Prediction on U.S. 2020 Presidential Election

Boyuan Cui, Zhaocheng Li

November 2, 2020

Abstract

Key Words: U.S. 2020 Election, Popular Vote, Post-stratification, Logistic Regression

This report is an analysis to predict the U.S. 2020 federal election. The analysis employs the post-stratification technique by dividing the data into 838,113 cells according to different age, gender, state, race, education, and income. The final result of popular vote proportion prediction is 39.27% for Biden and 38.57% for Trump, and the electoral vote is 301 for Biden and 237 for Trump.

Introduction

The U.S. presidential election has always been a popular topic of the world for every four year, and as the election day comes closer, we are all looking forward to see who will become the next U.S. president. This report is interested in predicting the U.S. 2020 election popular vote for each candidate (Joe Biden and Donald Trump). We will use the post-stratification technique in our analysis.

Data

We extract the survey data from Democracy Fund + UCLA Nationscape [**1**] and the census data from IPUMS USA [**2**]. We remove all the examples with N/A data from these two datasets.

The survey data is collected on June 25, 2020 [**1**], and the census data is collected in 2018 [**2**].

Model

Model Specifics

We choose the logistic regression model to estimate the outcome of the presidential election. Because we are trying to find out whether the voter will vote for Trump or Biden in the 2020 election, the outcome is binary, so the logistic regression model is more appropriate and suitable in this case.

We are using six predictor variables in our model, which are age, race, gender, income level and education level. These are the representative information in both census and survey dataset. To express the model mathematically, we have:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age} + \beta_2 x_{male} + \beta_3 x_{state} + \beta_4 x_{race} + \beta_5 x_{education} + \beta_6 x_{income} + \epsilon$$

where p represents the proportion of voters who will vote for Trump/Biden, depending on which model we are using. Value of ϵ represents the residual of our fit.

Parameter β_0 represents the intercept of the model, which, in this case, indicates the value of logit function $\log(\frac{p}{1-p})$ for a 0 year old Alabama female with 3rd grade or less education level and less than $15000 income.

Parameter β_1 represents the change in logit value by increasing age by 1.

Parameter β_2 represents the change in logit if the example is male. $x_{male} = 1$ if subject is male, 0 if subject is female. No other possible values

Parameter β_3 represents the impact of state on the logit value. We use Alabama as a reference, so all change in logit is comparing with Alabama

Parameter β_4 represents the impact of race on the logit value. We use white as the reference.

Parameter β_5 represents the change in logit value by increasing education level by 1. Reference table of education level can be found at Appendix Table 3.

Parameter β_6 represents the change in logit value by increasing income level by 1. Reference table of income level can be found at Appendix Table 4.

Finally, we will fit two models, one for Trump and one for Biden, and we will compare the result to predict the winner of 2020 US presidential election.

Detail of the fitted model can be found at Appendix Table 5 and 6

Post-Stratification

To estimate the popular vote for Biden and Trump, we will employ a post-stratification analysis.

To use the post-stratification technique, we will first divide the data into cells based on different age, gender, state, race, education, and income.

The reason why we choose these variables is that, we think these factors(age, gender,etc) are the basic voter's information which are being included in both census and survey dataset, so it might be easier to have access to these factors. Besides, the USA is a melting pot that consists of immigrants and people with various culture background from all over the world, so race plays an important role in the election and it should be considered into the analysis. Finally and the most importantly, historically, all the states in America could be classified into 3 categories: blue states, red states and swing states. The majority voters in blue states will choose Democratic party, while red states support Republican party. Two parties normally have similar amount of votes in swing states. Therefore, we think that state is also a crucial and interesting factor to determine the result of election; hence, we include it in the analysis.

Next, we apply the model fitted above to each cell to estimate the popular vote proportion for each candidate. To finish up our post-stratification analysis, we will calculate the weighted mean $\hat{y}^{PS}$ of all cells for each candidate with the formula below:

$$\hat{y}^{PS} = \frac{\sum_j N_j \cdot \hat{y_j}}{\sum_j N_j}$$

where $N_j, \hat{y_j}$ are the total count and model estimate of each cell, respectively.

The result can be found at Table 1 in the Result section.

Additional Analysis

In addition to the U.S. national popular vote, We also perform a post-stratification analysis by state. We first remove the variable state from the model and employ the same post-stratification technique on the new

model. For the result, instead of two numbers of popular vote proportion in the previous analysis, we have predicted the popular vote proportion of each candidate for every state, which is illustrated by Fig. 1, 2, 3 in Appendix.

