Understanding Liberals vote base*

STA304 - Winter 2025 - Assignment 2

Ahnaf Alam

March 13, 2025

Table of contents

1 Introduction

Election surveys are important for understanding voter preferences, predicting election outcomes, and informing public policy. Survey polls are also tools for political campaigns to identify where their support is highest and where they need to focus efforts to get better election results. In this report, we aim to understand which regions support the Liberal Party of Canada and where they need to improve to gain a greater share of the vote. Specifically,

*For reproducibity, please visit https://github.com/AhnafAlam1/cdn_election we estimate the proportion of total Liberal support across Canada and run a logistic regression to differentiate Liberal popularity across different regions. We also incorporate age into this model to understand how age influences the Liberal vote share. Our data comes from the 2019 Canadian Federal Election Study (Stephenson et al. 2020), specifically focusing on web data. The analysis examines respondents' age, province, and their reported likelihood of voting, aiming to predict the likelihood of voting for the Liberal Party in the 2019 elections. Age and provincial differences are critical factors in election campaigns, and identifying these differences can help the Liberal Party efficiently allocate resources and focus on demographics and regions where they stand the best chance of winning.

The 2019 Canadian Federal Election Study is an extensive dataset that includes responses from both phone and web surveys. Despite the different collection methods, both datasets contain comparable variables on voter intention and demographic information. This dataset offers a valuable opportunity to assess the reliability of each method and explore potential biases.

This report aims to identify age and province-related differences in Liberal vote share. The findings can inform future political campaigns by identifying strengths and weaknesses, helping the party target specific demographics, and developing policies that can resonate with groups that may not traditionally support Liberal ideas. The subsequent sections outline the data preparation process (Section 2), model details (Section 3) present visual comparisons of Liberal vote share, and analyze reported voting intentions (Section 4), and discuss potential biases in the method (Section 5).We end the report with a discussion on the use of generative AI (Section 6) and ethical consideration (Section 7).

2 Data

The data used in this analysis comes from the 2019 Canadian Federal Election Study (Stephenson et al. 2020). This study examines data collected from web sources, focusing on three variables: province, likelihood of voting in the election, and political party. I use province as the stratified variable.

Figure 1 shows the number of participants in the web survey from each Canadian province. While the distribution may appear skewed, with participants from Ontario and Quebec comprising a large portion of the data, this reflects the actual population distribution across Canada. Ontario has nearly 12,000 data entries, which aligns with its status as Canada's most populous province. Conversely, the territories have fewer than 1,000 entries combined, reflecting their smaller population size. Since the data reflects population distribution, stratifying by province ensures that each province's data is analyzed in proportion to its actual population. Without stratification, provinces with larger sample sizes, such as Ontario and Quebec, could disproportionately influence the overall results. Stratification solves this problem by ensuring that each region contributes appropriately to the analysis.

Data cleaning was limited to the variables used in this study. Specifically, province, likely_to_vote, and political_party were re-encoded from numerical values to their corresponding categories. For example, the raw data coded "Ontario" as "22"; in the cleaned version, "Ontario" is coded as "Ontario."

The analysis was conducted using R (R Core Team 2023), a statistical programming language. The tidyverse package (Wickham et al. 2019) was used for data cleaning, and ggplot2 (Wickham 2016) was used for visualizations. I also used the survey package (Lumley 2010) for modeling and relied on broom (Robinson, Hayes, and Couch 2023) and ggplot2 for aesthetic improvements.

