Is political affiliation a significant explanatory variable in predicting per-county COVID-19 case rates in the United States?

Abstract

It is safe to say that there is currently no public health issue more pressing than the COVID-19 pandemic, and understanding the factors that affect the spread of the disease can save lives. In the United States, the response to the pandemic has been heavily politicized with citizens facing conflicting sources of information about the virus. In this work, we evaluate the effect of this politicization through statistical analysis of two different periods of time; six months before the vaccine was available, and six months after the vaccine was available. For each period, we first use the best subsets variable selection technique to inform our construction of a parsimonious reduced linear regression model that predicts COVID-19 case rates using demographic data at the county level. We then add the variable of interest—the percentage of total votes in a county that went to Joe Biden—to create our full model. Next, we compare the difference between the two models using an Extra Sum of Squares test. Our results show that in both periods, the percentage of votes for Joe Biden is a significant explanatory variable in predicting COVID-19 case rates.

Background

The COVID-19 pandemic has been devastating throughout the USA and the rest of the world. According to Johns Hopkins University's COVID-19 tracker, by 8th December 2021, a total of 49.5 million people in the US had tested positive for COVID-19 at some point since the first case was identified in January 2020, and 793,000 people had succumbed to the disease. With the increasing polarization of the American political climate, we believe that it may be interesting to look at how a county's political affiliation affects its COVID-19 case rates.

COVID-19 has now been around for almost two years, and in this time, a large proportion of the American population has received vaccines. It is not hard to see that vaccination rates heavily influence per-county COVID-19 case rates, with counties getting the vaccine earlier and faster seeing significant decreases in the number of positive cases. In order to control for this, we have split our timeline into two distinct periods: from November 1, 2020 to February 28, 2021, and from May 1, 2021 to October 31, 2021 in order to show the effect. Our choice of split is explained by the fact that COVID-19's surge towards its first peak in the US began in November 2020, and started trickling down towards the end of February, which marks the beginning of the vaccine drive. For our analyses, we look at cumulative cases in each of these periods—we will refer to them as Period 1 and Period 2, respectively. We consider the months of March and April to constitute a 'transition period' where the rate of vaccination was the highest, the end of which marks an inflection point after which the rate of vaccination declines substantially: as of May 22, 2021, 57% of adults had received at least 1 dose of the vaccine, as reported by CDC's COVID-19 Vaccination Coverage Report. Eliminating this two month period of rapid change allows us to better estimate the impact of the vaccine in absence of more comprehensive data.

Building on previous research such as Kim, et al. (2020), we hypothesize that political affiliation is a significant explanatory variable in predicting per-county COVID-19 case rates.

Methods

Data Preparation

The dataset used in our analysis was created from three separate datasets. To obtain county-level demographic and economic data, we used the 2017 American Community Survey. To find COVID-19 case data, we used county-level data from the New York Times Covid-19 data Github repository. Finally for election data, we used the County Presidential Election Returns from the MIT Election Data and Science Lab (MEDSL). A significant amount of data manipulation was required to extract cumulative COVID-19 case rates for each of the two periods. We then used Federal Information Processing System(FIPS) codes, which are unique numerical identifiers for counties, to successfully merge these three datasets, using the merge function in RStudio. The resulting dataset includes key county statistics for the 2020 presidential election, COVID cases and deaths, and demographic and economic figures. The final dataset (before variable creation) has 3106 observations on 48 variables.

Variables

The primary response variable in our study for each of the two periods is the COVID-19 case rate (cumulative cases divided by county population) for that particular period. Our explanatory variable of interest is the number of votes for Joe Biden as a percentage of total votes cast in a county (hereinafter referred to as 'PercentageBiden'). We consider this to be our measure of a county's political affiliation since total votes almost completely consist of votes to Joe Biden or Donald Trump, with a negligible amount of votes going to third-party candidates. The remaining explanatory variables are 25+ county-level demographic indicators, including but not limited to per-capita income, voter population, and distribution of the population by race (See Appendix for more details). We created several new variables using one or more of these demographic indicators, such as proportion of total population eligible to vote and employment rate, for use in our analysis. We plotted the distributions of each of our explanatory variables in order to check for normality. A large number of variables were heavily skewed (to be expected of country-wide demographic data) but we accounted for this using non-linear transformations wherever required.

