Predicting Probability of Participants' Vote of Conservative and Liberty in 2025 Canadian election*

Jingyi Shen

April 13, 2024

ojjojojojljk,mkmk.

Table of contents

*Code and data supporting this analysis are available at: https://github.com/CSCmaster/Final-Project

1 Introduction

1.1 Background

The Canadian electoral process, epitomizing democratic principles, empowers citizens to elect parliamentarians who shape Parliament's composition and, consequently, the government formation. In the recent 2021 federal election, Justin Trudeau's Liberal Party secured a majority in the 44th Parliament. Voters' primary concerns include the cost of living, healthcare, climate change, post-pandemic recovery, and immigration, with dissatisfaction towards Trudeau's government prevalent. Trudeau's popularity lags behind UK's Poilievre. The government's ambitious immigration plan aims to admit 500,000 migrants annually by 2025, albeit facing a backlog. The Conservative Party, with a focus on economic prosperity, emphasizes attracting highly skilled immigrants while addressing concerns like living costs and housing issues(Canada, n.d.).

1.2 Hypothesis

Based on our analysis of historical trends, governmental policies, and electoral procedures, our hypothesis aligns with the prevailing expectation that the Conservatives will emerge as the likely winners in the 2025 election. **We anticipate that the Conservatives will likely secure the highest number of votes, with the Liberals expected to closely follow in second place.** Our research indicates that the vote-counting process in Canada significantly influences the outcome, as voters do not directly elect a specific prime minister candidate. Initially, all votes are tallied based on first preferences. If a candidate obtains 50% of the vote, they win. If not, the candidate with the fewest votes is eliminated(Canada, n.d.). Then, votes

from supporters of the eliminated candidate are redistributed to their second-choice candidate. This iterative redistribution continues until one candidate achieves 50% of the vote. Hence, our objective is to forecast the election probabilities for the Conservative and Liberal parties and compare which is higher(n.d.a).

1.3 Terminology

In this paper, I used binary logistic regression models to forecast the likelihood of individuals affiliating with the Conservative or Liberal parties. Binary logistic regression is a statistical technique tailored for predicting binary outcomes, where the variable under consideration has two possible results. In our study, these outcomes represent the probability of individuals aligning with each of the two parties. Each party's likelihood will be predicted through separate logistic regression models, with the response variables indicating whether individuals vote for the corresponding party or not. The predictors used in these models include age, province, education, religion, and income before tax. The variable "sex" will be initially included but later removed after AIC model testing, as further elaborated in the model selection section. This regression approach is particularly advantageous when there's a nonlinear relationship between independent and dependent variables, especially in scenarios with categorical dependent variables like ours. Additionally, we will incorporate post-stratification in our analysis to enhance accuracy. Post-stratification helps refine precision estimates by ensuring that the sample is more representative of the entire population.

2 Data

2.1 Raw Data

In both datasets, we've focused on standardizing variables such as age, sex, province, education, religion, and pre-tax income. Conducting a census typically involves higher expenses and more time since it requires reaching out to every individual in the population. Surveys, however, offer a more cost-effective and quicker alternative, especially for large populations, by collecting data from a select group that represents the larger whole.

Census Data: Census data refer to information systematically gathered from the entire population of a defined area at a specific point in time. The census data we have comes from the General Social Survey (GSS), conducted on August 12, 2022. We've made necessary adjustments to ensure it matches our survey data. Data collection aims to provide a detailed snapshot of demographic, economic, and social characteristics.The GSS is a nationwide survey program designed to collect data on societal trends, behaviors, and attributes across the entire population, offering insights into the changing dynamics of communities.

Survey Data: Survey data, are collected from a sample of the population rather than every individual. This method is often used when it's impractical or unnecessary to include

everyone. Survey data aim to infer the characteristics of the larger population based on the responses of the sample. The survey data, obtained from the Canada Election Study (CES) in 2023, includes responses from over 37,000 participants. The CES focuses specifically on voter behavior, attitudes, and the electoral process within the Canadian context. Unlike the GSS, which aims for a comprehensive overview, the CES targets specific topics of interest to derive insights from a segment of the population.

2.2 Data analysis Tools

In this research,the arrow(Richardson et al. 2024)package significantly impacts the way large datasets are handled and processed. It provides high-performance reading and writing of data in the Arrow file format, including support for Parquet files, which are highly efficient for storing and querying large datasets. Tidyr(Wickham, Vaughan, and Girlich 2024) helps in tidying data, meaning it makes it easier to structure datasets so that they are straightforward to work with. It provides functions to transform data into a tidy format, where each variable forms a column, each observation forms a row, and each type of observational unit forms a table. The mass package(Venables and Ripley 2002), It provides a wide range of statistical techniques including linear and nonlinear modeling, statistical tests, time series analysis, classification, and clustering. The package is known for its functions to fit generalized linear models, among many other tools. With summarytools(Comtois 2022), users can easily generate frequency tables, descriptive statistics summaries, cross-tabulations, and more. It makes exploratory data analysis more efficient and is particularly useful for preliminary data analysis, ensuring that researchers and analysts can understand their data before moving on to more complex analyses. Data analysis was conducted using the R programming language (R Core Team 2022), renowned for its open-source nature and robust statistical analysis capabilities. Visualization complexities were addressed with the ggplot2 package (Wickham 2016), which supports the creation of intricate graphics. For data manipulation, dplyr(Wickham et al. 2022) was employed, providing a streamlined grammar that simplifies dataset filtering, summarization, and reorganization. Fast and efficient data importing was achieved through the use of the readr package(Wickham, Hester, and Bryan 2022). The process of generating this report was seamlessly managed by knitr(Xie 2014), facilitating the embedding of R code directly within the text.

2.3 Variable Description

2.3.1 Survey Data Variable

Age: The cps21_age field from the survey data denotes participants' ages at the time they completed the General Social Survey. This variable requires no alterations but should be renamed to "age" to ensure consistency with the census data's corresponding variable.

sex: In the survey dataset, cps21_genderid identifies 9474 respondents as male, 11370 as female, 90 as non-binary, and 34 as other. Given the relatively small number of non-binary and other responses, distribute these respondents between male and female categories based on existing proportions, with females at 54.22% and males at 45.18%.

province: The survey data's cps21_province includes three additional provinces not present in the census data, creating discrepancies. To align the datasets, remove Northwest Territories, Nunavut, and Yukon from the survey data. Moreover, convert the numeric province labels in the survey data into their corresponding categorical names for analytical compatibility.

Education: The survey data's cps21_education, featuring twelve education levels, should be consolidated into three categories for analysis. Group No schooling through "Don't know/ Prefer not to answer" as "Limited Education"; "Some secondary/ high school" through "Some university as Some Education"; and degrees from "Bachelor's degree" onwards as "Highly Educated".

Religion: The cps21_religion variable indicates a respondent's religious affiliation. A response of '1' denotes atheism (categorized as 'NO'), whereas any other response signifies a religious affiliation (categorized as 'YES').

