Use of MRP Model Predicted Liberal Party would win Canadian Election

YeHao Zheng

December 21, 2020

GitHub repository

https://github.com/trtrgfh/sta304-final_project

Abstract

The 2020 US election has been a trending topic with its new record of number of people voted, and as its neighbour, people start to wonder how the 2019 Canadian Federal Election would have been different if more people had voted. In this paper, we'll be using a MRP model to predict which party would win the election with age, sex, education, marital status, and province as predictors. The result will be a close match between the Liberal Party, and the Conservative Party, but the Liberal Party would win eventually. Keywords: MRP model, forecasting, Canada Election 2019, Liberal Party, Conservative Party.

Introduction

For countries like Canada and the US, election has always been an important event that every citizens should be a part of. Especially With the number of voters reached a new high in the US, meaning more people are taking the elections more seriously. It got me wonder if more people voted in the 2019 Canadian Federal Election, how would the result be different. Would the Conservative Party overcome the obstacles and take the win or would the Liberal Party further secure its place. Since the break out of Covid-19, different countries have reacted differently. Some citizens are not satisfy with the acts their government has done, and many of them are realizing the importance of being part of the election. If people can vote again in the 2019 Canadian Federal Election, the result might actully be different. In order to find the probability that whether the Liberal Party would still win the election, I'll be looking at the age, sex, education, marital status of the voters, and which province the voters are currently living in. For the layout of the rest of the report, I'll use multiple predictors to investigate the wining chance of the Liberal Party based on the cleaned CES and GSS datase. The method I use will be the multilevel regression with poststratification model which is a popular way to adjust a sample population to better analyse a target population. In the Results section, the result of whether the Liberal Party would win the election will be provided, and some graph will be shown to support the outcome. Finally, the advantage of the MRP model, and some concerns about the model will be discuss at the end.

Data

```
## [1] "<table class=\"Rtable1\">\n<thead>\n<tr>\n<th class=\"grouplabel\"></th>\n<th colspan=\"1\" cla
```

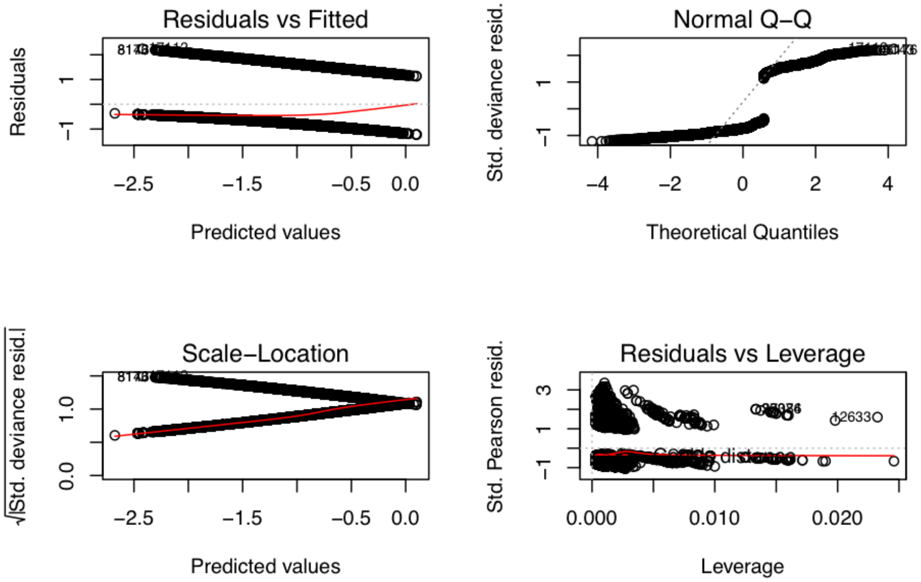
The survey data use in this paper is from the 2019 online survey of Canadian Election study. The data consists of 37822 cases and 634 variables(Canadian Election Study, 2019), and is mean to gather the attitude and opinions of Canadian for the 2019 Canadian Election. The census data is from the Canadian general social surveys (GSS) which is provided by CHASS. For the survey data, I would like to predict whether the Liberal Party of Conservative Party would win the election, and I've chosen the variable sex, age, province, educ, and marital_status as predictors. For the response variable, cps19_votechoice is used, and by mutating a new column where 1 is the votes for Liberal and 0 is the votes for other parties, the number of votes for Liberal can be counted. After removing all the NA in this dataset, we left with 31099 cases, and from the above baseline characteristics table, we see that 8823 of them would vote for Liberal. In this paper, the population is all Canadians, and the frame is people who intend to vote for Liberal, and the sample is Canadian who filled out the online survey from the Canadian Election Study. Using the sample, we'll predicted the vote intention for all Canadians.

Model

```
## 
## Call:
## glm(formula = vote_liberal ~ agegroup + sex + province + educ + 
##     marital_status, family = "binomial", data = survey_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2210  -0.8533  -0.7361   1.3603   2.2387  
## 
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                        -1.45920    0.07421 -19.664  < 2e-16 ***
## agegroup35 to 55                    0.01396    0.03569   0.391 0.695696    
## agegroup55 to 70                    0.14445    0.03832   3.769 0.000164 ***
## agegroupabove 70                    0.26984    0.05170   5.220 1.79e-07 ***
## sexMale                            -0.00374    0.02671  -0.140 0.888643    
## sexOther                           -0.22350    0.16358  -1.366 0.171835    
## provinceBritish Columbia            0.66054    0.06023  10.968  < 2e-16 ***
## provinceManitoba                    0.60954    0.07754   7.861 3.82e-15 ***
## provinceNew Brunswick               0.91363    0.09419   9.699  < 2e-16 ***
## provinceNewfoundland and Labrador   1.32081    0.10111  13.063  < 2e-16 ***
## provinceNova Scotia                 1.16353    0.08448  13.773  < 2e-16 ***
## provinceOntario                     0.99861    0.05011  19.930  < 2e-16 ***
## provincePrince Edward Island        0.98045    0.19331   5.072 3.94e-07 ***
## provinceQuebec                      0.82652    0.05365  15.405  < 2e-16 ***
## provinceSaskatchewan               -0.26738    0.10326  -2.589 0.009614 ** 
## educCollege degree                 -0.43983    0.03808 -11.551  < 2e-16 ***
## educHigh school                    -0.40222    0.03311 -12.147  < 2e-16 ***
## educLess than high school          -0.59366    0.06893  -8.612  < 2e-16 ***
## educMaster or doctorate degree      0.07091    0.04207   1.685 0.091899 .  
## educNo schooling                   -0.59607    0.28575  -2.086 0.036976 *  
## marital_statusLiving common-law    -0.14632    0.05872  -2.492 0.012711 *  
## marital_statusMarried              -0.10051    0.05097  -1.972 0.048640 *  
## marital_statusSeparated            -0.01605    0.08225  -0.195 0.845320    
## marital_statusSingle, never married 0.01672    0.05682   0.294 0.768631    
## marital_statusWidowed              -0.09522    0.08126  -1.172 0.241290    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
```

```
##     Null deviance: 37096  on 31098  degrees of freedom
## Residual deviance: 35971  on 31074  degrees of freedom
## AIC: 36021
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 36229.32
```



```
## Start:  AIC=36131.76
## vote_liberal ~ agegroup + sex + province + educ + marital_status
##
##                  Df Deviance   AIC
## - marital_status  5    35991 36120
## - sex             2    35973 36121
## <none>                 35971 36132
## - agegroup        3    36011 36153
## - educ            5    36277 36406
## - province        9    36684 36787
##
## Step:  AIC=36120.23
## vote_liberal ~ agegroup + sex + province + educ
##
##            Df Deviance   AIC
## - sex       2    35993 36109
## <none>           35991 36120
## - agegroup  3    36032 36142
## - educ      5    36295 36391
## - province  9    36709 36780
##
## Step:  AIC=36109.22
## vote_liberal ~ agegroup + province + educ
```

```
##
##            Df Deviance   AIC
## <none>           35993 36109
## - agegroup  3    36035 36132
## - educ      5    36298 36381
## - province  9    36711 36769
```

```
##
## Call:
## glm(formula = vote_liberal ~ agegroup + province + educ, family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2226  -0.8479  -0.7456   1.3748   2.2203
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.523608   0.054934 -27.735  < 2e-16 ***
## agegroup35 to 55                 -0.001216   0.034184  -0.036 0.971630
## agegroup55 to 70                  0.124497   0.034907   3.567 0.000362 ***
## agegroupabove 70                  0.243636   0.047197   5.162 2.44e-07 ***
## provinceBritish Columbia          0.663284   0.060189  11.020  < 2e-16 ***
## provinceManitoba                  0.609142   0.077517   7.858 3.90e-15 ***
## provinceNew Brunswick             0.908110   0.094125   9.648  < 2e-16 ***
## provinceNewfoundland and Labrador  1.317491   0.101050  13.038  < 2e-16 ***
## provinceNova Scotia               1.163783   0.084415  13.786  < 2e-16 ***
## provinceOntario                   1.002388   0.050064  20.022  < 2e-16 ***
## provincePrince Edward Island      0.975515   0.193224   5.049 4.45e-07 ***
## provinceQuebec                    0.826910   0.053255  15.527  < 2e-16 ***
## provinceSaskatchewan             -0.267034   0.103222  -2.587 0.009682 **
## educCollege degree               -0.441343   0.037978 -11.621  < 2e-16 ***
## educHigh school                  -0.394708   0.032941 -11.982  < 2e-16 ***
## educLess than high school        -0.584210   0.068739  -8.499  < 2e-16 ***
## educMaster or doctorate degree    0.068183   0.041985   1.624 0.104384
## educNo schooling                 -0.593436   0.285933  -2.075 0.037946 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 37096  on 31098  degrees of freedom
## Residual deviance: 35993  on 31081  degrees of freedom
## AIC: 36029
##
## Number of Fisher Scoring iterations: 4
```

