

Predicting the USA 2020 Federal Election

A Generalized linear model used to determine the popular vote of the 2020 election.

Prinsa Gandhi

November 2, 2020

Abstract

This report will be analyzing the Democracy Fund + UCLA nationscape “Full Data Set” to predict the popular vote of the 2020 American Federal Election with post-stratification performed using the American Community Surveys (ACS) data of 2018. This report will use model selection techniques to determine a prediction model for the binary response variables “*vote_trump*” and “*vote_biden*”, and try to predict the outcome using the ACS data. The datasets have been cleaned and used for exploratory data analysis, to aid in creating the model. The results are very interesting, and it would be very interesting to compare these results to the federal election results.

The code and data supporting this analysis is available at:

<https://github.com/username1-p/Elections-Model>

Introduction

I used R to perform my analysis on the survey data (R Core Team (2020)). The data used to create my model is Democracy Fund + UCLA Nationscape ‘Full Data Set’ (Tausanovitch, Chris and Lynn Vavreck). This dataset had many interesting variables and information that could be used. In order to test my data and obtain my prediction result, I performed post-stratification with a data extract from IPUMS USA, with variables that I used in my model (IPUMS USA, University of Minnesota, www.ipums.org). I created a model to determine the votes for Donald Trump, and a model to obtain the votes for Joe Biden, and the results are very interesting.

Model

In my analysis, the goal is to predict the popular vote outcome for the 2020 American Federal Election (<https://www.usa.gov/voting>). I will describe the model used, and the post-stratification method used in order to obtain my estimates.

Model Specifics

The model I have chosen to use is a generalized linear model which is a binary logistic regression model. I created 2 different models, in order to predict a result for response variables *vote_trump* or *vote_biden*. I decided to choose this model because my outcome variable in the model is a binary response variable which I had given value of either 0 or 1, and thus, a logistic regression would make most sense. My binary response variable is dependent on some predictor variables which I obtained using model selection techniques. Before

model selection, I determined which variables I had to use which were present in my survey data and census data, because they had to match in order to perform post-stratification. I performed both AIC and BIC model selection, and determined the most significant variables. I compared models using the ANOVA in order to determine the model with the least residuals, and to determine significant predictors one-by-one after the AIC and BIC selection techniques. The link used in this model is the logit link. The assumptions for this model are satisfied. The cases are independent in both the survey and census data, since each person was interviewed once. The dependent variable is distributed by a binomial distribution. The model I chose is shown below:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{\text{AgeGroup}} + \beta_2 x_{\text{Race}} + \beta_3 x_{\text{hispanic}} + \beta_4 x_{\text{state}} + \beta_5 x_{\text{education}} + \beta_6 x_{\text{ForeignBorn}} + \beta_7 x_{\text{gender}}$$

which is modeling

$$\text{logit}(\pi_i)$$

the log-odds of the probability of success for the response variable Y(either vote_trump or vote_biden) dependent on the explanatory variables such as gender, and the rest of the variables shown. The β_0 is the model intercept, and the β_1, \dots, β_7 are the model coefficients for each of the predictors.

Here are my model results shown in a condensed format. It is shown in complete form in the appendix. Table 1-Logistic Model for Vote_Trump

```
broom::tidy(logmodeltrump)
```

```
## # A tibble: 77 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept) -0.375    0.981    -0.383    0.702
## 2 stateAL     -0.166    0.760    -0.218    0.827
## 3 stateAR      0.132    0.786     0.167    0.867
## 4 stateAZ     -0.380    0.744    -0.511    0.609
## 5 stateCA     -0.797    0.730    -1.09     0.275
## 6 stateCO     -0.474    0.756    -0.628    0.530
## 7 stateCT     -1.47     0.782    -1.89     0.0594
## 8 stateDC     -0.769    0.881    -0.873    0.382
## 9 stateDE     -0.930    0.841    -1.11     0.269
## 10 stateFL    -0.368    0.731    -0.504    0.615
## # ... with 67 more rows
```

