Analysis of 2019 Canada General Election using Logistic Regression and Post-Stratification to identify the difference if 'everyone' voted

YUEHAO HUANG

07/12/2020

Abstract

In this paper I am going to build a logistic regression model based on the CES and a post-stratification dataset to identify how the 2019 Canadian Federal Election would have been different if 'everyone' had voted. The data are gathered from CESR2019 package. My hypothesis is that the if everyone voted, Justin Trudaue wont be able to continue be the PM of Canada with even more support from people. I used Post-stratification method based on provinces and then used logistic regression to create the model. The model predicted that the both liberals and conservative will gain more votes but liberals will have a slight lead in total hence confirming my hypothesis. Even if everyone in Canada voted, Justin Trudeau will still serve his next 4 years as PM of Canada, but Tories will lose the popularity from the Canadians as Trudeau will have 0.12% more popular votes than Tories.

Note

Code and data can be accessed via github at https://github.com/EroSkulled/STA304/tree/main/FINAL, licensed under MIT.

Keywords

- 2019 Canada General Election Study
- 2019 Canadian Election Result
- Voters
- Post-Stratification
- Observational study
- Liberal Party
- Conservation Party
- All citizens voted
- Difference in results

Background

The 2019 Canadian election ended with the liberals continue to rule the country even with the conservatives got more popularity. In this study we will build a logistic regression model based on the CES data and a post-stratification dataset from Government of Canada to identify how the 2019 Canadian Federal Election would have been different if 'everyone' had voted.

Introduction

The 2019 Canadian election was a close one for our Prime Minister Justin Trudeau to continue serve this country. From the CES2019 data set, we we could knock out an interesting question: what if everyone in Canada who is eligible to vote, voted? Will there be any differences? Will conservative still be the popular party that people would vote for? It is worthy to note that Throughout Canada's long history, the CES has been a rich source of observational data on Canadians' political behavior and attitudes,as observational data is often more reliable.

In this study, the outcome(dependent) variable is whether people will vote for the liberal party or conservative party. The treatment(Independent) variable is the gender of people. The province they currently live in will be used as strata for us to do post stratification process on the data.

A stratified dataset will be used to investigate how different people tend to vote differently based on different provinces they are from and which gender they are to make conclusion on the impact to the result of the election. In the Methodology section, I describe the study data and model that was used to perform the logistic regression analysis.Results of the MRP model analysis are provided in the Results section, and inferences of this data along with conclusions are presented in the conclusion section. Further development including but not limit to the improvements and Next steps are written in the second part of the dissuasion section.

