

# Forecasting the 2025 Canadian Federal Election Outcomes: A Study of Demographic Factors Using Logistic Regression Models Refined through Post-Stratification

STA304 - Fall 2023 -Assignment 2

Yi Kuang

2024-07-05

## Introduction

Elections are a cornerstone of democratic systems, allowing citizens to exercise their right to choose their political representatives. Though the Canadian federal election of 2021 saw Justin Trudeau's re-election, his government failed to gain a majority in the House of Commons. Indeed, the Liberal and Conservative parties were very close in terms of popular vote share in 2021 (CBC News, 2021). Furthermore, Trudeau's decision to call an early federal election also stirred various opinions among the public and politicians. This backdrop sets the stage for a potentially competitive contest in the upcoming 2025 federal election, with the Liberal and Conservative parties as the primary contenders (Swigart et al., 2020).

To forecast the dynamics of the upcoming election, our study employed data from the 2017 General Social Survey (GSS) and the 2021 Canadian Election Study (CES) as predictive tools. The GSS census contains a comprehensive set of social, economic, and demographic variables, while the CES dataset offers data on voters' attitudes toward the current political parties. We assume that each participant in the census or survey provided information reflective of their true beliefs. The synthesis of these datasets allows us to analyze voters' inclinations from different social backgrounds towards various political parties. Therefore, we can forecast the outcome of the upcoming 2025 federal election and offer plausible insights to the political parties participating in the election. While focused on Canada, this study also offers practical insights for election analysis in federated nations such as the US, thereby enriching a broader discourse on potential electoral reforms and strategies.

The purpose of this study lies in its use of the GSS and CES datasets to predict the outcome of the next federal election in 2025 between the Liberal and Conservative parties (Kushner et al., 2009). Through the employment of logistic regression modeling for the Liberal Party and the Conservative Party respectively, we aim to investigate the voting patterns for each party. Additionally, to ensure the reliability of our prediction, we have also applied the post-stratification method to align our analysis with a statistically representative demographic. Later, by evaluating the results of the two models, we are able to infer the party that is most likely to secure a greater number of votes in the 2025 election. However, before introducing our later implementation, we hypothesize that in 2025, the Conservative Party will have a greater probability of winning the federal election. This hypothesis is fueled by our intuition about people's attitudes toward the current ruling party.

# Data Collection & Data Cleaning

## Data Collection

As noted in the introduction, our study employed data from the 2017 General Society Survey (GSS), which we deemed as the census for our study, and the 2021 Canadian Election Study (CES), which we deemed as the survey. The CES provides us with information on the potential voter’s backgrounds and, most importantly, their motivation to vote, including factors like their educational levels, marital status, provisional location, and other general personal information. The GSS primarily concerned with the data on living conditions, encompassing a broader spectrum of the population but with a narrower scope of questions on the census question. Both datasets are open-source (Canadian Election Study, 2021; CHASS Data Centre, 2022), offering a wealth of data: the GSS with 20,602 observations and 81 variables and CES with 20,968 observations and 1,062 variables.

In this study, the logistic regression models will be built on our survey data from the CES dataset, supplemented by the post-stratification method on the CES dataset. This entire process necessitates harmonizing variables across these two datasets. For our very first step, we reasoned that, by the time of 2025, the age would be a variable that sets as a threshold for voting, namely, being at least 18 years old in 2025, along with the variable, citizenship, where those who are eligible to vote should be Canadian citizens. With two relatively sensible datasets at hand, we then compared these two datasets and came across four instrumental variables that are relevant to our election prediction, namely, sex, province, marital status, and education. Among them, we chose the variable sex, as it has been a politically controversial topic and a deeper dive into this variable would yield more insights into this election dynamics (Politics & Gender, 2021). To further navigate this variable, our study focuses solely on biological sex, setting aside broader social constructs of gender identity. In addition, marital status is chosen for our study because it is largely associated with an individual’s life stability. For example, a married person might vote for a party that benefits family stability and community-focused policies. There are also studies suggesting the importance of education in fostering participation in voting (Current Psychology, 2022). Therefore, education might be of great interest to investigate. On top of it, the results of the 2021 federal election showed political divides across different provinces, which inevitably highlights the importance of different provinces in our consideration of election prediction (CBC News, 2021). Here, in our study, our primary interest was among the Liberal and the conservative parties, since these two parties combined garnered over half of the entire votes in 2021 with notably similar outcomes (CBC News, 2021). Therefore, we created two variables for voting outcomes for each of these parties. After all, we use age and citizenship to screen for eligible voters and we adopt sex, education, province, and marital status as our four independent variables in later logistic regression models to predict our two dependent variables for the Liberal party and the Conservative party.

## Data Cleaning

Since there is a discrepancy in the time those who filled out the survey and the time of the upcoming election, in our very first step of data cleaning, we filtered the observations on the age variable by adding an additional 8 years for GSS people and an additional 4 years for CES people to check whether the age is equal to or greater than 18. In addition, we only retained observations that recognized their citizenship status.

After screening through age and citizenship to ensure the basic requirements for federal voting, we moved on to align our four dependent variables.

While the census dataset categorizes ‘sex’ into man, woman, and other, the survey dataset’s corresponding variable, gender, encompasses a wide range of gender identities. To reconcile these datasets, we adopt a pragmatic approach to removing all the non-binary gender responses in the survey. This decision was made to avoid misrepresenting the non-binary group of people and inappropriately imputing one category for them that was not theirs. Therefore, gender was cleaned up to have two categories - male and female. Next, since the census data does not include any observations for the three territories in the province variable, we keep only the observations from the ten provinces in the two datasets to facilitate comparisons.

For marital status, we align data by redefinition of the living status. We categorized participants according to their current living status: if a participant lives with someone else (living common-law), we considered him/her married; in addition, we consider all other cases of living alone as single. For our last independent variable, educational levels, we categorized participants in both datasets into the following levels: “less than high school”, “high school and its equivalency”, “profession and non-university certificate”, “bachelor’s”, and “degree above bachelor’s”. The bar to our standard in categorization is high, where certain observations with some knowledge in certain programs, which is primarily the case in our survey dataset, would be categorized at one level below. Specifically, in our census dataset, “trade certificate or diploma” and “University certificate or diploma below the bachelor’s level” are categorized as “profession and non-university certificate” and the rest of the categories correspond to our initial set of categories. In our survey dataset, those obvious labels are categorized according to our initial set of categories; however, we categorized those with “some university” into “high school and its equivalency”, and those with “some technical background” into “less than high school”. Apart from the independent variables, we extracted the Liberal and Conservative parties from the survey dataset as two variables for our later dependent variable prediction and rendered the levels to be binary: “Yes” for voting; “No” for not voting. Lastly, we refined and aligned the two datasets by removing observations that contained missing values in the six variables of our interest, where those who were not willing to disclose their answers were also included, and by harmonizing the name of each variable and corresponding level. Here, the census had 13117 rows and 4 columns, and the survey had 14309 rows and four columns.