Income_before_tax: Convert the numerical cps21_income_numbe, indicating total household income before taxes, into categories for logistic regression analysis. Define "Lower Middle Class to Poor" as incomes $<$ \$50,000, "Middle Class" as incomes between \$50,000 and \$124,999 and "Upper Middle Class to Wealthy" as incomes \$125,000.

vote_liberal: The cps21_votechoice variable records intended voting behavior. For logistic regression, recode responses to a binary format, where selecting '1' indicates an intention to vote for the Liberal Party ('1'), and any other selection is recoded as '0'.

vote_conservative: for the Conservative Party, recode cps21_votechoice to '1' for respondents who choose '2', indicating a preference for this party, and '0' for all other selections.

2.3.2 Census Data Variable

Age Adjustment: In the census data, age is presented as a decimal number. For consistency with the survey data, round this value to the nearest whole number.

Gender Representation: The gender classification in the census data, listed as either 'Male' or 'Female', aligns perfectly with the survey data, requiring no modifications.

Provincial Data: The listing of provinces in the census data, including "Newfoundland" and "Labrador", "Nova Scotia", "Quebec", "Saskatchewan", "Ontario", "Alberta", "British Columbia", "Prince Edward Island", "New Brunswick", and "Manitoba", matches that of the survey data, making no further adjustments necessary.

Educational Levels: The census data sorts educational attainment into eight categories, with one category for missing information. Align these with the survey data's education classifications by grouping "Less than high school diploma or its equivalent" and missing data as "Limited Education"; "High school diploma or a high school equivalency certificate", "Trade certificate or diploma", "College, CEGEP, or other non-university certificate or diploma", "University certificate or diploma below the bachelor's level" as "Some Education"; and both "Bachelor's degree", and any university certification above a bachelor's degree as "Highly Educated".

Religious Affiliation: The "religion_has_affiliation" variable in the census data identifies if an individual has a religious affiliation, is uncertain, claims no religious affiliation, or if the response is missing. For alignment with the survey data, consolidate into two groups: categorize as `YES` for those with a religious affiliation, uncertain, or missing responses (assuming missing responses indicate reluctance to disclose religious affiliation), and `NO` for those without any religious affiliation.

Income Categories: The Income_respondent field in the census dataset uses 7 income brackets to classify respondents' earnings. To harmonize with the survey data's "income_before_tax" categories, reclassify these into three broader groups. Designate incomes of "$125,000 and more" to the "Upper Middle Class to Wealthy" group. Categorize incomes of "$50,000 to $74,999", "$75,000 to $99,999", and "$100,000 to $124,999" as "Middle Class". Assign "Less than $25,000" and "$25,000 to $49,999" incomes to the "Lower Middle Class to Poor" group.

2.4 Sample of cleaned Data

2.4.1 Data Summary Measures

Table 1: Summary of Numerical Variables Across Census and Survey Data

Variable	Minimum	Q1	Median	Average	Q3	Maximum
Age in Census Data	15	37	54	52.180	67	80
Age in Survey Data	18	36	53	51.300	66	97
Liberal Votes in Survey	0	0	0	0.267	1	1
Conservative Votes in Survey	0	0	0	0.249	0	1

Table 1 summaries the census and survey data reveals age distributions with the census data capturing a narrower age range (15 to 80 years, with a median of 54) compared to the broader age span in the survey data (18 to 97 years, with a median of 53). The average ages for the census and survey populations are 52.18 and 51.3 years, respectively. Voting data from the survey indicate that a quarter of the respondents are inclined towards the Liberal Party, as shown by an average of 0.267, while a slightly lower average of 0.249 suggests Conservative

support; both parties have 50% of the population not voting for them, with a median of 0. The upper bounds for Liberal and Conservative votes are 1, meaning some respondents indicated a vote for these parties, while the third quartile for Conservative votes indicates less overall support compared to the Liberal Party.

2.4.2 Data Visualization

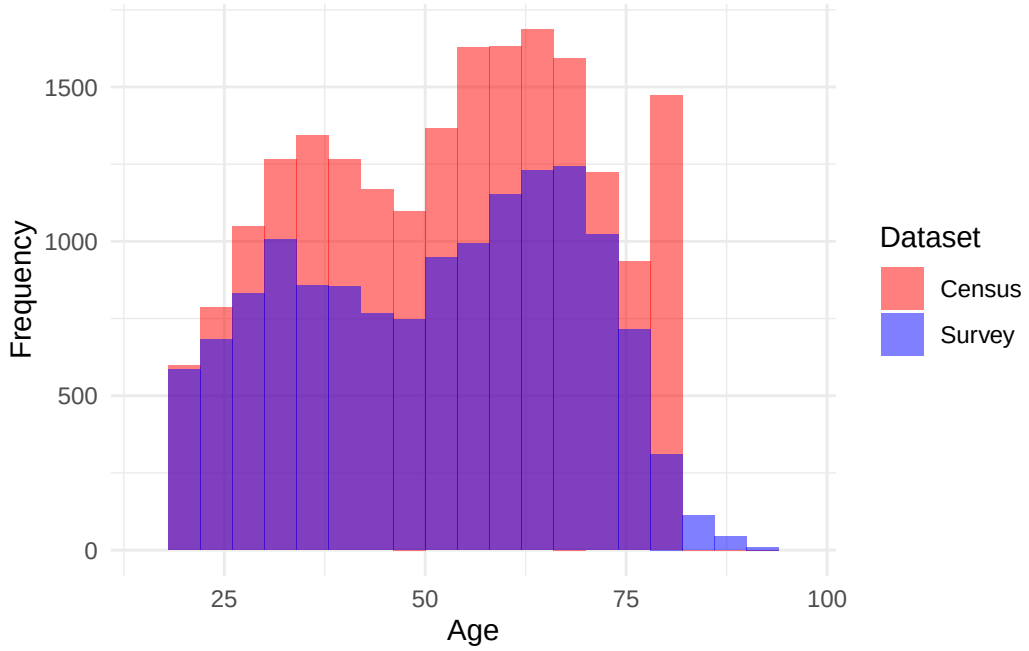

Figure 1

Figure 1 displays the age distribution from both census (in red) and survey (in blue) datasets. The census distribution exhibits a unimodal peak around the 50-60 year age range and a sharp decline beyond 80 years, suggesting a concentration of middle to late-middle-aged respondents. Conversely, the survey dataset portrays a wider distribution, extending towards older ages with a noticeable skew to the right, indicating the presence of older individuals up to 97 years old. The regions of overlap appear purple, indicating age ranges common to both datasets, with the census data generally showing higher frequencies in most age categories. This comparison highlights the distinct age profiles captured by the two sources, with a significant representation of middle-aged individuals in both, with the survey dataset capturing a broader demographic.

