SOC6707 Intermediate Data Analysis, Winter 2021

Assignment 2

Due date: 5 April 11:59pm ET

Details

There are **100 points** in total.

You will need to submit both your answers to the questions and accompanying R code. You should submit:

- your R Markdown file; and
- the knitted PDF resulting from your R Markdown file.

Please submit both files via Quercus.

Remember to:

- Label the answers to each question
- Label any graphs clearly with suitable axis labels and titles
- Comment your code so that it is easy to understand

Overview

This assignment relates to analyzing patterns in who intended to vote Liberal, as reported in the 2019 Canadian Election Survey. You will be using a cleaned, subsetted version of the 2019 Canadian Election Survey (`ces2019`) and median household incomes derived from the census (`census_income`).

Question 1 (5 points)

Read in the two datasets and join them together, so that the median household income by province is contained in the CES dataset. You should be able to do the join using code that looks something like that shown below.

```
ces <- ces %>% left_join(census)
```

Question 2 (10 points)

- Make a new binary variable called `vote_liberal` which is equal to 1 if respondents intended to vote Liberal and 0 otherwise
- Make new variables called `log_hh_income` and `log_median_hh_income` which is the log of the two relevant variables.
- Once you have created the logged income variables, create two more variables, which are standardized versions of the logged income variables. You can do this using the code below:

```
# standardized log hh income
ces$log_hh_income_c <-  (ces$log_hh_income - mean(ces$log_hh_income))/sd(ces$log_hh_income)

# standardized log median hh income
ces$log_median_hh_income_c <-  (ces$log_median_hh_income - mean(ces$log_median_hh_income))/
  sd(ces$log_median_hh_income)
```

Make sure these new variables are are saved in your `ces` dataset.

Question 3 (20 points)

With the aid of graphs and discussion, tell me three interesting things about patterns in who intends to vote Liberal.

Question 4 (40 points)

a)

Using `glmer`, fit the following model:

$$Y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit} p_i = \alpha_{j[i]} + \beta x_i$$

$$\alpha_j \sim N(\mu, \sigma^2)$$

where

- Y_i refers to whether or not individual i intends to vote Liberal
- The index j refers to province/territory
- x_i refers to standardized logged household income (`log_hh_income_c`)

Interpret the coefficient on x_i.

b)

Plot the $\hat{\alpha_j}$'s for each province/territory, also showing +/- 1 standard error.

Here's some code to help get the estimates extracted from the model:

```
alphas <- coef(mod1)[[1]]["(Intercept)"]
alphas <- alphas %>% mutate(province = rownames(alphas))
ses <- attr(ranef(mod1)[[1]],  "postVar")[,,1:11]
alphas <- alphas %>% mutate(se = ses) %>%
  rename(alpha = `(Intercept)`) %>%
  mutate(province = fct_reorder(province, alpha))
```

c)

What's the expected probability of a respondent in Ontario with mean (log) income voting Liberal, based on the model fit in part a)?

Question 5 (25 points)

a)

Using `glmer`, fit the following model:

$$Y_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}p_i = \alpha_{j[i]} + \beta x_i$$
$$\alpha_j \sim N(\gamma_0 + \gamma_1 z_j, \sigma^2)$$

where everything is as before and z_j is the standardized log median income at the province/territory level (`log_median_hh_income_c`).

Interpret the coefficient on z_j.

b)

Discuss briefly the estimated coefficients on income at different levels in the model fit in part b). Why is this counter-intuitive/surprising? What does it mean?