Table 2-Logistic Model for Vote_Biden

```
broom::tidy(logmodelbiden)
```

```
## # A tibble: 77 x 5
##   term      estimate std.error statistic p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept) -1.20     1.05    -1.14     0.254
## 2 stateAL      0.266    0.856     0.311    0.756
## 3 stateAR     -0.342    0.894    -0.383    0.702
## 4 stateAZ      0.391    0.841     0.465    0.642
## 5 stateCA      0.734    0.828     0.886    0.376
## 6 stateCO      0.423    0.852     0.497    0.619
## 7 stateCT      1.13     0.861     1.32     0.188
```

```
## 8 stateDC      1.45      0.950      1.53      0.127
## 9 stateDE      1.09      0.913      1.19      0.233
## 10 stateFL     0.484     0.830     0.584     0.559
## # ... with 67 more rows
```

Post-Stratification

Post-stratification is a statistical technique which allows a model to be used on a new dataset, and allows estimates to be corrected according to their sampling weights. Then, an estimate is calculated for the entire population. The estimate can be calculated for the prediction model, which can be very useful in modeling behaviour or other various outcomes, and determining an estimate to make important future decisions. In my analysis, this allows me to estimate the proportion of voters who will vote for Donald Trump and in the second calculation, the proportion of voters who will vote for Joe Biden.

I will be using a post-stratification technique which allows me to use my model on the census dataset to predict estimates for certain groups that I have chosen. I will use the weight of each estimate for each group and determine the population level estimate, and this is very useful to allow a model to be used for prediction purposes and weight each estimate. This means bias will decrease because underrepresented groups will be adjusted for with the sample weight, and also leads to smaller variance. This leads to a more accurate prediction.

In this model, first I split the cells by age group, gender, state, race, education, whether Hispanic or not, and whether foreign born or not. I will partition the census data into demographic cells by grouping them based on my predictors which are age group, gender and state. These groups of these cells allow an estimate for each possible group (each datapoint/row) by using my model to obtain and estimate, and then I weight the estimate by its relative proportion within the whole population and aggregate all estimates to population level estimates. I weight each estimate by its population size of the bin, and find the sum of all these values for each bin. Finally, I divide this summation by the total population count.

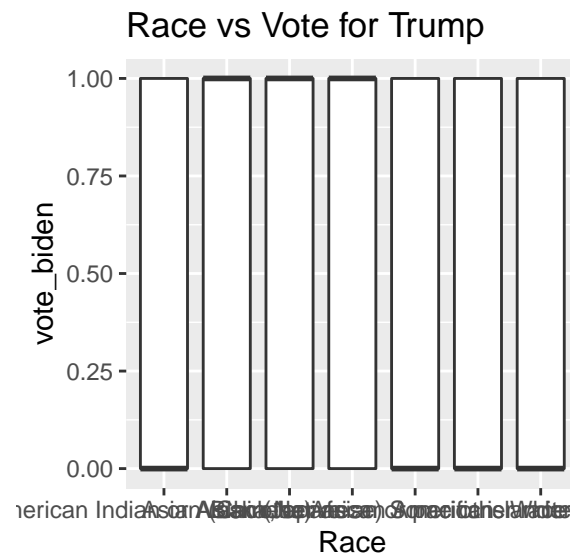
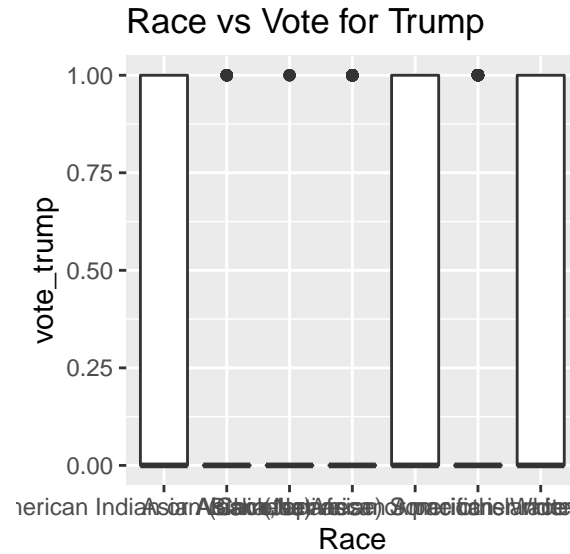
Grouping the Cells

I decided to split my cells based on the variables that I used in my model, because they are all extremely significant variables statistically as well as based on real life outcomes. I chose the three most significant variables to group the cells. For example, state is important because it is shown that certain states are more likely to vote for certain parties. The age group is important, because different ages might have different perspectives and would vote based on what issues are important to them. The gender is important because many voters may want to vote based on issues relating to gender such as gender equality, and this would influence voter outcome as well. Thus, my cell groupings are based on age group, gender, and state because I thought these variables were the most significant to group.

