How the Immigration in Canada Affected the Voter Turnout Rate

Chien-Che Hung

10/14/2020

Note:

- Here is the link to get the source code for cleaning the raw GSS data:
  - https://github.com/frankkhung/gss_data/blob/main/data_cleaning/gss_cleaning-1.R
  - Follow the instruction in the file to get the GSS 2013 Survey raw data
  - Download the files in this folder to get the dictionaries needed to tidy the data:
    * https://github.com/frankkhung/gss_data/tree/main/data_cleaning
- Here is the link to get the source code of this report:
  - https://github.com/frankkhung/gss_data/blob/main/data_analysis.Rmd
  - Please also download the files in this folder to make this file reproducible.
    * https://github.com/frankkhung/gss_data/tree/main/inputs

Abstract

As international immigration accounted for most of Canada's population growth, we want to explore how this phenomenon affected civil participation in 2013. This report uses a multilevel logistic regression model to analyze how a sense of belonging and birthplace contribute to the voter turnout rate in the Federal election. As people born in Canada would have different levels of sense of belonging than people born outside of Canada, our analysis shows that a sense of belonging contributes a little to the voter turnout rate. On the other hand, the single fact of whether the respondent born in Canada or not negatively affected the voter turnout rate. Additionally, different provinces would affect the voting participation rate.

Introduction

This report is going to analyze the data obtained from the 2013 General Social Survey on Social Identity. The survey topics range from social networks, civil participation, birthplace to a sense of belonging and trust. Since there are varying topics that could be covered, we will discuss whether the birthplace and sense of belonging would affect the respondents voting turnout rate. Statistics Canada suggests that two-third of the population growth was accounted for by international immigrants in 2006. The report from CIC News points out that 82% of the population growth in 2019 came from immigration in Canada (Thevenot, 2020). We want to explore whether the upsurge in immigration would affect the participation rate in the Federal Election. The report structure includes explaining the data, interpreting the model that we use, presenting the results, and, most importantly, discussing the results.

Data

This data covers varying topics, for example, social networks, civic participation, pride, the main activity of respondent, birthplace, and well-being. With these categories, analysts can easily combine different variables to conduct their desirable analysis. A meaningful analysis can be carried out with combinations of these results and other background information and activities. The weakness of this data would be that there are 790 variables in total from different categories. It would be tedious for analysts to go over each variable one by one or find the most significant features from the dataset. Also, the data only includes respondents from ten provinces, excluding residents of the Yukon, Northwest Territories, and Nunavut. By doing this, the survey and data would not be inclusive enough. For this report, it considers variables that would be related to the election. The topic would be focusing on the social participation in voting, including federal election, provincial election, and municipal election.

Several questions allow write-in responses, and it will be coded into a category in the question that fits the response. By doing this, it is possible that the survey would lose some information or even categorize the response into the wrong category. However, the advantage of the survey is relatively apparent. It touches on very detailed information for the respondents. Also, since this survey is conducted through computer-assisted telephone interviewing and electronic questionnaires, it is customizable for respondents who are not eligible for some activities, such as voting.

This General Social Survey program surveys the ten provinces from June 2013 to March 2014, who were above 15-year-old. These surveys were asked through interviews via computer-assisted telephone interviewing (CATI) and electronic (Internet) questionnaire (EQ). Unlike previous years' surveys conducted through Random Digit Dialing, respondents in this 2013 survey were reached by telephone numbers available to Statistics Canada and the Address Register. For Electronic Questionnaire, during the phone interview, if the respondents agreed to complete the survey one, the interviewer would send them emails containing the survey and its access code.

As we mentioned above, the population was the ten provinces, which Yukon, Northwest Territories, and Nunavut were excluded. For the frame for this survey, it was created from different components. The first component was the list of telephone numbers to Statistics Canada from telephone companies or Census data. The second one was the Address Register, which contains all dwellings in the ten provinces. For the sample size, the targeted sample size was 31973. However, the number of respondents who answered was 27695.

