Analysising social factors influence Canadian voters' choices*

Decoding CES 2021 Vote Data by using A Multilevel Logistic Regression Model

Yichen ji

April 2, 2024

1) what was done, 2) what was found, and 3) why this matters (all at a high level). .

Table of contents

[*] Code and data are available at: https://github.com/Selinayichenji/Canadian_election_prediction.git.

1 Introduction

This report utilizes the 2021 Canadian Election Study data and the "General Social Survey, Cycle 31: Families, Public Use Microdata File Documentation and User's Guide" (Statistics Canada, 2019) to solve the emergent dilemma of diminishing vote participation rate. The primary purpose of this report is to use the data set to predict future voter turnout by differentiating throughout the province(note: not include Quebec, Yukon, Northwest Territories, and Nunavut).

Based on previous voting analysis, the main factors affecting voting include age, education level, family status(represented by the number of children), and economic well-being(represented by the family income). ("Factors Associated with Voting," 2015) Thus, the examination will focus on demographic variables, including age, gender, family income, education level, and the number of children in the vote rate corresponding to Canada's three main political parties. The three main political parties are Liberal, Conservative, and NDP(Canada's New Democrats), which occupied around 90 percent of the election 2010s. ("Canadian Federal Election Results," n.d.) The report will use logistic regression on different parties to model the participation rate. Whether the voting for particular parties or not will be recorded binary, as 1 or 0. Under this method, we assume the authenticity of our source, 2021 Canadian Election Study data, to build up the model.

The importance of this study is to reveal the complexities behind the recent decline in voter participation rates. By exploring the complex demographic factors and the voting rate of different political parties, this report could provide insights into the advantages of each of the three main parties during the election. According to past data, where liberals always take up the majority of the voting, we will assume the liberals will win the overall election. (Simon Fraser University, n.d.) Meanwhile, this report will examine how various demographic elements correlate with political affiliations, aiming to understand the challenges facing Canadian democracy comprehensively. In addition, the findings can not only predict the democratic participation rate throughout different states but also serve as a potential solution to the diminishing voting participation rate corresponding to each province, which can strengthen democracy throughout the nation.

1.1 estimand

The paper introduces the basic information contains data source, methodology, variables and measurements in Section 2. The visualization and analysis of the tendency differences in moderation by people across three social factors are presented in Section 4. And in Section 5, we discuss what we have learned, our understanding of the world, the limitations, and the next steps of our research. **?@sec-app** is for appendix and **?@sec-ref** is listed all references in this paper.

2 Data

The survey data was sourced from the "2021 Canadian Election Study" (Stephenson, Harell, Rubenson, & Loewen, 2022).

2.1 Analysis Tools

Our paper applies using the statistical programming language R (R Core Team 2023). Besides the programming tool, we also employ the following packages: readr (**?**), ggplot2 (**?**), dplyr (**?**), tidyverse (**?**), MetBrewer (**?**), knitr (**?**), tidyr(**?**), kableExtra(**?**) and grid(**?**).

2.2 Variables

> **Age**: include five different groups: 18-29 years old, 30-44 years old, 45-59 years old, 60-74 years old, and above 74 years old. We manipulate the age for census data, which has different years than the survey data.
>
> **vote_liberal/vote_conservative/vote_NDP**: the voting result on whether to vote for three different parties. (Liberal, Conservative, or NDP) The result of voting or not is recorded in binary.
>
> **Gender**: only include two types of gender, female and male, excluding non-binary or others
>
> **education_level**: including six different education levels(Less than high school, High school, Non-University, University certificate below the bachelor, Bachelor's degree, Above the bachelor).

family__income: including six different groups. (“Less than 25,000”, “25,000 to 49,999”, “50,000 to 74,999”, “75,000 to 99,999”, “100,000 to 124,999”, “125,000 and more”)

Province: 9 different states and excluding Quebec, Yukon, Northwest Territories, and Nunavut.

children__number: the number of children within the family, which includes only six levels (0,1,2,3,4 and 4+ children)

Table 1: Sample of the Cleaned Data

age	vote_liberal	vote_conservative	vote_NDP	gender	education_level	f
18-29	0	0	1	Female	University certificate below the bachelor	$
45-59	0	0	1	Female	Above the bachelor	$
45-59	0	0	1	Female	University certificate below the bachelor	$
60-74	0	1	0	Male	University certificate below the bachelor	$
45-59	1	0	0	Male	Non-University	$

