Will the Liberals Win Again?

STA304 - Assignment 2

Peush Gomes

November 24, 2022

Introduction

Every couple years Canadians gather around from across the country to take part in the vote for the Canadian Federal Election. And with each new election to come around comes a lot of statistical data that can be analyzed. The benefit to all this data that is collected is that it provides areas to analyze and create models from that can help us actually predict the results of the future federal elections to come.

Using collected data from the lead up to the 2019 Canadian Federal Election we can create various models to predict the outcome of the next election. By using different variables like age, sex, province of residence, if the person was born in Canada, and their current marital status.

Our hypothesis for today is, do the provinces have half the proportional vote towards the liberal party? Will they win the next election?

Data

There are two different data sets that we apply the use of. For the census data we are using the 2013 General Social Survey. This data was collected through the use of conducting interviews on individuals over the age of 15 in Canada's ten provinces. These interviews were conducted from June 2013 and March 2014 via computer assisted telephone interviewing (CATI) and electronic questionnaires.

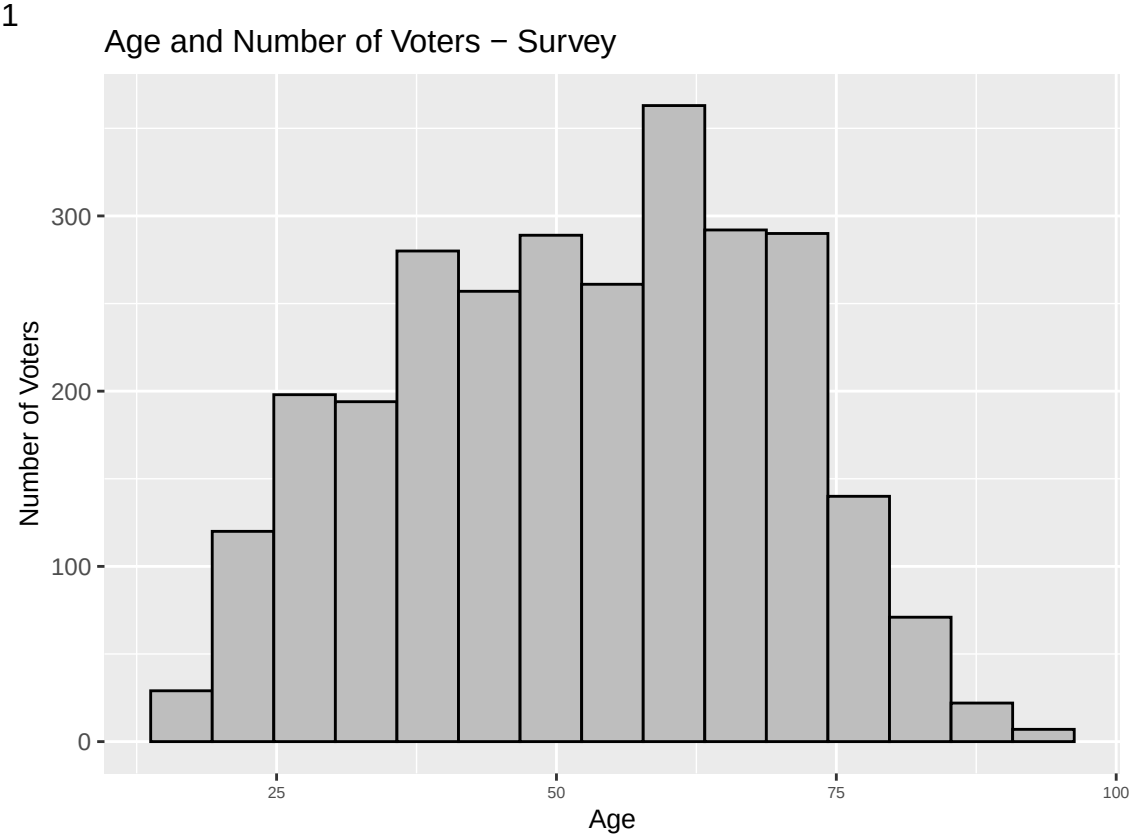
The second data set that we apply the use of is the 2019 Canadian Election Study. We use this data set for the survey data. This survey was done over the phone, both wireless telephone and land line numbers between September and October 2019.

We have 6 different important variables in our data sets that we are working with. The variable `age` is the age of the person when data was collected. `vote_liberal` is the variable that we are trying to predict, 1 corresponds to voting for liberal, 0 is against. `sex` is the sex of the person, this is either male or female, as no other sex was represented in the data sets. `province` is one of the 10 Canadian provinces, depended on province of residence. `place_birth_canada` is either 1 or 0, 1 meaning that the person is born in Canada, 0 is that they are not. The final variable is `marital status` representing the current marital status of the person, ranging from 'Single, Never married,' to 'Widowed'

Mean Age Survey	Mean Age Survey (Census)	Highest # of people in Province (Survey)	Highest # of people in Province (Census)	Second Highest # of people in Province (Survey)	Second Highest # of people in Province (Census)
50.89	52.18	Ontario: 807	Ontario:5621	BC:804	Quebec:3822

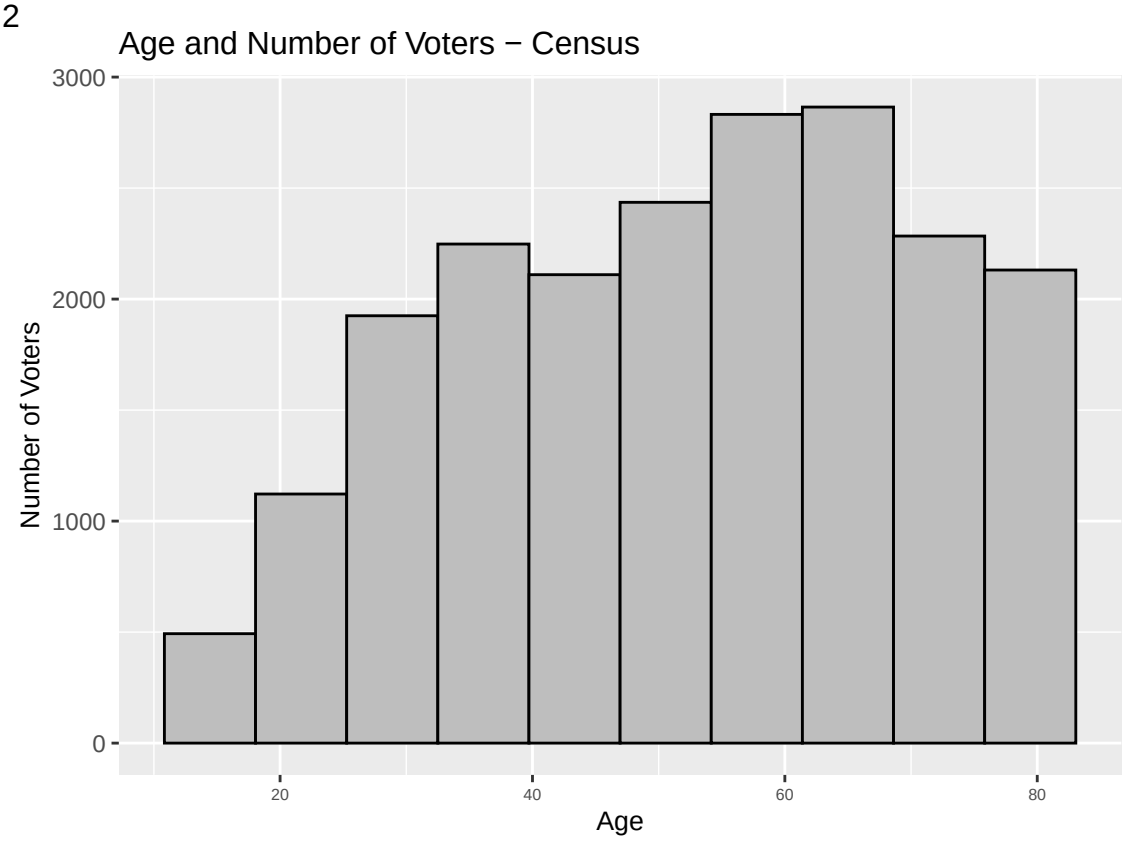
```
survey_data %>%
  ggplot(aes(x = age)) +
    geom_histogram(color="black", fill="gray", bins = 15) +
    labs(x = "Age",
         y = "Number of Voters",
         title = "Age and Number of Voters - Survey",
         tag = 'Figure 1') +
    theme(title = element_text(size = 10),
          axis.text.x  = element_text(size=6),
          axis.title.x = element_text(size=10))
```

Figure 1



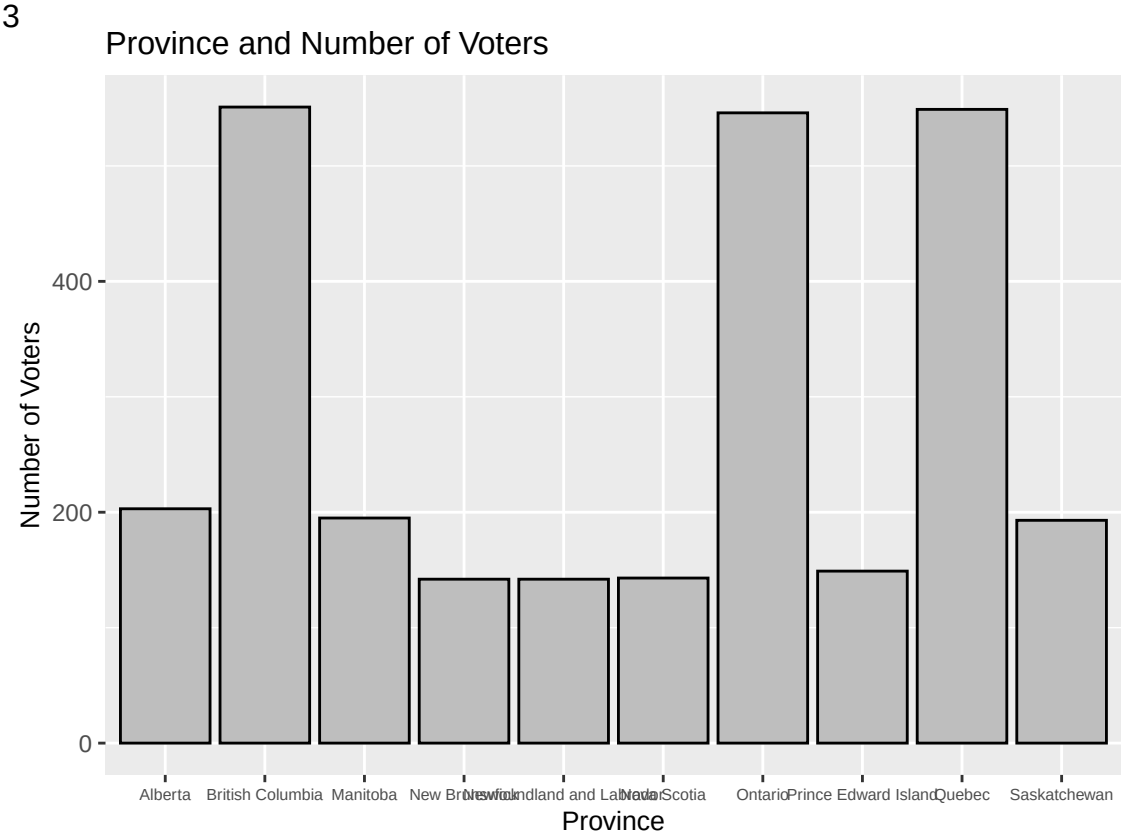

```
census_data %>%
  ggplot(aes(x = age)) +
    geom_histogram(color="black", fill="gray", bins = 10) +
    labs(x = "Age",
         y = "Number of Voters",
         title = "Age and Number of Voters - Census",
         tag = 'Figure 2') +
    theme(title = element_text(size = 10),
          axis.text.x  = element_text(size=6),
          axis.title.x = element_text(size=10))
```

Figure 2



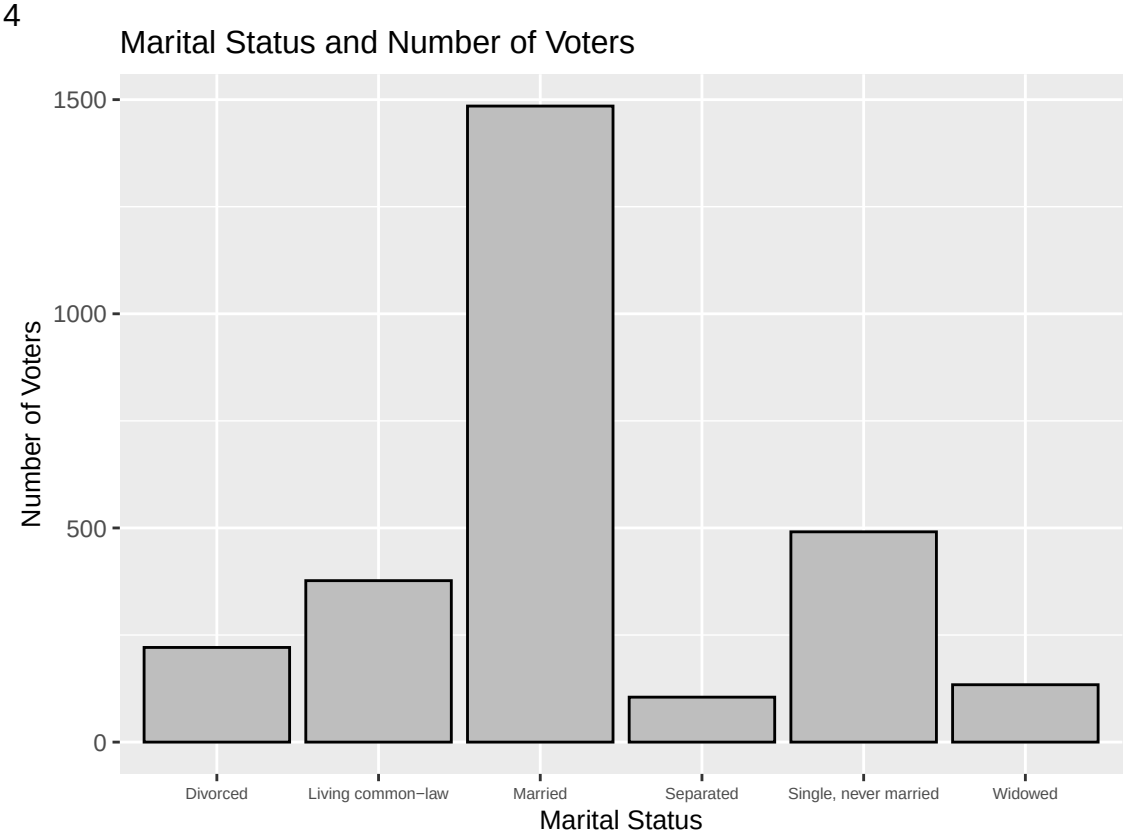

```
survey_data %>%
  ggplot(aes(x = province)) +
    geom_bar(color="black", fill="gray") +
    labs(x = "Province",
         y = "Number of Voters",
         title = "Province and Number of Voters",
         tag = 'Figure 3') +
    theme(title = element_text(size = 10),
          axis.text.x  = element_text(size=6),
          axis.title.x = element_text(size=10))
```

Figure 3




```
survey_data %>%
  ggplot(aes(x = marital_status)) +
    geom_bar(color="black", fill="gray") +
    labs(x = "Marital Status",
         y = "Number of Voters",
         title = "Marital Status and Number of Voters",
         tag = 'Figure 4') +
    theme(title = element_text(size = 10),
          axis.text.x  = element_text(size=6),
          axis.title.x = element_text(size=10))
```

Figure 4



In the table, we have some clear differences between our survey data and our census data. The mean age are similar, and the highest number of individuals for both in Ontario. However, when looking at the second highest number of individuals, in our survey, we have British Columbia with 804, right behind Ontario. But when looking at the Census data, BC is no where near the numbers of Ontario, and in second is actually Quebec. This may have a impact on our model which we will base on the survey data.

Figure 1 consists of a histogram showing the the number of voters per age, from the survey. The histogram clearly show a unimodal design, as well as a relative center around 50 years of age. There is a slight right skew that begins before 75 years old, representing that for the survey there are significantly less voters in their elder years compared to the younger.

Figure 2 shows very similar data, its now showing the number of voters per age, from the census data. The histogram is again uni modal in shape, its center being around the same mark of 50 years of age, but is showing slight left skew this time. As there are more number of voters in the upper more elder ages in the census when comparing to Figure 1.

Figure 3 shows the number of voters per province from the survey data. This shows that there were more correspondents to the survey from the provinces of British Columbia, Ontario, and Quebec. With those provinces having the largest amount of voters, with the Maritime provinces having the least of them all.

Figure 4, is also from the survey data. This data showing the number of voters per martial status. This shows that most of the voters are married with the second highest number of voters belonging to the marital status of single. This is disregarding the large amount of NAs that are present for this variable, all of which will be cleaned and removed. The lowest amount of voters for a status group is that of Separated, which has lower numbers than Widowed.

Methods

As our goal is to predict the outcome if a province were to vote for the liberals or not, we will conduct a create a logistic regression model. This model will be created based on the survey data, and then will be post stratified onto the census data.

Model Specifics

The logisitic regression model that will be fitted onto the survey data is a model with many predictor variables.

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 x_{age}+\beta_2 x_{Male}+\beta_3 x_{BC}+\beta_4 x_{MN}+\beta_5 x_{NB}+\beta_6 x_{NL}+\beta_7 x_{NS}+\beta_8 x_{ON}+\beta_9 x_{PEI}+\beta_{10} x_{QB}+\beta_{11} x_{SW}$$
$$+\beta_{12} x_{canadian\ born} + \beta_{13} x_{Divorced} + \beta_{14} x_{commonlaw} + \beta_{15} x_{seperated} + \beta_{16} x_{single} + \beta_{16} x_{Widowed}$$

Where p represents the probability of the intrest of voting for liberals to occur, β_0 represents the initial value taken with the log odds. β_1 represents the change in log odds for every one unit changed in age. β_2, represents the change in log odds if the individual is male. β_3 to β_{11} all represents the change in log odds depending on if the individual is a resident of a certain province. For example, if the individual is from Ontario, then the model will use $\beta_8 x_{ON}$, if they are from Alberta is will not use any of the betas from 3 to 11. β_{12}, represents the change in log odds if the individual is Canadian born. Then finally β_{13} to β_{16} are all changes in log odds depending on the martial status of the individual. If the individual is single, it will use β_{16}, and if they are Married it will not use any of the betas from 13 to 16.

Post-Stratification

The poststratification process is a process in where we fit a model based on a smaller data set, usually this is a dataset from a survey. We then use this model and apply it to a much bigger data set, in this case a census. We estimate the prediction for what we are looking for, for each individual cell. For this case, we are looking for the result of voting liberal for the provinces, so they our are cells. We extrapolate from our survey, and use this prediction for our entire population.

In order to estimate the proportion of voters we are predicting the proportion first on the survey, basing the prediction on age, sex, province, birth in Canada, and marital status. That prediction for each individual from our survey is then multiplied by our cell proportion of division total, to get a version of that prediction that is more inline with the entire population as it is now more weighted correctly

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

Results

```
# Creating the Model
model <- glm(vote_liberal ~ age + sex + province + place_birth_canada + marital_status, data=survey_data

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

```
##
## Call:
## glm(formula = vote_liberal ~ age + sex + province + place_birth_canada +
##     marital_status, data = survey_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.47036  -0.27290  -0.20176  -0.03586   0.98299
```

```
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       0.0595612  0.0559624   1.064 0.287281
## age                               0.0019547  0.0005553   3.520 0.000439 ***
## sexMale                          -0.0305414  0.0162192  -1.883 0.059799 .
## provinceBritish Columbia          0.1221052  0.0345968   3.529 0.000423 ***
## provinceManitoba                  0.1495699  0.0422652   3.539 0.000408 ***
## provinceNew Brunswick             0.1332261  0.0460795   2.891 0.003867 **
## provinceNewfoundland and Labrador 0.2136183  0.0460521   4.639 3.67e-06 ***
## provinceNova Scotia               0.2037531  0.0460320   4.426 9.95e-06 ***
## provinceOntario                   0.2458141  0.0346209   7.100 1.57e-12 ***
## provincePrince Edward Island      0.2248277  0.0454823   4.943 8.14e-07 ***
## provinceQuebec                    0.1585544  0.0348053   4.555 5.45e-06 ***
## provinceSaskatchewan              0.0621729  0.0423179   1.469 0.141894
## place_birth_canada               -0.0870821  0.0223186  -3.902 9.77e-05 ***
## marital_statusLiving common-law   0.0363486  0.0367544   0.989 0.322768
## marital_statusMarried             0.0047018  0.0305949   0.154 0.877873
## marital_statusSeparated          -0.0150706  0.0500718  -0.301 0.763452
## marital_statusSingle, never married  0.0509208  0.0363578   1.401 0.161461
## marital_statusWidowed            -0.0640444  0.0466206  -1.374 0.169634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1765823)
##
##     Null deviance: 513.55  on 2812  degrees of freedom
## Residual deviance: 493.55  on 2795  degrees of freedom
## AIC: 3125.2
##
## Number of Fisher Scoring iterations: 2
```

Province	# of People	Proportion of voting for liberal
Alberta	1708	0.085
British Columbia	2509	0.22074189
Manitoba	1180	0.23481835
New Brunswick	1318	0.21080144
Newfoundland and Labrador	1086	0.28747136
Nova Scotia	1412	0.28360752
Ontario	5585	0.34408096
Prince Edward Island	697	0.30080802
Quebec	3804	0.24886306
Saskatchewan	1147	0.14008389

From fitting our model onto the census data, and after poststratification. We arrive at some interesting results. Based on our model, no province is more than half as likely to vote for the liberal party in the next election. The low proportions are very expected as Alberta and the prairies have always voted against the liberals in heavy majority. Alberta has the lowest proportion with 0.085. The highest belonging to Ontario, with 0.344. However, it feels a little low for provinces like the Ontario and maritime provinces, as those have been seen to vote liberal more often. Ontario has usually been a even ground.

Conclusions

We have come to a conclusion for our analysis and data. The hypothesis was asking if the provinces have half the proportional vote towards the liberal party. To come to a conclusion for the hypothesis, we created a logistic regression model, based on predictor variables from our survey data. And then used the method of poststratification to bring those predictive survey results to a population level, and find predictions for the provinces as a whole. The results, were not surprising in some areas, and more surprising in some others.

The key results, is that each province based on our predictive model, had less than half of the proportion to vote for the liberal party in the next election. For provinces such as the prairies, a low proportion of liberal party voting is always expected. However, the proportion predicted does seem low for the Atlantic provinces, based on previous election results. And a bit low for Ontario, which has always been more of a middle ground for the conservative party and the liberal party.

There may have been some weakness to this method and analysis that we might not have considered. Our model, might have been over complicated. However, we do not fully know until we redo this analysis with many other different models. So for the future, it would be wise to revisit this hypothesis again. Then we could make different models either using more predictors or less, we could also make predictive models based on more previous election results. And conduct comparision between different models trained on different election years, to see if that changes any of our outcomes we have come to today.

Overall, with the model fit and hypothesis answered. That non of the 10 provinces have at least half of proportion to vote for the liberal party. We conclude that from our current model, that they will not win the next election.

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)