Then we apply the winner-takes-all rule of the U.S. election system by distributing all the state's electoral vote to the candidate with higher popular vote at that state (except for Main and Nebraska which use proportional vote distribution rule) **[3]**. Finally we add up all the electoral votes for each candidate. The result is illustrated at Table 2 in the Result section.

Results

Table 1: Total Popular Vote(%)

Biden	Trump
39.26929	38.56663

Table 2: Total Electoral Vote

Biden	Trump
301	237

We estimate the popular vote proportion for Donald Trump is 38.567% and that for Joe Biden is 39.269%. This is based off our post-stratification analysis modelled by logistic regression, which accounted for age, gender, race, state, income, education.

For our additional analysis, we estimate the electoral vote for Donald Trump is 237 and that for Joe Biden is 301. This is based off our post-stratification analysis of states modelled by logistic regression, which accounted for age, gender, race, income, education. Detailed prediction of each state's electoral vote can be found at Table 7 in Appendix.

Discussion

Summary

In our report, we employ the post-stratification technique to predict the popular vote proportion of Biden and Trump. We first clean the census data by splitting the examples into 838,113 cells based on their age, gender, state, race, education level, and income level, which are our explanatory variables. Next we build two logistic regression models with our explanatory variables and use them to estimate the popular vote proportion of Biden and Trump for each cell. Ultimately, with each cell's estimate, we calculate the weighted mean of these estimates, which is our predicted popular vote of Biden and Trump.

We also conduct an analysis on post-stratification by state to find out the electoral vote of each candidate. The result is illustrated by graph Fig. 1, 2, 3 in Appendix.

Conclusion

We predict that the popular vote in 2020 election is 39.269% for Joe Biden and 38.567% for Donald Trump. Although Biden has a higher popular vote, it does not necessarily mean that he is going to win, because in

the US election system, the popular vote does not determine the winner, and actually candidate with less popular vote has become the president several times in the history [**3**].

What actually determines the winner is the electoral vote. Each state has certain number of electoral votes, and if one candidate has more state popular vote than the other, the leading candidate takes all the electoral vote of that state [**3**]. Therefore, the result in Table 2 determines the winner. As we can see, Joe Biden has more electoral votes than Donald Trump has ($301 > 237$), so we predict that Joe Biden will win the 2020 presidential election.

Weakness

One of the major drawback of our analysis is that the survey data is outdated. The survey is conducted at June 25, 2020, but the election day is at November 3, 2020. Multiple major events had happened during this period, such as Donald Trump getting infected by COVID-19 and two presidential debates. These events could potentially change American people's opinion and hence affect the vote result.

Another weakness is that we include all people in our analysis, including people less than 18 years old, who are not eligible to vote [**4**]. They do not contribute to the final vote outcome of our prediction. However, we do not have enough data to estimate the popular vote of the entire American Population, so removing these examples will shrink our sample size, and hence we could have even more biased prediction. There is a trade-off here between removing irrelavent data and large sample size.

The other weakness would be the small cell counts. When we divide the data into cells, there are many cells with small counts. These cells could bring uncertainties to the prediction and hence become problematic.

Next Steps

Now that we have analyzes the popular vote through a frequentist approach, we can complicate the model to minimize the uncertainty brought by both the small cell counts and small sample size. One of the common way to deal with uncertainties is the Bayesian approach. We may consider to build a Bayesian logistic model for deeper analysis.

References

[**1**] Tausanovitch, Chris and Lynn Vavreck. 2020. Democracy Fund + UCLA Nationscape, October 10-17, 2019 (version 20200814). Retrieved from https://www.voterstudygroup.org/publication/nationscape-data-set.

