Different thoughts on the result of 2019 Canadian Federal Election

Zhongfan Sun 1004031193

Due: Dec 22, 2020

Github: https://github.com/Jonassun144/304finalassignment

keywords

Canadian election, prediction, MLR, CES, post stratification

Abstract

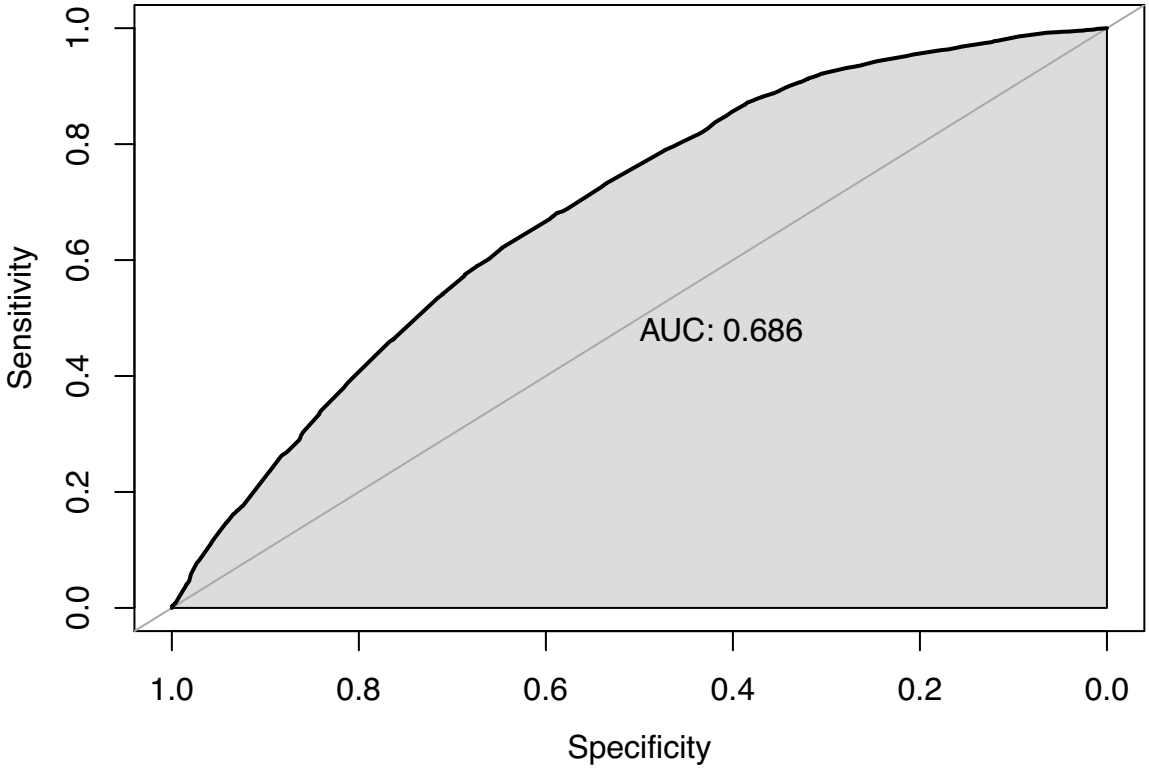
In this paper, I used the CES dataset built a multilevel regression model and a logistic regression model for the 2019 Canadian Federal Election. Then used the gss data simulate a post-stratification to predict the result of the 2019 Canadian Federal Election. Although there was some error, I finally successfully predict the actual result of the 2019 Canadian Federal Election in the end.

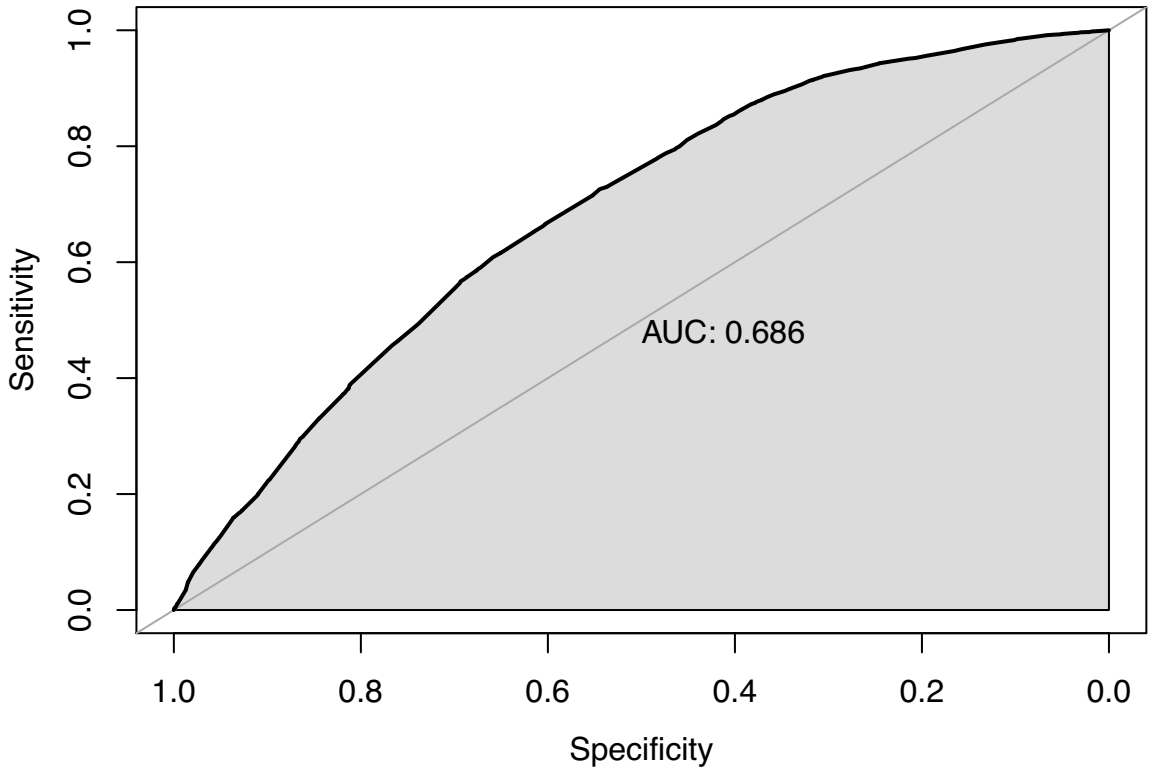
Introduction

I chose option B for my final assignment in assignment 3 our group successfully predict the result of the 2020 American Federal Election. Was that a coincidence or it was an accurate prediction? To find out, in this assignment, I will try to predict the result of the 2019 Canadian Federal Election and compare my result to the actual result of the 2019 Canadian Federal Election that everyone had voted for. This is important because, imagine if people can predict the result of the federal election, this will be a big step in the statistic field. To start with, the federal election is one of the most important events that happens to a country because, the decisions made by the government or the president will affect the future of the country in many ways and it of course will affect individuals as well in their daily life, therefore, everyone should vote. Next, it is worth knowing that the Canadian federal election system works differently compared to the American Federal Election system. Canada has 338 ridings, each riding is an area and one riding also count as one seat in the House of Commons, so there are 338 seats in the House of Commons, and they are sitting by the Member of Parliament or MPs for short which are the representatives selected by the people from the riding. There are 5 political parties in Canada an MP can from any of those parties. During the Federal Election, people don't vote for the people to run for Prime Minister directly, instead, they vote for the MP that in the same political party as the Prime Minister. If a party has the most seats in the House of Commons then they will win the election, and the leader of that party will become the Prime Minister. The party with the second place will become the opposition party or the Official Opposition.

In this assignment, I will try to predict the 2019 Canadian Federal Election by using data from the CES. I will simulate 2 models and then compare them to see which one is better. For the first model, I will do simple logistic regression and then I will use multilevel regression to group up observations by age and sex. The data I will be using is from the CES and the gss census dataset for post-stratification. First, I cleaned the data by using the code from the problem set one then, I filtered the data with people who are certain that they

will vote because in this way I can increase the certainty and accuracy of modeling and that leftover with people whose voting choice is Liberal party or Conservative party. All people from the Liberal party will vote for Justin Trudeau and people from the Conservative party will vote for Andrew Scheer. Furthermore, I narrow down gender for only men or women to match the values in the census dataset and simulate the multilevel regression model in R.






```
Call:
glm(formula = cps19_votechoice ~ cps19_gender + cps19_education +
    group_age + cps19_province, family = "binomial", data = survey_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8707  -1.1873  -0.4846   1.0294   2.0976

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -0.50978    0.15181  -3.358 0.000785 ***
cps19_genderMale                    -0.33532    0.03532  -9.493  < 2e-16 ***
cps19_educationelementaryschoollower -0.07556    0.19595  -0.386 0.699796
cps19_educationhighschool           -0.51776    0.04748 -10.905  < 2e-16 ***
cps19_educationnoanswer             -0.45743    0.50806  -0.900 0.367930
group_age21 to 35                   -0.44307    0.14908  -2.972 0.002959 **
group_age36 to 50                   -0.66188    0.14736  -4.492 7.06e-06 ***
group_age51 to 65                   -0.69952    0.14603  -4.790 1.67e-06 ***
group_ageAbove 65                   -0.62448    0.14716  -4.244 2.20e-05 ***
cps19_provinceBritish Columbia       1.35472    0.07314  18.522  < 2e-16 ***
cps19_provinceManitoba               1.02696    0.09309  11.032  < 2e-16 ***
cps19_provinceNew Brunswick          1.73740    0.12447  13.959  < 2e-16 ***
cps19_provinceNewfoundland and Labrador  2.08960    0.14547  14.364  < 2e-16 ***
cps19_provinceNorthwest Territories  2.92306    0.79648   3.670 0.000243 ***
cps19_provinceNova Scotia            2.17958    0.12063  18.068  < 2e-16 ***
cps19_provinceNunavut                1.74455    0.59024   2.956 0.003120 **
cps19_provinceOntario                1.56794    0.05859  26.760  < 2e-16 ***
cps19_provincePrince Edward Island   2.08700    0.28841   7.236 4.61e-13 ***
cps19_provinceQuebec                 2.06852    0.06721  30.778  < 2e-16 ***
cps19_provinceSaskatchewan          -0.02023    0.11677  -0.173 0.862443
cps19_provinceYukon                  1.25242    0.58329   2.147 0.031780 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20743  on 14962  degrees of freedom
Residual deviance: 18963  on 14942  degrees of freedom
AIC: 19005

Number of Fisher Scoring iterations: 4
```

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['
 Family: binomial  ( logit )
Formula: cps19_votechoice ~ (1 | cell) + cps19_education + cps19_province
   Data: survey_data

     AIC      BIC   logLik deviance df.resid
 19028.9  19158.3  -9497.4  18994.9    14946

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0383 -1.0008 -0.3501  0.8407  2.8566

Random effects:
 Groups Name        Variance Std.Dev.
 cell   (Intercept) 0.07255  0.2693
Number of obs: 14963, groups:  cell, 10

Fixed effects:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -1.19035    0.10350 -11.501  < 2e-16 ***
cps19_educationelementaryschoollower -0.08230    0.19623  -0.419 0.674918
cps19_educationhighschool           -0.51552    0.04747 -10.859  < 2e-16 ***
cps19_educationnoanswer             -0.49202    0.50838  -0.968 0.333132
cps19_provinceBritish Columbia       1.34717    0.07317  18.411  < 2e-16 ***
cps19_provinceManitoba               1.02475    0.09309  11.008  < 2e-16 ***
cps19_provinceNew Brunswick          1.73343    0.12450  13.923  < 2e-16 ***
cps19_provinceNewfoundland and Labrador  2.08426    0.14563  14.312  < 2e-16 ***
cps19_provinceNorthwest Territories  2.90401    0.80408   3.612 0.000304 ***
cps19_provinceNova Scotia            2.17622    0.12067  18.035  < 2e-16 ***
cps19_provinceNunavut                1.71550    0.59264   2.895 0.003795 **
cps19_provinceOntario                1.56580    0.05859  26.725  < 2e-16 ***
cps19_provincePrince Edward Island   2.08136    0.28861   7.212 5.53e-13 ***
cps19_provinceQuebec                 2.06633    0.06721  30.744  < 2e-16 ***
cps19_provinceSaskatchewan          -0.01963    0.11680  -0.168 0.866516
cps19_provinceYukon                  1.26250    0.59060   2.138 0.032545 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Methodology

In this analysis, we used logistic regression models to simulate the prediction of which candidate that each observation would vote for, specifically between Donald Trump and Joe Biden. Logistic regression should be applied when investigating the relationship between a binary dependent variable and other independent predictor variables. We have simulated two models, and by comparing both, we will select the model that could better represent the population. For model #1, we applied a simple logistic regression model and for model #2, we used multilevel regression to group observations by cells.

The reason for building two models is because I will compare these two models in the end by their AIC and AUC scores. With the low AIC score, the model will be more accurate. The AUC score is from 0 to 1 if the AUC is closer to 1 that means this model is more correct. From the code we can see that the AIC for model 1 is 19005 and AIC for model 2 is 19028 this means model 1 is more accurate. The AUC for model 1 is 0.686 and the AUC for model 2 is 0.685, so model 1, the logistic regression is better than model two because it is more accurate and more correct than model 2. Therefore, I will use model 1 to predict the 2019 Canadian Federal Election.

Individual Level:

$$log(\frac{y_i}{1-y_i}) = \beta_{0i} + \beta_{gen} * x_{gen} + \beta_{pro} * x_{pro} + \beta_{gage} * x_{gage} + \epsilon$$

In this model we look at each observation individually, the y_i represents people in the i^{th} group whose vote choice is the Liberal party, and they will vote for Justin Trudeau. Next, β_0 is the i^{th} group that intercept point y_i. Furthermore, β_i also is the coefficients of the corresponding explanatory variable. So, if there is one unit increase in the variable there will be a β_i increase in log-probability that will vote for the Liberal party or in other words Justin Trudeau.

Level 2: Group Level:

$$\beta_{0i} = r_{00} + r_{0i} * W_j + u_{0i}$$

In this model, I grouped up all observations by cells which means they will be group up by age and gender. The r_{0i} in the equation is the log-probability of i^{th} group of observations that vote for the Liberal party. Next, r00 is a constant variable it is the intercept point with the dependent variable in this model. Finally, both ϵ and u_{0j} is the expected error in each model, because they are both following a normal distribution that means their mean is 0.

```
sum(groupby_age_gender$alp_predict)
```

```
[1] 0.3630248
```

Post-Stratification

In this part, I did the post-stratification analysis to estimate the percentage of people who vote for Justin Trudeau and Andrew Scheer. I grouped up observations by groupage and sex then created another variable called cells. One example in cells could be females 20 to 35 or male 65 or over. To win this election Justin Trudeau should win the majority of seats in the House of Commons. I used the census dataset to predict the result, first I count the number of province and education observations within each cell group. Next, I use faction to predict and got the result of the 2019 Canadian Federal Election. However, keep in mind that the result of this prediction, is the proportion of the total number of seats that a political party has in the House of Commons and there are five political parties in Canada and the party that gets the most proportion of seats wins.

Results

Result of the post-stratification prediction

Finally, from the post-stratification analysis the yhatˆps is 0.363 this means the result of the 2019 Canadian Federal Election from my prediction is the Liberal Party has majority seats of 36.3 percent of seats which is 122 seats in the House of Commons. Therefore, the Liberal party leader Justin Trudeau will be the Canadian Prime Minister and win the election. According to Wikipedia, the actual result of the 2019 Canadian Federal Election is the Liberal party won 39.5 percent seats in the House of Commons which is 184 seats and the Conservative party won 99 seats. Although there expects some error from my prediction, the Liberal party still has the majority of seats in the house of commons. Therefore, this model successfully predicts the actual result of the 2019 Canadian Federal Election.

Discussion

In this assignment, I simulated two models, one logistic regression model, and one multilevel regression model, by comparing the AIC and AUC values of those two models shows that the first model is more accurate with a lower AIC score, and higher correctness with an AUC score that closer to one. Then I use post-stratification to simulate the voting and made a prediction. Finally, I got the yhatˆps equal to 0.363 which means the Liberal party has the most seats in the House of Commons and wins the 2019 Canadian Federal Election.

However, there is a weakness of this model when I comparing to the actual result of the 2019 Canadian Federal Election there are some errors that need to be considered. One of the reason can be I removed three other parties from the beginning to simulate the logistic regression model, this is not good because we should count all of the parties, maybe people from other party but still vote for Justin Trudeau or Andrew Scheer then I will be losing count of those people. Furthermore, by applying the logistic regression model after I only can get one yhatˆps value and five parties are running for the Federal Election so I could not calculate the number of seats that each party has in the House of Commons.

References

1. Canadian federal election. (2020). Retrieved 22 December 2020, from https://en.wikipedia.org/wiki/2019_Canadian_federal_

2. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset] http://www.ces-eec.ca/

3. SDA. (2017). General social survey on Family (cycle 31). Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm