Final Report - NYP Data Science Team 8

Introduction

The Covid-19 Pandemic infected millions of people worldwide, leaving many with compromised long-lasting health effects, and has killed millions. It is crucial to find effective methods of combating the virus to help people and fight future viruses. Out of the many policies implemented to combat the Pandemic, we must find which practices were effective. Given this, we seek the most effective policy for addressing Covid-19. Our team has decided to investigate the following refined question based on the assigned task: Which policy among public information campaigns, testing requirements, mass gathering restrictions, mask mandates, and vaccine mandates, has led to the greatest reduction in COVID-19 growth rate of new infections in the United States since the first policy on COVID-19 was instituted?

We used COVID-19 infections because we believe that the dominant policies implemented during the Pandemic directly affect the number of infections. Positive COVID-19 infections are recorded in accredited sources such as the CDC is measuring the Pandemic's growth. We chose five dominant policies implemented during the COVID-19 Pandemic to study: public information campaigns, testing mandates, mass gathering restrictions, mask mandates, and vaccine mandates. Instead of targeting specific groups, as other policies did, such as school closings, and travel restrictions, these five dominant policies were implemented in some form throughout the United States and targeted most Americans. Given this, we believe these five policies have a broader effect on combating the COVID-19 Pandemic.

Data

Our data is a combination of data sets from The Centers for Disease Control and Prevention and The Oxford Covid-19 Government Response Tracker (OxCGRT), Blavatnik School of Government, and the University of Oxford. The data contains daily infections, multiple policies tracked at a state level, and the date of implementation. It contains daily infections of 50 states in the U.S from January 1, 2020, to November 1, 2021, as well as the levels of the dominant policies, tracked at a state level (from 0 to 4 or 5, 0 represents no policy, the large number represents rugged policy). We chose this data as it has the policies at the state level we desired for our analysis such as the date of implementation and policy intensity. The merits of this data set are that it is combined from accredited sources and is publically available. A limitation of our data selection is that we do not have data on every actual positive COVID-19 case in the United States, as not everyone gets tested when feeling ill. Due to this, we can only get data on people who received a COVID-19 test.

Data Preprocessing

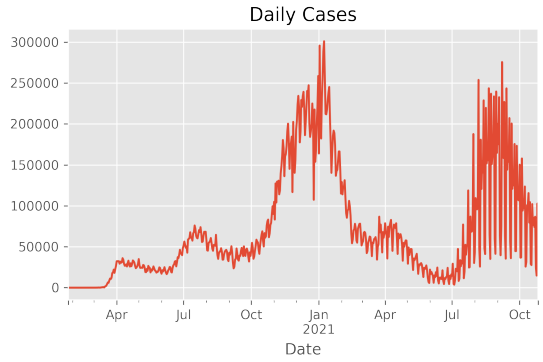
We combined the two data sets from the locations described and only kept the policies we desired. This data cleaning process resulted in 30,000 rows and 6 columns as we are studying 5 policies and 1 response variable (infections). As we want to study the five dominant policies, we removed other lines containing other policies and only kept the lines of date, state names, daily

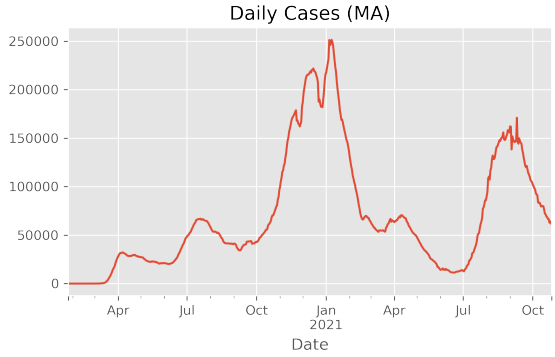
new infections and the policies at state level. Every policy has a different number of levels and the level of policy was at least 1 since the epidemic outbreak. To simplify the analysis, we assigned 0 to level-0 and level-1, and 1 to higher levels.

Statistical Analysis

Assumptions

We are assuming that the number of COVID-19 cases are only related to its state policies. We are assuming that people are not crossing state lines during the epidemic, meaning we will not consider interstate effects. We are also assuming that every district in a state has the same effect of policy implementation. At the beginning of the Pandemic, we assumed that a minority of the population has been infected and that herd immunity has not yet been achieved (i.e., at least 60% of the population has acquired immunity by vaccination or infection). Under these assumptions, the growth in the number of infections is proportional to itself, which means cases were growing exponentially. We use the growth rate of new infections, $g_{it} = log(x_{i,t+\Delta t}) - log(x_{it})$, to define effectiveness.





