General Prediction of winner Liberal party in the 2019 Canadian Federal Election

Kaicheng Huang

2020/12/21

Code and data supporting this analysis is available at:https://github.com/kennethhh123/STA-304-final-proj

Abstract

As I interested in whether the Canadian Election Study Data sets can be used to predict the result of election, and whether some variable in the data relate to people vote choice. I will create a multilevel regression model with post-stratification to predict the winner political party of the 2019 Canadian Federal Election. Through my analysis, I have found that the Liberal party won the election by prediction, which is the same as the fact. The result proved the Federal Election result can be predicted by the model, and have high accuracy compare with the true vote choice.

Keywords

Canadian Election Study Data sets,General Social Survey Data sets, Multilevel regression model, 2019 Canadian Federal Election, Liberal Party

Introduction

The Canadian general election held every four years, and it is the most important political event in Canada, which concerned by Canadian citizens, immigration, or even other countries. The 2019 Canadian Federal Election is the 43rd Canadian general election, was held on October 21, 2019. Thus, the general prediction of this election can figure out relevance between the properties or thoughts of people and their vote choice.

Multilevel regression with post stratification is a statistical technique used for correcting model estimates for known differences between a sample population, and a target population. In this report, I will use MRP to discern if there is a causal link between the properties or thoughts of people and their vote choice.

In the Data section, two data sets will be used to investigate how MRP could be used to make inference n the causal link between the properties or thoughts of people and their vote choice (Section 1). In regards to the Methodology section, these two data sets will be cleaned can then matched up, and the MRP model was used to perform the prediction result (Section 2). Prediction result analysis and model validation result are provided in the Result section (Section 3), and inferences of this data along with conclusions are presented in Conclusion section (Section 4).

The main interest of this report is that how election result can be predicted by the variables in the data sets. Two data sets have been cleaned and matched up, and then used to create MRP model to predict the people vote choice toward the federal election. The result shows that the Liberal party will win the election, which is the same as fact. Lastly, the report is important as it shows that the result can be predict by variables in the CES data, and these data are relevant to people vote choice.

Methodology

Data

Two data sets have been used in this report. The first data set is the Canadian Election Study Data Set (2019) that obtained from cesR package, and the second data set is the General social Survey Data Set (2017) that obtained from CHASS website. During the process of cleaning data, in order to match up variables in both data sets, I mutated column names and outputs, and filter out variables that don't matched in both data sets, and omit the missing observations lastly. After data cleaning, I denoted CES data as survey_data and GSS data as census_data to proceed further analysis. Thus, the sample population is survey_data and the target population is census_data. The variables I chose are "Age", "Sex", "Health" and "Province" as these variables provide basic information of people. For those non-response, I denote them as missing observations and omitted. Furthermore, the key feature of these two data sets is that these data sets are the data in Canada, which have strength to provide high relevance to the Canada federal election. However, these data sets have some weakness, one is that the time of two data sets are different (2019 vs 2017), another is that not every variable in the data sets match perfectly. For example, the variable "Province" has 13 kind of outputs in survey_data and only 10 kind of outputs, which need to filter 3 outputs out. Although it sill effect the accuracy of prediction, the effect is not large since these outputs doesn't contain much observations. Regarding the variables in the data set that I used, "Age" stand for the age of respondents, which is integer; "Sex" stand for the gender of respondents, which has two kind of observations Male and Female; "Health" stand for the health of respondents, which measures in 5 different levels; and the last variable "Province" stand for the province of resident of the respondents, which represent the different province in Canada. There is a similar variable in the data set but I didn't use it, which is "Region". The reason why I didn't choose this variable is that I think "Province" is more detail and representative than "Region".

survey_data overview

sex	age	health	province	vote_party
Female	50	Very good	ON	0
Female	72	Fair	ON	0
Female	28	Very good	ON	0
Male	59	Excellent	BC	0
Female	60	Very good	ON	1

census_data overview

age	sex	province	health
15	Female	AB	Excellent
15	Female	AB	Very good
15	Female	BC	Excellent
15	Female	BC	Fair
15	Female	BC	Very good

Through these cleaned data sets, survey_data will be use to create MRP model and census_data will be use to do post stratification in order to predict the winner of election.

