Political Affiliation and it's Association with Mask Wearing Habits

Demian Rivera, Samuel Gomez, Juan Ramirez

August 5th, 2020

Contents

1 Introduction

In this research paper, we study whether political affiliation affects the percentage of the U.S. population wearing masks. Our question of interest is the following:

Does political affiliation affect population behaviors related to wearing face protection/masks during the COVID-19 pandemic?

Answering this question is pivotal, given the current situation of the COVID-19 pandemic. As economies begin to reopen, more and more local businesses start to welcome higher volumes of customers. This increase in community interaction will undoubtingly put citizens more at risk of coming into contact with others infected by COVID-19. A casual understanding of the political leaning impact on mask-wearing can aid state and local officials in designing and implementing safer practices for citizens.

Given our results from following the well studied and agreed upon causal effect of wearing masks on the decrease of COVID-19 spread (if done well and as recommended by experts), we propose that political affiliation has a statistical significance in predicting the population's mask-wearing behavior.

It is important to note that in this paper, the sum of republican, democrat, and no preference or other equal the total (100%). We have divided independents according to their political leaning into the beforementioned republican and democrat groups. We have not used the subgroup that has identified themselves as having no preference or other. We found that running our analysis, whether focused on democrat or focused on republican, produced very similar statistical significance and practicality (the coefficient signs are opposite). For ease of interpretability, we concentrate on Republican party affiliation, which includes citizens who identify Republican and independents with republican leaning throughout the rest of this paper. Our dataset also consists of the democrat variable (including democrat leaning) if the reader wants to study further the effect of the rest of the political leaning spectrum.

2 Exploratory Data Analysis

2.1 Data Cleaning

We found a number of minor problems with the dataset that needed to be addressed prior to starting our analysis. We first noticed that the dataset had two rows for Arizona. Each had some duplicate data but other fields had different numbers. It was initially unclear what row was correct but noticed cases and deaths per 100k was the same. We therefore calculated new total deaths and total cases from the population provided. We also corrected deaths per 100,000 for the states of Colorado and New York as mentioned in our EDA section below.

Major takeaways from our EDA:

- There are a handful of states that did not mandate wearing face masks during the timeframe the data was sampled–July 6th, 2020. The peak at zero days in the bimodal `maskdays` histogram represents this observation. States without mask-wearing mandates include Idaho, Iowa, Missouri, Montana, Oklahoma, South Carolina, South Dakota, Tennessee, and Wisconsin. All of these states strongly supported the Republican party in the 2016 election, with the exception of Wisconsin, which just marginally supported the Republican party in the 2016 election.

- The number of days since a mask-wearing decree was implemented (`maskdays`) and the percent of the population wearing masks (`maskwear`) has a positive correlation of 0.645. The positive correlation between the two variables implies that as the number of days since a mask-wearing mandate increases, so does the percentage of Americans wearing masks.

- There is a strong negative correlation of -0.671 between the proportion of republican political affiliation (`repu`) for each state and the percentage of the state's population wearing masks (`maskwear`). This negative correlation implies that as republican party affiliation increases, the percent of the population wearing masks decreases.

- A weak negative correlation of -0.384 exists between political affiliation (`repu`) and the number of days since wearing masks was mandated (`maskdays`). Perhaps, this can be explained by COVID-19 arriving later in more conservative states and/or more conservative states responding slower with mask-wearing implementation.

- There is a weak negative correlation between deaths per 100k (`d100k`) and political party affiliation (`repu`). Presumably, republican states are less populated and tend to be rural hence not transmitting as easily compared to higher populated areas with more human interaction.

- We noticed that both New York and Colorado state deaths per 100k seemed to be suspiciously high. We calculated the death rate per 100k people and found that New York's death rate is 163 instead of the original 354 given in our dataset; likewise, Colorado experienced 32 deaths per 100k, not 299.

- We debated about keeping the District of Columbia in the dataset given its differences to most other states. DC has the lowest reported amount of republicans or leaning republican with only 11% and the sixth-highest death rate. We decided to keep DC as it provides valuable information about an important area in our country.

