Biden to Win Popular Vote in Landslide

A Forecast of the 2020 US Election

Isaac Ehrlich

02 November 2020

Abstract

On the heels of a widely mispredicted 2016 US presidential election, politicians, statisticians, and voters alike are looking for more accurate forecasts as we approach what is perhaps the most consequential election in recent US history. In an attempt to avoid errors of unrepresentative samples seen in past research, we analyze the Nationscape Data Set, a survey on political opinions, and use the American Community Survey, a key source of information on the American populus, to post-stratify this data, achieving a representative analysis of current political opinion and behavior. After applying a multilevel regression to our data, we predict that Biden will win the popular vote of the upcoming election in a landslide, capturing 83% of the vote.

Keywords: Forecasting, US 2020 Election, Trump, Biden, Multilevel Regression with Post-Stratification

Introduction

From the errors of Literary Digest in 1936 to the mispredicted outcome of the 2016 US presidential election, surevyors and forecasters have often struggled to accurately poll, extrapolate, and predict election winners. Often, as was the case in 1936, this is a result of non-representative samples, a common blight on statisticians' attempts for unbiased surveying and prediction, in and outside of politics. However, post-stratification, a method where responses in smaller surveys are weighed based on population statistics, is a promising method of reducing sampling biases. By weighting responses in accordance to their likely effect among the entire population, analyses that would previously be misleading can be completed on non-representative data sets.

As elections are a common attracter of non-representative sampling, as well as a common cause for statistical modelling, the upcoming US Presidential election is a promising subject for this technique. In this analysis, we apply a multilevel logistic regression with age, race, and gender as predictors, to the Nationscape Data Set, a survey conducted in order to gauge political opinions across the US (Tausanovitch and Vavreck 2020). However, in order to ensure the reliability of this data, we use the American Community Survey (ACS), an annual large-scale survey, hosted on the IPUMS database, meant to provide regular updates to housing data collected from the US Census every ten years (Ruggles et al. 2020).

In this paper, we discuss in detail, the data used in our analyses, as well as the multilevel logistic regression model applied to the data set, our results, which indicate that Biden will win 83% of the popular vote, and finally a discussion on the efficacy of this method and success of our model. Furthermore, detailed code and analysis of this model can be found at https://github.com/imehrlich/STA304