Results

Here you will include all results. This includes descriptive statistics, graphs, figures, tables, and model results. Please ensure that everything is well formatted and in a report style. You must also provide an explanation of the results in this section.

Figure 1



While analyzing figure 2, it can be seen that Race is quite an important predictor in determining whom one will vote for. It seems that Race is not impacting votes for Biden, but it is very much impacting the votes for Trump.

Please ensure that everything is well labelled. So if you have multiple histograms and plots, calling them Figure 1, 2, 3, etc. and referencing them as Figure 1, Figure 2, etc. in your report will be expected. The reader should not get lost in a sea of information. Make sure to have the results be clean, well formatted and digestible.

Discussion

Here you will summarize the previous sections and discuss conclusions drawn from the results. Make sure to elaborate and connect your analysis to the goal of the study.

Weaknesses

Here we discuss weaknesses of the study, data, analysis, etc. You can also discuss areas for improvement.

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).

Appendix

Here is my full model summary.

Model 1-Logistic Regression for Vote_Trump

```
summary(logmodeltrump)
```

```
##
## Call:
## glm(formula = vote_trump ~ state + education + foreign_born +
##      gender + Race + hispanic + agegrp, family = "binomial", data = survey_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6450  -1.0273  -0.5435   1.1382   2.6614
##
## Coefficients:
##                                     Estimate Std. Error z value
## (Intercept)                    -0.37543     0.98113  -0.383
## stateAL                       -0.16592     0.76039  -0.218
## stateAR                        0.13152     0.78589   0.167
## stateAZ                       -0.38034     0.74387  -0.511
## stateCA                       -0.79734     0.73004  -1.092
## stateCO                       -0.47431     0.75569  -0.628
## stateCT                       -1.47402     0.78197  -1.885
## stateDC                       -0.76937     0.88086  -0.873
## stateDE                       -0.92988     0.84108  -1.106
## stateFL                       -0.36823     0.73122  -0.504
## stateGA                       -0.06084     0.74440  -0.082
## stateHI                       -0.60686     0.84692  -0.717
## stateIA                       -0.66759     0.77738  -0.859
## stateID                       -0.03456     0.81943  -0.042
## stateIL                       -0.65030     0.73626  -0.883
## stateIN                       -0.52327     0.74981  -0.698
## stateKS                       -0.22560     0.78253  -0.288
## stateKY                       -0.28426     0.75666  -0.376
## stateLA                       -0.14443     0.76500  -0.189
## stateMA                       -1.30447     0.75849  -1.720
```