- In Figure 2, a deeper dive into the death rate shows us that the highest death rates are mainly on the east coast. The top six states are New Jersey, New York, Connecticut, Massachusetts, Rhode Island, and the District of Columbia.
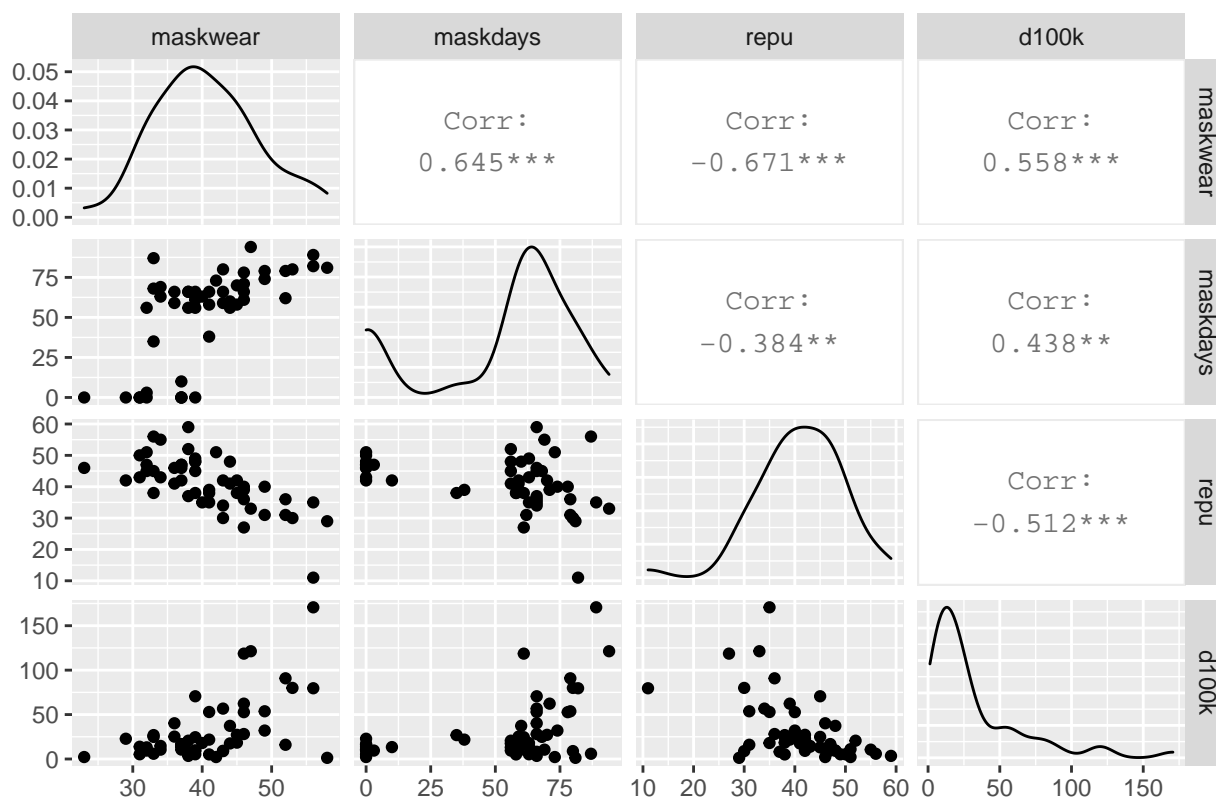
Optimal Model variable relationships

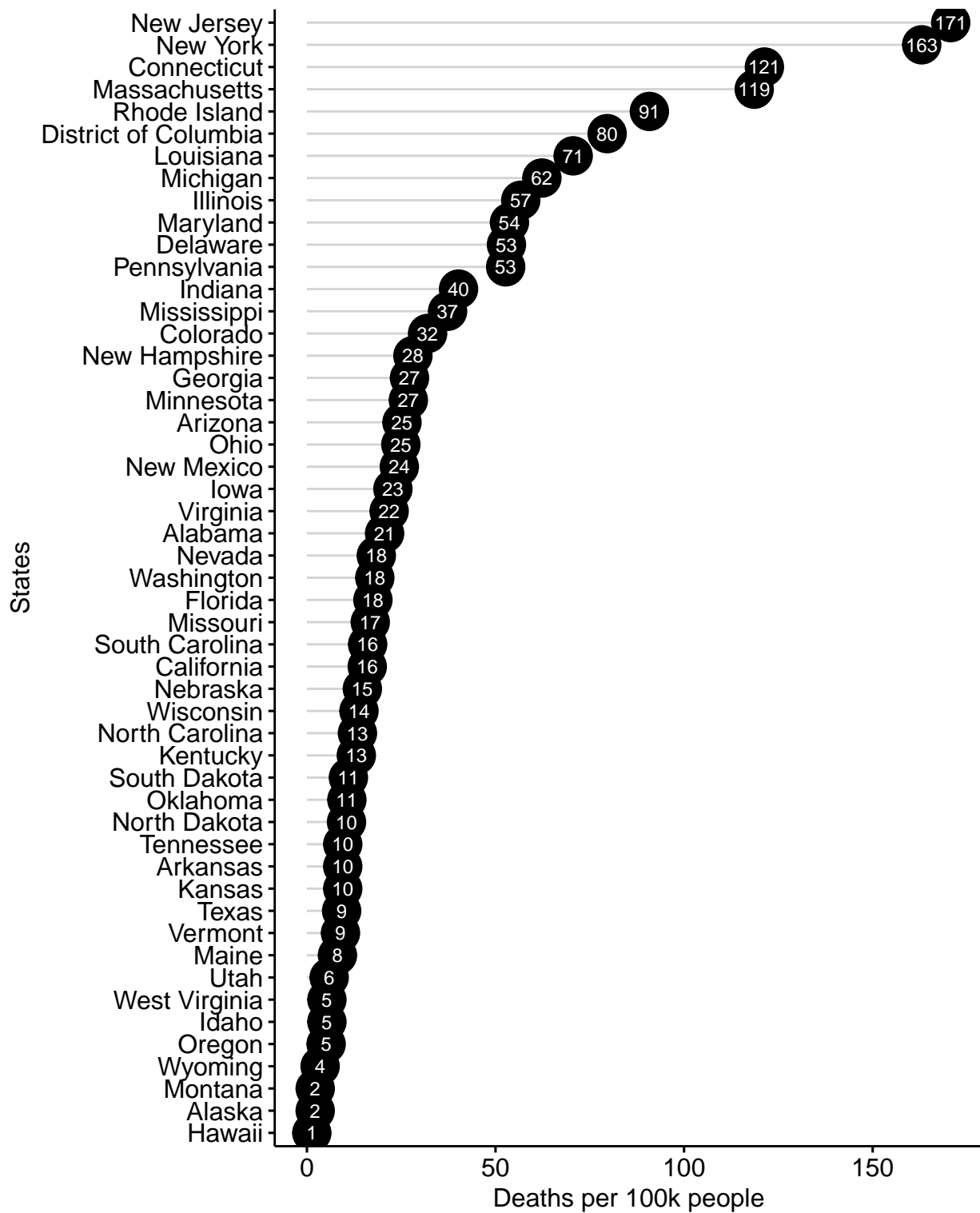Figure 1: Correlation plot for variables of interest

Figure 2: Highest to lowest states based on deaths per 100k

2.2 Data Transformation

Although most of our variables were unaltered or transformed in any major way, we saw the need to add new variables based on the dates provided. For example, the calculation for `masksdays` starts with the date the dataset was pulled (July 6th, 2020) and detracts the date of the mask-wearing mandate for public-facing employees. As a result, we can use the net number of days the policy was in effect (as of the dataset date). In addition, we transformed decimal data points (reported as percentages) to integers by multiplying by 100. This did not change any of our statistics but rather improved interpretability.

