Differences in 2019 Canadian Federal Election if 'everyone' had voted

Yanrong Huo 1004720965

December 20th

Github link

https://github.com/huoyanro/Final-report.git

Abstract

This report is to detect if all eligible voters voted, wether there is any difference to the result of 2019 Canadian Federal Election. Two datasets have been used to perform multiple linear regression model with post-stratification and to do further analysis. Based on the result of analysis, a predicted voting result will be provided and the difference will be detected.

Keywords

2019 Canadian Federal Election, Data cleaning, Logistic regression model, Post-stratification, $\hat{Y}^{PS}$

Introduction

From an article 'Compulsory Voting' published by the International Institute for Democracy and Electoral Assistance (International IDEA), it is known that in some countries, voting at elections has been made compulsory and has been regulated in the national constitutions and electoral laws [IDEA, 2020, para.1], however it is not mandatory in Canada. According to CBC News, with 17.9 million people getting to the polls, turnout for Canada's 2019 general election was 65.95% of eligible voters [CBC News, 2019, para.1], which means 34.05% of eligible Canadian voters have not been involved. With 34.05% of votes missing, perhaps the result of the 2019 Canadian Federal Election could have been different. In this case, this project will determine whether there are any change/difference in results if all eligible voters voted.

Two datasets will be used for this investigation. The first one is the Canadian Election Study (CES) dataset which is a series of large-scale surveys that have been conducted in the lead-up to, and immediately after, each Canadian federal election since 1965 [CES package], this dataset will be used as the survey data. And the second one is from Statistics Canada called 'Highest level of educational attainment (general) by sex and selected age groups' which will be used as the census data. This dataset provides education data from the Census of Population, for various geographic levels and census years [Government of Canada, S, 2017, para.1]. With these two datasets, a multiple linear regression model with post-stratification will be performed starting from creating cells based on demographics from the census dataset. And in the end, YPS will be used for further analysis.

In the methodology section, details of the model performed using multiple linear regression and post-stratification will be shown. Results of $\hat{Y}^{PS}$ and any tables or pictures with statistical analysis will be shown in the results section. Conclusions drawn from this project will be presented in the discussion section.

