STA304-Final

Shlok Somani

12/21/2020

Github Repo :

Abstract

In an ideal democracy, every voter has a voice that is heard. What happens when the essence of this system is removed? What happens when it is put back? These are the questions this paper aims to address through this analysis and statistics-based thought experiment. Assuming a 100% voter participation rate in the 2019 Canadian Federal Election, logistic regression models are applied along with factors such as age, gender, province and education to predict the voter choice between the liberal and conservative party.

To further this application, we leverage the data gathered within the 2016 General Social Survey on Families to post-stratify and predict the popular vote results. Drastic differences in results go on to support the initial predictions as presented in the report.

Keywords

Multilevel Regression, Post-Stratification, 2019 Canadian Federal Election, Survey, Census

Introduction

Voting, the lifeblood of democracy, to this day remains an issue in Canadian politics. Despite several efforts such as social media marketing and the 'I voted' movement, voter turnout remained at a 77% in the 2019 Canadian federal election(1). It has been proven, time and time again, that the demographic that opts out of voting, in fact, the most influential group as they represent the neutral group of the political spectrum. In order to understand this phenomena further, this paper will utilize MRP as a statistical technique to transform national opinion survey results into local estimates to examine how the result could have been different if all the eligible voters have voted.

Methodology

Data

Survey Data

(Stephenson et al.,2020)

Table 1: Liberal Data

	0 (N=3816)	1 (N=2163)	Total (N=5979)	p value
province				< 0.001
Alberta	602 (15.8%)	141 (6.5%)	743 (12.4%)	
British Columbia	557 (14.6%)	266 (12.3%)	823 (13.8%)	
Manitoba	190 (5.0%)	94 (4.3%)	284 (4.7%)	
New Brunswick	74 (1.9%)	63 (2.9%)	137 (2.3%)	
Newfoundland and Labrador	66 (1.7%)	53 (2.5%)	119 (2.0%)	
Northern Canada	5 (0.1%)	6 (0.3%)	11 (0.2%)	
Nova Scotia	96 (2.5%)	96 (4.4%)	192 (3.2%)	
Ontario	1374 (36.0%)	1017 (47.0%)	2391 (40.0%)	
Prince Edward Island	12 (0.3%)	13 (0.6%)	25 (0.4%)	
Quebec	628 (16.5%)	375 (17.3%)	1003 (16.8%)	
Saskatchewan	212 (5.6%)	39 (1.8%)	251 (4.2%)	

Table 2: Conservative Data

	0 (N=4007)	1 (N=1972)	Total (N=5979)	p value
province				< 0.001
Alberta	292 (7.3%)	451 (22.9%)	743 (12.4%)	
British Columbia	553 (13.8%)	270 (13.7%)	823 (13.8%)	
Manitoba	175 (4.4%)	109 (5.5%)	284 (4.7%)	
New Brunswick	98 (2.4%)	39 (2.0%)	137 (2.3%)	
Newfoundland and Labrador	83 (2.1%)	36 (1.8%)	119 (2.0%)	
Northern Canada	10 (0.2%)	1 (0.1%)	11 (0.2%)	
Nova Scotia	160 (4.0%)	32 (1.6%)	192 (3.2%)	
Ontario	1635 (40.8%)	756 (38.3%)	2391 (40.0%)	
Prince Edward Island	22 (0.5%)	3 (0.2%)	25 (0.4%)	
Quebec	868 (21.7%)	135 (6.8%)	1003 (16.8%)	
Saskatchewan	111 (2.8%)	140 (7.1%)	251 (4.2%)	

Census Data

The data of the 2016 GSS (General Social Survey) was collected through CATI1 (Computer Assisted Telephone Interviews). Respondents were interviewed in either English or French. Phone calls were made from 9:00 a.m. to 9:30 p.m. Mondays to Fridays, 10:00 a.m. to 5:00 p.m. on Saturday and 1:00 p.m. to 9:00 p.m. on Sunday. Each record in the survey frame was assigned to a stratum within its province. A simple random sample without replacement of records was next selected in each stratum. A minimum sample size for each stratum is needed to ensure an acceptable estimation for every stratum. Once that minimum is reached, the rest would be distributed to the strata that would balance the precision of both nation-level and stratum-level estimates. The amount of people who responded to this survey, which is the sample population, is 19,609. However, it is important to note that the response rate of this survey was 50.8%.