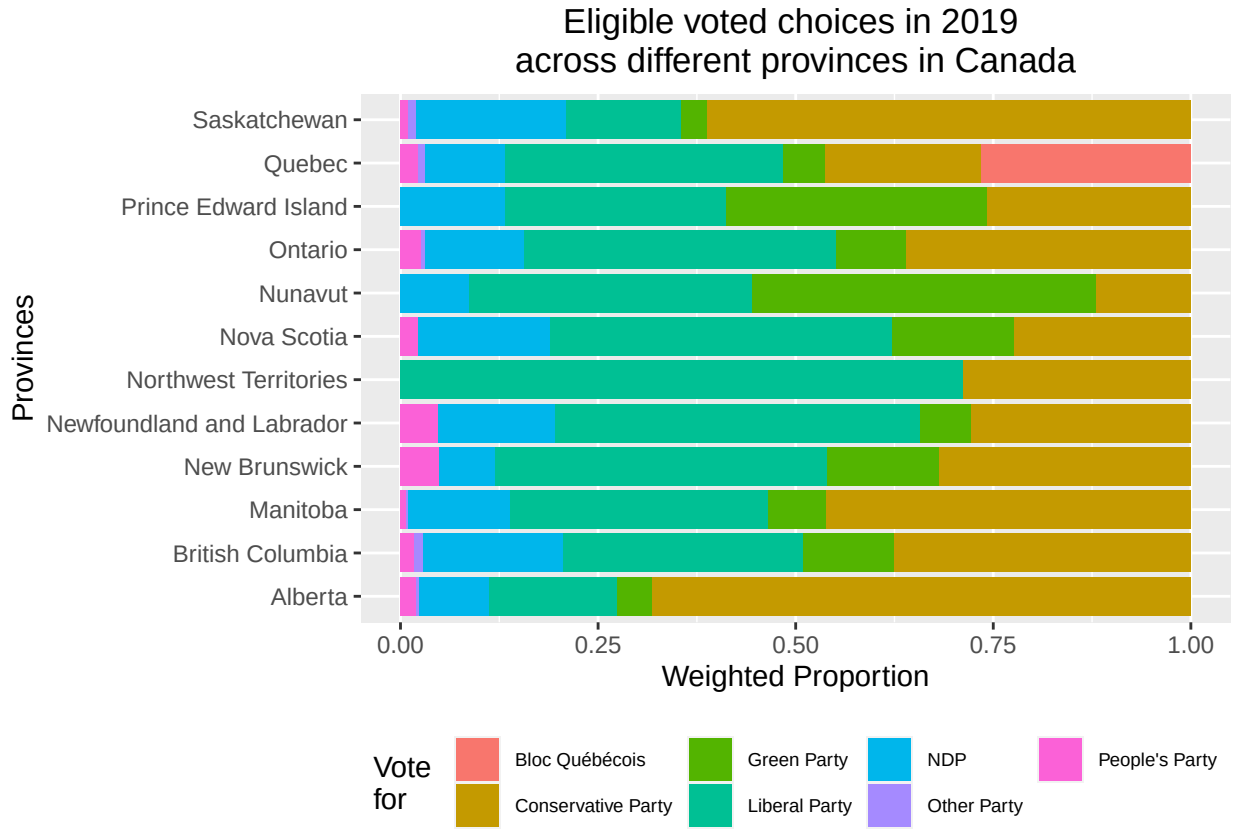
Methodology

Data

The data for creating the model is from 2019 Canadian election study - web survey. See reference section for authors and links to the data. This dataset used a weight system to to ensure that the data is representative of the population. This data consists of two sections: one before the election, named CPS, and one after, named PES. Among them there are 33905 high quality entries in CPS and 8313 high quality entries in PES. We will only use those entries for high accuracy results. Note above named entries have excluded those whose province and postal code did not match or having other issues. For the post-stratification dataset, Government of Canada post newest census data on Canada.ca and accessible via https://www.statcan.gc.ca/

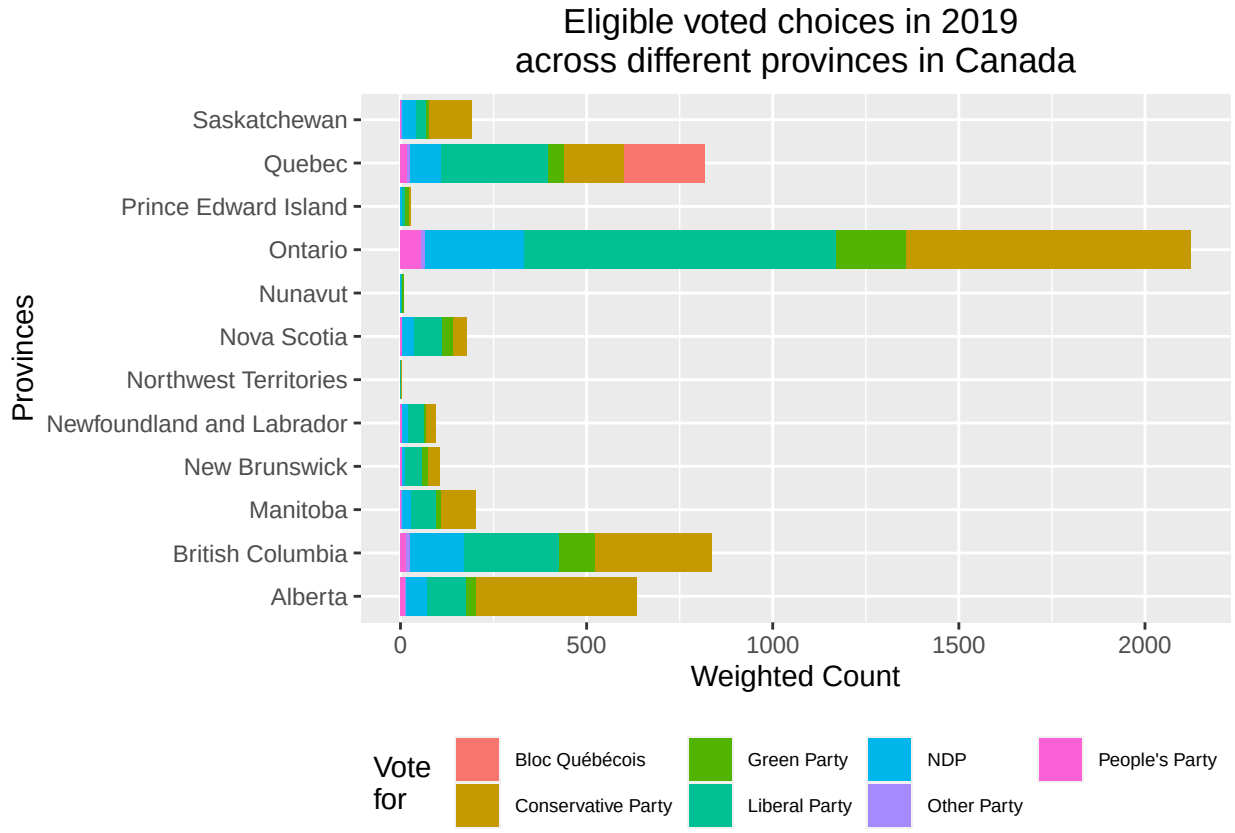
Note Canada is not a direct democracy system[1], due to the first-past-the-post electoral system, the number of seats won by a party determines who wins — not the number of overall votes. The vote here only indicates the popularity a party have, or popular vote.

