Informative Title Name

STA304 - Assignment 2

GROUP NUMBER: ADD YOUR NAMES HERE

May 28, 2021

Introduction

The year 2021 is closing in on halfway and soon enough we will see 2022. That leaves 1 year 10 months and 16 days till the next Canadian federal election. The next federal election will decide the House of Commons to the 44th Canadian Parliament. The House of Commons comprises of multiple political parties, but the ones that take center stage are The Liberal Party, The Conservative Party, Bloc Québécois, The New Democratic and The Green Party. Through this paper we try to predict the overall popular vote of the 2023 Canadian federal elections.

The importance of forecasting Canadian federal elections has been relevant throughout history. The party that forms the federal government represents the voice, concerns and values of Canadians for the next 4 years. Prediction analysis allows political parties to understand, plan and develop strategies that popularize their values, beliefs and ideas of implementing and running the country. It provides an insight into voting patterns observed by province, age, income etc. and other demographics of Canadians who are eligible to vote. Predication analysis further gives an insight into ethical, social and economical changes and beliefs that align with the general Canadian population and thus the most popular party of their choice. It allows Canadians to exercise their right to vote and demand for change upon reflecting the decisions, policies and welfare acts of the past federal government, such as evaluating government response and strategies during the COVID-19 pandemic. This paper focuses on predicting the most popular party of choice, and in doing so, it determines the morals, codes and beliefs of the 27 million Canadians who are eligible to vote[1].

This paper predicts the overall popular vote of the 2023 federal elections using a Multilevel Logistic Regression Model with Post-stratification. The Multilevel Logistic Regression model is built from 'Canadian Election Study, 2019, Phone Survey' data (CES_2019 data) which analysis responses of Canadian to question related to their preferred political party and also gathers demographic information(such as age, sex, occupation etc.) of the respondents. Post-stratification is done using the General Social Survey(GSS) Data found at University of Toronto's CHASS data repository. Furthermore, this paper recognizes and justifies age, education, income level and province as meaningful predictors to the popular party of choice of eligible Canadians using the CES_2019 survey data which is post-stratified to the general voter population. Therefore, research question of this paper is focused on forecasting the overall popular vote for the next federal election based on the age, education, income of a voter in each province, for the six aforementioned political party's. We hypothesize that due to the popularity of The Liberal Party in Canada and the fact that they formed government in the 2015, 2019 federal elections [2], most provinces will be dominated by it, however some provinces may also see a neck-to-neck competition between the Conservative Party and the Liberal party which has seen an upward trend in its following, especially in the past two elections [2].

Data

Data Description

The CES_2019 survey data contains about 4021 rows and 278 variables. It is an annual survey, with data being collected between 2019-09-10 and 2019-11-21. It targets Canadian Citizens and Permanent Residents over 18 years compiling a rich set of data about Canadians' demographics, opinions and thought process on a wide variety of social, economic, and political issues.

The GSS dataset corresponds to the census dataset used in this research paper. It contains 20602 observations and 81 variables. The target population includes all non-institutionalized persons 15 years of age and older, currently residing in the 10 provinces of Canada. This survey was conducted from February 2nd to November 30th, 2017.[3] It uses telephone numbers registered with Statistics Canada's Address Register to primarily collect demographic information on the target population through telephonic means. It is important to note that this dataset does not include any information about an individuals political preferences, but rather, it focuses on collecting information related to the demographic attributes of individuals.

Data Cleaning

The Survey data and Census data, both needed to be cleaned to identify important variables and run the prediction analysis.

The statistical language `R version 4.0.2` was used to clean and analyze the data. Important packages include : *tidyverse, ggplot and dplyr*. Important functions include *select()*, which allows selection/deselection of variables in an r data frame, *mutate()* which allows to create new coloumns by manipulating existing columns, *filter()* which filters column values based on specified conditions and *rename()* which renames a coloumn/variables name. Since the paper implements post-stratification, the process of data cleaning included - selecting chosen variables from the 2 datasets and then matching variables within the two datasets similar variables have the same levels/categories/bins. To implement this, *case_when()* function was used to specify, change and match variables between the two datasets.

Important variables chosen from the survey data that are also present in the census data include

- Popular_party : The name of the respondents preferred party.
- Age : Age of the respondent
- Education : level of education of the respondent
- Income : Household/personal income of the respondent
- Province : Province of residence of the respondent.

