ADSP 31014 Statistical Models for Data Science

Course Project Part 2 - Hira Stanley

Business Problem

Political Analysts would like to understand what factors drove the results of the 2020 U.S. Presidential election (specifically the percentage of votes for Joe Biden in each county), such as demographics, prior election results, and the impact of COVID.

Model Selection

The nature of this data set allows for either logistic or linear regression models to be performed. With logistic, we could predict a binary outcome of the 2020 election. However, because we also have access to percentage of vote share for the 2020 winner in each county, it would be more helpful in the context of the problem to do a linear regression so we can see how the other factors affect the vote share rather than only the final outcome.

Linear Regression Model

We build a linear regression model using a cleaned dataset of 3,021 rows. The predicted variable is percentage20_Joe_Biden. The model formula is:

percentage20_Joe_Biden ~ factor(state)*NonWhite + case_pct*Black + turnout_difference + Income + Men_pct + Transit

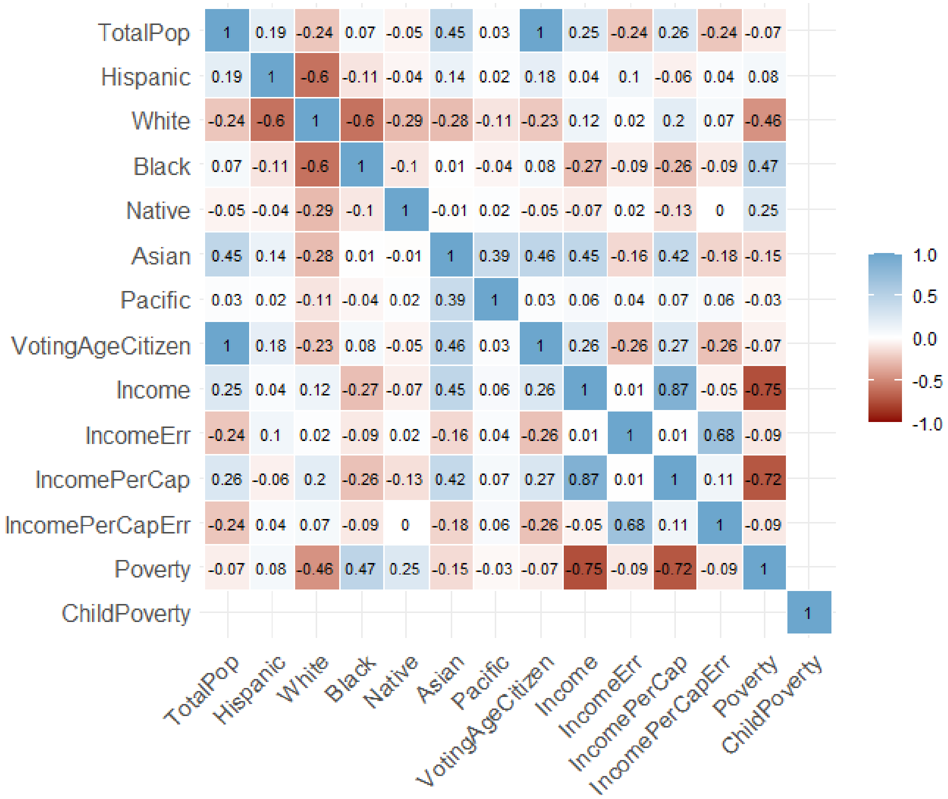
This model has 107 parameters with a R^2 of 0.76. See Appendix 1 for model summary.

Multicollinearity and Feature Engineering

We divide our 54 potentially explanatory variables into five sections to investigate and omit any columns that represent multicollinearity with each other or our predicted variable:

- Election results: The glaring multicollinearity is the 2016 election results with 0.98 correlation to our predicted variable. The simplest model would just be a linear relationship between percentage16_Hillary_Clinton and our predicted variable, which offers a r-squared of 0.95, but does not help us understand any other factors. We also omit total_votes2020, as it is multicollinear with total_votes16 (0.96) and turnout_difference (0.78).
- Demographic: TotalPop and VotingAgeCitizen have perfect correlation of 1, we remove TotalPop because in this context, the vote share should represent only those who are eligible to vote. There are 5 columns representing racial demographics, and none stand out with high multicollinearity that would inhibit explainability. We feature engineer a column called NonWhite, which sums all racial categories that are not White to represent a percentage of the population. A heatmap of correlation showed other potential multicollinearities, that allow us to reduce the 6 income-related columns to just income.
- Commute: Transit would offer some explainability without high correlation.
- Employment: Knowing the context of the dataset, we just keep Unemployment.
- Others: We previously converted COVID cases and deaths into a percentage of the population in that county (case_pct and death_pct) to not be multicollinear with TotalPop or VotingAgeCitizen. We also previously converted gender in a similar way to just represent the male population as a percentage (Men_pct).

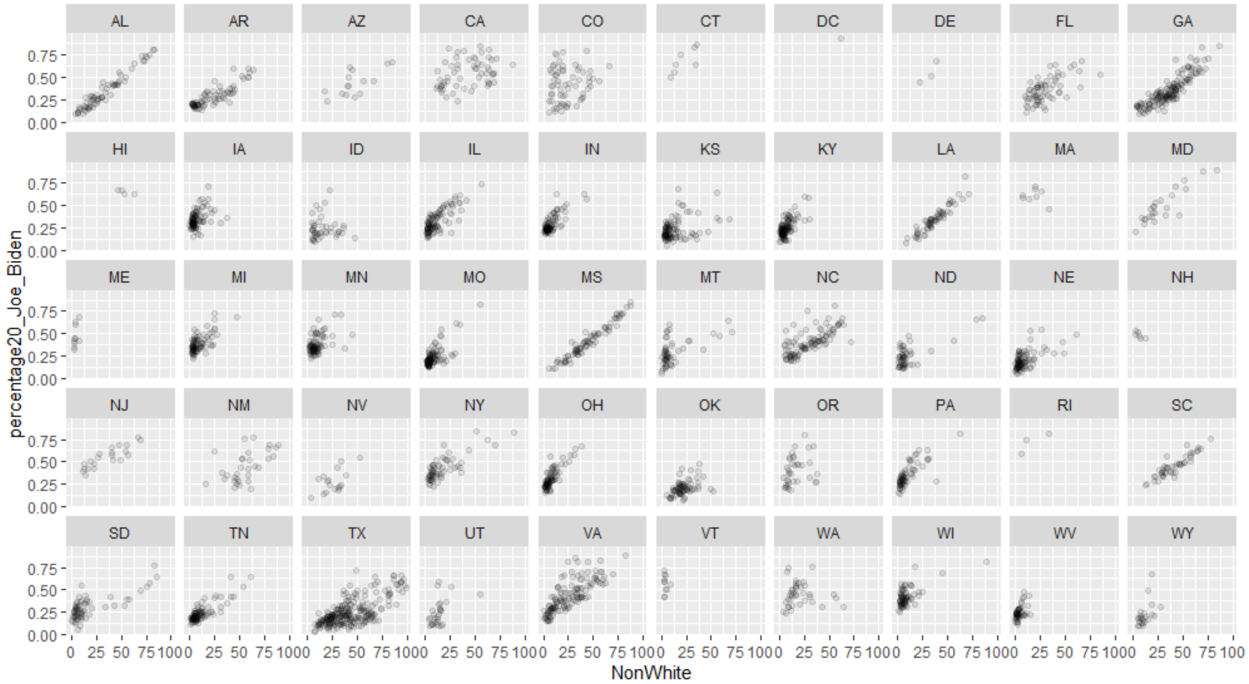
Correlation Matrix Heatmap of Demographic Data (remainder in Appendix 2)



	TotalPop	Hispanic	White	Black	Native	Asian	Pacific	VotingAgeCitizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty
TotalPop	1	0.19	-0.24	0.07	-0.05	0.45	0.03	1	0.25	-0.24	0.26	-0.24	-0.07	
Hispanic	0.19	1	-0.6	-0.11	-0.04	0.14	0.02	0.18	0.04	0.1	-0.06	0.04	0.08	
White	-0.24	-0.6	1	-0.6	-0.29	-0.28	-0.11	-0.23	0.12	0.02	0.2	0.07	-0.46	
Black	0.07	-0.11	-0.6	1	-0.1	0.01	-0.04	0.08	-0.27	-0.09	-0.26	-0.09	0.47	
Native	-0.05	-0.04	-0.29	-0.1	1	-0.01	0.02	-0.05	-0.07	0.02	-0.13	0	0.25	
Asian	0.45	0.14	-0.28	0.01	-0.01	1	0.39	0.46	0.45	-0.16	0.42	-0.18	-0.15	
Pacific	0.03	0.02	-0.11	-0.04	0.02	0.39	1	0.03	0.06	0.04	0.07	0.06	-0.03	
VotingAgeCitizen	1	0.18	-0.23	0.08	-0.05	0.46	0.03	1	0.26	-0.26	0.27	-0.26	-0.07	
Income	0.25	0.04	0.12	-0.27	-0.07	0.45	0.06	0.26	1	0.01	0.87	-0.05	-0.75	
IncomeErr	-0.24	0.1	0.02	-0.09	0.02	-0.16	0.04	-0.26	0.01	1	0.01	0.68	-0.09	
IncomePerCap	0.26	-0.06	0.2	-0.26	-0.13	0.42	0.07	0.27	0.87	0.01	1	0.11	-0.72	
IncomePerCapErr	-0.24	0.04	0.07	-0.09	0	-0.18	0.06	-0.26	-0.05	0.68	0.11	1	-0.09	
Poverty	-0.07	0.08	-0.46	0.47	0.25	-0.15	-0.03	-0.07	-0.75	-0.09	-0.72	-0.09	1	
ChildPoverty														1

Interaction Terms

Categorical field state interacts with percentage of the population being NonWhite.



Useful and Less Useful Terms

Interaction terms:

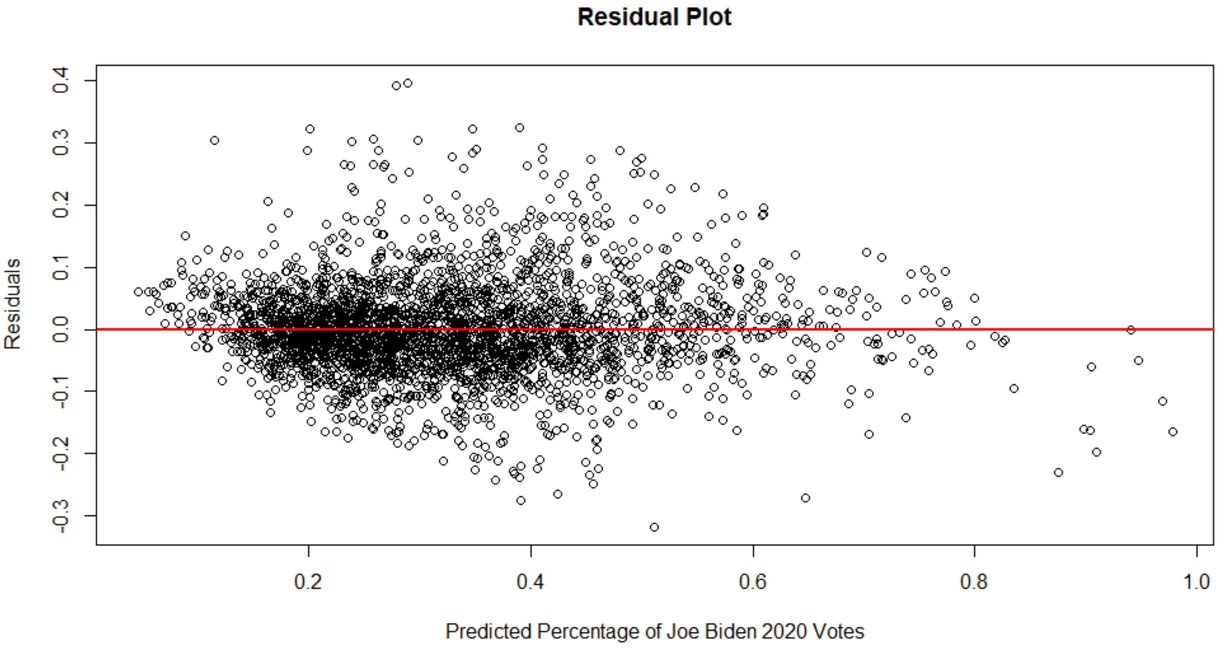
- State interacting with the NonWhite percentage of the population significantly impacts the response variable. The p-values that are less than 0.05 in16 states indicate that the effect of NonWhite varies by state, with both positive and negative relationships observed.
- COVID cases (case_pct) when interacting with race had p-value 9.73e-05 and the final model has adjusted R^2 of 0.76. When input into the model separately without interaction, the adjusted R^2 is 0.75. This means the interaction term is more useful in providing a nuanced understanding of the data.