Methodology

Data

```
## # A tibble: 6 x 620
##   cps19_StartDate     cps19_EndDate       cps19_ResponseId cps19_consent
##   <dttm>              <dttm>              <chr>            <fct>
## 1 2019-09-13 08:09:44 2019-09-13 08:36:19 R_1OpYXEFGzHRUp~ I consent to~
## 2 2019-09-13 08:39:09 2019-09-13 08:57:06 R_2qdrL3J618rxY~ I consent to~
## 3 2019-09-13 10:01:19 2019-09-13 10:27:29 R_USWDAPcQEQiMm~ I consent to~
## 4 2019-09-13 10:05:37 2019-09-13 10:50:53 R_3IQaeDXy0tBzE~ I consent to~
## 5 2019-09-13 10:05:52 2019-09-13 10:32:53 R_27WeMQ1asip2c~ I consent to~
## 6 2019-09-13 10:10:20 2019-09-13 10:29:45 R_3LiGZcCWJEcWV~ I consent to~
## # ... with 616 more variables: cps19_citizenship <fct>, cps19_yob <fct>,
## #   cps19_yob_2001_age <fct>, cps19_gender <fct>, cps19_province <fct>,
## #   cps19_education <fct>, cps19_demsat <fct>, cps19_imp_iss <chr>,
## #   cps19_imp_iss_party <fct>, cps19_imp_iss_party_7_TEXT <chr>,
## #   cps19_imp_loc_iss <chr>, cps19_imp_loc_iss_p <fct>,
## #   cps19_imp_loc_iss_p_7_TEXT <chr>, cps19_interest_gen_1 <dbl>,
## #   cps19_interest_elxn_1 <dbl>, cps19_v_likely <fct>, cps19_v_likely_pr <fct>,
## #   cps19_votechoice <fct>, cps19_votechoice_7_TEXT <chr>,
## #   cps19_votechoice_pr <fct>, cps19_votechoice_pr_7_TEXT <chr>,
## #   cps19_vote_unlikely <fct>, cps19_vote_unlikely_7_TEXT <chr>,
## #   cps19_vote_unlike_pr <fct>, cps19_vote_unlike_pr_7_TEXT <chr>,
## #   cps19_v_advance <fct>, cps19_v_advance_7_TEXT <chr>, cps19_vote_lean <fct>,
## #   cps19_vote_lean_7_TEXT <chr>, cps19_vote_lean_pr <fct>,
## #   cps19_vote_lean_pr_7_TEXT <chr>, cps19_2nd_choice <fct>,
## #   cps19_2nd_choice_7_TEXT <chr>, cps19_2nd_choice_pr <fct>,
## #   cps19_2nd_choice_pr_7_TEXT <chr>, cps19_not_vote_for_1 <fct>,
## #   cps19_not_vote_for_2 <fct>, cps19_not_vote_for_3 <fct>,
## #   cps19_not_vote_for_4 <fct>, cps19_not_vote_for_5 <fct>,
## #   cps19_not_vote_for_6 <fct>, cps19_not_vote_for_7 <fct>,
## #   cps19_not_vote_for_8 <fct>, cps19_not_vote_for_9 <fct>,
## #   cps19_not_vote_for_7_TEXT <chr>, cps19_fed_gov_sat <fct>,
## #   cps19_party_rating_23 <dbl>, cps19_party_rating_24 <dbl>,
## #   cps19_party_rating_25 <dbl>, cps19_party_rating_26 <dbl>,
## #   cps19_party_rating_27 <dbl>, cps19_party_rating_28 <dbl>,
## #   cps19_lead_rating_23 <dbl>, cps19_lead_rating_24 <dbl>,
## #   cps19_lead_rating_25 <dbl>, cps19_lead_rating_26 <dbl>,
## #   cps19_lead_rating_27 <dbl>, cps19_lead_rating_28 <dbl>,
## #   cps19_cand_rating_23 <dbl>, cps19_cand_rating_24 <dbl>,
## #   cps19_cand_rating_25 <dbl>, cps19_cand_rating_26 <dbl>,
## #   cps19_cand_rating_27 <dbl>, cps19_cand_rating_28 <dbl>,
## #   cps19_lr_scale_bef_1 <dbl>, cps19_lr_parties_1 <dbl>,
## #   cps19_lr_parties_2 <dbl>, cps19_lr_parties_3 <dbl>,
## #   cps19_lr_parties_4 <dbl>, cps19_lr_parties_5 <dbl>,
## #   cps19_lr_parties_6 <dbl>, cps19_lr_scale_aft_1 <dbl>,
## #   cps19_lead_int_113 <fct>, cps19_lead_int_114 <fct>,
## #   cps19_lead_int_115 <fct>, cps19_lead_int_116 <fct>,
## #   cps19_lead_int_117 <fct>, cps19_lead_int_118 <fct>,
## #   cps19_lead_int_119 <fct>, cps19_lead_int_120 <fct>,
## #   cps19_lead_strong_113 <fct>, cps19_lead_strong_114 <fct>,
## #   cps19_lead_strong_115 <fct>, cps19_lead_strong_116 <fct>,
## #   cps19_lead_strong_117 <fct>, cps19_lead_strong_118 <fct>,
## #   cps19_lead_strong_119 <fct>, cps19_lead_strong_120 <fct>,
```

```
## #   cps19_lead_trust_113 <fct>, cps19_lead_trust_114 <fct>,
## #   cps19_lead_trust_115 <fct>, cps19_lead_trust_116 <fct>,
## #   cps19_lead_trust_117 <fct>, cps19_lead_trust_118 <fct>,
## #   cps19_lead_trust_119 <fct>, cps19_lead_trust_120 <fct>,
## #   cps19_lead_cares_113 <fct>, cps19_lead_cares_114 <fct>,
## #   cps19_lead_cares_115 <fct>, cps19_lead_cares_116 <fct>, ...
```

In this project, two datasets have been used, the first one is the Canadian Election Study (CES) dataset as mentioned in the introduction, which is an online survey with population Canadian citizens and permanent residents, aged 18 or older [Laura B., 2019, p.5]. This dataset includes 620 variables and 37822 interviews (observations). Five variables from this dataset have been selected to make survey data, they are cps19_age, cps19_gender, cps19_province, cps19_education and cps19_votechoice where cps19_age describes respondent age in years, recoded based on their year of birth [Laura B., 2019, p.23], cps19_gender describes sex of interviewers, cps19_province describes provinces or territories interviewers are currently living in, cps19_education describes the highest level of education interviewers have completed and cps19_votechoice describes the parties interviewers think they will vote for.

```
## Rows: 252
## Columns: 20
## $ Geographic.code                                                                <int> .
## $ Geographic.name                                                                <chr> .
## $ Global.non.response.rate                                                       <dbl> .
## $ Data.quality.flag                                                              <int> .
## $ Age                                                                            <chr> .
## $ Sex                                                                            <chr> .
## $ Total...Highest.certificate..diploma.or.degree..2016.counts.                   <int> .
## $ No.certificate..diploma.or.degree..2016.counts.                                <int> .
## $ Secondary..high..school.diploma.or.equivalency.certificate..2016.counts.       <int> .
## $ Apprenticeship.or.trades.certificate.or.diploma..2016.counts.                  <int> .
## $ College..CEGEP.or.other.non.university.certificate.or.diploma..2016.counts.    <int> .
## $ University.certificate.or.diploma.below.bachelor.level..2016.counts.           <int> .
## $ University.certificate..diploma.or.degree.at.bachelor.level.or.above..2016.counts. <int> .
## $ Total...Highest.certificate..diploma.or.degree....distribution.2016.           <int> .
## $ No.certificate..diploma.or.degree....distribution.2016.                        <dbl> .
## $ Secondary..high..school.diploma.or.equivalency.certificate....distribution.2016. <dbl> .
## $ Apprenticeship.or.trades.certificate.or.diploma....distribution.2016.          <dbl> .
## $ College..CEGEP.or.other.non.university.certificate.or.diploma....distribution.2016. <dbl> .
## $ University.certificate.or.diploma.below.bachelor.level....distribution.2016.   <dbl> .
## $ University.certificate..diploma.or.degree.at.bachelor.level.or.above....distribution.2016. <dbl> .
```

The second dataset is highest level of educational attainment (general) by sex and selected age groups from Statistics Canada, it provides counts and percentage distributions for various geographic levels by highest level of educational attainment, sex and selected age groups for the 2016 Census [Government of Canada, S, 2017, para.2]. This dataset includes 20 variables and 252 observations. Ten variables have been selected to make census data (other variables abandoned are either very similar to those variables kept or not corresponding to variables in survey data), after pivoting, 5 variables have been kept including age, sex, geographic.name, education and count. Age, sex, geographic.name and education describe exactly same as cps19_age, cps19_gender, cps19_province and cps19_education, count describes the number of interviewers under same age, sex, living and education backgrounds conditions. Admittedly, this is not a dataset with large number of observations, its variables are pretty clear and appropriate.

About data cleaning, for survey data, the variable gender has been modified by transferring 'A woman' to 'Female', 'A man' to 'Male'; the variable age has been modified by transferring real age (numbers) to age groups (factors) and the variable education has been transferred from long sentences to short phrases. For census data, all variables except for 'count' have been modified by transferring to same content as in survey data.

