Science and Technology and Social Development Proceedings Series

Article

2024 International Conference on Education, Economics, Management, and Social Sciences (EMSS 2024)

Prediction of Canadian Federal Election Results Based on Multilevel Regression and Post-Stratification

Xiang Lai [1,*]

[1] University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada
* Correspondence: Xiang Lai, University of Toronto, 27 King's College Circle, Toronto, Ontario M5S 1A1, Canada

Abstract: In democratic countries like Canada, elections provide eligible citizens (aged 18 or older) the opportunity to vote and elect their leader. Since different political parties have distinct ideologies, election outcomes have significant societal impacts, making election result predictions crucial. This study aims to predict whether the Liberal Party will maintain its victory in the 2025 Canadian federal election using a multilevel regression model combined with post-stratification. The data for this research comes from the 2021 Canadian Election Study (CES) and the General Social Survey (GSS), with the cleaned datasets including variables such as age, gender, education, and province. Through the constructed multilevel logistic regression model and post-stratification adjustments, the results show that approximately 26.63% of Canadian citizens will vote for the Liberal Party in the next Canadian federal election. This prediction aligns with the hypothesis that the Liberal Party will not win the upcoming federal election. However, some variables in the model are not statistically significant, and the data is somewhat outdated. Future research should consider incorporating more variables and updated data.

Keywords: election prediction; multilevel regression; post-stratification; Canadian federal election; Liberal Party

Published: 03 October 2024

1. Introduction

In democratic nations such as Canada, elections offer eligible citizens (aged 18 or older) the opportunity to vote and elect their leader. The leader represents a particular party, and since different parties have different political ideas, the result of an election has important consequences for society; consequently, predicting the election result is necessary.

Among the several political parties in Canada, the Liberal and Conservative parties are the largest, with the Liberal Party currently holding the majority. This report aims to predict whether the Liberal Party will maintain its winning in the next Canadian federal election (tentatively 2025) using a multilevel regression model with post-stratification method. The hypothesis is the Liberal Party will not win in the upcoming Canadian federal election.

2. Data

In this report, the General Social Survey (GSS) is the "census" data and data from the 2021 Canadian Election Study (CES) is the "survey" data.

Initially, there are 20,602 observations and 81 variables in the census dataset while there are 20,968 observations and 1,062 variables in the survey dataset. As the variable names and formats are not well set, so the two datasets are cleaned in order to be consistent to apply the methods later.

For the census data, the observations that are not eligible to vote (age less than 18 and not Canadian citizen) are removed. Meanwhile, the variables that highly associated with election result are kept in the dataset; they are 'age', 'sex', 'province', 'education', and 'citizenship_status'. Furthermore, observations contain missing values are removed to avoid bias. After the cleaning process, there are 18,818 observations and 5 variables in the census dataset.

For the survey data, based on the project goal,'cps21_votechoice' which represent if the respondent will vote to the Liberal Party and the variables that represent the similar information as the cleaned census data variables ('cps21_age', 'cps21_genderid', 'cps21_province', 'cps21_education', and 'cps21_citizenship') are selected. Then, the values in the variables are converted from number to their real meanings. Meanwhile, those observations with responses "Don't know" are removed to avoid bias. Furthermore, the varoables are renamed as 'vote_liberal','age','gender','province', 'education', and 'citizenship_status'. After the cleaning process, there are 14,468 observations and 6 variables in the census dataset.

In order to make the two datasets consistent, several adjustments are made. First, in the census dataset,'age' is rounded to the nearest whole number. Meanwhile, in the survey data, there has a "Non-binary" category in the 'gender' variable and it differs from the 'sex' variable in the census data. The "Non-binary" category has proportion 0.43% and is removed to transform the variable into a binary 'sex' variable. Furthermore, the 'province' variable in the survey data has three extra categories ("Northwest Territories", "Nunavut", and "Yukon") compared to the census data, they all have very small proportions (0.06%, 0.03%, and 0.14%), so these categories have removed for consistency. Lastly, the 'education' variable has very different categories in the two datasets, and they both been re-categorized into five levels: "< High School", "High School",

"College", "Bachelor", and "> Bachelor". These adjustments have made the variables' names, formats, and values consistent across the two datasets (see Table 1).

