Informative Title Name

STA304 - Assignment 2

69: Woojin Park and David Pham

May 28, 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

<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.

<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.>] The original .csv file 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.

<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.>

<Include a description of the important variables.> 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. If the variable is equal to -9 or -8, this means that the respondent did not know their income, or refused to answer the question (respectively).

Here is a table of all the variables, along with their respective means and standard deviations of the sample:

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

Variable Name	Mean	Standard Deviation
age	50.89	16.836
income	80330.776	111472.24

As a side note, the spread of income is incredibly high because many people did not answer this question; hence, there is some variability in this response. Furthermore, it is nonsensical to find the mean and spread of categorical data, so that has been omitted. Instead, here are some tables to display the frequency of the variables:

```
table(survey_data$gender)
```

```
##
##    0    1
## 1749 2272
```

<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:>

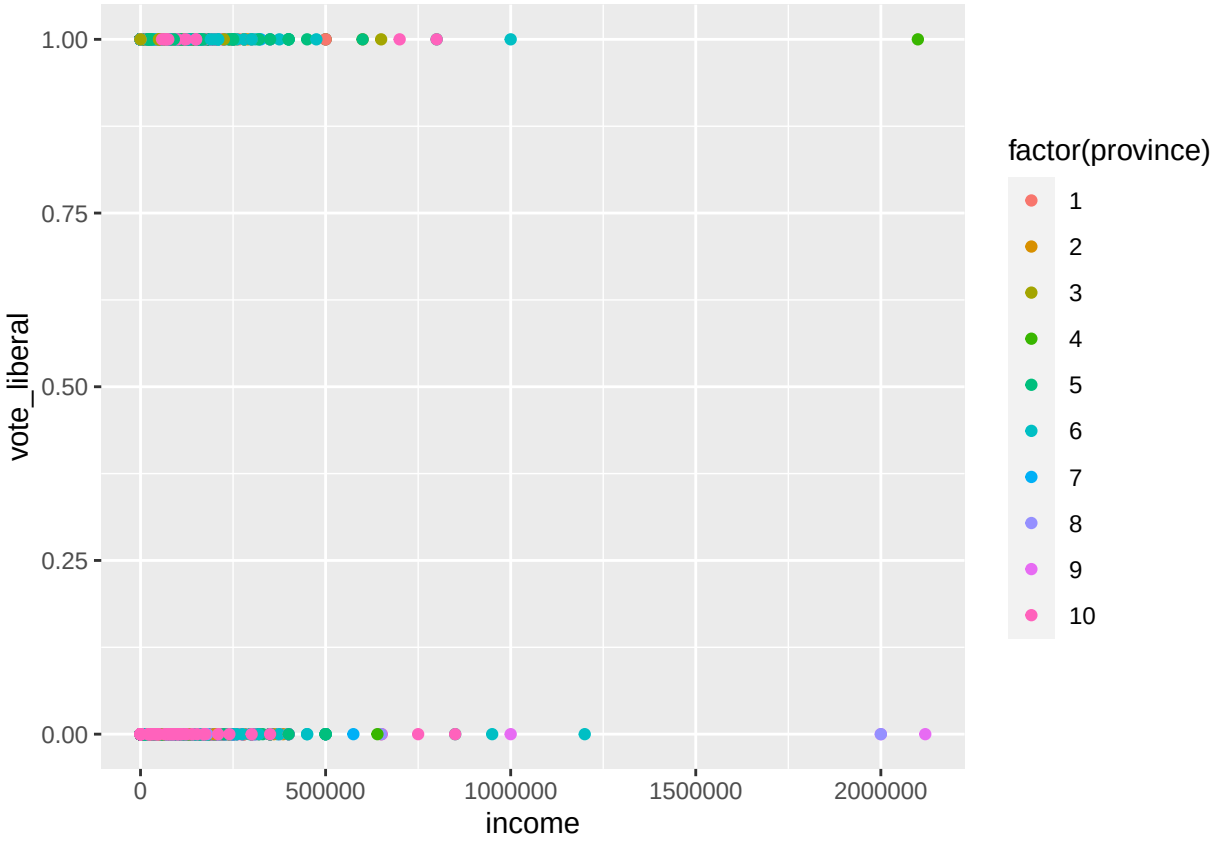
$$y = \beta_0 + \beta_1 x_{age} + \epsilon$$

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

```
# Creating the Model
model <- glm(vote_liberal ~ age + province + income + education, data=survey_data, family = binomial)

# back1 = step(model, direction = "backward")

ggplot(survey_data, aes(x = income, y = vote_liberal, color = factor(province))) + geom_point()
```




```
# Model Results (to Report in Results section)
summary(model)
```

```
##
## Call:
## glm(formula = vote_liberal ~ age + province + income + education,
##     family = binomial, data = survey_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.202  -0.770  -0.633  -0.370   2.450
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.37e+01   1.85e+02   -0.07   0.9410
## age          8.15e-03   2.37e-03    3.43   0.0006 ***
## province2   -5.17e-02   2.28e-01   -0.23   0.8204
## province3   -8.23e-02   2.30e-01   -0.36   0.7202
## province4   -3.24e-01   2.36e-01   -1.37   0.1694
## province5   -2.86e-01   1.84e-01   -1.55   0.1211
## province6    5.24e-02   1.80e-01    0.29   0.7708
## province7   -4.64e-01   2.25e-01   -2.07   0.0389 *
## province8   -1.18e+00   2.59e-01   -4.57  4.8e-06 ***
## province9   -1.42e+00   2.66e-01   -5.33  9.7e-08 ***
## province10  -5.72e-01   1.86e-01   -3.07   0.0021 **
```

```
## income       8.60e-07   3.27e-07    2.63   0.0086 **
## education-8  1.26e+01   1.85e+02    0.07   0.9457
## education1   1.35e+01   1.85e+02    0.07   0.9419
## education2   1.13e+01   1.85e+02    0.06   0.9511
## education3   1.22e+01   1.85e+02    0.07   0.9474
## education4   1.19e+01   1.85e+02    0.06   0.9486
## education5   1.21e+01   1.85e+02    0.07   0.9477
## education6   1.20e+01   1.85e+02    0.06   0.9483
## education7   1.20e+01   1.85e+02    0.06   0.9483
## education8   1.24e+01   1.85e+02    0.07   0.9466
## education9   1.26e+01   1.85e+02    0.07   0.9459
## education10  1.28e+01   1.85e+02    0.07   0.9449
## education11  1.23e+01   1.85e+02    0.07   0.9468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4254.7  on 3936  degrees of freedom
## Residual deviance: 4062.0  on 3913  degrees of freedom
##   (84 observations deleted due to missingness)
## AIC: 4110
##
## Number of Fisher Scoring iterations: 12
```

```
# OR
# broom::tidy(model)

### Don't show the results/output here...
```

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)