During data preprocessing, we observed that daily infections are highly unstable and fluctuant due to testing errors (see Daily Cases Figure). Applying the moving average method can solve this issue by making the infection numbers more continuous (see DailyCases (MA) Figure). Therefore, here we define x_{it} as the average infection cases from (t-3) day to (t+3) day in state i.

In the label formula $g_{it} = log(x_{i,t+\Delta t}) - log(x_{it})$, Δt is taken into consideration because it takes time for a policy to show impacts on new infections after it is implemented. We first chose Δt=21, then we changed Δt to be 3, 7, 14, 30 later in the sensitivity analysis, and searched for patterns of change.

We need to take the logarithm of daily new infections and it makes no sense when the number of infections is 0. Therefore we add 1 to daily new infections to account for this. Due to the huge fluctuation range (around 30000) of the daily new cases, the influence of adding 1 can be neglected.

Models

Model 1: Log-linear Model

In order to estimate the effect of the five policies, we constructed a log-linear model with g (growth rate of new infections) as the dependent variable and the levels of policies as the independent variables. The log-linear model constructed is as follows:

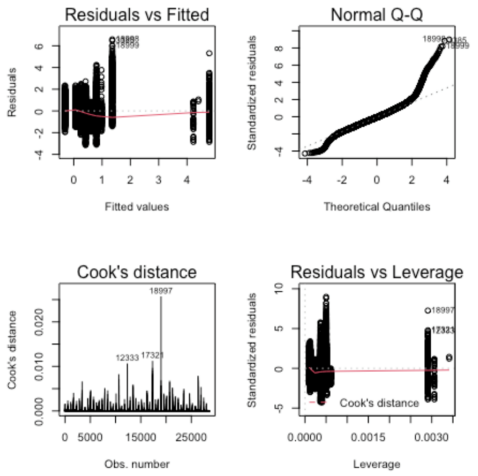
$$log(x_{i,t+\Delta t}) - log(x_{it}) = \delta_0 + \sum_p \quad \delta_{ipt}\Delta g_p + \varepsilon_{it},$$

where δ_0 is the intercept term (i.e., the initial growth rate if no policy was implemented), $\delta_{ipt} = 1$ if state i had policy p in place in time t and $\delta_{ipt} = 0$ otherwise. Δg_p is the estimate for the average effect of policy p. ε_{it} is the random error term and we assume it's normal distributed with mean 0 and variance σ^2.

The coefficients of five policies are shown in Table 1. The coefficients are all negative which means these policies all contribute to reducing the COVID-19 growth rate of new infections. Among them, the absolute value of the coefficient of Public information campaign is greater than others, the second effective policy is restriction on gathering policy, and testing policy, facial covering and vaccination policy have similar effectiveness.

Table 1

	Estimate	
(Intercept)	4.78427	***
Restrictions.on.gatherings	-0.56117	***
Public.information.campaigns	-3.42277	***
Testing.policy_Level	-0.38196	***
Facial.Coverings	-0.35557	***
Vaccination.policy	-0.36907	***
R-squared	0.4114	



The summary table tells us more about the effectiveness of this model and the effects of the five policies we are studying. The R-squared value of 0.4114 means that the model is not the best fit, since a higher R-squared such as .7 would indicate the model is a good fit for the data while a R-squared of less than .3 would indicate that this model is not a good fit for the data. The table also indicated that all of the five policies were statistically significant,which means that the five policies are effective in reducing COVID-19 infections.

The Residual vs Fitted values in Residual Plots 1 shows the heteroscedasticity of the model as the residuals and the square root of standardized residuals increase with the increase of the fitted values. The Normal Q-Q Plot illustrates that the residuals are not normally distributed and violate the normality assumption, because under ideal conditions those points should fall on the straight line in the diagram. It is telling us the model may not be the best option. According to the last two

plots, the data has many outliers as the Cook's Distance and leverage of some points are much larger than $\frac{4}{n}$ (i.e., 0.00013). These are telling us this model has a lack of fit and is far from perfect. But considering we used only a small part of policies, a lack of fit is unavoidable. We continue to use other models to check if these models are telling us the same conclusion.

Model 2: Hierarchical Model

We think the condition in different states may be different, so based on Model 1, we use the hierarchical linear model to explore differences in initial growth rates of new infections in different states. We add binary variables for different states to the intercept term. In other words, different states have different intercepts due to their own characteristics. The equation is as follows:

$$log(x_{i,t+\Delta t}) - log(x_{it}) = \delta_{0i} + \sum_{p} \delta_{ipt}\Delta g_p + \varepsilon_{it},$$

where $\delta_{0i}$is the initial growth rate of state i, the meaning of other terms are the same as Model 1.