For this paper, the model is performed in the statistical language R (R Core Team, 2020). For the predictors used in this model, age was selected insetead of age group because older population tend to be more invovled in the election. Most young people are not really interested in politics. I have also chosen sex as a categorical predictor. Typically, there are more male involved in politics, but in this survey data, there are actually more female. It would be interesting to see how sex would change the result. Then I've chosen province as a categorical predictor. Just like the US, each province in Canada also have their own favorite party. For example, people in Alberta are more likely to vote for the Conservative Party. Therefore, province is an important factor to predict the result. Moreover, the categorical predictor educ which represents the education leven of the respondents also play a big part in the election. People with different eduction would have different opinions on the speeches and commitments made by the parties. Lastly I've chosen the categorical predictor marital_status. People with different marital_status might also have different thought on the party's acts. A party might be more supportive on having more people getting babies or it can give out more funds to single mothers. For the response variable, I've created a binary variable where 1 is the votes for the Liberal Party, and 0 is the votes for other parties. I've also created another binary variable where 1 is the votes forthe Conservative Party, and 0 is the votes for other parties. Since these two party are the top two competitors in the election, the party with a higher vote percentage would be the predict winner of the election. The model used is the generalized linear model, and the model would be apply to the post-stratified data. Let x_1, X_2, X_3, X_4, X_5 represent agegroup, sex, province, educ, and marital_status respectively. Let Y_i represent the \$iˆ{th} \$response variable observation- the binary variable of Liberal Party or Other parties. Then the model would look like:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon_i$$

The beta coefficients is the slope of the predictors and the epsilon is the error term for the i^{th} observation. This model clearly find the significant level of the predictors, but the variables used in this model are all categorical which means the regression line would probably not follow a linear trend. This might make the result not as accury. By looking at the summary table, we see that most of the p-values are smaller than the significant level of 0.05 which means most of the predictors are significant in terms of predicting the election result. However, the AIC and BIC values are so big that it indicates this model is not well fitted. Using the backward elimination with BIC, the new fitted model is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i$$

where x_1, X_2, X_3 represent agegroup, educ, and province respectively.

Result

```
## # A tibble: 1 x 1
##   predict_liberal
##             <dbl>
## 1           0.284
```

```
## # A tibble: 1 x 1
##   predict_conservative
##                  <dbl>
## 1                0.274
```

From the above graphs, we see that the Liberal Party tend to be more popular among female, and the number of female and male respondent who are willing to vote for the Conservative Party is very close. For the education level of the respondents, more high educated people intend to vote for the Liberal Party. People with high school diploma and college degree are more likely to vote for the Conservative Party. For the marital status, more single and never married respondents are willing to vote for Liberal Party, and more married respondents are willing to vote for the Conservative. For the provinces the respondents are currently living in, the people in Alberta are significantly more favored in the Conservative Party, and the Liberal Party gets a lot more votes from people in Ontario and Quebec. Finally, there's not much difference in the age distribution for both party. The Liberal Party gets a bit more votes from people at age 35 or less. Then, modelled by a generalized linear model with predictors age, sex, educ, province, marital status, and based off the post-stratification analysis of the proportion of voters in favour of the Liberal and Conservative Party. we estimate the final result that the Liberal Party would get a proportion of about 28.37 percents of the votes, and the Conservative Party would get a proportion of about 27.39 percent of the votes. Since these two party have the most votes among all other parties, the Liberal Party would win the election in this close match with the Conservative Party.

