Predicting The 2020 American Federal Election Results

Cindy Gao(1005223410), Yanrong Huo(1004720965), Aiting Zhang(1004926066)

2 November 2020

Code and data supporting this analysis is available at: https://github.com/aiting zhang47/STA304-Group33-ProbelmSet3.git

Cleaning survey data

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## -- Column specification --------------------------------------------------------
## cols(
##   YEAR = col_double(),
##   SAMPLE = col_double(),
##   SERIAL = col_double(),
##   CBSERIAL = col_double(),
##   HHWT = col_double(),
##   CLUSTER = col_double(),
##   STATEFIP = col_double(),
##   STRATA = col_double(),
##   GQ = col_double(),
##   HHINCOME = col_double(),
##   PERNUM = col_double(),
##   PERWT = col_double(),
##   SEX = col_double(),
##   AGE = col_double(),
##   RACE = col_double(),
##   RACED = col_double(),
##   CITIZEN = col_double(),
##   EDUC = col_double(),
##   EDUCD = col_double()
## )
## [1] "51 to 60"   "31 to 40"   "61 to 70"   "20 or less" "21 to 30"
## [6] "above 80"   "41 to 50"   "71 to 80"
## [1] "41 to 50"   "71 to 80"   "51 to 60"   "21 to 30"   "31 to 40"
## [6] "61 to 70"   "above 80"   "20 or less"
```

```
## [1] "Male"   "Female"
```

```
## [1] "Female" "Male"
```

```
## [1] "Other race"                        "Black, or African American"
## [3] "White"                             "other asian or pacific islander"
## [5] "Chinese"                           "Japanese"
## [7] "American Indian or Alaska Native"
```

```
## [1] "White"                             "Black, or African American"
## [3] "other asian or pacific islander"   "Chinese"
## [5] "Japanese"                          "Other race"
## [7] "American Indian or Alaska Native"
```

```
##  [1] "$250,000 and above"   "$45,000 to $49,999"   "$200,000 to $249,999"
##  [4] "$175,000 to $199,999" "$80,000 to $84,999"   "$150,000 to $174,999"
##  [7] "$125,000 to $149,999" "$65,000 to $69,999"   "$35,000 to $39,999"
## [10] "$95,000 to $99,999"   "$30,000 to $34,999"   "$20,000 to $24,999"
## [13] "$25,000 to $29,999"   "$15,000 to $19,999"   "$85,000 to $89,999"
## [16] "$100,000 to $124,999" "Less than $14,999"    "$40,000 to $44,999"
## [19] "$60,000 to $64,999"   "$50,000 to $54,999"   "$70,000 to $74,999"
## [22] "$55,000 to $59,999"   "$90,000 to $94,999"   "$75,000 to $79,999"
```