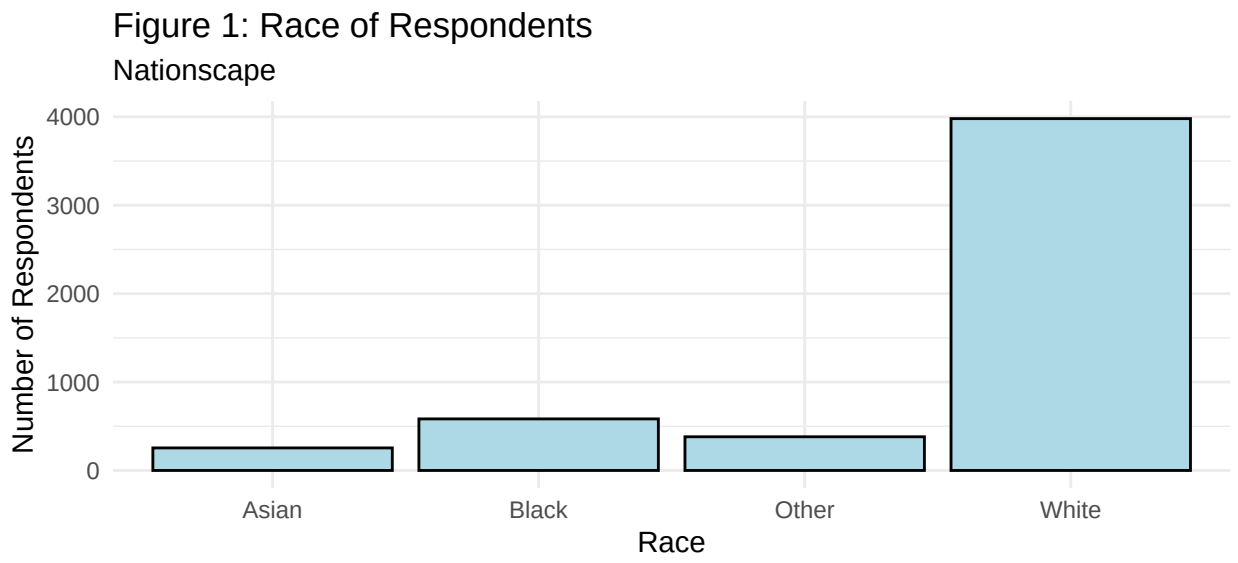
Data

There are two data sets used in this forecast of the 2020 US Presidential Election: survey data collected by the Democracy Fund and UCLA Nationscape on the topic of voting habits and political inclinations, and a data set on the general population of the United States, adapted from the American Community Surveys (ACS) conducted by IPUMS. The following sections describe these data sets and how they were used.

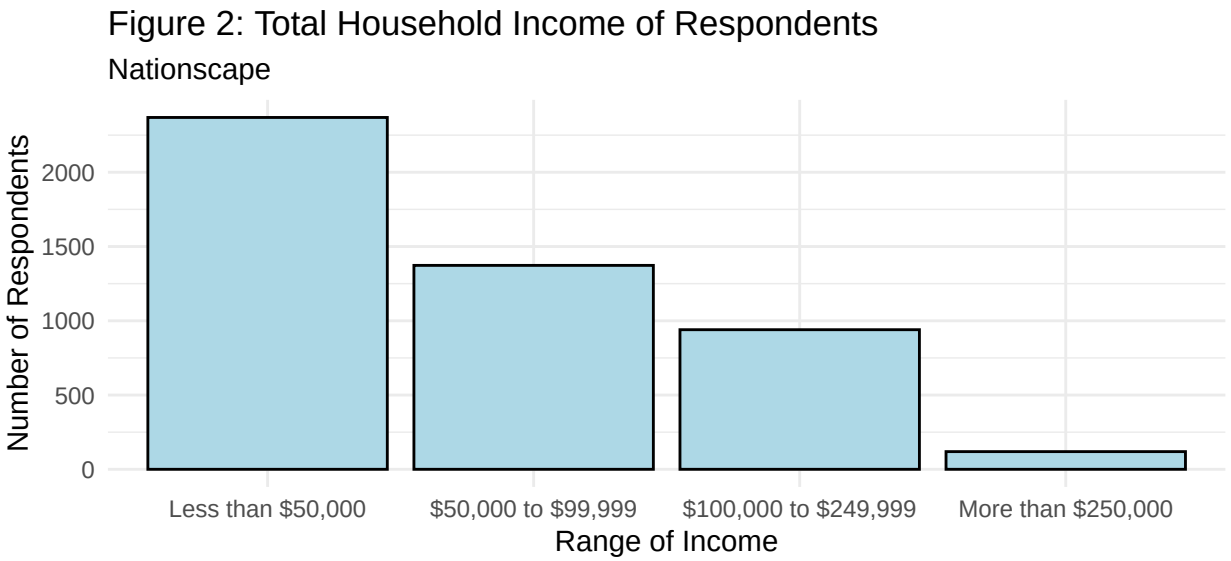
Nationscape Data by the Democracy Fund and UCLA Nationscape

The Nationscape Data Set is a product of the collaboration between the Democracy Fund and Political Scientists at UCLA, created with the goal of surveying political opinion from a wide range of locations and demographics across the United States (Tausanovitch and Vavreck 2020). As it is concerned with potential voters, the population of this dataset is all eligible US voters, and the survey obtains its samples from Lucid, a survey research platform (Tausanovitch, Vavreck, and Democracy Fund 2020). Targetting over 500,000 responses, which together are a representative sample of the United Stated population eligible to vote in 2020, the survey interviews roughly 6,250 participants each week(Tausanovitch, Vavreck, and Democracy Fund 2020). While participants are allowed to omit answers to certain questions, such as household income, the data is treated before release, and participants who were deemed to 'speed through' the survey were excluded prior to data release, minimizing the data cleaning required in this analysis (Tausanovitch, Vavreck, and Democracy Fund 2020). In this analysis, we use responses from the most recent release of this survey, occuring on June 25, 2020. This release contained just under 6,500 responses.

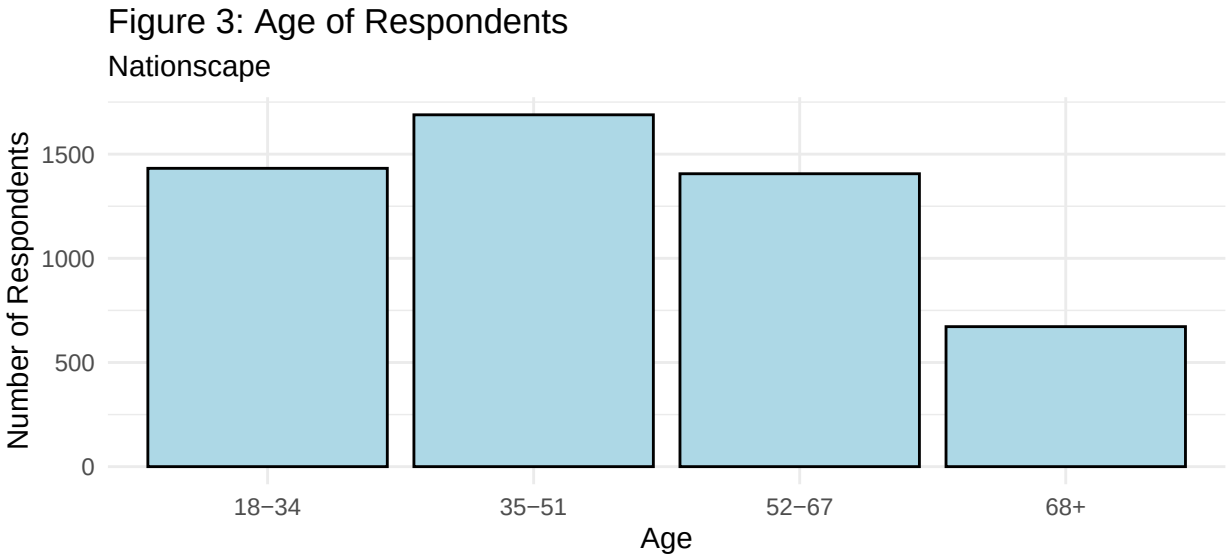
Although the data is treated prior to release, for the purposes of our analysis, we used the statistical programming language R (R Core Team 2020), to create several new variables, either to break continuous variables into classes for easier grouping, or in order to match this data set to the data recorded in the ACS. There were three main variables in this data set to which adjustments were made. First, in order to match the ACS, specific qualifiers were removed from the race variable (e.g. Asian (Chinese) was modified to Asian), leaving us with four factors for race: 'White,' 'Black,' 'Asian,' and 'Other.' The distribution of the remaining factors is shown in Figure 1.



Second, income classes were combined to form larger ranges in order to cater towards a higher number of samples in each grouping. While the original data contained 25 different categories for household income, we reduced this to four classes: 'Less than $\$50,000'$,'$\$50,000$ to $\$100,000$','$\$100,000$ to $\$250,000$,'and 'More than $\$250,000$' The distribution of these modified income classes is shown in Figure 2.



Finally, the age of participants was also modified to be represented by generation. The exact splits in the ages were based off of generational guidelines released by Gallup, which have also been used in previous research on voting models, such as Auerbach, Ghitza, and Gelman (2020). These age groups are '18 to 34', '35 to 52', '52 to 68', and 'over 68'. The counts of these resultant groupings can be seen in Figures 1-3.



American Community Surveys by IPUMS

The American Community Survey is a data set on American demographics, population, and household variables compilled and published annually on IPUMS, a database for American censuses (Ruggles et al. 2020). Meant to provide annual updates on the decennial census, the ACS considers all housing units (and the people living inside them) as their population, and use the Master Address File (MAF), the Census Bureau's housing databse, as their frame for contacting respondents (United States Census Bureau 2020). Their annual sample size contains 3,000,000 participants, with over 2,000,000 final interviews conducted each year (United States Census Bureau 2020). Despite the high non-response, the size, frequency, and range of information recored in the ACS make it a useful contribution to US Census data. As we attempt to link this data set to a political survey in our analysis, we focus on information useful for predicting voting habits, such as gender, race, and state of residence.

As with the Nationscape data, key variables in the data set were modified in order to create easily separable groupings. The same variables and classes were constructed in the ACS data as with the Nationscape data. The resultant distributions are seen in Figures 4-6.

Figure 4: Race of Respondents

ACS

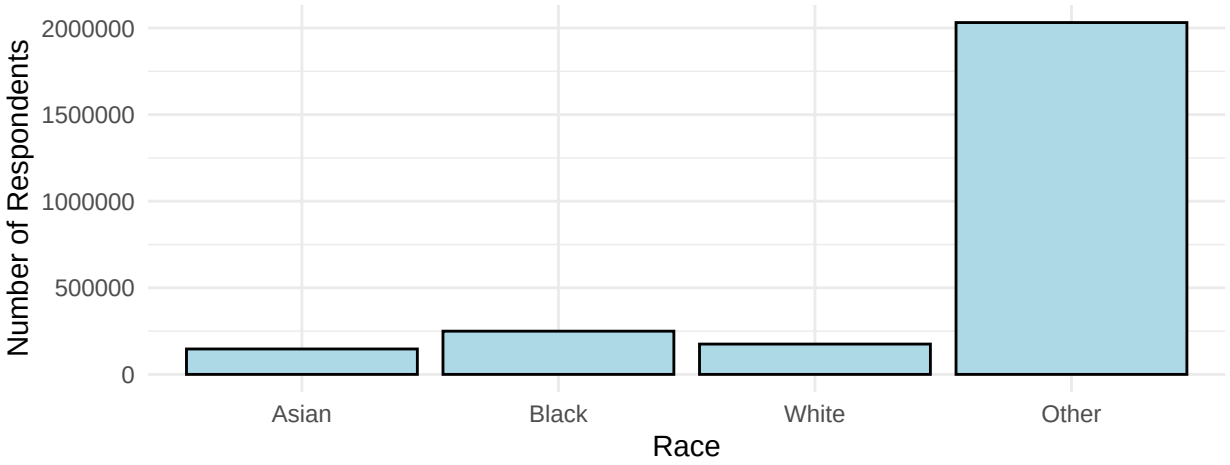
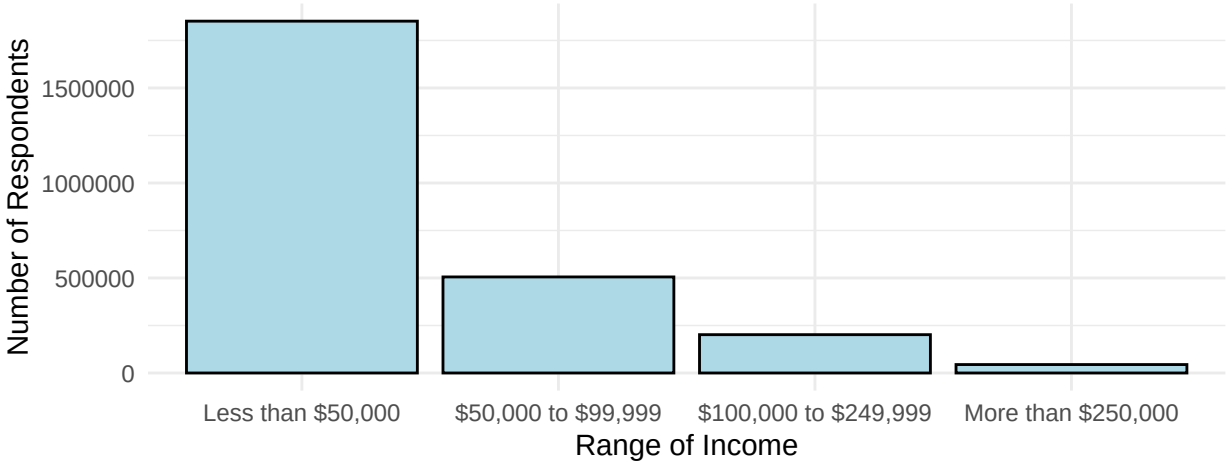
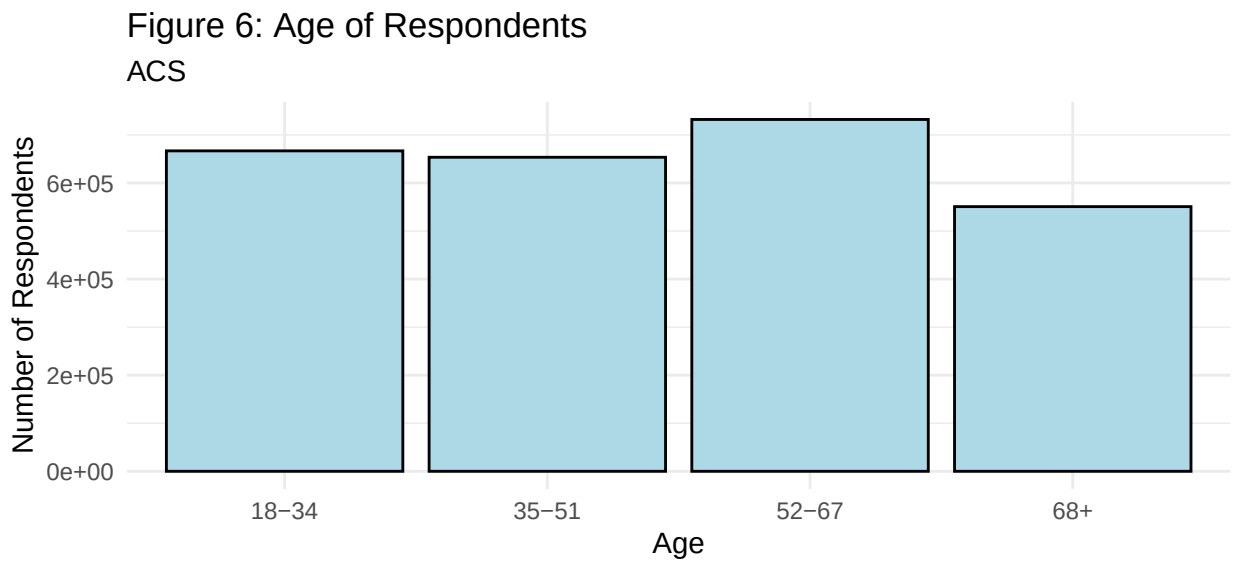