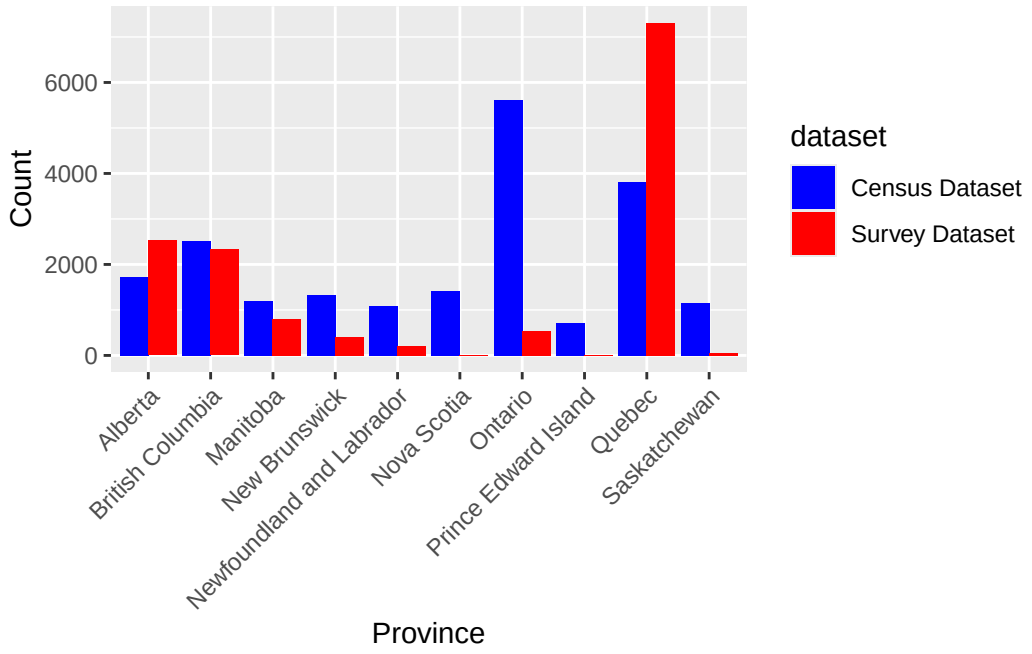

Figure 2: Comparison of Province Variable Between Survey and Census Data

Figure 2 contrasts the provincial distribution of respondents from both survey and census datasets, with provinces along the x-axis and respondent counts on the y-axis. The survey data, shown in blue, reaches its highest point in Ontario, indicating a significantly larger number of respondents compared to other provinces in the survey. This peak towers above all other survey data points and also surpasses the census data count for Ontario, which suggests a disproportionate representation or potentially an over-sampling in the survey for Ontario. In contrast, the lowest point for the survey dataset appears in Prince Edward Island, with very few respondents, paralleled by a similarly low count for this province in the census dataset, shown in red. Notably, the census data peaks in Quebec, where the red bar indicates the highest frequency among all provinces for the census, while Alberta's census count is at its lowest, barely rising above the baseline. This visual comparison reveals not only the variations in population coverage between the two types of data but also significant regional differences within the datasets, which could reflect varied sampling methods or demographic distributions.

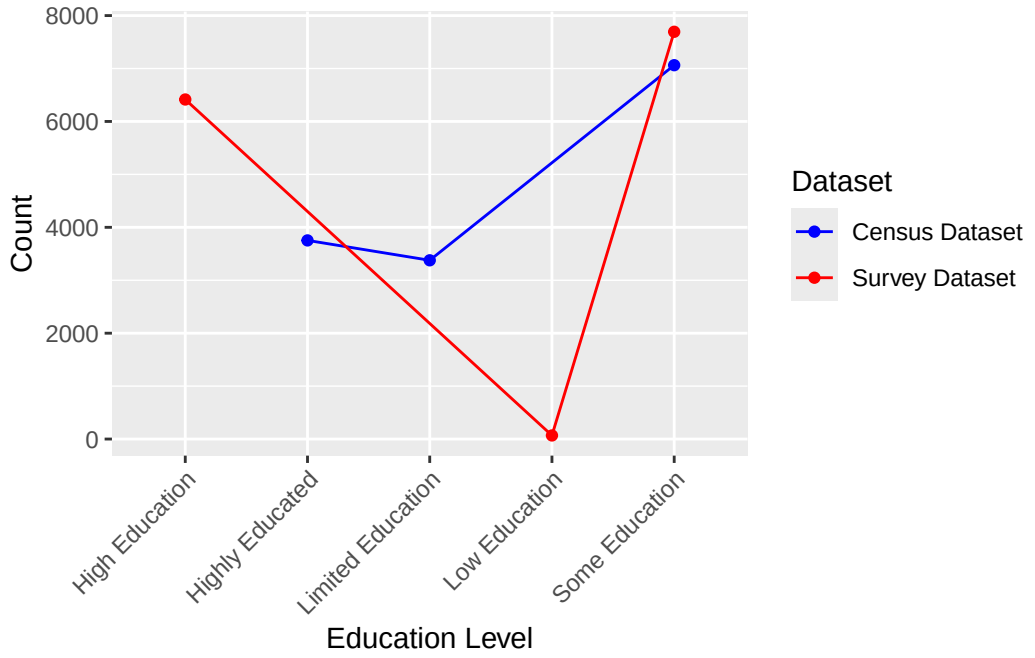

Figure 3: Line Chart Comparison of Education Levels Between Survey and Census Data

Figure 3 presented compares the distribution of education levels across two different datasets, one from a survey and another from a census, differentiated by colors—red for the survey and blue for the census. Education levels are divided into three categories along the x-axis: "Highly Educated", "Limited Education", and "Some Education" in census dataset, with the y-axis displaying the count of individuals within each category. While the education level is divided into "Low Education", "Some Education" and "High Education" in the survey dataset. A noticeable trend is the inverse relationship between the two datasets from "Highly Educated" to "Limited Education": the count for the Census dataset declines while that for the Survey dataset rises, suggesting a disparity in the representation of education levels between the two data sources.

This discrepancy highlights potential differences in sampling methods, response rates, or the representativeness of each dataset relative to the underlying population. The most pronounced difference is at the "Limited Education" level, where the Survey dataset shows significantly lower counts compared to the Census dataset, indicating a possible overrepresentation of individuals with limited education in the censaus or an underrepresentation in the survey. The data suggest that demographic, socioeconomic, or regional factors may influence the education level distribution in each dataset, and further statistical analysis could be warranted to determine the significance and cause of these differences. The differences observed in the chart might also reflect methodological variations in data collection or inherent biases within the sampling frames used.