Table 1: Summary for Variables in the Cleaned Census Dataset

Variable	Level	Count	Proportion
Province	Ontario	3266	0.25
Province	Quebec	2590	0.20
Province	British Columbia	1574	0.12
Province	Alberta	1089	0.08
Province	Nova Scotia	946	0.07
Province	New Brunswick	868	0.07
Province	Saskatchewan	816	0.06
Province	Manitoba	803	0.06
Province	Newfoundland and Labrador	743	0.06
Province	Prince Edward Island	422	0.03
Education	High school diploma or a high school equivalency certificate	4625	0.35
Education	Bachelor’s degree (e.g. B.A., B.Sc., LL.B.)	3443	0.26
Education	Less than high school diploma or its equivalent	2925	0.22
Education	College, CEGEP or other non-university certificate or diploma	2124	0.16
Sex	Female	6790	0.52
Sex	Male	6327	0.48
Marital Status	Married	7089	0.54
Marital Status	Single	6028	0.46

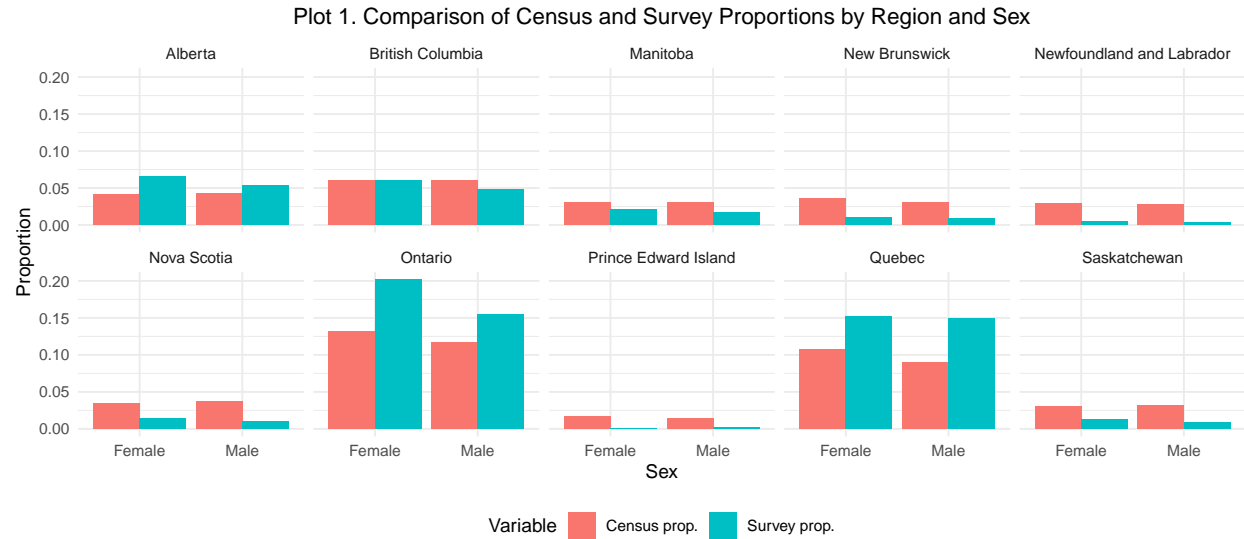
Table 2: Variables in the Cleaned Survey Dataset Summary

Variable	Level	Count	Proportion
Liberal	No	10459	0.73
Liberal	Yes	3850	0.27
Conservative	No	10710	0.75
Conservative	Yes	3599	0.25
Province	Ontario	5101	0.36
Province	Quebec	4306	0.30
Province	Alberta	1706	0.12
Province	British Columbia	1551	0.11
Province	Manitoba	543	0.04
Province	Nova Scotia	352	0.02
Province	Saskatchewan	306	0.02
Province	New Brunswick	268	0.02
Province	Newfoundland and Labrador	134	0.01
Province	Prince Edward Island	42	0.00
Education	Bachelor’s degree (e.g. B.A., B.Sc., LL.B.)	4214	0.29
Education	High school diploma or a high school equivalency certificate	3352	0.23
Education	College, CEGEP or other non-university certificate or diploma	3125	0.22
Education	University certificate, diploma or degree above the bachelor’s level	2037	0.14
Education	Less than high school diploma or its equivalent	1581	0.11
Sex	Female	7809	0.55
Sex	Male	6500	0.45
Marital Status	Married	8967	0.63
Marital Status	Single	5342	0.37

Table 1 and Table 2 display the categories of each variable in the two datasets and their respective proportions. Firstly, Table 2 outlines voter preferences for the Liberals, Conservatives in 2021. The Liberals and Conservatives are in close competition, with the Liberals holding a marginal lead of only 2% more votes than the Conservatives. Moreover, Tables 1 and 2 show that the populations of both Ontario and Quebec increased by about 10% from the 2017 census data to the 2021 survey data, suggesting a tendency for people in Canada to migrate to more densely populated provinces. This trend also implies that it might be more strategic for political parties to focus their campaign efforts in these two provinces.

As highlighted in the Introduction, the level of education plays a significant role in influencing voters’ engagement with elections. Between 2017 and 2021, there was a notable decrease in the percentage of people with a high school degree or less (from 57% to 34%). This change in the education variable offers insight into its impact on voting patterns. The distribution of sex is fairly balanced in both datasets. However, there is a notable difference in marital status between the two datasets. In the census data, the proportion of married individuals is slightly higher than that of single individuals. In contrast, in the survey data, where individuals living in common-law relationships are classified as married, the proportion of married individuals is significantly higher than that of single individuals.

This analysis underscores the importance of demographic shifts and their potential influence on voting behaviors, highlighting the need for political parties to adapt their strategies to these changing dynamics.



Plot 1 compares census and survey data on the proportion of men and women in each province. For both census and survey data, the ratio of men to women does not always show similarity across different provinces. Notably, Ontario has about 5% more women than men. This provides evidence to look at the sex ratio within each province afterwards and reweight it using post-stratification.

## Methods

In our analysis, we had employed two statistical methods, logistic regression model and Post-stratification.

### Logistic Regression Model

Broadly speaking, the logistic model is a powerful tool for predicting binary outcomes, such as “Yes” or “No.” It was actualized by transforming the relationship between the independent and dependent variables into a logit function, where the probability of the event of interest can be calculated, lying between 0 and 1.

In this study, we aimed to adopt the logistic regression model to predict the 2025 election outcome between the Liberal and Conservative Parties, based on the demographic factors mentioned in the survey data. Two sets of logistic regression models were fitted, each modeling the relationship between our four demographic factors and the voting inclination for or against a particular party. In our study, the dependent variable for our two models was the inclination to vote for the Liberal Party and the Conservative Party, respectively, while the independent variables remained the same across these two models.