Model

By definition, Multilevel regression with post stratification is a statistical technique used for correcting model estimates for known differences between a sample population, and a target population. Since Multilevel models are particularly appropriate for research designs where data for participants are organized at more than one level (Fidell, Barbara G. Tabachnick, Linda S. 2007), I used the multilevel regression model here to figure out the proportion of each political party in the Canadian Federal Election, R studio is the software I used to create this model. The model created by four predictor which are Age“,”Sex“,”Health" and “Province”, and I used these predictor to calculate the vote proportion of each political party.

Here is the logistic equation of the model:

$$log(p/1-p) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{sex} + \beta_3 x_{province} + \beta_4 x_{health}$$

Which each notation was defined below: p representing the proportion of voters supports the left wing,
β_0 representing the the intercept of the model,
β_1 representing the log odds based on age,
β_2 representing the log odds based on sex,
β_3 representing the log odds based on province,
β_4 representing the log odds based on health.

In regards to the variables in the model, chose variable “Province” rather than similar variable “Region” can let the model become more significance. Using “Age” rather than “Age_group” can also make model become more significance as “Age” is numerical variable.

Through the model validation, the p-value shows that all variables are significant except “health. Thus, The alternate model can be the reduced model by only have variables”age“,”sec" and “province” try to get more significant. Although the reduced model maybe be better in AIC values, it has too less variables. Thus, the reduced model may not representative. Also, the origin model can be compared with the reduced model by checking AIC and p-value.

The alternate model can be the reduced model by only have variables “age”, “sec” and “province”. Even the reduced model maybe be better in AIC values, it has too less variables. Thus, the reduced model may not representative.

Post Stratification

Post-stratification is a common technique in survey analysis for incorporating population distributions of variables into survey estimates (Little, R., 1993). Thus, post stratification analysis is a effective method to investigate the estimate of variables. In this report, continue with the multilevel regression model, I used census_data to calculate the estimate of relative variables, and then figure out the higher vote proportion political party.

Results

```
##
## Call:
## glm(formula = vote_party ~ age + sex + health + province, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0618  -0.9004  -0.7718   1.3941   2.1693
##
```