Important variables chosen from the survey data were *popular party*, *age*, *education*, *income* and *province*. They correspond to *q35, q2, q61, q69, q4* of the original data respectively. Important variables from the census data were *age*, *education*, *income_family* and *province*. This was done using the *select()* function. Then due to the different naming conventions used in the datasets, *rename()* function was used to match the names between the data sets variables.Then, using data dictionaries of both survey data and the census data, numeric variables where changed to their string representation. For example, the *province* column in the survey data referred *1* as *Ontario*, the survey data was modified to replace it. This was done using *mutate()* and *case_when()* functions for both the data sets. Next, in order match different levels of variables, *mutate()* and *case_when()* were used. For example, to match age in the survey data and census data the following categories/bins were defined: *ages20to34*,*ages35to49*, *ages50to64*, *ages65to79*, *ages80to94*, *ages95to102*, so that every respondent in the survey data and census data exclusively fall into one of the defined categories. Similarly, bins were created for other variables to match factors between the survey and census data.

- Bins for age: ages20to34, ages35to49, ages50to64, ages65to79, ages80to94, ages95to102,

- Bins for education: Less than high school diploma or its equivalent, High school diploma or a high school equivalency certificate, College, CEGEP or other non-university certificate or di..., Trade certificate or diploma, University certificate or diploma below the bachelor's level, Bachelor's degree (e.g. B.A., B.Sc., LL.B.), University certificate, diploma or degree above the bach..., Dont Know, Refusal, Valid Skip.
- Bins for Income: Less than $25,000, $25,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $ 124,999, $125,000 and more, Dont Know, Refusal, Valid Skip.

Data Cleaning also includes dealing with missing data from the two datasets. The following snippet shows the missing values in the survey data followed by the missing values in the census data.

```
## popular_party           age      province     education income_family
##             0            57             0             0             0

##           age      province     education income_family
##           965             0           341             0
```

In the survey data, missing values represent 0.01% of the number of entries, therefore, they were replaced by the mode of the *age* column. In the census data, missing values of *age* and *education* represented an even lesser percentage of the number of entries, therefore it was considered safe to drop them.

After implementing the above mentioned procedures, a glimpse of the survey data and the census data are provided below.

```
head(survey_data)
```

```
##            popular_party        age province
## 1      The Liberal party ages50to64   Quebec
## 2 The Conservative party  age35to49   Quebec
## 3      The Liberal party ages20to34   Quebec
## 4      The Liberal party ages20to34   Quebec
## 5      The Liberal party  age35to49   Quebec
## 6 The Conservative party ages75to90   Quebec
##                                                    education
## 1                  Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
## 2 University certificate or diploma below the bachelor's level
## 3                  Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
## 4 University certificate or diploma below the bachelor's level
## 5  University certificate, diploma or degree above the bach...
## 6 High school diploma or a high school equivalency certificate
##           income_family
## 1 $100,000 to $ 124,999
## 2    $75,000 to $99,999
## 3     Less than $25,000
## 4 $100,000 to $ 124,999
## 5    $75,000 to $99,999
## 6    $25,000 to $49,999
```

```
head(census_data)
```

```
## # A tibble: 6 x 4
##   age       education                                 income_family      province
```

```
##   <fct>     <fct>                                      <fct>              <fct>
## 1 ages50to~ High school diploma or a high school eq~ $25,000 to $49,999 Quebec
## 2 ages50to~ Trade certificate or diploma             $75,000 to $99,999 Manitoba
## 3 ages50to~ Bachelor's degree (e.g. B.A., B.Sc., LL~ $75,000 to $99,999 Ontario
## 4 ages75to~ High school diploma or a high school eq~ $100,000 to $ 124~ Alberta
## 5 ages20to~ College, CEGEP or other non-university ~ $50,000 to $74,999 Quebec
## 6 ages50to~ High school diploma or a high school eq~ $50,000 to $74,999 Quebec
```

This paper primarily focuses on analyzing the popular vote of the 6 Political parties including The Liberal Party, Conservative Party, NDP, Bloc Quebecois, Green Party and The Peoples' Party. The survey data however, contains information about all the party's in a single column - *popular_party*. To proceed with our analysis, we created 6 different iterations of the survey data called *survey_data_liberal*, *survey_data_conservative*, *survey_data_ndp*, *survey_data_bloc*, *survey_data_green* and *survey_data_peoples*. Every iteration had the same information as the original survey data but with the *popular_party* column was modified. For example, the *survey_data_liberal* dataframe created a new variable *vote_liberal* that marked "1" if an individual thinks the liberal party is the most popular and thus has a chance to win, "0" if the individual is not sided with the liberal party or if they skipped/refused to reply to the respective survey question.