```
##  [1] $75,000 to $79,999   $175,000 to $199,999 $65,000 to $69,999
##  [4] Less than $14,999    $80,000 to $84,999   $40,000 to $44,999
##  [7] $100,000 to $124,999 $20,000 to $24,999   $60,000 to $64,999
## [10] $15,000 to $19,999   $30,000 to $34,999   $150,000 to $174,999
## [13] $85,000 to $89,999   $90,000 to $94,999   $45,000 to $49,999
## [16] $200,000 to $249,999 $95,000 to $99,999   $25,000 to $29,999
## [19] $35,000 to $39,999   $125,000 to $149,999 $50,000 to $54,999
## [22] $55,000 to $59,999   $70,000 to $74,999   $250,000 and above
## 24 Levels: Less than $14,999 $15,000 to $19,999 ... $250,000 and above
```

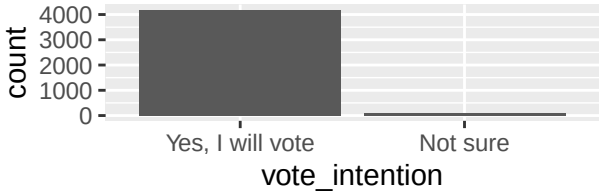
```
## [1] "Completed some college, but no degree"
## [2] "High school graduate"
## [3] "Doctorate degree"
## [4] "Completed some high school"
## [5] "Associate Degree"
## [6] "Masters degree"
## [7] "3rd Grade or less"
## [8] "Middle School - Grades 4 - 8"
```

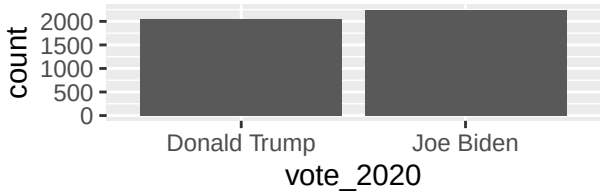
```
## [1] "Associate Degree"
## [2] "High school graduate"
## [3] "Completed some college, but no degree"
## [4] "Masters degree"
## [5] "Completed some high school"
## [6] "Doctorate degree"
## [7] "3rd Grade or less"
## [8] "Middle School - Grades 4 - 8"
```

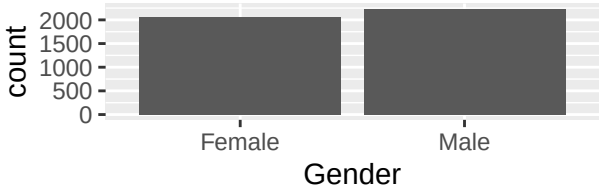
```
##  [1] "AL" "AK" "AZ" "AR" "CA" "CO" "CT" "DE" "DC" "FL" "GA" "HI" "ID" "IL" "IN"
## [16] "IA" "KS" "KY" "LA" "ME" "MD" "MA" "MI" "MN" "MS" "MO" "MT" "NE" "NV" "NH"
## [31] "NJ" "NM" "NY" "NC" "ND" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT"
## [46] "VT" "VA" "WA" "WV" "WI" "WY"
```

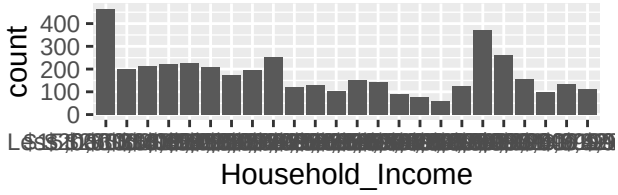
```
##  [1] "WI" "VA" "TX" "WA" "MA" "CA" "NC" "MD" "FL" "WV" "OH" "NY" "KY" "IN" "IA"
## [16] "SC" "MN" "GA" "PA" "NJ" "AZ" "IL" "OR" "MI" "CT" "MO" "CO" "DC" "NM" "TN"
```

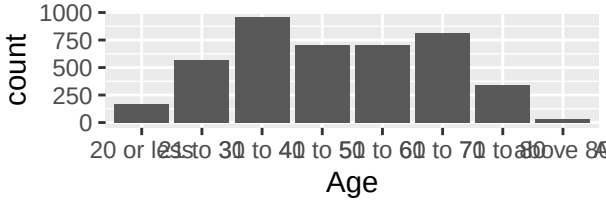
```
## [31] "OK" "HI" "MT" "VT" "UT" "NE" "NH" "NV" "ME" "ID" "LA" "MS" "KS" "AL" "AR"
## [46] "SD" "DE" "WY" "ND" "RI" "AK"
```

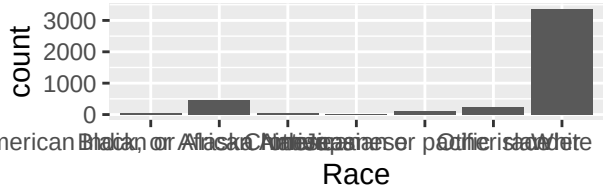


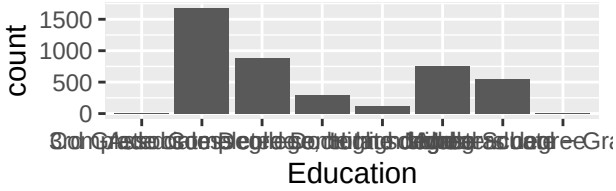


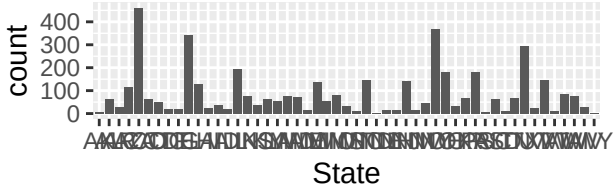