Statistical Analysis

To test our hypothesis, we first looked at the data for the first period and used best subsets to inform our selection of demographic variables in creating a parsimonious reduced model to predict COVID-19 case rates(See Appendix A). We created a matrix correlation plot (See Appendix B) to check for possible multicollinearity of variables, and excluded variables with a high degree of correlation from our model in order to find the most parsimonious model possible. Additionally, we tested for possible interactions between variables and included the significant interaction effects in our model (for example, the interaction between income per capita and the proportion of the population that was Black). We also looked closely at the normal probability plot of residuals, the residuals vs fitted values plot, and the residuals vs each explanatory variable. Based on these plots, we chose to include some of the explanatory variables in logarithmic specifications if they exhibited a strong rightward skew.

Next, we created the full model by adding our explanatory variable of interest (PercentageBiden) to the reduced model. Finally, we conducted an Extra Sum of Squares F-test to determine whether the addition of PercentageBiden led to a significant increase in the R^2 of the model, which would indicate a strong contribution of this variable to the fit of the regression model. We repeated this entire process for the data from Period 2.

Results

For Period 1, our reduced model was statistically significant at the 0.0001 level with an adjusted R^2 of 0.18. This low value for the adjusted R^2 was not surprising, since our data does not include explanatory variables such as comorbidities, and mask compliance rates which are important in explaining a large amount of the variation in COVID-19 case rates. The final full model yielded an adjusted R^2 of 0.21. The increase in the adjusted R^2 upon the addition of our variable of interest (PercentageBiden), suggests that it is a statistically significant variable in predicting COVID-19 case rates. This was solidified by our Extra Sum of Squares test (See

Appendix C) for the full model versus the reduced model, which yielded an extremely small p-value less than 0.00001. The coefficient of PercentageBiden in this model was negative, indicating that an increase in percentage of votes for Joe Biden is correlated to a decrease in the COVID-19 case rates for a county.

For Period 2, our reduced model was once again statistically significant at the 0.01 level with an adjusted R^2 of 0.26. The full model yielded an adjusted R^2 of 0.31, this time the difference in the two models being even higher than in Period 1. The Extra Sum of Squares test for Period 2 (See Appendix C) once again returned an extremely small p-value less than 0.00001. The coefficient of PercentageBiden in this model was also negative, indicating a similar correlation to Period 1.

Discussion

Our results show that in both periods, the variable PercentageBiden provides a significant contribution to the regression model that is highly unlikely to be caused by random chance. The model for Period 1 shows a notable jump in the model R^2 upon the addition of PercentageBiden, indicating that there is some association between a county's political affiliation and their COVID-19 case rates. In addition, the increase in R^2 after the addition of PercentageBiden to the Period 2 model is greater than what we see in Period 1.

Since the primary difference between the two periods—by design—is the increase in vaccination rates, our results also hint towards a possible relationship between a county's political affiliation and their vaccination rates. However, given that the coefficients for PercentageBiden depend upon the other variables in the model, we need to interpret the model with caution. That being said, both models have negative coefficients, indicating that counties with a higher percentage of votes for Biden tend to have lower Covid-19 case rates. We believe this would be a very interesting topic for future research.

Some potential sources of bias in our study are the lack of data on comorbidities, mask compliance, and more extensive vaccination statistics. In addition, because this is an observational study and not a randomized experiment, we can only establish correlation and not causation. However, based on our findings, we believe that it is safe to conclude that within the United States, political affiliation is indeed a significant explanatory variable in predicting per-county COVID-19 case rates, with counties casting a higher percentage of votes for Joe Biden seeing comparatively lower case rates.

References

COVID-19 Dashboard. Center for Systems Science and Engineering (CSSE), Johns Hopkins University (JHU). (Accessed December 8, 2021). https://coronavirus.jhu.edu/map.html.

COVID-19 vaccine doses administered. Our World in Data. (Accessed December 8, 2021). https://ourworldindata.org/grapher/cumulative-covid-vaccinations?country=~USA.

Diesel J, Sterrett N, Dasgupta S, et al. *COVID-19 Vaccination Coverage Among Adults — United States, December 14, 2020–May 22, 2021*. MMWR Morb Mortal Wkly Rep 2021;70: 922–927. DOI: http://dx.doi.org/10.15585/mmwr.mm7025e1.

Kim, et. al. *Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States.* Proceedings of the National Academy of Sciences. (September, 2020). https://doi.org/10.1073/pnas.2007835117.

Appendix A

Best Subsets Informed Reduced Models

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   -2.859e+00  2.881e-01  -9.923  < 2e-16 ***
IncomePerCap                  -1.375e-05  2.795e-06  -4.918 9.20e-07 ***
log(Black + 1)                -3.091e-02  9.995e-02  -0.309  0.75717    
Voters_by_pop                 -1.325e+00  2.277e-01  -5.820 6.50e-09 ***
log(Asian + 1)                -9.441e-02  1.596e-02  -5.914 3.71e-09 ***
White                          3.301e-03  6.550e-04   5.039 4.95e-07 ***
log(Hispanic + 1)              5.413e-02  1.052e-02   5.144 2.86e-07 ***
log(Native + 1)                5.957e-02  1.093e-02   5.448 5.49e-08 ***
Production                     8.105e-03  1.410e-03   5.748 9.94e-09 ***
Drive                          8.586e-03  1.816e-03   4.727 2.38e-06 ***
Carpool                       -5.703e-03  2.775e-03  -2.055  0.03992 *  
MeanCommute                   -8.681e-03  2.115e-03  -4.104 4.17e-05 ***
log(Walk + 1)                 -4.953e-02  1.999e-02  -2.478  0.01326 *  
employment_rate                1.407e+00  2.479e-01   5.675 1.51e-08 ***
PrivateWork                   -9.569e-04  1.371e-03  -0.698  0.48539    
SelfEmployed                  -5.471e-03  2.439e-03  -2.243  0.02496 *  
IncomePerCap:log(Black + 1)    5.401e-06  1.481e-06   3.647  0.00027 ***
log(Black + 1):Voters_by_pop   1.462e-01  1.133e-01   1.290  0.19715    
log(Black + 1):MeanCommute    -1.528e-03  1.023e-03  -1.494  0.13535    
log(Black + 1):employment_rate -3.441e-01  1.316e-01  -2.614  0.00899 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3308 on 3086 degrees of freedom
Multiple R-squared:  0.1848, Adjusted R-squared:  0.1798 
F-statistic: 36.82 on 19 and 3086 DF,  p-value: < 2.2e-16
```

Figure A1: Period 1 Reduced Model.

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.4481676  0.4764070   0.941 0.346920
Voters_by_pop                 0.5433563  0.1451514   3.743 0.000185 ***
log(IncomePerCap)            -0.3721983  0.0500448  -7.437 1.32e-13 ***
Construction                  0.0053056  0.0019889   2.668 0.007679 **
Drive                         0.0038222  0.0015036   2.542 0.011072 *
log(Walk + 1)                -0.2526823  0.0729350  -3.464 0.000538 ***
employment_rate              -1.6374882  0.3058027  -5.355 9.20e-08 ***
PrivateWork                   0.0057045  0.0013738   4.152 3.38e-05 ***
log(Black + 1)               -0.2340563  0.0602616  -3.884 0.000105 ***
log(Asian + 1)               -0.0574289  0.0164769  -3.485 0.000498 ***
log(Walk + 1):employment_rate  0.3011826  0.1650574   1.825 0.068141 .
PrivateWork:log(Black + 1)    0.0027175  0.0007969   3.410 0.000658 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3448 on 3094 degrees of freedom
Multiple R-squared:  0.2611, Adjusted R-squared:  0.2585
F-statistic: 99.39 on 11 and 3094 DF,  p-value: < 2.2e-16
```

Figure A2: Best Subsets Plot for Period 2 Reduced Model.

Appendix B

Matrix Correlation Plot for explanatory variables

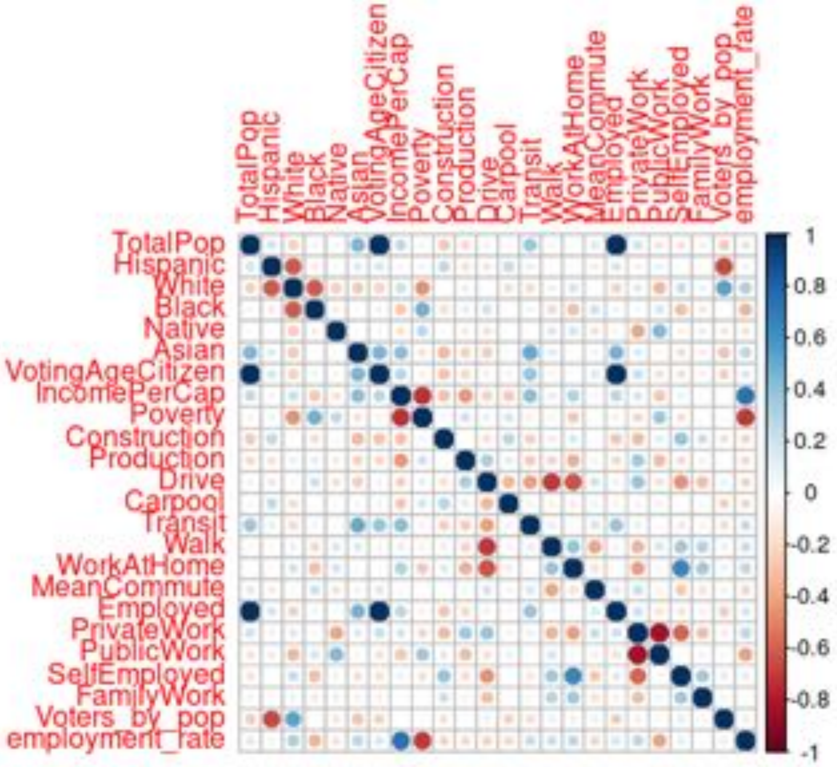
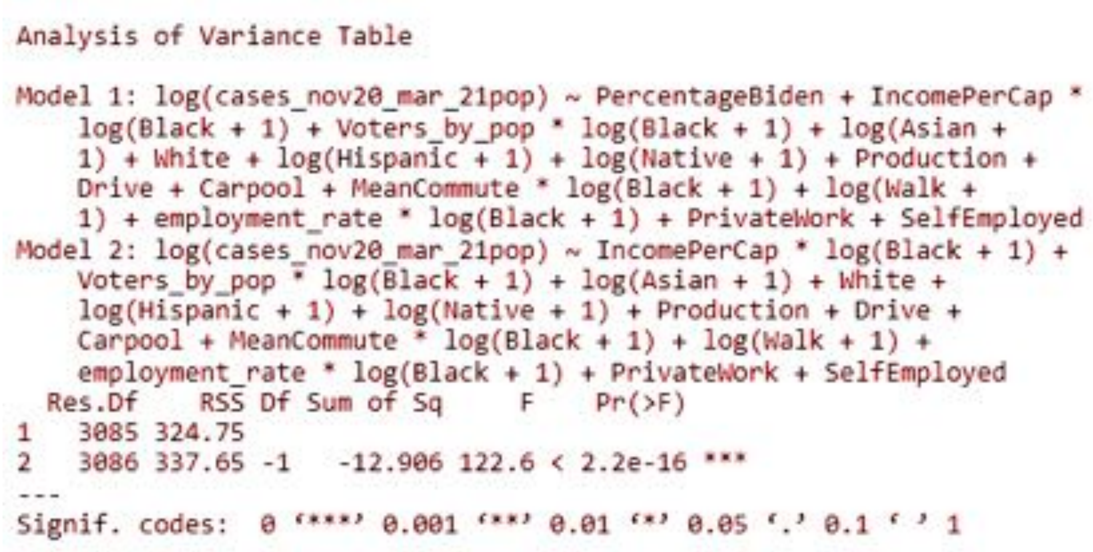


Figure B1: Matrix Correlation Plot for quantitative explanatory variables.