[**2**] 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

[**3**] Presidential Election Process. (2020). In USA.gov. Retrieved November 2, 2020 from https://www.usa.gov/election

[**4**] Who Can and Can't Vote in U.S. Elections. (2020). In USA.gov. Retrieved November 2, 2020 from https://www.usa.gov/who-can-vote

Appendix

Fig.1 Popular Vote of State Alabama−Kansas (Alphabetically)

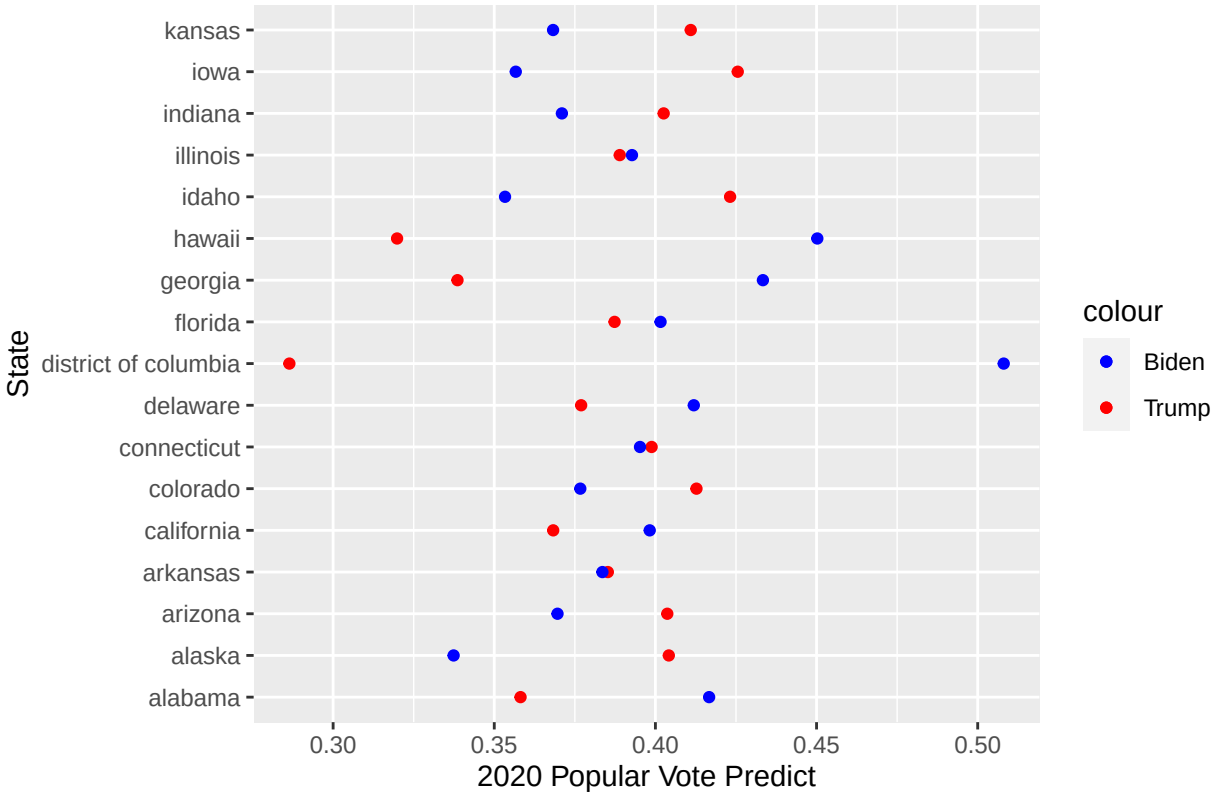

Fig.2 Popular Vote of State Kentucky−North Carolina (Alphabeticall

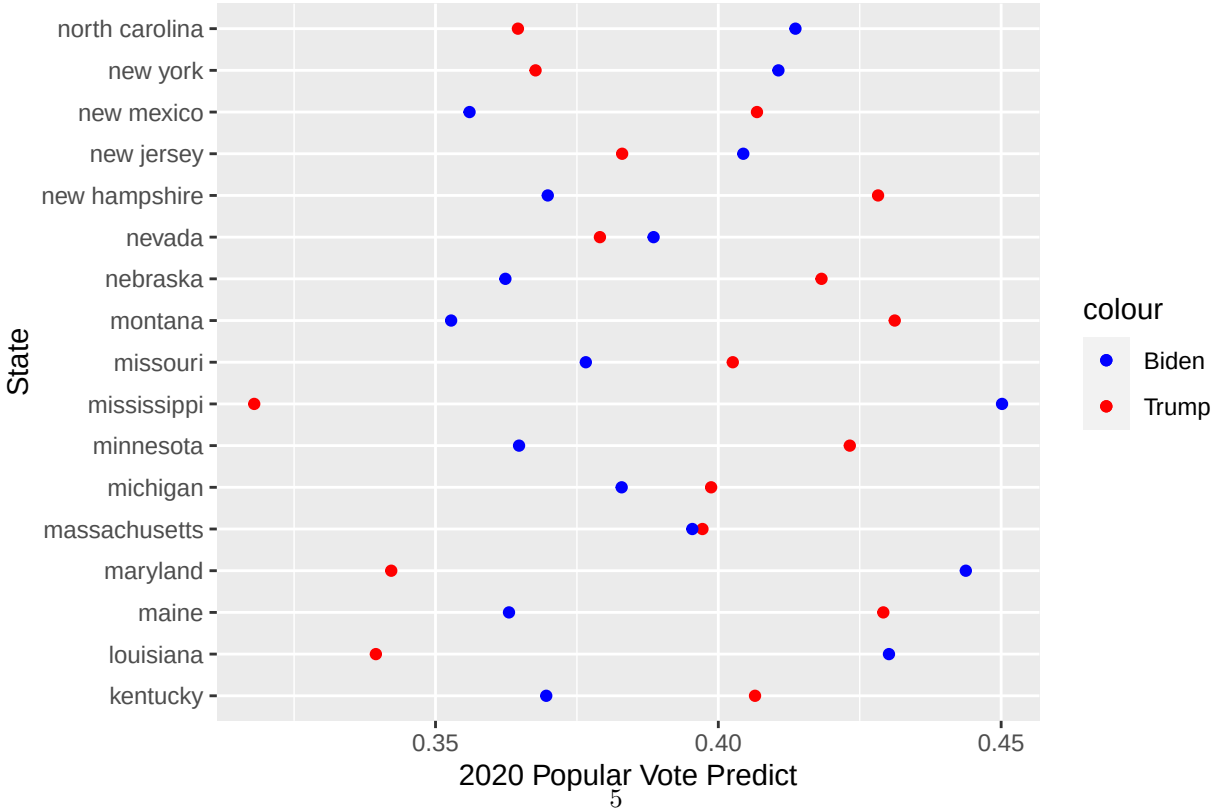

Fig.3 Popular Vote of State North Dakota−Wyoming (Alphabetically)