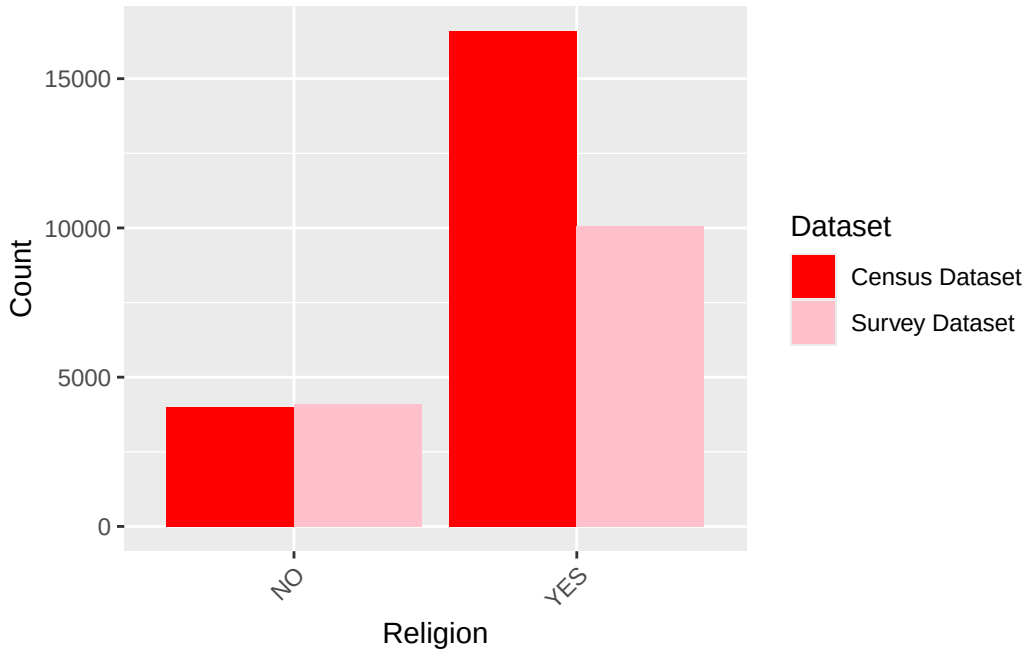

Figure 4: Bar Chart Comparison of Religion Across Survey and Census Data

Figure 4 delineates the reported religious affiliations in both the census and survey datasets, differentiated by color intensity, with the census depicted in a darker red and the survey in a lighter pink. The x-axis is categorized into 'No' for non-affiliated individuals and 'Yes' for those with religious affiliation. A striking contrast is observed in the 'Yes' category, where the census data's count towers over that of the survey, indicating a greater number of religiously affiliated respondents. Conversely, for those reporting no religious affiliation, the survey's lighter pink bar is comparable in height to the census's, suggesting a more even distribution between the two datasets. This discrepancy underscores a pronounced difference in the representation of religious affiliation between the two data sources, potentially alluding to divergent demographic characteristics or response patterns within the datasets.

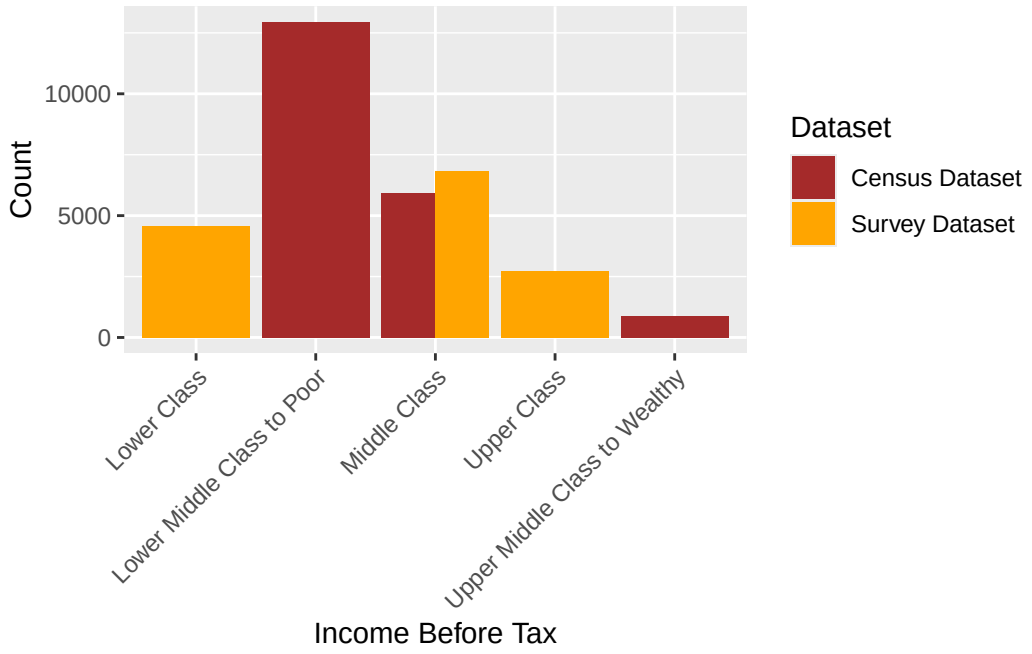

Figure 5: Bar Chart Comparison of Income Before Tax Between Survey and Census Data

Figure 5 compares the distribution of socioeconomic groups before tax as represented in the census and survey datasets, with income hierarchy arrayed along the x-axis and the number of individuals in each bracket indicated on the y-axis. Both datasets show a similar count in the 'Upper Middle Class to Wealthy', 'Upper Class', 'Lower Middle Class to Poor' and 'Lower Class' category. The census dataset, indicated by the significantly taller deep red bar as opposed to the shorter orange bar of the survey dataset. The 'Middle Class' brackethas a higher representation in the survey data but by a less pronounced margin. The colors—deep red for the census and orange for the survey—aid in the visual differentiation between the two datasets, enabling a direct comparison of how each income level is represented within them, and suggest a higher census capture of lower-income brackets compared to the survey data.

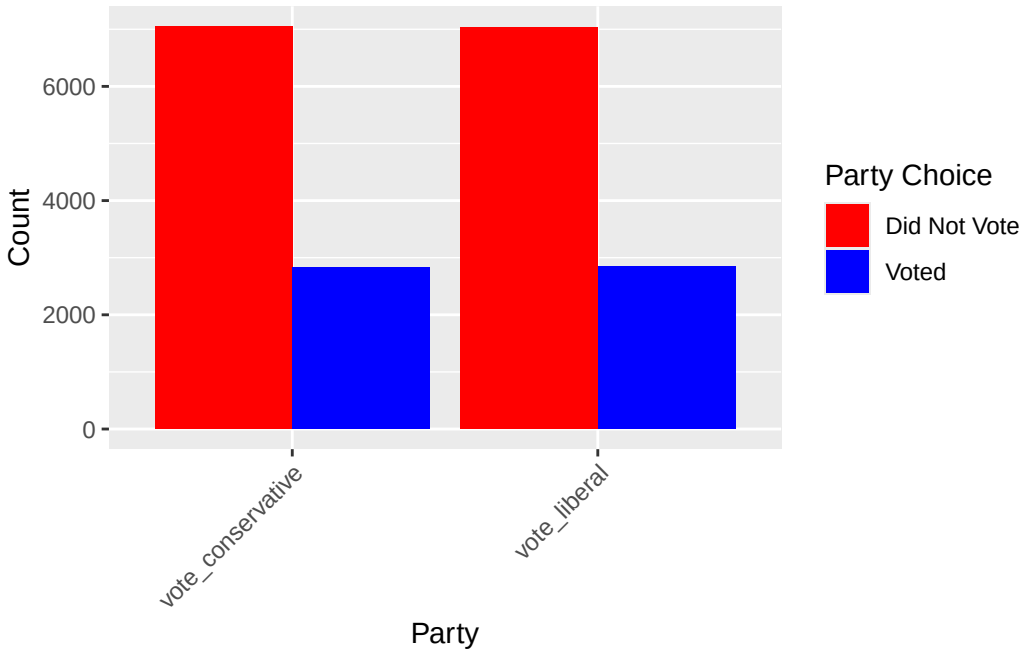

Figure 6: Bar Chart Comparison of Voting Preferences

Figure 6 contrasts the voting preferences for the Conservative and Liberal parties, delineating counts of participants who voted (blue bars) against those who did not (red bars). Notably, there is a substantial number of non-voters in the Conservative category, exceeding their Liberal counterparts, suggesting a larger disenchantment or mobilization challenge within the Conservative base. Despite this, the voter turnout (blue bars) is fairly balanced between the parties, indicating a competitive engagement among those who participated in the vote. The overall higher presence of red bars signifies a general trend of voter disengagement, which may warrant further investigation through regression analysis and post-stratification to grasp the full scope of electoral dynamics and the broader implications for the Canadian democratic process.

