Informative Title Name

STA304 - Assignment 3

GROUP NUMBER: ADD YOUR NAMES HERE

November 5, 2021

Introduction

<Here you should have a few paragraphs of text introducing the problem, getting the reader interested/ready for the rest of the report.>

<Introduce terminology.>

<Highlight hypotheses.>

<Optional: You can also include a description of each section of this report as a last paragraph.>

Data

Data Collection Process

The Data is a collection of two datasets. The first dataset contains data from the General Social Survey on Family (cycle 31) on 2017. Canada's General Social Survey (GSS) program conducts annual survey covering one topic in depth (ciation). As such, this dataset contains mostly contains information pertaining to families. However, we will later investigate some common variables between this dataset and the Canada Election Study (CES) dataset. This will give us a list of factors in the general population that could be potentially associated to political affiliations

the Canada Election Study (CES) data which was collected in 2019. This data was collected from a questionnaire delivered to the people living in Canada through the Computer Assisted Telephone Interviewing(CATI). Phone calls were made potential interviewees during both the day and evening for every weekday. These questions asked some personal information such as their age and also political opinions such as "what is your opinion on Justin Trudeau?" (citation)

Data cleaning

<Type here a summary of the cleaning process (**only add in stuff beyond my original gss_cleaning.R code**). You only need to describe additional cleaning that you and your group did.>]

<Remember, you may want to use multiple datasets here, if you do end up using multiple data sets, or merging the data, be sure to describe this in the cleaning process and be sure to discuss important aspects of all the data that you used.>

The regression model will be constructed, will be applied to both datasets, therefore only variables that appeared in both the datasets could be used. One way was examine if there were common terms present in the columns for both the dataset. Unfortunately most the columns the CES dataset were just question numbers. However, each column also had a label that stated the question itself. For example column "q2" had the label "IN what year were you born?". Hence we collected common words that appeared in the columns of the GSS data and the labels of the CES data to find possible topics that were common in both datasets.

Then for each topic we search the columns for both the dataset to see if any two column were describing indentical or similar variables. These are the variables we found:

- Age of the person was recorded in both data with the same column names
- Gender was recorded in CES data as q3 and sex was recorded in the GSS data. Here we assumed the people who reported their gender to be male or female also would have the same sex as their gender. Hence we removed all genders that were not male or female and renamed the q3 column in the CES data to sex
- Both the datasets recorded the province a person lived in. To make the variable name clear, q4 in CES dataset was renamed to province
- Both datasets recorded how the level of importance for religion. This was column q63 in the CES data. The categorical values in both the datasets were the same with the exception of an additional value "Refused" being present in the CES data. We removed rows containing this value from the CES data and renames columns q63 to "religion_importance" to have the same column names in both datasets. Furthermore, we also removed missing rows with missing values in this column for both the datasets
-

<Include a description of the important variables.>

<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.>

Methods

<Include some text introducing the methodology, maybe restating the problem/goal of this analysis.>

Model Specifics

<I will (incorrectly) be using a linear regression model to model the proportion of voters who will vote for Donald Trump. This is a naive model. I will only be using age, which is recorded as a numeric variable, to model the probability of voting for Donald Trump. The simple linear regression model I am using is:>

$$y = \beta_0 + \beta_1 x_{age} + \epsilon$$

<Where y represents the .... β_0 represents....>

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)