Model

In this project, a multilevel logistic regression model is built by R to predict the probability of response variable, which, in our case, is the probability that an eligible voter votes for the Liberal party. About the variable vote choice, there are 7 parties in the real data, however, it is known that the Liberal party and the Conservative party have major votes and the winner must comes from these two, so other parties except for these two are just deleted. Because a multilevel regression model is performed, cells needs to be partitioned first. In this model, cells are created using variables gender and age, there are 8 cells in both survey and census data. Predictors in this model include education, age and gender, the variable province and cells will be performed as a random coefficient . The formula of this model is as follows: $\mathbf{log}(\frac{\hat{\mathbf{ProbLiberal}}}{1-\hat{\mathbf{ProbLiberal}}})$ = -0.73015 + a_j - 0.50438**EducationCollege** - 0.59872**EducationHigh school** - 0.04322**EducationNo certificate** - 0.09499**EducationSome university** - 0.15064**Age35 to 44** - 0.28625**Age45 to 54** - 0.47871**Age55 to 64** - 0.01775**GenderMale** Where $\hat{\mathbf{ProbLiberal}}$ is the predicted probability that an eligible voter votes for the Liberal party. We can get the probability from the log-odds $\mathbf{log}(\frac{\hat{\mathbf{ProbLiberal}}}{1-\hat{\mathbf{ProbLiberal}}})$, which is the response variable of this model, through a little bit of basic mathematics. In this model, -0.73015 is the coefficient baseline (intercept) which means if all variables stay 0, the log-odds will be -0.73015.a_j is a random coefficient which will be presented later. If the voter is female, age 25 to 34, a_j will equal to 6.079349e-04; if the voter is female, age 35 to 44, a_j will equal to 2.859314e-04; if the voter is female, age 45 to 54, a_j will equal to -8.022212e-05; if the voter is female, age 55 to 64, a_j will equal to -3.620303e-04; if the voter is male, age 25 to 34, a_j will equal to 3.171194e-04; if the voter is male, age 35 to 44, a_j will equal to -4.942502e-05; if the voter is male, age 45 to 54, a_j will equal to -1.506599e-04; if the voter is male, age 55 to 64, a_j will equal to -1.779541e-04. GenderMale is a dummy variable which means that if the voter is a female, this variable will equal to 0; if the voter is a male, this variable will equal to 1. Also, EducationCollege, EducationHigh school, EducationNo certificate, EducationSome university, Age35 to 44, Age45 to 54 and Age55 to 64 are dummy variables, if the voter satisfies their conditions, these variables will equal to 1, otherwise, they will equal to 0. For example, if the voter is a female, age 46, has no certificate, her log-odds will be -0.73015 - 8.022212e-05 - 0.04322 - 0.28625 = - 1.0597.

For this part, two models was built first, they have partitioned same cells, one is the model that has been described above, the other one uses education and cells as random coefficient a_j. The reason why the above model has been chosen is that the ROC curves of both models have been plotted, AUC (area under the curve) of the above model is 0.692, which means there is 69.2% probability that this model predict the true result. But AUC of the other model is 0.688, which is smaller, thus worse.

Post-Stratification

The project is continued by conducting a post-stratification analysis to predict the probability for voting the Liberal party instead of the Conservative party. Post-stratification means that the weights are adjusted so that the weighted totals within mutually exclusive cells equal the known population totals [Kolenikov, S, 2016, para.2]. This method is used because it can decrease bias resulting from nonresponse and underrepresented groups in the population. Through research, it is known that, according to the election policy of Canada, there are 13 provinces in Canada with 338 electoral districts, every districts will have to decide one winner for the election and votes for this party. After all electoral districts have been voted, the party who gets the most votes will be the final winner.

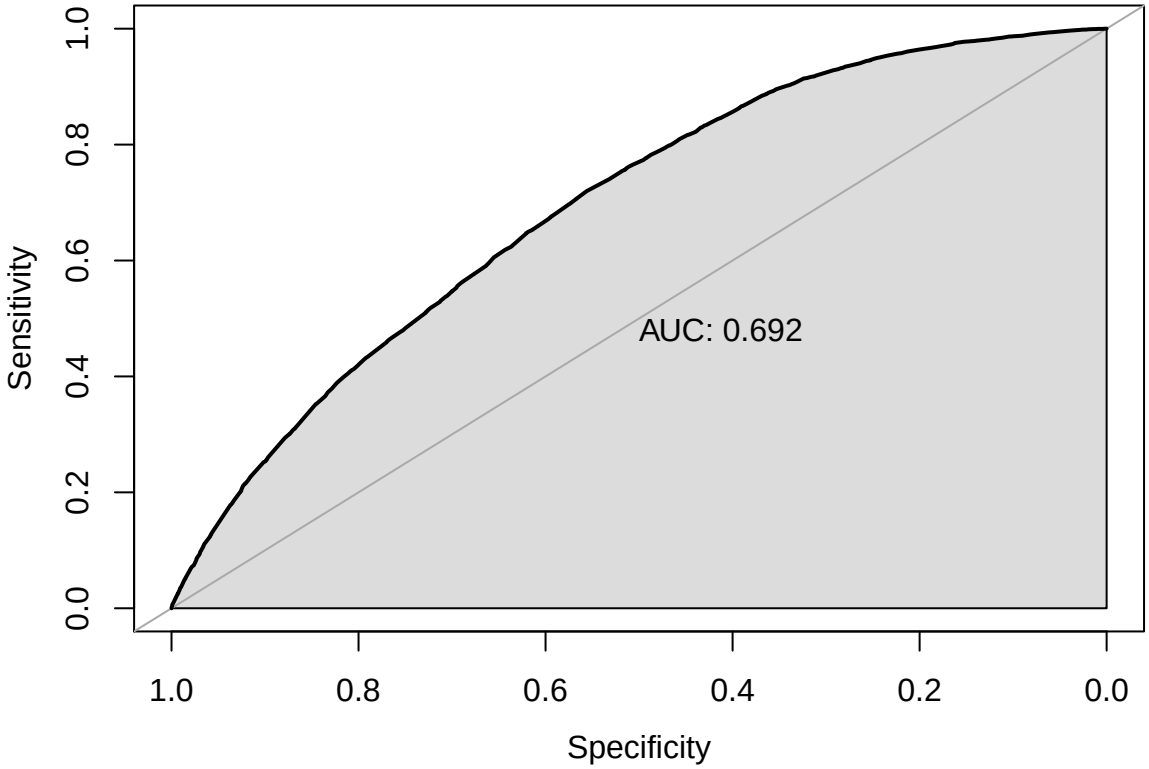
In this project, a logistic regression model has been performed using survey data, it will be applied to the census data to predict the probability of the vote in each province. When we get the probabilities, multiply them with the number of electoral districts of each province, thus the predicted number of votes for both Liberal and Conservative party in each province can be obtained. We add them up separately, the party who gets more votes will be the winner. The variables count (which describes the number of voters under certain conditions) and province in census data will be used to predict votes in each province.

Then we sum up the multiple of number of voters under certain conditions and their predicted probability voting for Liberal party and divide this summation by total population to obtain $\hat{Y}^{PS}$, this process can be summarized by formula: $\frac{\sum_j N_j \hat{y_j}}{\sum_j N_j}$ where N_j is the number of voters under certain conditions, that is, the

variable count; $\hat{y_j}$ is the predicted probability voting for Liberal party.

Results

```
## Setting levels: control = Conservative Party, case = Liberal Party
## Setting direction: controls < cases
## Area under the curve: 0.6924
```



In this part, ROC curve is plotted and thus AUC (area under the curve) is obtained. AUC equals to 0.692, which means that there is 69.2% probability that this model predict the true result. Since the datasets used are real, 69.2% is a pretty high score meaning that the logistic regression model we performed is good.

After the whole process of post-stratification analysis, the predicted number of electoral districts that will vote for Liberal party is 214, that is, 63.3% of electoral districts will vote for it. The predicted number of electoral districts that will vote for Conservative party is 124, that is, 36.7% of electoral districts will vote for it.