To provide a clearer sense of our model, we present a plausible mathematical representation:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 M + \beta_2 Single + \beta_3 ON + \beta_4 QC + \dots + \beta_{12} LessHigh + \dots + \beta_{15} AboveBach$$

- **p** represents the probability of a voter choosing a particular party, which is embedded in a logit function  $\log \frac{p}{1-p}$ . In this case, for our two models, this could be the probability of either voting for the Liberal Party or voting for the Conservative Party.
- **Sex** is divided into two levels, male and female. Here, in the formula, male and female are represented by the dummy variable attached by  $\beta_1$ . If this dummy variable turns out to be 1, then a male is initiated.
- **Marital Status** could be boiled down into either married or single, reflecting an individual’s current living status in our study. In the formula, it categorizes the dummy variables attached by  $\beta_2$ , where if this dummy variable turns out to be 1, then the observation is single.
- **Education** is divided into five levels fully described in the data section and the model incorporates four of these levels as dummy variables associated with coefficients from  $\beta_{12}$  to  $\beta_{15}$ . Here, in the formula as a representation of the mechanism of our model, four different levels of education were represented by the dummy variables attached by from  $\beta_{12}$  to  $\beta_{15}$  where if the variable attached to  $\beta_{12}$  represents “less than high school”, the variable attached to  $\beta_{13}$  represents “high school and its equivalency”, the variable attached to  $\beta_{14}$  represents “profession and non-university certificate”, the variable attached to  $\beta_{15}$  represents “bachelor’s”, and if all these were zero, then the value of the remaining level, “degree above bachelor’s”, is initiated in our formula, serving as a reference category for education.
- **Province** represents the geographic location of the voter, categorized based on province. In the formula, it is represented by the dummy variables associated from  $\beta_3$  to  $\beta_{11}$ , where the very same mechanism from the Education variable applies.
- $\beta_0$  is the intercept, representing the log-odds when all independent variables are zero and it gives us the baseline of the log-odds of the outcome independent of our independent variables.
- $\beta_1$  to  $\beta_{15}$  are coefficients associated with each variable. Each of these signifies the unit change in the log-odds for a voter choosing a particular party for a one-unit change in the corresponding variable, while keeping other variables constant.

These coefficients are of great interest as they help us understand how each independent variable influences the log-odds of a voter voting for a particular party.

## Post-stratification

Post-stratification allows us to weight the results of our data in relation to the population to better represent the patterns of the entire population. In our study, different combinations or “cells” are stratified by our independent variables of interest. Within each cell, representing a specific demographic group, a prediction (or technically, an estimation) based on our logistic regression model is made, and the weight is calculated according to the variables stratified, which is then applied to the result yielded by the model in each cell.

With 10 different provinces, 5 education level, 2 types of sex and 2 marital status, we have  $10 \times 5 \times 2 \times 2 = 200$  cells. Each cell has different weight assigned according to the proportion of associated variables of our survey in relation to our census. We believe, in doing so, a more accurate prediction will be made.

A mathematical representation of the post-stratification is shown as follows:

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

- $N_j$ : It represents the population size of this subgroup or cell  $j$  in the overall population, where this subgroup is defined by certain features, for example, sex, province, marital status, and education in our study.
- $\hat{y}_j$ : The predicted or our estimated probability of vote for a particular party (i.e. calculated by the amount of those who voted over the overall amount of response) for a particular subgroup or cell after some calculations of the logistic regression model.
- $\sum N_j \hat{y}_j$ : The sum of the predicted or our estimated probability of vote for a particular party for each stratum. Here, each prediction of the stratum is weighted by the respective stratum’s population size.
- $\sum N$ : The total population size of our census across different strata.
- $\hat{y}^{PS}$ : This is the predicted or our estimated probability of vote

## More Clarity in the Process

1. We firstly built our logistic regression model as fitted by the survey data on the four aforementioned demographic variables and the corresponding voting inclinations. Here, two sets of models were built different in the dependent variable, the voting inclination of either the Liberal Party or of the Conservative Party, while keeping the independent variables the same.
2. We then grouped our census people into different cells based on their demographic characteristics of our interest and we calculated the sum of how many cells in total.
3. Based on our grouping in step 2, we predicted or estimated probability of vote in each cell by our logistic regression model which built by our survey data in step 1. On top of it, we multiply our prediction and the population size in that specific cell.
4. We then ungrouped the people and divided the sum of the results from the multiplication across different cells from step 3 by the total population size of the census.
5. The last four steps could be used to build a model for either the Liberal Party vote prediction model or the Conservative Party vote prediction model. Either way, we would like to repeat it for the second prediction model for another party.
6. We then compared the probability of these two parties and recorded the results in our later result section.

After all these, we are able to have a more accurate prediction of the vote by post stratification technique after fitting the logistic regression model, compared to solely fitting two models and comparing results from these two models. The leverage in accuracy is marked by weighing predictions on people with different backgrounds according to its proportion to the overall census data, a more comprehensive view on people's demographics.

### **Assumptions**

However, before all the proposed methodology, we would like to check some assumptions on it, primarily our logistic regression model and later post stratification technique.

We aim to have two binary outcomes, one for the Liberal Party and another for the Conservative Party. Therefore, logistic regression fits our intention. Also, since all variables are categorical, we do not need to consider the cases of linearity in the logit for continuous variables. On top of it, each voter's choice of vote is independent of one another, and the variables are independent of each other as well, given the obvious literal significance of the independent variables. Thus, there is an absence of multi-collinearity. Since our considerations are on categorical variables, there is also not much worry about strong outliers. By satisfying all four assumptions of logistic regression, we might safely proceed with our analysis.

In addition to the logistic regression model, we might want to assume the representativeness of our data for both the survey and the census and the relevance of our stratification variables to our interest in the study. Therefore, a post-stratification technique is safe to proceed with.



## Results

Since we adopted two methodologies in this study, we might want to examine the results of these two methods in this section.

### Logistic Regression Model

Two sets of summary tables were produced for the logistic regression models predicting the voting probability of the Liberal Party and the Conservative Party respectively.

Interestingly, the reference categories were shared across these two models, where the level “Bachelor’s” was set as the reference category for the education variable, the level “Married” was set as the reference category for the marital status variable, the level “Alberta” was set as the reference category for the province variable, and the level “female” was set as the reference category for the sex variable.

In the summary table of the Liberal Party and the Conservative Party, we held a 0.05 statistical significance level for rejecting or adopting the meaningfulness (i.e., the odds of voting for this level are similar to the reference category) of the corresponding variable compared to the reference category. Based on comparing the p-value against this 0.05 threshold, if the corresponding p-value is less than or equal to 0.05, we might want to reject it; if it is greater than 0.05, we might want to adopt it. By contrasting the p-value against this threshold, we could better interpret the validity of the corresponding level imputed in the model compared to the reference category, ultimately, in relation to the inclination to vote for a particular party.