Figure 2 shows the results of the model: there is the same general pattern between policies as displayed in the log-linear model 1. The black line is the mean estimate of Δg. The box is the 95% confidence interval. It showed negative coefficients for the fixed effects of the model indicating a reduction in positive daily tests.

The residual vs fitted value plot (Fig.3) shows the similar pattern as the Model 1. But the R-squared of this model is 0.477, which is bigger than that of Model 1, so we think this model is more reasonable.

Fig.2 Hierarchical model

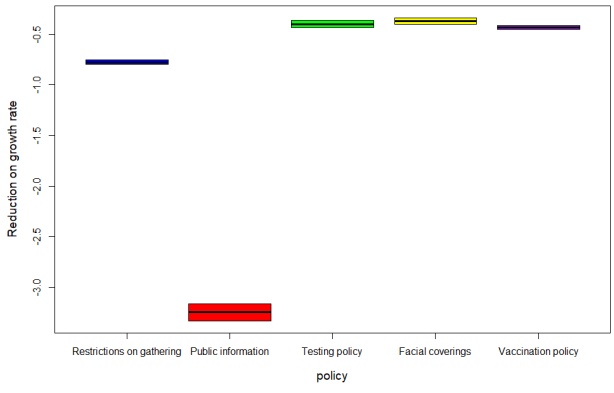
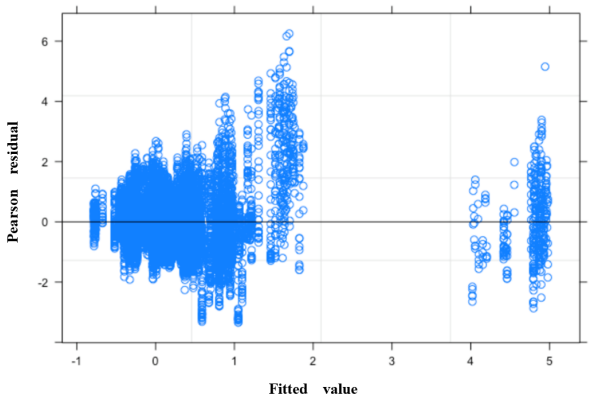