Less useful terms that we omitted either due to high p-values, or that their removal did not change the model's explanatory power (adjusted R^2).

- Unemployment had p-value 0.04
- Lat had p-value 0.26

Residuals

Plot of the response variable vs. residuals shows the model is a good fit, with the residuals randomly scattered around the horizontal line.



Model Trade-offs

The model where state interacted with NonWhite and case_pct had a slightly higher R^2 at 0.77, but the three-way interaction created 204 parameters. In the trade-off between model complexity and achieving a better fit, we choose the less complex one with 108 parameters.

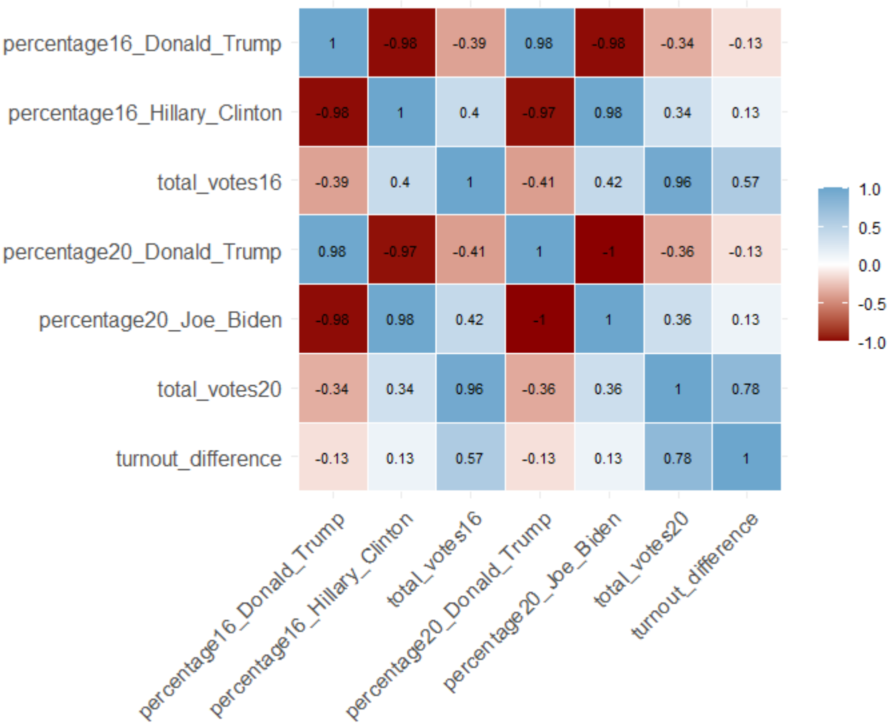
Heteroscedasticity is present in multiple explanatory variables. When we take the log of the response variable, it reduces the R^2 down to 0.70. Despite the violation of the homoscedasticity assumption, we prioritize prediction accuracy in this trade-off. More detail in Appendices 3 and 4.

Appendix 1: Final Model Summary

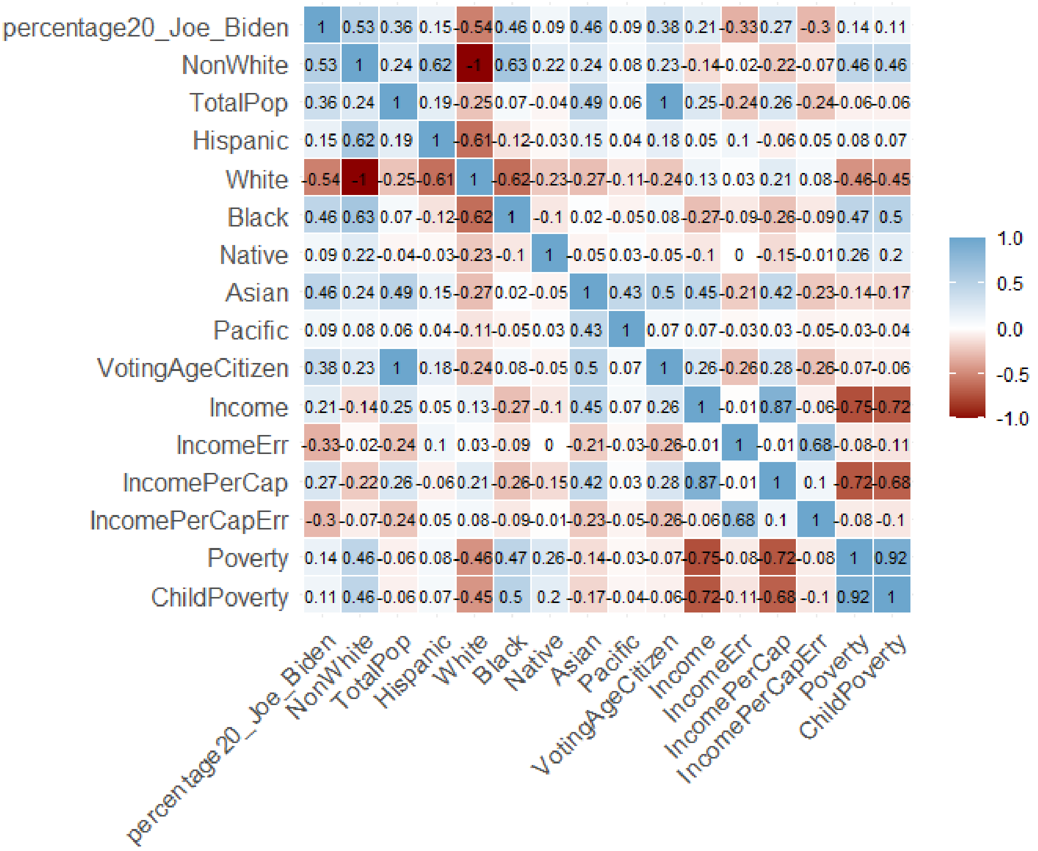
```
Call:
lm(formula = percentage20_Joe_Biden ~ factor(state) * NonWhite + 
    case_pct * Black + turnout_difference + Income + Men_pct + 
    Transit, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.31771 -0.04380 -0.00571  0.03599  0.39612 

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                4.060e-01  3.806e-02  10.667  < 2e-16 ***
factor(state)AR            1.146e-01  2.289e-02   5.004 5.95e-07 ***
factor(state)AZ            4.532e-02  5.893e-02   0.769 0.441999    
factor(state)CA            3.280e-01  3.045e-02  10.773  < 2e-16 ***
factor(state)CO            2.267e-01  2.641e-02   8.583  < 2e-16 ***
factor(state)CT            3.261e-01  7.152e-02   4.559 5.35e-06 ***
factor(state)DC           -2.071e-02  8.452e-02  -0.245 0.806447    
factor(state)DE            9.797e-02  2.141e-01   0.458 0.647212    
factor(state)FL            1.270e-01  2.758e-02   4.604 4.33e-06 ***
factor(state)GA           -2.923e-03  2.259e-02  -0.129 0.897044    
factor(state)HI            1.037e+00  3.439e-01   3.016 0.002587 ** 
factor(state)IA            2.062e-01  2.146e-02   9.607  < 2e-16 ***
factor(state)ID            1.591e-01  2.692e-02   5.910 3.83e-09 ***
factor(state)IL            1.480e-01  2.123e-02   6.972 3.85e-12 ***
factor(state)IN            1.264e-01  2.155e-02   5.863 5.06e-09 ***
factor(state)KS            9.636e-02  2.109e-02   4.570 5.09e-06 ***
factor(state)KY            1.167e-01  2.125e-02   5.493 4.29e-08 ***
factor(state)LA           -7.022e-02  3.358e-02  -2.091 0.036632 *  
factor(state)MA            4.392e-01  6.809e-02   6.450 1.30e-10 ***
factor(state)MD            1.480e-01  3.447e-02   4.294 1.82e-05 ***
factor(state)ME            2.211e-01  6.196e-02   3.568 0.000365 ***
factor(state)MI            2.420e-01  2.252e-02  10.744  < 2e-16 ***
factor(state)MN            2.111e-01  2.247e-02   9.395  < 2e-16 ***
factor(state)MO            1.050e-01  2.091e-02   5.022 5.42e-07 ***
factor(state)MS           -3.268e-02  2.711e-02  -1.205 0.228136    
factor(state)MT            1.500e-01  2.207e-02   6.793 1.32e-11 ***
factor(state)NC            1.680e-01  2.360e-02   7.120 1.36e-12 ***
factor(state)ND            1.056e-01  2.205e-02   4.786 1.78e-06 ***
factor(state)NE            5.492e-02  2.099e-02   2.617 0.008926 ** 
factor(state)NH            3.920e-01  6.827e-02   5.742 1.03e-08 ***
factor(state)NJ            1.912e-01  4.196e-02   4.557 5.41e-06 ***
factor(state)NM           -2.018e-03  5.002e-02  -0.040 0.967814    
factor(state)NV           -1.142e-02  5.172e-02  -0.221 0.825270    
factor(state)NY            2.717e-01  2.377e-02  11.427  < 2e-16 ***
factor(state)OH            1.231e-01  2.192e-02   5.617 2.12e-08 ***
factor(state)OK            2.312e-02  2.952e-02   0.783 0.433595    
factor(state)OR            2.248e-01  3.116e-02   7.214 6.90e-13 ***
factor(state)PA            1.629e-01  2.263e-02   7.198 7.75e-13 ***
factor(state)RI            3.878e-01  7.860e-02   4.934 8.52e-07 ***
factor(state)SC            7.413e-02  3.739e-02   1.982 0.047534 *  
factor(state)SD            1.501e-01  2.160e-02   6.948 4.55e-12 ***
factor(state)TN            1.160e-01  2.094e-02   5.543 3.24e-08 ***
factor(state)TX           -8.246e-02  2.173e-02  -3.795 0.000151 ***
factor(state)UT            1.440e-01  5.686e-02   2.533 0.011377 *  
factor(state)VA            1.219e-01  2.272e-02   5.364 8.77e-08 ***
factor(state)VT            4.690e-01  6.330e-02   7.410 1.64e-13 ***
factor(state)WA            3.214e-01  2.882e-02  11.153  < 2e-16 ***
factor(state)WI            2.801e-01  2.139e-02  13.095  < 2e-16 ***
factor(state)WV            1.253e-01  2.446e-02   5.123 3.21e-07 ***
factor(state)WY           -3.563e-02  3.934e-02  -0.906 0.365090    
NonWhite                   4.939e-03  6.065e-04   8.145 5.58e-16 ***
```

```
case_pct                 -8.977e-01  1.418e-01  -6.329 2.85e-10 ***