```
## Rows: 138,783
## Columns: 8
## $ PERWT            <dbl> 16, 56, 22, 16, 74, 75, 108, 22, 107, 28, 88, 13, ...
## $ Gender           <chr> "Male", "Male", "Male", "Male", "Female", "Female"...
## $ CITIZEN          <dbl> 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, 2, 1, 1, 2, 1,...
## $ Age              <chr> "51 to 60", "31 to 40", "61 to 70", "51 to 60", "2...
## $ Race             <chr> "Other race", "Black, or African American", "Black...
## $ Household_Income <chr> "$250,000 and above", "$250,000 and above", "$250,...
## $ Education        <chr> "Completed some college, but no degree", "High sch...
## $ State            <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "A...
```

Model

In this report, we will be building a multilevel logistic regression model with post-stratification to predict the overall popular vote of the 2020 American Federal Election. First, we created a multilevel logistic regression model with post-stratification off of the provided information from the Democracy Fund + UCLA Nationscape's Full Data Set's ns20200625.dta sample dataset and from the American Community Survey's 2018 5-year census. A multilevel logistic regression model with post-stratification was chosen for this study because the American Federal Election is very complex, and involves many different aspects and factors. Therefore, to have a better prediction of the outcome, it is better to gather more information, and from the chosen model, we are able to build a model where both level one/individual variables and level two/group variables are being considered.

Model Specifics

To build a multilevel logistic regression model to help predict which candidate has the overall popular vote in the current 2020 American Federal Election, we chose nine predictor variables. The first two selected predictor variables were "registration", and "vote_intention", which respectively tells us whether the person is eligible to vote and if they have any intentions to vote. This is important as only the people who are eligible and who may have an intention to vote will actually factor into this study, therefore, those who are not eligible and who have sternly stated they are not voting were removed from the data. Another important predictor variable that was extracted was "vote_2020", which essentially is the variable of interest, which is who the person will vote for in the 2020 American Federal Election. Here, another cleaning was performed, and only those who were voting for either Donald Trump or Joe Biden remained, and those who were undecided or were voting for someone else were removed. This is because currently the two main strongest competitors are Donald Trump and Joe Biden, thus it will be better to focus our model on them. To further analyze and study this case, one's age, gender, ethnicity, household income, education level, and home state, were also included as predictor variables to help us see if these few factors had any impacts on one's final choice/vote. Together, from these nine predictor variables, we now have data on an individual level and at a group overall level. After that, a cleaning was also performed on the census data, so that the variables selected from the census data also matched with the selected predictor variables from the sample/survey dataset.

In our model, we used the Frequentist approach, and created cells and focused on gender, age, and race, which will be later explained below. With that, we ran our model, and it can be seen that

$$\beta_0$$

, the intercept is 0.703094, meaning when all other predictor variables do not exist, the probability of Donald Trump winning will be 70.31%. Next, all the estimated under fixed values below the intercept,

$$\beta_0$$

, are our

$$\beta_1$$

to

$$\beta_n$$

, which represents the slopes of our variables in each group.

To test whether we have a strong model or not, we performed an area under the curve test. The result from the area under the curve test was 0.7239, which means our multilevel logistic model is strong and pretty accurate as 0.7239 is far greater than 0.5 which represents blindly guessing as there are only two choices (Donald Trump or Joe Biden).

```
## [1] 16
```

```
## [1] 16
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## Loading required package: Rcpp

## Loading 'brms' package (version 2.14.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following object is masked from 'package:lme4':
##
##     ngrps

## The following object is masked from 'package:stats':
##
##     ar

##
## Attaching package: 'tidybayes'

## The following objects are masked from 'package:brms':
##
##     dstudent_t, pstudent_t, qstudent_t, rstudent_t

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

## Warning in commonArgs(par, fn, control, environment()): maxfun < 10 *
## length(par)^2 is not recommended.

## Warning in optwrap(optimizer, devfun, start, rho$lower, control = control, :
## convergence code 1 from bobyqa: bobyqa -- maximum number of function evaluations
## exceeded

## Warning in (function (fn, par, lower = rep.int(-Inf, n), upper = rep.int(Inf, :
## failure to converge in 10000 evaluations

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.687599 (tol = 0.002, component 1)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : Model is nearly unide
##  - Rescale variables?

## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## vote_2020 ~ (1 + Gender + Race | cell) + Household_Income + Education +
```

```
##     State + Age
##    Data: data_survey
##
##      AIC      BIC   logLik deviance df.resid
##   5518.8   6308.1  -2635.4   5270.8     4172
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3691 -0.9066 -0.2469  0.8489  5.5090
##
## Random effects:
##  Groups Name                                Variance Std.Dev. Corr
##  cell   (Intercept)                          0.9713  0.9856
##         GenderMale                           0.2162  0.4650   -0.21
##         RaceBlack, or African American       8.0828  2.8430   -0.44 -0.78
##         RaceChinese                          4.0707  2.0176   -0.13 -0.89  0.87
##         RaceJapanese                        21.5290  4.6399    0.64 -0.69  0.26
##         Raceother asian or pacific islander  0.8744  0.9351   -0.68 -0.53  0.88
##         RaceOther race                       1.9525  1.3973   -0.76 -0.47  0.92
##         RaceWhite                            0.7970  0.8928   -0.99  0.09  0.54
##
##
##
##
##
##   0.29
##   0.80 -0.13
##   0.69 -0.11  0.94
##   0.25 -0.57  0.76  0.83
## Number of obs: 4296, groups:  cell, 16
##
## Fixed effects:
##                                                Estimate Std. Error z value
## (Intercept)                                    0.703094   1.409997   0.499
## Household_Income$15,000 to $19,999             0.006759   0.191326   0.035
## Household_Income$20,000 to $24,999             0.224069   0.187439   1.195
## Household_Income$25,000 to $29,999             0.215269   0.185130   1.163
## Household_Income$30,000 to $34,999             0.199156   0.182129   1.093
## Household_Income$35,000 to $39,999             0.227379   0.188533   1.206
## Household_Income$40,000 to $44,999             0.035280   0.199148   0.177
## Household_Income$45,000 to $49,999             0.350908   0.190936   1.838
## Household_Income$50,000 to $54,999             0.351852   0.178756   1.968
## Household_Income$55,000 to $59,999             0.496840   0.227676   2.182
## Household_Income$60,000 to $64,999             0.183445   0.222913   0.823
## Household_Income$65,000 to $69,999             0.351353   0.241233   1.456
## Household_Income$70,000 to $74,999             0.155768   0.212160   0.734
## Household_Income$75,000 to $79,999             0.465120   0.214339   2.170
## Household_Income$80,000 to $84,999             0.060063   0.256545   0.234
## Household_Income$85,000 to $89,999             0.166900   0.269320   0.620
## Household_Income$90,000 to $94,999             0.329014   0.292556   1.125
## Household_Income$95,000 to $99,999             0.126558   0.228783   0.553
## Household_Income$100,000 to $124,999           0.690397   0.164688   4.192
## Household_Income$125,000 to $149,999           0.547244   0.179494   3.049
## Household_Income$150,000 to $174,999           0.411366   0.212693   1.934
```

