2020 American Federal Election Popular Vote Prediction

Kaan Gumrah 1002916473

November 2, 2020

2020 American Federal Election Popular Vote Prediction using a Linear Regression Model with Post-Stratification

Kaan Gumrah 1002916473

Nov 2, 2020

Model

This paper aims to predict the popular vote result of the 2020 American federal election that will take place on November 3, 2020. The prediction will be between the two major party candidates for the election; Donald J. Trump for the Republican Party and Joe Biden for the Democratic Party. Data sets from Integrated Public Use Microdata Series (IPUMS) USA (2018 5-year ACS) and Democracy Fund + UCLA Nationscape (Full Data Set) will be used to conduct the prediction. To conduct the popular vote prediction, we are employing a post-stratification technique. In the following sub-sections I will describe the model specifics and the post-stratification calculation.

Model Specifics

I will be using a linear regression model to model a particular voter's probability to vote for Donald J. Trump. I will use years of education received, which is recorded as a numeric variable, as the only explanatory variable in predicting the probability of voting for Donald J. Trump. Estimates for years of schooling are based on the duration of schooling at each level of education, and are derived according to the definitions of Oxford Poverty & Human Development Initiate (OPHI) and Institute of Macroeconomic Analysis and Development. This model was selected to understand and observe the effects of each additional year of education on the probability to vote for Republican Party candidate Donald J. Trump for the 2020 American federal election.The simple linear regression model I am using is:

$$y = \beta_0 + \beta_1 x_{yearsofeducation} + \epsilon$$

Where y represents the probability of a particular person who attained x years of schooling to vote for Donald J. Trump. β_0 is the intercept of the model, and is interpreted as the probability of voting for Donald J. Trump with no years of schooling. β_1 is the slope of the linear regression model, and it is interpreted as the expected increase in the probability of voting for Donald J. Trump for each additional year of schooling received.

Post-Stratification

Generally, samples are not representative enough to make inferences regarding the population of interest. Post-stratification tackles this problem by categorizing each observation under various bins, with each bin representing some proportion of the population, and extrapolating the expected probability of behaving in a certain way to the entire population. This technique allows us to make a good correction for poorly represented samples, deeming inference on the population level more accurate. I will create bins based on

different years of schooling received to estimate the proportion of voters who will vote for Donald J. Trump for the 2020 American Presidential Election. Using the described linear regression model, I will estimate the proportion of voters for bins that represent different years of education received. Lastly, I will weight each proportion estimate within each cell by the respective population size of that bin and sum those values and divide that by the entire population size.

```
# Here I will perform the post-stratification calculation
census_data$estimate <-
  model %>%
  predict(newdata = census_data)

census_data %>%
  mutate(alp_predict_prop = estimate*n) %>%
  summarise(alp_predict = sum(alp_predict_prop)/sum(n))
```

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.349
```

Results

```
# The Model
model <- lm(DonaldTrumpVote ~ years_of_schooling,
            data=survey_data)

# Model Results
 summary(model)
```

```
##
## Call:
## lm(formula = DonaldTrumpVote ~ years_of_schooling, data = survey_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4719 -0.3801 -0.3572  0.5969  0.7576
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.219476   0.032733   6.705 2.19e-11 ***
## years_of_schooling 0.011476   0.002256   5.088 3.72e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4853 on 6473 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.003983,   Adjusted R-squared:  0.003829
## F-statistic: 25.89 on 1 and 6473 DF,  p-value: 3.724e-07
```

According to our estimation, the proportion of voters in favour of voting for Donald J. Trump of the Republican Party to be 0.35. This estimation is based on the discussed post-stratification analysis of the proportion of voters in favour of Donald J. Trump, modeled by the discussed linear regression model that accounted for years of schooling received.

Discussion

By creating a linear regression model that controls for years of schooling received and extrapolating our results to the entire population with weighted corrections according to the post-stratification technique, we found that each additional year of education received increases the probability to vote for the Republican presidential candidate Donald J. Trump by roughly 1.1% for the 2020 American Presidential Election. According to our estimate of voter proportion that will vote for Donald J. Trump, which stands at 35%, we predict that Joe Biden of the Democratic Party will win the 2020 Presidential Election.

Weaknesses

This analysis tried to predict the result of the 2020 Presidential election by just looking at years of education received, however, there are many other variables that affect voter decision, including but not limited to age, race, gender, income, state, etc. None of these variables were included in the analysis, therefore the conducted analysis is not a strong predictor.

Additionally,

Next Steps

Here you discuss subsequent work to be done after this report. This can include next steps in terms of statistical analysis (perhaps there is a more efficient algorithm available, or perhaps there is a caveat in the data that would allow for some new technique). Future steps should also be specified in terms of the study setting (eg. including a follow-up survey on something, or a subsequent study that would complement the conclusions of your report).

In order to improve the accuracy of the analysis, other explanatory variables that account for other significant variables can be added to the model. Such additions can improve the the accuracy of the prediction.

References

Gelman, Andrew. "Statistical Modeling, Causal Inference, and Social Science." Statistical Modeling, Causal Inference, and Social Science: 10 Jan. 2020, 9:13am, statmodeling.stat.columbia.edu/2020/01/10/linear-or-logistic-regression-with-binary-outcomes/.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [ACS 2018]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D010.V10.0

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, 2020(version 20200625)

Oxford Poverty & Human Development Initiative,"Training Material For Producing National Human Development Reports",2011, http://www.ophi.org.uk/wp-content/uploads/OPHI-RP-29a.pdf-2011.pdf

Kraigher, Tomaž. "Average Years of Schooling." Institute of Macroeconomic Analysis and Development. https://www.umar.gov.si/fileadmin/user_upload/publikacije/dr/07/ml/aMLPovpstletsolanja.pdf