Prediction of 2020 US election using Logistic Regression & Poststratification

Eung Kyu Kim, Dong Kyu Kim, Jiwon Chai

November 2nd, 2020

Model

Our model focused on major factors that might affect the election vote 2020 in the USA. Logistic regression and post-stratification methods were used to predict the outcome of the 2020 American federal election. To begin with, from the survey data we have selected several variables. The chosen variables are age and state. And we have used a logistic function to determine the probability of winning the election. In addition, post-stratification methods were used with census data. Census data included all the ages. However, we have removed people who are under 19 from the data for accuracy. As a result, we could have determined the approval rating of two candidates and the potential winner of the election.

Model Specifics

We will be using logistic regression for both Donald Trump and Joe Biden to model the relationship between the binary variables: ages / states and the dependent variable (vote for trump).

Figure 1, shows:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{AL} + ... + \beta_k x_k$$

beta_0 represents the intercept of the model and is the log of the odds when the age is 0 regardless of the state. beta_1 represents the slope of the logistic model, one unit increase in age, we expect 0.0171 increase in the probability of voting for Donald Trump. Similarly from beta_2 to beta_k, it indicates the change in the probability of voting for Donald Trump at Ith state.

For the sake of simplicity, we have calculated a person who lives in Alabama with the age of 50 voting for Donald Trump.

$$log(\frac{p}{1-p}) = -0.692424 + (0.017153 * 50) - (0.260172 * 1) + (0.062354 * 0) + ... - (0.466067 * 0)$$

After doing some math and isolating p, it gives that this person has a probability of 47.628% voting for Donald Trump.

$$log(\frac{p}{1-p}) = -0.82937 - (0.006639 * 50) - (0.7647 * 1) + (0.259325 * 0) + ... - (0.318721 * 0)$$

This gives us the value of p, which is 0.40212. This shows the probability of 40.212% voting for Biden. As a result, in the state AL with age 50. Donald Trump has a higher approval rating.

Post-Stratification

In the post-stratification analysis, our group has decided to calculate the y_ps value for each candidate: Donald Trump, Joe Biden. For finding the y_ps value, the cells were first created based on different ages greater than 18 because the voting age is greater than 18 in the US. Then, it was sub-divided by the state because the voting system in the US is determined by the number of won elections in each state. After finding

that, the total number of the same division of cells was counted. Then, the estimate of the logarithmic odds was predicted. Since we know that the general logistic regression equation, the estimate was calculated by:

$$\frac{e^{logodds.estimate}}{(1+e^{logodds.estimate})}$$

After finding the estimate values and population size of the jth cell, it can be plugged into the equation below to find the y_ps value.

$$\widehat{y}_{PS} = \frac{\sum N_j * \widehat{y}_j}{\sum N_j}$$

Then, the two values were compared to find which candidate is predicted to win in which state. After that, the number of won states was counted to find which candidate will be elected.

Results

In logistic regression, the positive value of age estimate indicates the older you are the higher probability of voting for Donald Trump. In Figure 1, the age estimate with a positive value of 0.171 indicates older people have a higher probability of voting for Donald Trump. On the other hand, the negative age estimate value with -0.006 shown in figure 2, indicates younger people will likely vote for Joe Biden.

Furthermore, for post-stratification, the calculated Yˆps values for each state are shown in figure 6. Trump_predicted represents the probability of voting for Trump and Biden_predicted presents the probability of voting for Biden. The person with a higher value gets points for each state. The result of each state is also shown in figure 6. As a result, among 50 states, Biden gets 24 states' votes compared to Trump with 26 states according to figure 7. Since the US election depends on who gets more state votes, we can anticipate that Donald Trump has a higher probability of winning the election.

Discussion

Our goal was to predict the winner of the 2020 US presidential election. We used two data sets: survey data is from IPUMS and census data is from Voter Study Group. We obtained this data and went through a data cleaning process. We filtered variables so we only work with useful data. We used two variables, 'State' and 'age'. We picked the 'State' variable because the state is one of the most important variables in the US election system, and we picked the 'age' variable because certain age groups tend to support specific leaders. We created two logistic regression models, one for vote_trump and one for vote_biden. After doing so, we used the summary() function to get all the estimates. Then we calculated each State's probability of voting Trump and voting Biden using a post stratification method. We used census data for this. We then compared each result to come up with the final winner.