Table 1. Adjusted Variables for Consistency Across Datasets.

Variable	Description
age	discrete numerical (whole number); age of the respondent
sex	categorical with 2 levels ("Female" and "Male"); sex of the respondent
province	categorical with 10 levels ("Alberta", "British Columbia", "Manitoba", "New Brunswick", "Newfoundland and Labrador", "Nova Scotia", "Ontario", "Prince Edward Island", "Quebec", and "Saskatchewan"); current province of the respondent
education	categorical with 5 levels ("< High School", "High School", "College", "Bachelor", and "> Bachelor"); highest education of the respondent
vote_liberal	binary with 2 levels (0 and 1); if the respondent will vote for Liberal (yes = 1, no = 0)

Next, the characteristics of the categorical variables in the two datasets are explored.

Table 2. Proportion Table of Sex (For Each of the Two Datasets).

	Female	Male
Census	54.66%	54.66%
Survey	54.71%	45.29%

Table 2 shows that the proportions of the sex category in the two datasets are approximately the same, with 55% female and 45% male.

Table 3. Proportion Table of Province (For Each of the Two Datasets).

	AB	BC	MB	NB	NL	NS	ON	PE	QC	SK
Census	8.26%	11.98%	5.66%	6.66%	5.52%	7.15%	27.04%	3.47%	18.74%	5.52%
Survey	11.87%	10.82%	3.79%	1.87%	0.94%	2.47%	35.70%	0.29%	30.12%	2.12%

Table 3 shows that the proportions of each province category in the two datasets are not the same and most of the respondents in the survey dataset are from Ontario and Quebec.

Table 4. Proportion Table of Highest Education (For Each of the Two Datasets).

	< High School	High School	College	Bachelor	> Bachelor
Census	13.90%	24.54%	23.21%	29.58%	8.78%
Survey	2.36%	12.89%	30.47%	40.06%	14.22%

Table 4 shows that the proportions of each education category in the two datasets are not the same and the survey dataset has more higher-educated respondents.

Table 5. Proportion Table of If Vote for Liberal.

Liberal	Not Liberal
73.19%	26.81%

In the survey dataset, 73.19% of the respondents will vote for the Liberal Party, while the remaining 26.81% of the respondents will not. Then, the joint distributions for whether to vote for the Liberal party and the variables (sex, province, and education) are explored to determine if they are factors associated with voting for the Liberal Party.

Table 6. Frequency Table for Sex and If Vote for Liberal.

	Female	Male
Not Liberal	5820	4745
Liberal	2078	1792

Table 6 shows that different sex categories have different opinions on voting for the Liberal Party, which indicates that it is a factor associated with voting for the Liberal Party.

Table 7. Frequency Table for Province and If Vote for Liberal.

	AB	BC	MB	NB	NL	NS	ON	PE	QC	SK
Not Liberal	1368	1156	413	178	83	232	3483	32	3354	266
Liberal	346	406	134	92	53	124	1671	10	994	40

Table 7 shows that different province categories have different opinions on voting for the Liberal Party, which indicates that it is a factor associated with voting for the Liberal Party.

Table 8. Frequency Table for Education and If Vote for Liberal.

	< High School	High School	College	Bachelor	> Bachelor
Not Liberal	269	1435	3415	4079	1367
Liberal	71	426	984	1704	685

Table 8 shows that different education categories have different opinions on voting for the Liberal Party, which indicates that it is a factor associated with voting for the Liberal Party.

