All citizens voted for 2019 survey

Feixue Han

2020/12/20

#Abstract

The Canadian Election Survey shows Canadians' altitude of life and choice of candidates before and after the election. According to the survey, we discuss there will be a difference if everyone voted in 2019 Canadian Election. Moreover, by using the logistic model and post-stratification, the result will be predicted according to province, age and sex.

#Introduction

In 2019 October 13th, the 43rd Canadian federal election was held, and Justin Trudeau became the new Prime Minster Canada. As a member of the Liberal Party, Justin Trudeau won 157 seats from the government. [6] There are several Parties attended this election, which is the Liberal Party, Conservative Party, Bloc Qu, New Democratic Party, Green Party and People's Party. The Conservative Party and Liberal Party are the two majority parties in this election. The Liberal Party of Canada is the longest-serving and oldest active federal political party in Canada. The party has dominated federal politics for much of Canada's history. [7,8] In the 2019 federal election, they lost the popular vote [cite9] and their majority in the House of Commons, winning 157 seats, but they remained the largest party in the House. In Canada, in 37,802,043 people who have right to vote, 66% of the population voted for the election. However, there are still nearly 33% of the people who did not vote. [10] It is important to motivate Canadian citizens to vote for the election because the result will be different. Thus, through this survey we explore the result of the 2019 Canadian election if everyone voted.

In this report, the survey data was chosen from Canadian Election 2019 online survey (CES) and Education Highlight Tables (2016 Census) [5,11]. The CES has recorded Canadians' political behaviour and altitudes and the preference on the key political issues. The Education Highlight Tables is the latest version in the Statistic Canada website; thus, we can assume that no changes happened until 2019. The datasets have categories of age, sex, education, province and vote choice. The MRP model was built based on the online survey data and the post-stratification was based on the Education Highlight table. While cleaning the data, only the person with the answer "certainly to vote" were kept for further research. The answers without choosing specific parties were also filtered. If all the people voted in the election, the supporting rate of Liberal Party will increase a lot.

Some biases still exist in this survey. The answers with "certainly vote" were kept so that there will be biases in the result. The sample size of the survey data is much smaller than the census which cause the limit conditions in model. Also, the cell can be divided to more parts so that the result can be more accurate. The data sets were completed a few months before the election thus, the information is not fresh enough. In addition, the data set is lack of some conditions such as marriage, income, health and so on. In the future, the model should be improved. Survey data should have larger sample size and it should be the latest version. To make the result more accurate, Bayesian approach model can be applied to predict the result. By comparing the result, the better one can be used to report. In conclusion, when more people vote for the Canadian election, Liberal Party will have a higher supporting rate. At the same time, the model can be used to predict election result, but it still needs to be improved.

#data

The survey data is chosen from 2019 Canadian Election Study. The online Survey was conducted to document the attitudes of Canadians during the 2019 election period. The tradition of Canadian Election Studies started in 1965. [9] In the survey data, it included sex, age, education, vote choice and so on. At first, the data set had 37822 cases and after cleaning, 6030 cases were kept. The non- responded answers were filtered. This may cause biases. The census data is from website of statistic Canada. In this dataset it includes age, province, sex and highlighted education column. In the model, two predictors (sex and province) and cell were selected as predictors. The binary model was built to predict the election result.

According to figure1, the vote rate of the citizens is very similar. In figure 2, the box plot for sex shows that men prefer to vote for Liberal Party and women prefer to vote for Conservative Party. For figure3, the Conservative Party gets more votes than the Liberal Party. But the overall number of people who support for Liberal and Conservative Party is similar.