Also, through the method mentioned above about obtaining $\hat{Y}^{PS}$, it is calculated that $\hat{Y}^{PS} = 0.4949419$ which means there are 49.5% eligible voters vote for Liberal party and 50.5% eligible voters vote for Conservative party based off the post-stratification analysis modeled by the multilevel logistic regression model we performed.

Discussion

Summary

In this project, two datasets (Canadian Election Study and Highest level of educational attainment (general) by sex and selected age groups) are used as survey data and census data to perform multilevel logistic regression model with post-stratification to predict the result of 2019 Canadian Federal Election and see if there is any difference between the predicted and actual results.

Conclusion

As mentioned above, $\hat{Y}^{PS} = 0.4949419$ which means that there are 49.5% eligible voters vote for Liberal party, in this case, the winner of the election would be the Conservative party. And the predicted proportion of electoral districts that will vote for Liberal party is 63.3% which means the winner of the election would be the Liberal party. It can be seen that 49.5% is different from 63.3% and that will lead to a difference in the result of election. However, that is pretty reasonable, since $\hat{Y}^{PS}$ represents the predicted proportion of votes of voters but 63.3% represents the predicted proportion of votes of electoral districts. If all voters in a province votes for Conservative party but that province has only 1 electoral district, then no matter how many voters are there in that province, the Conservative party will only get 1 vote from that province. In this case, 63.3% is the real proportion of votes Liberal party will get, but $\hat{Y}^{PS}$ doesn't mean as much as the number 63.3%.

In this case, we can get that after the whole process of the multilevel logistic regression model with post-stratification analysis, the predicted winner of 2019 Canadian Federal Election would be the Liberal party, just same as the actual result. There is no differences in 2019 Canadian Federal Election if 'everyone' had voted.

Weakness

Talking about the limitations in this study, first, the census data used is not large enough, there are only 252 observations in the raw data, after the process of data cleaning, the real census data used for further analysis has only 6 variables and 520 observations. When the number of observations is not big enough, the analysis will not be random enough and that will lead to a wrong result. Second, the model is chosen from two based on AUC, however, if AIC and BIC are compared, it can be seen that AIC and BIC of the model abandoned are smaller which make it a better model, at the same time, AIC and BIC of the model chosen are bigger which make it worse. Third, about the predictors of the model chosen, some of them (e.g EducationNo certificate, GenderMale) have very large p-value which means they are bad predictors for predicting the probability of votes, however, that cannot be abandoned in this model. And finally, when we do data cleaning, for the variable vote choice, because we want to perform a logistic regression model, we must make it a dummy variable that has only two outputs, in this case, only Liberal party and Conservative party are kept, all other parties are removed. That is an appropriate method for logistic model but not perfect. In the future, if other kinds of model that can analyze multiple output are learned, they will fit better in this study.

Next Steps

For future improvements, first, I should try to find a larger census data with more observations to make the analysis and prediction more random thus get a more accurate result of voting. If it's possible, a post-hoc analysis can be done to improve the prediction in future election. Also, I should try find other predictors that fit better (have smaller p-value in the model) and try more models to see if there is a better model with smaller AIC and BIC values and bigger AUC.

Reference

1. Compulsory Voting. (n.d.). Retrieved December 09, 2020, from https://www.idea.int/data-tools/data/voter-turnout/compulsory-voting
2. Canadian election drew nearly 66% of registered voters | CBC News. (2019, October 22). Retrieved December 09, 2020, from https://www.cbc.ca/news/canada/voter-turnout-2019-1.5330207
3. Government of Canada, S. (2017, November 27). Education Highlight Tables, 2016 Census. Retrieved December 09, 2020, from https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/edu-sco/index-eng.cfm
4. Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection. [dataset]

5. Kolenikov, S. (2016, August 01). Post-stratification or non-response adjustment?: Published in Survey Practice. Retrieved December 19, 2020, from https://www.surveypractice.org/article/2809-post-stratification-or-non-response-adjustment

6. List of Canadian federal electoral districts. (2020, November 16). Retrieved December 18, 2020, from https://en.wikipedia.org/wiki/List_of_Canadian_federal_electoral_districts