2.3 Sample of cleaned data

2.4 Survey data summary

All detailed tables include summary data of 6 variables are in Section A age, gender, income, education level, province, children number

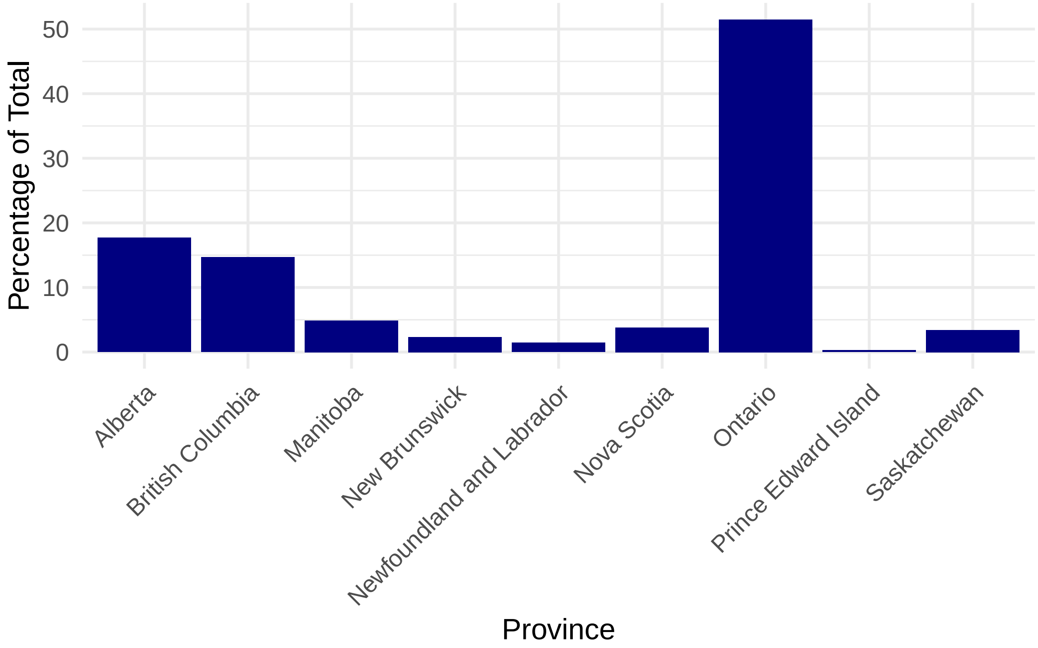
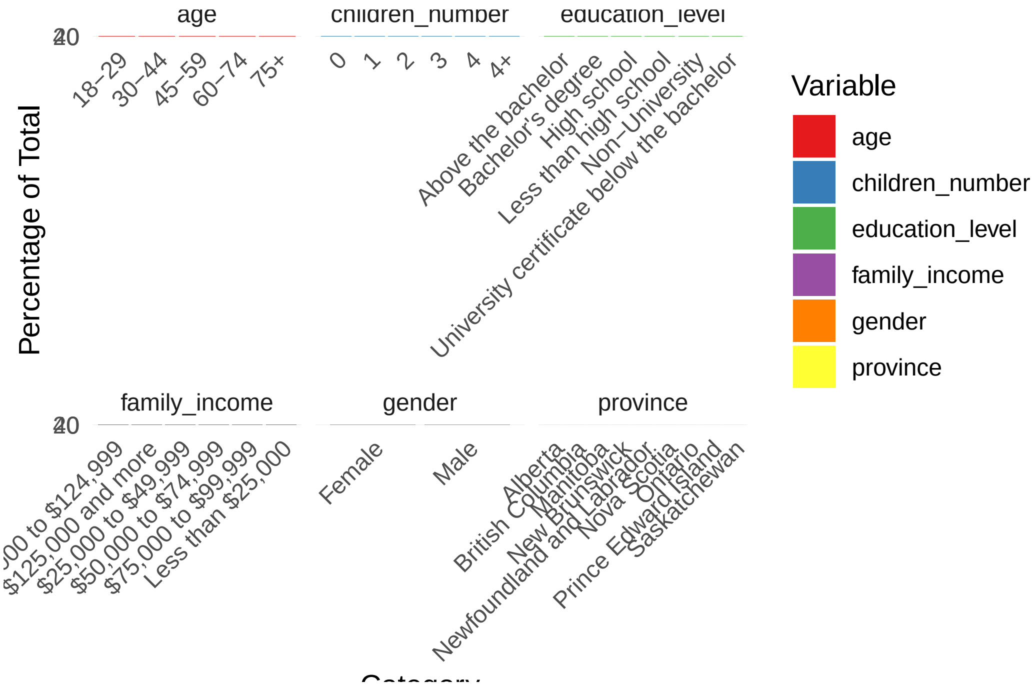


Figure 1: Population Distribution by Province in CES2021

```r
library(dplyr)
library(ggplot2)

# Define a function to calculate percentages for each category within a variable
calculate_percentage <- function(data, category) {
  data %>%
    count(!!sym(category)) %>%
    mutate(percentage = n / sum(n) * 100) %>%
    rename(Category = !!sym(category)) %>%
    mutate(Variable = category)  # store the variable name
}

# Apply the function to each of the variables you are interested in
category_list = c("age", "gender", "family_income", "education_level", "province", "children_
percentages <- lapply(category_list, calculate_percentage, data = analysis_data)

# Combine the list of data frames into one data frame
percentages_df <- bind_rows(percentages)

# Generate the plot
combined_plot <- ggplot(percentages_df, aes(x = Category, y = percentage, fill = Variable)) +
  geom_bar(stat = "identity") +
  facet_wrap(~Variable, scales = "free_x") + # allows each facet to have its own x axis
  labs(x = "Category", y = "Percentage of Total") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set1")

# Print the combined plot
combined_plot
```



2.5 Measurement

The CES 2021, consisting of an online cohort of 20,968 participants from the broader Canadian populace, was facilitated via the Leger Opinion panel. This study was structured in two distinct stages: the campaign period survey (CPS) and the post-election survey (PES). The CPS phase was further divided into three segments – "CPS", "CPS Modules", and "CPS Oversample", which were later consolidated into one dataset. Regarding the PES, all individuals who participated in the CPS were invited back for this subsequent survey after the election. Out of these, 15,069 individuals completed the PES, marking a response rate of 72%. Respondents needed to be aged 18 or over and Canadian citizens or permanent residents in order to participate.

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix C.

3.1 Multilevel Logistic Regression Model

$$
\begin{aligned}
y_i|\pi_i &\sim \text{Bern}(\pi_i) && (1)\\
\text{logit}(\pi_i) &= \beta_0 + \alpha^{\text{age}}_{a[i]} + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{education level}}_{e[i]} + \alpha^{\text{family income}}_{f[i]} + \alpha^{\text{children number}}_{c[i]} + \alpha^{\text{state}}_{s[i]} && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\alpha^{\text{age}}_{a} &\sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5 && (4)\\
\alpha^{\text{gender}}_{g} &\sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2 && (5)\\
\alpha^{\text{education level}}_{e} &\sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5, 6 && (6)\\
\alpha^{\text{family income}}_{f} &\sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, 3, 4, 5, 6 && (7)\\
\alpha^{\text{children number}}_{c} &\sim \text{Normal}(0, 2.5) \text{ for } g = 1, 2, ..., 6 && (8)\\
\alpha^{\text{state}}_{s} &\sim \text{Normal}(0, \sigma^2_{\text{state}}) \text{ for } s = 1, 2, ..., 9 && (9)\\
\sigma_{\text{state}} &\sim \text{Exponential}(1) && (10)
\end{aligned}
$$

We run the model in R using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. The `stan_glm` function from `rstanarm`

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

$$log(\frac{p_{ij}}{1-p_{ij}}) = \beta_{0j} + \beta_1 x_{ij\ age} + \beta_2 x_{ij\ gender} + \beta_3 x_{ij\ education\ level} + \beta_4 x_{ij\ family\ income} + \beta_5 x_{ij\ children\ number}$$

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

In the model, j is the province and i is the individual voter indicator.

p represents the probability of the event of interest occurring, which is the voter's probability of voting for any specific party. p_{ij} means the voting of any specific party probability of ith voter in jth province.

$log(\frac{p_{ij}}{1-p_{ij}})$ means the log of odds of voting for any specific party.

β_0 represents the intercept of the model and is the log of odds of voting for any specific party when the individual is at a certain age, with a certain gender, education level, family income and children number.

β_{0j} stands for jth province effect on the log of odds of voting for any specific party, and we assume the intercept β_{0j} in different provinces is random and follows a normal distribution with mean μ_{β_0} and variance $\sigma^2_{\beta_0}$.

β_1 represents the slope of age level in the model, then for any voter, one unit increase in age, we expect a β_1 increase in log odds of voting for any specific party. β_2 represents the average difference in log odds of voting for any specific party between Males and Females for a certain age, education level, family income and children number.