2.5 Measurement

In this paper, there are two data set, the survey dataset and census dataset. The survey dataset is utilized from Canadian Election Study(CES) while the census dataset is utilized from General Social Survey(GSS).

Canadian Election Study gathered a sample of 20,968 individuals from across Canada using the Leger Opinion panel, ensuring representation from various regions while maintaining a balance in gender and age demographics within each area. The sample aimed for an equal distribution of 50% males and 50% females, with age groups targeted as follows: 28% between 18-34 years, 33% between 35-54 years, and 39% 55 years and older. The geographical regions included Atlantic Canada, Quebec, Ontario, Western Canada, and the Territories, with quotas set to reflect their population proportions (7% for Atlantic, 23% for Quebec, 38% for Ontario, and 32% for the West) and an open acceptance for any participants from the Territories. In Quebec, the language preference aimed for 80% French-speaking and 20% English-speaking participants, with a goal of 10% French-speaking participants in both the Atlantic region and across Canada. Eligibility for participation required being at least 18 years old and a Canadian citizen or permanent resident(n.d.b).

The General Social Survey focuses on non-institutionalized individuals aged 15 and above residing in Canada's ten provinces. Within each selected household, a random individual is chosen by an application to answer the survey following the initial household roster completion. General Social Survey's questionnaire was developed through rigorous research and broad consultations with data users. The Statistics Canada Questionnaire Design Resource Center (QDRC) conducted qualitative testing in four cities to identify effective questions and those requiring adjustments. Based on the testing, QDRC prepared a comprehensive report with suggestions, which were integrated into the final survey design to enhance clarity and effectiveness.The survey employs a cross-sectional sample survey methodology, utilizing a combined frame of landline and cellular phone numbers from Census data, administrative records, and Statistics Canada's dwelling database. This approach improves coverage of households linked to phone numbers. The sampling strategy is stratified, with stratification occurring at the level of provinces and census metropolitan areas (CMA), and it relies on probability sampling techniques. Information is collected from a single individual aged 15 or older per household, without the use of proxy responses.(n.d.c)

I made several adjustment to the raw dataset to parepare for the analyzed dataset. In the process of cleaning the survey data, several transformative steps were taken to ensure its suitability for analysis. Initially, the age variable was directly adopted as is from `cps21_age`. Gender distribution was then randomized based on predefined proportions for male and female respondents, aiming for a balanced representation. Province information, originally in code form, was converted to textual names for clarity, with any undefined codes being assigned as `NA` to exclude incomplete records. Educational background was categorized into three broad levels: "Limited Education", "Some Education", and "Highly Educated", based on specific ranges of `cps21_education` codes. Religious affiliation was simplified to a binary "YES"/"NO" based

on the `cps21_religion` variable. Income before tax was segmented into three distinct brackets to reflect socioeconomic status. Political preferences were distilled into binary indicators for liberal and conservative vote choices. The dataset was further refined by removing records with missing province information, focusing the analysis on a selected set of variables: age, sex, province, education, religion, and income before tax. This streamlined dataset was then saved in both CSV and Parquet formats, marking the completion of the data cleaning phase and ensuring a well-structured and analysis-ready dataset.

In cleaning the census data, approach was employed to ensure the data's accuracy and relevance for analysis. The age variable was rounded to the nearest whole number to standardize age data across the dataset. Education levels were categorized into three distinct groups: "Limited Education" for those with less than a high school diploma or equivalent, including missing data; "Some Education" for individuals with high school diplomas, trade certificates, or college diplomas below the bachelor's level; and "Highly Educated" for those holding a bachelor's degree or higher. Religious affiliation was simplified to a binary "YES" for those with any religious affiliation or uncertain/missing responses, and "NO" for those explicitly without religious affiliation. Income levels were stratified into three categories: "Lower Middle Class to Poor" for incomes below $50,000, "Middle Class" for incomes between $50,000 and $124,999, and "Upper Middle Class to Wealthy" for incomes of $125,000 and above. This process streamlined the dataset by focusing on essential variables for sociodemographic analysis—age, sex, province, education, religion, and income before tax. The cleaned data was then saved in both CSV and Parquet formats to facilitate accessibility and further analysis.

3 Model

We utilized the rstanarm package for constructing the generalized linear model(Goodrich et al. 2020), the rstanarm package serves as a user-friendly interface in R for fitting Bayesian regression models using the Stan probabilistic programming language. It enables users to effortlessly construct various types of regression models, including linear, logistic, and Poisson regression, while leveraging the power of Bayesian methods for inference. By seamlessly integrating with Stan, rstanarm provides efficient algorithms for model fitting and accurate estimates of uncertainty. Additionally, the package offers diagnostic tools for assessing model convergence and sensitivity to prior specifications, making it an invaluable resource for Bayesian data analysis in R.

3.1 Method

In forecasting the outcomes of the forthcoming Canadian federal election, statistical techniques like regression analysis, selection of variables, and post-stratification play a pivotal role. Our dataset has some missing variables, but these were deemed negligible in terms of their impact on our predictive model. Initially, the dataset contained 20,921 records, which would have been

reduced to 14,544 if entries with missing values were excluded. To maintain the robustness and fullness of our dataset, we opted to retain these incomplete records.

The first step in our modeling approach involved applying logistic regression to estimate the likelihood of a vote being cast for the Liberal or Conservative, based on a set of explanatory variables. Key metrics such as model coefficients and p-values were scrutinized to understand the influence of each predictor. Upon reviewing the p-values, the AIC stepwise selection method was chosen to refine the model further, aiming to lower the AIC score for a more parsimonious fit.

The last stage incorporated post-stratification to gauge the distribution of voter preferences across the identified political factions. This involved segmenting the populace according to our model's predictors and calculating the voting proportions for each party within these segments.