As an example to the above procedure, we show a glimpse of *survey_data_liberal*:

```
head(survey_data_liberal)
```

```
##          age province
## 1 ages50to64   Quebec
## 2  age35to49   Quebec
## 3 ages20to34   Quebec
## 4 ages20to34   Quebec
## 5  age35to49   Quebec
## 6 ages75to90   Quebec
##                                                  education
## 1                 Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
## 2 University certificate or diploma below the bachelor's level
## 3                 Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
## 4 University certificate or diploma below the bachelor's level
## 5  University certificate, diploma or degree above the bach...
## 6 High school diploma or a high school equivalency certificate
##           income_family vote_liberal
## 1 $100,000 to $ 124,999            1
## 2    $75,000 to $99,999            0
## 3     Less than $25,000            1
## 4 $100,000 to $ 124,999            1
## 5    $75,000 to $99,999            1
## 6    $25,000 to $49,999            0
```

Data analysis

To further understand the importance of our analysis, we present some summary statistics. These summary statistics dive into the reason behind using age, education and income in our paper.

Analysis of Age

```
## # A tibble: 6 x 7
##   age         liberal_votes Conservative_votes ndp_votes Bloc_votes Green_votes
##   <chr>               <dbl>              <dbl>     <dbl>      <dbl>       <dbl>
## 1 age35to49             577                335        34          0           5
## 2 ages20to34            339                219        36          1           7
## 3 ages50to64            609                396        21          2           1
## 4 ages65to74            353                216         4          1           4
## 5 ages75to90            213                133         3          1           2
## 6 ages90to102            13                  6         0          0           1
## # ... with 1 more variable: Peoples_votes <dbl>
```

The above chunk provides an insight into the importance of age vs the party of choice and is modeled from the survey data. It groups age and sums the votes given by each category of age to different partys'. The liberal party and the conservative party have the most votes in every age category compared to other parties.

Analysis of Education

```
## # A tibble: 9 x 7
##   education      liberal_votes Conservative_vo~ ndp_votes Bloc_votes Green_votes
##   <chr>                  <dbl>            <dbl>     <dbl>      <dbl>       <dbl>
## 1 Bachelor's de~           657              307        16          0           0
## 2 College, CEGE~           106               96         7          0           2
## 3 Don't know                 2                2         2          0           1
## 4 High school d~           312              283        28          3           8
## 5 Less than hig~            31               22         4          1           1
## 6 Refusal                    1                0         0          0           0
## 7 Trade certifi~           387              347        24          1           6
## 8 University ce~           178              115        13          0           2
## 9 University ce~           430              133         4          0           0
## # ... with 1 more variable: Peoples_votes <dbl>
```

The above chunk provides an insight into the importance of education vs the party of choice and is modeled from the survey data again. It groups education and sums the votes given by each category of education to different partys'. Again, the liberal party and the conservative party have the most votes in every education category compared to other parties.

Analysis of Income

```
## # A tibble: 8 x 7
##   income_family  liberal_votes Conservative_vo~ ndp_votes Bloc_votes Green_votes
##   <chr>                  <dbl>            <dbl>     <dbl>      <dbl>       <dbl>
## 1 $100,000 to $~           222              126         9          0           0
## 2 $125,000 and ~           511              300        17          0           1
## 3 $25,000 to $4~           235              153        11          0           6
```

```
## 4 $50,000 to $7~           285              203         9          0          2
## 5 $75,000 to $9~           231              132         9          0          1
## 6 Don't know               173              109        19          1          5
## 7 Less than $25~           180              101        17          4          5
## 8 Refusal                  267              181         7          0          0
## # ... with 1 more variable: Peoples_votes <dbl>
```

The above chunk provides an insight into the importance of income vs the party of choice and is modeled from the survey data again. It groups income and sums the votes given by each category of the income variable to different partys'. Yet again, the liberal party and the conservative party have the most votes in every category of the income variable compared to other parties.

Methods

Model Specifics

The goal of this research paper is to predict the popular vote of the upcoming 2023 federal elections by province using a multilevel logistic regression model with post-stratification.

The independent variables *age*, *income*, *education* and *province* were selected to create multilevel logistic regression model since a paper posted by Statistics Canada [5] suggests that these variable affect voter turn out rates as well as reflect a persons beliefs, ethics and morals which conclusively reflects their political preferences and expectation towards a party.

To the create the multilevel logistic regression model we will be using age, income, education and province as independent variables which will predict the popular vote. The independent variables province will be used as a group level variable that will model the intercepts of models with respect to different provinces.

The data used for this analysis is the cleaned survey data which is further modified into *survey_data_liberal*, *survey_data_conservative*, *survey_data_ndp*, *survey_data_bloc*, *survey_data_green* and *survey_data_liberals*. This means we have 6 different models that use the same independent and dependent variables but are made from different datasets. To be precise, we use :

- A model to predict whether an individual votes for the liberal party based on their age, education, income and province made using *survey_data_liberal*, with dependent variable *vote_liberal*.
- A model which predicts whether an individual votes for the conservative party based on the same variables but made using *survey_data_conservative*, with dependent variable *vote_conservative*.
- A model which predicts whether an individual votes for the NDP party based on the same variables but made using *survey_data_ndp*, with dependent variable *vote_ndp*
- A model which predicts whether an individual votes for the Bloc Quebecois party based on the same variables but made using *survey_data_bloc*, with dependent variable *vote_bloc*.
- A model which predicts whether an individual votes for the Green party based on the same variables but made using *survey_data_green*, with dependent variable *vote_green*.
- A model which predicts whether an individual votes for the Peoples party based on the same variables but is made using *survey_data_peoples*, with dependent variable *vote_peoples*.

Since each model uses the same independent variables, the general form of multilevel logistic regression model is given by:

$y = \beta_1 x_{age} + \beta_2 x_{education} + \beta_3 x_{income} + \beta_4 x_{1|province} + \epsilon$

Where,

- y represents the log odds of one of the dependent variable : *vote_liberal*, *vote_conservative*, *vote_ndp*, *vote_bloc*, *vote_green*, *vote_peoples*.

- β_0 represents the intercept of the model
- β_1 represents the coefficient for age
- β_2 represents the coefficient for education
- β_3 represents the coefficient for income
- β_4 represents the coefficient for group variable province
- ϵ represents the error in the model

Post-Stratification

In order to estimate the popular vote of each province, we use the statistical technique of post stratification. Under this technique, the census dataset is split into mutually exclusive bins based on demographics of the population. Then we estimate a response or quantity of interest for each cell. This is followed by aggregating the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population or the census data. Post-stratification is particularly useful since it allows us to adjust non-representative samples to better analyze opinions and other survey responses.

The following formula is used for post-stratification: $\hat{y}^{PS} = \frac{\sum N_j \cdot \hat{y_j}}{\sum N_j}$

- $\hat{y}^{PS}$ is the estimate in each cell after weighting it by its relative portion.
- $\sum N_j$ is the population size of the j^{th} cell based off demographics.

To fulfill the purpose of this paper, we use age, education, income and province as demographic variables. We create bins splits using different levels of age, education, income and province in the census data. This means each cell is mutually exclusive and only includes those individuals that belong to a single category of age, a single category of education, a single category income and a province (Categories of each of these variables is described in the data section). Next, we estimate N_j by summing the number of individuals in each cell/bin. To estimate the popular vote in each province, we estimate the proportion of those groups of cells/bins that belong a particular province. This is done by dividing each N_j by the sum of all $N_j's$ that belong to a particular province. Finally, each of the 6 different models is applied to the post-stratification data to estimate the 6 different $\hat{y}^i$, which calculates the log odds of the probability of voting for a particular party (depending on the model used). Finally, $\hat{y}^{PS}$ is estimated by taking the product of the $\hat{y}^i$ and the estimated proportion of each cell grouped by province for each model.

Since the census data and the survey data do not exactly match in levels they use to measure province of residence, there are certain assumptions we had to make during the development of our model. Yukon, NW territories, Nunavut which are measure in the census data have been market as "Unkown" in the survey data and therefore, we consider only the left over 10 Canadian provinces - (Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, British Columbia.

<To put math/LaTeX inline just use one set of dollar signs. Example: $\hat{y}^{PS}$ >

$$include.your.mathematical.model.here.if.you.have.some.math.to.show$$

All analysis for this report was programmed using `R version 4.0.2`.

Results

<Here you present your results. You may want to put them into a well formatted table. Be sure that there is some text describing the results.>

<Note: Alternatively you can use the `knitr::kable` function to create a well formatted table from your code. See here: https://rmarkdown.rstudio.com/lesson-7.html.>

<Remember you can use `r` to use inline R code.>

<Include an explanation/interpretation of the visualizations. Make sure to comment on the appropriateness of the assumptions/results.>

Conclusions

<Here you should give a summary of the Hypotheses, Methods and Results>

<Highlight Key Results.>

<Talk about big picture.>

<Comment on any Weaknesses.>

<End with a concluding paragraph to wrap up the report.>

Bibliography

1. Mapleleafweb.com. 2021. Voter Turnout in Canada | Mapleleafweb.com. [online] Available at: http://www.mapleleafweb.com/features/voter-turnout-canada [Accessed 28 May 2021].

2. En.wikipedia.org. 2021. 2019 Canadian federal election - Wikipedia. [online] Available at: https://en.wikipedia.org/wiki/2019_Canadian_federal_election [Accessed 28 May 2021].

3. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

4. 2019 Canadian federal election - [online] Available at: https://en.wikipedia.org/wiki/2019_Canadian_federal_election [Accessed 28 May 2021].