Table 3: Summary Table of the Logistic Regression Model on the Liberal Party

Coefficients	Estimate	p
Intercept	-1.226222	< 2e-16
Education: College, CEGEP or other non-university certificate or diploma	-0.425003	8.48e-15
Education: High school diploma or a high school equivalency certificate	-0.289710	4.30e-08
Education: Less than high school diploma or its equivalent	-0.421840	1.59e-09
Education: University certificate, diploma or degree above the bachelor’s level	0.121450	0.03690
Martial: Single	0.007033	0.86041
Province: British Columbia	0.331271	7.95e-05
Province: Manitoba	0.260884	0.02589
Province: New Brunswick	0.745341	1.81e-07
Province: Newfoundland and Labrador	0.990642	1.33e-07
Province: Nova Scotia	0.779896	1.02e-09
Province: Ontario	0.638335	< 2e-16
Province: Prince Edward Island	0.272699	0.45954
Province: Quebec	0.176187	0.01271
Province: Saskatchewan	-0.477100	0.00823
Sex: Male	0.055541	0.15090

From the summary table of the Liberal Party (Table 3.), it is evident that the odds of voting at the provincial level of “Prince Edward Island” might not be statistically different from “Alberta” in our model, since its p-value is significantly greater than our threshold, which is also the case for the marital status level of “Single” compared to “Married”, and for the sex level of “Male” compared to “Female”, as they were compared against their corresponding reference category. As can be inferred, sex and marital status respectively did not differentiate significantly in voting behaviours. From the summary table of the Conservative Party (Table 4.), in addition to the provincial level of “Prince Edward Island” as previously suggested in the table of the Liberal Party, the odds of voting at another provincial level “Saskatchewan” might not be statistically different from the reference province “Alberta”. Overall, no significant difference compared to the reference category does not suggest no practical significance. One should consider a broader set of socio-political factors upon evaluating these levels.

Table 4: Summary Table of the Logistic Regression Model on the Conservative Party

Coefficients	Estimate	P Value
Intercept	-0.62793	< 2e-16
Education: College, CEGEP or other non-university certificate or diploma	0.32350	6.00e-09
Education: High school diploma or a high school equivalency certificate	0.23396	2.84e-05
Education: Less than high school diploma or its equivalent	0.33112	1.94e-06
Education: University certificate, diploma or degree above the bachelor’s level	-0.14934	0.0262
Martial: Single	-0.42454	< 2e-16
Province: British Columbia	-0.77413	< 2e-16
Province: Manitoba	-0.46798	1.27e-05
Province: New Brunswick	-1.22529	5.83e-13
Province: Newfoundland and Labrador	-1.03519	2.85e-06
Province: Nova Scotia	-1.05983	1.84e-13
Province: Ontario	-0.57893	< 2e-16
Province: Prince Edward Island	-0.53986	0.1149
Province: Quebec	-1.38767	< 2e-16
Province: Saskatchewan	0.06811	0.5946
Sex: Male	0.55895	< 2e-16

Furthermore, the estimates of the coefficients also provide some insights into our understanding of voting behaviours. The magnitude of the estimate indicates the strength of the relationship to the voting behaviour. The positive estimate indicates an increase in the odds of voting for a particular party and vice versa. Interestingly, most of the levels in the variables were in the opposite direction of their counterparts. Also, as can be seen from the conservative table, it should be noted that being male and being in married status, the odds of voting for the conservative party are significantly high as compared to their counterparts. This might suggest a potential political inclination of the conservative party toward this group of people. Province-wise, Quebec showed the greatest opposition among other provinces against the conservative party as suggested in the table of the conservative party with an estimate of -1.38767, following New Brunswick with an estimate of -1.22529, while Newfoundland and Labrador showed the greatest support among other provinces for the liberal party as suggested in the table of the liberal party with an estimate of 0.990642. This might suggest a potential policy-related influence that resonates with those local voters, where those in Newfoundland and Labrador, Quebec, and New Brunswick might not be favored by the conservative policies. In addition, it is clear from both tables that, compared to “Bachelors”, only those with degrees above the bachelor’s level showed favor towards the Liberal Party, while others leaned more towards the Conservative Party, which might suggest potential alignment of educational levels with political inclination.

### Post-Stratification

Table 5: Prediction on the Conservative and the Liberal Party in the 2025 Election

Party	Prediction
Conservative	0.2625759
Liberal	0.2623956

After building the models, we proceeded to the post-stratification technique to more accurately represent the overall context. As described in the method section, we have the probability of each party getting elected in the upcoming 2025 federal election, as detailed in the table (Table 5.).

As the results from the post-stratification on the probability calculated by the logistic regression model, the conservative party has a 0.02% more percentage probability over the liberal party. It might seem a

possible result, given the already competitive nature in these two parties and that the data set we used only captured the Canadian demographics and their inclination for voting in 2021 and in 2017, as encapsulated by the survey dataset and census dataset we employed in the model, which shows a time gap in explaining the upcoming election in 2025. The difference is not overwhelming significant, but a small difference might result in significant shift in Canada's political status. The constantly evolving societal trends might play a significant role in these couple of years and we might expect more perspectives on people's views for the upcoming election in 2025.

Table 5 shows the results of the probability of the predicted voting outcome from the two logistic regression models weighted by the post stratification method. As can be seen, the conservative party shows marginally more probability over the liberal party.

# Conclusion

This study adopted demographic variables to predict the likelihood of winning the Canadian federal election in 2025 between the Liberal and Conservative Parties. We first built two logistic regression models using four aforementioned informative variables from the survey data. We then weighted the models using the post-stratification technique according to the census data to make a plausible prediction.

As it turned out, the Conservative Party would have a marginally higher vote share in 2025, with a difference to the Liberal Party of about 0.02% in the vote share. This outcome underscores the influence of demographic factors such as sex, education, marital status, and province in voting preferences as we included in our model. The prediction points towards a potential win for the Conservative Party in the upcoming 2025 election which aligns with our initial hypothesis, which was based more on, admittedly, intuition than empirical evidence.

Though there is an alignment of our results with our initial hypothesis, it is important to note that data-driven decision-making should be taken seriously in politics. It is because relying on data helps us minimize intuitive biases that influence decisions affecting a family, a community, a province, and an entire nation.

Overall, this study is subject to certain limitations that, if addressed, could refine our predictions further:

1. The census and survey datasets date back to 2017 and 2021 respectively. The time gap might suggest our study did not capture the most recent change in vote sentiment and the demographic shift, thereby yielding relatively non-representative datasets to be studied.
2. In the Data Section, recategorizing different levels to fit the levels of the variables in the census might oversimplify the intricate dynamics of the demographics. A more nuanced categorization would render more representative datasets to be investigated.
3. Other variables might have some correlational relationship with our prediction which we failed to incorporate into our model. This might lead our prediction towards a biased outcome.
4. Our study involves removing the observations that contained missing values. While this certainly provides a cleaner dataset that is easy to study, this might eliminate valuable observations that might provide insights into our study.

For future studies, we suggest the following:

1. Future studies should definitely base the study on a more recent dataset. A more recent dataset would be more reflective of the most recent demographics and people's attitudes toward different parties, thereby offering more representative values.
2. Future studies might want to adopt more granular categorizations of the data. A more refined version of categorization might represent more nuances within the demographic groups, which helps build a more robust model.
3. More research on plausible independent variables that are relevant to the study should be conducted, so that more patterns associated with the outcome of our interest would be discovered.
4. The employment of an applicable imputation method to account for the missing values would give us a wider range of data to be used in our study.
5. Consideration should be given to more advanced statistical models which would explain more sophisticated relationships within different variables. For example, a Multinomial Logistic Regression is capable of predicting beyond binary outcomes.