Modle

By employing techniques such as post-stratification, we create a model that predicts the popular vote outcome if the voter turnout was 100% in the 2019 Canadian federal election. In the following subsections, the model specifics and the post-stratification calculations are explained in further detail.

Model-specific

Using two logistic regression models, the proportion of voters who choose to vote for the Liberal and Conservative party respectively will be identified. Since logistic regression is a statistical model that uses logistic function to model a binary response variable, it is an appropriate prediction tool to determine whether the voter would vote liberal or conservative. Furthermore, age, gender, province and education will act as predictor variables due to their high level of influence on voter behavior. Below are the two models one for the liberal and the other is for conservative.

$$log(\frac{\hat{p}_{Liberal}}{1-\hat{p}_{Liberal}}) = \alpha_0 + \alpha_1 x_{age} + \alpha_2 x_{sex} + \alpha_3 x_{education} + \alpha_4 x_{province} \quad (1)$$

$$log(\frac{\hat{p}_{Conservative}}{1-\hat{p}_{Conservative}}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{education} + \beta_4 x_{province} \quad (2)$$

Post-Stratification

In order to estimate the popular vote results for the Liberal and Conservative Party given a 100% voter turnout, we need to perform a post-stratification analysis. A post-stratification analysis finds the sum of the same models with different predictions, multiplies the population-related with said predictor, divides that by the sum of the population related to each predictor. This process is then completed by creating cells based on each explanatory variable in the model - age groups, education, gender and province. Using the model described in the previous subsection, these analyses allow us to estimate the proportion of voters in each predictor bin. Continuing on this journey, we will then weight each proportion estimate (within each bin) by the respective population size of that bin and sum those values and divide that by the entire population size.

$$\hat{y}^{PS}_{Liberal} = \frac{\sum N_j \hat{y}_i}{\sum N_j}$$

$$\hat{y}^{PS}_{Conservative} = \frac{\sum N_j \hat{y}_i}{\sum N_j}$$

Results

In the previous section, we use four variables including age, sex, education level and provinces lived, to build two multilevel regression models for the Liberal and the Conservative parties. By applying the census data to the multilevel regression models and calculating the post-stratification, we get the predicted proportion of people who would vote for the Liberal or the Conservative in each province, which is also the predicted popular vote results in each province if there was a 100% turnout and the respective $\hat{y}$'s will be

$$\hat{y}^{PS}_{Libreal} = 0.3461472$$
$$\hat{y}^{PS}_{Conservative} = 0.3155943$$

Now we take a deeper dive into individual predictor variables and how they affected the results.

In Figure 1 we see the Liberal V.S Conservative proportions based on the gender.

In Figure 2 we see the Liberal V.S Conservative proportions based on the province.