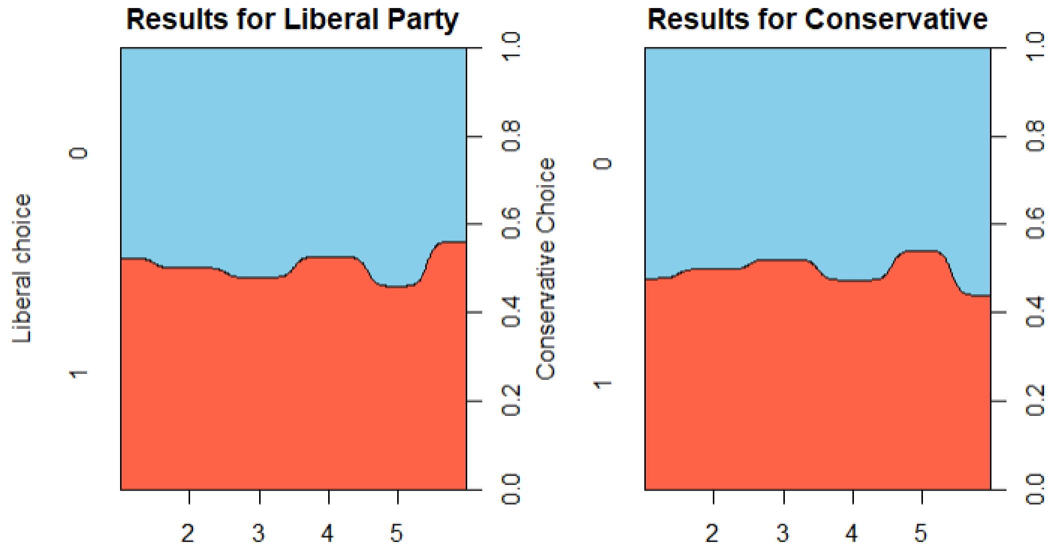


Figure 1

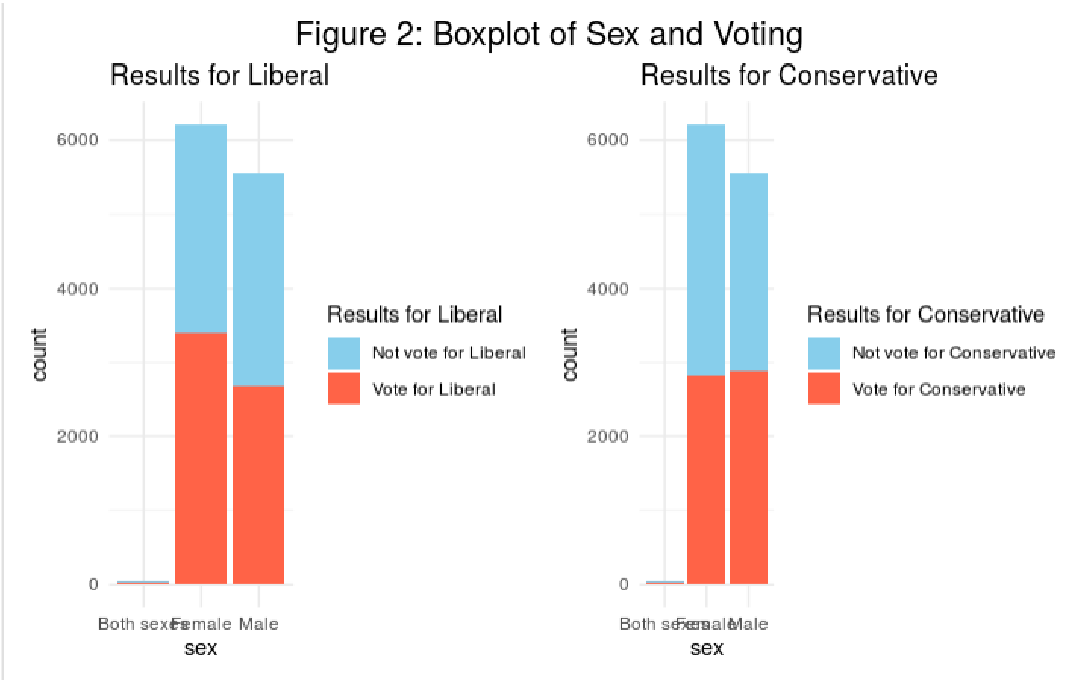


Figure 2

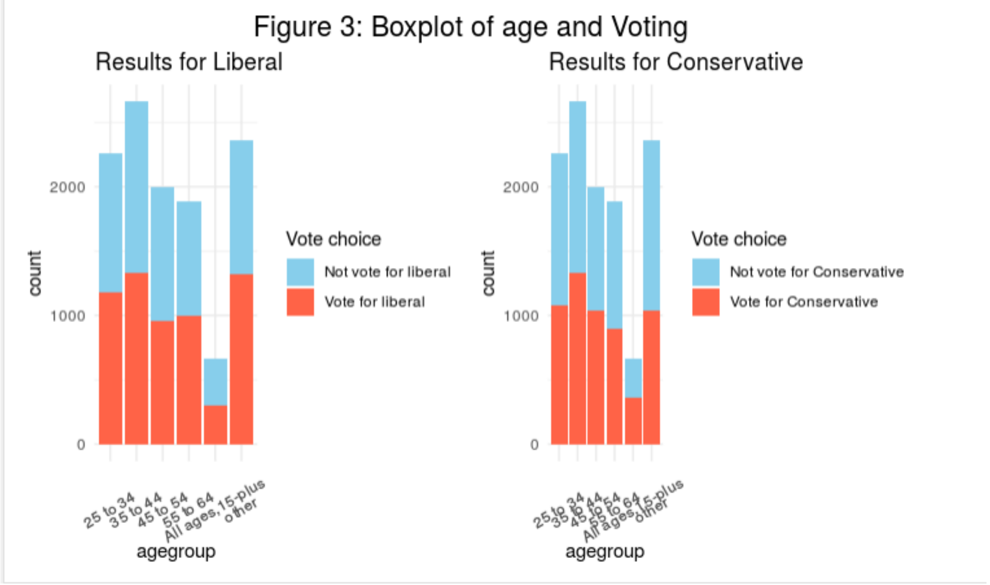


Figure 3

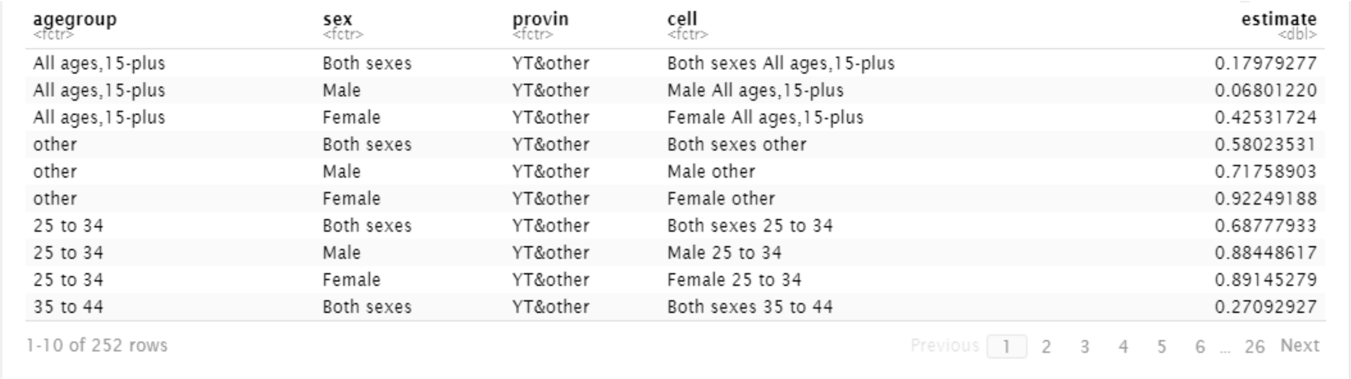
#post-stratification

The multilevel regression and post-stratification (MRP) techniques were used to determine relationship between the variables which were chosen. The post-stratification method combined with the linear logistic regression process. At first, the model of vote choice, vote liberal and vote conservative were made to predict the vote result among the large sample. The reason to choose MRP model is that this survey is about demographic. Four predictors were used to predict according to the model. By combining MRP and the predictors, results from different population can be estimated.

However, because of the insufficient sample of the survey data, bias will be existing in the result. Other to that, it can still be chosen as the predict model.

