Predicting Joe Biden Win of US Popular Vote for 2020 Election

Maxwell Garrett

November 2nd 2020

Model

We are interested in predicting the popular vote and electoral vote outcome for the 2020 US election. We will model this result below, using two seperate models: one model with votes for Biden as the observed outcome (coding the observed value as 1 if the vote was for Biden, and 0 otherwise), and the other model with votes for Trump as the observed outcome. We will use post-stratification to ensure the model is representative of the US population. We will discuss 2 models throughout, model 1 which has the response variable vote_trump and model 2 which has the response variable vote_biden.

Model Specifics

We will be using a logistic regression model to model the log-odds of voting for Donald Trump. We will also use an similar model with the same predictors to model the log-odds of voting for Joe Biden. We will use the following variables as predictors in the models: age (a numeric variable), gender (a categorical variable), white (a categorical variable), and state (a categorical variable). The logistic regression model equation is below:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{gend=M} + \beta_3 x_{st=AL} + ... + \beta_{53} x_{st=WY} + \beta_{54} x_{white=white} + \epsilon$$

Where p represents the proportion of voters who will vote for Donald Trump in model 1, and Joe Biden in model 2. β_0 is the intercept of the model, and is the log-odds of voting Trump/Biden (dependent on model) at age of 0, gender female, non-white race individual and lives in state Alaska. β_1 represents the change in log-odds per unit of age. β_2 represents the change in log-odds if an individual is male. β_3 to β_{53} represent the change in log-odds if an individual is from a given state compared to the reference state of Alaska. β_{54} represents the change in log-odds if an individual's race is white, compared to all other races (classified as non-white).

Post-Stratification

We will use our above models to estimate the proportion of voters who will vote for Donald Trump, and then Joe Biden. We will create a table, with each cell showing the estimated proportion of interest (i.e. either proportion of voters for Trump or proportion of voters for Biden) for each combination of age, gender, state, and race (white vs non-white). These proportions are estimated using the models defined above, and plugging in specific values for the predictors age, gender, state, and white. We will then use the estimated probability in each cell, and weight it by the proportion of the population that cell represents using census data. Finally, we will sum all weighted values and divide by the total population size (represented by all cells combined).

Secondly, We will use the models to estimate the percentage of electoral votes each candidate will receive. We will do this by estimating the proportion of each cell that will vote for Biden or Trump. Then, we group all cells by their states and sum all estimated proportions for each state. Then, we will divide each proportion by the sum of population in each state. If the percent of population in a state is greater than 50%, We will add that state's number of electoral votes to a sum. We will divide the sum of electoral votes won by the total electoral votes in the US. The final result is the percent of electoral votes expected to be won by a candidate.

Results

Table 1: Popular vote predictions for Biden and Trump

candidate	popular_vote	lower_popular_vote	upper_popular_vote
Donald Trump	0.4031327	0.3239753	0.4822901
Joe Biden	0.4146553	0.3353002	0.4940104

In table 1, we see that model 1, modeling vote_trump, predicts Trump receiving 40.31% of the popular vote. We also see that model 2, modeling vote_biden, predicts Biden receiving 41.47% of the popular vote. We as well have that Trump's popular vote is between 32.40% and 48.23% with 95% confidence. Biden's popular vote is between 33.53% and 49.40% with 95% confidence

Table 2: Electoral vote predictions for Biden and Trump

candidate	lower_electoral_vote	mid_electoral_vote	upper_electoral_vote
Donald Trump	0.0000000	0.0706320	0.4795539
Joe Biden	0.0055762	0.0650558	0.5762082

In table 2, we see that model 1, modeling vote_trump, predicts Trump receiving 7.06% of the electoral vote. The upper and lower bounds of a 95% confidence interval around this estimate are also included. The upper bound is 48.00% and the lower bound is 0.00%.

We also see that model 2, modeling vote_biden, predicts Biden receiving 6.51% of the electoral vote. The upper and lower bounds of a 95% confidence interval around this estimate are also included. The upper bound is 57.62% and the lower bound is 0.557%.

Discussion

Above, we created two logistic models, one representing the log-odds of Biden winning and the other Trump winning, and used these models to predict the proportions of subsets of the population voting for either candidate. We added these subsets of the population up and divided by the total population to estimate the popular vote for the United States, as a whole. We also estimate the confidence intervals of the percent of the electoral vote won by predicting the winner of the popular vote for each state using our models. We then add up the number of electoral votes won for a given state if the popular vote for that state indicates a given party receiving more than 50% of the vote.

We estimate Donald Trump receiving 40.31% of the popular vote and Joe Biden receiving 41.47% (table 1). We also find a 95% confidence interval of the percent of electoral vote won by each candidate. We estimate

Donald Trump receiving between 0% to 48% of the electoral vote with 95% confidence. We also estimate Joe Biden receiving between 0.6% to 57.62% of the electoral vote with 95% confidence. Based on the above results, we predict that Joe Biden will win the election because he is predicted to receive more of the popular vote, and the confidence interval of his electoral vote percent prediction includes a higher upper bound when compared with the confidence interval of Trump's electoral vote percent prediction (i.e. Trump's upper bund is below 50% whereas Biden's is above 50%).

Weaknesses

One weakness of this study is that there is high variability in the estimates produced. This can be seen in the 95% confidence intervals on the estimates displayed in table 1 and 2. We see that the confidence intervals overlap for each candidate. A solution to this could be to collect more survey data on the voting intentions of Americans to decrease the potential error associated with the data. Another weakness in the study is the modification of the race variable from a categorical variable with multiple values to a binary variable. This was done due to differences in how the census data reported race and how the survey data reported race. In the census data, respondents were allowed to select that they belonged to multiple races and other categories that were not present in the survey data. To combine the datasets together, race was simplified, resulting in lost information. To solve this issue, more work could be done in identifying which race categories correspond and collecting more complete race data related to voters.

Next Steps

There are multiple possible avenues for future work on this analysis. We would likely benefit from a greater sample size in the survey data on voting intention. The greater sample size would help reduce error in the estimates. It would also be helpful in future to connect more variables between census data and survey data as this would help produce more subsets of the population for post-stratification. The additional predictor variables would also help remove some variation in the proportion estimates. It might also be helpful to see how closely the survey data from pre-election resembles the survey data collected from election day polls. It would help show us how accurate or inaccurate our collected voter data is.

References

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Rohan Alexander, and Sam Caetano (2020). Survey and Strat Cleaning Code. Retrieved October 31, 2020.

Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from [URL].

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

United States Electoral College Votes by State. (2016). Retrieved November 01, 2020, from https://www.britannica.com/topic/United-States-Electoral-College-Votes-by-State-1787124

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Appendix

Model 1: full model output for Donald Trump log-odds.

term	estimate	std.error	statistic	p.value
(Intercept)	-1.4974679	0.7453101	-2.0091877	0.0445172
age	0.0126468	0.0016730	7.5594692	0.0000000
genderMale	0.4669004	0.0545990	8.5514432	0.0000000
stateAL	-0.4477632	0.7745496	-0.5780949	0.5632000
stateAR	-0.1429650	0.7983478	-0.1790761	0.8578779
stateAZ	-0.6623158	0.7602820	-0.8711449	0.3836751
stateCA	-1.0000664	0.7465151	-1.3396466	0.1803603
stateCO	-0.7350104	0.7715451	-0.9526473	0.3407688
stateCT	-1.6933434	0.7973557	-2.1236988	0.0336953
stateDC	-0.9181082	0.8818398	-1.0411281	0.2978161
stateDE	-1.1315010	0.8516103	-1.3286606	0.1839600
stateFL	-0.6850736	0.7476946	-0.9162480	0.3595369
stateGA	-0.4501154	0.7584370	-0.5934777	0.5528615
stateHI	-0.4110775	0.8514070	-0.4828213	0.6292226
stateIA	-0.9007704	0.7924313	-1.1367174	0.2556564
stateID	-0.3170792	0.8308620	-0.3816268	0.7027382
stateIL	-0.9243482	0.7524362	-1.2284738	0.2192691
stateIN	-0.7425592	0.7660848	-0.9692912	0.3323999
stateKS	-0.4763207	0.7975293	-0.5972454	0.5503435
stateKY	-0.5041097	0.7724290	-0.6526291	0.5139954
stateLA	-0.4651305	0.7778806	-0.5979458	0.5498761
stateMA	-1.5116723	0.7746087	-1.9515303	0.0509940
stateMD	-0.8418505	0.7742213	-1.0873512	0.2768816
stateME	-1.0335225	0.8730132	-1.1838567	0.2364698
stateMI	-0.9606996	0.7587900	-1.2660942	0.2054794
stateMN	-0.6247763	0.7814247	-0.7995349	0.4239803
stateMO	-0.7682870	0.7657837	-1.0032690	0.3157311
stateMS	-0.4543849	0.8067694	-0.5632153	0.5732883
stateMT	-0.6612404	0.8914581	-0.7417515	0.4582379
stateNC	-0.6567509	0.7558254	-0.8689187	0.3848916
stateND	-0.4716296	1.0933436	-0.4313644	0.6662034
stateNE	-0.8263027	0.8617008	-0.9589206	0.3375987
stateNH	-0.9467072	0.8815803	-1.0738752	0.2828786
stateNJ	-0.8246065	0.7569086	-1.0894400	0.2759599
stateNM	-1.5986176	0.8812294	-1.8140766	0.0696659
stateNV	-0.5855416	0.7825360	-0.7482616	0.4543024
stateNY	-0.8375806	0.7478735	-1.1199496	0.2627352
stateOH	-0.8546353	0.7520691	-1.1363786	0.2557981
stateOK	-0.3370533	0.7865374	-0.4285280	0.6682668
stateOR	-1.0107334	0.7748905	-1.3043563	0.1921121
statePA	-0.6601841	0.7525192	-0.8772988	0.3803244
stateRI	-1.3469243	1.0149594	-1.3270720	0.1844849
stateSC	-0.3439754	0.7691544	-0.4472125	0.6547217
stateSD	-0.5484344	0.9009986	-0.6086962	0.5427258