3.2 Model Set-up

We define our Bayesian logistic regression model for the probability of an individual voting for the Liberal/Conservative party as follows:

- y_i: Binary outcome of voting for the Liberal/Conservative party (1) or not (0)
- π_i: Probability of voting for the Liberal/Conservative party

$$y_i|\pi_i \sim \text{Bernoulli}(\pi_i) \tag{1}$$

$$\text{logit}(\pi_i) = \beta_0 + \alpha^{age}_{a[i]} + \alpha^{gender}_{g[i]} + \alpha^{province}_{p[i]} + \alpha^{religion}_{r[i]} + \alpha^{income}_{m[i]} + \alpha^{education}_{e[i]} \tag{2}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$

$$\alpha^{age}_a \sim \text{Normal}(0, \sigma^2_{age}) \quad \text{for } a \text{ in } 1, 2, \ldots, A \tag{4}$$

$$\alpha^{gender}_g \sim \text{Normal}(0, \sigma^2_{gender}) \quad \text{for } g \text{ in } Male, Female \tag{5}$$

$$\alpha^{province}_p \sim \text{Normal}(0, \sigma^2_{province}) \quad \text{for } p \text{ in } BC, AB, MB, NB, NL, NS, ON, PE, QC, SK \tag{6}$$

$$\alpha^{religion}_r \sim \text{Normal}(0, \sigma^2_{religion}) \quad \text{for } r \text{ in } YES, NO \tag{7}$$

$$\alpha^{income}_m \sim \text{Normal}(0, \sigma^2_{income}) \quad \text{for } m \text{ in } LowerClass, MiddleClass, UpperClass \tag{8}$$

$$\alpha^{education}_e \sim \text{Normal}(0, \sigma^2_{education}) \quad \text{for } e \text{ in } LowEducation, SomeEducation, HighEducation \tag{9}$$

$$\sigma^2_{age} \sim \text{Exponential}(1) \tag{10}$$

$$\sigma^2_{gender} \sim \text{Exponential}(1) \tag{11}$$

$$\sigma^2_{province} \sim \text{Exponential}(1) \tag{12}$$

$$\sigma^2_{religion} \sim \text{Exponential}(1) \tag{13}$$

$$\sigma^2_{income} \sim \text{Exponential}(1) \tag{14}$$

$$\sigma^2_{education} \sim \text{Exponential}(1) \tag{15}$$

3.3 Model justification

3.4 Model Coefficients Interpretation

3.5 Model Equation

We will deploy binary logistic regression analyses to examine the likelihood of individuals casting their vote for the Conservative party or Liberal party. After conducting a thorough investigation into variables that influence voting preferences, we have chosen six key predictors for our models. These include province, age, sex,pre-tax income, religion, and education. Among these, age stands out as the sole continuous variable, while the rest are categorical. Consequently, the framework for our model is structured as follows.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gender} + \beta_3 x_{porvince} + \beta_4 x_{religion} + \beta_5 x_{income} + \beta_6 x_{education}$$

p is the probability of an individual casting their vote for a certain political party (Liberal or Conservative).

β_0(Intercept): This is the base log odds of an individual voting for the political party when all the explanatory variables are at their reference levels.

β_1(age): This coefficient represents the change in the log odds of voting for the political party for each additional year of age. A positive coefficient indicates that as age increases.

β_2(gender): This coefficient indicates the difference in the log odds of voting for the political party for males compared to females (the reference category). A positive coefficient suggests that males are more likely to vote this party than females.

β_3(province): Each of these coefficients (e.g. $\beta_B C$, $\beta_O N$, etc) measures the change in the log odds of voting for the political party for individuals living in each respective province compared to the reference province. Each coefficient reflects the unique effect of residing in that province on voting behavior.

β_4(religion): This coefficient captures the change in the log odds of voting this political party if an individual professes a religious belief (1) versus no religious belief (0, the reference category).

β_5(income): This coefficients reflect the differences in the log odds of voting for the political party based on income levels.

β_6(education): Similar to income, these coefficients represent the effect of education level on the log odds of voting for this political party.

3.6 Post-Stratification

Poststratification is a technique used to refine survey data analysis by breaking down the overall population into distinct segments based on specific attributes. Each segment undergoes an independent assessment. This method involves adjusting the weights of survey responses and computing estimates within each distinct group after data collection. This approach proves particularly advantageous when significant variations exist among certain variables within the survey data, such as the 'province' variable in our study, which exhibits considerable diversity. By accounting for these variations, poststratification enhances the accuracy and reliability of statistical estimates. We utilize this technique to project the voter shares for the Liberal or Conservative parties in the upcoming election. Moreover, we are interested in comparing the outcomes from the initial model to those derived after applying AIC stepwise selection, leading us to calculate two versions of the poststratified estimator, denoted as $\hat{y}^{PS}$, for each political party. The preliminary models categorize the populace using six factors: age, sex, province, education, religion, and pre-tax income. Meanwhile, the AIC-adjusted models exclude 'sex' as a variable. We employ the subsequent mathematical expression to compute the estimator $\hat{y}^{PS}$.

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

In our analysis, $\hat{y}_j$ represents the estimated voting proportion within each subgroup.

N_j denotes the number of individuals within each of these subgroups.

For every subgroup, we first determine its respective proportion estimate. Subsequently, we combine these to ascertain the aggregate voting proportion as predicted by the initial model and also that of the refined model, following the prescribed formula.

4 Result

4.1 Analysis of AIC Model

Table2: Summary the residuals for Conservative Party:

Table 2: Summary of Residuals by Party

Party	Mean_Residual	Median_Residual	SD_Residual	IQR_Residual
Liberal	-0.0001223	0	0.6407023	0
Conservative	0.0002151	0	0.6392760	0

Table3: Summary the coefficient for Liberal Party:

Table 3: Coefficients of the Liberal Model

Term	Estimate
(Intercept)	-1.9993713
age	0.0123667
sexMale	-0.1039276
provinceBritish Columbia	0.3352822
provinceManitoba	0.2709619
provinceNew Brunswick	0.7690145
provinceNewfoundland and Labrador	1.0365775
provinceNova Scotia	0.5395444
provinceOntario	0.8160051
provincePrince Edward Island	-1.7791478
provinceQuebec	0.6458862
provinceSaskatchewan	0.3014459

Term	Estimate
educationLow Education	-0.4533101
educationSome Education	-0.4369064
religionYES	0.0243150
income_before_taxMiddle Class	0.2928792
income_before_taxUpper Class	0.3734962

Table 4: Coefficients of the Conservative Model

Term	Estimate
(Intercept)	-2.1963740
age	0.0185920
sexMale	0.0075194
provinceBritish Columbia	-0.7457189
provinceManitoba	-0.4910314
provinceNew Brunswick	-1.2206592
provinceNewfoundland and Labrador	-1.0069936
provinceNova Scotia	-0.4656448
provinceOntario	-0.9771379
provincePrince Edward Island	-2.3451817
provinceQuebec	-0.6087460
provinceSaskatchewan	-0.4597632
educationLow Education	0.3115370
educationSome Education	0.3175267
religionYES	0.4480938
income_before_taxMiddle Class	0.4080200
income_before_taxUpper Class	0.6162018

4.2 Plot of binary logistic regression models

5 Discussion

Appendix

6 Additional data details

7 Model details

7.1 Posterior predictive check

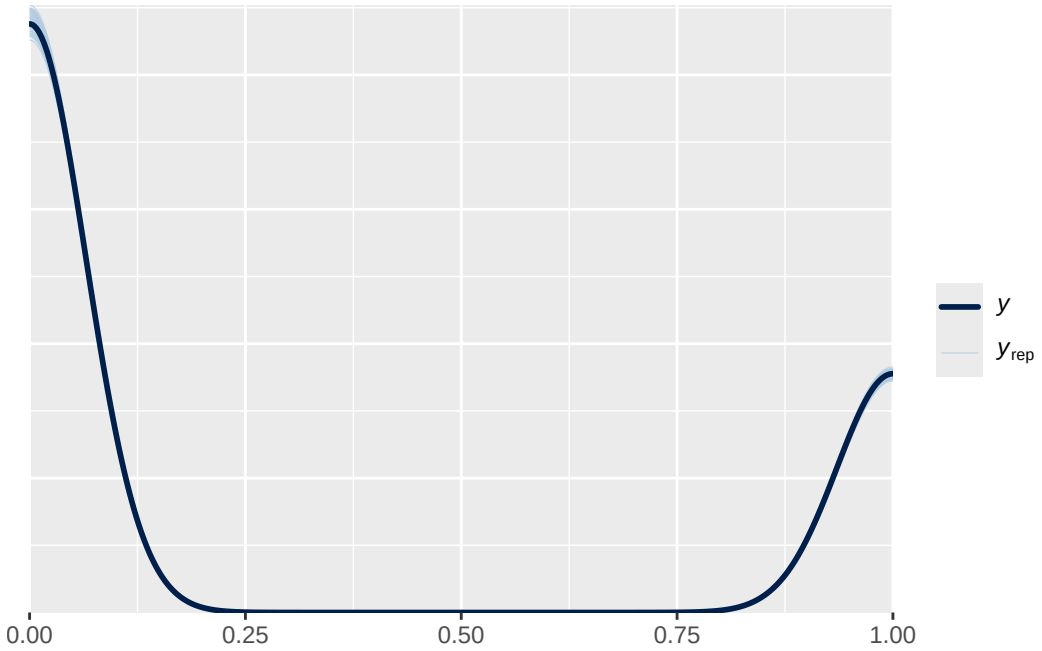

Figure 7: Posterior Predictive Check: Observed vs. Simulated Vote Counts in the Liberal Model

Figure 7 illustrates a posterior predictive check for the liberal model, comparing observed vote counts with those generated by the model's simulations. The dark solid line represents the observed data, while the lighter blue lines represent simulated vote counts from the posterior distribution. The close tracking of the simulated lines to the observed line suggests that the model can effectively replicate the patterns seen in the real vote counts. This indicates a good model fit, as the simulations encompass the range of observed data, providing confidence in the model's predictions. The alignment of the simulations with the observed data points to the adequacy of the model's structure and the priors in capturing the underlying trends of the voting behavior it aims to represent. The graph also visualizes the variance in the simulations, reflecting the model's probabilistic nature and the inherent uncertainty in predicting complex phenomena such as voting behavior.

Histogram of posterior predictive distributions from the liberal model. This plot compares the observed counts of the number of votes for the Liberal party (indicated by the dark line) with the simulated counts from the posterior predictive distribution (shown as histograms). The overlap between the observed counts and the predictive distribution suggests how well the model captures the data. Areas where the observed counts fall outside the bulk of the simulated distributions may indicate model misfit.

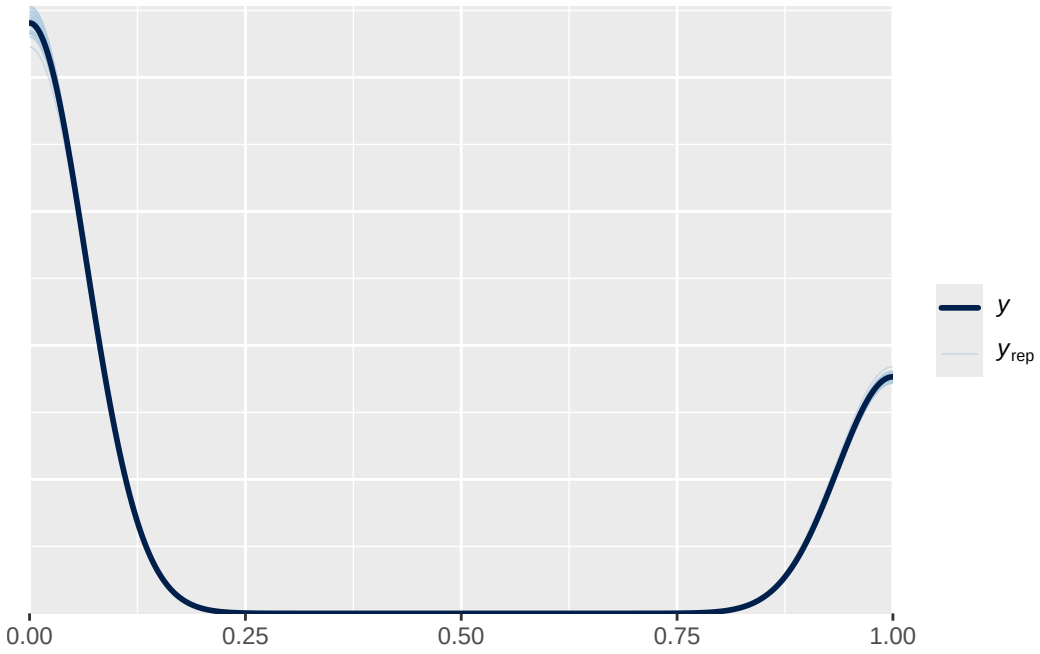

Figure 8: Posterior Predictive Check: Observed vs. Simulated Vote Counts in the Conservative Model

Figure 8 performs a posterior predictive check for a conservative model by overlaying observed vote counts against simulated data derived from the model's posterior distribution. The observed data is denoted by a solid line, while the simulations from the posterior are indicated by lighter lines. The congruence between these two sets of lines—where the simulated data traces the shape and trend of the observed data closely—suggests that the model has a high level of accuracy in reflecting the observed outcomes. This indicates a successful fit of the model to the data, affirming the appropriateness of the selected priors and the model's structure. The spread of the lighter lines around the darker one provides insight into the inherent uncertainty of the model's predictions, which is an expected characteristic of predictive modelling in complex, real-world situations.

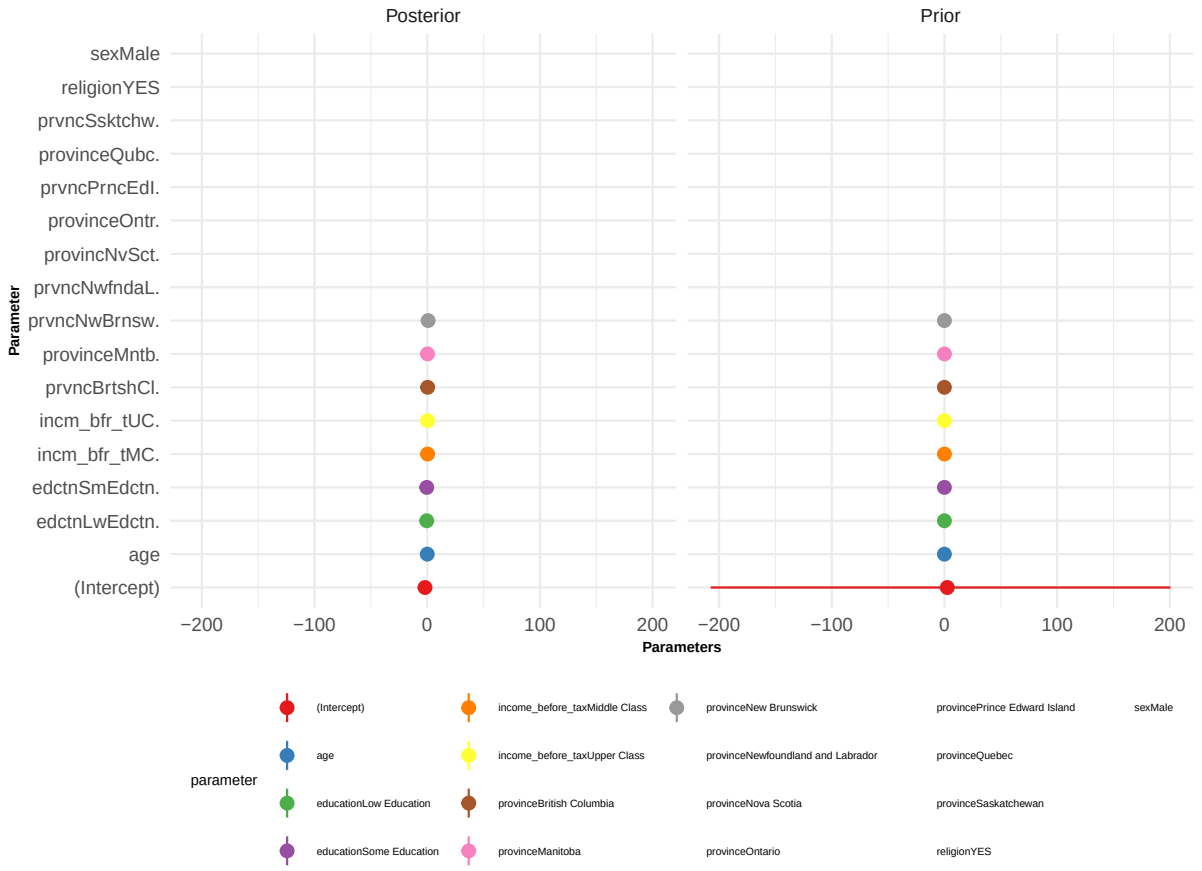

Figure 9: Comparison of Posterior and Prior Distributions in the Liberal Model

Figure 9 portrays the posterior and prior distributions for parameters in a Bayesian liberal model, with a stark divergence in the intercept suggesting significant data-driven adjustments from the model's baseline estimation. Parameters representing demographic and geographical factors are plotted as colored dots along an axis, indicating median values and the extent of uncertainty. While the priors and posteriors largely overlap, indicating consistent estimates for many parameters, the intercept's noticeable shift underscores the data's impactful role in refining the model, particularly highlighting the updated understanding of the model's fundamental starting point.

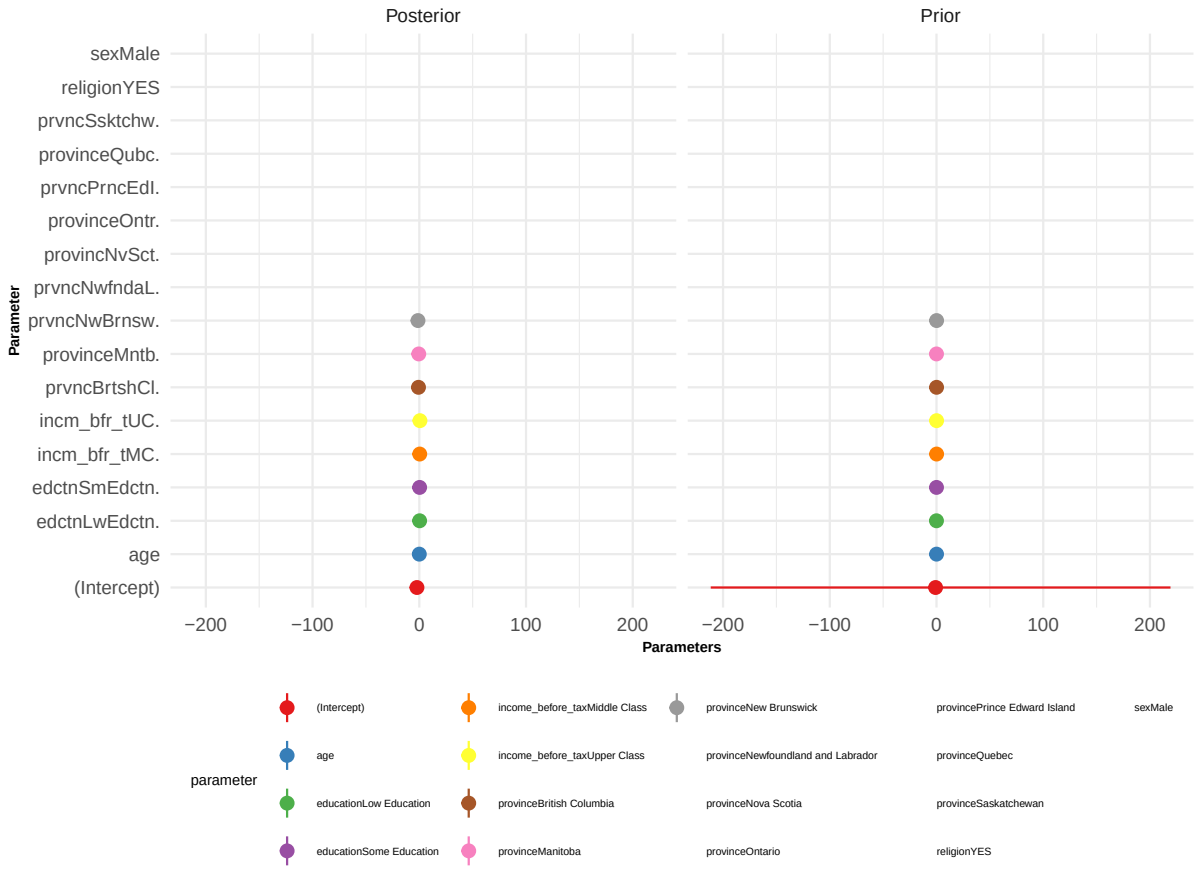

Figure 10: Comparison of Posterior and Prior Distributions in the Conservative Model

Figure 10 presents a side-by-side comparison of the median and spread of prior and posterior distributions for the parameters in a Bayesian model, applied to the Conservative context. Key findings include the posterior distribution of the "age" parameter mirroring its prior, which validates the initial assumption about its influence. In contrast, the intercept and sigma parameters show slight shifts, indicating the data's role in refining these estimates. The placement of medians for other parameters—like sex, religion, and province—demonstrates the data's impact in confirming or adjusting prior beliefs. Overall, the alignment of posterior distributions with priors confirms the model's accuracy and the informative value of the data in capturing the effects on the dependent variable, suggesting a robust model fit.

7.2 Diagnostics

References

n.d.a. *Elections Step by Step | Elections Canada's Civic Education*. https://electionsanddemocracy.ca/canadas-elections/canadas-election-process/elections-step-step.

———. n.d.c. *NORC*. https://gss.norc.org/.

———. n.d.b. *Canadian Election Study*. http://www.ces-eec.ca/.

Canada, Elections. n.d. "Past Elections." – *Elections Canada*. https://www.elections.ca/content.aspx?section=ele&dir=pas&document=index&lang=e.

Comtois, Dominic. 2022. *Summarytools: Tools to Quickly and Neatly Summarize Data*. https://github.com/dcomtois/summarytools.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2020. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

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

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2022. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.