Appendix C

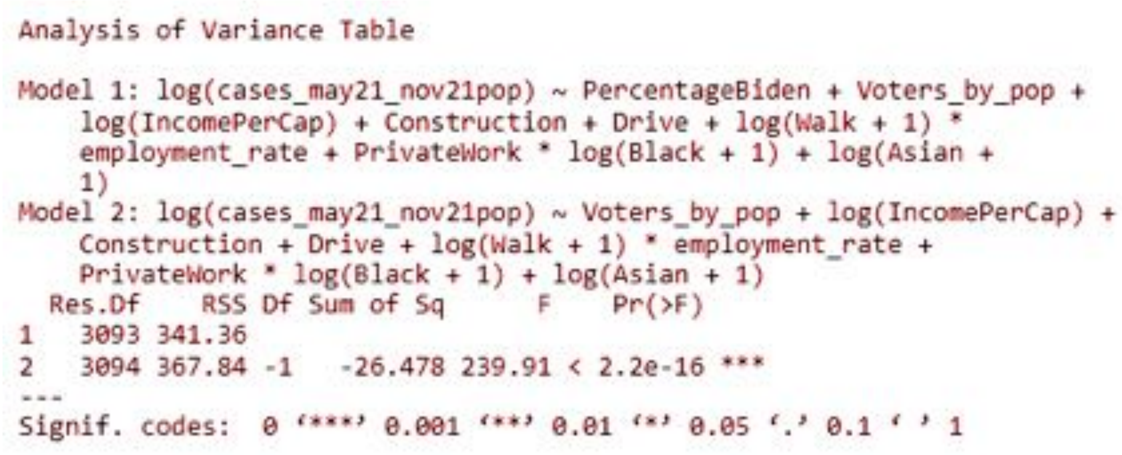
Extra Sum of Squares



```
Analysis of Variance Table

Model 1: log(cases_nov20_mar_21pop) ~ PercentageBiden + IncomePerCap *
    log(Black + 1) + Voters_by_pop * log(Black + 1) + log(Asian +
    1) + White + log(Hispanic + 1) + log(Native + 1) + Production +
    Drive + Carpool + MeanCommute * log(Black + 1) + log(Walk +
    1) + employment_rate * log(Black + 1) + PrivateWork + SelfEmployed
Model 2: log(cases_nov20_mar_21pop) ~ IncomePerCap * log(Black + 1) +
    Voters_by_pop * log(Black + 1) + log(Asian + 1) + White +
    log(Hispanic + 1) + log(Native + 1) + Production + Drive +
    Carpool + MeanCommute * log(Black + 1) + log(Walk + 1) +
    employment_rate * log(Black + 1) + PrivateWork + SelfEmployed
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   3085 324.75
2   3086 337.65 -1   -12.906 122.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure C1: Extra Sum of Square Test for full model vs reduced model, Period 1.



```
Analysis of Variance Table

Model 1: log(cases_may21_nov21pop) ~ PercentageBiden + Voters_by_pop +
    log(IncomePerCap) + Construction + Drive + log(Walk + 1) *
    employment_rate + PrivateWork * log(Black + 1) + log(Asian +
    1)
Model 2: log(cases_may21_nov21pop) ~ Voters_by_pop + log(IncomePerCap) +
    Construction + Drive + log(Walk + 1) * employment_rate +
    PrivateWork * log(Black + 1) + log(Asian + 1)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3093 341.36
2   3094 367.84 -1   -26.478 239.91 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure C2: Extra Sum of Square Test for full model vs reduced model, Period 2.