3 Models

3.1 Our model

We have arrived at three appropriate models that help us answer the research question, and further understand some of the possible interactions between variables.

Given the scale of our dependent variable, we are focused on explaining a sizable portion of variability and not the variation in its entirety.

This paper is a starting point. We propose a future more in-depth study on the subject and the potential applicability of these insights and statistics on other related policies with respect to COVID-19 and other widespread airborne diseases.

Warning on data/analysis

It is imperative to conduct a more in-depth and rigorous study based on panel data in order to size the impact of policies on behavior appropriately. Panel and time-series data starting in 2018 would equip us to have better causal evidence and effects of policies over time. It would also bring into light evidence of time lag between policy implementation and impact.

We are interested in measuring the effect political affiliation has on the state-level population's response to mandates and ordinances related to wearing masks by public-facing employees. Although we present three models for this purpose, we propose to focus on Model2 as it best combines parsimony, validity, and interpretation.

3.2 Main dependent variable for all three models

maskwear is our dependent variable. Although the mandates are imposed on businesses, there is an undeniable causal effect on the general population. We care to measure this effect on the regular population as we believe it is, in fact, of a much higher impact on the contagion rate of COVID-19 than the decree or mandate on businesses alone. Therefore, we created a new variable using results from a question that was part of a survey conducted by YouGov in June of 2020. The survey question was: "In the past seven days, how often have you worn a mask on your face when outside your home?" the responses were then categorized and turned into percentages. It is important to emphasize this variable is a result of a large sample and is not population data, as is the original dataset.

More on the survey: The Yahoo! News survey was conducted by YouGov using a nationally representative sample of 1,507 U.S. adult residents interviewed online between June 24-25, 2020. This sample was weighted according to gender, age, race, and education based on the American Community Survey, conducted by the U.S. Bureau of the Census, as well as the 2016 Presidential vote, registration status, geographic region, and news interest. Respondents were selected from YouGov's opt-in panel to be representative of all U.S. residents. The margin of error is approximately 3.6%.

3.3 Model 1: Base Model (bivariate only)

We started with a bivariate regression model using only the percent of the population that is republican or republican leaning.

vo: Percent of people wearing masks $= \beta_0 + \beta_1$(percent of population that is republican/leaning republican) + u

1) Percent of population that is republican: The percent of a state's population that is either affiliated with the republican party or is independent and leans republican. The variable is provided by a 2018 Gallup poll. Washington DC was not available in the state breakout from Gallup, so we used another reputable source's (Pew Research) number to complete the dataset.

Base Model Summary Statistics:

```
t test of coefficients:


            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 65.005424   3.685576 17.6378 < 2.2e-16 ***
repu        -0.591930   0.087384 -6.7739 1.486e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "R-squared: 0.451, Adjusted R-squared: 0.44, AIC: 324.6"
```

By interpreting our summary statistics result, we can see a high statistical significance of the overall model (p-value well below 0.001). We might also interpret the Adjusted R^2 by saying that 0.59% of the variability in the percent of people wearing masks can be attributed to the percentage change in political leaning. As is typical with human behavioral statistics, there are most likely other underlying variables not included in this model that address other facets of the reason behind people wearing masks. Model 2 addresses two additional variables to try to further aid in explaining this.

3.4 Model 2: Best interpretability and parsimonious model

v1: Percent of people wearing masks $= \beta_0 + \beta_1$(percent of population that is republican/leaning republican) $+ \beta_2$(net days since mask mandate) $+ \beta_3$ (deaths per 100,000 people) + u

In addition to political leaning, we now control for days mask-wearing mandates have been in place as well as deaths per 100,000 people.

2) Net days since mask mandate: We intuitively include this is a regressor. Even though the mask-wearing mandates were only placed on businesses and only for public-facing employees, it is reasonable to assume these mandates would also have an effect on non-government driven policies and habits at the household level.
We use net days from the day each state declared a mandate on wearing masks for public-facing employees up to 7/6/20. Some states did not declare a mandate and published a state-wide recommendation instead. Other states did neither. For the purposes of this analyis we assume that states which did not formally mandate mask-wearing have a total number of days equal to zero.

3) Deaths per 100,000: Generally, we would not include an outcome variable as a regressor, even if the outcome is not what we are measuring. The logic behind that is that if wearing masks decreases death, then we would have a problem of endogeneity. We use the variable in two different ways here: We do not use it as a measure of death (or outcome) but as a measure of fear. We believe that as the death percent gets higher, more people are likely to know someone who has died. These events would incentivize people to wear masks more often. Including deaths per 100,000 allows us to hold that variable constant so we can better focus on the true impact of political leaning.

Model 2 Summary Statistics:

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 49.482293   5.119970  9.6646 9.52e-13 ***
maskdays     0.103363   0.028858  3.5817 0.0008069 ***
repu        -0.374629   0.102379 -3.6592 0.0006387 ***
d100k        0.033912   0.026042  1.3022 0.1991964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] "R-squared: 0.647, Adjusted R-squared: 0.625, AIC: 306"

[1] "VIF:"

```
maskdays     repu     d100k
1.285619 1.423862 1.493390
```

The inclusion of the two additional regressors, days since the mask-wearing mandates and deaths per 100,000, makes our model stronger. Both `maskdays` and `repu` have very high statistical significance, and also improves our Adjusted R^2, explaining ~64% of the variability in percent of people wearing masks. The VIF is low for all variables, so there is no concern about collinearity.

It is also important to note that the associative change in political leaning on the independent variable is much lower than in the original model. This is evidence that political leaning may be hiding the effects of other potential interactions or variables. We can now say that a 1% change in political leaning has a negatively correlated effect on population wearing masks of 0.37%, ceteris paribus.

3.5 Model 3: Sacrifice in statistical significance and parsimony for a minimal increase in variance explainability.

v2: Percent of people wearing masks $= \beta_0 + \beta_1$(net days since mask mandate) $+ \beta_2$ (percent of population that is republican/leaning republican) $+ \beta_3$(deaths per 100k) $+ \beta_4$(percent of population living in an apartment) $+$ u

To further explain variation in our dependent variable maskwear, we add a fourth variable percent_apartment.

4) Percent of people living in apartments: The percentage of people living in apartments is another variable added to the original dataset. The data comes from a survey conducted by the American Community Survey–it is regarded as a premier source for detailed U.S. population and housing information. We are incorporating this variable into Model 3 because we want to understand the population density of our samples more precisely. Although the original dataset provides population density data, we felt this variable was more accurate. The population density data included in the dataset did not account for uninhabitable land and, therefore, seems to be an ineffective representation of densely populated living communities. Our thought process led us to believe that more densely populated areas have a higher percentage of the people living in apartments than in less densely populated areas.

Model 3 Summary Statistics:

```
t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)      45.125265   6.733602  6.7015 2.558e-08 ***
maskdays          0.101017   0.029047  3.4777  0.001116 **
d100k             0.031769   0.026175  1.2137  0.231044
repu             -0.304635   0.125911 -2.4194  0.019555 *
percent_apartment 0.141009   0.147379  0.9568  0.343684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] "R-squared: 0.656, Adjusted R-squared: 0.626, AIC: 306.8"

[1] "VIF:"

        maskdays           d100k              repu percent_apartment
        1.295458        1.507838          2.157137          1.939212
```

The net mandate mask days (`maskdays`) stays consistently statistically significant. We expected this outcome because the longer a mandate has been in place, the more likely individuals are to wear masks. Adding the percent of people living in apartments regressor to the model causes the statistical significance of the republican affiliation to become less significant. This can be observed due to the change in the t-value–from -3.65 in Model 2 to -2.42 in this model. Since the `repu` regressor does remain statistically significant, our claim that it is a good predictor for the percent of the population wearing masks is reinforced. The added regressor `percent_apartment` is not statistically significant and, therefore, does not improve the prediction of the dependent variable.

The adjusted R^2 value of this model is slightly higher than Model 2. Although a marginal increase of variability is explained by Model 3, it comes with a sacrifice in parsimony. The slight increase in R^2 does not justify abandoning the simplicity of the previous model. Furthermore, the lack of statistical significance in some of the variables does not contribute to a favorable model. In conclusion, we prefer Model 2 over Model 3 because of its simple interpretation, high R^2 value, and direct connection to our research question.

4 Assessment of CLM Assumptions

We will focus on **Model 2** for the remainder of our analysis. This model provides the most accurate representation of the question we are trying to answer and contains variables directly related to our research question.

Model 2

$$maskwear = \beta_0 + \beta_1 maskdays + \beta_2 repub + \beta_3 d100k + u$$
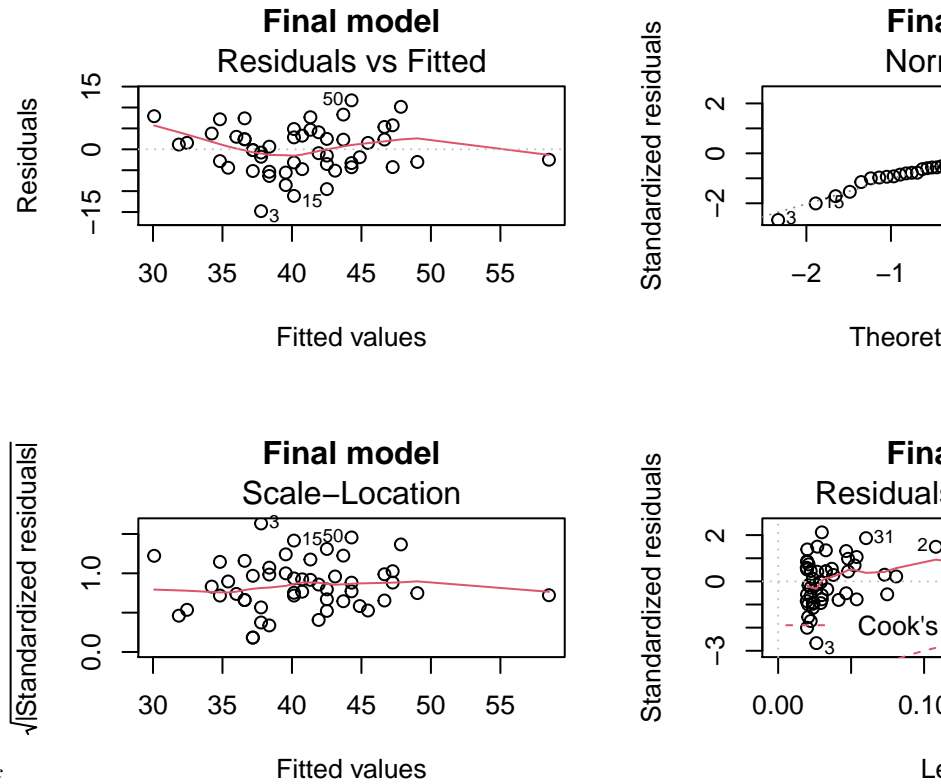
Plot for Model 2-1.pdf Plot for Model 2-1.pdf

4.1 CLM 1 Linear population model

In linear regression, the term "linear" refers to a model being linear in the parameters (beta). This means no parameter appears as an exponent or is multiplied or divided by another parameter. Moreover, a model doesn't need to include linear covariates. For example, individual predictor values can have nonlinear relationships with the response variable. In some models, predictor variables are transformed to obtain the line of best fit. Typical transformations used in linear regression include logs, squares, and sometimes square roots. Based on this requirement, Model 2 is linear in nature and does not include variable transformation. This characteristic is desirable for a more straightforward interpretation of how the dependent variable responds to changes in the regressors/independent variables.

4.2 CLM 2 Random Sampling

To determine random sampling for the overall mode each variable of Model 2 will be examined separately. Model 2 contains four variables collected from three different sources.

1. maskwear: The percent of the U.S. population wearing masks - YouGov survey.

2. repu: The rercent of the U.S. population that is Republican of leaning Republican - Gallup poll.

3. d100k: COVID-19 death rate in the U.S. reported to the CDC by state - CDC COVID-19 data tracker.

4. maskdays: The number of days since each state mandate wearing face masks in public - CDC COVID-19 data tracker.

`maskwear:`

Percent of the population wearing masks comes from a YouGov survey. The survey states that the respondents were selected from the YouGov panel by sample matching. The YouGov panel consists of a large and diverse set of over a million respondents. Each who have volunteered to complete surveys online and who regularly receive invitations to do so. Matching is intended to make the individuals who complete the survey represent the U.S. population. Since this dataset is from opt-in panels (non-probability samples), the samples are not independent in a true sense. But for the purposes of this evaluation, we will assume that the sample matching technique helps mirror the distribution of the sample to that of the U.S. population and therefore assume i.i.d.

`repu:`

The percentage of residents from each state that is Republican or leaning Republican comes from a Gallup survey. The survey was conducted in 2018 and randomly sampled 75,669 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia. Respondents were surveyed via telephone interviews. A minimum quota of 70% cellphone respondents and 30% landline respondents, with an additional minimum quota by time zone within each region, were used to build a sample of the adult population. For results based on the total sample of national adults, the margin of sampling error is ± 1 percentage point at a 95% confidence level. Based on these sampling conditions, we assume the sample is i.i.d.

`d100k:`

The deaths per 100k are from the CDC COVID-19 data tracker report. The rate was calculated by COVID-19 related deaths that were reported to the CDC from each state. The data for deaths is based on a subset of cases where case-level data was available. The CDC validates each case through a confirmation process with each jurisdiction. We assume that the death per 100k is a statistic of each state that meets i.i.d sampling since there is not enough information about the sampling methods to provide evidence otherwise.

`maskdays:`

The amount of days since masks were declared mandatory in each state comes from the CDC COVID Data Tracker. This variable meets the random sampling criteria since it is a statistic representing the entire population for each state. The number of days each state made wearing masks in public mandatory does not fluctuate within individual states; it applies to the whole population of the state. Using each state's statistic to make a statement about the U.S. population is a plausible assumption for i.i.d.

Takeaway

Overall all of Model 2 variables fulfill the random sampling criteria. The CDC did not provide much information related to the sampling method for deaths per 100k, which is the one variable that is most questionable. We assume random sampling for our data. Since there are no signs of clustering, we will proceed with running the regression. As a side note, we would be more considered about random sampling if we were using COVID-19 cases, cases per 100k, etc. since most citizens testing for COVID-19 are doing so because they have reason to believe they were exposed. This type of sample would not meet random sampling criteria.
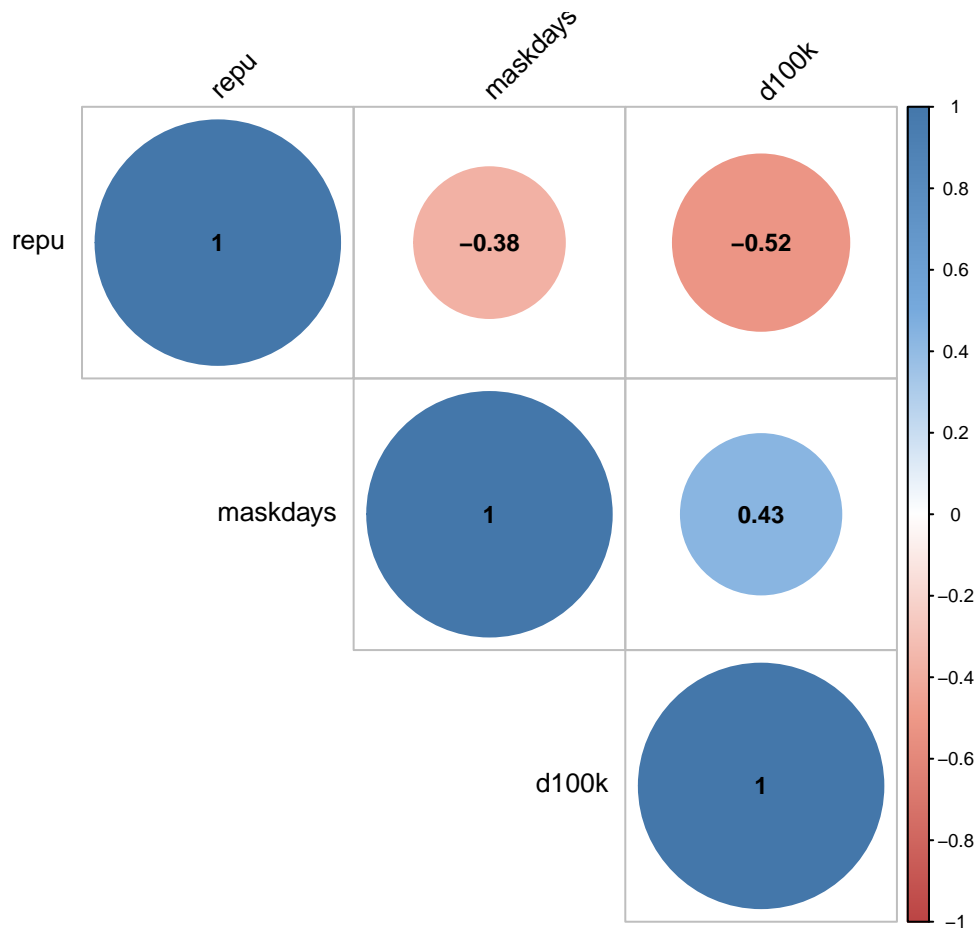
4.3   CLM 3 No perfect multicollinearity