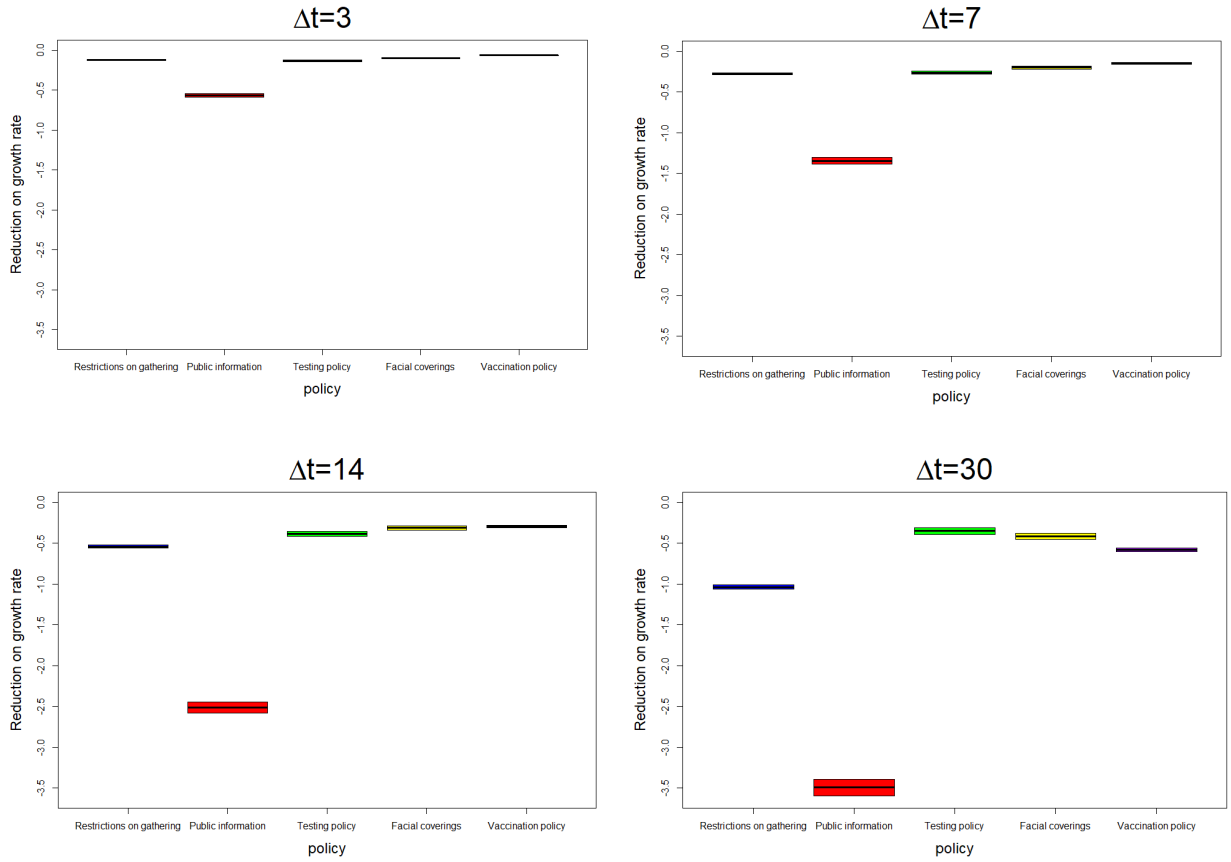
Fig.3 Pearson residual vs Fitted values



Sensitivity Analysis

We want to check if the results would change when we take different Δt. So we change the dependent variable with various Δt : A model with Δt of 3 and a model with Δt of 7 both show negative coefficients and therefore reduce new cases. However Δt=7 shows larger coefficients

and a larger spread between these coefficients. Exploring this further we did Δt of 14 and Δt of 30 shown below. As we increased the Δt the coefficients become more negative and the spread between them larger. These models with different Δt show the same general pattern between policies, that Public information campaign has the greatest effect on the growth rate of new infections, gathering restrictions shown to be the second most effective policy and then other three policies.



Interpretation

The Log Linear Model indicated that all the policies were effective in a reduction of cases. As the Public Information campaigns had the largest negative coefficient, it had the strongest effect on infections based on that model. When trying different levels of Δt all show that Public Information campaigns still have the largest reduction in growth rate. The Hierarchical Model also indicated that public Information campaigns had the greatest impact on the reduction of COVID-19 daily cases. Through the analysis of our models we believe that Public Information campaigns are the most effective in policy among mass gathering restrictions, mask mandates, public information, testing mandates and vaccine mandates, led to the greatest reduction in COVID-19 growth rate of new infections in the United States since the first policy on COVID-19 was instituted.

Conclusion

Our team was given the task of determining the most effective policy for addressing the COVID-19 Pandemic. We decided to define effectiveness as a decrease in daily positive COVID-19 cases. After considering policies that were implemented, we decided to analyze public information, testing mandates, mask mandates, vaccine mandates, and gathering restrictions because these were the most prevalent measures implemented to mitigate the spread of the virus. We obtained data about infection rates, policy implementations. We acquired these from the CDC and the University of Oxford and began our statistical analysis. We built a log-linear model and a hierarchical model to study policy effectiveness. After consideration of our models constructed, this team has determined that the most effective policy in addressing the COVID-19 Pandemic is Public Information Campaigns.

Limitations

We did face limitations in our data as we only have observational data available and not all people who were infected with the COVID-19 Virus received a positive test. We face a broader limitation that there could be a development of brief herd immunity before any COVID variants are mutated in the public. Given this, we can only base our analysis on policy implementation and recorded positive cases. Therefore the effectiveness of mask mandates can only be considered by positive covid cases and not the actual amount of COVID infections throughout the population. Our sensitivity analysis is limited by models not producing a high R-squared value. Sensitivity analysis did show changes in r squared values; they still were not high. This may imply that the model is not extremely effective in predicting cases.

Works Cited

Al-Betar, M.A., Alyasseri, Z.A.A., Awadallah, M.A. et al. Coronavirus herd immunity optimizer (CHIO). Neural Comput & Applic 33, 5011–5042 (2021). https://doi.org/10.1007/s00521-020-05296-6

CDC. (n.d.). CDC Covid Data tracker. Centers for Disease Control and Prevention. Retrieved from https://covid.cdc.gov/covid-data-tracker/#datatracker-home.

Centers for Disease Control and Prevention. (2021, May 24). *Covid-19 vaccinations in the United States,jurisdiction*. Centers for Disease Control and Prevention. Retrieved November 1, 2021, from https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc.