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. 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.

Data

<Type here a paragraph introducing the data, its context and as much info about the data collection process that you know.> 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

Table 1: Means and standard deviations of all variables in the survey data.

Variable Name	Mean	Standard Deviation
age	50.89	16.836

- `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.

Here is the mean and standard deviation of all respondents' age:

This is the only numeric variable, as it is nonsensical to find the mean and spread of categorical data; so that has been omitted.

<Include a description of the numerical summaries. Remember you can use `r` to use inline R code.>

```
# Use this to create some plots. Should probably describe both the sample and population.
```

<Include a clear description of the plot(s). I would recommend one paragraph for each plot.>

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.

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}

- $\beta_i x_{provincei}, i = 2, 3, ..., 10$ 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

<In order to estimate the proportion of voters. . . ..>

<To put math/LaTeX inline just use one set of dollar signs. Example: $\hat{y}^{PS}$ >

$$include.your.mathematical.model.here.if.you.have.some.math.to.show$$

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

Results

<Here you present your results. You may want to put them into a well formatted table. Be sure that there is some text describing the results.>

<Note: Alternatively you can use the `knitr::kable` function to create a well formatted table from your code. See here: https://rmarkdown.rstudio.com/lesson-7.html.>

<Remember you can use `r` to use inline R code.>

<Include an explanation/interpretation of the visualizations. Make sure to comment on the appropriateness of the assumptions/results.>

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)