```
## Household_Income$175,000 to $199,999            0.938765   0.254484   3.689
## Household_Income$200,000 to $249,999            1.300426   0.249136   5.220
## Household_Income$250,000 and above              0.868512   0.242882   3.576
## EducationAssociate Degree                      -0.359880   0.826711  -0.435
## EducationCompleted some college, but no degree -0.162014   0.828039  -0.196
## EducationCompleted some high school             0.152774   0.834879   0.183
## EducationDoctorate degree                      -0.096425   0.854607  -0.113
## EducationHigh school graduate                   0.210098   0.828485   0.254
## EducationMasters degree                        -0.425934   0.832187  -0.512
## EducationMiddle School - Grades 4 - 8           0.302663   1.131301   0.268
## StateAL                                        -1.311936   1.140703  -1.150
## StateAR                                        -1.061976   1.189016  -0.893
## StateAZ                                        -1.427320   1.124256  -1.270
## StateCA                                        -2.046180   1.110889  -1.842
## StateCO                                        -1.791162   1.138751  -1.573
## StateCT                                        -2.549272   1.153243  -2.211
## StateDC                                        -2.079216   1.228264  -1.693
## StateDE                                        -2.415684   1.213173  -1.991
## StateFL                                        -1.609501   1.112028  -1.447
## StateGA                                        -1.197092   1.124994  -1.064
## StateHI                                        -2.041249   1.237292  -1.650
## StateIA                                        -1.885447   1.159914  -1.626
## StateID                                        -0.828752   1.224243  -0.677
## StateIL                                        -1.957143   1.117003  -1.752
## StateIN                                        -1.648258   1.133065  -1.455
## StateKS                                        -1.140635   1.169342  -0.975
## StateKY                                        -1.819429   1.137649  -1.599
## StateLA                                        -1.566107   1.145167  -1.368
## StateMA                                        -2.574309   1.136538  -2.265
## StateMD                                        -1.835419   1.137512  -1.614
## StateME                                        -2.123944   1.220396  -1.740
## StateMI                                        -1.897843   1.120931  -1.693
## StateMN                                        -1.766630   1.142660  -1.546
## StateMO                                        -1.693306   1.131468  -1.497
## StateMS                                        -0.901350   1.182661  -0.762
## StateMT                                        -1.523021   1.242551  -1.226
## StateNC                                        -1.649575   1.120161  -1.473
## StateND                                         9.517917  81.210279   0.117
## StateNE                                        -2.039686   1.236676  -1.649
## StateNH                                        -1.925255   1.229910  -1.565
## StateNJ                                        -1.735010   1.121356  -1.547
## StateNM                                        -2.534449   1.238782  -2.046
## StateNV                                        -1.206754   1.151835  -1.048
## StateNY                                        -1.857475   1.112496  -1.670
## StateOH                                        -1.708880   1.118096  -1.528
## StateOK                                        -1.251535   1.169369  -1.070
## StateOR                                        -1.942197   1.135307  -1.711
## StatePA                                        -1.502817   1.117134  -1.345
## StateRI                                        -2.621485   1.424304  -1.841
## StateSC                                        -1.003539   1.141947  -0.879
## StateSD                                        -1.285797   1.268092  -1.014
## StateTN                                        -1.093331   1.138688  -0.960
## StateTX                                        -1.245607   1.112808  -1.119
## StateUT                                        -1.476390   1.191514  -1.239
```

```
## StateVA                                          -1.922414   1.121587  -1.714
## StateVT                                          -4.069317   1.525673  -2.667
## StateWA                                          -2.208739   1.130742  -1.953
## StateWI                                          -2.165943   1.132635  -1.912
## StateWV                                          -1.356050   1.175026  -1.154
## StateWY                                          -2.082158   1.814352  -1.148
## Age21 to 30                                       0.530035   0.253586   2.090
## Age31 to 40                                       0.976969   0.239718   4.076
## Age41 to 50                                       0.960454   0.277038   3.467
## Age51 to 60                                       1.135840   0.248415   4.572
## Age61 to 70                                       0.894372   0.263611   3.393
## Age71 to 80                                       1.113156   0.291398   3.820
## Ageabove 80                                       1.194968   0.447664   2.669
##                                                  Pr(>|z|)
## (Intercept)                                      0.618026
## Household_Income$15,000 to $19,999               0.971818
## Household_Income$20,000 to $24,999               0.231922
## Household_Income$25,000 to $29,999               0.244912
## Household_Income$30,000 to $34,999               0.274181
## Household_Income$35,000 to $39,999               0.227800
## Household_Income$40,000 to $44,999               0.859388
## Household_Income$45,000 to $49,999               0.066087 .
## Household_Income$50,000 to $54,999               0.049029 *
## Household_Income$55,000 to $59,999               0.029093 *
## Household_Income$60,000 to $64,999               0.410539
## Household_Income$65,000 to $69,999               0.145258
## Household_Income$70,000 to $74,999               0.462827
## Household_Income$75,000 to $79,999               0.030006 *
## Household_Income$80,000 to $84,999               0.814890
## Household_Income$85,000 to $89,999               0.535449
## Household_Income$90,000 to $94,999               0.260751
## Household_Income$95,000 to $99,999               0.580142
## Household_Income$100,000 to $124,999             2.76e-05 ***
## Household_Income$125,000 to $149,999             0.002298 **
## Household_Income$150,000 to $174,999             0.053103 .
## Household_Income$175,000 to $199,999             0.000225 ***
## Household_Income$200,000 to $249,999             1.79e-07 ***
## Household_Income$250,000 and above               0.000349 ***
## EducationAssociate Degree                        0.663334
## EducationCompleted some college, but no degree 0.844876
## EducationCompleted some high school              0.854806
## EducationDoctorate degree                        0.910165
## EducationHigh school graduate                    0.799810
## EducationMasters degree                          0.608774
## EducationMiddle School - Grades 4 - 8            0.789057
## StateAL                                          0.250098
## StateAR                                          0.371774
## StateAZ                                          0.204238
## StateCA                                          0.065485 .
## StateCO                                          0.115738
## StateCT                                          0.027069 *
## StateDC                                          0.090492 .
## StateDE                                          0.046458 *
## StateFL                                          0.147797
```

