Project proposal

Approximately Normal: Jenny Huang, Bella Larsen, Jessie Bierschenk, Aidan Gildea

20 October 2020

```r
library(tidyverse)
library(broom)
library(patchwork)
```

```r
elections <- read_csv("data/data_full.csv")
```

Section 1. Introduction

In our project, we plan to investigate different factors and characteristics that appear to be involved or related to voters in the U.S. We are using information about voting behaviors in the U.S. over the past 14 years [1]. These data are sourced from IPUMS (an organization that provides census and survey data) and the American Statistical Association (ASA) [2, 3]. These data include 28 variables on more than 640,000 cases in the U.S. The data collected contains voter characteristics such as age, geographic location, sex, race, marital status, employment, citizenship, ethnicity, education, and voting history and tendencies [1]. With the upcoming November 2020 election, such information is particularly relevant in exploring what factors may be related to U.S. - Americans voting or not. By exploring factors most significant in relation to voting attendance, it is possible to further investigate ways to possibly increase voter turnout for future elections (such as the 2020 U.S. Presidential Election).

Motivation for our research comes from literature on the factors that are most important in predicting voting. In an MIT Election data study, researchers describe how understanding voter turnout is important when observing the particular tendencies of certain groups of people as well as factors that motivate individual U.S. citizens to vote [4]. The study highlights general assumptions of voter turnout predictions, noting how higher turnout rates tend to be related to individuals with the following traits: married, white, female, higher education, higher income, older age [4]. The article also addresses how reform may be able to increase voter turnout [4]. In another study done by Harvard graduate student Anthony George Fowler, voter turnout and its implications and repercussions are further examined in the U.S. as well as Australia and Mexico [5]. The study explores the 2008 U.S. election and addresses partisan gaps, voter knowledge (how politically-informed a voter is), and race as main variables of interest in exploring voter tendencies in the U.S. [5]. Both of these studies provide motivation for further and continued investigation into voter data and statistics – especially today in anticipation for the November 2020 election.

We are interested in predicting whether a person voted or not (and whether they were registered to vote), based on a list of predictor variables including sex, age, marital status, veteran status, citizenship status (native born or naturalized citizen), whether or not someone is Hispanic or Latinx, employment status and more (described in more detail in Section 2). Our proposed research question is: do voter turnout and registration rates depend on these predictors? Which predictors are more impactful than others? We are also interested in looking at voter turnout over time. We will use the predictor (year) to see if there are changes in voter turnout by demographics over time, or perhaps compare models from different time periods to determine how voter trends have changed over time. We hypothesize that the significant predictors of voter turnout will include age, level of education, whether they voted in previous elections, and race. Based on historical patterns, people in older age categories tended to vote more than people in younger age categories and people with higher levels of education tended to vote more frequently than people with lower levels of education [6]. Finally, if a person has voted previously, we predict they will be more likely to vote again compared to someone who has not voted previously.

Section 2. Data description

```
elections_clean <- elections %>%
  mutate(metro = if_else(METRO == 2, "Metro", "Not Metro/Unknown")) %>%
  mutate(sex = case_when(SEX == 1 ~ "Male",
                         SEX == 2 ~ "Female")) %>%
  mutate(marst = case_when(MARST == 1 ~ "Married",
                           MARST == 2 ~ "Married",
                           MARST == 4 ~ "Divorced/Separated",
                           MARST == 3 ~ "Divorced/Separated",
                           TRUE ~ "Not Married/Other")) %>%
  mutate(veteran = if_else(VETSTAT == 2, "Yes", "No/Uknnown")) %>%
  mutate(citizen = case_when(CITIZEN == 5 | CITIZEN == 9 ~ "No/Unknown",
                             CITIZEN == 1 | CITIZEN == 2 | CITIZEN == 3 ~
                               "Native Born",
                             CITIZEN == 4 ~ "Naturalized")) %>%
  #We should probably exclude people who are not citizens from our analysis
  #because they are not able to vote
  mutate(hispanic_status = if_else(HISPAN == 0 | HISPAN == 901 | HISPAN == 902,
                             "Not Hispanic/Unknown", "Hispanic/Latinx")) %>%
  mutate(employed = if_else(LABFORCE == 2, "Yes", "No/Unknown")) %>%
  mutate(highest_education = case_when(EDUC99 == 0 | EDUC99 == 1 ~
                                         "None/Unknown",
                               EDUC99 == 4 | EDUC99 == 5 | EDUC99 == 6 |
                                 EDUC99 == 7 | EDUC99 == 8 | EDUC99 == 9
                               ~ "Some High School",
                               EDUC99 == 10 ~ "High School Degree/GED",
                               EDUC99 == 11 ~ "Some College",
                               EDUC99 == 12 | EDUC99 == 13 | EDUC99 == 14
                               ~ "Associate Degree",
                               EDUC99 == 15 ~ "Bachelors Degree",
                               EDUC99 == 16 ~ "Masters Degree",
                               EDUC99 == 17 ~ "Professional Degree",
                               EDUC99 == 18 ~ "Doctorate Degree")) %>%
  mutate(current_student = case_when(SCHLCOLL == 5 | SCHLCOLL == 0 ~
                                       "No/Unknown",
                                     SCHLCOLL == 1  ~ "High School Full Time",
                                     SCHLCOLL == 2 ~ "High School Part Time",
                                     SCHLCOLL == 3 ~ "College Full Time",
                                     SCHLCOLL == 4 ~ "College Part Time")) %>%
#of people 16-24
  mutate(race = case_when(RACE == 820 | RACE == 830 | RACE == 830 | RACE == 819
                          | RACE == 804 | RACE == 805 | RACE == 806 |
                            RACE == 807 | RACE == 808 | RACE == 810 |
                            RACE == 811 | RACE == 812 | RACE == 813 |
                            RACE == 814 | RACE == 815 | RACE == 816 |
                            RACE == 817 | RACE == 818 | RACE == 803 |
                            RACE == 801 | RACE == 802 | RACE == 830 ~
                            "2 or more races",
                          RACE == 100 ~ "White",
                          RACE == 200 ~ "Black",
                          RACE == 651 | RACE == 809 | RACE == 652 | RACE == 650
                          ~ "Asian or Pacific Islander",
                          RACE ==  300 ~ "Native American",
                          RACE == 999 | RACE == 700 ~ "Other/Unknown")) %>%
  #Need to go over these race categorizations as a group!
  mutate(why_not_vote = case_when(VOWHYNOT == 10 ~ "Inconvenience",
                                  VOWHYNOT == 4 ~ "Interest",
                                  VOWHYNOT == 7 ~ "Political",
                                  VOWHYNOT == 9 | VOWHYNOT ==  6 |
                                    VOWHYNOT == 5 | VOWHYNOT == 2 ~
                                    "Logistical",
                                  VOWHYNOT == 1 ~ "Physically Unable",
                                  VOWHYNOT == 8 ~ "Registration Issues",
                                  VOWHYNOT == 3 ~ "Forgot")) %>%
#Reason why eligible voter did not vote
  mutate(why_not_reg = case_when(VOYNOTREG == 8 | VOYNOTREG == 3 ~
                                   "Not Eligible",
                                 VOYNOTREG == 7 | VOYNOTREG == 6 ~
                                   "Not Interested/Vote Won't Matter",
                                 VOYNOTREG == 5 ~ "Language Barrier",
                                 VOYNOTREG == 4 ~ "Physically Unable",
                                 VOYNOTREG == 2 | VOYNOTREG == 1 ~
                                   "Lacked Info/Missed Deadline")) %>%
  mutate(voting_method = case_when(VOTEHOW == 1 ~ "In Person",
                                   VOTEHOW == 2 ~ "Mail-In")) %>%
  mutate(voting_time = case_when(VOTEWHEN == 2 ~ "Early",
                                 VOTEWHEN == 1 ~"Voting Day")) %>%
  mutate(voted = case_when(VOTED == 1 ~ 0,
                           VOTED == 2 ~ 1)) %>%
  mutate(registered = case_when(VOREG == 1 ~ 0,
                           VOREG == 2 ~ 1)) %>%
  mutate(how_registered = case_when(VOREGHOW == 8 ~ "Online",
                                    VOREGHOW == 7 ~ "Polling Place",
                                    VOREGHOW == 6 ~ "Registration Drive",
                                    VOREGHOW == 5 | VOREGHOW == 2 |
                                      VOREGHOW == 1 ~
                                      "Govt Office/Public Agency",
                                    VOREGHOW == 4 ~ "School/College/Hospital",
                                    VOREGHOW == 3 ~ "By Mail")) %>%
  select(metro, sex, marst, veteran, citizen, hispanic_status, employed,
         highest_education, current_student, race, why_not_vote, why_not_reg,
         voting_method, voting_time, voted, registered, how_registered, YEAR,
         STATEFIP, AGE)

elections_clean <- elections_clean %>%
  mutate(metro = factor(metro),
         sex = factor (sex),
         marst = factor(marst),
         veteran = factor(veteran),
         citizen = factor(citizen),
         hispanic_status = factor( hispanic_status),
         employed = factor(employed),
         highest_education = factor(highest_education),
         current_student = factor(current_student),
         race = factor(race),
         why_not_vote = factor(why_not_vote),
         why_not_reg = factor(why_not_reg),
         voting_method = factor(voting_method),
         voting_time = factor(voting_time),
         voted = factor(voted),
         registered = factor(registered),
         how_registered = factor(how_registered))
```

```
observations <- as.tibble(count(elections_clean)) %>%
  mutate(variables = ncol(elections_clean))
observations %>%
  kable()
```

n	variables
643429	20

The data set was originally coded with different numerical values which would have made analysis difficult. Therefore, we recoded the data to make it more intelligible by using the codebook provided with the data download on Google Drive. [7] This was done in the data cleaning code section above, recoding the numerical values into string identifiers or binary numerical identifiers. The new dataset, `elections_clean`, is too large to load into the data folder and push to github, so we have included the raw dataset in the data folder, and the cleaned dataset can be attained by running the above code. In our data set `elections_clean`, there are 643,429 observations, each of which is a U.S. individual who participated in the survey over the past 14 years [1]. There are 20 variables in the data set, ranging from the race to the education of the respondents. The data is collected from samples of individuals that are surveyed in the U.S. every two years after November elections [1].

The primary response variables of our regression analysis will be `voted` and `registered`. `voted` is a categorical variable that identifies whether or not a respondent voted in the most recent November election [8]. Similarly, `registered` is a categorical variable that identifies whether or not a respondent registered to vote for the most recent November election [8]

To predict the odds of someone voting or registering to vote, we plan to consider predictor variables such as sex, age, marital status, veteran status, citizenship status (native born or naturalized citizen), whether or not someone is Hispanic or Latinx, employment status, whether someone lives in a metropolitan area, highest education level attained, whether someone is a current student, race, and the method of voter registration. In addition to these predictor variables, we are also interested in investigating the impact of variables such as year and voting time (early or voting day) [8]. Furthermore, we are interested in looking at variables that describe the reasons why someone did not register to vote and why someone did not vote to see whether these impact voting behavior in future elections, as this dataset includes elections over the last 14 years [8]. These data were originally collected from surveys conducted by the United States Census Bureau and the Bureau of Labor Statistics in a "Voting and Registration Supplement" of the Current Population Survey that occurs every two years after the November elections [1].

Section 3. Glimpse of data

```
glimpse(elections_clean)
```

```
## Rows: 643,429
## Columns: 20
## $ metro             <fct> Not Metro/Unknown, Not Metro/Unknown, Not Metro/U...
## $ sex               <fct> Male, Male, Female, Male, Female, Female, Male, M...
## $ marst             <fct> Divorced/Separated, Married, Married, Married, Ma...
## $ veteran           <fct> No/Uknnown, No/Uknnown, No/Uknnown, No/Uknnown, N...
## $ citizen           <fct> Native Born, Native Born, Native Born, Native Bor...
## $ hispanic_status   <fct> Not Hispanic/Unknown, Not Hispanic/Unknown, Not H...
## $ employed          <fct> Yes, No/Unknown, No/Unknown, Yes, Yes, No/Unknown...
## $ highest_education <fct> Some College, High School Degree/GED, Some High S...
## $ current_student   <fct> No/Unknown, No/Unknown, No/Unknown, No/Unknown, N...
## $ race              <fct> White, White, White, White, White, Black, Black, ...
## $ why_not_vote      <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ why_not_reg       <fct> NA, NA, NA, NA, NA, NA, Lacked Info/Missed Deadli...
## $ voting_method     <fct> In Person, NA, In Person, In Person, In Person, I...
## $ voting_time       <fct> Voting Day, NA, Voting Day, Voting Day, Voting Da...
## $ voted             <fct> 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1...
## $ registered        <fct> NA, 1, NA, NA, NA, NA, 0, 0, NA, NA, NA, NA, 0, N...
## $ how_registered    <fct> Govt Office/Public Agency, Govt Office/Public Age...
## $ YEAR              <dbl> 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2...
## $ STATEFIP          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ AGE               <dbl> 38, 70, 61, 60, 60, 37, 50, 38, 34, 31, 80, 42, 1...
```

References

[1]https://thisisstatistics.org/wp-content/uploads/2020/09/2020-Fall-Data-Challenge-DatasetFAQ-PDF.pdf

[2] https://ipums.org/what-is-ipums

[3] https://thisisstatistics.org/falldatachallenge/

[4] https://electionlab.mit.edu/research/voter-turnout

[5]https://dash.harvard.edu/bitstream/handle/1/11156810/Fowler_gsas.harvard_0084L_10773.pdf?sequence=3&isAllowed=y

[6] http://www.electproject.org/home/voter-turnout/demographics

[7] https://drive.google.com/file/d/1q4k2w6PXV8IIlsuiFYpElOlG23DNr7I6/view

[8]https://thisisstatistics.org/wp-content/uploads/2020/09/2020-Fall-Data-Challenge-Dataset-101-Glossary-v2.pdf