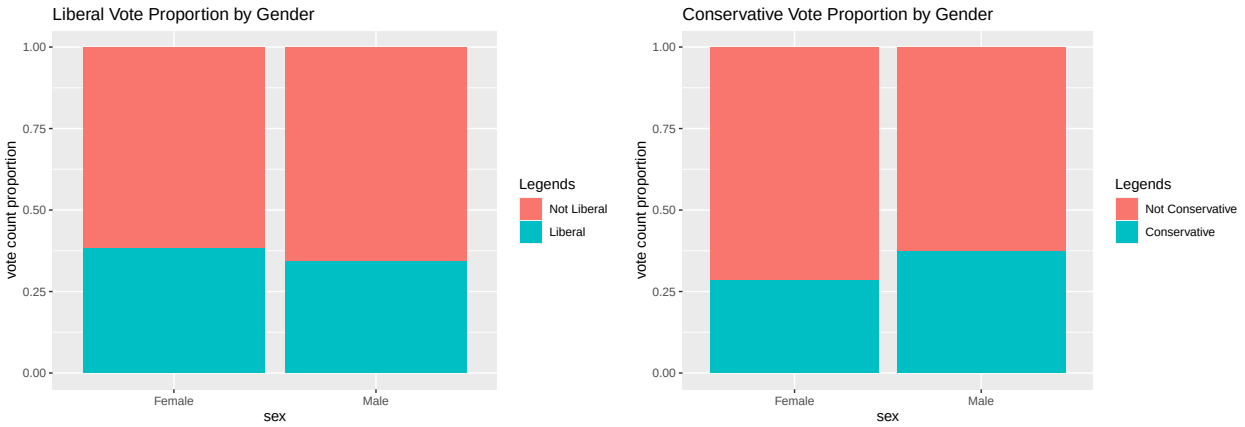

Figure 1: Vote Proportion by Gender

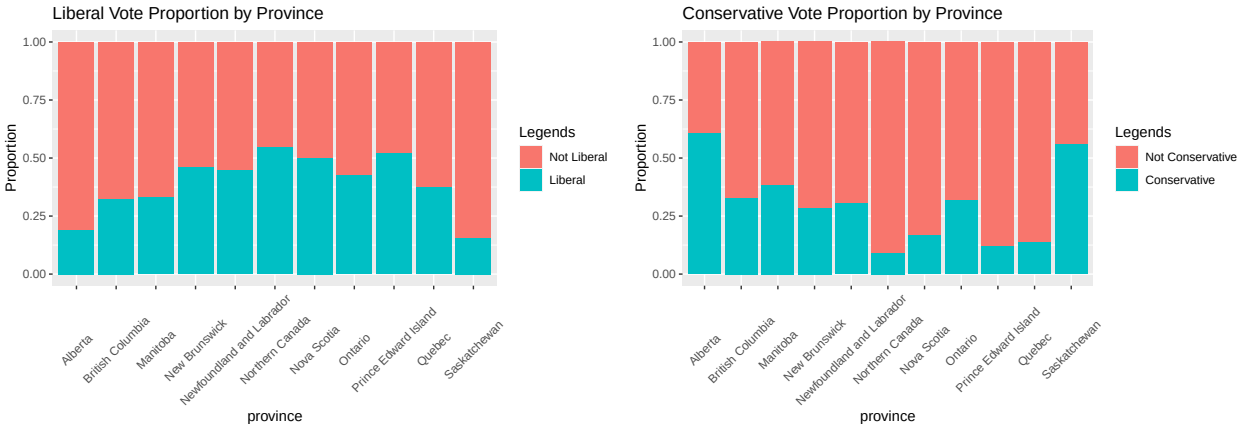

Figure 2: Vote Proportion by Province

Discussion

Per the aforementioned analysis strategy, the survey data provided by CES was leveraged and furthered by post-stratification analysis on GSS provided census data to correctly estimate the number of people voting either liberal or conservative.

It is clear from Figure 2 that more females voted for Liberals than male and more male vote for conservative than female However, male and female both voted for non conservative and non liberal majority of the times. This is an anomaly as female voted majority of the time for non liberal does that mean they voted for conservative or vice versa. We see the same trend in all the predictor variables This is further discussed in the weakness section.

In Figure 3 provinces like Manitoba and NL had 100% turnout then conservatives would have gotten less seats. Similarly NL would have gotten more liberal seats given 100% turnout (Federal election 2019 live results). North Canada would have gotten more conservative seats.

Given the above results of our analyses, it is clear that the lack of voter participation has a large impact on the results of an election. In comparison to the actual results of the 2019 Canadian federal election, which gave liberals 46% of the seats and 36% to the Conservative, our results show definitive changes: if there was a 100% voter turnout, liberal would have 35% of the seats and conservative would have 32%. While consistent with actual results, our results show that despite everyone voting, the election would not have resulted in a majority government. It is important to note that this does not invalidate our research, in fact, it goes to show that if all voters were accounted for, results would be balanced - hence representing a variety of opinions and beliefs embodied by the Canadian population.