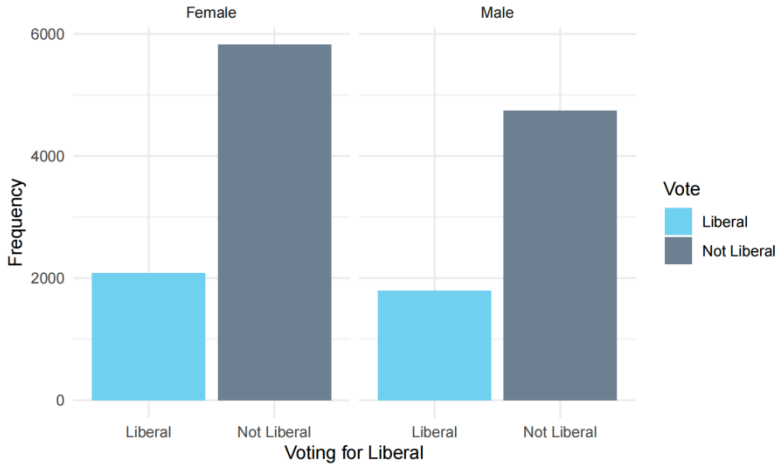

Figure 1. Split Bargraph of Voting for Liberal by Sex.

Figure 1 shows that the sex category is associated with voting for the Liberal Party because the bar heights for "Liberal" in each category are approximately the same, but the height for "Not Liberal" in males is lower than in females, which means males tend to vote for the Liberal Party based on the survey data.

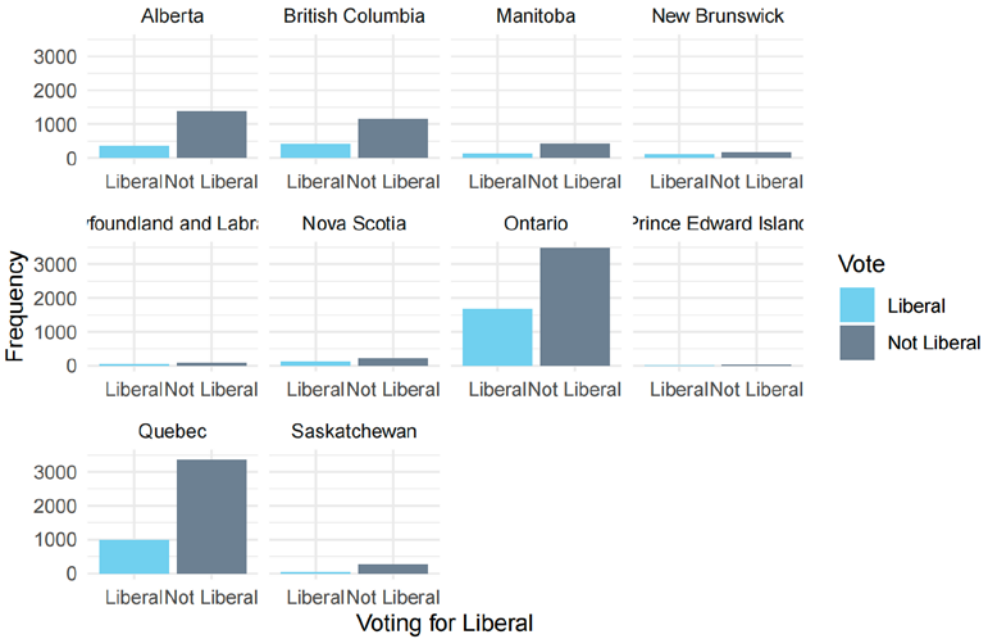

Figure 2. Split Bargraph of Voting for Liberal by Sex.

Figure 2 shows that the province category is associated with voting for the Liberal Party because the bar distributions are very different across provinces, and Ontario citizens tend to vote for the Liberal Party based on the survey data.