[1]See how democracy in Canada works posted on Canada.ca: https://www.canada.ca/en/democratic-institutions/services/democracy-canada.html.



Plot.1

Plot.1 Above is a proportional plot showing the make up percentage for each party people would vote for in each province. This indicates how the votes are distributed in each province. It is clear that conservatives and Liberals are two main rivals on the stage and has significant differences based on provinces.



Plot.2

Plot.2 is a weighted vote count plot. This indicates where most of the votes are from in this survey. It is clear that from province to province there is a significant difference on the number of votes as well. However, the post-stratification method will be able to solve those bias problems later.

Note the data in both plots are already weighted[2]. Both plots have the same number of results of 5483 with no empty response or response with "Do not want to disclose" option. While Liberal Party and Conservative party are competing each other for the seats, Liberals has some advantage all over conservative party except for Alberta province where conservative party has a significant lead.

We are interested in predicting the popular vote outcome of the 2019 Canadian Federal Election using Population Census data from Canada.ca (For citation please see reference). To do this we are employing a post-stratification technique. In the following section(s) I will describe the data and model specifics as well as the post-stratification calculation.

For this data I created new vote_Liberal and vote_Conservative variables(cleaned from variable cps19_votechoice), containing only 1 and 0 indicating whether people will vote for the liberals or not, to be our response variable. Then use gender(cleaned from variable cps19_province) as the treatment variable. At last, we will perform post-stratification on the census data set based on province to better predict the result.

Some preview of the data structure is listed in the following table:

[2]This dataset used a weight system to to ensure that the data is representative of the population.

province	gender	income	vote_Liberal	vote_Conservative
British Columbia	Males	75000	0	1
Ontario	Females	200000	1	0
Ontario	Males	210000	0	0
Ontario	Females	50000	1	0
Alberta	Males	150000	0	1
Ontario	Males	75000	0	1

Table.1

Model

We will be using a logistic regression model to model the proportion of Canadians who will vote for the Liberal Party in 2019 The reason we use this model is because we are interested in whether The liberals will be elected to be the ruling party by ALL Canadians in 2019, by cleaning the data and adding a variable with response only containing 1 and 0, indicating whether people will vote for Liberals Party or not, it is perfect to use logistic regression with binomial family for our prediction. Our model has one predictor. It is: gender[3]. which is recorded as a categorical variable, to model the probability of people voting for Liberal Party or not. The logistic regression model we are using is:

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

Where $\beta_0 = -0.47668$, $\beta_1 = -0.21337$, and ϵ is the error term.

y represents the log proportion value of people who will vote for Liberals vs. against them β_0 represents the intercept of the model, meaning the probability of a female would have a log value of -0.47668 for in the log proportion possibility value of voting for the Liberals against voting for some other party. Additionally, β_1,represents the slope of the model for male. For example, for any male, we would expect a 0.5382 increase in the log proportional probability value. Finally, β_2 is the slope for unemployed people. For instance, Male would have a -0.21337 change in numbers in the log proportional probability (Please see next section(s) for details).

We also have a second model for predicting people will vote for conservatives or not. It has the exact same response and treatment variables but different values, model is listed as following equation:

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

Where $\beta_0 = -0.84472$, $\beta_1 = -0.51144$, and ϵ is the error term.

Results

The model I used is listed in the following table:

Table.2

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4766800	0.0404829	-11.774835	0.0000000
genderMales	-0.2133664	0.0564664	-3.778642	0.0001577

As we can see we have a very small p-value meaning gender is very significant to our model.