Figure 5: Total Household Income of Respondents

ACS





Model

In US elections, there are only two candidates that realistically vie for the presidency in each election cycle. Therefore, we apply a logistic regression model, since logistic regression models are capable of using multiple predictors to determine a binomial distribution; in this case a Donald Trump or a Joe Biden victory. Therefore, we propose the following logistic regression model:

$$\hat{y}_{victory} = \text{logit}(\beta_0 + \beta_{gender} x_{gender} + \beta_{race} x_{race} + \beta_{age} x_{age})$$

where $\hat{y}_{victory}$ is a binomial variable where 1 denotes a Joe Biden victory and 0 denotes a Donald Trump victory, x_{gender} represents the gender of respondent, x_{race} represents the race of the respondent, and x_{age} represents the age group of the respondent. In other words, we are using respondents' gender, race, and age to predict the outcome of the election. It is important here to note that many other additional predictors, such as state, income, party affiliation, and past voting record may be useful and effective predictors as well. However, as we are post-stratifying this data, in an effort to minimize risk of creating over-specified groups with minimal counts, we have attempted to keep our model as simple as possible while maintaining accurate prediction results.

Post-Stratification

A key part of this model and analysis is the post-stratification conducted between our two data sets. Post-stratification is the process of weighing samples based on their proportion relative to the population. This is a key method to use when there is valid concern for non-response or other sampling bias, as this weighting adjusts for over or under sampling of groups. In this specific instance, we assume that the ACS data expresses a more accurate distribution to the true population, and we therefore use this data to adjust weights in the Nationscape data on the basis of gender, race, age, and total income.

Results

Table 1 displayes the results of the logistic regression model prior to post-stratification.

After post-stratification, we can see further detailed results on how different respondents are likely to vote. In the interest of brevity, the full table is shown in the appendix, however, in order to highlight key findings, Table 2 displays examples of differences observed among respondets of different race.

Table 1: Logistic Regression Model of Nationscape Survey

Predictor	Coefficient	SE	t	p
Intercept	1.26	0.145	8.68	<0.001
Gender: Male	-0.41	0.059	-6.87	<0.001
Race: Black	1.18	0.188	6.26	<0.001
Race: Other	-0.33	0.174	-1.90	0.057
Race: White	-0.92	0.142	-6.48	<0.001
Age: 35-51	-0.49	0.078	-6.34	<0.001
Age: 52-68	-0.40	0.081	-4.91	<0.001
Age: 68+	-0.35	0.100	-3.47	<0.001

Table 2: Comparing Logistic Regression Estimates by Race

Gender	Age	Race	Estimate
Female	18-34	White	0.3448771
Female	18-34	Black	2.4422156
Female	18-34	Asian	1.2627881
Female	18-34	Other	0.9316479
Male	52-67	White	-0.4596460
Male	52-67	Black	1.6376925
Male	52-67	Asian	0.4582650
Male	52-67	Other	0.1271247

Similarly, Table 3 displays examples of estimates when holding all variables other than 'Age' constant.

Table 3: Comparing Logistic Regression Estimates by Age

Gender	Age	Race	Estimate
Female	18-34	White	0.3448771
Female	35-51	White	-0.1479366
Female	52-67	White	-0.0532440
Female	68+	White	-0.0008312
Male	18-34	Black	2.0358136
Male	35-51	Black	1.5429998
Male	52-67	Black	1.6376925
Male	68+	Black	1.6901053

Finally, Table 4 shows exampels of estimates when holding all variables other than 'Gender' constant.

Table 4: Comparing Logistic Regression Estimates by Gender

Gender	Age	Race	Estimate
Female	68+	Asian	0.9170798
Male	68+	Asian	0.5106778
Female	18-34	Other	0.9316479
Male	18-34	Other	0.5252458
Female	18-34	Black	2.4422156
Male	18-34	Black	2.0358136
Female	35-51	White	-0.1479366
Male	35-51	White	-0.5543387

After predicting the results using the post-stratified data, we see that Biden is expected to win 83% of the popular vote.

Discussion

Before getting into the results and output of the regression model, it may be worthwhile, if for no other reason than to check the sampling bias of the Nationscape data, to compare the proportion of responses from the Nationscape data to the ACS data. Tables 5-7 show how demographic factors compare across data sets.

Table 5: Respondent's Race Across Data Sets

Race	Nationscape	ACS
Asian	0.049	0.056
Black	0.112	0.096
Other	0.073	0.067
White	0.766	0.780

Table 6: Respondent's Race Across Data Sets

Gender	Nationscape	ACS
Female	0.497	0.516
Male	0.503	0.484

Table 7: Respondent's Race Across Data Sets

Age Group	Nationscape	ACS
18-34	0.275	0.256
35-51	0.325	0.251
52-67	0.270	0.281
68+	0.129	0.212

Tables 5-7 confirm the need for post-stratification when conducting such analyses. While the proportions of distribution of race is similar across both data sets, we see a disparity between the samples of the two data sets increase for the other two factors. Broadly, this affirms the position that post-stratification is a useful tool in improving accuracy and efficacy of non-representative samples.