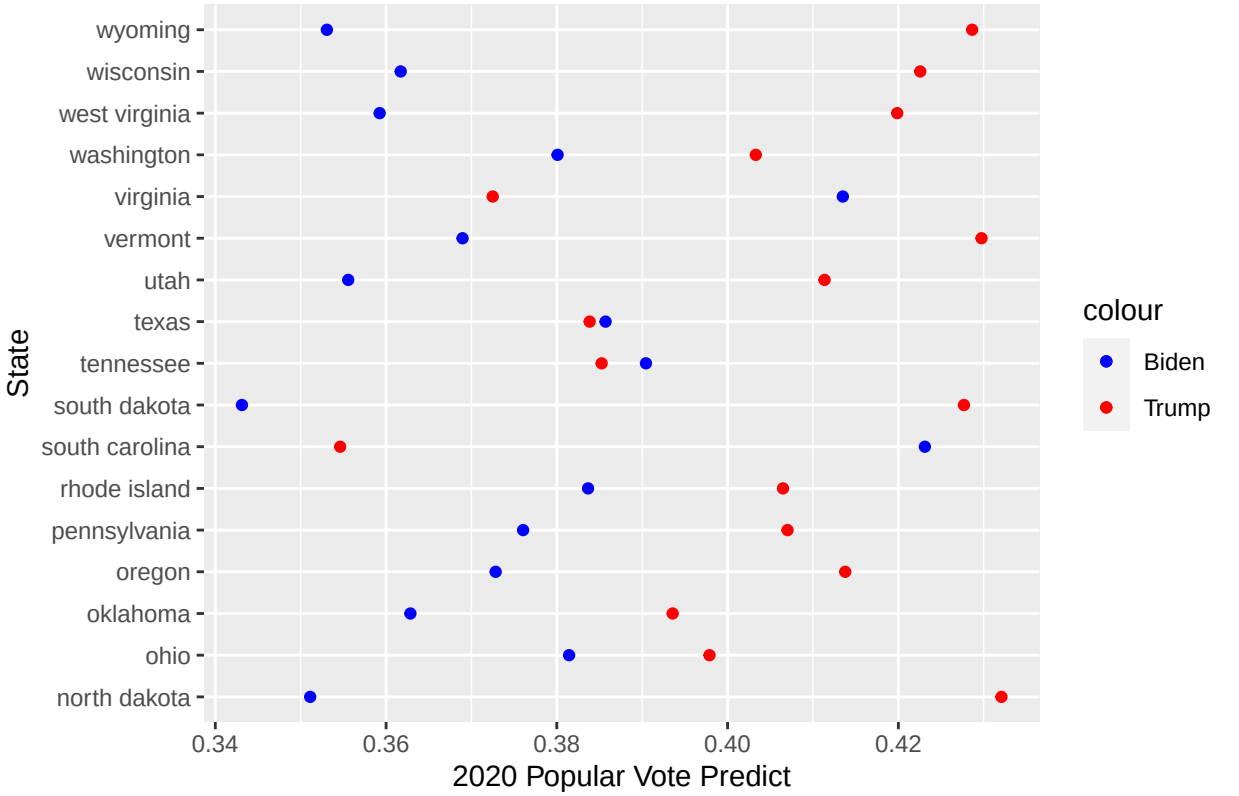

Table 3: Educational Level Reference

Education	Level
3rd Grade or less	0
Middle School - Grades 4 - 8	1
Completed some high school	2
High school graduate	3
Other post high school vocational training	4
Completed some college, but no degree	5
Associate Degree	6
College Degree (such as B.A., B.S.)	7
Completed some graduate, but no degree	8
Masters degree	9
Doctorate degree	10

Table 4: Income Level Reference

Income	Level
Less than $14,999	0
$15,000 to $19,999	1
$20,000 to $24,999	2
$25,000 to $29,999	3
$30,000 to $34,999	4
$35,000 to $39,999	5

Income	Level
$40,000 to $44,999	6
$45,000 to $49,999	7
$50,000 to $54,999	8
$55,000 to $59,999	9
$60,000 to $64,999	10
$65,000 to $69,999	11
$70,000 to $74,999	12
$75,000 to $79,999	13
$80,000 to $84,999	14
$85,000 to $89,999	15
$90,000 to $94,999	16
$95,000 to $99,999	17
$100,000 to $124,999	18
$125,000 to $149,999	19
$150,000 to $174,999	20
$175,000 to $199,999	21
$200,000 to $249,999	22
$250,000 and above	23

Table 5: Popular Vote Logistic Model for Biden

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9419781	0.2706737	-3.4801238	0.0005012
age	0.0000238	0.0018053	0.0131978	0.9894700
gendermale	-0.3898601	0.0585259	-6.6613231	0.0000000
statealaska	-0.3974091	0.8743319	-0.4545288	0.6494482
statearizona	0.0757999	0.3001263	0.2525601	0.8006081
statearkansas	-0.7122854	0.4408362	-1.6157598	0.1061463
statecalifornia	0.3325145	0.2571577	1.2930373	0.1959982
statecolorado	0.2650644	0.3417465	0.7756171	0.4379751
stateconnecticut	0.7173094	0.3585766	2.0004357	0.0454532
statedelaware	0.5938857	0.5169337	1.1488625	0.2506127
statedistrict of columbia	0.8810897	0.5418468	1.6260864	0.1039313
stateflorida	0.0868004	0.2626675	0.3304575	0.7410543
stategeorgia	-0.1332083	0.2976720	-0.4475003	0.6545139
statehawaii	0.1099782	0.5600919	0.1963573	0.8443305
stateidaho	-0.4943665	0.4900360	-1.0088372	0.3130527
stateillinois	0.3636498	0.2757723	1.3186596	0.1872829
stateindiana	0.1239729	0.3197440	0.3877254	0.6982192
stateiowa	0.4931024	0.3891476	1.2671345	0.2051072
statekansas	-0.2630137	0.4141205	-0.6351139	0.5253541
statekentucky	0.5945792	0.3466546	1.7151920	0.0863100
statelouisiana	0.0088148	0.3542057	0.0248861	0.9801458
statemaine	0.8893228	0.5390716	1.6497304	0.0989981
statemaryland	0.2753674	0.3301236	0.8341342	0.4042054
statemassachusetts	0.6747936	0.3155060	2.1387662	0.0324546
statemichigan	0.3519064	0.2950150	1.1928424	0.2329311
stateminnesota	0.4912230	0.3566473	1.3773355	0.1684086
statemississippi	-0.3634771	0.4216916	-0.8619500	0.3887150
statemissouri	0.2744821	0.3197039	0.8585510	0.3905883
statemontana	0.0864297	0.5834595	0.1481332	0.8822376