Model

Since we desire to explore whether the sense of belonging or respondents' birthplace would affectpeople's voting turnout rate, our model's response variable would be binary. As in the Data Section mentioned, the respondents were surveyed from different provinces (Stratum), which means that our observations might not be independent. As we have repeat measures and heterogeneity across the provinces, we would be ignoring essential variations. To explore the provinces' variance, we will use the Bayesian Multilevel (Hierarchical) Regression model to analyze the data. In the multilevel model, the variances among observations are separated into Group-Level Effect and Population-Level Effects. Group-Level Effect would be the variance across the provinces that GSS surveyed, and Population-Level Effects would be the effects of the model's variables: Sense of Belonging and Born In Canada or Not. In other words, we can say that this model is the compromise between the extremes:

1. No pooling: provinces pose no effects and independent on the voting turnout rate
2. Completely pooling: provinces have the same turnout rate

The Bayesian Multilevel Model utilizes the application of Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

We can have the posterior estimation on the intercepts, coefficients, and variations because of the prior information and the likelihood that we have on the response variable.

$$
\begin{aligned}
&P(\beta_{sense}, \beta_{born}, \alpha, \alpha_{prov}, \sigma^2_{prov}|y) \\
&\propto P(y|\beta_{sense}, \beta_{born}, \alpha)P(\alpha_{prov}|\sigma^2_{prov})P(\beta_{sense})P(\beta_{born})P(\sigma^2_{prov})
\end{aligned}
$$

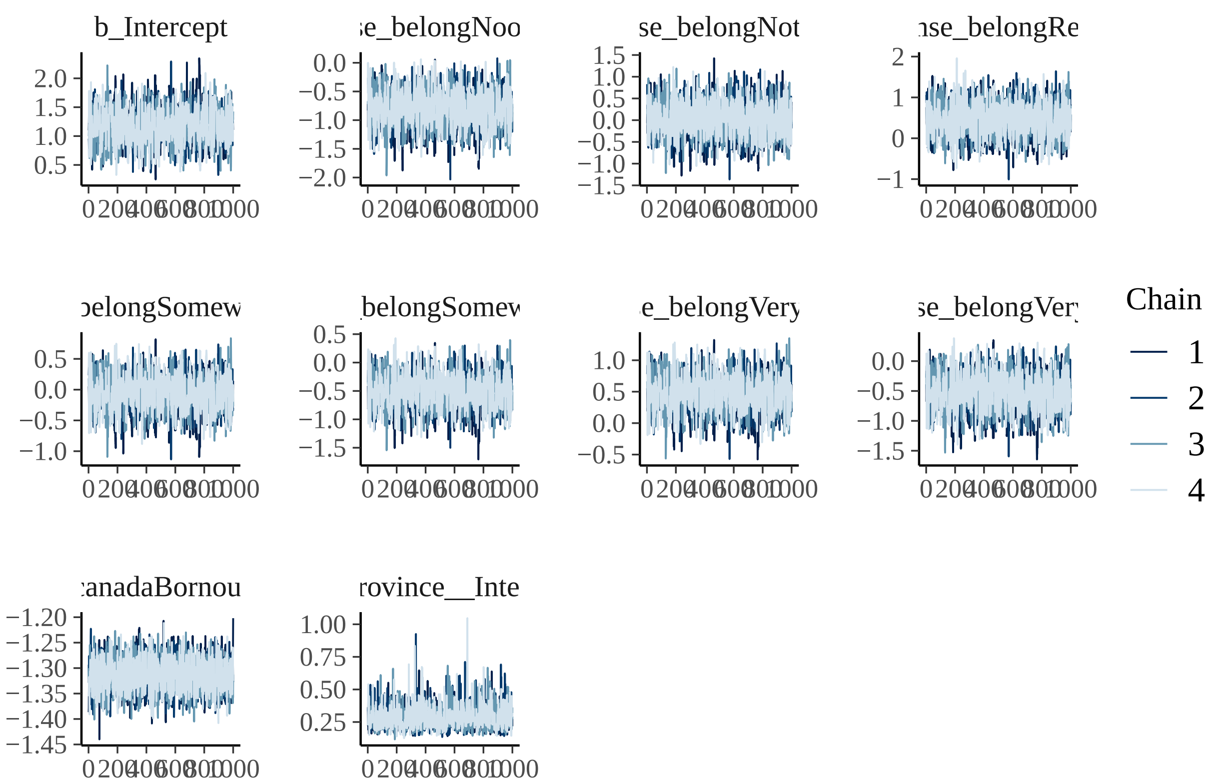
As we previously mentioned, in the model described below, we use a sense of belonging and whether born in Canada as fix intercepts and hierarchically model the effect of provinces.

$$
\begin{aligned}
Y_i|p_i &\sim Bernoulli(p_i), i = 1, ..., n \\
logit(p_i) = log(\frac{p_i}{1-p_i}) &= \alpha + \alpha_{province[i]} + \beta_{sense} belong + \beta_{born} born \\
\alpha_{province} &\sim student(0, 2.5) \\
\beta_{sense} &\sim Normal(0, 10) \\
\beta_{born} &\sim Normal(0, 10)
\end{aligned}
$$

For the priors that we pick are the weakly informed priors (Stan-Dev, 2020). As the coefficients in logistic regression could be a broad range, so we make the priors for the coefficients generally distributed with the variance of 10 and 0 mean. The standard deviation for the provinces would be in student-T distribution with 0 mean and 2.5 standard deviations. By having weakly informed prior, we can prevent our data from being too sensitive to our prior, especially with those provinces with fewer samples.

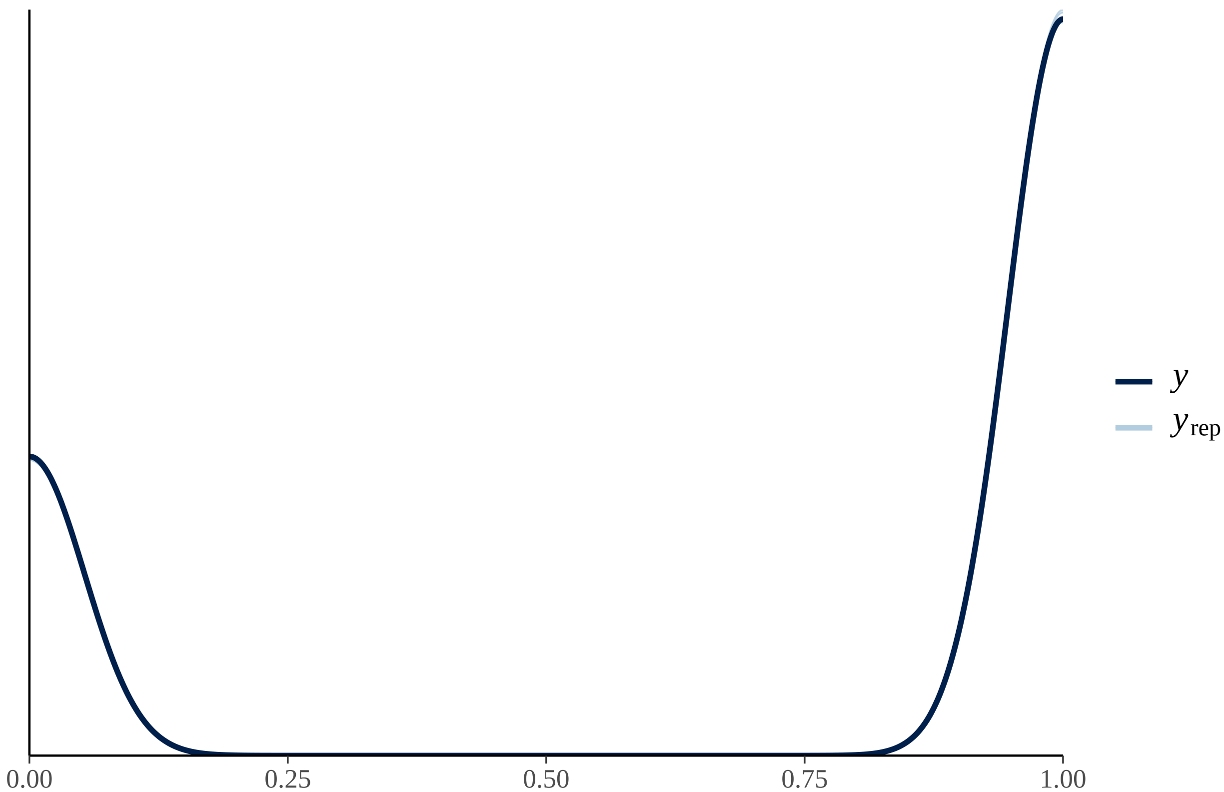
To make sure this is a reasonable model, we want to check some diagnosis. Firstly, we want to see if the model converges. From the chart below (Graph 1),we can see that no divergences are showing in each estimate in different chains that we run in the model.

Graph 1: Trace plot for Divergence



Secondly, we want to check if this plot has a proper fit. From the chart (Graph 2) we can see that the model has a remarkably similar density plot as the observations. Typically, for other Generalized Linear Models, one of the assumptions involves constant error variance. However, since the model we are using here is logistic regression and error variance is not a parameter in the Bernoulli distribution, we will not consider this assumption.

Graph 2: Model Fitness



The brms package runs this entire model and plots in R. It fits Bayesian generalized linear and nonlinear multivariate multilevel models using 'Stan" for full Bayesian Inference (https://cran.r-project.org/web/packages/brms/index.html). The following section will present and discuss the results of this model.

Result

Here are the general model and the fitness of the results.

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: fed_elec_voted ~ 1 + sense_belong + born_in_canada + (1 | province)
##    Data: voting (Number of observations: 26007)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~province (Number of levels: 10)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.29      0.09     0.17     0.49 1.00      914     1465
##
## Population-Level Effects:
##                                   Estimate Est.Error l-95% CI u-95% CI Rhat
## Intercept                             1.19      0.29     0.63     1.75 1.00
## sense_belongNoopinion                -0.80      0.30    -1.39    -0.23 1.00
```

```
## sense_belongNotstated               0.02      0.38    -0.69     0.78 1.00
## sense_belongRefusal                 0.46      0.38    -0.28     1.21 1.00
## sense_belongSomewhatstrong         -0.07      0.28    -0.64     0.47 1.00
## sense_belongSomewhatweak           -0.52      0.29    -1.09     0.03 1.00
## sense_belongVerystrong              0.46      0.28    -0.11     1.00 1.00
## sense_belongVeryweak               -0.54      0.30    -1.13     0.04 1.00
## born_in_canadaBornoutsideCanada    -1.32      0.03    -1.37    -1.26 1.00
##                                Bulk_ESS Tail_ESS
## Intercept                          1249     1473
## sense_belongNoopinion              1307     1706
## sense_belongNotstated              1729     2137
## sense_belongRefusal                1632     1869
## sense_belongSomewhatstrong         1227     1234
## sense_belongSomewhatweak           1230     1388
## sense_belongVerystrong             1234     1193
## sense_belongVeryweak               1304     1389
## born_in_canadaBornoutsideCanada    3349     2008
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

It presents the Group-Level (Province) Effects and Population-Level Effects (Sense of Belonging and Born In Canada or Not). The estimation in the Population-Level Effects is shown in Log-Odds. If it is positive, it means that it positively affects the odds and vice versa. With the assistance of the distribution of each coefficient, we can see the importance of them. The ones that touch on zero indicate that there are some chances that they could be zero. Here, we can see that the intercept (respondents who did not state their sense of belonging level) and whether born in Canada from the Population-Level Effects are far from 0. It means that they have a certain degree of influence on the response variable.

Since we are assessing the effects in multilevel, we want to see how much variation that the province contributes to the model by finding the Intraclass Correlation Coefficient with the formula.

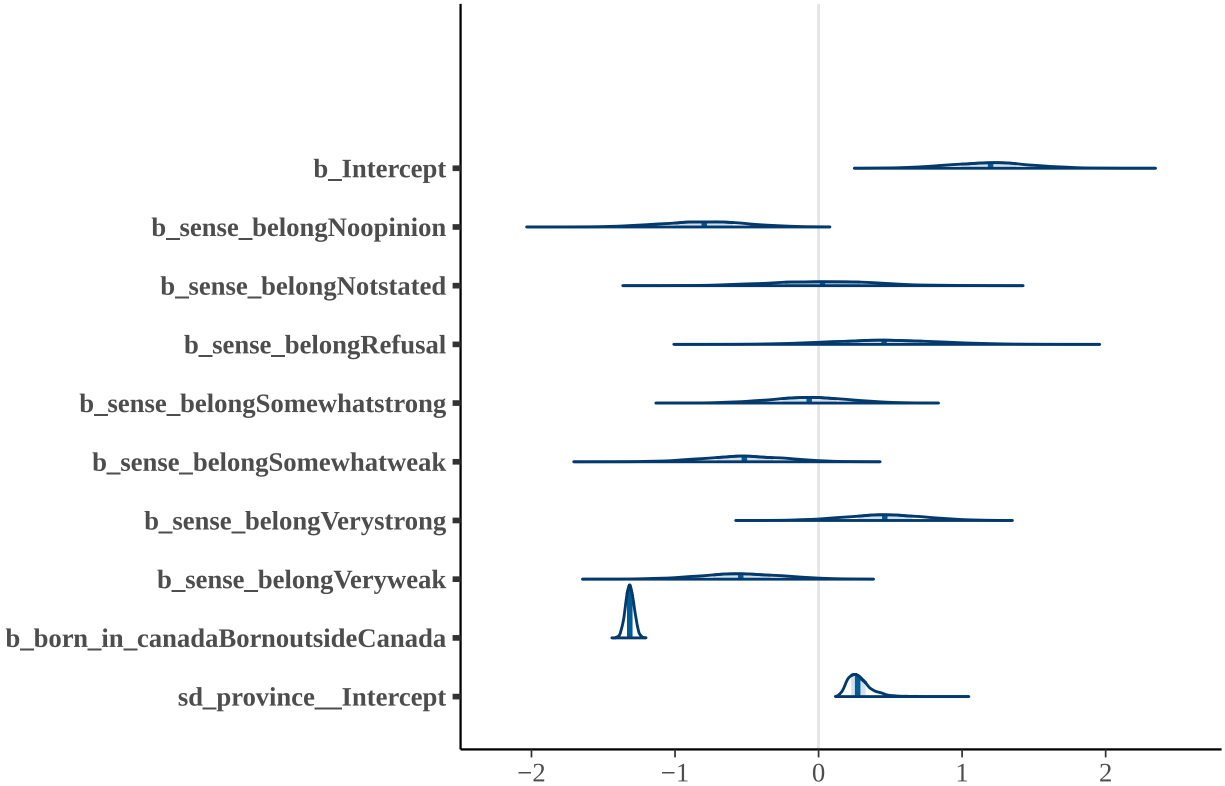
$$ICC = \frac{\sigma^2_{group}}{\sigma^2_{group} + \sigma^2_{population}}$$

The variance for logistic distribution is $\frac{\pi^2}{3}$. From the result, we can say that provinces contribute 2.4% of the variance to the model.

Table 1: ICC for province

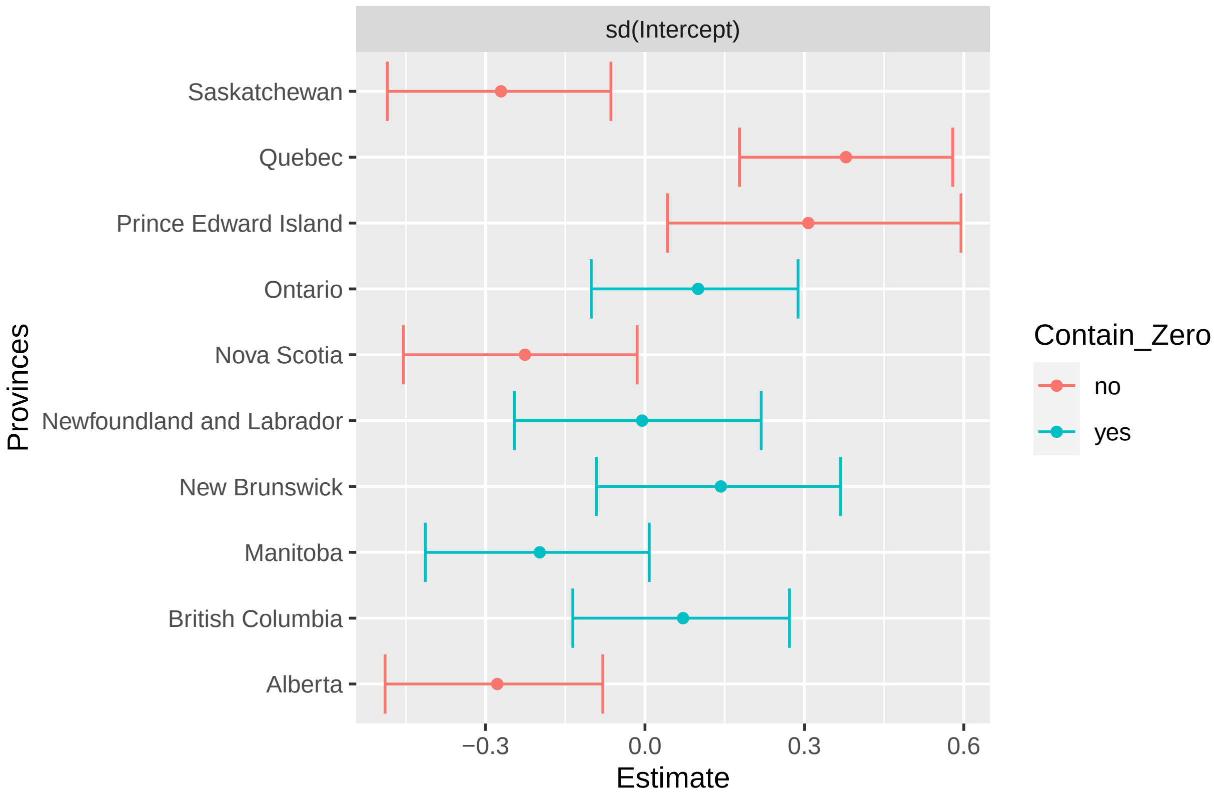
	Probability
Intraclass Correlation Coefficient	0.024

Graph 3: Distribution Plot for Parameter Posterior E



From the broader view from the distribution graph, we can see that the provinces' variation will not be zero since the distribution does not include zero, which means that provinces have effects on the election turnout. As we look closer into the Group-Level Effects, the plot below (Graph 4) demonstrates the variations each province contributed.

Graph 4: Variation in Each Province



If the estimated variation credible interval contains zero, it means that the variation regarding the response variable could be undetectable. On the other hand, if the credible interval does not contain zero, being in the province would make a difference in the outcome.

We use logistic regression as our model to fit the binary response data. Due to this reason, we can calculate the mean probability of each case by predicting each combination of cases and calculate the inverse logit scale. The below table (Table 2) shows the probability of different cases.

Table 2: Mean Probability for each case

	Probability
sense_dont_in	0.916
sense_noop_in	0.596
sense_notstate_in	0.771
sense_refuse_in	0.839
sense_somestrong_in	0.754
sense_someweak_in	0.754
sense_strong_in	0.663
sense_weak_in	0.839
sense_dont_out	0.469
sense_noop_out	0.284
sense_notstate_out	0.475
sense_refuse_out	0.583
sense_somestrong_out	0.452
sense_someweak_out	0.452
sense_strong_out	0.345
sense_weak_out	0.583

Discussion

Result Discussion

Graph 3, even though some of the credible intervals for coefficient estimate include 0, the coefficients have the possibility of posing no effects to the voting turnout. However, we will still take a look at the effects of each of them. In the table below, we can see that the respondents who did not know about their feeling regarding the sense of belonging to Canada positively influence the log-odds. Here we define odds as in the ratio between the amounts staked by the parties to a bet. People who did not have any opinion, somewhat strong opinion, and somewhat weak opinion have a negative relationship with the log-odds of the probability of voting in the Federal Election.

Table 3: Coefficient Estimates

	Estimate	Est.Error	Q2.5	Q97.5
sense_belongDontknow	1.192	0.291	0.625	1.749
sense_belongNoopinion	-0.802	0.301	-1.393	-0.229
sense_belongNotstated	0.024	0.377	-0.691	0.781
sense_belongRefusal	0.459	0.383	-0.284	1.211
sense_belongSomewhatstrong	-0.069	0.281	-0.635	0.471
sense_belongSomewhatweak	-0.517	0.288	-1.091	0.032
sense_belongVerystrong	0.460	0.280	-0.105	1.001
sense_belongVeryweak	-0.540	0.297	-1.132	0.045
born_in_canadaBornoutsideCanada	-1.315	0.030	-1.374	-1.257

Noticeably, the respondents born outside of Canada pose the most substantial adverse effect on the Federal Election voting turnout. It means that people born outside of Canada, residing in Canada while the survey was conducted, did not vote during the election. The below table shows the probability of voting in each situation with the mean probability of them. We can see that people born in Canada have a way higher probability of voting inthe Federal Election. (TABLE 4)

Table 4: Mean Probability for each case

	Born In	Born Out
sense_dont	0.916	0.469
sense_noop	0.596	0.284
sense_notstate	0.771	0.475
sense_refuse	0.839	0.583
sense_somestrong	0.754	0.452
sense_someweak	0.754	0.452
sense_strong	0.663	0.345
sense_weak	0.839	0.583
Mean Probability	0.767	0.456

Interestingly, as we look at people born outside of Canada's response to the sense of belonging, most of them responded in the "Strong" category. (Table 5) This result indicates that there are underlying reasons why this group of people has negative effects on voting.

Table 5: Counts for Sense of Belonging Response (Born Outside of Canada)

sense_belong	n
Don't know	41
No opinion	125
Not stated	40
Refusal	31
Somewhat strong	2556
Somewhat weak	343
Very strong	5892
Very weak	94

Graph 4 shows the variation caused in the Group-Level (provinces). Specifically, residents in Quebec, Saskatchewan, and Alberta would be most likely to vote while living in Ontario, Newfoundland, Labrador, New Brunswick, Manitoba, and British Columbia could suggest being in these provinces would not be the incentives of why the residences vote.

To conclude the findings in a broader context, a sense of belonging to Canada possibly would not contribute to voting turnout. On the other hand, the respondents' birthplace creates a massive difference between whether people would vote. For different provinces in Canada, some of the provinces would influence the decision of whether their residents should vote or not.

Weakness and Caveats

There are some possible weaknesses and caveats in the model. From the analysis above, we can see that most of the answers from the sense of belonging question could not affect the voting turnout. It could also be the weakly informed prior has too many effects on the sense of belonging variable. Even though we could get a considerable amount of information from this data and analysis, this research topic is limited to Canada and the Federal Election. Specifically, the data is collected by a stratified simple random sample, and each stratum has a different situation, as we have shown above. It would be hard and infeasible to generalize this conclusion regarding a sense of belonging to other countries. However, as immigration is happening in other countries, the massive difference in whether born-in-Canada could be utilized in other analyses.

Future Work

POLYAS, an online voting company, states that there is an increasing division between different voting groups. One evidence is that in General Election in the UK in 2015, 75% of the upper-middle class and the middle class voted, whereas only 56% of the working class and the non-working class voted. (Increase Voter Turnout: Create a Sense of Belonging 2017) This evidence from POLYAS shows that the salary level or education level in this GSS 2013 Social Identity Survey could contribute to the probability of voting turnout rate. If we only conclude from the analysis that this report has done, a sense of belonging would not be a significant social participation factor. On the other hand, there is a significant disparity between whether the respondents were born in Canada. We could dig more into this group of people, and policymakers could also shift their ideologies towards the people born outside of Canada to increase the voter turnout rate.

Reference

Data Citation

Data

- Statistics Canada (2013). Canadian General Social Surveys Cycle 27: Social Identity 2013 (Version 2)

Data Cleaning Code Adapted From:

- Alexander R. and Caetano S. (2020) gss_clearning-1.R (Version 1)

Report Citation

- Gabry J, Mahr T (2020). "bayesplot: Plotting for Bayesian Models." R package version 1.7.2, <URL: https://mc-stan.org/bayesplot>.
- Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr
- Hadley Wickham (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Increase Voter Turnout: Create a Sense of Belonging. (2017, January 27). Retrieved October 19, 2020, from https://www.polyas.com/increase-voter-turnout/sense-of-belonging
- Immigration and growth. Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. (2016, October 7). https://www150.statcan.gc.ca/n1/pub/11-402-x/2010000/chap/imm/imm01-eng.htm.
- JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. URL https://rmarkdown.rstudio.com.
- Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01
- Paul-Christian Bürkner (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Stan-Dev. stan-dev/stan. GitHub. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.
- Thevenot, S. (2020, July 20). Immigration to drive Canada's population growth as global birth rates fall. CIC News. https://www.cicnews.com/2020/07/immigration-to-drive-canadas-population-growth-as-global-birthrates-fall-0715086.html.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.