As for the model, estimates seem to confirm research on voter behaviour according to demographic (e.g Black voters are less likely to vote for Trump). Tables 2-4 provide clear evidence that [(1)]minority voters are more

likely to vote for Biden than white voters, [2]young voters, across gender and race are more likely to vote for Biden than older voters, and [3]female voters, across age and race are more likely to vote for Biden than male voters. These results support the general trend in voter behaviour seen within the US as well as across the world.

One of the troubles, perhaps, with the model is the margin of victory with which Biden is predicted to win. While the optomists among us may not want to quarrel with this outcome, recent polling numbers, as well as precedent set in the last two centuries of elections, suggests that this outcome is unlikely. A contribution to a possible explanation for this can be made by reffering to Tables 2-4 once more. Across these tables, it can be observed that while the magnitude of negative estimates, estimates that favor Trump are relatively low, the positive estimates for many groups of voters are quite high. This indicates that the model is not confident that the entire group that it has denoted as favourable to Trump will indeed vote this way, as opposed to the strong prediction in favor of Biden.

Weaknesses

The margin of victory this model has predicted for Joe Biden is certainly a result of several shortcomings in this model. First, while post-stratification is a useful tool to avoid sampling bias and decrease variance, we were afraid to overcomplicate groupings of the sample, and thus may have oversimplified the model. Furthermore, the absence of numerical variables may have made it difficult for the model to express nuances in voter behaviour.

Additionally, while this model estimates the winner of the popular vote, this is not necessarily a good estimate for the winner of the overall election. The US presidential election is performed through an electoral college, where points are attributed to candidates based on their performance in each state. As recently as the 2016 election, the winner of the popular vote was not the winner of the presidential election, and therefore a stronger forecast of the election would focus on races in individual states as opposed to the nation as a whole.

Appendix

Table 8: Full Table of Estimates by Demographic Groups

Gender	Age	Race	Estimate
Female	18-34	White	0.3448771
Female	18-34	Black	2.4422156
Male	52-67	White	-0.4596460
Male	18-34	White	-0.0615250
Female	52-67	White	-0.0532440
Male	35-51	White	-0.5543387
Male	52-67	Black	1.6376925
Female	68+	Black	2.0965073
Male	52-67	Other	0.1271247
Female	68+	White	-0.0008312
Male	35-51	Black	1.5429998
Female	52-67	Black	2.0440945
Female	35-51	White	-0.1479366
Female	35-51	Black	1.9494019
Male	18-34	Black	2.0358136
Male	68+	White	-0.4072332
Male	68+	Black	1.6901053
Male	18-34	Other	0.5252458
Female	35-51	Other	0.4388341
Male	18-34	Asian	0.8563861
Female	18-34	Asian	1.2627881
Male	35-51	Other	0.0324321
Male	35-51	Asian	0.3635723
Female	18-34	Other	0.9316479
Female	68+	Other	0.5859396
Female	52-67	Asian	0.8646670
Female	35-51	Asian	0.7699744
Female	68+	Asian	0.9170798
Male	52-67	Asian	0.4582650
Female	52-67	Other	0.5335268
Male	68+	Other	0.1795375
Male	68+	Asian	0.5106778

References

Auerbach, Jonathan, Yair Ghitza, and Andrew Gelman. 2020. "A Generational Voting Model for Forecasting the 2020 American Presidential Election." http://www.stat.columbia.edu/~gelman/research/unpublished/2020prediction.pdf.

Auguie, Baptiste. 2017. *GridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. *IPUMS USA: Version 10.0 ACS 1-Year Data.* VMinneapolis, MN: IPUMS. doi:10.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. *Nationscape Data Set.* Democracy Fund and UCLA

Nationscape. voterstudygroup.org/publication/nationscape-data-set.

Tausanovitch, Chris, Lynn Vavreck, and Democracy Fund. 2020. "Democracy Fund + UCLA Nationscape Methodology and Representativeness Assessment." https://www.voterstudygroup.org/uploads/reports/Data/NS-Methodology-Representativeness-Assessment.pdf.

United States Census Bureau. 2020. *American Community Survey: Design and Methodology Report.* United States Census Bureau. https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. doi:10.21105/joss.01686.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.