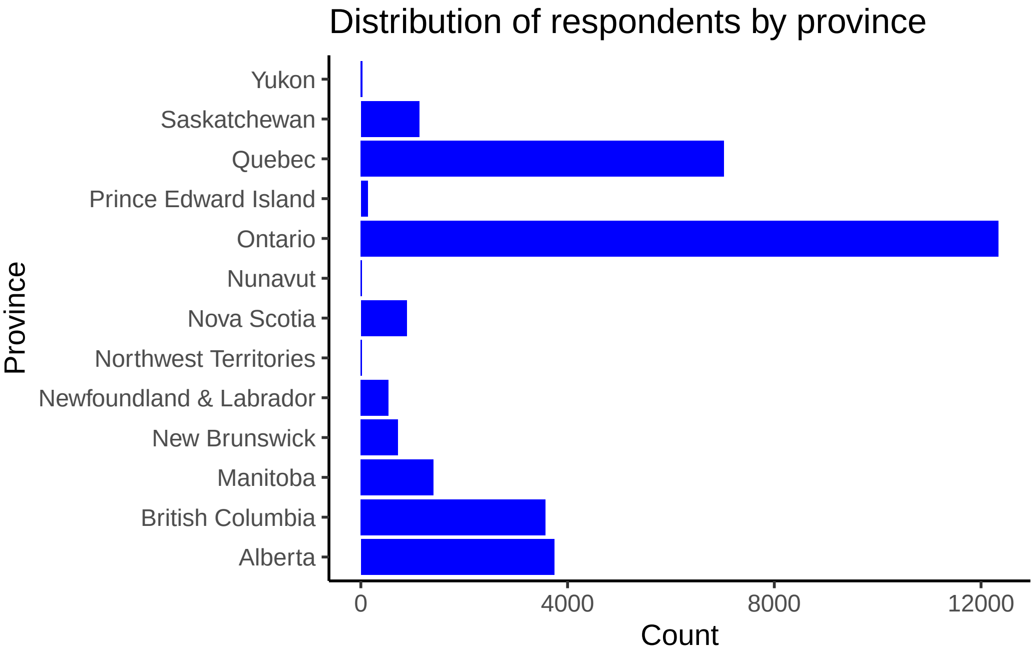

Figure 1: Number of participants in the survey from each Canadian province

3 Methods

The logistic regression formula for the model is:

$$\log\left(\frac{P(\text{vote for Liberals})}{1-P(\text{vote for Liberals})}\right) = \beta_0 + \beta_1 \cdot \text{province}_{\text{British Columbia}}$$
$$+ \beta_2 \cdot \text{province}_{\text{Manitoba}} + \beta_3 \cdot \text{province}_{\text{New Brunswick}}$$
$$+ \cdots + \beta_{13} \cdot \text{province}_{\text{Yukon}} + \beta_{14} \cdot \text{age}$$

In this model, the dependent variable represents the log odds of voting for the Liberal Party. The independent variables include a set of dummy variables for Canadian provinces and territories, as well as respondents' age. The intercept (β_0) represents the log odds of voting Liberal in the reference province when the respondent is 18 years old, the legal voting age in Canada. Each coefficient for the province dummy variables (β_1 to β_{13}) represents the difference in log odds of voting Liberal in that province compared to the reference province. A positive coefficient indicates a higher likelihood of supporting the Liberals relative to the reference province, while a negative coefficient indicates a lower likelihood. The coefficient for age (β_{14}) reflects the change in log odds of voting Liberal for each additional year of age. A positive coefficient would suggest older individuals are more likely to vote Liberal, whereas a negative coefficient would suggest younger individuals are more likely to support the party.

The formula for the expected proportion of Liberal votes is

$$\hat{p} = \frac{\text{Total Liberal Votes}}{n}$$

where the numerator is the total number of respondents who indicate they will likely vote Liberal, and the denominator is the total sample size. The 95% confidence interval for this proportion is calculated using the following formula:

$$\hat{p} \pm z_{0.975} \sqrt{\sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}(1-\hat{p})}{n_h}}$$

In this formula, the term $1 - \frac{n_h}{N_h}$ is the finite population correction, which adjusts for cases where a large portion of a stratum's population is sampled. Here, n_h is the sample size in stratum h (e.g., a particular province), and N_h is the population size for that stratum. Population data is obtained from Statistics Canada.

A wider confidence interval reflects greater uncertainty about the true proportion of Liberal support. This increased uncertainty may result from smaller sample sizes, greater variability within strata, or larger finite population corrections. A wider interval ultimately makes it more difficult to predict the actual vote share with confidence.

4 Results

Table 1 shows the 95% confidence interval for the proportion of expected votes for the Liberal Party. An estimated 28.4% of the survey population intends to vote for the Liberals. The 95% confidence interval ranges from 27.9% to 28.8%. This means we are 95% confident that the true proportion of Liberal support falls within this interval. The narrow range of the confidence interval suggests a relatively precise estimate, with some uncertainty.

Table 2 presents the results of the logistic regression analysis. The model designates Alberta as the reference category, meaning all β coefficients represent the log-likelihood of voting for the Liberal Party relative to Alberta. The intercept (β_0) is significant and negative, indicating a low baseline likelihood of voting Liberal in Alberta.

The regression results do not show uniform support for the Liberals across Canada but instead suggest regional differences. Compared to Alberta, most provinces and territories have positive and significant β coefficients, indicating a higher log-likelihood of voting for the Liberals. Specifically, Ontario, Quebec, British Columbia, Newfoundland & Labrador, Nova Scotia, New Brunswick, Manitoba, Prince Edward Island, Northwest Territories, and Nunavut all show significantly higher Liberal support. Saskatchewan is the only province with a lower likelihood of voting Liberal than Alberta, suggesting strong opposition to the party. Yukon is the only province where the difference from Alberta is not statistically significant. These results highlight a clear regional divide, with the Liberals performing significantly better in central and eastern Canada while struggling in the Prairie provinces.

The model also suggests a significant negative relationship between age and Liberal support. The negative and significant β coefficient indicates that as age increases, the log-likelihood of voting Liberal decreases. However, the estimated magnitude of this effect is small ($\beta =$ -0.0037), meaning the impact of age on voting Liberal is negligible. For each additional year of age, the log-odds of voting Liberal decreases by 0.0037, indicating a gradual rather than drastic decline.

Figure 2 displays the odds ratios from the logistic regression model, providing a clearer visualization of the relative likelihood of voting Liberal. The graph highlights several key points. Newfoundland & Labrador has the highest odds ratio, with voters there being nearly 3.6 times more likely to vote Liberal than those in Alberta. Ontario, Canada's most populous province, is almost 2.8 times more likely to support the Liberals. By contrast, Saskatchewan's odds ratio falls below 1, confirming its strong opposition to the party. These figures reinforce the regional and demographic differences in Liberal support across Canada.

Table 1: Table showing proportion and 95% confidence interval for Liberal vote

	Proportion of vote for Liberals	95% Confidence Interval
Liberals	0.284	(0.279, 0.288)

5 Discussion

In this report, we estimate how the proportion of Liberal votes varies across regions and age groups in Canada. The results suggests a regional divide in Liberal support, with the Prairie provinces showing significantly lower support than Eastern Canada. We also find a significant

Table 2: Table showing the proportion and 95% confidence interval for Liberal votes across Canada

```
|term                      | estimate| std.error| statistic| p.value| conf.low| conf.high|
|:-------------------------|--------:|---------:|---------:|-------:|--------:|---------:|
|(Intercept)               |  -1.5405|    0.0597|  -25.8003|  0.0000|  -1.6576|   -1.4235|
|British Columbia          |   0.6717|    0.0595|   11.2821|  0.0000|   0.5550|    0.7883|
|Manitoba                  |   0.6047|    0.0768|    7.8720|  0.0000|   0.4541|    0.7552|
|New Brunswick             |   0.9291|    0.0927|   10.0228|  0.0000|   0.7474|    1.1108|
|Newfoundland & Labrador   |   1.2888|    0.0993|   12.9839|  0.0000|   1.0943|    1.4834|
|Northwest Territories     |   1.2446|    0.4490|    2.7717|  0.0056|   0.3645|    2.1247|
|Nova Scotia               |   1.1619|    0.0833|   13.9552|  0.0000|   0.9987|    1.3251|
|Nunavut                   |   1.2588|    0.4505|    2.7940|  0.0052|   0.3757|    2.1418|
|Ontario                   |   1.0324|    0.0496|   20.8297|  0.0000|   0.9353|    1.1296|
|Prince Edward Island      |   0.9520|    0.1919|    4.9606|  0.0000|   0.5758|    1.3281|
|Quebec                    |   0.8320|    0.0527|   15.7780|  0.0000|   0.7287|    0.9354|
|Saskatchewan              |  -0.2651|    0.1023|   -2.5924|  0.0095|  -0.4655|   -0.0647|
|Yukon                     |   0.4542|    0.4311|    1.0535|  0.2921|  -0.3908|    1.2992|
|cps19_yob                 |  -0.0037|    0.0008|   -4.7291|  0.0000|  -0.0052|   -0.0021|
```