While doing post-stratification, sex, age, province was chosen as the predictor. Then the cell which includes sex and province were created to prepare for creating models. The vote.liberal and vote.conservative probability were predicted below.

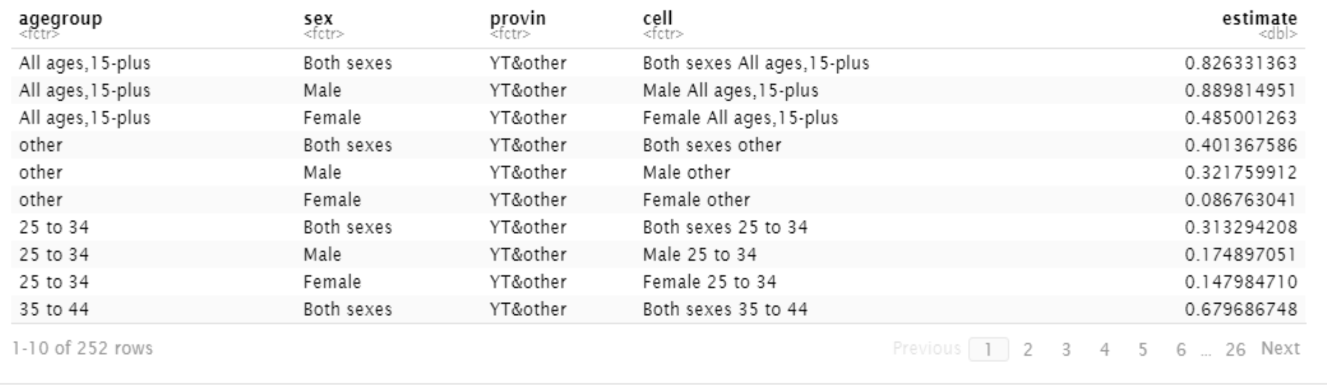
#Liberal



agegroup <fctr>	sex <fctr>	provin <fctr>	cell <fctr>	estimate <dbl>
All ages,15-plus	Both sexes	YT&other	Both sexes All ages,15-plus	0.17979277
All ages,15-plus	Male	YT&other	Male All ages,15-plus	0.06801220
All ages,15-plus	Female	YT&other	Female All ages,15-plus	0.42531724
other	Both sexes	YT&other	Both sexes other	0.58023531
other	Male	YT&other	Male other	0.71758903
other	Female	YT&other	Female other	0.92249188
25 to 34	Both sexes	YT&other	Both sexes 25 to 34	0.68777933
25 to 34	Male	YT&other	Male 25 to 34	0.88448617
25 to 34	Female	YT&other	Female 25 to 34	0.89145279
35 to 44	Both sexes	YT&other	Both sexes 35 to 44	0.27092927

1-10 of 252 rows Previous 1 2 3 4 5 6 ... 26 Next

#conservative



agegroup <fctr>	sex <fctr>	provin <fctr>	cell <fctr>	estimate <dbl>
All ages,15-plus	Both sexes	YT&other	Both sexes All ages,15-plus	0.826331363
All ages,15-plus	Male	YT&other	Male All ages,15-plus	0.889814951
All ages,15-plus	Female	YT&other	Female All ages,15-plus	0.485001263
other	Both sexes	YT&other	Both sexes other	0.401367586
other	Male	YT&other	Male other	0.321759912
other	Female	YT&other	Female other	0.086763041
25 to 34	Both sexes	YT&other	Both sexes 25 to 34	0.313294208
25 to 34	Male	YT&other	Male 25 to 34	0.174897051
25 to 34	Female	YT&other	Female 25 to 34	0.147984710
35 to 44	Both sexes	YT&other	Both sexes 35 to 44	0.679686748

1-10 of 252 rows Previous 1 2 3 4 5 6 ... 26 Next

#model

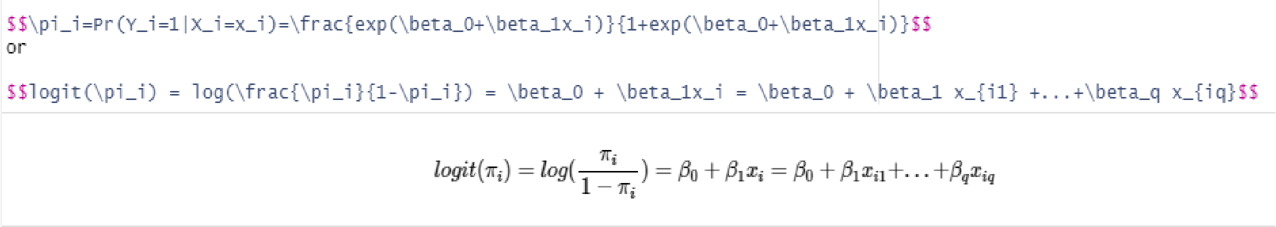
The purpose of our study is to determine whether the result of the Canadian election could affect the actual vote outcome. Multilevel regression and post-stratification technique were used for this analysis. In the following sub-sections I will describe the model specifics and the calculation for the post-stratification process.

#model specifies

In the beginning of the model, the linear logistic regression was used to analysis the result of the election. The model is used to predict the proportion of voters who will vote for Liberty Party (Justin Trudeau) and the proportion of voters who will vote for Conservative Party (Andrew Scheer) separately.

For both models, four same factors were used: age, sex, education, and province. Only age is a numeric model and others are categorial variables. To predict the result of the election, two model were used in total. The Akaike information criterion (AIC) was used for testing which model is better for fitting our data. AIC uses the number of independent variables that are used to make a model and the maximum likelihood estimate of the model to get a value[11]. The maximum likelihood estimate shows how accurate the calculated. By comparing the AIC for the multi regression model, the AIC for the two model is 19651 and 19418 separately. The data with smaller AIC tells us it has more accurate result. Therefore, there will be a better model by comparing the AIC value.

Two binary logistic regression models were used to run for the result who will vote for Liberty Party or Conservative Party base on the survey 2019 CES web data (cite). The equation below is the binary logistic regression model:



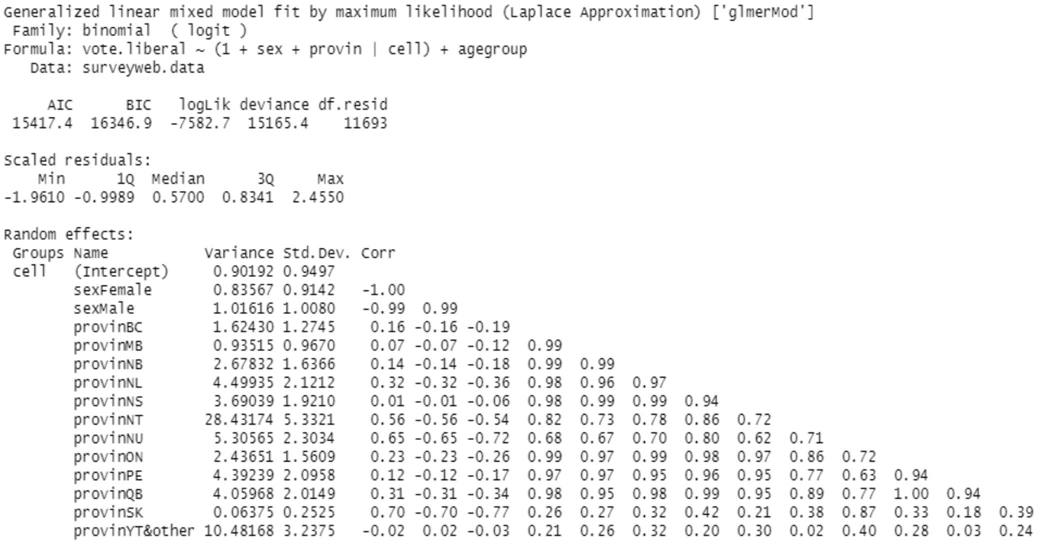
```
$$\pi_i=Pr(Y_i=1|X_i=x_i)=\frac{exp(\beta_0+\beta_1x_i)}{1+exp(\beta_0+\beta_1x_i)}$$
or

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1x_i = \beta_0 + \beta_1 x_{i1} +...+\beta_q x_{iq}$$
```

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1x_i = \beta_0 + \beta_1 x_{i1} +...+\beta_q x_{iq}$$

We assume that Y_i is a binary response variable for i = 1,...n and takes on value 0 or 1 with P(Y_i = 1) = pi_i. Suppose X is a set of explanatory variables, x_i is the observed value of the explanatory variables for observation i = 1,...q. From the above formula, we can also get: $$\frac{\pi}{1-\pi}=e^{\beta_0}e^{\beta_1x_1}...e^{\beta_q x_q} $$

Then the $\beta_0$ is the baseline odds and $\beta_1$ can be interpreted as holding predictors constant, a one-unit increase in $x_1$ increases the probability of voting for the Liberal Party or Conservative Party by a factor of $e^{\beta_1}$.

# result for Liberal Party



```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: vote.liberal ~ (1 + sex + provin | cell) + agegroup
   Data: surveyweb.data

     AIC      BIC   logLik deviance df.resid
 15417.4  16346.9  -7582.7  15165.4    11693

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9610 -0.9989  0.5700  0.8341  2.4550

Random effects:
 Groups Name          Variance Std.Dev. Corr
 cell   (Intercept)    0.90192 0.9497
        sexFemale      0.83567 0.9142   -1.00
        sexMale        1.01616 1.0080   -0.99  0.99
        provinBC       1.62430 1.2745    0.16 -0.16 -0.19
        provinMB       0.93515 0.9670    0.07 -0.07 -0.12  0.99
        provinNB       2.67832 1.6366    0.14 -0.14 -0.18  0.99  0.99
        provinNL       4.49935 2.1212    0.32 -0.32 -0.36  0.98  0.96  0.97
        provinNS       3.69039 1.9210    0.01 -0.01 -0.06  0.98  0.99  0.99  0.94
        provinNT      28.43174 5.3321    0.56 -0.56 -0.54  0.82  0.73  0.78  0.86  0.72
        provinNU       5.30565 2.3034    0.65 -0.65 -0.72  0.68  0.67  0.70  0.80  0.62  0.71
        provinON       2.43651 1.5609    0.23 -0.23 -0.26  0.99  0.97  0.99  0.98  0.97  0.86  0.72
        provinPE       4.39239 2.0958    0.12 -0.12 -0.17  0.97  0.97  0.95  0.96  0.95  0.77  0.63  0.94
        provinQB       4.05968 2.0149    0.31 -0.31 -0.34  0.98  0.95  0.98  0.99  0.95  0.89  0.77  1.00  0.94
        provinSK       0.06375 0.2525    0.70 -0.70 -0.77  0.26  0.27  0.32  0.42  0.21  0.38  0.87  0.33  0.18  0.39
        provinYT&other 10.48168 3.2375  -0.02  0.02 -0.03  0.21  0.26  0.32  0.20  0.30  0.02  0.40  0.28  0.03  0.24
```

```
  0.54
Number of obs: 11819, groups:  cell, 18

Fixed effects:
                        Estimate Std. Error z value          Pr(>|z|)    
(Intercept)              -1.2606     0.1747  -7.216 0.000000000000537 ***
agegroup35 to 44         -0.1044     0.2113  -0.494             0.621    
agegroup45 to 54         -0.3133     0.2375  -1.319             0.187    
agegroup55 to 64          0.2030     0.2285   0.888             0.374    
agegroupAll ages,15-plus  0.0879     0.2943   0.299             0.765    
agegroupother             0.2799     0.1991   1.406             0.160    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) a35t44 a45t54 a55t64 aAa,15
agegrp35t44 -0.475                            
agegrp45t54 -0.763  0.491                     
agegrp55t64 -0.796  0.339  0.571              
aggrpAa,15- -0.412  0.075  0.381  0.454       
agegroupthr -0.768  0.272  0.619  0.682  0.376
optimizer (Nelder_Mead) convergence code: 4 (failure to converge in 10000 evaluations)
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 8 negative eigenvalues
failure to converge in 10000 evaluations
```

#for Conservative Party

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: vote.conservative ~ (1 + sex + provin | cell) + agegroup
   Data: surveyweb.data

     AIC      BIC   logLik deviance df.resid 
 15418.3  16347.9  -7583.2  15166.3    11693 

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-2.4282 -0.8346 -0.5719  0.9979  1.9421 

Random effects:
 Groups Name          Variance Std.Dev. Corr                                                        
 cell   (Intercept)    1.13895 1.0672                                                               
        sexFemale      1.08212 1.0403   -1.00                                                       
        sexMale        1.15377 1.0741   -0.99  0.99                                                 
        provinBC       1.65540 1.2866    0.07 -0.07 -0.11                                           
        provinMB       0.97493 0.9874   -0.05  0.05 -0.01  0.99                                     
        provinNB       2.76521 1.6629    0.04 -0.04 -0.08  0.99  0.99                               
        provinNL       4.52336 2.1268    0.17 -0.17 -0.23  0.98  0.97  0.97                         
        provinNS       3.80016 1.9494   -0.07  0.07  0.03  0.98  0.99  0.99  0.94                   
        provinNT      19.86259 4.4567    0.49 -0.49 -0.48  0.82  0.73  0.79  0.85  0.73             
        provinNU       5.38833 2.3213    0.35 -0.35 -0.43  0.66  0.67  0.69  0.79  0.59  0.66       
        provinON       2.48051 1.5750    0.11 -0.11 -0.15  0.99  0.97  0.99  0.98  0.97  0.86  0.71 
        provinPE       4.30693 2.0753    0.04 -0.04 -0.09  0.97  0.97  0.96  0.96  0.96  0.76  0.61  0.94 
        provinQB       4.10715 2.0266    0.19 -0.19 -0.22  0.98  0.96  0.98  0.99  0.95  0.89  0.75  1.00  0.94 
        provinSK       0.07034 0.2652    0.39 -0.39 -0.46  0.35  0.36  0.40  0.50  0.29  0.41  0.92  0.41  0.26  0.47
        provinYT&other 7.28798 2.6996   -0.13  0.13  0.12  0.37  0.40  0.46  0.35  0.44  0.23  0.49  0.44  0.19  0.41


  0.56
Number of obs: 11819, groups:  cell, 18

Fixed effects:
                         Estimate Std. Error z value             Pr(>|z|)    
(Intercept)               1.31403    0.12574  10.450 <0.0000000000000002 ***
agegroup35 to 44          0.04546    0.18293   0.249                0.804    
agegroup45 to 54          0.27066    0.21308   1.270                0.204    
agegroup55 to 64         -0.27462    0.20610  -1.332                0.183    
agegroupAll ages,15-plus -0.13951    0.27808  -0.502                0.616    
agegroupother            -0.33253    0.21476  -1.548                0.122    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) a35t44 a45t54 a55t64 aAa,15
agegrp35t44 -0.459                            
agegrp45t54 -0.706  0.483                     
agegrp55t64 -0.712  0.380  0.477              
aggrpAa,15- -0.175  0.018  0.316  0.082       
agegroupthr -0.757  0.232  0.504  0.386  0.408
optimizer (Nelder_Mead) convergence code: 4 (failure to converge in 10000 evaluations)
boundary (singular) fit: see ?isSingular
failure to converge in 10000 evaluations
```

By using the binary logistic mode, two summary statistic tables were created. The first table is about the summary of the Liberal Party model while the second table is about the summary of Conservative Party model. In the Pr category, the age in 15 plus has the highest value which is 0.765 while the other age groups have relatively large p-value. The P-value for the age in other range is lower. The maximum z value for the age group is 1.406 and the minimum is -7.216 which is from other age group and all age above 15 group. They are half positive and half negative. For the second table, in the pr category, the highest value is 0.804 in 35 to 44 age group and the lowest one is in the other age group. The maximum z value is 1.270 for the age group 45 to 54 and the minimum is -1.548 which is from other age group. According to the graph, men is more likely to vote for the Liberal Party while women prefer the Conservative Party. There is not much difference in the boxplot of age group.

#discussion

According to the result of the model, p-value plays an important role on telling whether it is a null hypothesis or not. P-value often help to indicate whether the factors have correlation with the predictor. When the p-value is over 0.05, the null hypothesis is supported, and the result is not significant. If the p-value is less than 0.05, it rejects the null hypothesis so there may be a correlation between that factor and our response variable. There is a stronger evidence that it rejects the null hypothesis when it gets smaller. While calculating model for the Liberal Party, the cell includes age and sex, and the predictors are sex and province. For the p-value of Liberal party, the intercept is smaller than 0.05 while the age group above 15 is 0.765 which is larger than 0.05. Thus, it has no significant effect on the result. While calculating model for the Conservative Party, the cell includes age and sex, and the predictors are sex and province. For the Conservative Party, the p- value for the intercept is less than 0.05 which has a significant effect on the result. However, the p-value for age 35-44 is 0.621 which is larger than 0.05. The result has no significant effect on the result.

#weekness

While cleaning the data, only the answers with “certain to vote” were kept. Since the non-response answers were removed so that the biases exist in the result. The sample size of the survey data is not sufficient. The sample size is much smaller than the census data. As a result, bias will be produced in the result. Otherwise, the division of the cell is not detailed enough. While creating the cell using survey data section, the cell only included the sex and age. If the cell included more variables, the result will be more accurate than before. Also, the limited computing techniques is a weakness in the analysis. While creating the logistic model, there are not many conditions included in the model. Because of the limited computing technique and sample size, bias will be reduced a lot. This survey was completed a few months before the Canadian election. Thus, it has been a long time until now. Moreover, the data is somehow lacking predictors such as marriage, income etc. If more variables can be added into the data set, the result could be more accurate.

#next step

In the future, the model and dataset can be improved. The sample of the survey data can be increased. Also, more categories can be added to the data set such as: income level, marriage and so on. The date of the survey should be more closed to the election date. Moreover, Bayesian approach model can be applied to predict the result. By comparing the result, the better one can be used to report.

#reference

[1] Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich

[2] R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R

Foundation for Statistical Computing. https://www.R-project.org/.

[3] François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” Journal of Open Source Software 4

(43): 1686. https://doi.org/10.21105/joss.01686.

[4] Zhu, Hao. 2020. KableExtra: Construct Complex Table with ’Kable’ and Pipe Syntax. https://CRAN.Rproject.

org/package=kableExtra.

17

[5]- 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

- Stephenson, Laura, Allison Harrel, Daniel Rubenson and Peter Loewen. Forthcoming. 'Measuring Preferences and Behaviour in the 2019 Canadian Election Study,' Canadian Journal of Political Science.

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

[6] Zimonjic, Peter (October 22, 2019). "Liberals take losses but win enough in Quebec and Ontario to form minority government". Canadian Broadcasting Corporation. Retrieved October 21, 2019.

[7] Rodney P. Carlisle (2005). Encyclopedia of Politics: The Left and the Right. SAGE Publications. p. 274. ISBN 978-1-4522-6531-5.

[8] Donald C. Baumer; Howard J. Gold (2015). Parties, Polarization and Democracy in the United States. Taylor & Francis. pp. 152–. ISBN 978-1-317-25478-2.

[9] Canada, Elections. "Election Night Results - National". enr.elections.ca. Retrieved October 7, 2020.

[10] Voter Turnout at Federal Elections and Referendums, 1867-2008

[11] Stephenson, Laura B; Harell, Allison; Rubenson, Daniel; Loewen, Peter John, 2020, "2019 Canadian Election Study - Phone Survey", https://doi.org/10.7910/DVN/8RHLG1, Harvard Dataverse, V1, UNF:6:eyR28qaoYlHj9qwPWZmmVQ== [fileUNF]