Mixed Effect Logistic Regression With Poststratification to Predict the Total Popular Vote of the Upcoming Canadian Federal Election

GROUP 39: Charlotte Carlyle, Zachary Chan, Harrison Jones, Justin Yu

November 24, 2022

Contents

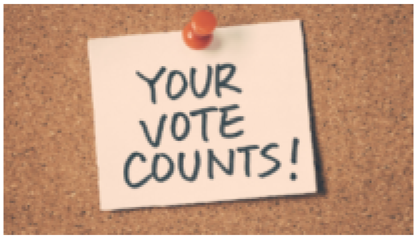


Figure 1: Vote Image Sourced from Children's Treatment Network

1 Introduction

Representative democracy has been how Canada has been governed for generations.This form of government has been the cornerstone for the expansion of political and social rights, and the continuous participation of the citizenry is without question required for the system to have any legitimacy. Therefore, the continued growth of political participation is what would be desired for every election cycle. This generates both more accurate public representation in government and displays the expansion of political rights as greater numbers of people migrate to Canada and become productive citizens.[7] However, there are signs that voter participation has started to fall. The previous two general elections saw voter turnout drop by 4.5% and 1.7% respectively. Even recently, the 2022 Ontario provincial election saw the lowest voter turnout in history, with a drop of more than 13% compared to the previous election.[2] There are several reasons why voter turnout has been falling recently, including a lack of interest in politics, health, and more recently the COVID-19 pandemic.[9] If this trend continues, it could be a concern for Canadian politics. Additionally, in order to gain more votes in the future, parties can target groups that have a low voter turnout to increase the overall popular vote and their own popular vote. Canadian elections function by allowing each citizen one vote. Citizens vote in their riding, which is based on their geographical area of residence.

The goal of this analysis is to predict the popular vote for all candidates for the upcoming Canadian Federal election that is currently scheduled for 2025. The two datasets used are a census from 2016, as well as a survey of people who voted in the 2019 election. Both of these datasets provide information on age, location, gender, citizenship status, and various other variables which might be used to provide an estimate of the total voter turnout. A multi-level logistic regression model with poststratification will be used to provide an estimate of the total voter turnout. This analysis will answer the question:

What is the total number of people that will vote in the next Canadian Federal election?

Based on the recent trends in voter turnout and population growth; **our hypothesis is that the overall popular vote will be 17,000,000, at approximately the same level as the 17,034,243 in 2021.** [8]

2 Data

The data collected from our analysis comes from the 2016 Canadian General Social Survey (GSS) and the 2019 Canadian Election Study (CES) phone survey. Both the 2016 GSS data and the 2019 CES data were collected from different websites. The first dataset represents all participants of the 2016 Canadian General Social Survey which is also known as the census data for the year. It includes numerous characteristics of participants such as their age, income, and education. The second dataset represents participants of the 2019 CES phone survey which surveys participants on their personal background as well as various political opinions.

Before beginning this analysis we needed to clean our dataset. This started with the cleaning up of the 2019 CES data. The first step in this process was filtering only for the questions from the survey that were going to be used for our analysis. Questions 1, 2, 3, 4, 10, and 69 were selected. Question 1 refers to the citizenship status of individuals, Q2 is the birth year, Q3 is the sex identification, Q4 is which province individuals are living in, Q10 is the certainty of individuals to vote and Q69 is the house hold income. Next, answers to questions that had either the options of "refused" or "don't know" were removed from our selected survey questions. This was done because these answers are not strong or helpful indications of one's political perspectives. Removing answers that were marked as NA was also necessary for cleaning the data since these responses do not indicate or provide any leverage for our analysis. The next step was to remove the option of "No" from question 1 which asks if the individual is a Canadian citizen. This was done because our analysis is regarding Canadian elections in which only Canadian citizens are allowed to participate. Next, we filtered out responses where the birth year is later than 2001 to ensure that all responses are 18 or older and were eligible to vote in the 2019 election. We then converted birth year to age by subtracting the birth year from 2019. This will ensure that the survey data is consistent with the census data which contains age and not birth year. Next, we removed option 3, "other", from the third question which asks which gender the individual identifies as. This was done because there was only 1 response of "other" out of thousands and it was therefore not significant. Lastly, we adjusted the possible answers for question 10 to either 1 or 0 in which 1 represents that the individual is either certain or likely to vote and 0 represents that the individual is unlikely to vote or will not vote. This was done in order to simplify whether or not a person will vote. Next, we created a new variable for income from the survey data. Income from the survey was split into 6 categories based on $25,000 (CAD) increments with the lowest category being less than $25,000 and the highest category being more than $125,000.

Next, we cleaned the 2016 GSS data. For our analysis, we chose variables caseid, age, province, education, family income, and sex. Therefore, the cleaning process for this dataset was to select these variables and remove other variables that are not being used in this analysis. We then removed all the empty observations found in the selected variables.A description of these variables will be provided next.

Table 1: Generalized Important Variables Used in the Study.

Variable	Description
Citizen	Whether the individual is a Canadian Citizen or not
Age	The age of individuals
Sex	The sex identification of individuals
Province	The current living location of individuals
Income	The household income levels of individuals
Vote Certainty	How likely individuals are going to vote

From the 2016 census data, the important variables being used in this analysis will be age, province, sex, and income. Age is a numerical variable and represents the age of the individual from the census. This variable was rounded to the nearest integer in order to display ages more clearly and to be consistent with the survey data. Province is a categorical variable that represents the Canadian province or territory the individual lived in while participating in the 2016 census. However, the data did not contain any respondents from the territories, thus only the provinces are represented in the census data. This means that there are only 10 possible options for province. The variable sex is a binary categorical variable that takes on the values male or female. Income represents the level of income of the respondent's family in intervals such as less than $25,000 or $50,000 to $74,999. There are a total of six intervals for income, each of which covers a $25,000 (CAD) interval.

From the 2019 phone survey The first variable, age, which was derived from birth year, gives us an understanding of the person's age and is a numerical variable. The second variable used is a binary categorical variable, sex, which represents the gender the respondent identifies with and takes on the values of male or female. The third variable, province, represents the province or territory the respondent resides in, which includes a list of all provinces and territories in Canada as options to choose from. However, similar to the census data, the only responses are for provinces, which means there are 10 possible categories for this variable. The fourth variable vote certainty, represents how likely the individual is to vote. This is used as our dependent/response variable in the regression model. It takes on a binary outcome of 1, the respondent will vote, or 0, the respondent will not vote. The fifth variable, income, is the total annual household income (CAD) of the survey respondent. Income is split into the same 6 categories based on $25,000 increments as it was in the census data.

Table 2: Summary of the Variables in the Phone Survey Separated by Vote_Certainty (0 - Unlikely to Vote, 1 - Likely to Vote).

Variable	**N**	**0**, N = 126	**1**, N = 2,793
Citizen, No. (%)	2,919	126 (100%)	2,793 (100%)
Age, Median (IQR)	2,919	44 (32, 59)	50 (38, 63)
Sex, No. (%)	2,919		
1		82 (65%)	1,606 (58%)
2		44 (35%)	1,187 (42%)
Province, No. (%)	2,919		
Alberta		9 (7.1%)	185 (6.6%)
British Columbia		29 (23%)	548 (20%)
Manitoba		8 (6.3%)	184 (6.6%)
New Brunswick		3 (2.4%)	142 (5.1%)
Newfoundland and Labrador		9 (7.1%)	135 (4.8%)
Nova Scotia		7 (5.6%)	146 (5.2%)
Ontario		23 (18%)	539 (19%)
Prince Edward Island		8 (6.3%)	139 (5.0%)
Quebec		22 (17%)	580 (21%)
Saskatchewan		8 (6.3%)	195 (7.0%)
Income, Median (IQR)	2,919	57,500 (27,250, 102,250)	80,000 (45,000, 140,000)

The numerical summary above contains the response of the dependent variable vote certainty in the cleaned survey data. Based on this statistic, 96% of the GSS respondents should have voted in the 2019 Canadian Federal election. This number seems quite high, particularly considering that we know only 63.2% of the eligible voting population voted in the 2021 election.[1]

Table 3: Numerical Summary Table for Ages in Study

Estimates	Age_Survey	Age_Census
Min.	18.0	18.0
1st Qu.	37.0	34.8
Median	50.0	46.6
Mean	50.1	46.3
3rd Qu.	63.0	57.7
Max	95.0	80.0
Std.	16.0	14.2
N	2919.0	12089.0

Table 3 shows the numerical summary of the ages of individuals in both the Census and Phone survey. We can see that the two sets of data are relatively similar, with the survey having less than 4 times the amount of observations seen in the census. The minimum age for individuals is 18 for both datasets, which is to be expected given that Canadians must be at least 18 in order to be eligible to vote. Additionally, we can see that there is very little skew in this data given that the mean and median are relatively the same for both datasets.

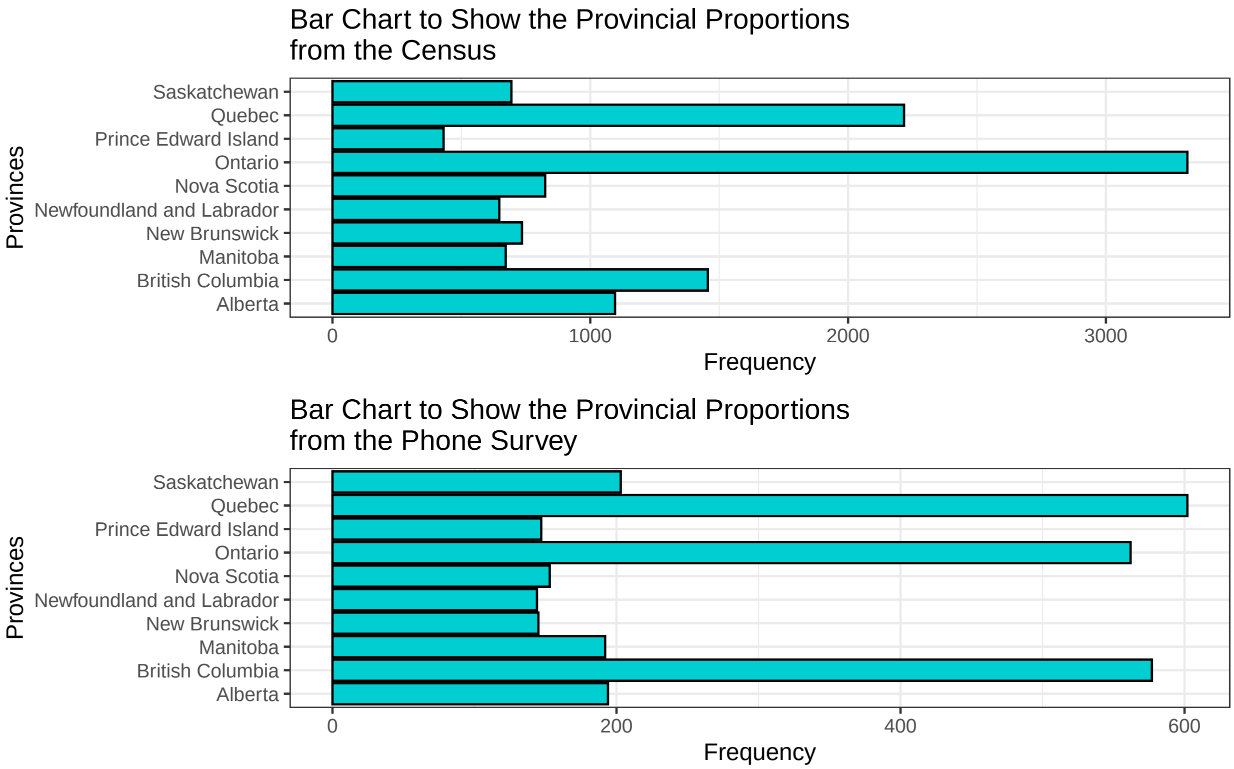
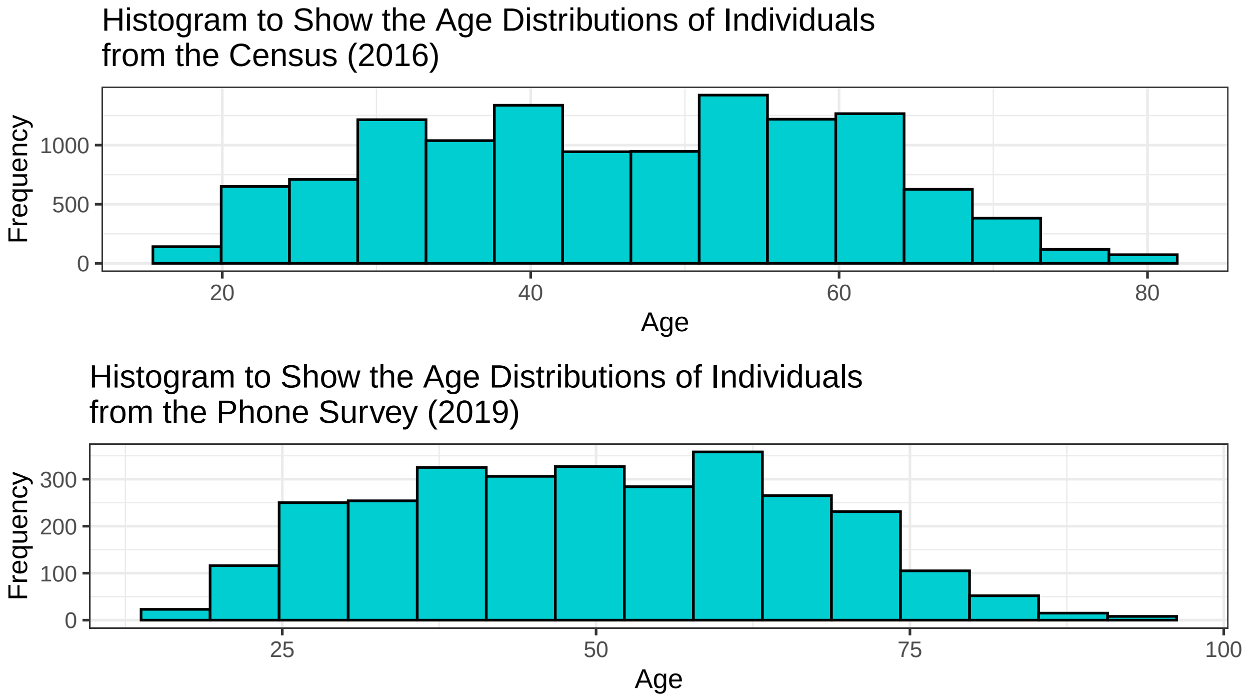


Figure 2: Bar Chart to Show the Provincial Proportions from the Phone Survey

Figure 2 above shows the comparative distribution of the provinces in the census and the survey. The top bar chart uses data from the census data while the second uses the survey data. Each chart captures all ten Canadian provinces. In the census data, we can see that an overwhelming majority of respondents resided in Ontario, while Prince Edward Island had the lowest number of respondents. In the survey data, the bottom bar chart, we can see that British Columbia had the largest number of respondents, while New Brunswick had the lowest number of respondents.

Similar to the distribution of sex in the survey, we can see that certain provinces are over or underrepresented in the survey. For example, Quebec is overrepresented by having the largest number of respondents in the survey data but only the 3^{rd} largest in the census data.


Histogram to Show the Age Distributions of Individuals from the Census (2016)

Histogram to Show the Age Distributions of Individuals from the Phone Survey (2019)

Figure 3: Histogram to Show the Age Distributions of Individuals from the Phone Survey

Figure 3 above shows the comparative age distribution between the survey data and the census data. The top histogram displays the age distribution for census respondents. We can see that the data is concentrated in the center with a small number of respondents near the age of 18 and a small number of respondents older than 70. The bottom histogram shows the age distribution for the survey data. The survey data shows a similar distribution to the census data with slightly less variation in the buckets toward the center of the histogram. Notably, the survey captures older age groups than the census, which has a maximum of 80. These histograms are also a visual representation of the fact that there is limited skew in the data for age. Additionally, when compared to sex and province, the survey more accurately captures the distribution of the census data.

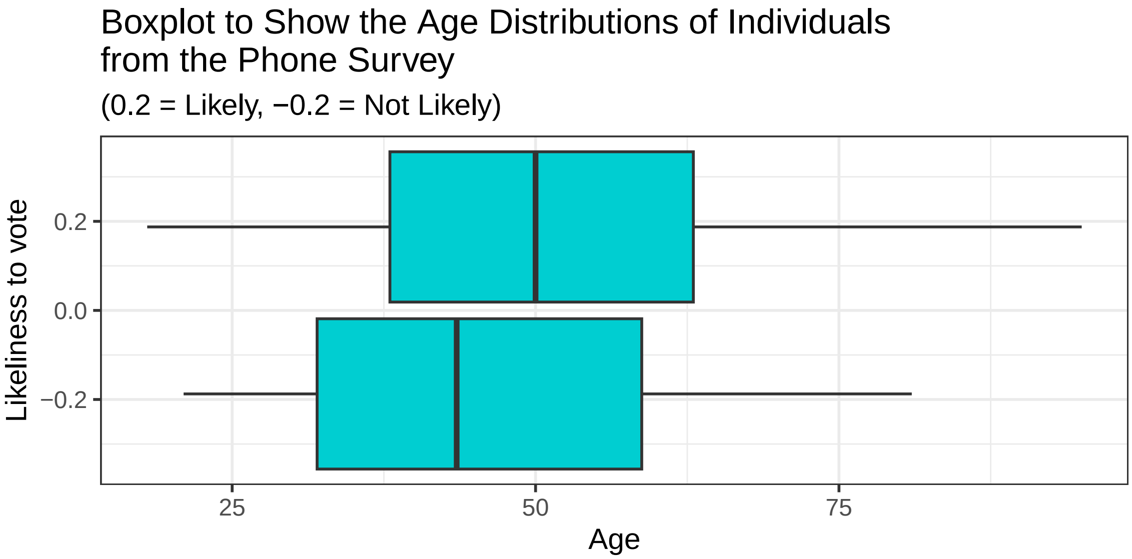


Figure 4: Boxplot to Show the Age Distributions of Individuals from the Phone Survey

Figure 4 above shows the age distribution among the respondents in the survey data based on whether or not they will vote. When the data is split in this manner, we can see the differences in the spread of these two groups. The median age of respondents that are likely to vote is higher than the median age of respondents that are not likely to vote. This observation makes sense given the typical belief that younger people are less likely to vote. Additionally, this will also set a reasonable expectation for what our final model should result in regarding a respondent's likelihood to vote based on age.

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

3 Methods

Our overall analysis will involve a multilevel regression with poststratification. This can be broken down into two components, first, we will perform a multi-level logistic regression and then post-stratify based on the regression results to determine overall voter turnout.

3.1 Multi-level Logistic Regression

We will be using a multi-level logistic regression model to model voter turnout for different strata. A logistic regression model uses a binary response variable, meaning that there are only two options. In determining voter turnout, the response variable is whether or not an individual will cast a ballot in the upcoming Canadian Federal election. A logistic model makes the most sense in this scenario since our variable of interest is a binary outcome, where 1 means that an individual will cast a ballot and 0 means that they will not. A multi-level model is useful to consider unobserved group characteristics, or random effects, as well as fixed effects at the group level. In determining the multi-level model, gender, province, and income will be used as random effects variables while birth year will be used as the only fixed effects variable. The model used is shown below:

$$log(\frac{p}{(1-p)}) = \beta_0 + \beta_1 x_{Age} + v_{Sex} + v_{Province} + v_{Incomek}$$

In this model, $\beta_0 + \beta_1 x_{Age}$ consider the fixed effects variables, while $v_{Sex} + v_{Province} + v_{Incomek}$ consider the random effects variables. The logistic regression model predicts log-odds, $log(\frac{p}{(1-p)})$, where $\frac{p}{(1-p)}$ is the odds and p is the probability that a respondent will vote. The logistic regression model shows a linear relationship for log odds but not for odds.

The assumptions of this analysis include the assumptions of the logistic regression model. The first assumption is that the outcome of this regression is binary, which we know to be true with whether or not a ballot is cast. The second assumption of logistic regression is linearity in the logit for continuous variables, meaning that there is a constant increase in the log odds for every increase in a predictor, holding all else equal, regardless of its starting value. The next assumption is the absence of multicollinearity between the explanatory variables.

This means that there is no high correlation between explanatory variables and the explanatory variables are independent. The last assumption is a lack of extreme outliers that influence the data. Outliers are observations that differ greatly from the majority of the other data points. It is necessary to assume that they do not exist in our data because they would skew the model and cause our results to be inaccurate.

3.2 Post-Stratification

In this analysis, a poststratification will be used, which is a technique used to improve model estimates by adjusting the weights for greater efficiency of its estimators. It is useful because it helps utilize samples that may have been under or over-sampled, therefore making the overall sample a better representation of our target. As shown in the data section, the census and survey data have an uneven distribution among strata. For example, the survey has more males than females while the census has more females than males. We can use post-stratification to adjust the weights from the sample data to the appropriate weights from the census data. The equation used for poststratification is below.

$$\hat{y}^{PS} = \frac{\sum N_j \hat{y}_j}{\sum N_j}$$

Where;

- $\hat{y}^{PS}$ is the variable of interest for the population, the probability of an individual casting a ballot in the election
- $\hat{y}_j$ is the estimate of the probability of casting a ballot for the j^{th} cell determined from the logistic regression model
- N_j is the population size of the j^{th} cell based on the demographics of the cell

In order for this model to be appropriate, we will assume that the size of each cell, N_h is known based on the population data. Additionally, we will assume that the breakdown of cells is an accurate way to represent the population.

The bins that will be used in the poststratification process are based on sex, province, and income. This resulted in a total of 120 cells. Sex is used because females tend to vote more than males. The choice to vote varies across provinces because individuals may feel that they have more of a say in who is voted into the House of Commons based on their population to number of ridings ratio. For example, Ontario is underrepresented by population while PEI is overrepresented. Finally, income is used to determine cells for two reasons. Those with low income may have difficulty accessing polls while those with high income may feel they have a greater stake in certain political issues such as taxation. Another factor that was considered was employment status. However, this information was not in the census data and could only be closely represented by occupation, which does not truly indicate employment status. Education was also considered; however, the education-related variables in the survey and census data were not compatible with each other. As a result, there would not be an appropriate to post-stratify the education in the survey to what is in the census.

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

4 Results

The results from the mixed effects logistic regression is shown below:

Table 4: Numeric Summary Obtained from Mixed Effect Model Built

	Vote_Certainty		
Predictors	Odds Ratios	CI	p
(Intercept)	8.50	4.37 – 16.55	<0.001
Age	1.02	1.01 – 1.03	<0.001
Random Effects			
σ2	3.29		
τ00 Province	0.00		
τ00 Incomek	0.16		
τ00 Sex	0.03		
N Sex	2		
N Province	10		
N Incomek	6		
Observations	2919		
Marginal R2 / Conditional R2	0.031 / NA		
AIC	1025.9		
BIC	1055.8		

Above is the table displaying the properties of each predictor. For the fixed effect predictors (Age, and Intercept) the results show great statistical significance for the model, meaning we have strong evidence of linearity between them and log-odds. The random effect predictors (Province, Incomek, and Sex) show the variance of each of their respective strata. I.e. the average of how deviation each group has from the overall average before stratifying.

Table 5: Summary Table for Census Individuals Probability of Going to Vote

Estimates	Probability_to_Vote
Min.	0.8553533
1st Qu.	0.9475134
Median	0.9612961
Mean	0.9565591
3rd Qu.	0.9707166
Max.	0.9868980
SD	0.0201763

The results of the regression model when applied to the cleaned census data show that the weighted mean probability that an eligible voter will cast the vote on election day is 0.956. The fixed effect variables display very significant evidence supporting a linear relationship between the variables and the log-odds determined through the logistic regression. The random effect variables of income and gender display a small but noticeable variance, which suggests that differences in log-odds do exist between the various strata of these groups. Though an interesting observation is that the province random effects variable doesn't show any variance in the model. This was very unexpected, as it implies that no variation exists in the log-odds depending on which province a person lives in that isn't already explained by another random effects predictor. This would suggest that province was potentially a redundant predictor for the model and it was not necessary to include.

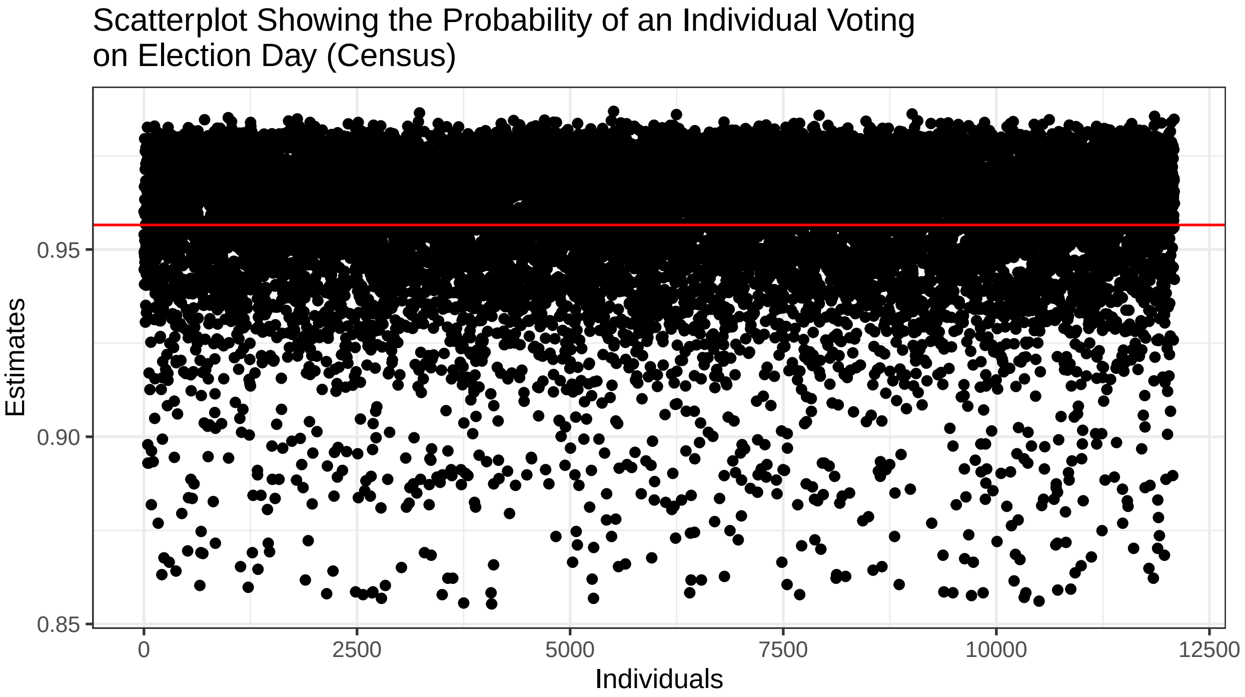


Figure 5: Scatterplot Showing the Probabilities of Census Individuals Voting on Election Day

Redline = Weighted Average Probability of Voting. Above is the plot of each individual in the census data and their probability to vote. The graph displays a moderate skew toward the lower end of the probability. Albeit, the weighted mean is still the preferred probability despite the skewness given the small difference between the two values.

With the census data given, we observed 12,314 observations. Therefore we can predict that 11,772 (12,314 * 0.956) Canadians will vote in the polls in the election scheduled for 2025. To estimate the total popular vote across the entire population, we can use a simple population growth estimator based on the annual growth rate of the number of adult Canadian citizens from 2016 to 2021. (0.6626%).[3] Using the total popular vote and voter turnout rate of the 2015 federal election we can estimate that Canada had 25,754,634 adult citizens at the time of the election.[1] Applying the growth rate we estimate that Canada will have 26,619,267 adult citizens in 2025. **Thus our model would predict that a total of 25,448,019 votes will be cast in the 2025 election.**

These results, however, don't seem reasonable when considering real-world examples of voter turnout. There is essentially no democracy in the world where voter turnout is realistically above 90%. Previous Canadian elections in 2015, 2019, and 2021 had turnout rates of 68.3%, 67%, and 63.2% respectively[1]. The potential reasons for this unreasonable result are numerous, albeit we can make some haphazard theories. The relatively small amount of data itself may have caused greater variation in the predictors. The heavily skewed response variable toward one outcome could have been another predictor. Additionally, the complexity of the model may have caused it to overfit the survey data and therefore create an unreliable model for poststratification.

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

5 Conclusions

From the introduction, recall our stated hypothesis;

"Based on the recent trends in voter turnout and population growth, our hypothesis is that the overall popular vote will be 17,000,000, at approximately the same level as the 17,034,243 in 2021."[8]

In an attempt to test the above-stated hypothesis, we created a mixed-effect logistic regression model and then post-stratified based on the regression results to determine overall voter turnout. In the end, the model constructed contained one fixed-effect predictor and three random-effect predictors. The summary of the model suggested that the predictors likely have a linear relationship with the log-odds probability of whether or not a citizen would vote based on the survey data, thus our assumptions held true.

The result of the poststratification of the model onto the census data sample gave us a weighted mean probability of 0.956 that an eligible citizen would vote in the upcoming election and that 25,448,019 ballots will be cast. Based on this, the results that we obtained are not reasonable when real-world voter turnout is considered. The relatively small amount of data used in the model construction serves as a drawback. Additionally, the poststratification was not completed on the entire census, but only a small portion of it, and therefore we cannot assume with a high degree of confidence that the sample of the census data is an accurate representation of the population of Canada. This serves as a drawback in how applicable our model truly is.

However, despite the obvious drawbacks and unreasonable results of the model, we can make useful conclusions regarding future attempts at estimations. Possible ways to improve the methods include: finding surveys with greater amounts of data, using less one-sided response variables, using simpler models with fewer predictors, using a survey that has better overlap with census questions, and using a larger sample of the census data or perhaps the entirety of the census data if we possess enough computing power for poststratification.

6 Bibliography

1. Wikimedia Foundation. (2022, October 3). 2021 Canadian federal election. Wikipedia. Retrieved November 24, 2022, from https://en.wikipedia.org/wiki/2021_Canadian_federal_election
2. Wikimedia Foundation. (2022, October 28). 2022 Ontario general election. Wikipedia. Retrieved November 24, 2022, from https://en.wikipedia.org/wiki/2022_Ontario_general_election
3. Working age population: Aged 15-64: All persons for Canada. FRED. (2022, November 14). Retrieved November 24, 2022, from https://fred.stlouisfed.org/series/LFWA64TTCAM647S#0
4. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
5. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
6. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
7. Day, J. (2021, March 4). Representative democracy and government: Definition & future. Liberties.eu. Retrieved November 24, 2022, from https://www.liberties.eu/en/stories/representative-democracy/43508
8. Elections Canada (n.d.). Election results - national validated results. National Results Election Result Help. Retrieved November 24, 2022, from https://www.elections.ca/enr/help/national_e.htm
9. Statistics Canada. (2022, February 16). Table 2 reasons for not voting by age group and sex, 2021 federal election . Reasons for not voting by age group and sex, 2021 federal election.Retrieved November 25, 2022, from https://www150.statcan.gc.ca/n1/daily-quotidien/220216/t002d-eng.htm

Packages Used

1. tidyverse https://www.tidyverse.org/packages/
2. grid https://CRAN.R-project.org/package=grid
3. gridExtra https://cran.r-project.org/web/packages/gridExtra/index.html
4. knitr::kable https://rmarkdown.rstudio.com/lesson-7.html
5. brms https://cran.r-project.org/web/packages/brms/index.html
6. lme4 https://cran.r-project.org/web/packages/lme4/index.html
7. gtsummary https://cran.r-project.org/web/packages/gtsummary/index.html

7 Appendix

Table 2:

```r
# Use this to calculate some summary measures.
survey_data_clean %>%
  select('Citizen', 'Age', 'Sex', 'Province', 'Vote_Certainty', 'Income') %>%
tbl_summary(
    by = Vote_Certainty, # split table by var1
    missing = "no" # don't list missing data separately
  ) %>%
  add_n() %>% # add column with total number of non-missing observations
  # add_p() %>% # test for a difference between groups
  # bold_p(t = 0.05) %>%   # highlight signficant p
  modify_header(label = "**Variable**") %>% # update the column header
  bold_labels() %>%
  add_stat_label(
    label = all_categorical() ~ "No. (%)"
  )
```

Variable	**N**	**0**, N = 126	**1**, N = 2,793
Citizen, No. (%)	2,919	126 (100%)	2,793 (100%)
Age, Median (IQR)	2,919	44 (32, 59)	50 (38, 63)
Sex, No. (%)	2,919		
1		82 (65%)	1,606 (58%)
2		44 (35%)	1,187 (42%)
Province, No. (%)	2,919		
Alberta		9 (7.1%)	185 (6.6%)
British Columbia		29 (23%)	548 (20%)
Manitoba		8 (6.3%)	184 (6.6%)
New Brunswick		3 (2.4%)	142 (5.1%)
Newfoundland and Labrador		9 (7.1%)	135 (4.8%)
Nova Scotia		7 (5.6%)	146 (5.2%)
Ontario		23 (18%)	539 (19%)
Prince Edward Island		8 (6.3%)	139 (5.0%)
Quebec		22 (17%)	580 (21%)
Saskatchewan		8 (6.3%)	195 (7.0%)
Income, Median (IQR)	2,919	57,500 (27,250, 102,250)	80,000 (45,000, 140,000)

Table 3:

```r
# Use this to calculate some summary measures.
Estimates <- c("Min.", "1st Qu.", "Median", "Mean",
               "3rd Qu.", "Max", "Std.", "N")
Age_Survey <- c(round(summary(survey_data_clean$Age), 1),
                round(sd(survey_data_clean$Age), 1), nrow(survey_data_clean))
Age_Census <- c(round(summary(census_data_clean$age), 1),
                round(sd(census_data_clean$age), 1), nrow(census_data_clean))


table3 <- tibble(Estimates, Age_Survey, Age_Census)

knitr::kable(table3)
```

Estimates	Age_Survey	Age_Census
Min.	18.0	18.0
1st Qu.	37.0	34.8
Median	50.0	46.6
Mean	50.1	46.3
3rd Qu.	63.0	57.7
Max	95.0	80.0
Std.	16.0	14.2
N	2919.0	12089.0

Table 4:

```r
### constructing the summary table
Estimates <- c("Min.", "1st Qu.",  "Median", "Mean", "3rd Qu.", "Max.","SD")
Probability_to_Vote <- c(summary(census_data_clean1$estimate),
                         sd(census_data_clean1$estimate))

table4 <- tibble(Estimates, Probability_to_Vote)

knitr::kable(table4)
```

Estimates	Probability_to_Vote
Min.	0.8553533
1st Qu.	0.9475134
Median	0.9612961
Mean	0.9565591
3rd Qu.	0.9707166
Max.	0.9868980
SD	0.0201763

Figure 2:

```
surv <- survey_data_clean %>%
  ggplot(aes(y = Province))+
  geom_bar(fill = "darkturquoise", color = "Black")+
  labs(x = "Frequency", y = "Provinces",
  title = "Bar Chart to Show the Provincial Proportions
from the Phone Survey")+
  theme_bw()

cens <- census_data_clean %>%
  ggplot(aes(y = province))+
  geom_bar(fill = "darkturquoise", color = "Black")+
  labs(x = "Frequency", y = "Provinces",
  title = "Bar Chart to Show the Provincial Proportions
from the Census")+
  theme_bw()

grid.arrange(cens, surv)
```

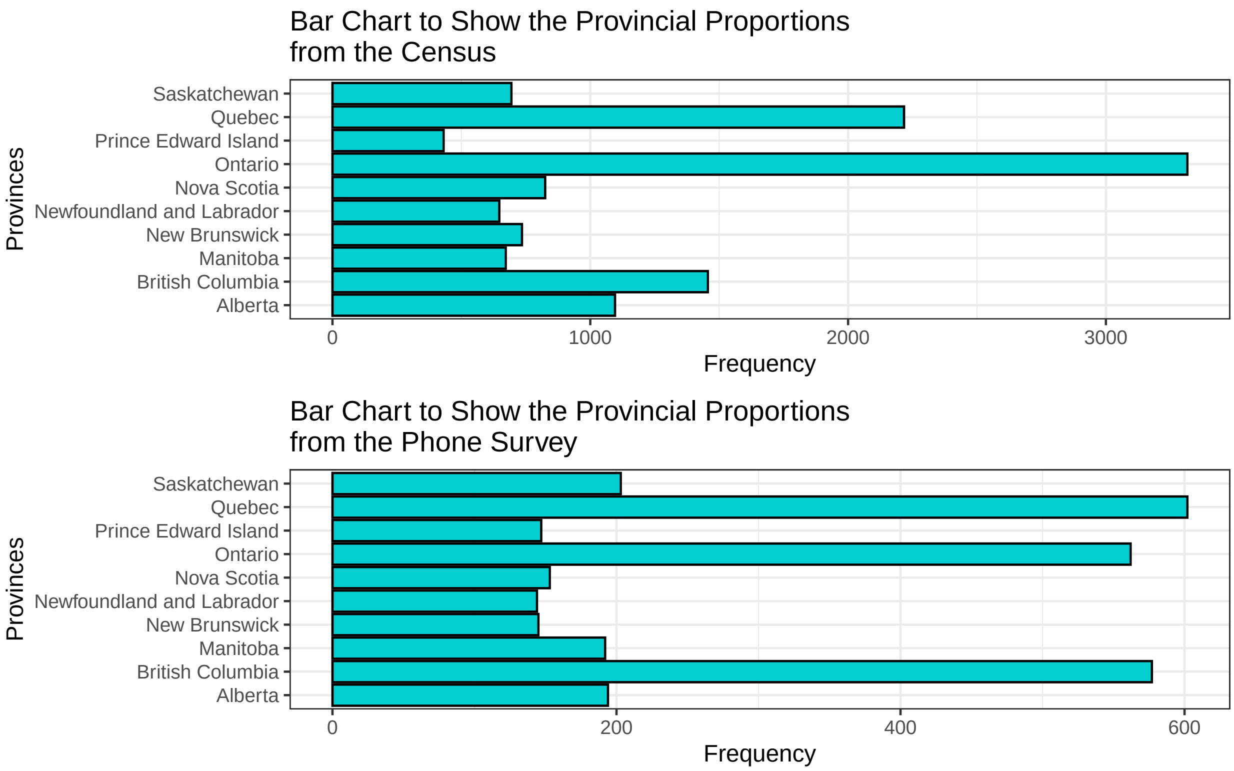


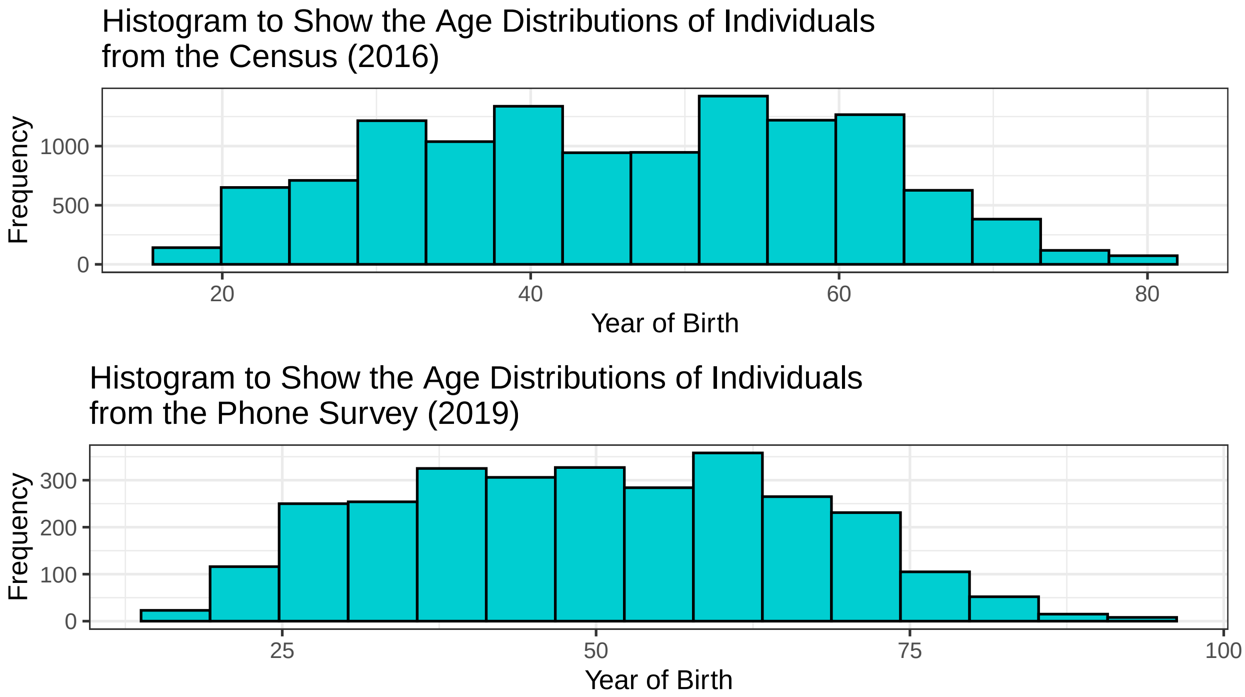
Figure 3:

```
graph1 <- survey_data_clean %>%
  ggplot(aes(x = Age))+
  geom_histogram(fill = "darkturquoise", color = "Black", bins = 15)+
  labs(x = "Year of Birth", y = "Frequency",
  title = "Histogram to Show the Age Distributions of Individuals
from the Phone Survey (2019)")+
  theme_bw()

census_data_clean <- census_data_clean %>%
  mutate(year = round(age))


graph2 <- census_data_clean %>%
  ggplot(aes(x = year))+
  geom_histogram(fill = "darkturquoise", color = "Black", bins = 15)+
  labs(x = "Year of Birth", y = "Frequency",
  title = "Histogram to Show the Age Distributions of Individuals
from the Census (2016)")+
  theme_bw()

grid.arrange(graph2, graph1)
```



Histogram to Show the Age Distributions of Individuals
from the Census (2016)

Histogram to Show the Age Distributions of Individuals
from the Phone Survey (2019)

Figure 4:

```
survey_data_clean %>%
  ggplot(aes(x = Age, group = Vote_Certainty)) +
  geom_boxplot(fill = "darkturquoise")+
  labs(title = "Boxplot to Show the Age Distributions of Individuals
from the Phone Survey",
       x = "Age",
       y = "Likeliness to vote",
       subtitle = "(0.2 = Likely, -0.2 = Not Likely)")+
  theme_bw()
```

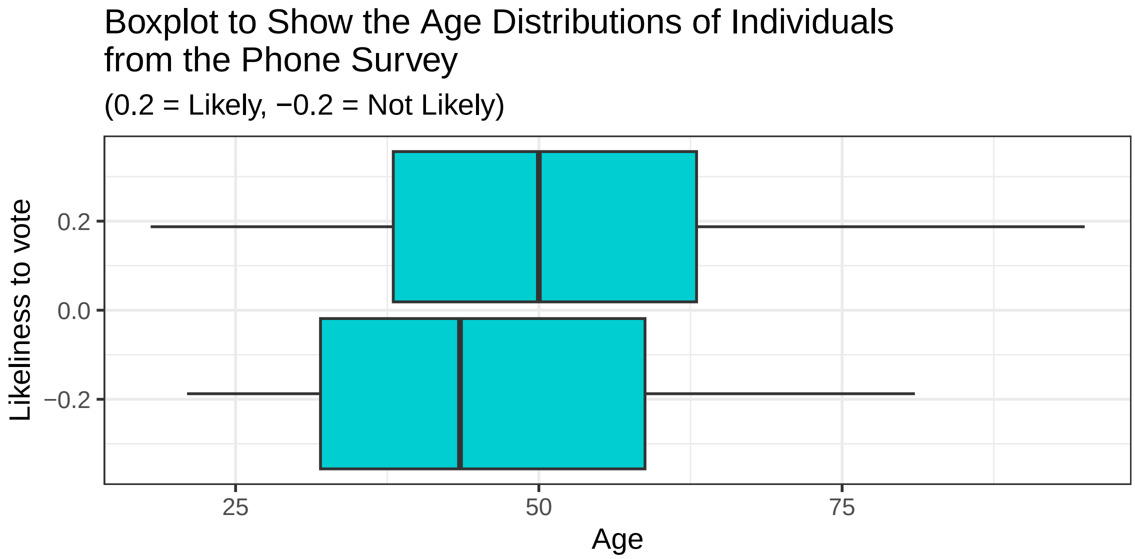
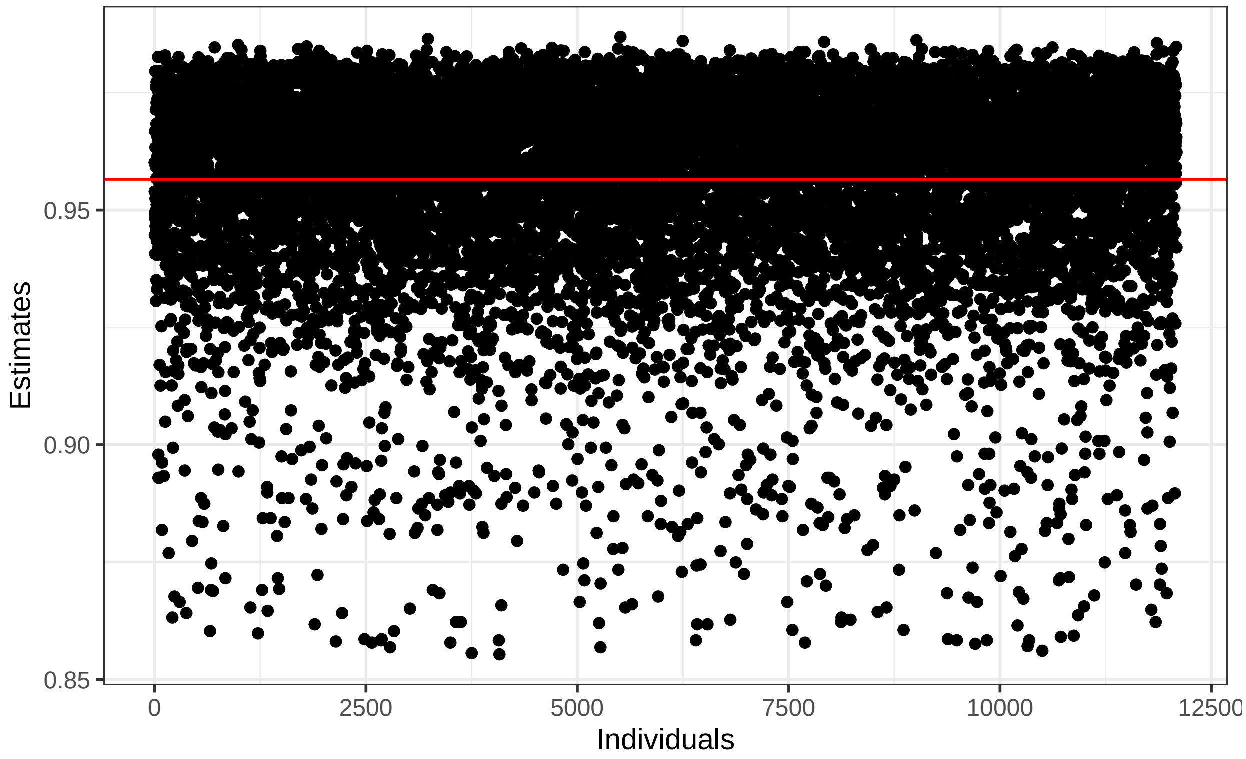


Figure 5:

```
### Scatterplot
census_data_clean1 %>%
  ggplot(aes(x = (1:nrow(census_data_clean1)), y= estimate))+
  geom_point()+
  geom_hline(yintercept = mean(census_data_clean1$estimate), colour = "Red")+
  labs(y = "Estimates", x = "Individuals",
       title = "Scatterplot Showing the Probability of an Individual Voting
on Election Day (Census)")+
  theme_bw()
```



Scatterplot Showing the Probability of an Individual Voting
on Election Day (Census)