```
## StateGA                                    0.287289
## StateHI                                    0.098990 .
## StateIA                                    0.104055
## StateID                                    0.498437
## StateIL                                    0.079750 .
## StateIN                                    0.145755
## StateKS                                    0.329337
## StateKY                                    0.109757
## StateLA                                    0.171444
## StateMA                                    0.023510 *
## StateMD                                    0.106628
## StateME                                    0.081794 .
## StateMI                                    0.090437 .
## StateMN                                    0.122088
## StateMO                                    0.134509
## StateMS                                    0.445978
## StateMT                                    0.220303
## StateNC                                    0.140853
## StateND                                    0.906701
## StateNE                                    0.099080 .
## StateNH                                    0.117498
## StateNJ                                    0.121805
## StateNM                                    0.040764 *
## StateNV                                    0.294786
## StateNY                                    0.094989 .
## StateOH                                    0.126417
## StateOK                                    0.284500
## StateOR                                    0.087132 .
## StatePA                                    0.178547
## StateRI                                    0.065689 .
## StateSC                                    0.379512
## StateSD                                    0.310601
## StateTN                                    0.336971
## StateTX                                    0.262996
## StateUT                                    0.215313
## StateVA                                    0.086526 .
## StateVT                                    0.007648 **
## StateWA                                    0.050778 .
## StateWI                                    0.055837 .
## StateWV                                    0.248476
## StateWY                                    0.251132
## Age21 to 30                                0.036603 *
## Age31 to 40                                4.59e-05 ***
## Age41 to 50                                0.000527 ***
## Age51 to 60                                4.82e-06 ***
## Age61 to 70                                0.000692 ***
## Age71 to 80                                0.000133 ***
## Ageabove 80                                0.007600 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation matrix not shown by default, as p = 88 > 12.
## Use print(x, correlation=TRUE)  or
```

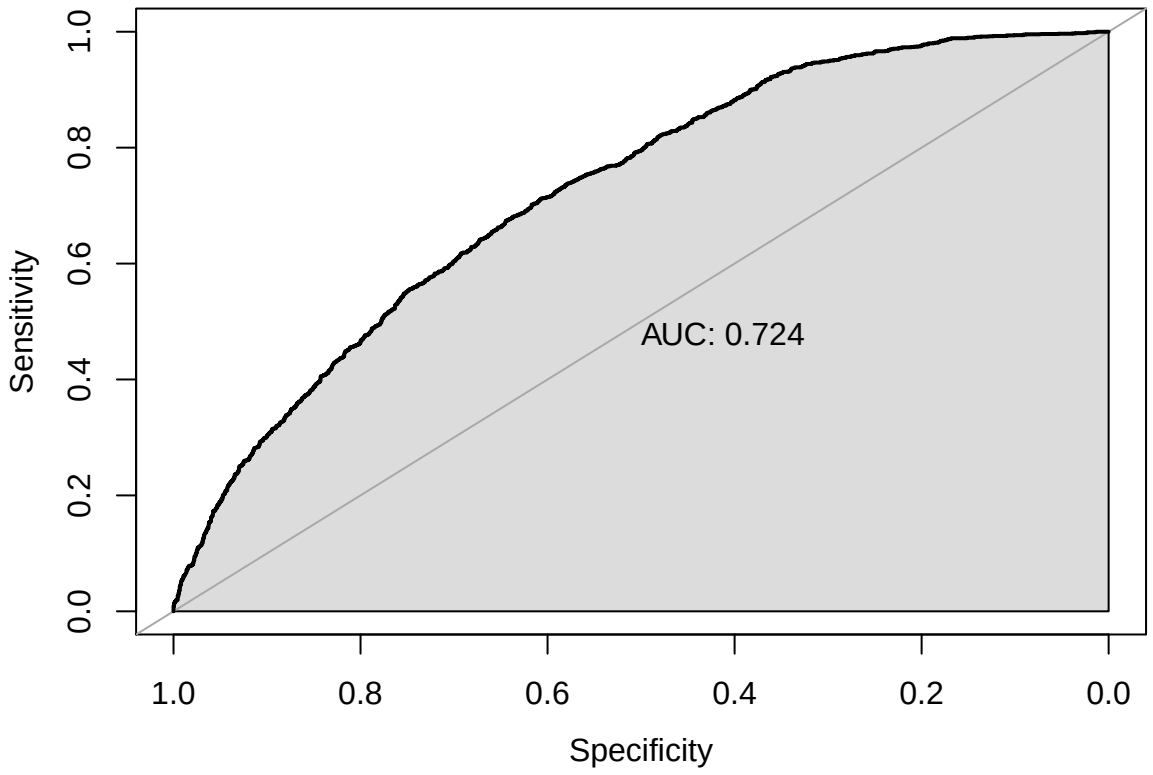
```
##     vcov(x)        if you need it

## convergence code: 0
## Model failed to converge with max|grad| = 0.687599 (tol = 0.002, component 1)
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
## failure to converge in 10000 evaluations

## Setting levels: control = Joe Biden, case = Donald Trump

## Setting direction: controls < cases

## Area under the curve: 0.7239
```



Post-Stratification

Post-stratification is used to reduce bias and error in our prediction. To perform post-stratification, we need to create cells based on demographics from the census population, and then apply the sample/survey estimate probabilities into each cell to estimate each response variable. Each estimate from each cell will differ, and thus we will have to combine them together by weighing them respectively to their proportions in the census population; this is known as the post-stratification estimate,

$$y^{PS}$$

. In our study, we created a cell of one's age and gender, because we believe that these two factors heavily influence how one thinks and how one will vote. For example, the boomer generation's mindset and generation Z's mindsets are completely different and will most likely cause differences in their votes. Gender plays a huge role in American politics, as Donald Trump is known for gender inequality and that he is more biased towards men, therefore men may have a higher probability of voting for Donald Trump than women. Along with that, another factor we also chose to heavily focus on was race in our model. This is because currently, racial injustice has become a very big issue in America, and it has thrown the entire country in uproar. Unlike Joe Biden, Donald Trump is not someone who actively supports racial justice, therefore we believe

this is currently an important factor to consider, as most likely Joe Biden will win the votes of the people of colour.

Factoring In The Electoral College

After computing the respective probabilities of both Joe Biden and Donald Trump votes, we applied their probabilities to the census dataset, and obtained the approximate amount of votes that they are each predicted to receive, with the person's weights already included. However, these are not our final numbers and/or proportions because in the United States of America, the Federal Election is not won by the majority of votes. Instead, the USA has a system called the "Electoral College", where each state has a fixed amount of electoral votes, totaling 538 electoral votes in across the country. To win the election a majority of 270 electoral votes are needed. How it works is that, whichever candidate wins the state, that candidate will receive all the electoral votes from that state and the loser will receive none, thus creating the problem of not having the candidate with the majority of the citizens votes becoming the new president. This is an important factor to consider, as this is how Donald Trump was able to win the 2016 American Federal Elections when Hillary Clinton had more citizen votes. To make our prediction more accurate, we decided to implement and simulate the Electoral College in our model.

```
## # A tibble: 1 x 1
##   alp_predict
##         <dbl>
## 1       0.460
```

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

```
## Rows: 2
## Columns: 2
## $ Winner <chr> "Donald Trump", "Joe Biden"
## $ Votes  <dbl> 295, 243
```

#Save the final datasets

Results

The calculation post-stratification estimate,

$$y^{PS}$$

, for our model was 0.4603834. This result is strictly based on citizen votes and not yet including the electoral votes. This means that after weighing the voters in each of our cells group, our aggregated prediction of the probability of Donald Trump winning is approximately 46.04% and the probability of Joe Biden winning is approximately 53.96%. This is a prediction strictly based on the nine predictor variables chosen for our model, and after dividing them into cells of age and gender.

Discussion

Throughout this analysis, we created and performed a multilevel logistic regression model to minimize the biases and errors when predicting the outcome of the 2020 American Federal Election. We used the post-stratification method to merge and use data from both our sample data set and our census, which represents the population. This way, we have real information coming in from the census, which will then result in a more accurate prediction and outcome of our model. With our model, focusing mainly on, one's age, gender and race, we computed a post-stratification estimate,

$$y^{PS}$$

, of 0.4603834 for Donald Trump and 0.5396166 for Joe Biden.

Near the beginning of this study, we also drew a barplots for each graph measuring their respective counts. It can be seen that in the one's voting intention barplot, Joe Biden was shown to lead through our sample data. From simply calculating

$$y^{PS}$$

, and not including the electorial votes, we can see that that is indeed the case, as Joe Biden had a

$$y^{PS}$$

of 53.96. However, in the end, with all the data combined together, and with the inclusion of the Electoral College, our final prediction is that Donald Trump will win the 2020 American Federal Election by roughly 55% (295/538) and Joe Biden will lose by roughly 45%. With both of these results, these show that the American Federal Election system is a very complex system, and that there are many different factors and aspects that may play an influential role in the final result of the election.

Weaknesses

As it can be seen, the American Federal Election is an incredibly complex system with many different and complicated factors influencing it. To create this study and prediction model, we had to simplify the situation more and focus on the main factors affecting the election. Therefore, there definitely exists some weaknesses in our model that could potentially influence our final result in the wrong direction.

One obvious factor to keep in mind is that the census used in this study was from 2018, thus meaning that the data is already two years old. Since those two years, there now exists new eligible voters, and voters may have different thoughts and opinions than what they did in 2018. Therefore, the first weakness is that our census is not up to date. Following that, in the beginning, when selecting and cleaning our predictor variables, a lot of small information was deleted. For example, in our model, we only considered people who will vote for Donald Trump or Joe Biden and deleted anyone who was undecided and anyone who was going to vote for another candidate. This definitely could have swayed the results in another direction as there was a possibility that at the last minute, people decided to vote and/or switch to vote for either Donald Trump or Joe Biden. There also is a possibility that the majority of the people who were undecided ended up voting for Joe Biden, and thus potentially resulting in the opposite of our model's predicted winner, Donald Trump. Along with that, another potential weakness is that we had to simplify our datasets and model so that it could actually run in R. As stated before, we did this by mainly only focusing on gender, age and race, as we thought they had significant influences on the election outcome.Although it was proved that we have a strong model through the area under the curve, these are simply based on our opinions, and there may be an even stronger models out there with different predictor variables that may just simply result in a completely different conclusion than ours.

Next Steps

In the future, to further improve our model, we can start by tackling our weaknesses. Finding a more recent and more up to date census or annual population survey will result in a much more reliable and accurate result. Luckily for us, another way to improve our model is to compare our model's predicted results with the actual results that will occur on November 3, 2020. Not only can we see if we have predicted the correct winner or not, but we can also see how accurate our data actually is. We will have the chance to see if our nine predictor variables did indeed heavily impact the votes, or not. In our modern day and age, we will have access to fresh and reliable data detailing everything that will happen in the actual election, thus we can also directly see and analyze what we predicted correctly and what we missed when creating our model.

References

1. Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/downloads?key=412bda07-e177-4cd0-92fa-fe6ff4739dd2.

2. Team, M. (n.d.). U.S. CENSUS DATA FOR SOCIAL, ECONOMIC, AND HEALTH RESEARCH. Retrieved November 02, 2020, from https://doi.org/10.18128/D010.V10.0

3. United States Electoral College Votes by State. (n.d.). Retrieved November 02, 2020, from https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124

4. List of All 50 US State Abbreviations. (n.d.). Retrieved November 02, 2020, from https://abbreviations.yourdictionary.com/articles/state-abbrev.html