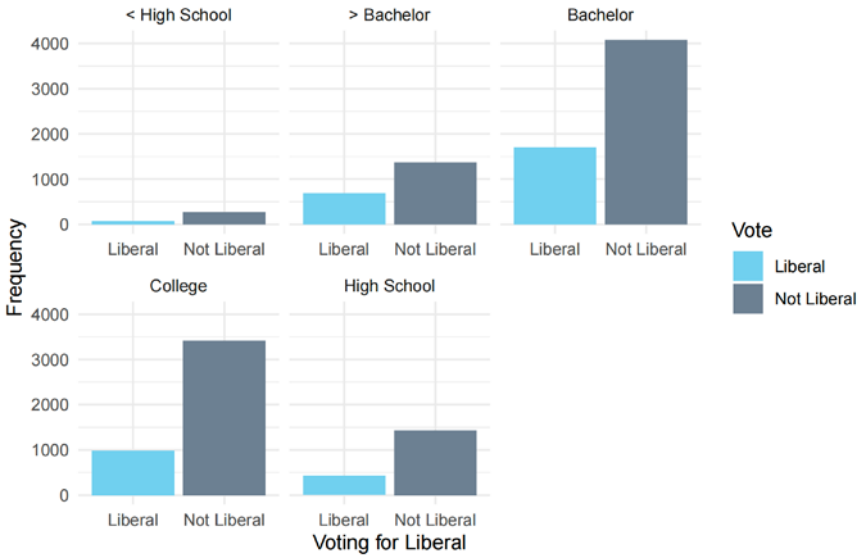

Figure 3. Split Bargraph of Voting for Liberal by Education.

Figure 3 shows that the education category is associated with voting for the Liberal Party because the bar distributions are very different across education levels, and higher-educated citizens (those with "Bachelor" and "> Bachelor" degrees) tend to vote for the Liberal Party based on the survey data.

3. Methods

Based on the research goal of predicting whether the Liberal Party will maintain its winning streak in the next Canadian federal election, a Multilevel regression model with a post-stratification method are used.

Multilevel refers to the categorization of variables into multiple hierarchies, there are individual level (level 1) and group level (level 2). In this case, the individual level has 'age', 'sex', and 'education' variables while the group level has the 'province' variable. The data is divided into several cells based on these combinations and each of them has a estimate value. Then, post-stratification is used to adjust the weights of estimates for each cell, making the sample more representative of the actual population.

4. Model Specifics

Since the response variable is binary, so Multilevel Logistic Regression is used. The individual level has 'age' 'sex' and 'education' variables while the group level has the 'province' variable. Assume the assumptions for logistic regression are all satisfied that the relation among logit and predictors are linear; there is no multicollinearity; and there is no strong influential outliers.

The individual level regression model is

$$\log\frac{p}{1-p} = \beta 0 + \beta 1\ age\ + \beta 2\ I(Male) + \beta 3\ I(>\ Bachelor) + \beta 4\ I(Bachelor) + \beta 5\ I(College) + \beta 6\ I(HighSchool)$$

where

- p represents the probability of voting for Liberal Party;
- β1 represents the log odds change of voting for Liberal Party as age increase by one unit, keeping other predictors unchanged;
- β2 represents the log odds change of voting for Liberal Party as sex changes from "< High School" to "> Bachelor", keeping other predictors unchanged;
- β3 represents the log odds change of voting for Liberal Party as sex changes from "< High School" to "Bachelor", keeping other predictors unchanged;

• $\beta 4$ represents the log odds change of voting for Liberal Party as sex changes from "< High School" to "College", keeping other predictors unchanged; and

• $\beta 5$ represents the log odds change of voting for Liberal Party as sex changes from "< High School" to "High School", keeping other predictors unchanged.

The group level regression model is

$$\beta 0 = r00 + r01W1 + r02W2 + \cdots + r08W8 + r09W9$$

where

• W1, W2 , W3 , W4 , W5 , W6 , W7 , W8 , W9 represent the indicating whether the province is "British Columbia", "Manitoba", "New Brunswick", "Newfoundland and Labrador", "Nova Scotia", "Ontario", "Prince Edward Island", "Quebec", and "Saskatchewan", respectively;

• r01 represents the intercept change in individual level model as province changes from "Alberta" to "British Columbia";

• r02 represents the intercept change in individual level model as province changes from "Alberta" to "Manitoba";