After going through logistic regression, we got a positive estimate(0.0171) for the 'age' variable in the vote_trump model and a negative estimate(-0.0066) for 'age' variable in the vote_biden model which implies that the older you are, the more one will support Trump and the younger you are, the more one will support Biden. After going through the post-stratification calculation, we could compare the winning states between two candidates Trump and Biden. According to the 'compare' data in the appendix, Trump wins in 26 states while Biden wins in 24 states. Since the candidate who gets more states will win the election, according to our model and calculations, Trump will win the 2020 US presidential election.

The data we are using could be somehow biased. The data we are using(survey data and census data) are not representing the total population, they are only representing the sample population. Also, while most of the age groups have more than 30,000 population size, as age past 80, the sample size is smaller than the average which is less representative. Therefore the age group representation could be biased.

Weaknesses

Our biggest weakness was that we only used two predictors which are 'age' and 'state'. There are many other predictors that could possibly affect one's choice of which candidate to support such as ideology, gender,

and race. However, we did not consider those predictors because we wanted our prediction to be based on the age and the previous vote result in order to make our prediction as clear as possible. Therefore, our prediction could be not accurate enough in real-life situations. The other weakness is that our error range is not clear. Since we only used a few variables, the prediction is not accurate enough to be applied to the real world prediction. Therefore, our error range is not accurate and this unclearness brings difficulties when predicting the result in competitive states.

Next Steps

We will first look at the actual election result to evaluate our prediction and figure out how to improve our prediction for the next election. If our prediction is wrong, we will have to include many more various variables such as gender, ideology, education, and income to get a more accurate prediction that could be applied to real-world situations. Then we will look at the specific data from the actual 2020 US election to know what variables are most useful for the election prediction. We will create multiple different models with different variables and see which model had the closest prediction with the actual result so we can use it for the next prediction.

References

IPUMS USA. (2020). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. IPUMS USA. https://usa.ipums.org/usa/index.shtml.

Voter Study Group. (2020, September). Nationscape Data Set. https://www.voterstudygroup.org/publication/nationscape-data-set.

Appendix

Figure 1

```
# Model Number 1

model_glm <- glm(vote_trump ~ age + as.factor(state) ,
            data= survey_data, family= "binomial")
summary(model_glm)
```

```
##
## Call:
## glm(formula = vote_trump ~ age + as.factor(state), family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5294  -1.0275  -0.8449   1.2617   1.9960
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.692424   0.717676  -0.965   0.3346
## age                 0.017153   0.001717   9.993   <2e-16 ***
## as.factor(state)AL -0.260172   0.752943  -0.346   0.7297
## as.factor(state)AR  0.062354   0.786968   0.079   0.9368
## as.factor(state)AZ -0.299221   0.734284  -0.408   0.6836
## as.factor(state)CA -0.755673   0.719773  -1.050   0.2938
## as.factor(state)CO -0.268699   0.750022  -0.358   0.7202
```

```
## as.factor(state)CT -1.256973   0.772029  -1.628   0.1035
## as.factor(state)DC -0.533283   0.859744  -0.620   0.5351
## as.factor(state)DE -0.909517   0.822778  -1.105   0.2690
## as.factor(state)FL -0.496820   0.721231  -0.689   0.4909
## as.factor(state)GA -0.402808   0.732714  -0.550   0.5825
## as.factor(state)HI -1.057408   0.829572  -1.275   0.2024
## as.factor(state)IA -0.517504   0.774546  -0.668   0.5040
## as.factor(state)ID  0.149348   0.828347   0.180   0.8569
## as.factor(state)IL -0.666255   0.726673  -0.917   0.3592
## as.factor(state)IN -0.337182   0.743083  -0.454   0.6500
## as.factor(state)KS -0.143318   0.776685  -0.185   0.8536
## as.factor(state)KY -0.231655   0.751668  -0.308   0.7579
## as.factor(state)LA -0.341532   0.756509  -0.451   0.6517
## as.factor(state)MA -1.217763   0.750715  -1.622   0.1048
## as.factor(state)MD -0.723565   0.749449  -0.965   0.3343
## as.factor(state)ME -0.484301   0.869518  -0.557   0.5775
## as.factor(state)MI -0.576944   0.732610  -0.788   0.4310
## as.factor(state)MN -0.306089   0.756827  -0.404   0.6859
## as.factor(state)MO -0.468265   0.742187  -0.631   0.5281
## as.factor(state)MS -0.425227   0.786252  -0.541   0.5886
## as.factor(state)MT -0.107083   0.885952  -0.121   0.9038
## as.factor(state)NC -0.481205   0.730676  -0.659   0.5102
## as.factor(state)ND  0.609456   1.422581   0.428   0.6683
## as.factor(state)NE -0.744423   0.856606  -0.869   0.3848
## as.factor(state)NH -0.429161   0.860098  -0.499   0.6178
## as.factor(state)NJ -0.500349   0.731110  -0.684   0.4937
## as.factor(state)NM -1.427896   0.875409  -1.631   0.1029
## as.factor(state)NV -0.358027   0.761312  -0.470   0.6382
## as.factor(state)NY -0.437763   0.721057  -0.607   0.5438
## as.factor(state)OH -0.552595   0.726965  -0.760   0.4472
## as.factor(state)OK -0.375081   0.769730  -0.487   0.6261
## as.factor(state)OR -0.664622   0.750325  -0.886   0.3757
## as.factor(state)PA -0.305520   0.726682  -0.420   0.6742
## as.factor(state)RI -1.424325   1.068495  -1.333   0.1825
## as.factor(state)SC -0.037528   0.746287  -0.050   0.9599
## as.factor(state)SD  0.008220   0.884400   0.009   0.9926
## as.factor(state)TN -0.053140   0.746699  -0.071   0.9433
## as.factor(state)TX -0.180766   0.721951  -0.250   0.8023
## as.factor(state)UT -0.497630   0.782789  -0.636   0.5250
## as.factor(state)VA -0.605148   0.731822  -0.827   0.4083
## as.factor(state)VT -1.873420   1.051775  -1.781   0.0749 .
## as.factor(state)WA -0.626785   0.741520  -0.845   0.3980
## as.factor(state)WI -0.851776   0.746057  -1.142   0.2536
## as.factor(state)WV  0.100907   0.800206   0.126   0.8997
## as.factor(state)WY -0.466067   1.419546  -0.328   0.7427
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7359.1  on 5445  degrees of freedom
## Residual deviance: 7168.7  on 5394  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 7272.7
```

```
##
## Number of Fisher Scoring iterations: 4
```

Figure 2

```
# Model Number 2

model_glm2 <- glm(vote_biden ~ age + as.factor(state) ,
            data= survey_data, family= "binomial")
summary(model_glm2)
```

```
##
## Call:
## glm(formula = vote_biden ~ age + as.factor(state), family = "binomial",
##     data = survey_data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.956  -1.096  -0.907   1.224   1.736
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.832937   0.820092  -1.016   0.3098
## age                 -0.006639   0.001677  -3.960 7.49e-05 ***
## as.factor(state)AL   0.764700   0.851458   0.898   0.3691
## as.factor(state)AR   0.259325   0.892570   0.291   0.7714
## as.factor(state)AZ   0.708048   0.835304   0.848   0.3966
## as.factor(state)CA   1.197030   0.821552   1.457   0.1451
## as.factor(state)CO   0.853673   0.848555   1.006   0.3144
## as.factor(state)CT   1.499018   0.857291   1.749   0.0804 .
## as.factor(state)DC   1.636631   0.945905   1.730   0.0836 .
## as.factor(state)DE   1.302229   0.901334   1.445   0.1485
## as.factor(state)FL   0.902210   0.823193   1.096   0.2731
## as.factor(state)GA   0.957745   0.832979   1.150   0.2502
## as.factor(state)HI   1.585984   0.904704   1.753   0.0796 .
## as.factor(state)IA   0.848556   0.869216   0.976   0.3289
## as.factor(state)ID   0.074236   0.944108   0.079   0.9373
## as.factor(state)IL   1.032716   0.827293   1.248   0.2119
## as.factor(state)IN   0.785524   0.842982   0.932   0.3514
## as.factor(state)KS   0.495176   0.877376   0.564   0.5725
## as.factor(state)KY   0.896138   0.850018   1.054   0.2918
## as.factor(state)LA   0.855585   0.853922   1.002   0.3164
## as.factor(state)MA   1.401095   0.842111   1.664   0.0962 .
## as.factor(state)MD   1.262708   0.845350   1.494   0.1353
## as.factor(state)ME   1.266853   0.951240   1.332   0.1829
## as.factor(state)MI   1.149865   0.832501   1.381   0.1672
## as.factor(state)MN   1.208293   0.853917   1.415   0.1571
## as.factor(state)MO   0.753871   0.841632   0.896   0.3704
## as.factor(state)MS   0.864147   0.879064   0.983   0.3256
## as.factor(state)MT   0.788189   0.973375   0.810   0.4181
## as.factor(state)NC   1.078519   0.831041   1.298   0.1944
## as.factor(state)ND -11.430257 187.234587  -0.061   0.9513
## as.factor(state)NE   0.738462   0.936586   0.788   0.4304
```

```
## as.factor(state)NH   0.955521   0.945616   1.010   0.3123
## as.factor(state)NJ   0.906270   0.831603   1.090   0.2758
## as.factor(state)NM   1.295885   0.914625   1.417   0.1565
## as.factor(state)NV   0.864739   0.858087   1.008   0.3136
## as.factor(state)NY   1.012388   0.823034   1.230   0.2187
## as.factor(state)OH   0.902778   0.827960   1.090   0.2756
## as.factor(state)OK   0.321709   0.871441   0.369   0.7120
## as.factor(state)OR   0.935982   0.846921   1.105   0.2691
## as.factor(state)PA   0.592157   0.828793   0.714   0.4749
## as.factor(state)RI   1.514848   1.042005   1.454   0.1460
## as.factor(state)SC   0.334035   0.849643   0.393   0.6942
## as.factor(state)SD   0.136654   1.004892   0.136   0.8918
## as.factor(state)TN   0.605385   0.847294   0.714   0.4749
## as.factor(state)TX   0.602177   0.824346   0.730   0.4651
## as.factor(state)UT   0.184836   0.886024   0.209   0.8348
## as.factor(state)VA   1.192855   0.831548   1.434   0.1514
## as.factor(state)VT   2.865527   1.122553   2.553   0.0107 *
## as.factor(state)WA   0.969440   0.839420   1.155   0.2481
## as.factor(state)WI   1.151229   0.841711   1.368   0.1714
## as.factor(state)WV   0.512463   0.898526   0.570   0.5684
## as.factor(state)WY   0.318721   1.472843   0.216   0.8287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7491.2  on 5445  degrees of freedom
## Residual deviance: 7367.3  on 5394  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 7471.3
##
## Number of Fisher Scoring iterations: 11
```

Figure 3

```
# Data Frame 1
# comparing two values of prediction and showing who won in which state

compare <- data.frame(census_test$state, as.numeric(census_test$y_ps),
                      as.numeric(census_test2$y_ps2))

compare <- compare %>%
  rename(trump_predicted = as.numeric.census_test.y_ps.) %>%
  rename(biden_predicted = as.numeric.census_test2.y_ps2.) %>%
  mutate(who_will_win = ifelse(trump_predicted > biden_predicted, "Trump", "Biden"))

result <- compare %>%
  group_by( who_will_win ) %>%
  summarise(Republican = sum(who_will_win == "Trump"),
            Democratic = sum(who_will_win == "Biden"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Figure 4

```
#Calculating the yps values for each state for Trump

census_data$logodds_estimate <-
  model_glm %>%
  predict(newdata = census_data)

census_data$estimate <-
  exp(census_data$logodds_estimate)/(1+exp(census_data$logodds_estimate))

census_test <- census_data %>%
  mutate(y_ps_prop = estimate*n) %>%
  group_by(state) %>%
  summarise(y_ps = sum(y_ps_prop)/sum(n)) %>%
  rename(y_ps = y_ps)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Figure 5

```
#Calculating the y_ps values for Biden

census_data2$logodds_estimate2 <-
  model_glm2 %>%
  predict(newdata = census_data2)

census_data2$estimate2 <-
  exp(census_data2$logodds_estimate2)/(1+exp(census_data2$logodds_estimate2))

census_test2 <- census_data2 %>%
  mutate(y_ps_prop2 = estimate2 * n) %>%
  group_by(state) %>%
  summarise(y_ps2 = sum(y_ps_prop2)/sum(n)) %>%
  rename(y_ps2 = y_ps2)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Figure 6

```
#Table of predicting which candidate will win in which state by comparing y_ps values
as_tibble(compare)
```

```
## # A tibble: 50 x 4
##    census_test.state trump_predicted biden_predicted who_will_win
##    <chr>                       <dbl>           <dbl> <chr>
##  1 AK                          0.531           0.241 Trump
##  2 AL                          0.485           0.399 Trump
##  3 AR                          0.562           0.286 Trump
##  4 AZ                          0.474           0.386 Trump
##  5 CA                          0.357           0.509 Biden
##  6 CO                          0.473           0.424 Trump
```

```
##  7 CT                          0.261           0.580 Biden
##  8 DE                          0.339           0.529 Biden
##  9 FL                          0.434           0.429 Trump
## 10 GA                          0.443           0.448 Biden
## # ... with 40 more rows
```

Figure 7

```
#Table of which candidate is predicted to be elected
as_tibble(result)
```

```
## # A tibble: 2 x 3
##   who_will_win Republican Democratic
##   <chr>             <int>      <int>
## 1 Biden                 0         24
## 2 Trump                26          0
```