term	estimate	std.error	statistic	p.value
statenebraska	0.2105155	0.5273560	0.3991905	0.6897528
statenevada	-0.1722215	0.3569023	-0.4825453	0.6294186
statenew hampshire	0.3427903	0.5288636	0.6481639	0.5168790
statenew jersey	0.2204172	0.2881544	0.7649272	0.4443149
statenew mexico	0.4345255	0.4817466	0.9019795	0.3670677
statenew york	0.1558984	0.2626861	0.5934778	0.5528615
statenorth carolina	0.1323204	0.2903625	0.4557076	0.6486003
statenorth dakota	-1.1270454	1.1368327	-0.9913907	0.3214948
stateohio	0.1303136	0.2763867	0.4714902	0.6372907
stateoklahoma	-0.3524742	0.3891159	-0.9058336	0.3650239
stateoregon	0.3872812	0.3328526	1.1635215	0.2446180
statepennsylvania	-0.1444215	0.2807825	-0.5143538	0.6070047
staterhode island	0.5867482	0.7599333	0.7721049	0.4400523
statesouth carolina	-0.6847466	0.3569616	-1.9182638	0.0550776
statesouth dakota	-0.0113627	0.6029788	-0.0188444	0.9849653
statetennessee	-0.5421240	0.3393485	-1.5975435	0.1101446
statetexas	-0.3055834	0.2662489	-1.1477355	0.2510778
stateutah	-0.2927118	0.4391208	-0.6665859	0.5050367
statevermont	1.6936939	0.7074094	2.3942200	0.0166558
statevirginia	0.4243875	0.2883669	1.4716930	0.1411038
statewashington	0.3446415	0.3160892	1.0903300	0.2755678
statewest virginia	-0.0709783	0.4436691	-0.1599803	0.8728966
statewisconsin	0.3067046	0.3135543	0.9781544	0.3279980
statewyoming	-11.6078423	186.7329196	-0.0621628	0.9504332
raceamerican indian or alaska native	-0.2706323	0.2678403	-1.0104241	0.3122922
raceblack/african american/negro	1.3337996	0.0983338	13.5639930	0.0000000
racechinese	0.5135446	0.2526266	2.0328207	0.0420706
racejapanese	1.2526266	0.5626726	2.2262086	0.0260002
raceother asian or pacific islander	0.2757084	0.1640959	1.6801666	0.0929249
raceother race, nec	0.4278491	0.1187532	3.6028435	0.0003148
education_level	0.1115473	0.0164484	6.7816528	0.0000000
income_level	-0.0152221	0.0045531	-3.3432579	0.0008280

Table 6: Popular Vote Logistic Model for Trump

term	estimate	std.error	statistic	p.value
(Intercept)	-0.8536856	0.2817458	-3.0299855	0.0024457
age	0.0127916	0.0018793	6.8066316	0.0000000
gendermale	0.4626900	0.0607030	7.6221970	0.0000000
statealaska	0.5765868	0.8190811	0.7039434	0.4814680
statearizona	-0.2309092	0.3085189	-0.7484443	0.4541922
statearkansas	0.3610167	0.4233330	0.8527959	0.3937725
statecalifornia	-0.5723210	0.2694552	-2.1239934	0.0336707
statecolorado	-0.1527076	0.3529449	-0.4326669	0.6652568
stateconnecticut	-1.3239955	0.4045292	-3.2729293	0.0010644
statedelaware	-0.6989877	0.5518894	-1.2665357	0.2053214
statedistrict of columbia	-0.4072869	0.5670608	-0.7182422	0.4726080
stateflorida	-0.1831250	0.2736941	-0.6690865	0.5034403
stategeorgia	0.1150630	0.3135163	0.3670081	0.7136130
statehawaii	-0.3587817	0.5962841	-0.6016959	0.5473766
stateidaho	0.1318690	0.4686022	0.2814093	0.7783965

term	estimate	std.error	statistic	p.value
stateillinois	-0.6096251	0.2895515	-2.1054117	0.0352555
stateindiana	-0.3247377	0.3290451	-0.9869094	0.3236870
stateiowa	-0.7179694	0.4076081	-1.7614208	0.0781672
statekansas	0.2257868	0.4105606	0.5499475	0.5823554
statekentucky	-0.3778982	0.3581126	-1.0552496	0.2913112
statelouisiana	-0.0290270	0.3707321	-0.0782964	0.9375922
statemaine	-0.6354337	0.5544974	-1.1459633	0.2518103
statemaryland	-0.5814730	0.3502623	-1.6601071	0.0968929
statemassachusetts	-1.1668433	0.3450628	-3.3815389	0.0007208
statemichigan	-0.4186083	0.3072923	-1.3622479	0.1731197
stateminnesota	-0.2178716	0.3682857	-0.5915832	0.5541297
statemississippi	0.0885694	0.4476968	0.1978335	0.8431753
statemissouri	-0.2841194	0.3330740	-0.8530221	0.3936470
statemontana	-0.1792159	0.5869270	-0.3053462	0.7601025
statenebraska	-0.4135857	0.5388889	-0.7674786	0.4427970
statenevada	-0.0173129	0.3713476	-0.0466218	0.9628147
statenew hampshire	-0.5116359	0.5475203	-0.9344600	0.3500666
statenew jersey	-0.4578179	0.3008299	-1.5218497	0.1280468
statenew mexico	-1.2241233	0.5752669	-2.1279226	0.0333435
statenew york	-0.4525897	0.2739706	-1.6519646	0.0985418
statenorth carolina	-0.1936979	0.3031955	-0.6388548	0.5229174
statenorth dakota	0.0375958	0.8875052	0.0423613	0.9662107
stateohio	-0.3785330	0.2881326	-1.3137460	0.1889317
stateoklahoma	0.0679744	0.3833057	0.1773372	0.8592435
stateoregon	-0.6510612	0.3491381	-1.8647669	0.0622141
statepennsylvania	-0.3423712	0.2890928	-1.1842950	0.2362963
staterhode island	-0.9080129	0.8731804	-1.0398915	0.2983903
statesouth carolina	0.2782429	0.3486791	0.7979913	0.4248755
statesouth dakota	-0.0322989	0.5901369	-0.0547311	0.9563527
statetennessee	0.3077448	0.3383208	0.9096244	0.3630206
statetexas	-0.0500859	0.2756227	-0.1817190	0.8558033
stateutah	-0.7329260	0.4382014	-1.6725777	0.0944105
statevermont	-2.3921182	1.0739355	-2.2274318	0.0259184
statevirginia	-0.6283411	0.3049267	-2.0606297	0.0393384
statewashington	-0.4741097	0.3303567	-1.4351450	0.1512458
statewest virginia	-0.1445660	0.4347942	-0.3324929	0.7395171
statewisconsin	-0.6250408	0.3286516	-1.9018339	0.0571929
statewyoming	-12.7569942	187.0761147	-0.0681915	0.9456332
raceamerican indian or alaska native	-0.1082262	0.2594758	-0.4170955	0.6766086
raceblack/african american/negro	-2.0019083	0.1468025	-13.6367417	0.0000000
racechinese	-1.2893982	0.3377473	-3.8176416	0.0001347
racejapanese	-0.9293168	0.6057060	-1.5342704	0.1249631
raceother asian or pacific islander	-0.5095525	0.1804602	-2.8236277	0.0047484
raceother race, nec	-0.6299489	0.1343317	-4.6895034	0.0000027
education_level	-0.0358276	0.0169546	-2.1131447	0.0345884
income_level	0.0371305	0.0047203	7.8661945	0.0000000

Table 7: Electoral Vote of Each State

State	Trump	Biden
alabama	0	9
alaska	3	0
arizona	11	0
arkansas	6	0
california	0	55
colorado	9	0
connecticut	7	0
delaware	0	3
district of columbia	0	3
florida	0	29
georgia	0	16
hawaii	0	4
idaho	4	0
illinois	0	20
indiana	11	0
iowa	6	0
kansas	6	0
kentucky	8	0
louisiana	0	8
maine	3	1
maryland	0	10
massachusetts	11	0
michigan	16	0
minnesota	10	0
mississippi	0	6
missouri	10	0
montana	3	0
nebraska	3	2
nevada	0	6
new hampshire	4	0
new jersey	0	14
new mexico	5	0
new york	0	29
north carolina	0	15
north dakota	3	0
ohio	18	0
oklahoma	7	0
oregon	7	0
pennsylvania	20	0
rhode island	4	0
south carolina	0	9
south dakota	3	0
tennessee	0	11
texas	0	38
utah	6	0
vermont	3	0
virginia	0	13
washington	12	0
west virginia	5	0
wisconsin	10	0

State	Trump	Biden
wyoming	3	0