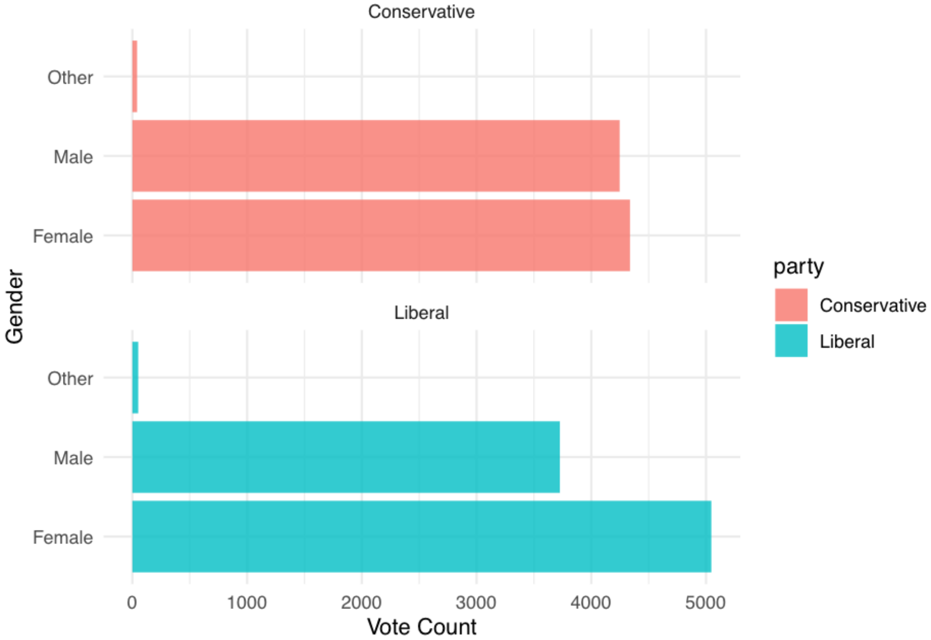


Figure 1: Popular vote perdiction by gender

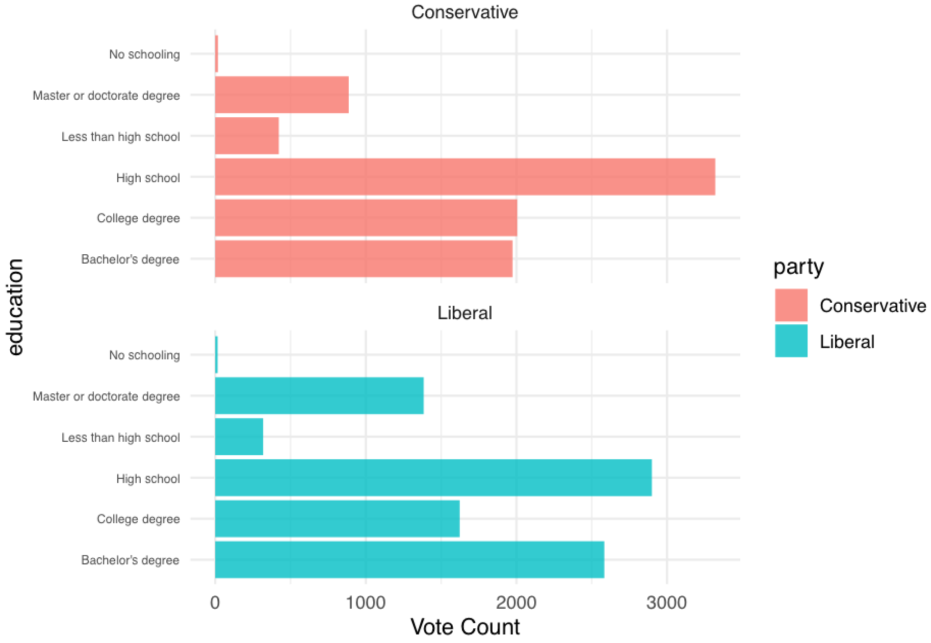


Figure 2: Popular vote perdiction by education

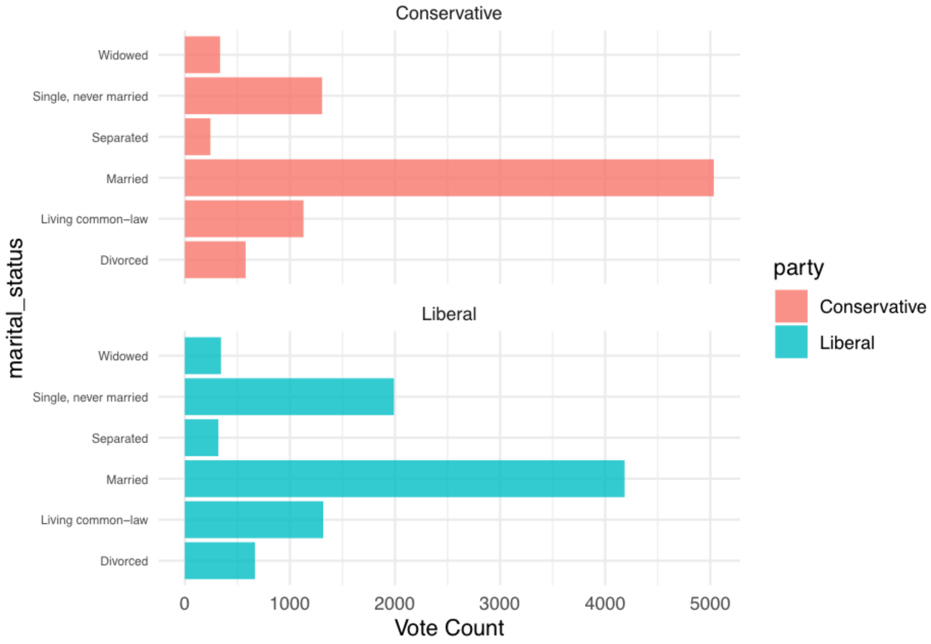


Figure 3: Popular vote perdiction by marital_status

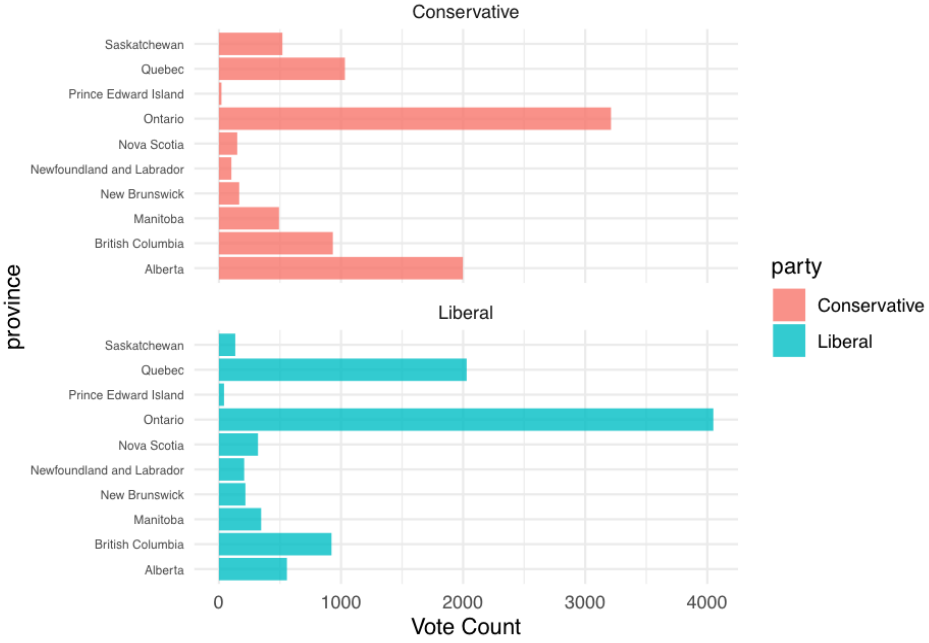


Figure 4: Popular vote perdiction by province

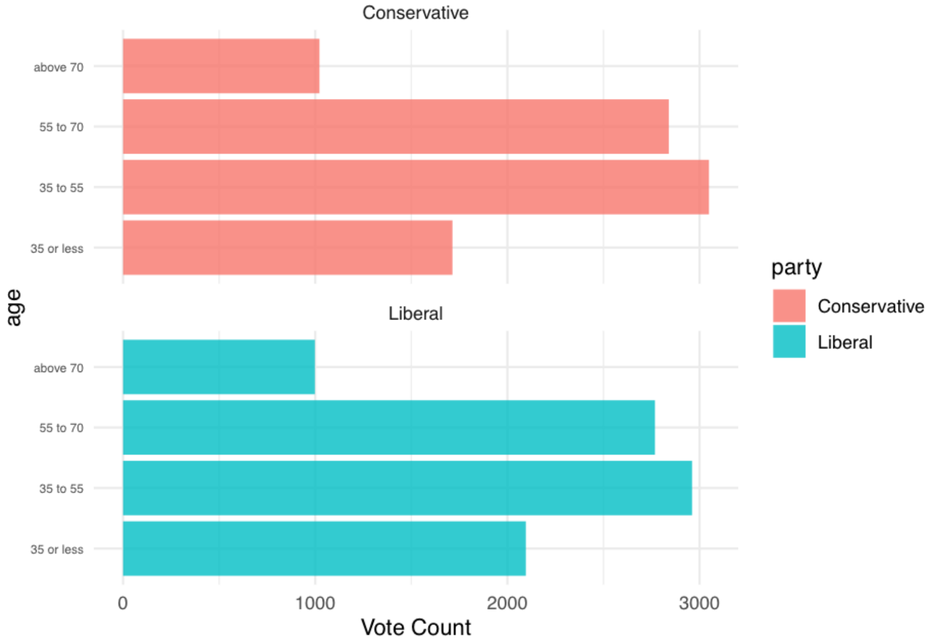


Figure 5: Popular vote perdiction by age

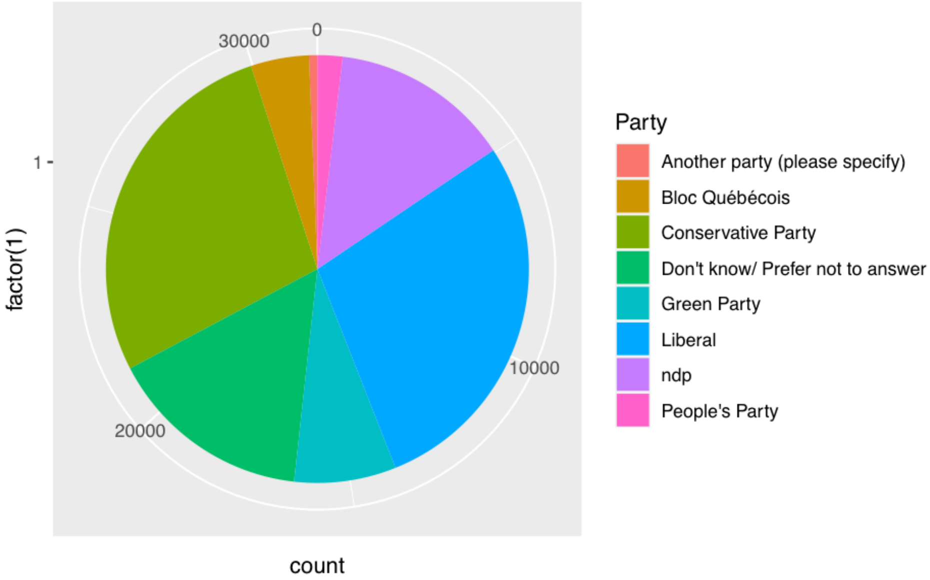


Figure 6: Popular vote perdiction by parties

Discussion

In this paper, we use the dataset from the 2019 online survey of Canadian Election study and the Canadian general social surveys (GSS), and we predicted the proportion of vote in favour of the Liberal and Conservative Party using a generalized linear model with post stratification. Then we get the result that the proportion of votes in favour of the Liberal Party is around 28.37 percent and the proportion of votes in favour of the Conservative Party is around 27.39 percent. Based on the result, we know that among 100 respondents, about 28 of them would vote for the Liberal Party and 27 of them would vote for the Conservative Party. In figure 6, we see that the Liberal Party and the Conservative Party have coverd more areas than any other parties, and their votes are actually too close that we couldn't tell who wins the election. Also, looking at figure 4, we further confirms that the geographical advantage of the parties that different province may favor different party. Futhermore, figure 3 has shown that education is also a big factor that effect the result such that more high educated people favour in the Liberal Party. Different parties have different opinions on managing the country, and one mistake made by the winning party might cause enormous damage to the country. Especially this year, many countries have made terrible calls on the reactions to the Covid-19 and their citizens have suffered greatly.

Weakness

One weakness of the analysis is that we predict the proportion of votes by a online survey, so the result might not be as accury. Also, getting more votes in the analysis does not mean the party will defintely win the election. There are many cases where a party wins the popular vote from survey but still lost the election. Another weakness is that I have only chose 5 predictors and there are a lot more factors that may effect the result of an election. The variable cps19_votechoice also have many NAs, and after removing them, we only have around 30000 cases left. For future improvements of the analysis, we can use a more updated census data. The census data used in this paper is produced in the year 2016. Also, we can add more predictors to the model and first finding out which one of them are significant to the result, and deleting the ones that are not as significant.

References

Alexander, R and Caetano, S. (2020). *01-data_cleaning-post-strat1*. University of Toronto

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Rich, B. (2020, November 25). Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html

Stastic Canada (2020, April). General Social Survey, Cycle 31 : Families Public Use Microdata File Documentation and User's Guide, fro https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

Data from, https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

Multiple Linear Regression Analysis. (n.d.). Retrieved October 19, 2020, from http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis

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

```
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agosti
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }

##
## To cite package 'table1' in publications use:
##
##   Benjamin Rich (2020). table1: Tables of Descriptive Statistics in
##   HTML. R package version 1.2.1.
##   https://CRAN.R-project.org/package=table1
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {table1: Tables of Descriptive Statistics in HTML},
##     author = {Benjamin Rich},
##     year = {2020},
##     note = {R package version 1.2.1},
##     url = {https://CRAN.R-project.org/package=table1},
##   }

##
## To cite package 'cesR' in publications use:
##
##   Paul A. Hodgetts and Rohan Alexander (2020). cesR: Access the CES
##   Datasets a Little Easier.. R package version 0.1.0.
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {cesR: Access the CES Datasets a Little Easier.},
##     author = {Paul A. Hodgetts and Rohan Alexander},
##     year = {2020},
##     note = {R package version 0.1.0},
##   }

##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }
```