, if we use age, province and income_family as our predictors, then we will group our data by these predictors. Each rows after grouping is cell/stratum and has its own combination of age group, province and family income level. Use summarise function to record the stratum size for each stratum.

$$\hat{y}^{PS} = \frac{\sum N_j \widehat{y}_j}{\sum N_j}$$

Then, fit our regression model for each stratum to get the estimates and weight these estimates by its sample size based on the formula above. $\widehat{y}_j$ is the estimated proportion of voting one specific party for jth stratum. N_j is the sample size for jth stratum. $\sum N_j$ is the total size in cleaned census data. Last but not least, $\hat{y}^{PS}$ is the overall weighed proportion of voting one specific party by Post-Stratification.

Since we have 5 regression models, we will get the proportion of voting the 5 parties by Post-Stratification. The party with the highest proportion will be predicted to win the election.

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

Results

Since the models for predicting all 5 major political parties will be the same, for simplicity, we will showcase our model selection process using the model for Liberal party only.

After we fit the initial model:

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}} + \beta_8 SEX_{\text{ij}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

The outcome of the model for liberal party is showed on this table:

Table 3: Summary of model 1 for liberal party

Parameter	Coefficient	95% CI	z	p	Effects
(Intercept)	-1.51	[-1.71, -1.31]	-14.66	< .001	fixed
age [25-30]	0.05	[-0.17, 0.28]	0.48	0.633	fixed
age [30-35]	0.37	[0.16, 0.58]	3.40	< .001	fixed
age [35-40]	0.51	[0.29, 0.72]	4.66	< .001	fixed
age [40-45]	0.41	[0.20, 0.62]	3.77	< .001	fixed
age [45-55]	0.47	[0.28, 0.65]	4.85	< .001	fixed
age [55-65]	0.57	[0.38, 0.75]	6.04	< .001	fixed
age [65 and above]	0.65	[0.47, 0.83]	7.12	< .001	fixed
sex [Male]	-0.03	[-0.10, 0.05]	-0.70	0.484	fixed
province:income_family	0.39				random

Parameter	Coefficient	95% CI	z	p	Effects
AIC	16535.3				
BIC	16611.1				
logLik	-8257.6				

We choose age and sex as the fixed effects and the combination of province and family income level as the random intercept. It is noticeable that the p-value for predictor sex is quite large, which means we have strong evidence that sex is not a significant predictor for the response. Thus, we drop the variable sex. And refit the model.

There is our reduced model look like:

$$log(\frac{p}{1-p}) = \beta_{0j} + \beta_1 I_{\text{ij}}^{\text{25-30age}} + \beta_2 I_{\text{ij}}^{\text{30-35age}} + \beta_3 I_{\text{ij}}^{\text{35-40age}} + \beta_4 I_{\text{ij}}^{\text{40-45age}} + \beta_5 I_{\text{ij}}^{\text{45-55age}} + \beta_6 I_{\text{ij}}^{\text{55-65age}} + \beta_7 I_{\text{ij}}^{\text{over65}}$$

$$\beta_{0j} \sim \text{Normal}(\mu_{\beta_0}, \sigma^2_{\beta_0})$$

The outcome of the reduced model for liberal party is showed on this table:

Table 4: Summary of reduced model for liberal party

Parameter	Coefficient	95% CI	z	p	Effects
(Intercept)	-1.52	[-1.72, -1.32]	-14.79	< .001	fixed
age [25-30]	0.05	[-0.17, 0.28]	0.47	0.641	fixed
age [30-35]	0.37	[0.15, 0.58]	3.38	< .001	fixed
age [35-40]	0.50	[0.29, 0.72]	4.63	< .001	fixed
age [40-45]	0.40	[0.19, 0.62]	3.73	< .001	fixed
age [45-55]	0.46	[0.27, 0.65]	4.81	< .001	fixed
age [55-65]	0.56	[0.38, 0.74]	6.00	< .001	fixed
age [65 and above]	0.65	[0.47, 0.82]	7.09	< .001	fixed
province:income_family	0.39				random
AIC	16533.8				
BIC	16602.0				
logLik	-8257.9				

After dropping variable sex, the model has lower AIC and BIC. Indicating that the modified model is a better fit to the data and may be a more suitable choice for analysis or prediction. So we choose the reduced model to be our final model.

With our final model determined, we then performed poststratification using population demographics. As described in the Methods Section, we poststratified our population with respects to age groups, provinces, and family incomes. Hence, we had a total of $8 * 6 * 12 = 576$ strata (8 age groups, 6 levels of family income, 12 provinces). The estimated means ($\widehat{y}_j$) for each stratum (j) were calculated, but due to the large number of strata, we will only include a glimpse of the voting estimates for each strata in the Supplementary Materials Section.

Then, we calculated the $\hat{y}^{PS}$ for the three major political parties, i.e., Liberal, Conservative, and NPD. The results were shown in the following table:

Political Party	Predicted Proportion of Votes
Liberal	0.2786074

Political Party	Predicted Proportion of Votes
Conservative	0.2610684
NDP	0.1934651

As can be seen from the table, our models predicted that there will be 27.8607392% voters voting for the Liberal Party, 26.1068369% voting for the Conservative Party, and 19.3465058% voting for the NDP. Together, the three major political parties take approximately 73.3140818% of the total votes, suggesting little likelihood that other parties will have votes higher than any of the three. Therefore, the prediction indicates that the Liberal Party is the most likely party to win the election because it has the largest predicted proportion of voters voting for them.

Our results are plausible both methodologically and practically. From the methodological perspective, we carefully selected our potential predictor variables and calibrated our model by removing redundant variable(s). Further, we employed multi-level regression method when building our model, which takes random variations at both individual and group levels (e.g., age groups), enabling precise pinpoints at the estimates. When generating prediction, we stratified the population with respect to all of the three predictor variables (i.e., age groups, provinces, and levels of income). With the thorough and refined methods, we are confident with our predictions.

More importantly, our predictions were consistent with the real-world situation. For example, results from 2019 Canadian Election were that the Liberal Party won the election with the largest proportion of votes (39.47%), followed by the Conservative Party (31.89%) and the NDP (19.71%). The trend showed in 2019 election results was identical to the trend in our prediction. Therefore, with calibrated statistical methods and real-world confirmations, our predictions should be of relatively high reliability and plausibility.

Table 6: Lowest 7 support rate for liberal party

province	family income	support rate
Saskatchewan	Less than $25,000	0.1474716
Alberta	Less than $25,000	0.1497756
Alberta	$25,000 to $49,999	0.1554937
Saskatchewan	$125,000 and more	0.1556505
Saskatchewan	$75,000 to $99,999	0.1615728
Manitoba	Less than $25,000	0.1693897
Saskatchewan	$50,000 to $74,999	0.1793330

Table 7: Top 7 support rate for liberal party

province	family income	support rate
Nova Scotia	$125,000 and more	0.4406798
Newfoundland and Labrador	$125,000 and more	0.3717751
Newfoundland and Labrador	Less than $25,000	0.3688540
New Brunswick	$75,000 to $99,999	0.3640932
Ontario	$125,000 and more	0.3371434
Ontario	$75,000 to $99,999	0.3323784
New Brunswick	$125,000 and more	0.3281568

From the table of Lowest 7 support rate for liberal party, we notice that people in Saskatchewan and Alberta are not a supporter of the Liberals, especially for the people with low income.

From the table of Top 7 support rate for liberal party, liberal party have highest support rate at Nova Scotia, New Brunswick and Ontario. It is noticeable that middle upper class and people with high income are more likely to vote liberal party.

Conclusions

In our analysis, we performed MRP with logistic regression to predict election results. Specifically, we hypothesized that age, sex, family income, and provinces will be potent predictors for forecasting voting results. We therefore constructed our model with these key predictors and dropped the redundant variable, sex, to simplify our model. For each model, we had a binary outcome variable, voting vs not voting. Since we were generating predictions for the three major political parties in Canada, we built three models for predicting their shares of votes respectively. According to our results, Liberal party is predicted to win the election with 27.8607392% votes, exceeding the votes for all other major political parties in Canada. Detailed demographic analyses suggested that people with middle or higher family income and people who live in Ontario, Nova Scotia, and New Brunswick were more likely to vote for the Liberal party.

Our work contributed to the election forecast in both statistical and political aspects. First, our work demonstrated an exemplified attempt to address the problems with complicated data sets like the election prediction datasets. Specifically, we employed MRP which accounts for variations at both individual and group levels and allows us to poststratify the population and to precisely model the votes from each stratum. This helps mitigate the influences of non-representative samples in generalizing predictions to population. Second, our work forecasted the proportions of votes for the three major political parties in Canada, contributing to the understanding and awareness of public opinions and requests.

However, there are some limitations with our prediction process:

- Throughout the data cleaning process, certain variables were dropped or merged. Specifically, we removed some variables with missing data, which may cause some biases. In order to connect survey data and census data, certain variables were merged. For example, the variable "family income" represented the income number for each individual in the survey data and income categories in the census data. In order to connect them, we mapped the income numbers from the survey data to corresponding income categories.
- The available ces-survey data is from 2021, while gss-census data dates are from 2017. Given our projection is for 2025, the lack of up-to-date data poses challenges, potentially leading to inaccuracies in forecasts related to population shifts or unforeseen events.
- Our predictive model relies on a limited set of predictors, namely "province," "income families," and "age." This simplicity may overlooks crucial predictors, such as education level.

Future work should address these limitations and explore more methods to generate predictions. For example, future work can add more predictor variables, including but not limited to education level, employment status, etc. At the same time, expand the result variables, such as adding political parties. This approach can provide insights into the trends and preferences of specific political parties. In terms of data update, in order to make the model more relevant to time, we will seek the latest survey data sets of multiple years. This can accurately reflect political dynamics, population changes, etc., and potentially improve the model's predictive accuracy. Frequentist-based methods are currently used, and we are keen to explore Bayesian methods in the future. This shift could provide a more flexible framework to adjust forecasts based on emerging data.

In conclusion, our analysis used MRP to model and predict election results. Our results suggested that Liberal party is likely to win the election with the highest share of votes, which were consistent with and supported by real-world situations. Nonetheless, we are aware of the limitations of our analysis and suggested ways to address these limitations in future analyses.

Bibliography

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

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

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

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

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

6. Belyadi, H., & Haghighat, A. (2021). Machine Learning Guide for Oil and Gas Using Python: a step-by-step breakdown with data, algorithms, codes, and applications. Gulf Professional Publishing.

7. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2021 Canadian Election Study. [dataset]

8. Downes, M., Gurrin, L. C., English, D. R., Pirkis, J., Currier, D., Spittal, M. J., & Carlin, J. B. (2018). Multilevel regression and poststratification: a modeling approach to estimating population quantities from highly selected survey samples. American journal of epidemiology, 187(8), 1780-1790.

9. Møller, B. (2003). National, societal and human security: Discussion—Case study of the Israel-Palestine conflict. In Security and Environment in the Mediterranean: Conceptualising Security and Environmental Conflicts (pp. 277-288). Berlin, Heidelberg: Springer Berlin Heidelberg.

10. Linzer, D. A. (2014). The future of election forecasting: More data, better technology. PS: Political Science & Politics, 47(2), 326-328.

11. Mario Canseco. (2023, September 7). Conservative Party Holds Six-Point Lead Over Liberals in Canada. Retrieved from Researchco: https://researchco.ca/2023/09/07/cdnpoli-sept2023/

Appendix

Generative AI Statement

We used the following generative artificial intelligence (AI) tool: ChatGPT by OpenAI, (ChatGPT 2023) Version 3.5. Accessed in September, 2023. We used it several times in our analysis report, which includes:

We used the tool in the introduction section of this assignment:
With the prompt of `What are 3 major political parties in canada?`
ChatGPT replied us: `Liberal Party and Conservative Party and the New Democratic Party`. Then we use this as one of the references to select our response variables.

We used the tool in the Data section of this assignment:
With the prompt of `How to use case_when() when I want everything else to be "Others"?`
ChatGPT replied us: it introduced case_when() function and showed us an example use of the case_when() function. The useful help is the line of code `TRUE ~ "Others"`, when we then used in our data cleaning to simplify our work.

We used the tool in the Results section of this assignment:
With the prompt of `How to add row headers to a kable?`
ChatGPT replied us: it showed us an example of using kable to add row names, with the key function `add_rownames()`

Supplementary Materials

Below is part of the table for the estimated votes from each of the stratum:

```r
glimpse(census_data_counts)
```

```
## Rows: 480
## Columns: 7
## $ age                   <chr> "18-25", "18-25", "18-25", "18-25", "18-25", "18~
## $ province              <chr> "Alberta", "Alberta", "Alberta", "Alberta", "Alb~
## $ income_family         <chr> "$100,000 to $ 124,999", "$125,000 and more", "$~
## $ n                     <int> 10, 29, 17, 13, 11, 16, 8, 33, 12, 9, 13, 30, 6,~
## $ estimate_liberal      <dbl> 0.1343807, 0.1920405, 0.1108544, 0.1388910, 0.14~
## $ estimate_Conservative <dbl> 0.32757259, 0.31340549, 0.24111398, 0.29887653, ~
## $ estimate_ndp          <dbl> 0.4177400, 0.3639308, 0.5139075, 0.4560058, 0.50~
```