[3]Note here gender relates to a physical condition of human to determine one's sex. We only use male and female here because on the official Canadian CENSUS, there is only two gender, Male and Female. For more information on gender please see the following paper from stanford: https://web.stanford.edu/~eckert/PDF/Chap1.pdf

Note before this model I also included income as an variable but the model summary indicates income variable is not significant(see Table.3) so i removed it and only have gender as the only variable in this logistic regression model.

Table.3

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4930491	0.0457545	-10.7759801	0.0000000
genderMales	-0.2157798	0.0565617	-3.8149455	0.0001362
income	0.0000002	0.0000002	0.7669593	0.4431057

This also apply to the model created for the conservative part as shown in Table.4 below

Table.4

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4766800	0.0404829	-11.774835	0.0000000
genderMales	-0.2133664	0.0564664	-3.778642	0.0001577

As we can see we have a very small p-value meaning gender is very significant to our model.

Next I will explain the meaning of result I got after applying post-sttratification.

Table.5

lib_predict
-0.5827043

con_predict
-0.5905789

By using the logistic regression model here, meaning the model will give us a log ratio of $log(p/1-p)$ on what party people would vote for and against them in this 2019 election instead of the percentage of people voting for Liberal or Conservative Party than others. However, some simple calculation would be more than enough to get this value.

Using the result from the Liberal party as an example, we estimate that the log proportion of voters in favor of voting for the Liberal Party and against them to be -0.5827 after applying post stratification as shown in Table.5. Then we take exponential on both sides since it is a log proportion, what we have is 0.5583 for the proportion of vote for vs. against the liberals. For example, Say there are 156 people, about 56 people would vote for the liberals and 100 would not. Hence we can interpret the percentage of people who would vote for the liberals based on the post stratification data will be 56/156 = 35.8%, meaning about 35.8% of people would vote for the Liberal Party, which accounted for genders and using provinces as strata, modeled by the Logistic regression model. We apply the same calculation to the prediction of conservative party. The model predicted a -0.5905789 log value of the proportion as shown in Table.5. Hence we get 59/159=35.65%, meaning about 35.65% of people will vote for the Conservative Party.

#Discussion ## Summary In order to predict the outcome of the 2019 Canadian Federal Election, the logistic regression model is used to predict whether the liberals will be able to retain their position. So our response variable will be whether people will vote for the liberal Party, using dummy variable 1 as vote for the liberals and 0 otherwise. For treatment variables, only gender is used and as shown in Table.1, Gender is a very significant variable in our model. After the regression model is built, we use post-stratification to partition the data into different cells based on provinces. We may obtain our expected outcome for each province. Weight each province cell by its relative proportion to population and sum them up to get the estimate for the entire data.

Conclusion

We estiamted 35.8% for people would like to vote for liberal party and 35.65% people would tend to vote for conservative party based on the entire population of Canada after applying the same calculation above if all

Canadians voted. Note historically there are 34.34% of people voted for the conservative party and 33.12% voted for the liberals in 2019. Our model predict that there will be about 2.7% more people choose to vote for the liberals during the past 2019 election if we have all Canadians voted and 1.31% more would choose conservatives party. Hence based on our post-stratification analysis of the proportion of voters in favor of the liberals party and the conservative party modeled by a logistic regression model,we can conclude that the liberals will overwhelm the conservatives in pupolar vote by 0.15% and would still be the ruling party of Canada.

Weakness & Nextstep

One of the weakness to note is that the census data is based on 2020 Canada Census Website and I was not able to separate the census data who is eligible to vote and those who is not[4] as you must be 18 on the election day to be able to legally vote[5].

Next, if possible, to do it in more small, doing post-stratification based on each small towns or cities would give us more precis result than what this model predicted the results based on Province. However, note the difficulty getting those data from the small towns and cities.

References

Canadian Census Data from statistic website of the Government of Canada https://www.statcan.gc.ca/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Online Survey", https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1

Wu C., Thompson M.E. (2020) Basic Concepts in Survey Sampling. In: Sampling Theory and Practice. ICSA Book Series in Statistics. Springer, Cham. https://doi.org/10.1007/978-3-030-44246-0_1

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis (2020). skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra

[4]This does not include Permanent Resident, only citizenship is included.

[5]Facts about voter registration, citizenship and voter ID https://www.elections.ca/content.aspx?section=med&dir=c76/citizen&document=index&lang=e