Informative Title Name

STA304 - Assignment 2

69: Woojin Park and David Pham

May 28, 2021

Introduction

In this report, we seek to predict the popular political party in the upcoming 2023 Canadian Federal Election. We will achieve this by constructing a suitable model using the 2019 Canadian Election Study phone (survey) data, and post-stratifying it with the General Social Survey (census) data.

Politics and democracy are crucially important aspects in Canada because they determine our government and help shape the society we live in (*). The analysis can predict which political party is more likely to win the election. Thus, people can have an idea of which candidate in the party will likely get elected by certain aspects we have used in our analysis. Furthermore, this analysis can set a global precedence for all democratic countries in the world and have them also employ these predictive models.

Predicting the next political term is a very interesting subject, especially in the United States of America. To elaborate, Professor Allan Lichtman, a professor at the American University in Washington, D.C has created a flawless predictive model using historical data to correctly predict the next president of the U.S.A; in fact, it is so flawless, it has correctly predicted every election since 1984 (*). Since we are using the GSS (Global Service Survey) data, there is a high chance the analysis in our report might occur.

On the political spectrum, we can generally divide peoples' views into a left (Liberal) view, and a right (Conservative) view. Liberals generally believe in governmental action that promotes equality for all, while conservatives believe in personal responsibility and freedom (*).

Finally, we present our research question:
Is there a correlation with age, gender, province, highest level of education, religion, and income as predictors for whether or not an individual votes Liberal?

The model in question will be a logistic regression model. By definition, logistic regression must come with assumptions, similarly to ordinary least squares. These will be the only assumptions we are making, and will be clarified in the Methods section.

Data

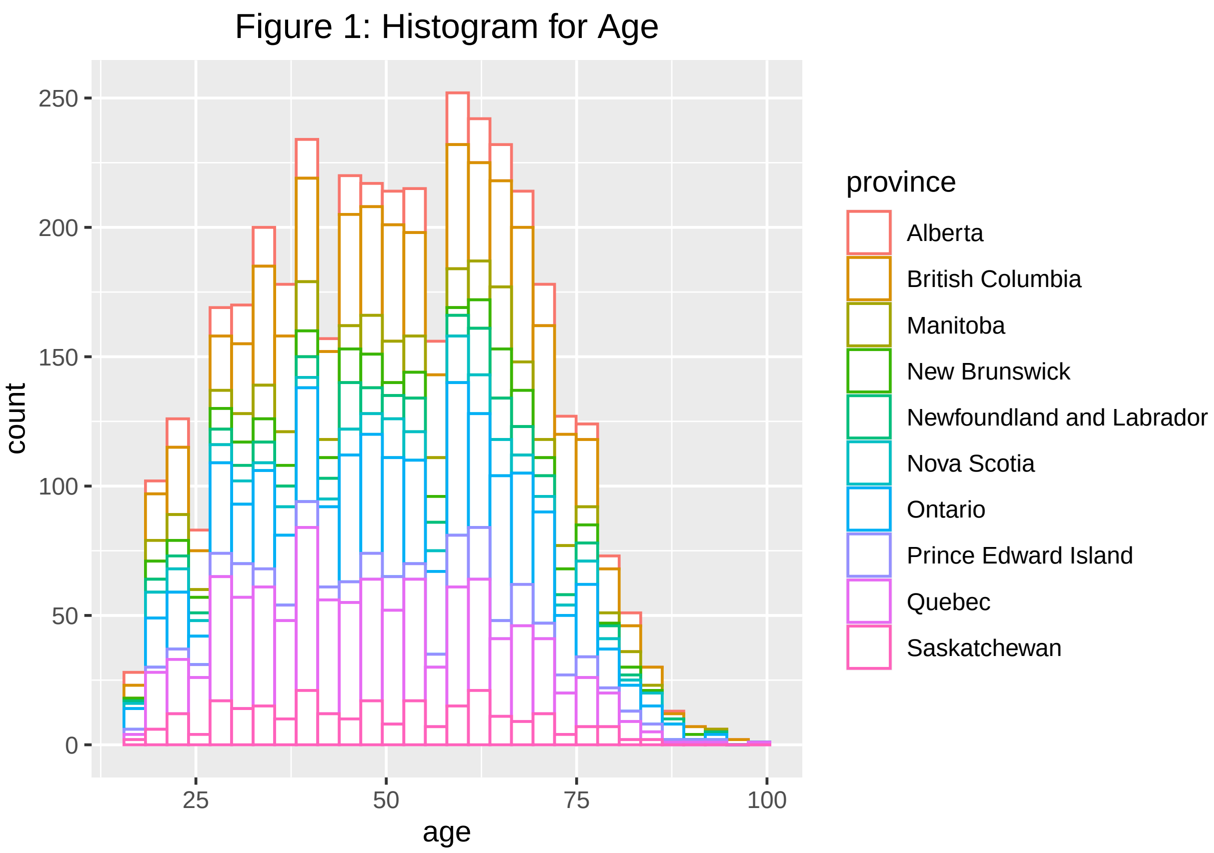
The data we will be using to create a regression model is from the 2019 Canadian Election Survey Phone data. The CES organization strives to learn more about Canadians' political opinions and behaviour and provide transparency about the thoughts of citizens (*). The information was initially created by calling eligible voting Canadians, and asking them about their thoughts on the political candidates and the state of the country.

The original .csv file for the CES data was quite messy and had many variables that were not used in this report. To clean the data, we first mutate the desired variables into distinguishable names that could be used in the model, such as age and gender. A few if/else clauses were used to make the data tidy. Next, we turn the appropriate variables into factor types so that the results can be properly interpreted. Finally, we select the variables that we thought were most appropriate for our model. A similar process was done for the GSS data.

In summary, these are the most important variables collected from the survey:

- `age`: Age based on year of birth (discrete, numeric variable).
- `gender`: The gender of the respondent (nominal, categorical variable).
- `province`: Indicates which province/territory the respondent is currently living in. The labels are defined as so:
  - (1) Newfoundland and Labrador
  - (2) Prince Edward Island
  - (3) Nova Scotia
  - (4) New Brunswick
  - (5) Quebec
  - (6) Ontario
  - (7) Manitoba
  - (8) Saskatchewan
  - (9) Alberta
  - (10) British Columbia
  - (There were no respondents from the Northwest Territories, Yukon, or Nunavut.)
- `education`: Indicates the highest level of education the respondent has received. The labels are defined as such:
  - (-9): Don't know
  - (-8): Refused
  - (1): No schooling
  - (2): Some elementary school
  - (3): Completed elementary school
  - (4): Some secondary/high school
  - (5): Completed secondary/high school
  - (6): Some technical, community college, CEGEP, College
  - (7): Completed technical, community college, CEGEP, College
  - (8): Some university
  - (9): Bachelor's degree
  - (10): Master's degree
  - (11): Professional degree or doctorate
- `religion`: The religion that the respondent follows. There are a plethora of religions out there in the world, so we will not be writing them all here. If you would like to see an exhaustive list of all the recorded responses, the CES has provided documentation for all survey data. For our report, we are interested in whether or not a respondent follows a religion. If they are religious, they will have a value of *1* in the data frame. Otherwise, the value for this variable is *0.* (cite)
- `income`: The income of a respondent before taxes. This variable has been turned into a factor variable and is organized into income brackets.

The mean age is 51, and the standard deviation is 16.836. This is the only numeric variable, as it is nonsensical to find the mean and spread of categorical data; therefore, we will omit the rest.



This histogram shows all ages of respondents who participated in the survey, colored by province. There are good distributions of respondents from multiple age groups, but we see that it is more condensed in the 30-60 range. Furthermore, within each province, the age distribution seems to divided fairly well; each province is fairly represented in the sample. With some hypothesis testing and confidence intervals, we could infer that the true mean lays somewhere around 50 as well.

All analysis for this report was programmed using `R version 4.0.2`.

Methods

Next, we will focus on the model used to predict whether or not a respondent will vote for the Liberal party. We believe that logistic regression is the most appropriate for this model because the response we are interested in consists of binary outcomes.

We have 4 main assumptions (*):

1. **Binary Response**: The response variable can only have two possible responses. This is necessary for logistic regression and is satisfied in our report, as we will see later on.

2. **Independence**: The observations must be independent of one another. This is satisfied due to the nature of the CES data.

3. **Variance Structure**: By definition of a binomial random variable, its variance is $np(1-p)$, implying that variance is highest when $p = 0.5$.

4. **Linearity**: The log of the odds ratio, $log(\frac{p}{1-p})$, must be a linear function of x. This is similar to linear regression, but from a logistic standpoint.

Model Specifics

We have used the Akaike Information Criterion (AIC) to help us remove unnecessary variables in our report. After running extensive summaries on the model, the predictors that the AIC kept were age, province, and income of the respondent. Our lowest value of the criterion was 4110. Hence, our final model looks like this:

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 x_{age}+\beta_2 x_{province2}+\beta_3 x_{province3}+\beta_4 x_{province4}+\beta_5 x_{province5}+\beta_6 x_{province6}+\beta_7 x_{province7}+\beta_8 x_{province8}$$

$$+\beta_9 x_{province9} + \beta_{10} x_{province10} + \beta_{11} x_{income2} + \beta_{12} x_{income3} + \beta_{13} x_{income4} + \beta_{14} x_{income5} + \beta_{15} x_{income6} + \epsilon$$

where:

- $log(\frac{p}{1-p})$ is the logit function with probability p
- β_0 is the intercept term
- $\beta_1 x_{age}$ is a term where the β_1 coefficient represents the change in log odds for every one unit increase in x_{age} (holding all other predictors constant)
- $\beta_i x_{provincei}, i = 2, 3, ..., 10$ are the terms representing province, where dummy variables have been encoded. If province i is being observed, all other values go to 0.
- $\beta_j x_{incomej}, j = 2, 3, 4, 5, 6$ are the terms representing income, where dummy variables have once again been implemented. If income bracket j is being observed, all other values go to 0. For values of j:
  - j = 1: Respondent makes less than $25,000.
  - j = 2: Respondent makes between $25,000 and $49,999
  - j = 3: Respondent makes between $50,000 and $74,999
  - j = 4: Respondent makes between $75,000 and $99,999
  - j = 5: Respondent makes between $100,000 and $124,999
  - j = 6: Respondent makes $125,000 or more
- ϵ is the error term in the model.

Post-Stratification

Multilevel regression with post stratification is a method to get weighted average estimate from all combination of attributes which is referred to as "cell" [4]. Each cell will be modeled by different variables. In our case we are going to use age, province and income to model each cell.

To calculate these weights the we are going to use the equation below

$$y^{ps} = \frac{\sum N_j \hat{y_j}}{\sum N_j}$$

Here $\hat{y_j}$ is the estimate of vote proportion at cell j. N_j is the population size of the jth cell.

We are going to estimate our y^{ps} (proportion of voters who are voting for demographics) for each cell by using multi-level modelling. Then we can calculate these weights at the population level by using the GSS data to predict the proportion of votes in the upcoming election.

The general work flow is as below.[5] 1. Read in the poll 2. Model the poll 3. Read in the poststratification data 4. Apply the model to the poststratification data

All analysis for this report was programmed using `R version 4.0.2`.

Results

Finally, we interpret our findings. To begin, we take a look at the summary of the GLM using the survey data, along with all associated variables, slope estimates, and p-values:

Table 1: Results of figures from Data section. [*]

Variable	Slope Estimate	p-value
Age	0.00836	0.00039
Province (B.C)	0.87737	0.00012
Province (MB)	0.95866	0.00024
Province (N.B)	1.11431	3.6e-05
Province (N&L)	1.37125	2.1e-07
Province (N.S)	1.33236	5.1e-07
Province (ON)	1.5276	8.1e-12
Province (PEI)	1.37905	1.7e-07
Province (QB)	1.18043	1.8e-07
Province (SAS)	0.19015	0.51271
Income Bracket 2	0.18886	0.1624
Income Bracket 3	0.28289	0.02356
Income Bracket 4	0.46032	0.00064
Income Bracket 5	0.19512	0.17298
Income Bracket 6	0.40145	0.00017

Age, every province but Saskatchewan, and the $75,000-$99,999 and $125,000+ income brackets were found to be statistically significant. To interpret this (*):

- For every one unit increase in age, the log-odds of voting for the Liberal party increases by 0.008.
- If you live any province except Saskatchewan or Alberta, the log-odds of voting Liberal increases by its respective slope estimate (with respect to Alberta as the base comparison).
- If a respondent makes $75,000-$99,999 or $125,000+ annually, the log-odds of voting Liberal increases by around 0.4, compared to a respondent that makes less than $25,000.

```
## # A tibble: 66 x 2
##      age liberal_predict
##    <dbl>           <dbl>
##  1    15           -1.71
##  2    16           -1.71
##  3    17           -1.70
##  4    18           -1.75
##  5    19           -1.75
##  6    20           -1.66
##  7    21           -1.62
##  8    22           -1.64
##  9    23           -1.57
## 10    24           -1.64
## # ... with 56 more rows
```

```
## # A tibble: 10 x 2
##    province                  liberal_predict
##    <fct>                               <dbl>
##  1 Alberta                             -2.35
##  2 British Columbia                   -1.45
##  3 Manitoba                           -1.38
##  4 New Brunswick                      -1.23
##  5 Newfoundland and Labrador          -0.978
##  6 Nova Scotia                        -1.01
##  7 Ontario                            -0.809
##  8 Prince Edward Island               -0.971
##  9 Quebec                             -1.18
## 10 Saskatchewan                       -2.16
```

```
## # A tibble: 6 x 2
##   income                liberal_predict
##   <fct>                           <dbl>
## 1 Less than $25,000               -1.43
## 2 $25,000 to $49,999              -1.22
## 3 $50,000 to $74,999              -1.15
## 4 $75,000 to $99,999              -1.00
## 5 $100,000 to $ 124,999           -1.27
## 6 $125,000 and more               -1.07
```

Conclusions

<Here you should give a summary of the Hypotheses, Methods and Results>

<Highlight Key Results.>

<Talk about big picture.>

<Comment on any Weaknesses.>

<End with a concluding paragraph to wrap up the report.>

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Juan Lopez-Martin, Justin H. Phillips. "Multilevel Regression and Poststratification Case Studies." Chapter 1 Introduction to MRP, 9 Feb. 2021, bookdown.org/jl5522/MRP-case-studies/introduction-to-mrp.html.

5. Alexander, Rohan. "Getting Started with MRP." Rohan Alexander, 3 Dec. 2019, rohanalexander.com/posts/2019-12-04-getting_started_with_mrp/.