## stateMD	-0.65353	0.75885	-0.861
## stateME	-0.86042	0.85862	-1.002
## stateMI	-0.71315	0.74256	-0.960
## stateMN	-0.37655	0.76607	-0.492
## stateMO	-0.55811	0.74989	-0.744
## stateMS	-0.08351	0.79748	-0.105
## stateMT	-0.47468	0.87622	-0.542
## stateNC	-0.39274	0.73973	-0.531
## stateND	-0.38032	1.06865	-0.356
## stateNE	-0.51917	0.85131	-0.610
## stateNH	-0.78912	0.86835	-0.909
## stateNJ	-0.57738	0.74054	-0.780
## stateNM	-1.35181	0.86719	-1.559
## stateNV	-0.24336	0.76908	-0.316
## stateNY	-0.62958	0.73149	-0.861
## stateOH	-0.62256	0.73546	-0.846
## stateOK	-0.30379	0.76928	-0.395
## stateOR	-0.78652	0.75842	-1.037
## statePA	-0.45704	0.73573	-0.621
## stateRI	-1.01105	1.00135	-1.010
## stateSC	-0.05576	0.75471	-0.074
## stateSD	-0.32318	0.88576	-0.365
## stateTN	-0.08291	0.75021	-0.111
## stateTX	-0.16630	0.73157	-0.227
## stateUT	-0.62444	0.78132	-0.799
## stateVA	-0.64922	0.74146	-0.876
## stateVT	-2.09020	1.05413	-1.983
## stateWA	-0.75274	0.75002	-1.004
## stateWI	-0.87676	0.75192	-1.166
## stateWV	-0.26262	0.79507	-0.330
## stateWY	-1.81583	1.33954	-1.356
## educationAssociate Degree	-0.84292	0.63366	-1.330
## educationCollege Degree (such as B.A., B.S.)	-0.72496	0.62948	-1.152
## educationCompleted some college, but no degree	-0.74407	0.62959	-1.182
## educationCompleted some graduate, but no degree	-0.63989	0.64226	-0.996
## educationCompleted some high school	-0.65342	0.63286	-1.032
## educationDoctorate degree	-0.13226	0.65207	-0.203
## educationHigh school graduate	-0.65670	0.63023	-1.042
## educationMasters degree	-0.56861	0.63298	-0.898
## educationMiddle School - Grades 4 - 8	-1.02739	0.78445	-1.310
## educationOther post high school vocational training	-0.51714	0.63782	-0.811
## foreign_born	0.37716	0.12493	3.019
## genderMale	0.41732	0.05647	7.390
## RaceAsian (Chinese)	-0.90584	0.38183	-2.372
## RaceAsian (Japanese)	-0.88276	0.63415	-1.392
## RaceBlack, or African American	-1.88301	0.26518	-7.101
## Raceother asian or pacific islander	-0.24788	0.28387	-0.873
## RaceSome other race	-0.32552	0.26182	-1.243
## RaceWhite	0.18670	0.23420	0.797
## hispanic	-0.34768	0.09454	-3.678
## agegrp26-35	0.47707	0.10873	4.388
## agegrp36-45	0.71015	0.10785	6.585
## agegrp46-55	0.85077	0.11369	7.483
## agegrp56-65	0.79505	0.11163	7.122

## agegrp66-75	0.74633	0.11928	6.257
## agegrp76-85	1.04649	0.19049	5.494
## agegrp86-95	1.52165	0.72519	2.098
##	Pr(> z)		
## (Intercept)	0.701981		
## stateAL	0.827274		
## stateAR	0.867089		
## stateAZ	0.609145		
## stateCA	0.274752		
## stateCO	0.530235		
## stateCT	0.059428	.	
## stateDC	0.382427		
## stateDE	0.268911		
## stateFL	0.614555		
## stateGA	0.934864		
## stateHI	0.473652		
## stateIA	0.390465		
## stateID	0.966357		
## stateIL	0.377098		
## stateIN	0.485263		
## stateKS	0.773115		
## stateKY	0.707157		
## stateLA	0.850256		
## stateMA	0.085465	.	
## stateMD	0.389119		
## stateME	0.316299		
## stateMI	0.336857		
## stateMN	0.623054		
## stateMO	0.456720		
## stateMS	0.916597		
## stateMT	0.587997		
## stateNC	0.595475		
## stateND	0.721923		
## stateNE	0.541965		
## stateNH	0.363481		
## stateNJ	0.435582		
## stateNM	0.119033		
## stateNV	0.751672		
## stateNY	0.389412		
## stateOH	0.397278		
## stateOK	0.692916		
## stateOR	0.299718		
## statePA	0.534463		
## stateRI	0.312644		
## stateSC	0.941104		
## stateSD	0.715218		
## stateTN	0.912003		
## stateTX	0.820176		
## stateUT	0.424164		
## stateVA	0.381245		
## stateVT	0.047382	*	
## stateWA	0.315560		
## stateWI	0.243607		
## stateWV	0.741164		


```

## stateWY 0.175239
## educationAssociate Degree 0.183442
## educationCollege Degree (such as B.A., B.S.) 0.249456
## educationCompleted some college, but no degree 0.237268
## educationCompleted some graduate, but no degree 0.319101
## educationCompleted some high school 0.301847
## educationDoctorate degree 0.839268
## educationHigh school graduate 0.297405
## educationMasters degree 0.369021
## educationMiddle School - Grades 4 - 8 0.190299
## educationOther post high school vocational training 0.417486
## foreign_born 0.002536 **
## genderMale 1.47e-13 ***
## RaceAsian (Chinese) 0.017676 *
## RaceAsian (Japanese) 0.163914
## RaceBlack, or African American 1.24e-12 ***
## Raceother asian or pacific islander 0.382535
## RaceSome other race 0.213758
## RaceWhite 0.425365
## hispanic 0.000235 ***
## agegrp26-35 1.15e-05 ***
## agegrp36-45 4.55e-11 ***
## agegrp46-55 7.26e-14 ***
## agegrp56-65 1.06e-12 ***
## agegrp66-75 3.92e-10 ***
## agegrp76-85 3.93e-08 ***
## agegrp86-95 0.035880 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 8619.4 on 6474 degrees of freedom
## Residual deviance: 7763.4 on 6398 degrees of freedom
## (4 observations deleted due to missingness)
## AIC: 7917.4
##
## Number of Fisher Scoring iterations: 5