## Generative AI Statement

I used ChatGPT Version available on Nov 23, 2023 (ChatGPT, n.d.). The interaction is rather scattered and not extensive at all. I used the tool primarily for inspiration for some aspects of the project and guidance on some methodologies. The full interaction with ChatGPT is in the appendix.

Frankly, it inspired me to some degree throughout the entire assignment. In the introduction section, I primarily asked about related ideas on political elections and voting, hoping for some insights in writing a beautiful introduction. Moving on, in the data part, I asked about the formulation of some R code and some mathematical formulas, which might also be the case in the later parts of the assignment. In the method part, instead of scratching slides for ideas, I asked about the logistic regression model and related concepts. The relatively more extensive use of it might show upon the conclusion section where I asked general questions regarding limitations and possible ways to address them. Apart from that, there are scattered uses for reformulating my sentences.

Please feel free to scan through.

## Bibliography

- 1.CBC. (2021). Dynamic Documents for R. Rmarkdown.rstudio.com. <https://rmarkdown.rstudio.com/docs/>
- 2.CES editor. (n.d.). 2021 Canadian Election Study – Canadian Election Study. <http://www.ces-eeec.ca/2021-canadian-election-study/>
- 3.CHASS Data Centre. (n.d.). Dc.chass.utoronto.ca. Retrieved November 24, 2023, from <https://dc.chass.utoronto.ca/myaccess.html>
- 4.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.
- 5.Grolemund, G. (2014, July 16). Introduction to R Markdown. Rmarkdown.rstudio.com. [https://rmarkdown.rstudio.com/articles\\_intro.html](https://rmarkdown.rstudio.com/articles_intro.html)
- 6.Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL <http://www.jstatsoft.org/v21/i12/>.
- 7.Hadley Wickham, Romain François, Lionel Henry, Kirill Müller and Davis Vaughan (2023). dplyr: A Grammar of Data Manipulation. <https://dplyr.tidyverse.org>, <https://github.com/tidyverse/dplyr>.
- 8.Hao Zhu (2021). kableExtra: Construct Complex Table with ‘kable’ and Pipe Syntax. <http://haozhu233.github.io/kableExtra/>, <https://github.com/haozhu233/kableExtra>.
- 9.H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- 10.Kushner, J., Siegel, D., & Stanwick, H. (1997). Ontario Municipal Elections: Voting Trends and Determinants of Electoral Success in a Canadian Province. Canadian Journal of Political Science, 30(3), 539–553.<https://doi.org/10.1017/s0008423900016000>
- 11.Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno and Christopher Barr (2021). openintro: Data Sets and Supplemental Functions from ‘OpenIntro’ Textbooks and Labs. <http://openintrostat.github.io/openintro/>, <https://github.com/OpenIntroStat/openintro/>.
- 12.OpenAI. (2023). ChatGPT (Feb 13 version) [Large language model]. <https://chat.openai.com>
- 13.R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- 14.Swigart, K. L., Anantharaman, A., Williamson, J. A., & Grandey, A. A. (2020). Working While Liberal/Conservative: A Review of Political Ideology in Organizations. Journal of Management, 46(6), 014920632090941. <https://doi.org/10.1177/0149206320909419>
- 15.Xie Y (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.43, .

## Appendix

### 1. The first six rows of the census dataset

```
## # A tibble: 6 x 81
##   caseid   age age_first_child age_youngest_child_under_6 total_children
##   <dbl> <dbl>         <dbl>                 <dbl>         <dbl>
## 1     1  52.7           27                     NA             1
## 2     2  51.1           33                     NA             5
## 3     3  63.6           40                     NA             5
## 4     4   80           56                     NA             1
## 5     5   28           NA                      NA             0
## 6     6   63           37                     NA             2
## # i 76 more variables: age_start_relationship <dbl>,
## #   age_at_first_marriage <dbl>, age_at_first_birth <dbl>,
## #   distance_between_houses <dbl>, age_youngest_child_returned_work <dbl>,
## #   feelings_life <dbl>, sex <chr>, place_birth_canada <chr>,
## #   place_birth_father <chr>, place_birth_mother <chr>,
## #   place_birth_macro_region <chr>, place_birth_province <chr>,
## #   year_arrived_canada <chr>, province <chr>, region <chr>, ...
```

### 2. The first six rows of the survey dataset

```
## # A tibble: 6 x 1,062
##   cps21_StartDate      cps21_EndDate      Duration__in_seconds_
##   <dtm>              <dtm>              <dbl>
## 1 2021-09-19 06:14:46 2021-09-19 06:28:25      818
## 2 2021-09-15 15:23:33 2021-09-15 15:46:57     1403
## 3 2021-08-20 09:44:55 2021-08-20 09:57:51      775
## 4 2021-09-14 18:26:56 2021-09-14 18:40:41      825
## 5 2021-09-17 17:24:04 2021-09-17 17:51:44     1660
## 6 2021-09-09 14:13:01 2021-09-09 14:35:14     1332
## # i 1,059 more variables: RecordedDate <dtm>, cps21_ResponseId <chr>,
## #   DistributionChannel <chr>, UserLanguage <chr>,
## #   cps21_consent_t_First_Click <dbl>, cps21_consent_t_Last_Click <dbl>,
## #   cps21_consent_t_Page_Submit <dbl>, cps21_consent_t_Click_Count <dbl>,
## #   cps21_consent <hvn_lbl>, cps21_captcha_t_First_Click <dbl>,
## #   cps21_captcha_t_Last_Click <dbl>, cps21_captcha_t_Page_Submit <dbl>,
## #   cps21_captcha_t_Click_Count <dbl>, cps21_citizenship <hvn_lbl>, ...
```

### 3. The full interaction with ChatGPT for this project:

User: I'm working on Assignment 2 for STA304, a statistics course, with the goal of researching the prediction of the 2025 Canadian election results. What do you think might influence the voting outcomes? I mean, citizens participating in the voting process, what characteristics of the voters might influence their voting behavior?

ChatGPT: Certainly, understanding the characteristics of individual voters is essential for predicting election outcomes. Here are some key voter characteristics that often influence voting behavior:

#### 1. Demographics:

- **Age:** Younger and older voters may have different priorities and political leanings.

- **Gender:** Gender can influence voting preferences, with variations in issues deemed important by different genders.
- **Ethnicity/Race:** Cultural and ethnic backgrounds can impact political affiliations and priorities.
- **Income and Socioeconomic Status:** Economic factors often play a role in shaping political preferences.

## 2. **Education:**

- Levels of education can influence political attitudes. Higher levels of education may correlate with specific policy preferences.

## 3. **Religion:**

- Religious beliefs and values can influence voters' stances on social and moral issues.