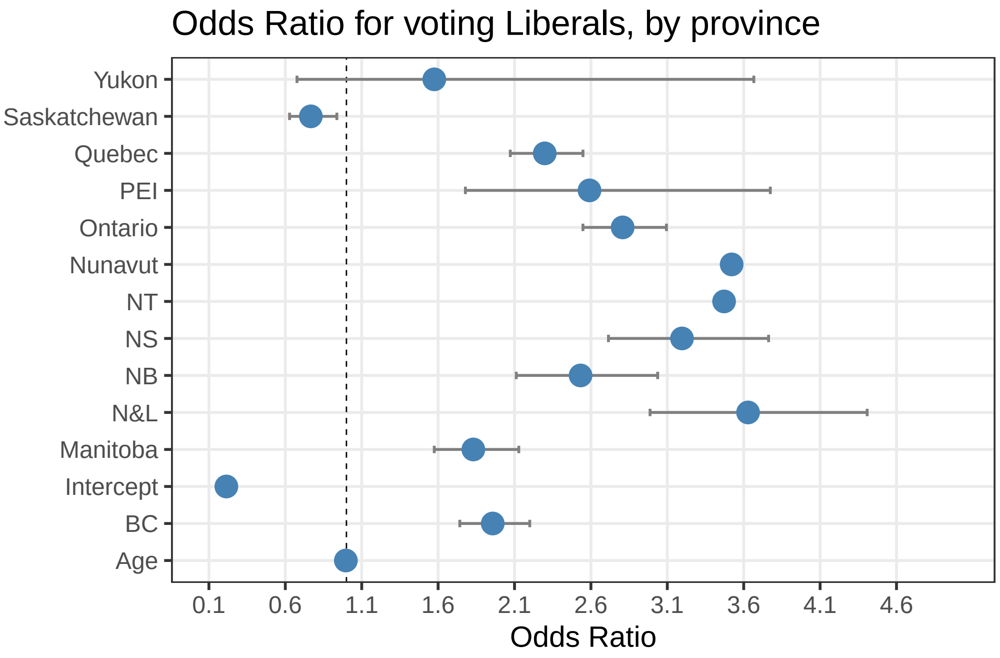

Figure 2

but negligible effect of age, with each additional year decreasing the log-likelihood of voting for the Liberals. However, the current model combines age with regional effects, meaning the individual effect of age on voting behavior is not captured. While the log-likelihood statistic shows the likelihood of voting Liberal across regions, the reasons behind these differences—whether due to age or other demographic factors—are not fully understood. Further analysis is needed to explore these effects.

Another potential issue is omitted variable bias. We did not include predictors such as education and income, which previous studies have shown to significantly impact voting behavior. For example, research suggests that education tends to shift individuals toward the right on the political spectrum (Meyer 2017), while income differences may influence party preferences (Polacko, Kiss, and Graefe 2022). The exclusion of these factors could lead to biased estimates of the relationship between region and voting preference, as well as affect the magnitude and direction of the results.

Another limitation comes from missing data in the survey responses. For instance, 16.5% of responses are missing for the preferred political party variable. We assume the data are missing at random (MAR), because there is a possibility that missing data could come from certain behavioral patterns linked to participant characteristics. We predict that individuals with lower education levels or younger participants may be less likely to engage in political surveys, leading to missing values in the voting behavior responses. However, we do not conduct any analysis to establish that this is the case in actuality. The current study excludes rows with missing values, which was not problematic as the dataset still contained a sufficient number of observations for robust analysis.