Black                     3.044e-03  4.521e-04   6.733 1.99e-11 ***
turnout_difference        1.620e-07  3.422e-08   4.735 2.30e-06 ***
Income                    2.634e-06  1.453e-07  18.125  < 2e-16 ***
Men_pct                  -8.661e-01  6.490e-02 -13.346  < 2e-16 ***
Transit                   9.703e-03  8.246e-04  11.767  < 2e-16 ***
factor(state)AR:NonWhite -2.859e-03  6.811e-04  -4.198 2.78e-05 ***
factor(state)AZ:NonWhite  1.477e-03  1.261e-03   1.171 0.241865    
factor(state)CA:NonWhite -3.795e-03  7.995e-04  -4.747 2.17e-06 ***
factor(state)CO:NonWhite -1.494e-03  9.124e-04  -1.637 0.101738    
factor(state)CT:NonWhite  2.934e-03  2.840e-03   1.033 0.301665    
factor(state)DC:NonWhite         NA         NA      NA       NA    
factor(state)DE:NonWhite  2.058e-03  6.560e-03   0.314 0.753737    
factor(state)FL:NonWhite -1.361e-03  8.136e-04  -1.673 0.094443 .  
factor(state)GA:NonWhite -3.642e-05  5.788e-04  -0.063 0.949830    
factor(state)HI:NonWhite -1.595e-02  6.469e-03  -2.466 0.013721 *  
factor(state)IA:NonWhite  1.509e-03  1.336e-03   1.130 0.258703    
factor(state)ID:NonWhite -3.838e-03  1.227e-03  -3.128 0.001779 ** 
factor(state)IL:NonWhite  2.576e-04  8.867e-04   0.290 0.771464    
factor(state)IN:NonWhite  2.788e-03  1.217e-03   2.292 0.021979 *  
factor(state)KS:NonWhite -1.527e-03  8.307e-04  -1.839 0.066062 .  
factor(state)KY:NonWhite  1.749e-03  1.373e-03   1.273 0.203002    
factor(state)LA:NonWhite  1.117e-03  8.530e-04   1.310 0.190413    
factor(state)MA:NonWhite -6.541e-03  3.232e-03  -2.024 0.043100 *  
factor(state)MD:NonWhite -2.328e-03  1.018e-03  -2.286 0.022314 *  
factor(state)ME:NonWhite  2.905e-02  1.367e-02   2.124 0.033740 *  
factor(state)MI:NonWhite -5.177e-04  1.160e-03  -0.446 0.655541    
factor(state)MN:NonWhite  3.195e-04  1.239e-03   0.258 0.796483    
factor(state)MO:NonWhite  7.140e-04  1.215e-03   0.588 0.556678    
factor(state)MS:NonWhite  1.292e-05  6.132e-04   0.021 0.983195    
factor(state)MT:NonWhite  9.951e-04  8.989e-04   1.107 0.268381    
factor(state)NC:NonWhite -2.317e-03  6.466e-04  -3.583 0.000345 ***
factor(state)ND:NonWhite  8.075e-04  8.813e-04   0.916 0.359568    
factor(state)NE:NonWhite  1.088e-03  9.635e-04   1.129 0.258790    
factor(state)NH:NonWhite -1.387e-02  9.639e-03  -1.439 0.150186    
factor(state)NJ:NonWhite -2.921e-03  1.114e-03  -2.622 0.008787 ** 
factor(state)NM:NonWhite  1.734e-03  1.004e-03   1.727 0.084292 .  
factor(state)NV:NonWhite  1.084e-03  1.795e-03   0.604 0.546041    
factor(state)NY:NonWhite -7.948e-03  1.006e-03  -7.902 3.84e-15 ***
factor(state)OH:NonWhite  4.054e-03  1.258e-03   3.224 0.001278 ** 
factor(state)OK:NonWhite -5.149e-04  1.136e-03  -0.453 0.650405    
factor(state)OR:NonWhite -3.386e-04  1.455e-03  -0.233 0.816016    
factor(state)PA:NonWhite  3.089e-04  1.047e-03   0.295 0.767959    
factor(state)RI:NonWhite  3.898e-03  3.672e-03   1.062 0.288522    
factor(state)SC:NonWhite -9.033e-04  8.756e-04  -1.032 0.302323    
factor(state)SD:NonWhite  1.584e-04  7.534e-04   0.210 0.833501    
factor(state)TN:NonWhite -2.062e-03  8.232e-04  -2.504 0.012317 *  
factor(state)TX:NonWhite  5.121e-04  6.611e-04   0.775 0.438616    
factor(state)UT:NonWhite  3.223e-04  2.180e-03   0.148 0.882485    
factor(state)VA:NonWhite -1.545e-03  7.028e-04  -2.198 0.027994 *  
factor(state)VT:NonWhite -1.351e-02  1.410e-02  -0.958 0.337939    
factor(state)WA:NonWhite -4.851e-03  1.080e-03  -4.493 7.31e-06 ***
factor(state)WI:NonWhite  3.008e-04  9.784e-04   0.307 0.758501    
factor(state)WV:NonWhite  2.749e-03  3.007e-03   0.914 0.360642    
factor(state)WY:NonWhite  7.235e-03  2.794e-03   2.590 0.009657 ** 
case_pct:Black            2.771e-02  7.101e-03   3.903 9.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0769 on 2915 degrees of freedom
Multiple R-squared:  0.7616,	Adjusted R-squared:  0.753 
F-statistic: 88.71 on 105 and 2915 DF,  p-value: < 2.2e-16
```

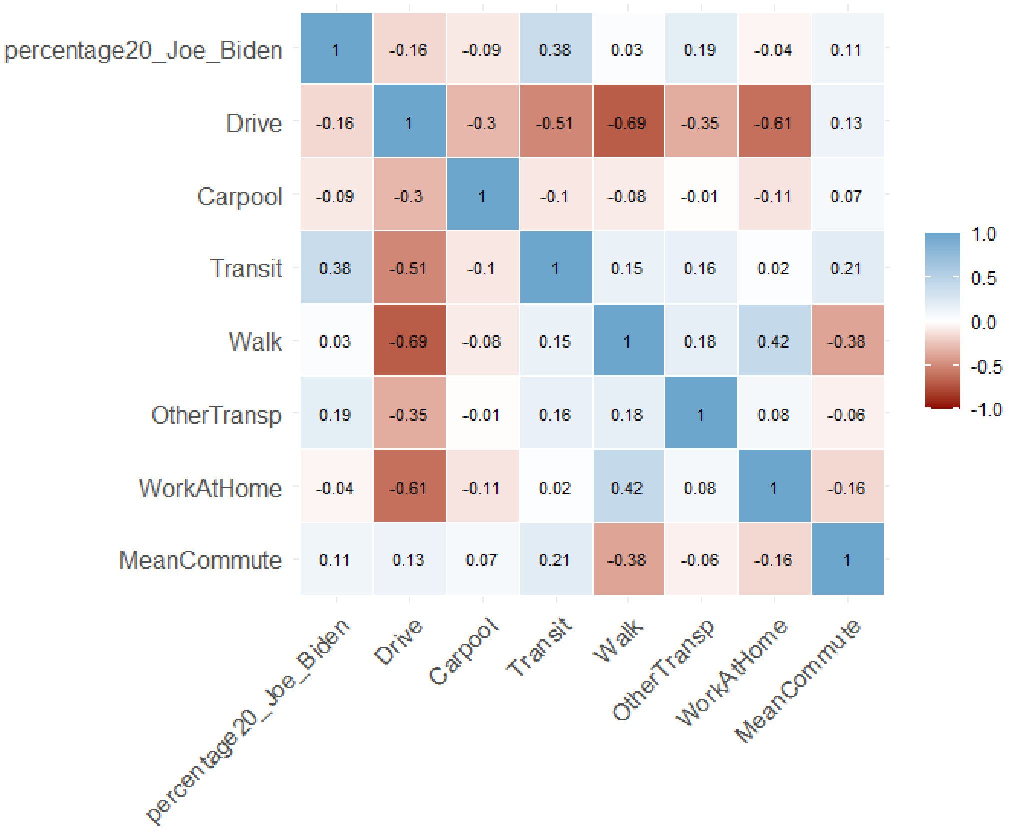
Appendix 2: Correlation Matrix Heatmaps

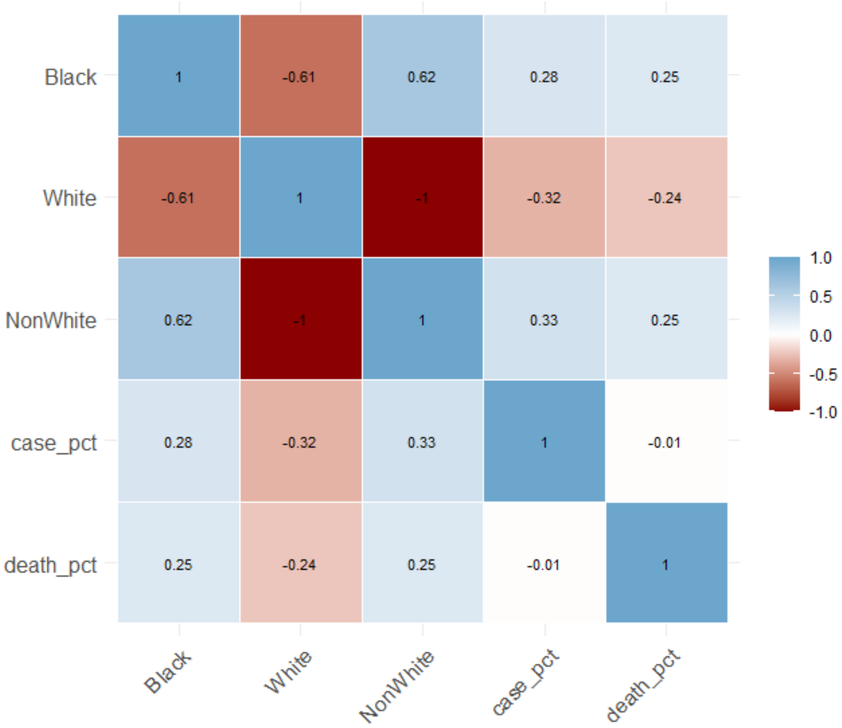


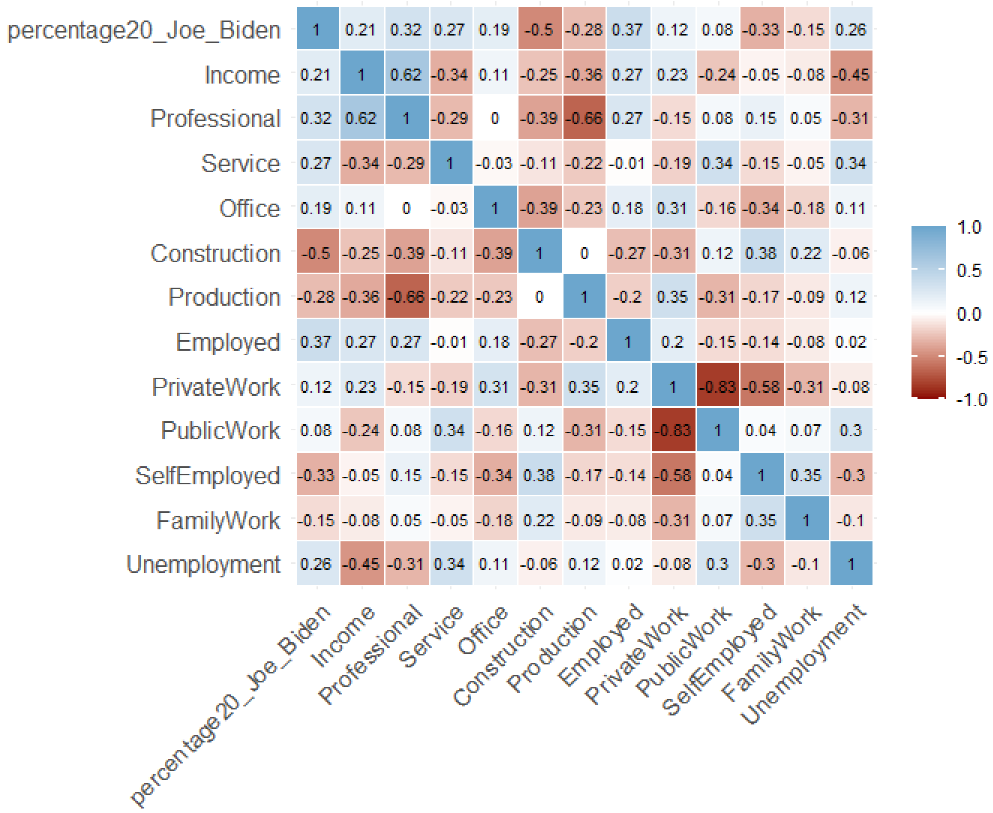
	percentage16_Donald_Trump	percentage16_Hillary_Clinton	total_votes16	percentage20_Donald_Trump	percentage20_Joe_Biden	total_votes20	turnout_difference
percentage16_Donald_Trump	1	-0.98	-0.39	0.98	-0.98	-0.34	-0.13
percentage16_Hillary_Clinton	-0.98	1	0.4	-0.97	0.98	0.34	0.13
total_votes16	-0.39	0.4	1	-0.41	0.42	0.96	0.57
percentage20_Donald_Trump	0.98	-0.97	-0.41	1	-1	-0.36	-0.13
percentage20_Joe_Biden	-0.98	0.98	0.42	-1	1	0.36	0.13
total_votes20	-0.34	0.34	0.96	-0.36	0.36	1	0.78
turnout_difference	-0.13	0.13	0.57	-0.13	0.13	0.78	1