• r03 represents the intercept change in individual level model as province changes from "Alberta" to "New Brunswick";

• r04 represents the intercept change in individual level model as province changes from "Alberta" to "Newfoundland and Labrador";

• r05 represents the intercept change in individual level model as province changes from "Alberta" to "Nova Scotia";

• r06 represents the intercept change in individual level model as province changes from "Alberta" to "Ontario";

• r07 represents the intercept change in individual level model as province changes from "Alberta" to "Prince Edward Island";

• r08 represents the intercept change in individual level model as province changes from "Alberta" to "Quebec"; and

• r09 represents the intercept change in individual level model as province changes from "Alberta" to "Saskatchewan".

5. Post-Stratification

Post-stratification is used to adjust the weights of estimates for each cell, making the sample more representative of the actual population. Based on the model, the 4 categorical variables are 'age', 'sex', 'education', and 'province' that have 78, 2, 5, and 10 categories respectively. In total, there are $78 \times 2 \times 5 \times 10 = 7800$ combinations which are the cells. Using these combinations, the estimated result:

$$\hat{y}^{ps} = \frac{\sum N_j * \hat{y}_j}{\sum N_j}$$

where $\hat{y}_j$ is the proportion of voting the Liberal party in each cell, N_j is the population size for the jth cell,

and εN_j is the population size. All analysis for this report was programmed using R version 4 .0 .2.

6. Results

Table 9. Summary Table for the Regression.

	β	SE(β̂)	p-value
(Intercept)	-1.9627	0.1948	< 2e-16
age	0.0115	0.0012	< 2e-16
I(Male)	-0.0284	0.0394	0.470906
I(> Bachelor)	0.6685	0.1429	2.91e-06
I(Bachelor)	0.5323	0.1382	0.000117

I(College)	0.1123	0.1397	0.421274
I(High School)	0.1318	0.1458	0.366013

Based on Table 9, the estimated equation for voting Liberal is $\log(\frac{\hat{p}}{1-\hat{p}}) = -1.9627+0.0115age-0.0284I(Male)+0.6685I(>Bachelor)+0.5323I(Bachelor)+0.1123I(College)+0.1318I(HighSchool)$.

Then, the post-stratification is used by applying the model to each cell in the census dataset and plug in the formula:

$$\hat{y}^{ps} = \frac{\sum N_j * \hat{p}_j}{\sum N_j}$$

The resulting value for the proportion of voting for Liberal is 0.2662524 that means 26.63% Canadian citizens will vote for the Liberal Party in the Canadian federal election.

7. Conclusions

In conclusion, based on the constructed multilevel logistic regression model and post-stratification method, 26.63% of Canadian citizens will vote for the Liberal Party in the upcoming Canadian federal election, which is consistent with the hypothesis that the Liberal Party will not win in the upcoming Canadian federal election.

In terms of weaknesses, in the regression model, the 'sex' variable and two categories of the 'education' variable are not significant at significance level $\alpha = 0.05$, which leads to bias in the estimated results. Meanwhile, there might be other variables associated with the election that can be considered in the model, such as socioeconomic status. Furthermore, the datasets are from 2021, and there have been many significant changes in Canadian society, such as COVID-19, which could cause errors in the estimated results. In future studies, it would be better to use more recent datasets and consider some other potential factors in the model.

References

1. Grolemund, G. (2014, July 16) Introduction to R Markdown. RStudio. https://rmarkdown.rstudio.co m/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.co m/docs/. (Last Accessed: April 4, 1991)
4. OpenAI. (2023). ChatGPT (September 13 version) [Large language model]. https://chat.openai.com/ch at (Last Accessed: September 13, 2023)
5. Çetinkaya-Rundel M, Diez D, Bray A, Kim A, Baumer B, Ismay C, Paterno N, Barr C (2022). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.4.0, https://CRAN.R-project.org/package=openintro.
6. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686https://doi.org/10.21105/joss.01686.
7. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
8. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.