```

Model 2-Logistic Regression for Vote_Biden

```
summary(logmodelbiden)
```

```

##
## Call:
## glm(formula = vote_biden ~ state + education + foreign_born +
##      gender + Race + hispanic + agegrp, family = "binomial", data = survey_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.1499  -0.9979  -0.7634   1.2069   2.1124
##
## Coefficients:

```

	Estimate	Std. Error	z value
## (Intercept)	-1.20235	1.05350	-1.141
## stateAL	0.26615	0.85568	0.311
## stateAR	-0.34200	0.89377	-0.383
## stateAZ	0.39084	0.84101	0.465
## stateCA	0.73366	0.82818	0.886
## stateCO	0.42338	0.85167	0.497
## stateCT	1.13387	0.86098	1.317
## stateDC	1.45157	0.95037	1.527
## stateDE	1.08940	0.91266	1.194
## stateFL	0.48435	0.82962	0.584
## stateGA	0.26419	0.83970	0.315
## stateHI	0.70068	0.91806	0.763
## stateIA	0.79154	0.86966	0.910
## stateID	-0.25231	0.92455	-0.273
## stateIL	0.59303	0.83338	0.712
## stateIN	0.43431	0.84726	0.513
## stateKS	0.16803	0.88181	0.191
## stateKY	0.72366	0.85280	0.849
## stateLA	0.41042	0.85840	0.478
## stateMA	1.01920	0.84630	1.204
## stateMD	0.64254	0.85016	0.756
## stateME	1.13309	0.94166	1.203
## stateMI	0.79196	0.83839	0.945
## stateMN	0.91337	0.85965	1.062
## stateMO	0.50919	0.84597	0.602
## stateMS	0.08670	0.88064	0.098
## stateMT	0.58440	0.96299	0.607
## stateNC	0.60952	0.83644	0.729
## stateND	-0.91008	1.36614	-0.666
## stateNE	0.16868	0.93794	0.180
## stateNH	0.85173	0.94987	0.897
## stateNJ	0.56760	0.83707	0.678
## stateNM	0.79481	0.91047	0.873
## stateNV	0.30488	0.86086	0.354
## stateNY	0.60176	0.82968	0.725
## stateOH	0.52854	0.83342	0.634
## stateOK	0.04910	0.86893	0.057
## stateOR	0.73399	0.85178	0.862
## statePA	0.17416	0.83479	0.209
## stateRI	1.11407	1.01228	1.101
## stateSC	-0.19831	0.85512	-0.232
## stateSD	0.24156	0.98811	0.244
## stateTN	-0.07933	0.85035	-0.093
## stateTX	0.05142	0.83048	0.062
## stateUT	-0.13279	0.88643	-0.150
## stateVA	0.71095	0.83726	0.849
## stateVT	1.93935	1.01576	1.909
## stateWA	0.70700	0.84452	0.837
## stateWI	0.80126	0.84545	0.948
## stateWV	0.33895	0.89545	0.379
## stateWY	0.07837	1.39437	0.056
## educationAssociate Degree	0.06956	0.62620	0.111
## educationCollege Degree (such as B.A., B.S.)	0.14156	0.62240	0.227