β_3 represents the average difference in log odds of voting for any specific party between groups with different education levels for a certain age, gender, family income and children number.

β_4 represents the average difference in log odds of voting for any specific party between groups with different family income levels for a certain age, gender, education level and children number.

β_5 represents the average difference in log odds of voting for any specific party between groups with different children number levels for a certain age, gender, education level and family income.

x_{ij} is a binary variable, it can only be 1 or 0 as all the variables we have are categorical, it also stands for the ith voter in jth province.

4 Results

4.1 Model Coefficients

5 Discussion

What is done in this paper? What is something that we learn about the world? What is another thing that we learn about the world? What are some weaknesses of what was done? What is left to learn or how should we proceed in the future?

Table 2: Coefficients of social factors influence vote for liberal party

	Vote for Liberal Party
(Intercept)	−1.143
	(0.262)
age30-44	0.507
	(0.139)
age45-59	0.750
	(0.143)
age60-74	0.911
	(0.147)
age75+	0.933
	(0.179)
genderMale	−0.027
	(0.078)
education_levelBachelor's degree	−0.028
	(0.118)
education_levelHigh school	−0.501
	(0.132)
education_levelLess than high school	−0.738
	(0.277)
education_levelNon-University	−0.720
	(0.130)
education_levelUniversity certificate below the bachelor	−0.188
	(0.158)
family_income$25,000 to $49,999	0.152
	(0.169)
family_income$50,000 to $74,999	0.240
	(0.162)
family_income$75,000 to $99,999	0.308
	(0.166)
family_income$100,000 to $124,999	0.174
	(0.175)
family_income$125,000 and more	0.385
	(0.168)
children_number1	0.088
	(0.112)
children_number2	−0.166
	(0.101)
children_number3	−0.245
	(0.134)
children_number4	−0.075
	(0.242)
children_number4+	−0.070
	(0.315)
Sigma[province × (Intercept),(Intercept)]	0.187
	(0.112)
Num.Obs.	3310
R2	0.060
R2 Marg.	0.042
Log.Lik.	−2066.102
ELPD	−2094.3
ELPD	[illegible]

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

B Additional data details

C Model details

C.1 Posterior predictive check

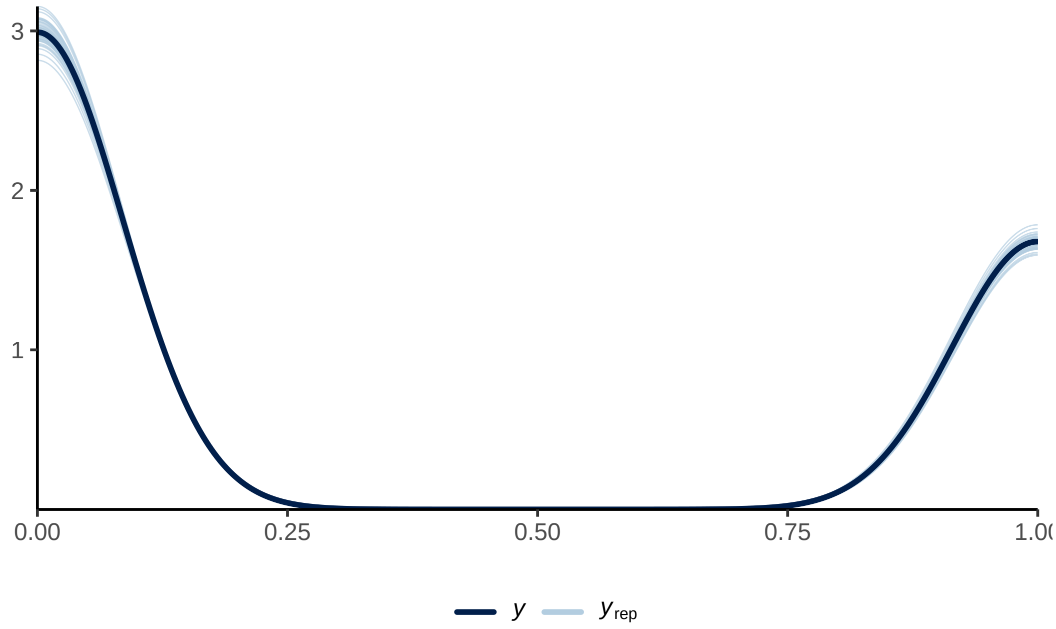


Figure 2: pp check

C.2 Diagnostics

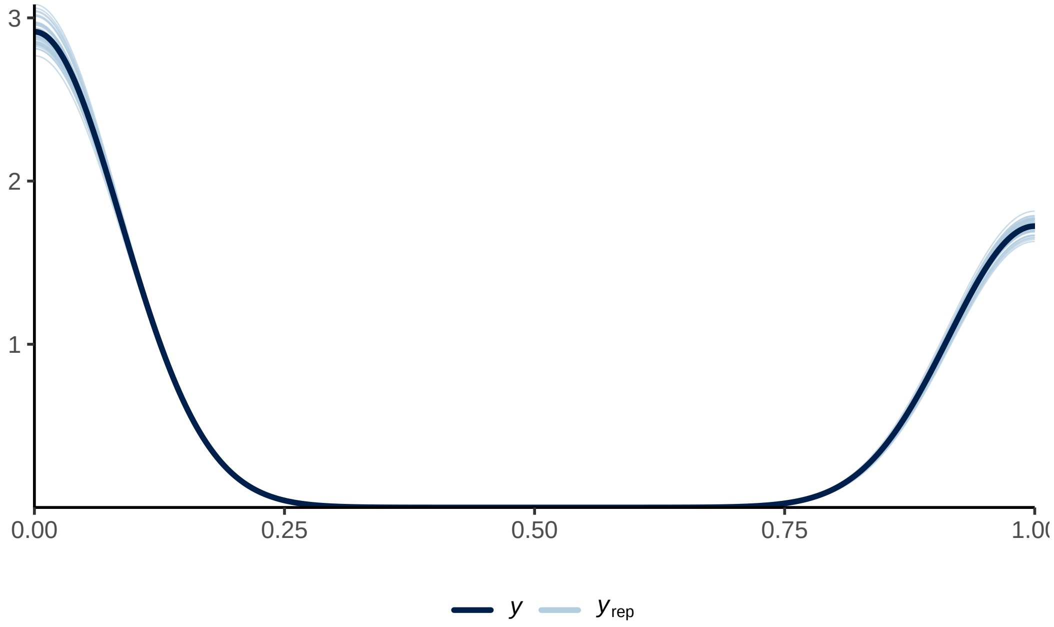


Figure 3: pp check

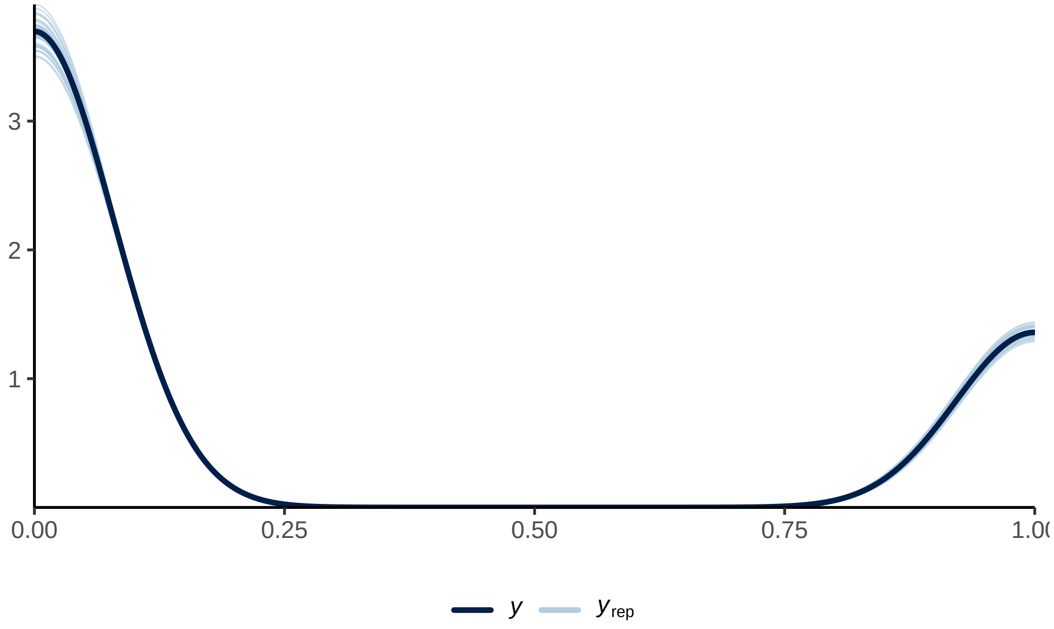


Figure 4: pp check

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.