## 4. **Geography:**

- Urban and rural voters may have different perspectives on issues such as infrastructure, environment, and social policies.

## 5. **Political Affiliation:**

- Past party affiliation and loyalty can strongly influence voting behavior.

## 6. **Ideological Beliefs:**

- Voters' political ideologies, such as liberal, conservative, or moderate, shape their preferences and decisions.

## 7. **Social Issues:**

- Views on social issues, such as abortion, LGBTQ+ rights, and immigration, can significantly impact voting choices.

## 8. **Perception of Candidates:**

- Voters' perceptions of a candidate's personality, trustworthiness, and leadership qualities can be decisive.

## 9. **Economic Concerns:**

- Personal economic circumstances and concerns about the economy at large can influence voting decisions.

## 10. **Media Consumption:**

- The sources and types of media individuals consume can shape their political opinions.

## 11. **Civic Engagement:**

- Previous voting history, participation in community activities, and overall civic engagement can indicate a voter's likelihood of turning out.

Remember, the interplay of these factors is highly complex, and individuals may prioritize these characteristics differently. It's also important to consider that public opinion can shift based on current events and the political climate.

User: Can you explain logistic regression model and post-stratification to the general science readers? Chat-GPT: Certainly! Let's break down both logistic regression and post-stratification in simple terms:



## **Logistic Regression:**

**Logistic Regression** is a statistical method used for predicting the probability of an event occurring. Unlike linear regression, which predicts a continuous outcome, logistic regression is specifically designed for binary outcomes, where the result is either 0 or 1 (e.g., success or failure, win or lose).

Here's a simple breakdown of how it works:

### **1. Outcome Probability:**

- Logistic regression models the relationship between one or more independent variables and the probability of a particular outcome.

### **2. Sigmoid Function:**

- Logistic regression uses a sigmoid (S-shaped) function to map the linear combination of input features to a range between 0 and 1. This ensures that the predicted probabilities fall within a valid probability range.

### **3. Log-Odds Transformation:**

- The logistic regression model works with log-odds, also known as the logit function. The log-odds are then transformed back into probabilities using the sigmoid function.

### **4. Coefficient Interpretation:**

- The coefficients in logistic regression represent the change in log-odds for a one-unit change in the corresponding independent variable.

In simpler terms, logistic regression helps us understand the relationship between input factors and the likelihood of a specific event happening. It's widely used in various fields, including medicine, finance, and social sciences, for predicting binary outcomes.

## **Post-Stratification:**

**Post-Stratification** is a technique used in survey sampling and statistics to improve the accuracy of estimates by adjusting for differences between the sample and the population it represents. Here's a straightforward explanation:

### **1. Stratification:**

- Initially, the population is divided into strata or groups based on certain characteristics (e.g., age, gender, region).

### **2. Random Sampling:**

- Within each stratum, a random sample is taken. This helps ensure representation from each subgroup in the final sample.

### **3. Post-Stratification:**

- After data collection, the sample is compared to the known population characteristics. If the sample doesn't perfectly reflect the population, post-stratification adjusts the weights of each observation to align the sample with the population proportions.

### **4. Improved Estimates:**

- By accounting for variations in the population, post-stratification improves the accuracy of estimates and reduces bias.

In simple terms, post-stratification helps correct for any imbalance in the sample by adjusting the weights of individuals within different subgroups. This ensures that survey results are more representative of the entire population, making the findings more reliable and applicable.

User: If we assume all the answers from the survey are true and the voters will not change their mind after finish the survey, what will be the limitation to this assignment? ChatGPT: Assuming that all the answers from the survey are true and that voters will not change their minds after completing the survey introduces several limitations to the assignment. Here are some key considerations:

**1. Static Nature of Responses:**

- People's opinions and political preferences can change over time due to evolving circumstances, new information, or changing social and economic conditions. Assuming that survey responses remain fixed may not capture the dynamic nature of political attitudes.

**2. External Events:**

- Unforeseen events, such as political scandals, economic crises, or major policy changes, can significantly impact voter preferences. If the survey does not account for the possibility of such events, the predictions may be less accurate.

**3. Response Bias:**

- Assuming that all survey responses are true may not account for potential biases in responses. Respondents might provide socially desirable answers, and there could be issues related to non-response bias or sample selection bias.

**4. Limited Predictive Power:**

- Even with accurate survey responses, predicting election outcomes involves complex dynamics influenced by various factors. The assignment might not consider all relevant variables or interactions that play a role in shaping electoral results.

**5. Homogeneity of Responses:**

- Assuming that all voters within a certain category (e.g., age group, income level) will have the same opinions and voting behavior oversimplifies the diversity of opinions within these groups. Individuals within the same demographic may still have varying political views.

**6. Incomplete Information:**

- Surveys typically capture a snapshot of opinions at a specific point in time. Without considering potential changes in voter sentiment leading up to the election, the assignment may lack a comprehensive understanding of the political landscape.

**7. Overreliance on Self-Reported Data:**

- Relying solely on self-reported data assumes that respondents accurately represent their true feelings and intentions. People might not always provide accurate information due to social desirability bias, memory recall issues, or other cognitive biases.

**8. Modeling Assumptions:**

- The assignment might be based on certain assumptions about the modeling technique (e.g., logistic regression) and the suitability of the chosen variables. If these assumptions are not met, the predictive power of the model could be compromised.

**9. Lack of Validation:**

- Without validation against real-world election results or other independent data sources, it's challenging to assess the accuracy and generalizability of the predictive model.

It's crucial to recognize these limitations and consider the uncertainties inherent in predicting election outcomes. Real-world political scenarios are highly dynamic, and various unforeseen factors can influence the final results.

User: no, we are using regression model and post-stratification to predict the 2025 election results. And we have 2 dataset, one is from survey, one is census data. Based on this additional information, can you explain more on the limitation? ChatGPT: Certainly, with the additional information that you are using a regression model and post-stratification with two datasets (survey and census data) to predict the 2025 election results, let's delve into some specific limitations associated with this approach:

**1. Assumption of Homogeneity:**

- The assumption that survey respondents within each stratum (defined by the census data) are homogeneous in their voting behavior may not hold true. Even within specific demographic groups, there can be substantial variation in political preferences.

**2. Sampling Bias:**

- If the survey sample is not representative of the entire population, the results may suffer from sampling bias. Certain groups may be overrepresented or underrepresented, leading to inaccurate predictions.

**3. Inaccuracy in Census Data:**

- Census data might not perfectly capture the current demographic and socioeconomic landscape. If there are inaccuracies or if the population has changed significantly since the last census, it could affect the precision of post-stratification.

**4. Limited Predictive Power of Census Variables:**

- Census data typically includes demographic variables, but it may lack more dynamic predictors that can influence voting behavior, such as recent economic changes or public sentiment shifts. The model's predictive power may be limited by the variables available in the census data.

**5. Assumption of Stationarity:**

- Assuming that the relationships between predictor variables and voting behavior remain constant over time (stationarity) may not hold true. Political landscapes are subject to change, and historical relationships may not accurately predict future outcomes.

**6. Data Quality and Missing Values:**

- Both the survey and census datasets may have missing values or data quality issues. Incomplete or inaccurate data can affect the reliability and validity of the regression model.

**7. Model Overfitting or Underfitting:**

- The regression model may suffer from overfitting (capturing noise in the data) or underfitting (oversimplifying the relationships). The model's performance depends on the balance between complexity and generalizability.

**8. Extrapolation Risks:**

- If the model is used to predict outcomes beyond the range of the observed data, there is a risk of extrapolation. Extrapolating too far into the future may lead to less reliable predictions.

**9. External Events and Shocks:**

- The model may not account for unforeseen events or external shocks that could significantly impact voter behavior. Political, economic, or social events occurring close to the election may not be adequately captured in the model.

**10. Validation Challenges:**

- It may be challenging to validate the model’s accuracy without real-world election results. The lack of a historical dataset with similar conditions makes it difficult to assess how well the model performs in a predictive sense.

## 11. Dynamic Survey Responses:

- Assuming that survey responses remain fixed may not consider the possibility of changing opinions or shifts in sentiment over time, especially as the election date approaches.

To mitigate these limitations, it’s crucial to thoroughly validate the model, use robust statistical techniques, and continually reassess the assumptions and data quality throughout the analysis. Additionally, incorporating real-time data and considering uncertainties in predictions can enhance the model’s reliability.

User: I am using R to create a table, but it appears on the top of one page, but I want this table to be exactly where I enter the code. Also I want it to be at the center. How can I edit my code? my code:  

```
{r, echo=FALSE, message=FALSE} prediction_data <- data.frame( Party = c("Conservative", "Liberal"), Prediction = c(0.2625759, 0.2623956) )
```

Party	Prediction
Conservative	0.2625759
Liberal	0.2623956

ChatGPT: To control the placement of the table in your R Markdown document, you can modify your code:  

```
prediction_data <- data.frame( Party = c("Conservative", "Liberal"), Prediction = c(0.2625759, 0.2623956) )
```

`knitr::kable(prediction_data, caption = "My Table Caption") %>% kable_styling(latex_options = c("hold_position", "center"))` In this modified code:

User: Because the winner is always from either the Liberal Party or the Conservative Party, can I focus my research only on these two parties and not consider other parties? ChatGPT: Certainly, focusing your research on the Liberal Party and the Conservative Party is a reasonable approach, especially if these two parties are consistently the main contenders in Canadian elections, and you have limited resources or specific research objectives. Here are some considerations for this decision:

### 1. Dominance of Major Parties:

- If historical data indicates that the Liberal Party and the Conservative Party are the dominant forces in Canadian elections, it makes sense to focus on these parties. This is particularly relevant if other parties have not been significant contenders in terms of winning seats or forming the government.

### 2. Resource Allocation:

- Focusing on two major parties allows you to allocate your research resources more efficiently. You can delve deeper into the dynamics, policies, and factors specific to these parties, leading to a more in-depth and focused analysis.

### 3. Simplification of Analysis:

- Analyzing a smaller set of parties simplifies the research process. It can make it easier to model and interpret results, especially if the factors influencing voter behavior are well-understood for these two parties.

### 4. Relevance to the Research Question:

- Consider whether your research question specifically pertains to the competition and dynamics between the Liberal Party and the Conservative Party. If your goal is to understand the factors influencing the outcomes between these two parties, narrowing your focus is appropriate.

### 5. Election Dynamics:

- Understand the historical context and election dynamics. If other parties occasionally become major players or if their influence is increasing, you may need to reconsider excluding them from your analysis.

#### 6. Generalizability:

- Be aware that findings and conclusions drawn from a study focused on two parties might not be directly applicable to elections involving a broader range of parties. Consider the generalizability of your results and whether they align with the broader political landscape.

#### 7. Potential for Future Expansion:

- While focusing on two major parties initially, keep in mind that you can expand your analysis in the future if the political landscape evolves, or if your research questions change.

Always ensure that your research approach aligns with the goals of your study and the characteristics of the political system you are analyzing. If the Liberal Party and the Conservative Party are the primary focus of your research, it's reasonable to concentrate your efforts on understanding the dynamics between these two major political entities.

User: Can you help me check the grammar and maybe make these paragraph more logistic?

ChatGPT: Certainly! Please provide the paragraph or text you'd like me to check for grammar and possibly make more logically structured, and I'll do my best to assist you. User: Elections are a cornerstone of democratic systems, allowing citizens to exercise their right to choose their political representatives. Though the Canadian federal election of 2021 saw Justin Trudeau's re-election, his government failed to gain enough seats in the House of Commons. Indeed, the Liberal and the Conservative parties were very close regarding popular vote share in 2021 (CBC News, 2021). Furthermore, Trudeau's move to call an early federal election also stirred up different voices from the public and politicians. This backdrop sets the stage for a potentially competitive contest in the later 2025 federal election, with the Liberal and the Conservative as the primary contenders (Swigart et al., 2020).

To forecast the dynamics of the coming election, our study employed the data from the 2017 General Social Survey (GSS) and the data from the 2021 Canadian Election Study (CES) as predictive tools. The GSS census contains a comprehensive set of social, economic, and demographic variables, while the CES dataset offers data on voter's attitudes toward the current political parties. The synthesis of these datasets allows us to analyze voters' inclinations from different social backgrounds towards different political parties. Therefore, we can forecast the outcome of the upcoming federal election in 2025 and offer plausible insights to the political parties participating in the election. While focused on Canada, this study also offers practical insights for election analysis in federated nations such as the US, thereby enriching a broader discourse of potential electoral reforms and strategies.

The purpose of this study lies in its use of the GSS and CES datasets to predict the outcome of the next federal election in 2025 between the Liberal and Conservative parties (Kushner et al., 2009). Through the employment of logistic regression modelling for the Liberal Party and the Conservative Party respectively, we aim to investigate the voting patterns for each party. Additionally, to ensure the reliability of our prediction, we have also applied the post-stratification method to align our analysis with a statistically representative demographic. Later, by evaluating the results of the two models, we are able to infer the party that is most likely to secure a greater number of votes in the 2025 election.

For this study, we assume that each participant in the census or survey provided information that was reflective of their true beliefs.

ChatGPT: Here's a revised version of your text with some improvements in grammar and logical structure:

Elections serve as a cornerstone of democratic systems, allowing citizens to exercise their right to choose their political representatives. Despite Justin Trudeau's re-election in the Canadian federal election of 2021, his government failed to secure enough seats in the House of Commons. The Liberal and Conservative parties were closely matched in terms of popular vote share in 2021 (CBC News, 2021). Furthermore, Trudeau's

decision to call an early federal election sparked various opinions among the public and politicians. This context establishes a potentially competitive landscape for the upcoming 2025 federal election, with the Liberal and Conservative parties as the primary contenders (Swigart et al., 2020).

To forecast the dynamics of the forthcoming election, our study utilized data from the 2017 General Social Survey (GSS) and the 2021 Canadian Election Study (CES) as predictive tools. The GSS census encompasses a comprehensive set of social, economic, and demographic variables, while the CES dataset provides information on voters' attitudes toward current political parties. The integration of these datasets enables the analysis of voter inclinations across various social backgrounds, offering insights into the potential outcomes of the 2025 federal election and informing the participating political parties. While the focus is on Canada, this study also provides practical insights for election analysis in federated nations like the US, contributing to a broader discourse on potential electoral reforms and strategies.

The primary objective of this study is to utilize the GSS and CES datasets to predict the outcome of the 2025 federal election between the Liberal and Conservative parties (Kushner et al., 2009). Employing logistic regression modeling for each party, we aim to investigate voting patterns. Additionally, to enhance the reliability of our predictions, we applied the post-stratification method, aligning our analysis with a statistically representative demographic. By evaluating the results of the two models, we can infer which party is most likely to secure a greater number of votes in the 2025 election.

In conducting this study, we operate under the assumption that each participant in the census or survey provided information that accurately reflects their true beliefs.

User: I will describe the post-stratification process in my assignment, can you give me a form that I can use to do this?

ChatGPT Certainly! Below is a template you can use to describe the post-stratification process in your assignment. You can customize it based on the specific details and nuances of your research.

#### Post-Stratification Process Description

##### Introduction:

Post-stratification is a statistical technique employed to refine and improve the accuracy of survey estimates by adjusting for discrepancies between the sample and the population it represents. The process involves categorizing the population into strata based on certain characteristics and then adjusting the survey data to align with the known population proportions within each stratum.

##### Steps in the Post-Stratification Process:

**Define Strata:** Identify key demographic or socio-economic variables for stratification. Common strata include age groups, gender, income levels, education levels, or other relevant characteristics. **Obtain Census Data:** Acquire reliable census data that provides the distribution of the chosen strata in the target population. This serves as the benchmark for the known proportions of each stratum. **Compare Sample to Census:** Compare the demographic distribution of the sample obtained through the survey to the corresponding distribution in the census data. Identify any disparities or differences. **Calculate Weighting Factors:** Calculate weighting factors for each stratum by determining the ratio of the proportion of the stratum in the census data to the proportion in the survey sample.  $\text{Weighting Factor} = \frac{\text{Proportion in Census}}{\text{Proportion in Survey}}$  **Sample Weighting Factor =** Proportion in Survey Sample Proportion in Census

**Apply Weights to Data:** Multiply the obtained survey data for each respondent by the calculated weighting factor for their respective stratum. This adjusts the sample data to align with the known population proportions.  $\text{Weighted Data} = \text{Survey Data} \times \text{Weighting Factor}$  **Weighted Data =** Survey Data  $\times$  Weighting Factor **Reanalyze Weighted Data:** Perform subsequent analyses, including regression modeling or other statistical techniques, using the post-stratified and weighted data to generate more accurate estimates and predictions. **Considerations:**

Ensure that the chosen strata are relevant to the research question and align with the characteristics influencing the variables of interest. Validate the post-stratification process by comparing the weighted sample to external benchmarks, if available. Document any assumptions made during the process and discuss the potential impact of these assumptions on the validity of the results. **Conclusion:**

Post-stratification enhances the representativeness of survey data, allowing for more reliable and accurate inferences about the target population. By accounting for demographic variations, this technique improves the generalizability of survey findings and contributes to the robustness of the research outcomes.

Feel free to adapt and expand on this template to suit the specifics of your assignment and the details of your post-stratification process.

User: This is my mathematics model of this regression model, I was wondering if I need to write the variable province like this? Or I need to list all of them as variable.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{education} + \beta_2 X_{marital.status} + \beta_3 X_{province} + \beta_4 X_{sex}$$

ChatGPT In your logistic regression model, the variable “province” is treated as a single categorical variable with multiple levels. The notation you provided suggests that you are representing it as a single variable with distinct categories.

However, if “province” has more than two categories, it’s common to use a technique called “dummy coding” or “one-hot encoding” to represent categorical variables in logistic regression. This involves creating binary (0/1) indicator variables for each category except one, which serves as the reference category. This ensures that the model doesn’t assume an inherent order among the different provinces.

Since province are binary indicator variables representing each province except the reference province. are the corresponding coefficients for each province. Make sure to choose one province as the reference category, and the omitted province will be captured in the intercept term. This approach allows the model to account for the categorical nature of the “province” variable without assuming an ordinal relationship among the provinces.

If you are using a statistical software package like R, Python, or others, they often handle this encoding automatically when fitting logistic regression models with categorical variables.

User: After using the logistic regression model and post-stratification, we have the result that conservative party has higher proportion, what can be reasonable reason for this result? For example, now the winner is liberal party, so next time maybe conservative party.

ChatGPT When your logistic regression model and post-stratification process indicate a higher proportion for the Conservative Party, even if the current winner is the Liberal Party, there could be several reasonable reasons for this result. Here are some possibilities:

**Demographic Shifts:** Changes in the demographic composition of the population over time could lead to shifts in political preferences. If the Conservative Party has stronger support in demographic groups that are growing or becoming more influential, it could result in a higher predicted proportion. **Policy Changes:** Public opinion can be influenced by changes in government policies and political decisions. If the Conservative Party is perceived as addressing key issues more effectively or if there is dissatisfaction with the Liberal Party’s policies, voters may shift their support. **Economic Conditions:** Economic factors play a crucial role in voter preferences. If the Conservative Party is perceived as having better economic policies or if there are positive economic developments associated with the party, it could contribute to higher predicted support. **Campaign Strategies:** Effective campaign strategies, messaging, and outreach efforts by the Conservative Party may resonate with a larger segment of the population, leading to a higher predicted proportion of support. **Political Events:** Unforeseen political events or developments may influence voter preferences. If events occur that favor the Conservative Party or if there is a shift in public sentiment due to external factors, it could contribute to the model’s prediction. **Regional Dynamics:** Political support can vary regionally. If the Conservative Party gains stronger support in key regions, even if the overall winner is the Liberal Party, the model may capture these regional variations and predict a higher proportion for the Conservative Party. **Dynamic Nature of Politics:** Political landscapes are dynamic, and voter preferences can change over time. The model may be reflecting a snapshot of current trends, and future elections may see shifts in support based on changing circumstances. **Popularity Cycles:** Political parties may experience cycles of popularity. It’s possible that the Conservative Party is currently in a phase of increased support, and this is reflected in the model’s predictions. **Sampling Variability:** The model predictions are based on survey data, and there

can be sampling variability. If the survey captures a particular subset of the population that aligns with characteristics favoring the Conservative Party, it could influence the results. Methodological Considerations: Review the methodology of your logistic regression model and post-stratification process. Ensure that the variables included, model assumptions, and post-stratification adjustments align with the dynamics of the political landscape. It's essential to interpret the model results with caution, considering the complexity of political systems and the numerous factors that can influence election outcomes. Additionally, understanding the limitations of your model and the potential impact of assumptions will contribute to a more nuanced interpretation of the results.