## educationCompleted some college, but no degree	-0.10528	0.62248	-0.169
## educationCompleted some graduate, but no degree	0.02335	0.63449	0.037
## educationCompleted some high school	-0.45147	0.62579	-0.721
## educationDoctorate degree	-0.27545	0.64517	-0.427
## educationHigh school graduate	-0.51027	0.62336	-0.819
## educationMasters degree	0.16360	0.62596	0.261
## educationMiddle School - Grades 4 - 8	-0.21707	0.74826	-0.290
## educationOther post high school vocational training	-0.22316	0.63106	-0.354
## foreign_born	0.28725	0.11117	2.584
## genderMale	-0.33568	0.05454	-6.154
## RaceAsian (Chinese)	0.97131	0.33593	2.891
## RaceAsian (Japanese)	1.39773	0.56751	2.463
## RaceBlack, or African American	1.57608	0.25625	6.151
## Raceother asian or pacific islander	0.73126	0.28309	2.583
## RaceSome other race	0.50665	0.26488	1.913
## RaceWhite	0.22782	0.24590	0.926
## hispanic	0.28001	0.08652	3.236
## agegrp26-35	-0.11317	0.09444	-1.198
## agegrp36-45	-0.15847	0.09546	-1.660
## agegrp46-55	-0.36200	0.10318	-3.508
## agegrp56-65	-0.04056	0.10041	-0.404
## agegrp66-75	0.10482	0.10886	0.963
## agegrp76-85	-0.21776	0.18916	-1.151
## agegrp86-95	-0.02368	0.72116	-0.033
##	Pr(> z)		
## (Intercept)	0.253748		
## stateAL	0.755773		
## stateAR	0.701980		
## stateAZ	0.642129		
## stateCA	0.375684		
## stateCO	0.619106		
## stateCT	0.187854		
## stateDC	0.126669		
## stateDE	0.232611		
## stateFL	0.559337		
## stateGA	0.753050		
## stateHI	0.445336		
## stateIA	0.362732		
## stateID	0.784930		
## stateIL	0.476714		
## stateIN	0.608231		
## stateKS	0.848882		
## stateKY	0.396123		
## stateLA	0.632564		
## stateMA	0.228472		
## stateMD	0.449774		
## stateME	0.228866		
## stateMI	0.344851		
## stateMN	0.288014		
## stateMO	0.547244		
## stateMS	0.921572		
## stateMT	0.543946		
## stateNC	0.466175		
## stateND	0.505304		

```

## stateNE 0.857275
## stateNH 0.369887
## stateNJ 0.497725
## stateNM 0.382684
## stateNV 0.723225
## stateNY 0.468273
## stateOH 0.525960
## stateOK 0.954940
## stateOR 0.388850
## statePA 0.834741
## stateRI 0.271090
## stateSC 0.816611
## stateSD 0.806873
## stateTN 0.925674
## stateTX 0.950632
## stateUT 0.880923
## stateVA 0.395800
## stateVT 0.056227 .
## stateWA 0.402497
## stateWI 0.343266
## stateWV 0.705040
## stateWY 0.955181
## educationAssociate Degree 0.911545
## educationCollege Degree (such as B.A., B.S.) 0.820087
## educationCompleted some college, but no degree 0.865687
## educationCompleted some graduate, but no degree 0.970642
## educationCompleted some high school 0.470634
## educationDoctorate degree 0.669418
## educationHigh school graduate 0.413019
## educationMasters degree 0.793821
## educationMiddle School - Grades 4 - 8 0.771736
## educationOther post high school vocational training 0.723619
## foreign_born 0.009771 **
## genderMale 7.53e-10 ***
## RaceAsian (Chinese) 0.003835 **
## RaceAsian (Japanese) 0.013782 *
## RaceBlack, or African American 7.72e-10 ***
## Raceother asian or pacific islander 0.009791 **
## RaceSome other race 0.055775 .
## RaceWhite 0.354198
## hispanic 0.001211 **
## agegrp26-35 0.230807
## agegrp36-45 0.096917 .
## agegrp46-55 0.000451 ***
## agegrp56-65 0.686279
## agegrp66-75 0.335575
## agegrp76-85 0.249637
## agegrp86-95 0.973800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 8809.5 on 6474 degrees of freedom

```

```
## Residual deviance: 8255.1  on 6398  degrees of freedom
##    (4 observations deleted due to missingness)
## AIC: 8409.1
##
## Number of Fisher Scoring iterations: 4
```

References

- Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [<https://www.voterstudygroup.org/downloads?key=46a716b2-7321-4fcf-9ee6-8987a584a253>].
- Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. <https://doi.org/10.18128/D010.V10.0>
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.