Appendix D

Normal Probability Plots and Residual vs Fit Plots

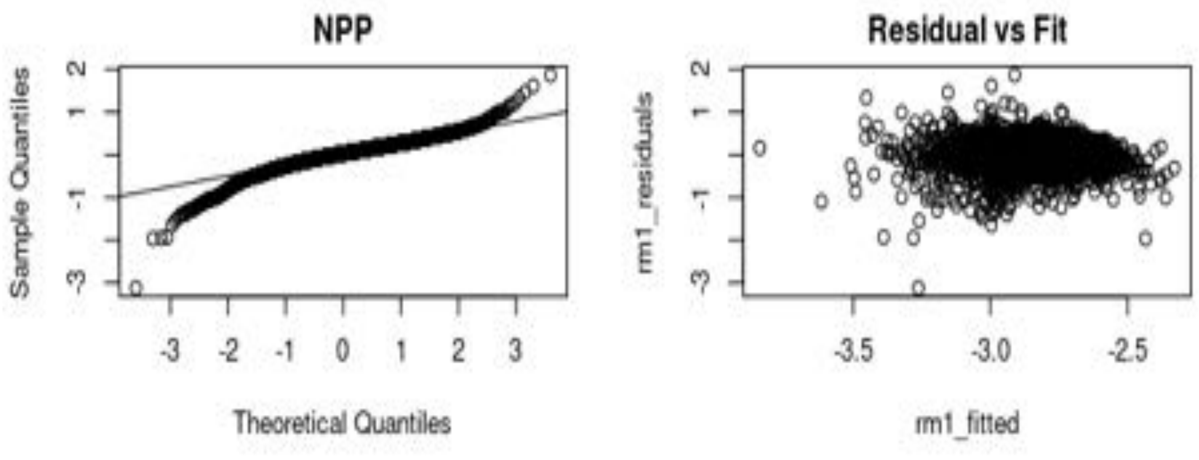


Figure D1: Normal Probability Plot and Residual vs Fit Plot, Period 1.

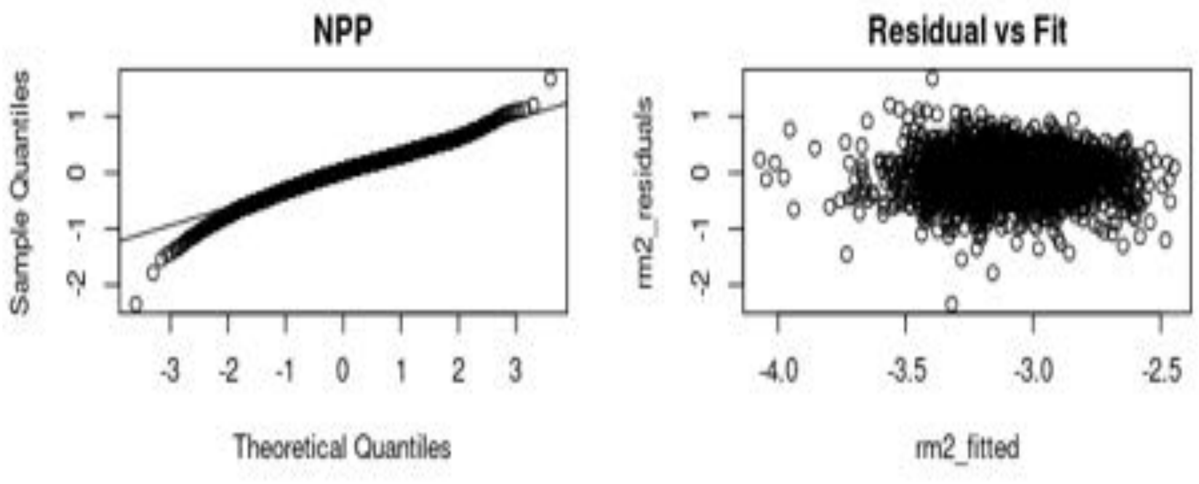


Figure D2: Normal Probability Plot and Residual vs Fit Plot, Period 2.