Informative Title Name

STA304 - Assignment 3

GROUP NUMBER: ADD YOUR NAMES HERE

November 5, 2021

Introduction

The Canadian Election Study is an annual survey of voting and other preferences and demographics which are thought to pertain to political behavior of Canadian voters.This survey is used to predict the overall popular vote of the next Canadian federal election (tentatively 2025) using a regression model with post-stratification.

Some of the important terminologies is the Majority government, which is the party that won the most number of votes and were elected to become the ruling government. Next, is the Minority party, it's the party with the least votes and they are elected to become the opposition to the majority party. Also, some of the parties mentioned below are Bloc Quebecois, Conservatives, Greens, Liberal, NDP, and People's Party.

The research question we would be investigating is that does age, religion and location enough to predict the number of votes for each party?

In this paper we will use post-stratification and multilevel regression to investigate if certain features of the population population such as: age, religion and location can be used to predict how many votes each party gets in a federal election

Data

Data Collection Process

The Data is a collection of two datasets. The first dataset contains data from the General Social Survey on Family (cycle 31) on 2017. Canada's General Social Survey (GSS) program conducts annual survey covering one topic in depth (citation). As such, this dataset contains mostly contains information pertaining to families. However, we will later investigate some common variables between this dataset and the Canada Election Study (CES) dataset. This will give us a list of factors in the general population that could be potentially associated to political affiliations

the Canada Election Study (CES) data which was collected in 2019. This data was collected from a questionnaire delivered to the people living in Canada through the Computer Assisted Telephone Interviewing(CATI). Phone calls were made potential interviewees during both the day and evening for every weekday. These questions asked some personal information such as their age and also political opinions such as "what is your opinion on Justin Trudeau?" (citation)

Data cleaning

The regression model will be constructed, will be applied to both datasets, therefore only variables that appeared in both the datasets could be used. One way was examine if there were common terms present in the columns for both the dataset. Unfortunately most the columns the CES dataset were just question numbers. However, each column also had a label that stated the question itself. For example column "q2" had the label "IN what year were you born?". Hence we collected common words that appeared in the columns of the GSS data and the labels of the CES data to find possible topics that were common in both datasets.

Then for each topic we search the columns for both the dataset to see if any two column were describing indentical or similar variables. These are the variables we found:

- Age: The age of the person was recorded in both data with the same column names.
- Sex: Gender was recorded in CES data and sex was recorded in the GSS data. Here we assumed the people who reported their gender to be male or female also would have the same sex as their gender. Hence we removed all genders that were not male or female and renamed the q3 column in the CES data to sex.
- Province: Both the datasets recorded the province a person lived in.
- religious importance: Both datasets recorded how the level of importance for religion. The categorical values in both the datasets were the same with the exception of an additional value "Refused" being present in the CES data. We removed rows containing this value from the CES data.
- Aboriginal: Both datasets recorded if a person was aboriginal or not. This was an option for question 66a in CES data which asked which ethnicity the person belonged to. q66a_15 recorded if the person belonged to an Aboriginal group. the value "(1) Selected" indicated the person identified as original and "(0) Not Selected". We filtered out other values as the indicated that the question was skipped or if the person was not sure. Then we mutated "(1) Selected" to "Yes" and "(0) Not Selected" to "No" to match the values of the variable in both dataset. Both datasets had values "Don't know" to indicate that a person was not sure. However, there very few rows with this value so we removed these rows.
- Education Level: Both datasets recorded the highest level of education completed by a person. However, these categories had slightly different names. For example, "University certificate, diploma or degree above the bach.." indicated that a person has a qualification above a bachelor's degree. In the CES data there are Master's degree as well as other degree above bachelor's degree. Hence we put the education level in both datasets into three groups: "Above Bachelor", "Below Bachelor" and "Bachelor".
- Household size: Both the datasets recorded household size.

After finding the common variables, we removed rows with missing values in these columns for both the datasets. Then we renamed the columns of the CES data to match the names of the columns of the GSS data. This was done so that models constructed on the CES data could also be used on the GSS data for post-stratification. Then we took the common variables and constructed two datasets. One dataset containing the rows from the CES data and the other dataset containing the rows of the GSS data. Futhermore, in the first dataset, voting data was added to the CES variables. This is the outcome of interest

Table 1: Variable description table

Variable Name	Description
sex	Sex of the person
province	Province that a person is currently residing in
education	Level of education divided into three categories: 'Above Bachelor', 'Below Bachelor' and 'Bachelor'
religion_importance	How important is their religion to a person: 'Not important at all', 'Somewhat important', 'Very important', 'Not very important' 'Don't know'
aboriginal	Is the person aboriginal
hh_size	How many people does a person share their living space with including themselves
age	Age of the person in years
vote	Party that the person will/might/has vote(d) for

Numerical Summaries

Table 2: Frequency for each category in the CES data

		Count
Sex	(1) Male	1057
	(2) Female	895
Province	(1) Newfoundland and Labrador	114
	(2) Prince Edward Island	102
	(3) Nova Scotia	106
	(4) New Brunswick	102
	(5) Quebec	367
	(6) Ontario	436
	(7) Manitoba	142
	(8) Saskatchewan	135
	(9) Alberta	152
	(10) British Columbia	296
Education	Above Bachelor	304
	Bachelor	534
	Below Bachelor	1114
Religion Importance	Don't know	4
	Not important at all	184
	Not very important	355
	Somewhat important	742
	Very important	667
Aboriginal	No	1889
	Yes	63
vote	Bloc Quebecois	74
	Conservatives	764
	Greens	163
	Liberal	673
	NDP	245
	People's Party	33

From the above table we can gain the following insights on the survey data:

- There seems to slightly more males and females in this dataset.
- Most of the people in the survey have an education level below a bachelor's degree.
- Most of the people in give a lot of importance or at least some importance to their religion.
- The large majority of the people in the survey are not Aboriginal. This perhaps suggests that this variable might not be of use for the model.
- It seems that the most popular party was the Conservative party followed closely by the Liberal party in 2019
- There does not seem to be an significant difference in the number of answered the survey for each province. This indicates that this survey represents the data of each province fairly well.

Table 3: Frequency for each category in the GSS data

		Count
Sex	Female	8961
	Male	7391
Province	Alberta	1352
	British Columbia	1807
	Manitoba	974
	New Brunswick	1225
	Newfoundland and Labrador	1030
	Nova Scotia	1283
	Ontario	3859
	Prince Edward Island	645
	Quebec	3193
	Saskatchewan	984
Education	Above Bachelor	1249
	Bachelor	2787
	Below Bachelor	12316
Religion Importance	Don't know	167
	Not at all important	3335
	Not very important	2708
	Somewhat important	5003
	Very important	5139
Aboriginal	No	15572
	Yes	780

In the census data the differences observed in the categorical variables are similar to the differences observed in the survey data, with the exception of Sex. There are more females than males in the census data. This indicates that there is an underrepresentation of females in the data survey data.

Table 4: Comparing Age variable in both dataset

	mean age	minmum age	maximum age	standard deviation in years
From GSS dataset	52.52854	15	80	17.70734
From CES Dataset	53.71721	18	100	16.60615

Even though the mean age is similar in both datasets, the magnitude of the min, max and standard deviation values indicates that age has a wider distribution in GSS dataset.

Table 5: Comparing Age variable in both dataset

	25th qauntile household size	median household size	75th quantile household size	min household size	max household size
From GSS dataset	1	2	3	1	6
From CES Dataset	2	2	4	1	15

From this table, we can see that household size thas similar distributions, however CES data seems to indicate that people in this dataset tend to have more household members. The max size 15 could be a potential outlier

Graphical Summaries

Methods

<Include some text introducing the methodology, maybe restating the problem/goal of this analysis.>

Model Specifics

<I will (incorrectly) be using a linear regression model to model the proportion of voters who will vote for Donald Trump. This is a naive model. I will only be using age, which is recorded as a numeric variable, to model the probability of voting for Donald Trump. The simple linear regression model I am using is:>

$$y = \beta_0 + \beta_1 x_{age} + \epsilon$$

<Where y represents the .... β_0 represents....>

Post-Stratification

<In order to estimate the proportion of voters.....>

<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. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)