Canadian Federal Election analysis using MRP model based on the GSS(CES) and CESR(model) data

Author: jinying Liu(10002987967)

Due Date: Dec. 22, 2020

Abstract

Since we have predicted the popular vote outcome of the 2020 American federal election, then I want to apply what we learn to the Canadian Federal Election. As the pattern we did in the PS3, I collected the data from CESR[i] to produce the model and use the data from GSS[ii] as the CES data for the prediction. The reason I use the GSS data is the Canadian census dataset from the "Statistics Canada" website has a very different structure, which means the structure would not be able to let me use the multivariate post-stratification method, then I use the GSS data in PS2 instead it.

Key word

Canadian Federal Election, Liberal Party of Canada, Conservative Party of Canada

Introduction

The federal election is always one of the most important things in Canada. It will change the Prime Minister and influencing the policy of government in the future. The MRP model is the perfect model to analyze an election for a country. As we learned in class, the standard variables in the election model has long been to adjust for gender, age, race, and education. Therefore, I plan to use the variables of individual level: age, sex, race, education level, income, and group level variables: province, to build the multivariable logistic (logit) regression model. The dependent variable would be the cps19_votechoice in CESR2019 which is federal party the candidate voted for. In the predicted data, the variables of age, income, and education level are binary outcomes. The data was extracted from the 2017 General Social Survey (GSS) on the Family, which is a nation-wide survey that taking in Canadians' social-demographic variables, as well as seizing the socio-economic trend over the country. Using the representative counts (GSS) to predict Canada's federal election if "everyone" votes. Before running the model, I suppose the result will be different from the real result of the federal election in 2019, which indicate the importance of turnout.

Methodology:

Data:

For the part of cleaning data for Post-Stratification, Firstly, put sex into male, female, and others. Secondly, make the province same for GSS and CESR2019 dataset. Thirdly, put language people speak at home into English, French, and others. Lastly, put the education level into 5 levels which are lower than high school, high school, lower than diploma, diploma and higher than diploma.

I was trying to add a term of income, but the scale of income for two data set are totally different. Also, I was trying to add a term to indicate individual circumstance of work, but I think it will have some correlation to the education and it is hard to clean the data. Then I did not use that two terms.

Model:

$$\log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{\text{province}} + \beta_4 x_{language} + \beta_5 x_{education}$$

where,

- p_i *represents the probability of voterswho will vote for* Liberal Party(or Conservative Party)
- x_{age} *is the age of* corresponding people X_{ii},
- x_{sex} *is the sex of* corresponding people X_{ii}. (1 *represents male*, 2 *represents female*, 3 *represents others*)
- x_{province} *is the province of* corresponding people X_{ii},
- $x_{language}$ *is the language usually speak at home of* corresponding people X_{ii},
- $x_{education}$ *is highest education of* corresponding people X_{ii},
- *all* β *is the coefficient in the*.

I make two model. One is for the Liberal Party and the other is for the Conservative Party. The only different is the P_i, first model is the probability voted for Liberal Party and second model is the probability voted for Conservative Party.

Post-Stratification

Essentially estimate $\frac{p_i}{1-p_i}$ for each cell (using multi-level modelling). Use demographics to "extrapolate" how entire population will vote.

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

Where,

y_j is the estimate in each cell.

N_j is the population size of the j_{th} cell based off demographics.

For my model:

The province is representing the group level in the model.

The age, sex, language, education is representing the individual level in the model.

Results:

The summary table for Liberal Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.733	0.177	-9.788	0.000
age	0.006	0.002	3.994	0.000
sexMale	-0.173	0.048	-3.576	0.000
sexOther	0.094	0.323	0.292	0.770
language_homeFrench	-0.226	0.098	-2.300	0.021
language_homeother	0.031	0.117	0.262	0.793
education	0.227	0.025	9.197	0.000

From the table above, we can see the other sex term and the other language term are not significant since the P-value is higher than the 0.05. This maybe is caused by me put all other type in sex or language into one term, but they may have different impact on the result.

We calculated from the poststratification conducted the 95% confidence interval around it, are presented in Table below. The code is in the r-code file.

Point estimate and margin of error of probability of voted for Liberal Party

Point.estimate	Margin.of.error
0.2725	0.1846

We can see the estimate result of the voted rate for the **Liberal Party** from my model is 27.25%

The summary table for Conservative Party

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.341	0.209	-6.404	0.000
age	0.008	0.002	5.154	0.000
sexMale	0.579	0.050	11.690	0.000
sexOther	-0.143	0.368	-0.389	0.697
language_homeFrench	-0.764	0.120	-6.387	0.000
language_homeother	0.426	0.115	3.693	0.000
education	-0.123	0.025	-4.944	0.000

From the table above, it is different from the first model. The only one not significant is the other sex term which mean different language speaker have significant different in the result of voted for Conservative Party.

We calculated from the poststratification conducted the 95% confidence interval around it, are presented in Table below. The code is in the r-code file.

Point estimate and margin of error of probability of Conservative Party

Point.estimate	Margin.of.error
0.3036	0.297

We can see the estimate result of the voted rate for the **Conservative Party** from my model is 30.36%

Discussion:

Summary:

I analyzed survey data at the individual level from CES poststratification census data from the GSS to predict two main party in 2019 Canadian federal election. I established a multivariate Logistic regression model and used it to calculate post-stratified estimates which my result shows that **Liberal Party** has voting rate of 27.25% and the **Conservative Party** has voting rate of 30.36%. The real result of 2019 Canadian federal election is quite similar to my result of **Conservative Party** has a little bit higher voting rate which is **Liberal Party** has voting rate of 33.12% and the **Conservative Party** has voting rate of 34.34%.[1]

Conclusions:

In this report, I am interested in predicting the popular voting rate for the 2019 Canadian federal election which I have used a logistic regression model and Post-Stratification method to estimate percentage of people who will vote for **Liberal Party** and **Conservative Party**. My analysis shows that the **Conservative Party** have a higher voting rate. However, that is not respect the result of the 2019 Canadian federal election, since the voting rate not same as the Seats of the party got. The result of the real voting rate in 2019 Canadian federal election is close for the **Liberal Party** and **Conservative Party**, but there is higher difference of voting rate in my model result. Therefore, if "everyone" votes the result may be differentiated.

Weakness & Next Steps:

The voting rate is not same as the result of the 2019 Canadian federal election. For example, in the real 2019 Canadian federal election, **Conservative Party** have a higher voting rate, but the **Liberal Party** won the election eventually. Build model for each province would be better, since the set will have difference among different province and the voting rate for each province may have some method to transfer to the set. Then we can calculate the probability of winning the federal election.

The power of our model is suboptimal since I only used five variables in our logistic regression model. It should be lots of factors that influence the voting rate.

The technique for calculating the margin of error can be more robust takes into the account of post-stratification weighting.

For the data part, it would be some response bias in the sampling since the data is all coming from mail.

[1] 1. https://en.wikipedia.org/wiki/2019_Canadian_federal_election

References

i. 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
LINK: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V

ii. Technology, A. (2020, April). Data Centre. Retrieved December 10, 2020, from http://dc.chass.utoronto.ca/myaccess.html

iii. R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

iv. The data in the summary, https://en.wikipedia.org/wiki/2019_Canadian_federal_election

v. Rohan, and Sam Caetano. "gss_cleaning.R", 7 Oct. 2020.
LINK: tellingstorieswithdata.com