	percentage20_Joe_Biden	NonWhite	TotalPop	Hispanic	White	Black	Native	Asian	Pacific	VotingAgeCitizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty
percentage20_Joe_Biden	1	0.53	0.36	0.15	-0.54	0.46	0.09	0.46	0.09	0.38	0.21	-0.33	0.27	-0.3	0.14	0.11
NonWhite	0.53	1	0.24	0.62	-1	0.63	0.22	0.24	0.08	0.23	-0.14	-0.02	-0.22	-0.07	0.46	0.46
TotalPop	0.36	0.24	1	0.19	-0.25	0.07	-0.04	0.49	0.06	1	0.25	-0.24	0.26	-0.24	-0.06	-0.06
Hispanic	0.15	0.62	0.19	1	-0.61	-0.12	-0.03	0.15	0.04	0.18	0.05	0.1	-0.06	0.05	0.08	0.07
White	-0.54	-1	-0.25	-0.61	1	-0.62	-0.23	-0.27	-0.11	-0.24	0.13	0.03	0.21	0.08	-0.46	-0.45
Black	0.46	0.63	0.07	-0.12	-0.62	1	-0.1	0.02	-0.05	0.08	-0.27	-0.09	-0.26	-0.09	0.47	0.5
Native	0.09	0.22	-0.04	-0.03	-0.23	-0.1	1	-0.05	0.03	-0.05	-0.1	0	-0.15	-0.01	0.26	0.2
Asian	0.46	0.24	0.49	0.15	-0.27	0.02	-0.05	1	0.43	0.5	0.45	-0.21	0.42	-0.23	-0.14	-0.17
Pacific	0.09	0.08	0.06	0.04	-0.11	-0.05	0.03	0.43	1	0.07	0.07	-0.03	0.03	-0.05	-0.03	-0.04
VotingAgeCitizen	0.38	0.23	1	0.18	-0.24	0.08	-0.05	0.5	0.07	1	0.26	-0.26	0.28	-0.26	-0.07	-0.06
Income	0.21	-0.14	0.25	0.05	0.13	-0.27	-0.1	0.45	0.07	0.26	1	-0.01	0.87	-0.06	-0.75	-0.72
IncomeErr	-0.33	-0.02	-0.24	0.1	0.03	-0.09	0	-0.21	-0.03	-0.26	-0.01	1	-0.01	0.68	-0.08	-0.11
IncomePerCap	0.27	-0.22	0.26	-0.06	0.21	-0.26	-0.15	0.42	0.03	0.28	0.87	-0.01	1	0.1	-0.72	-0.68
IncomePerCapErr	-0.3	-0.07	-0.24	0.05	0.08	-0.09	-0.01	-0.23	-0.05	-0.26	-0.06	0.68	0.1	1	-0.08	-0.1
Poverty	0.14	0.46	-0.06	0.08	-0.46	0.47	0.26	-0.14	-0.03	-0.07	-0.75	-0.08	-0.72	-0.08	1	0.92
ChildPoverty	0.11	0.46	-0.06	0.07	-0.45	0.5	0.2	-0.17	-0.04	-0.06	-0.72	-0.11	-0.68	-0.1	0.92	1


percentage20_Joe_Biden
1
-0.16
-0.09
0.38
0.03
0.19
-0.04
0.11
Drive
-0.16
1
-0.3
-0.51
-0.69
-0.35
-0.61
0.13
Carpool
-0.09
-0.3
1
-0.1
-0.08
-0.01
-0.11
0.07
Transit
0.38
-0.51
-0.1
1
0.15
0.16
0.02
0.21
Walk
0.03
-0.69
-0.08
0.15
1
0.18
0.42
-0.38
OtherTransp
0.19
-0.35
-0.01
0.16
0.18
1
0.08
-0.06
WorkAtHome
-0.04
-0.61
-0.11
0.02
0.42
0.08
1
-0.16
MeanCommute
0.11
0.13
0.07
0.21
-0.38
-0.06
-0.16
1
percentage20_Joe_Biden
Drive
Carpool
Transit
Walk
OtherTransp
WorkAtHome
MeanCommute
1.0
0.5
0.0
-0.5
-1.0

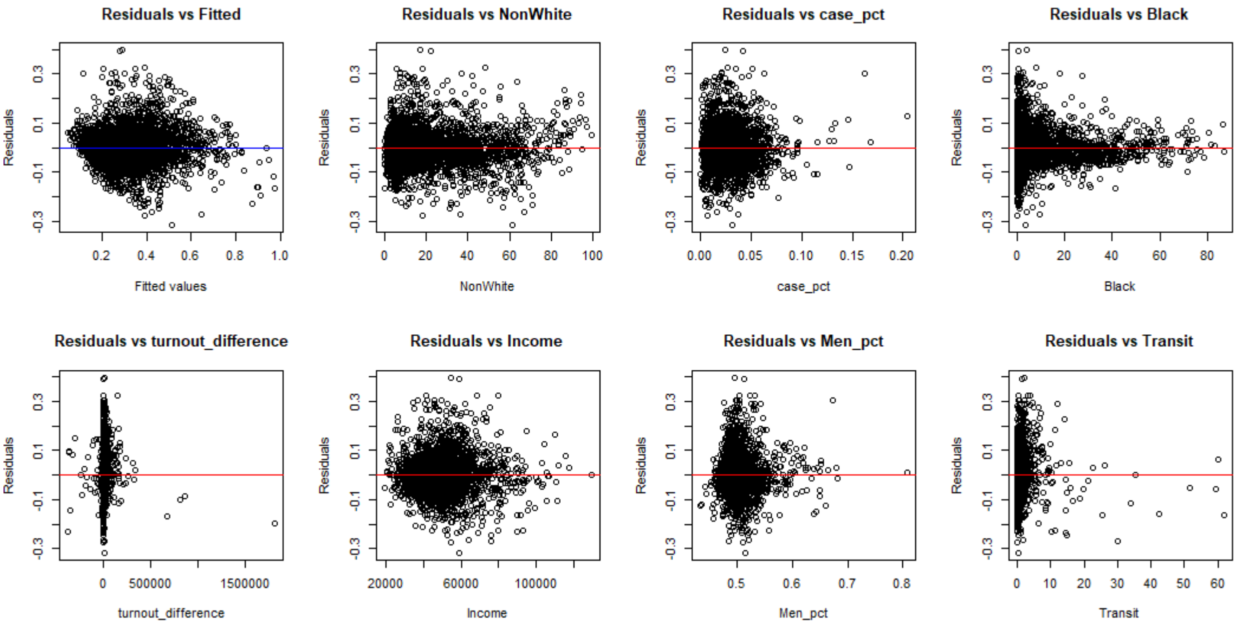

Black
1
-0.61
0.62
0.28
0.25
White
-0.61
1
-1
-0.32
-0.24
NonWhite
0.62
-1
1
0.33
0.25
case_pct
0.28
-0.32
0.33
1
-0.01
death_pct
0.25
-0.24
0.25
-0.01
1
Black
White
NonWhite
case_pct
death_pct
1.0
0.5
0.0
-0.5
-1.0