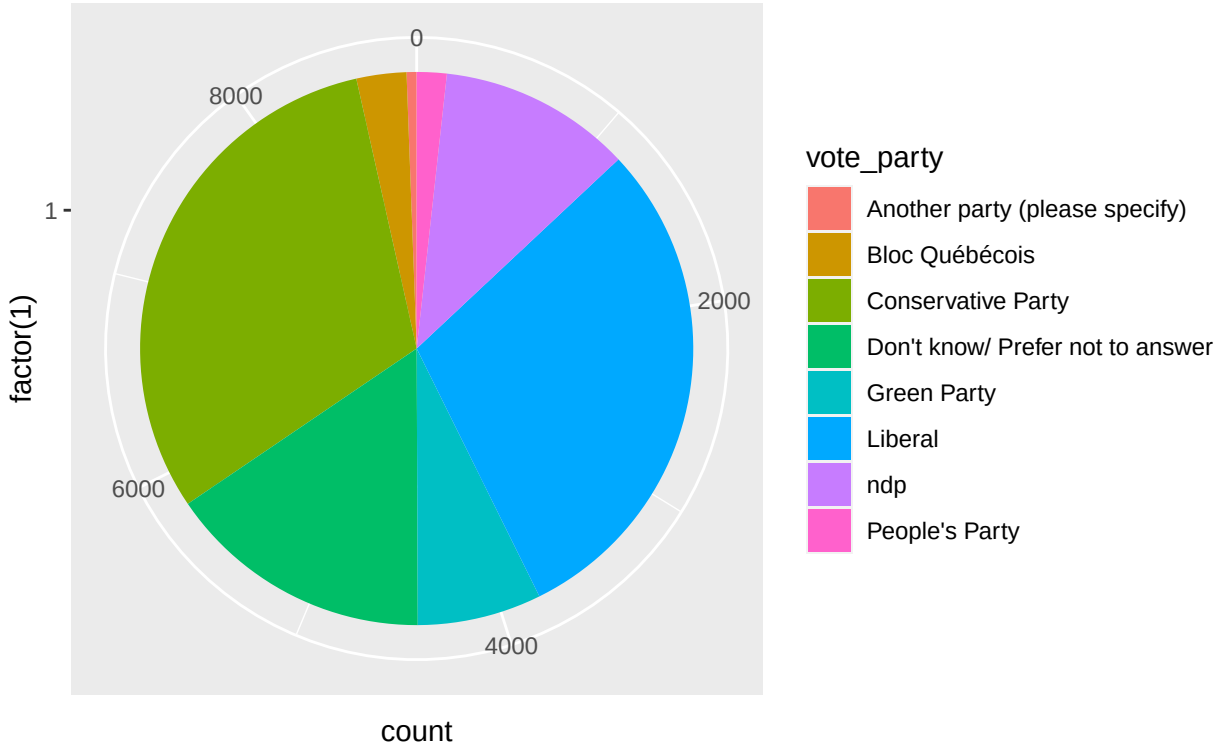
```
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.737420   0.126947 -13.686  < 2e-16 ***
## age              0.004455   0.001567   2.843  0.00446 **
## sexMale         -0.135821   0.048185  -2.819  0.00482 **
## sexOther         0.091023   0.321719   0.283  0.77723
## healthFair      -0.215296   0.070960  -3.034  0.00241 **
## healthOther     -1.157798   0.485148  -2.386  0.01701 *
## healthPoor      -0.200978   0.121501  -1.654  0.09810 .
## healthVery good -0.029394   0.064045  -0.459  0.64626
## provinceBC       0.606603   0.104226   5.820 5.88e-09 ***
## provinceMB       0.603474   0.135217   4.463 8.08e-06 ***
## provinceNB       0.891951   0.166334   5.362 8.21e-08 ***
## provinceNL       1.151877   0.174158   6.614 3.74e-11 ***
## provinceNS       1.051681   0.140313   7.495 6.62e-14 ***
## provinceON       1.047720   0.087951  11.913  < 2e-16 ***
## provincePE       0.773581   0.318357   2.430  0.01510 *
## provinceQC       0.859395   0.101149   8.496  < 2e-16 ***
## provinceSK      -0.315956   0.178820  -1.767  0.07725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10792  on 8874  degrees of freedom
## Residual deviance: 10506  on 8858  degrees of freedom
##   (1425 observations deleted due to missingness)
## AIC: 10540
##
## Number of Fisher Scoring iterations: 4


##
## Call:
## glm(formula = vote_party ~ age + sex + province, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0473  -0.9083  -0.7753   1.4061   2.1333
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.851601   0.116027 -15.958  < 2e-16 ***
## age          0.004983   0.001561   3.192  0.00141 **
## sexMale     -0.134609   0.048115  -2.798  0.00515 **
## sexOther     0.089978   0.320881   0.280  0.77916
## provinceBC   0.601414   0.104108   5.777 7.61e-09 ***
## provinceMB   0.599307   0.135098   4.436 9.16e-06 ***
## provinceNB   0.880606   0.166041   5.304 1.14e-07 ***
## provinceNL   1.145858   0.173820   6.592 4.33e-11 ***
## provinceNS   1.045023   0.140128   7.458 8.81e-14 ***
## provinceON   1.044339   0.087881  11.884  < 2e-16 ***
## provincePE   0.754752   0.318054   2.373  0.01764 *
## provinceQC   0.860426   0.101046   8.515  < 2e-16 ***
```

```
## provinceSK  -0.315497   0.178690  -1.766  0.07746 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10792  on 8874  degrees of freedom
## Residual deviance: 10527  on 8862  degrees of freedom
##   (1425 observations deleted due to missingness)
## AIC: 10553
##
## Number of Fisher Scoring iterations: 4
```