term	estimate	std.error	statistic	p.value
stateTN	-0.3870146	0.7656266	-0.5054875	0.6132165
stateTX	-0.4694338	0.7478782	-0.6276876	0.5302086
stateUT	-0.7940384	0.7978024	-0.9952820	0.3195991
stateVA	-0.8835207	0.7573484	-1.1665974	0.2433730
stateVT	-2.3036605	1.0654527	-2.1621424	0.0306072
stateWA	-0.9791454	0.7662667	-1.2778128	0.2013154
stateWI	-1.1025217	0.7679526	-1.4356639	0.1510980
stateWV	-0.5061873	0.8095176	-0.6252950	0.5317775
stateWY	-2.0498843	1.3445153	-1.5246270	0.1273522
whitewhite	1.2355996	0.0749435	16.4870895	0.0000000

Model 2: full model output for Joe Biden log-odds.

term	estimate	std.error	statistic	p.value
(Intercept)	-0.6048890	0.8312922	-0.7276490	0.4668285
age	0.0022633	0.0016142	1.4020487	0.1609007
genderMale	-0.3070043	0.0526264	-5.8336519	0.0000000
stateAL	0.6669004	0.8581395	0.7771468	0.4370721
stateAR	0.0248468	0.8940977	0.0277898	0.9778298
stateAZ	0.7807884	0.8450248	0.9239828	0.3554952
stateCA	1.0952367	0.8323333	1.3158631	0.1882200
stateCO	0.8101011	0.8554015	0.9470419	0.3436174
stateCT	1.5178446	0.8638821	1.7570044	0.0789171
stateDC	1.9701343	0.9511533	2.0713110	0.0383297
stateDE	1.3952087	0.9139070	1.5266418	0.1268501
stateFL	0.9114585	0.8338111	1.0931234	0.2743396
stateGA	0.7719117	0.8425232	0.9161904	0.3595670
stateHI	1.0445332	0.9098933	1.1479733	0.2509796
stateIA	1.1286168	0.8727721	1.2931403	0.1959625
stateID	0.1140835	0.9264526	0.1231402	0.9019961
stateIL	1.0392261	0.8373785	1.2410471	0.2145884
stateIN	0.7810170	0.8509593	0.9178077	0.3587195
stateKS	0.5243261	0.8838292	0.5932438	0.5530180
stateKY	1.0034396	0.8564558	1.1716187	0.2413502
stateLA	0.8478738	0.8601526	0.9857248	0.3242682
stateMA	1.3575876	0.8502909	1.5966155	0.1103514
stateMD	1.0289663	0.8534002	1.2057255	0.2279234
stateME	1.4536332	0.9426864	1.5420115	0.1230708
stateMI	1.1743199	0.8425167	1.3938239	0.1633708
stateMN	1.3418686	0.8634002	1.5541675	0.1201445
stateMO	0.8323505	0.8495549	0.9797489	0.3272101
stateMS	0.6324538	0.8807185	0.7181113	0.4726887
stateMT	0.9991035	0.9657734	1.0345114	0.3008972
stateNC	0.9981655	0.8403543	1.1877913	0.2349157
stateND	-0.7241380	1.3723939	-0.5276459	0.5977451
stateNE	0.6536239	0.9389763	0.6961026	0.4863646
stateNH	1.0762268	0.9521969	1.1302566	0.2583681
stateNJ	0.9677055	0.8411483	1.1504577	0.2499554
stateNM	1.1776197	0.9141966	1.2881471	0.1976948
stateNV	0.7830955	0.8632147	0.9071852	0.3643089

term	estimate	std.error	statistic	p.value
stateNY	1.0092087	0.8338497	1.2103005	0.2261636
stateOH	0.8853361	0.8375768	1.0570207	0.2905021
stateOK	0.1948442	0.8719293	0.2234634	0.8231749
stateOR	1.0628216	0.8559890	1.2416299	0.2143731
statePA	0.5475874	0.8390599	0.6526201	0.5140012
stateRI	1.2315258	1.0170573	1.2108715	0.2259446
stateSC	0.2416224	0.8577135	0.2817053	0.7781695
stateSD	0.6196992	0.9904595	0.6256684	0.5315325
stateTN	0.3412113	0.8532973	0.3998739	0.6892494
stateTX	0.4956334	0.8344693	0.5939504	0.5525453
stateUT	0.1215480	0.8900489	0.1365633	0.8913760
stateVA	1.1616324	0.8410175	1.3812225	0.1672106
stateVT	2.3110780	1.0155663	2.2756545	0.0228667
stateWA	1.0395565	0.8486950	1.2248883	0.2206173
stateWI	1.1737662	0.8493165	1.3820127	0.1669678
stateWV	0.6047325	0.8989706	0.6726944	0.5011417
stateWY	0.2168965	1.3925647	0.1557532	0.8762275
whitewhite	-0.7721365	0.0625635	-12.3416405	0.0000000