	percentage20_Joe_Biden	Income	Professional	Service	Office	Construction	Production	Employed	PrivateWork	PublicWork	SelfEmployed	FamilyWork	Unemployment
percentage20_Joe_Biden	1	0.21	0.32	0.27	0.19	-0.5	-0.28	0.37	0.12	0.08	-0.33	-0.15	0.26
Income	0.21	1	0.62	-0.34	0.11	-0.25	-0.36	0.27	0.23	-0.24	-0.05	-0.08	-0.45
Professional	0.32	0.62	1	-0.29	0	-0.39	-0.66	0.27	-0.15	0.08	0.15	0.05	-0.31
Service	0.27	-0.34	-0.29	1	-0.03	-0.11	-0.22	-0.01	-0.19	0.34	-0.15	-0.05	0.34
Office	0.19	0.11	0	-0.03	1	-0.39	-0.23	0.18	0.31	-0.16	-0.34	-0.18	0.11
Construction	-0.5	-0.25	-0.39	-0.11	-0.39	1	0	-0.27	-0.31	0.12	0.38	0.22	-0.06
Production	-0.28	-0.36	-0.66	-0.22	-0.23	0	1	-0.2	0.35	-0.31	-0.17	-0.09	0.12
Employed	0.37	0.27	0.27	-0.01	0.18	-0.27	-0.2	1	0.2	-0.15	-0.14	-0.08	0.02
PrivateWork	0.12	0.23	-0.15	-0.19	0.31	-0.31	0.35	0.2	1	-0.83	-0.58	-0.31	-0.08
PublicWork	0.08	-0.24	0.08	0.34	-0.16	0.12	-0.31	-0.15	-0.83	1	0.04	0.07	0.3
SelfEmployed	-0.33	-0.05	0.15	-0.15	-0.34	0.38	-0.17	-0.14	-0.58	0.04	1	0.35	-0.3
FamilyWork	-0.15	-0.08	0.05	-0.05	-0.18	0.22	-0.09	-0.08	-0.31	0.07	0.35	1	-0.1
Unemployment	0.26	-0.45	-0.31	0.34	0.11	-0.06	0.12	0.02	-0.08	0.3	-0.3	-0.1	1

Appendix 3: Heteroscedasticity in Model

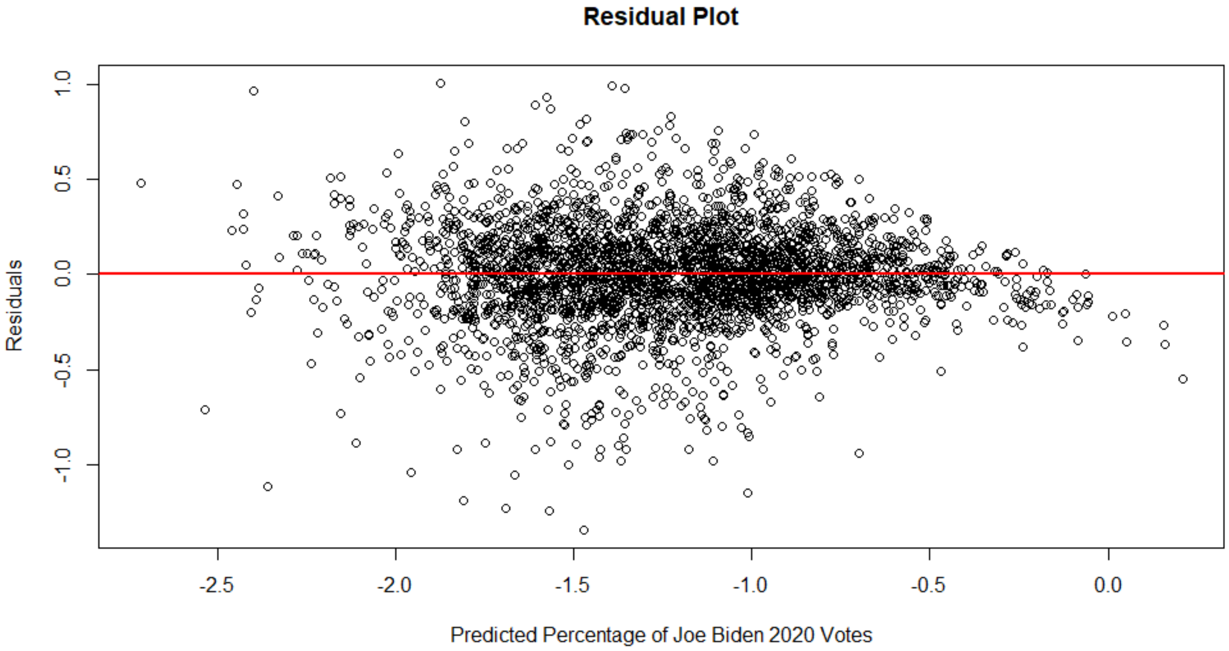
Heteroscedasticity looks present in multiple explanatory variables, including NonWhite, Black, turnout_difference, and Transit.



Appendix 4: Log Transformed Model

We take the log of the response variable to mitigate above heteroscedasticity. The new model has R^2 of 0.70 and the same 107 parameters.

log(percentage20_Joe_Biden) ~ factor(state)*NonWhite + case_pct*Black + turnout_difference + Income + Men_pct + Transit


Residual Plot
Residuals
Predicted Percentage of Joe Biden 2020 Votes