Figure 3: Correlation matrix to understand collinearity

```
## VIF for Model 2:

## maskdays     repu     d100k
## 1.285619 1.423862 1.493390
```

Based on the correlation matrix, this assumption is met for the covariates in Model 2. The result of the VIF (variance inflation factor) test is additional evidence that no perfect multicollinearity is valid. All three independent variables are sufficient to use in the model since there is no evidence of perfect multicollinearity.

4.4   CLM 4 Zero-conditional mean

Zero-conditional mean describes a population where all values of the predictors have an expected zero value error. Typically, this is the hardest assumption to meet for OLS regression. The spline curve on the residuals vs. fitted plot for Model 2 is not completely flat, which is an indication that zero-conditional mean is not met with the data in our model. However, this is an expected outcome after analyzing the `maskdays` variable. The histogram of `maskdays` gives a good visual representation of the data not being from a normal distribution. The outliers in the histogram come from the states that did not issue a mandate to wear masks in public. To fit a more normal distribution, we could have removed those states from the analysis, but in order to make a claim about the U.S. population, we chose to leave those states in the dataset. Removing them would have provided a more normal distribution and, therefore, a model closer to the zero-conditional mean assumption.

One area of concern is towards the right of the graph. There is one data point to the right of 50, which dictates the expectation. Examining the shape of the spline curve does not give sufficient reason to transform any of the variables.

To deal with not meeting this assumption, a weaker assumption of exogeneity can be satisfied. Exogeneity exists when the independent variables are not correlated with the error term. When building a casual model, it is more difficult to assume exogeneity. On the other hand, exogeneity comes with building an associative model. Since two of the three independent variables are practically significant, it can be assumed that these variables are not correlated with the error. Since deaths per 100k is used as a proxy for fear, we would recommend using a more effective measure for this emotion. Also, since there is no causal relationship between the dependent variable and independent variables, we can assume strict exogeneity.

4.5   CLM 5 Homoskedasticity

The residuals vs. leverage plot is a good tool to visually test for homoskedasticity. After analyzing the graph, it appears that the model meets this assumption. Moving from left to right on the graph, the band of vertical variation is mostly uniform. A more formal method of testing for homoskedasticity is running the Breush-Pagan test. The test returns a p-value of 0.91. This implies that there is not enough statistical significance to reject the null hypothesis of homoskedasticity. It is concluded that our model meets this assumption. Nevertheless, using robust standard errors when calculating variable coefficients is a good practice, and it will be used in our analysis. One consideration to keep in mind is that the Bresuch-Pagan test assumes heteroskedasticity is a linear function of the independent variable. It does not rule out a nonlinear relationship between the independent variable and the error variance.

4.6   CLM 6 Normality of Errors

The q-q plot gives detail about the normality of errors. After examining the q-q plot, the normality of errors appears normal. Plotting the histogram of the model's residuals is another visual check that confirms this claim. The statistic from the Shapiro-Wilks test does not provide statistical significance to reject the null hypothesis–which is the residuals are drawn from a population with a normal distribution. The caveat with having a small dataset is that it might be hard to obtain statistical significance for large deviations from normality. Based on the visual aids and statistical test results, it is concluded that this assumption is met.

5   Regression Table

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Sep 19, 2020 - 18:50:07

We used the Hetereoskedastic-Consistent robust standard errors to generate the independent variable coefficients in the results of our regression table.

We tested three different models to understand if political affiliation plays a role into the use of masks.

Model 1 contains one explanatory variable:

1. `repu` - the percentage of the population identifying Republican or leaning Republican for each state.

Model 2 contains three explanatory variables, two additional to Model 1:

2. `maskdays` - the number of days since a mandate was enforced.

3. `d100k` - deaths per 100k people.

Model 3 contains four variables, one additional to Model 2:

4. `percent_apartment` - the percent of people living in apartments.

Table 1: Summary of all models with Huber-White standard errors

	Dependent variable:		
	maskwear		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
maskdays		0.103***	0.101***
		(0.029)	(0.029)
repu	−0.592***	−0.375***	−0.305*
	(0.087)	(0.102)	(0.126)
percent_apartment			0.141
			(0.147)
d100k		0.034	0.032
		(0.026)	(0.026)
Constant	65.005***	49.482***	45.125***
	(3.686)	(5.120)	(6.734)
AIC	324.6	306	306.8
Observations	51	51	51
R^2	0.451	0.647	0.656
Adjusted R^2	0.440	0.625	0.626
Residual Std. Error	5.611 (df = 49)	4.591 (df = 47)	4.584 (df = 46)
F Statistic	40.237*** (df = 1; 49)	28.764*** (df = 3; 47)	21.929*** (df = 4; 46)

Note: *p<0.05; **p<0.01; ***p<0.001

The R^2 adjusted value improves from Model 1 through Model 3. However, the difference in R^2 adjusted between Model 2 and Model 3 is marginal, and adding the percent of people living in apartments did not benefit in answering our research question.

Our standard errors are low across each coefficient, which indicates we have low variance in our OLS estimators. The percent of the population identifying republican remains statistical throughout all the models. This indicates that it is a good variable for explaining the response of the model. The practical significance of a 1% increase in political affiliation is associated with a 0.37% change in the response is significant in the number of people wearing masks throughout the country. However, a change of 1% in political affiliation will most likely come slowly over time. If a change in political leaning could be measured and verified as changing in a reasonably short timeframe, it would be more practically significant.

6 Omitted Variables influencing our model

We have to look at omitted variable bias because we don't meet the Zero Conditional Mean assumption. This tells us that there are variables affecting our error term in our regression. It's important to note that this data is more like panel data. We have to consider that behavior and the explanatory variables we are using to describe mask-wearing are time-variant.

Being republican contains many intrinsic properties. We are going to attempt to explain through our omitted variable bias analysis. Ultimately, many of these variables seek to measure human behavior and the association with political belief, some of which can be measured, many of which would require a deeper dive into behavioral psychology and relationships to a political party.

6.1 Omitted variable Bias #1: Hours consuming conservative media

Throughout the pandemic, news sources such as Fox News and One America Network had a strong stance against mask-wearing. The pandemic has been ubiquitous in the news, and there were several months of lockdown during March in many states, so viewership of news was up. However, the proportion of overall republicans that watch Fox News and ONA on a regular basis is a small fraction of the population so the true effect of viewing these news sources on not wearing masks is low.

There is a positive correlation between viewing Fox News and ONA and being Republican and negative correlation between Fox News and mask-wearing hence negatively biasing the effect of being republican away from zero making it more statistically significant than it actually is. Again we believe the effect has a small magnitude of influence on the true coefficient relationship between being republican and mask-wearing.

6.2 Omitted Variable Bias #2: Level of education

It is reasonable to believe that people are more likely to follow recommendations from the government related to safety measures against the COVID-19 pandemic to the extent that three requirements are satisfied. First off, people must believe and have trust in their government and that the government is looking out for the best interest of the population as a whole. Second, they must also believe the risks associated with contracting COVID-19, specifically severe illness or death. Third, people must understand the scientific research and findings behind the effect of wearing protective face coverings has on the spread of the virus. The impacts of numbers two and three could be related to the level of education. This could be measured by average years of education, high school graduation rates, or college graduation rates. If what has been mentioned in previous literature regarding red states having a lower level of education holds true. It is possible that by omitting this variable, the republican leaning effect on habitual mask-wearing practices is more negative and therefore moves the coefficient away from zero.

6.3 Omitted Variable Bias #3: Level of rural vs. urban living

From the data we have to date, we can say that COVID-19 affected cities in a larger proportion than more rural areas. We believe that as contagion, illness, and death became more visible in urban areas, fear increased,

and more people started to wear masks as a preventive habit. We believe this effect is not captured properly by the population density variable that is available in the dataset as it does not make adjustments for livable land. One example is Alaska, where density is just over one person per square mile, but the percent of people living in rural areas is under 35% (depends on how rural living is defined). There are fifteen states where rural residents make up more than half the population. Republicans are governors of eleven of them. If the relationship of rural living and republican leaning holds, we might see the republican negative coefficient more negative than it should be, and therefore might be farther from zero than it would be if the omitted variable was captured effectively.
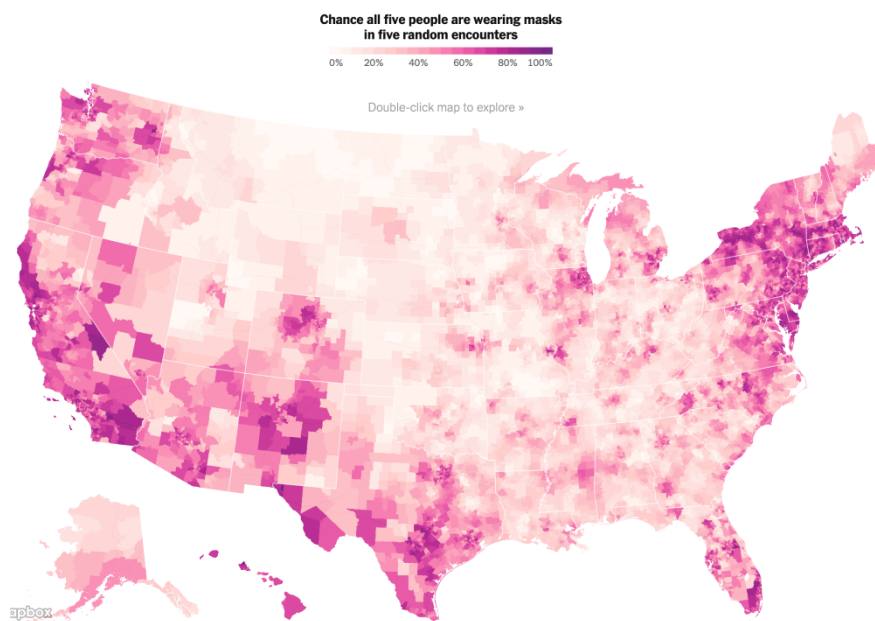

Figure 4: Wearing of Masks by County.