In comparison, the AIC for the mlr model is 10540 and for reduced model is 10553. Also, the variable "health" has p-value larger than 0.05, which whis variable doesn't include in the reduced model. Thus, almost all variables in the reduced model have less than 0.05 p-value.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.293
```

Graph 1: Vote Proprotion for each political party



According to the output of post stratification analysis, the alp_predict (0.2927) which is the predict vote proportion for the election winner. Also, from graph 1, we can see that Liberal party has most predict vote proportion, along with the second party is Conservative Party. Other parties like ndp just have few vote.

Discussion

Summary

Through the whole report, after importing the data sets Canadian Election Study Data (2019) and General Social Survey Data (2017), I cleaned and matched up these two data sets. Then I used multilevel regression to create a model with parameter "age", "sex", "health", and "province". As the variable "health" is not significant, I tried to create a reduced model that without variable "health". However, by comparing the p-value and AIC, the origin model is still better. Also, I used post stratification analysis to predict the winner political policy of 2019 election, and I got Liberal party win with 29.27% vote proportion, which is approach to the vote proportion by respondent.

In regards to the calculation of post stratification analysis, Liberal party has the highest vote proportion with 29.27%, which means we predicted that the Liberal party will won the 2019 election.In the model validation, our origin model has less AIC than reduced model, which means the origin model is better than the reduced model. But the p-value of variable "health" in the origin model shows that this variable is not significant. Lastly, when we look at the graph 1, Liberal party has the most predicted vote, which also conclude that we predicted Liberal won the election.

Conclusion

In conclusion, by using two data sets to create multilevel regression with post stratification, we successfully predicted that the Liberal party won the election by the highest vote proportion (29.27%).

Weaknesses

There are some weakness exist through the full report. Firstly, the data set I used to analysis is not ideal. To illustrate, these two data sets Canadian Election Study Data (2019) and General Social Survey Data (2017) are not perfectly relate due to time differ. The second point is that the model also have some weakness in variables choosing. As we can from the p-value of each variables, variable "health" is not really significant as its p-value larger than 0.05. Also, the variable "age" is not perfectly match, as well as the variable "province". Thus, in order to successfully use post stratification analysis, I filtered out the observation that not matched in both data sets. Although the observation filtered out were very few, it still have slight effect to the prediction. Lastly, since we compare our vote proportion prediction result with the respondent vote choice in the CES data set rather than the real vote proportion result for the 2019 Canadian Federal Election, the conclusion of prediction accuracy maybe differ. As we can see from this comparison, the real vote proportion for winner Liberal party is 33.12% (2019) rather than proportion in our data is around 28%. Under this circumstance, the model has less accuracy.

Next Steps

In order to improve the model as well as the study, there are some next steps to do. One point is that we can choose the analysis data under same year, such as find sample population and target population in a same year as our survey_data and census_data. Therefore, the data sets will have more relevant, the result of study will thereby be more accurate. In addition, when we build model, we should choose variables more strictly. Choosing variables that match up better in both data sets to prevent filtering out observation can also improve the accuracy. Also, we can add more variables to the model as predictors to make the model better. When the origin model has variable that is not significant, we can try to replace that variable with another rather than just reduce that variable. Thus, the new model will not be lack representative due to

too less variables. Also, when we try to improve to the model to get better result, we can use different methodology to see if other model whether better or not, such as replace MRP model with BRM model.

References

- General Social Survey Data (2017) http://www.chass.utoronto.ca/
- Census Data cleaning code rohan.alexander@utoronto.ca
- Strength of post stratification Fidell, Barbara G. Tabachnick, Linda S. (2007). Using multivariate statistics (5th ed.). Boston ; Montreal: Pearson/A & B.
- 2019 CES Online Survey 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 Documentation for the 2019 CES Online Survey can be accessed from here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DUS88V
- MRP desription Survantion (2018) https://www.survation.com/what-is-mrp/
- Result of 2019 Canada Federal Election https://www.elections.ca/res/rep/off/ovr2019app/home.html#3
- Post stratification Little, R. (1993). Post-Stratification: A Modeler's Perspective. Journal of the American Statistical Association, 88(423), 1001-1012. doi:10.2307/2290792
- Tidyverse package Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Knitr package Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

  Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

  Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595
- Janitor Package Sam Firke (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor