Predictive Polling*

The Impact of Demographics on Polling Outcomes

Raghav Bhatia

March 19, 2024

Model Construction

In this essay, we build a multiple logistic regression model given in the article "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results." (Cohn 2016) The model we specifically build is by Sam Corbett-Davies, Andrew Gelman and David Rothschild. Their aim was to predict the voting outcomes of their poll data, using a factor model that specifically utilized demographic factors such as race, gender and eduction. We replicate that approach by using the ces2020 dataset which contains voter outcomes and characteristics for Joe Biden and Donald Trump.

First, we download the ces dataset from the harvard dataverse database. The code for downloading the dataset is in 00-model_script. After downloading the dataset, we specifically use 3 variables as our input variables into the predictive model for voter outcomes:

- Gender
- Race
- Education

The code for cleaning and selecting the specific variables is given below:

```r
Voter_data_cleaned <-
  Voter_data_raw |>
  filter(votereg == 1, CC20_410 %in% c(1, 2)) |>
  mutate(
    voted_for = if_else(CC20_410 == 1, "Biden", "Trump"),
    voted_for = as_factor(voted_for),
    gender = if_else(gender == 1, "Male", "Female"),
    education = case_when(
```

*Code and data are available at: https://github.com/1raghav-bhatia/political-polling.git

```
      educ == 1 ~ "No HS",
      educ == 2 ~ "High school graduate",
      educ == 3 ~ "Some college",
      educ == 4 ~ "2-year",
      educ == 5 ~ "4-year",
      educ == 6 ~ "Post-grad"
    ),
    education = factor(education, levels = c("No HS", "High school graduate",
                                             "Some college", "2-year", "4-year",
                                             "Post-grad")),
    race = case_when(
      race == 1 ~ "White",
      race == 2 ~ "Black",
      race == 3 ~ "Hispanic",
      race == 4 ~ "Asian",
      race == 5 ~ "Native American",
      race == 6 ~ "Middle Eastern",
      race == 7 ~ "Two or more races",
      TRUE ~ NA_character_
    ),
    race = as_factor(race)
  ) |>
  select(voted_for, gender, education, race)
```

After cleaning the dataset, we use the demographic variables as inputs into our model and and voted_for as the response. We first select a random sample of the dataset. This allows us to preserve the characteristics for the model while speeding up calculations. Next, we regress gender, race and education on who the responded voted for. This allows us to have an historical precedent in order to base our predictions.

```
### Model data ####

## The example considers a sliced sample to improve the runtime of the model.

set.seed(853)

Voter_data_reduced <-
  Voter_data_cleaned |>
  slice_sample(n = 2000)

## Voter Outcomes Model
```

```
# This glm regresses voting outcome on  gender race and education

voter_outcomes_model <-
  stan_glm(
    voted_for ~ gender + race + education,
    data = Voter_data_reduced,
    family = binomial(link = "logit"),
    prior = normal(location = 0, scale = 2.5, autoscale = TRUE),
    prior_intercept =
      normal(location = 0, scale = 2.5, autoscale = TRUE),
    seed = 853
  )
```

Digression on Model Selection

Before we use the model to make predictions, we note that out model uses logitic regression as opposed to poisson regression and negative binomial regression as firstly, we only have 2 outcomes for the response. This means that a logitic regression is able to capture the response perfectly, and the multiple outcomes property of the other 2 models is not required. Poisson regression also has very restrictive assumptions such as an infinite population, which doesn't capture the reality of out situation. Similarly, negative binomial improves the applicability of the model, but isn't required for this simple model.

Polling Prediction

Using this model now, in line with Sam Corbett-Davies, Andrew Gelman and David Rothschild, we use it to make predictions for which respondents in our sample vote for Biden and Trump. We use the criterion that if a candidate as an estimate greater than 0.5, then they vote Biden. Our results are captured by the mean value of the 'model_prediction' variable, which is a high of 76.83%. Though our methodology was apt, such a high vote share for Biden is due to the innacuracies generated from the simplicity of our model.

References

Cohn, Nate. 2016. "We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results." *The New York Times*. https://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html.