Weakness

1. Only accounts for liberal versus conservative. While the two parties split the majority of the vote together, there are several other contenders also in the running during the federal elections, namely the NDP. The lack of this variable acts as a weakness in this analysis. This creates anomalies which were discussed above. Form Figure 1, majority of the people voted for non-liberal but that does not equate to majority of people voting for conservative as the votes for both the parties were binary and other parties were not considered.

2. The data of the 2016 GSS only had two option for the gender and the survey data has three options. Because of this the the other's option in the survey data was disregarded which does not depict the full picture of elections.

Next Steps

Based on the results of this model and analysis, we can take our research further in a few ways to enhance the utilization and accuracy of this concept.

1. Include more predictor variables to create a holistic model. While age, gender, province and education level make for a good base model to allow us to create an educated prediction regarding voter results, one of the best ways to improve such a study would be to include more predictor variables in the analysis. Factors such as ethnicity and household income are examples of variables that could provide significant detail to such an analysis.
2. Apply this model to other countries and their elections. To examine the accuracy and usefulness of this model and analyses, next steps can include a similar process on other countries' elections. Perhaps
3. Utilize such models to raise awareness on the importance of voting.

Refrences

1. https://www150.statcan.gc.ca/n1/daily-quotidien/200226/dq200226b-eng.htm
2. Stephenson, Laura B., Harell, A., Rubenson, D., & Loewen, Peter J. (2020). 2019 Canadian Election Study - Online Survey. https://doi.org/10.7910/DVN/DUS88V, Harvard Dataverse, V1
3. Faculty of Arts & Sciences, University of Toronto. (2016). General social survey on Canadians at Work and Home (cycle 30),
4. Computing in the Humanities and Social Sciences. http://www.chass.utoronto.ca/index.html
5. Wickham, H., et al. (2019, November 19). Welcome to the Tidyverse. Retrieved October 18, 2020, from https://tidyverse.tidyverse.org/articles/paper.html
6. Xie Y (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, from https://yihui.org/knitr/
7. Federal election 2019 live results, CBCnews, December 22, 2020, https://newsinteractives.cbc.ca/elections/federal/2019/results/
8. Wickham, H., et al. (2019, November 19). Welcome to the Tidyverse. Retrieved October 18, 2020, from https://tidyverse.tidyverse.org/articles/paper.html
9. Xie Y (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30, from https://yihui.org/knitr/

Appendix

```
model_liberal
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_liberal ~ (1 | age) + (1 | sex) + (1 | education) + (1 |
##     province)
##    Data: survey_data
##       AIC       BIC    logLik  deviance  df.resid
##  7566.318  7599.798 -3778.159  7556.318      5974
## Random effects:
##  Groups    Name        Std.Dev.
##  province  (Intercept) 0.54447
##  education (Intercept) 0.28657
##  age       (Intercept) 0.04722
##  sex       (Intercept) 0.14406
## Number of obs: 5979, groups:  province, 11; education, 6; age, 4; sex, 2
## Fixed Effects:
## (Intercept)
##     -0.6174
```

```
model_conservative
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: vote_conservative ~ (1 | age) + (1 | sex) + (1 | education) +
##     (1 | province)
##    Data: survey_data
##       AIC       BIC    logLik  deviance  df.resid
##  6965.208  6998.688 -3477.604  6955.208      5974
## Random effects:
##  Groups    Name        Std.Dev.
##  province  (Intercept) 0.7889
##  education (Intercept) 0.2859
##  age       (Intercept) 0.1129
##  sex       (Intercept) 0.3252
## Number of obs: 5979, groups:  province, 11; education, 6; age, 4; sex, 2
## Fixed Effects:
## (Intercept)
##     -0.9307
```