A study done by the New York Times through a global data and survey company called Dynata obtained 250,000 survey responses between July 2nd and July 14th across the United States, enough to provide county level resolution. They concluded similar to what we suspected that areas to the Northeast and West (more urban areas) wear more masks on average and mask wearing is lower in the Plains and parts of the South (more rural areas).

6.4 Omitted Variable Bias #4: Belief on infringement of civil liberties by mask wearing mandate

A common complaint that is heard overwhelmingly by republicans is that they don't want the government to mandate mask wearing and they believe it to be an infringement upon their civil liberties and a constraint on freedom of expression.

Believing mask wearing is an infringement of civil liberties is positively correlated with being republican and negatively correlated with wearing masks, this negatively biases the effect of being republican further away from zero in the negative direction (gaining statistical significance). This means that if we were to have added the percentage of those who believe mask wearing is an infringement on civil liberties the effect of being republican would still be negative but would move closer to zero. The reason we believe this to be a useful variable to look at, is because there is are diversified views within the republican party and this would help separate those that are more libertarian in ideology but associate themselves to the republican party, in essence partitioning the effect of being republican. The effect that this is having on being republican we believe to be of moderate magnitude. Many surveys showed this was a common answer when asking

republicans as to why they didn't wear masks.

6.5   Omitted Variable Bias #5: Fear of negative impact to personal health caused by COVID-19

When COVID-19 began to affect thousands worldwide, there was a major divide among US citizens as to whether or not the virus was lethal to one's well-being. The public portrayal of the virus's seriousness seemed to be downplayed by Republicans, and comments were made that COVID-19 is no more severe than the seasonal flu. Republicans, in general, seemed less fearful of contracting the disease. A survey conducted during the third week of June 2020 provides evidence that this was more than just perception but reality. The survey concluded that Republicans claim to have a lower degree of a concern than Democrats about contracting the virus.

Since a higher level of fear about potential adverse health impacts by COVID-19 can reasonably increase the percentage of the population wearing masks, it makes sense to analyze the effect fear might have on our model. Intuitively as the fear of contracting the virus increases, so does the percent of wearing masks. Based on the study mentioned above, Republicans are less fearful of the virus and, therefore, being republican is negatively correlated with fear. This result predicts that the fear coefficient on Republicans will scale away from zero (more negative) while gaining statistical significance.

6.6   Omitted Variable Bias #6: Religious beliefs and affiliation

Palm Beach County's emergency order to require mask wearing in public exempts people who cite religion as an excuse to disobey. The county is not alone and it might be evidence of a bigger trend of citizens not wearing masks in public due to their own religious beliefs. The Pew Research Center conducted a survey in 2014 with over 30,000 respondents (Republican, Democrat, No lean) and found that 90% of republicans (and republican leaning) either Believe in God: absolutely certain, or Believe in God; fairly certain versus 76% of democrat in the same beforementioned categories. If the relationship of high level of religious belief and republican leaning holds, we might see the republican negative coefficient more negative than it should be, and therefore might be farther from zero that it would be if the omitted variable was captured separately.
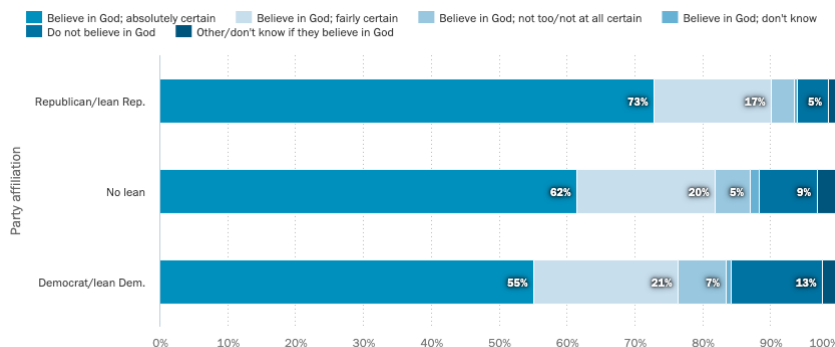

Figure 5: Level of Religious Belief by Political Affiliation.

Summary Table

Omited Variable Bias Effects				
OMVB	β_4	δ_2	**Effect on β_2**	**True Effect**
Conservative News	-	+	negative away from zero	Small Magnitude
Education Level	-	+	negative away from zero	Medium Magnitude
Rural %	-	+	negative away from zero	Small Magnitude
Civil Liberties	-	+	negative away from zero	Medium Magnitude
Fear	+	-	negative away from zero	Medium Magnitude
Religion	-	+	negative away from zero	Small Magnitude

Table 2: Summary table of omitted variable bias on Republican coefficient

7 Conclusion

We set out to answer the question on whether political affiliation affect population behaviors on decrees and mandates related to mask wearing during the COVID-19 pandemic. We specifically focused on whether political affiliation or leaning had an association with populations following government lead policies and initiatives related to the wearing of masks and face protection in public. Answering the question is of upmost importance as it would allow policy makers as well as health organizations to more effectively target communities depending on political leaning and beliefs.

Given our results, and of course, following the well-studied and agreed upon causal effect of wearing masks on the decrease of COVID-19 spread (if done well and as recommended by experts), we propose that political leaning (in this case republican) has a statistically significant association with percent of the population wearing masks, even when adjusted for deaths per 100,000 and days mask wearing policies have been in place. We can therefore say that a 1% decrease in percent of a state's population that is republican is associated with a 0.37% increase in mask-wearing at a high statistical significance (p-value = 0.0006).It is extremely important to note the difficulty that comes with working in the realm of behavioral psychology.

We would certainly like to see further studies that can go deeper into some of the omitted variables mentioned in this paper. In addition, it would be pivotal to have a better understanding of this phenomena via panel data analysis to, among other things, take into account serial effects of the dependent variable before we can confidently claim a causal effect.