A further issue arises with the population data obtained from StatCan. Their estimate includes ineligible voters, such as temporary residents, refugee claimants, and landed immigrants. This is problematic because approximately 6.2% of the Canadian population consists of temporary residents (Canada 2025b), which could result in statistics that do not accurately reflect the size of the eligible voting Canadian population. This could skew the estimates found in Table 1, leading to erroneous conclusions.

In future research, we aim to address these issues. First, we will include additional predictor variables in the analysis model. Previous literature highlights that education, income, and other factors can predict voting patterns. Including these variables will reduce bias and improve the robustness of the analysis. Second, we intend to investigate the origin of the missing data values and clean up population statistics to reduce biases in our report.

6 Generative AI Statement

Generative AI was used in this analysis. Specifically, ChatGPT served several purposes. First, I used it to edit and correct mechanical mistakes in my writing. It identified errors in punctuation, spelling, and grammar. Additionally, I used ChatGPT (OpenAI 2023) as a proofreader, asking it to identify sentences or paragraphs that might be unclear or difficult for readers to follow, including run-on sentences. Lastly, I relied on ChatGPT for LaTeX coding. In Section 3, I provide details on the models used in this report. The code for these models was generated by ChatGPT. As the sole author of this report, I proofread all text after running it through ChatGPT to ensure that it accurately reflects my ideas and does not contain any errors or hallucinations from the generative AI tool.

For workflow, I used two external sources for this report. First, I obtained population data from Statistics Canada (StatCan). I used population estimates from the fourth quarter of 2024. StatCan relies on census data, surveys, and administrative records such as tax and immigration data to produce these estimates (Canada 2025a). Lastly, I obtained the code for Figure 2 from Stack Overflow, an inquiry-based forum for coding questions. I altered their code to fit my data in order to generate Figure 2. The website link can be found here.

7 Ethics Statement

The analysis uses data from a secondary source that is publicly available. Therefore, we did not require Research Ethics Board (REB) approval for our analysis. However, this is human research, and there are a few ethical issues we considered in this study. First, this is observational research on humans, and there is a possibility that the information could be used to identify individuals, even when not associated with their names. However, this is not a concern for this study, as the dataset is large with over 37,000 total observations. Additionally, no personal information regarding participants' characteristics, other than age and gender, is reported. Lastly, individuals in the study had no reasonable expectation of privacy, as they consented to the use of this data for research purposes (Stephenson et al. 2020). Therefore, this analysis adheres to all the ethical guidelines required for using the data.

To ensure reproducibility, we have included all code and data on our GitHub page, with the link provided on Page 1. We have also included .qmd file containing all the code used to generate this paper. The GitHub page includes a data dictionary and information on the dataset to facilitate reproducibility. Additionally, all code chunks contain comments explaining their functions, and we have also documented all the packages used in this analysis, which can be found on Section 2.

8 Bibliography

Canada, Statistics. 2025a. "Table 17-10-0009-01 Population Estimates, Quarterly." 2025. https://doi.org/10.25318/1710000901-eng.

———. 2025b. "Table 17-10-0121-01 Estimates of the Number of Non-Permanent Residents by Type, Quarterly." 2025. https://doi.org/10.25318/1710012101-eng.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using r: A Guide to Analysis Using r*. John Wiley; Sons.

Meyer, Alexander G. 2017. "The Impact of Education on Political Ideology: Evidence from European Compulsory Education Reforms." *Economics of Education Review* 56: 9–23. https://doi.org/10.1016/j.econedurev.2016.11.003.

OpenAI. 2023. "ChatGPT." https://openai.com/chatgpt.

Polacko, Matthew, Simon Kiss, and Peter Graefe. 2022. "The Changing Nature of Class Voting in Canada, 1965–2019." *Canadian Journal of Political Science* 55 (3): 663–86. https://doi.org/10.1017/S0008423922000439.

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

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Stephenson, Laura B, Allison Harell, Daniel Rubenson, and Peter John Loewen. 2020. "2019 Canadian Election Study - Online Survey." Harvard Dataverse. https://doi.org/10.7910/DVN/DUS88V.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.