Effect of Stay-at-Home Orders on the Growth Rate of COVID-19 Cases

Abstract

The COVID-19 pandemic has led to many governments taking drastic measures to keep people from infection. One of the largest steps they have taken is implementing stay-at-home orders to deter the spread. The goal of this paper is to see if there is a significant difference in the rate of infection at the county level in the US before and after the order was put in place. In particular, using the random forest for classification as a main tool, we show that the number of days since the start date of the stay-at-home order is significant. The result is further confirmed using the classification tree and lasso regression. Based on these results, we conclude that the stay-at-home orders did help reduce new cases of COVID-19 in the US.

1. Background and Significance

Since COVID-19 was first reported on January 11, 2020 in Wuhan China, it has spread worldwide having devastating effects on the lives of millions. Because COVID-19 is caused by a novel coronavirus, humans have no antibodies and no vaccine. This has led to high initial infection rates which have helped the virus spread quickly across nations, states, and cities. Many hospitals, especially those in densely populated areas, quickly got overwhelmed. As a result, the need for medical equipment such as ventilators, personal protection equipment such as masks, gloves, and hand sanitizer skyrocketed. Furthermore, the quick spread of the coronavirus triggered a call for people to "flatten the curve". The purpose of the call was to limit the number of new cases so that hospitals would be able to have beds and resources for all COVID-19 patients.

In order to help "flatten the curve", many state governments in the US implemented stay-at-home orders (SAHOs). These orders mandated that all non-essential businesses close and that people stay home unless necessary. Although each state had different guidelines behind what constituted an essential business, the idea behind them remained the same. A reduced amount of unneeded interaction would help slow the spread of the virus and allow for hospitals to be able to treat more people. In addition, other measures, including wearing masks while in public and washing hands, were recommended or mandated in many states.

The goal of this paper is to see how effective these SAHOs were in reducing the amount of new COVID-19 cases by county. Because these orders have caused the closure of many businesses, some of which may never recover, the efficacy of these orders needs to be questioned. If there is seen to be no impact on the rate of new cases, the validity of these orders should be suspected as thousands of small businesses and people would be crippled with debt. However, if these SAHOs do have an effect, countless lives could be saved.

Our hypothesis is that the SAHOs do have a positive effect and will decrease the *average daily growth rate (ADGR) of total confirmed cases*, formally defined in the next section. We postulate that coupled with other factors, such as percent of population over 55 years old and movement changes compared to the prior year, the data should show a decrease in the rate of new cases. Lastly, we also examine if any other factors are better suited to predict the pre-SAHO-to-post-SAHO change to understand why some counties seem to be responding better than the others.

3. Methods

Data Collection

We gathered data from the 2010 Census, US Bureau of Economic Analysis (US BEA), John Hopkins University (JHU), and Google's COVID-19 Mobility Report, all focused at the US county level [1-4]. The JHU data for daily total confirmed cases per county was averaged over the span of the first day that the county had at least 100 cases until five days after the SAHO was in effect (accounting for the five-day incubation period of COVID-19 [5]), which we called "Pre-SAHO". Similarly, "Post-SAHO" takes the average of the daily total confirmed cases per county from five days after the SAHO until May 22. Finally, the quantitative response variable is given by

$$ADGR = Pre\text{-}SAHO - Post\text{-}SAHO.$$

We require that the counties have at least five days with over 100 total cases in the pre-SAHO period in order to stabilize the ADGR. In addition, for the classification methods, we convert the ADGR into a binary categorical variable, "Decreased Greatly", separated by the median ADGR

of 0.01305, a decrease of 0.01305 cases per day on average per county. Here, the value of 1 is assigned for a large decrease (> 0.01305) and 0 otherwise for at most a small decrease or increase.

We have chosen ADGR as our response variable to avoid any trivial results. For example, simply looking at the average number of COVID-19 cases would result in an uninteresting outcome of identifying the size of population as the main factor affecting the COVID-19 infection. By correctly accounting for the infection before and after the SAHO using ADGR, we can better identify factors affecting the COVID-19 infection rates.

From Google's COVID-19 Mobility Report, we have obtained percent changes in mobility from baseline (median value during January 3 – February 6, 2020) in six categories ("Retail and Recreation", "Grocery and Pharmacy", "Parks", "Transit Stations", "Workplaces", and "Residential") measured daily, averaged over all of Google's users tracked in those counties. For each category, we averaged the daily percentages for each county all days after the SAHO until May 22 as movement metrics to see how well citizens have adhered to the SAHO. These counties without movement data were removed from our training dataset.

Additionally, we calculated the number of days between the date of the first case and May 22 in the county ("Days Since First Case"), and between the date that the SAHO came into effect and May 22 ("Days Since SAH"). Moreover, from the Census, we collected information about size of population ("Total Population") and the percentage of population over 55 years of age ("Percent 55"). Finally, from the US BEA, we collected county per capita income ("Per Capita Income").

In total, 134 counties are included in our study. Out of these 134 observations, 100 of them were randomly selected to train our models. The remaining 34 observations were used for testing the models.

Random Forest Model

We use random forest (RF) as our main model to show how well we can predict the "Decreased Greatly" response variable. RF is a popular statistical method to obtain accurate prediction by aggregating results from multiple classification trees with different subsets of predictors. In the model, we use the 11 predictors described in the last two paragraphs related to population data, date of the SAHO, and adherence to the SAHO. For the RF parameter, 3 features per tree were chosen by using the square-rooting rule-of-thumb ($11^{0.5} \approx 3$). To determine the number of trees for the final RF model, we used the grid search by fitting the training dataset using 10 to 1000 trees with an increment of 10. Eventually, 640 is chosen as it minimizes the out-of-bag (OOB) error (**Figure 1, left**).

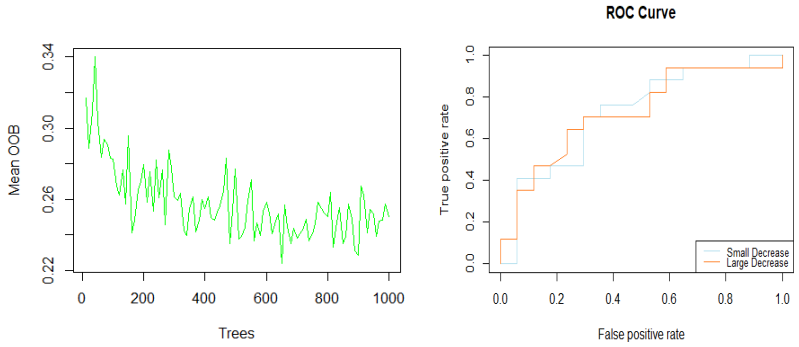

Figure 1: Changes in the OOB errors as a function of the number of trees (**left panel**) and the ROC curve (**right panel**).

Classification Tree (CT) and Lasso Regression Model

We supplemented the RF model above with a classification tree (CT) model to obtain more information on the predictors affecting the ADGR using the same training and testing data as the RF model. Due to its lack of robustness to small changes in the data, the CT model is not used as the main prediction tool. A similar analysis is also made using the lasso regression model for further confirmation. The additional analysis confirms that "Days Since SAH" is the most important predictor. All the results are reported in the appendix.

4. Results Using the Random Forest (RF) Model

The RF model described in the previous section has an OOB error of 22.4% (**Figure 1, left**) with a sensitivity rate of 78% and a specificity rate of 74%. These numbers suggest that the model is accurate enough to obtain the underlying variables affecting the "Decreased Greatly" variable (the binary version of ADGR). Specifically, the variable importance chart (**Figure 2**) demonstrates that "Days Since SAH" is extremely important, followed by the movement data ("Grocery and Pharmacy" and "Residential") and the percentage of population over 55 ("Percent 55") in predicting whether or not a county has successfully decreased the ADGR greatly.

A further analysis of our random forest is shown by the receiver operating characteristic (ROC) curve, demonstrating the tradeoff between sensitivity and specificity (**Figure 1, right**). Because our ROC curves are far from a 45-degree line, we conclude that our RF model is doing well at accurately predicting results.

5. Discussion

We conclude that days since the stay-at-home order (SAHO) is implemented is highly associated with large decreases in the average confirmed cases (ADGR). We also see from the variable importance chart that grocery and residential movement as well as percentage of people over 55 are also significant in determining the ADGR. That implies that not only how early SAHOs are implemented, but also how well people adhere to the SAHO (leading to higher "Residential" values), is highly associated with large decreases in ADGR. In addition, older age demographics are associated with large decreases, which could potentially be from less mobility from the elderly or from higher levels of caution from their susceptibility to severe symptoms.

On the other hand, the number of days since first case was not included in the RF, CT, and lasso regression model. That implies that the passage of time is not a significant factor of slowing the ADGR regardless of the policy change. This finding strengthens our conclusion that the SAHO is not only effective but is also necessary to reduce the spread of COVID-19.

The conclusion above may be limited due to several important assumptions we made on the data. First, we assumed that the COVID-19 infection rate follows a general trend without too many outliers. Next, we assumed that the Google movement data have no seasonal effects on movements over the year. Additionally, we assumed that removing observations with missing data do not significantly affect the overall outcome. Note that our dataset was reduced substantially by lack of reporting from certain states on movement data and by our choice to only include counties that had at least five days with over 100 cases before the SAHO. In other words, a more complete dataset may have led to less biased results as we omitted a relatively large segment of the US population.

For our future work, time series analysis on the ADGR, movement changes, and the cumulative infections could be effective. For example, intervention analysis with SAHO at different counties could be useful. We also believe that a longer time period for the infection rates should be examined in the future to retrospectively identify factors that contributed to "flattening the curve".

References

1. *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University*. GitHub. (2020). Retrieved from https://github.com/CSSEGISandData/COVID-19.

2. *Community Mobility Report*. COVID-19 Community Mobility Report. (2020). Retrieved from https://www.google.com/covid19/mobility/.

3. *Personal Income by County, Metro, and Other Areas*. Bea.gov. (2020). Retrieved from https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas.

4. *County Population by Characteristics: 2010-2018*. The United States Census Bureau. (2020). Retrieved from https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

5. Lauer, S., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172(9), 577—582. DOI: 10.7326/M20-0504.

APPENDIX

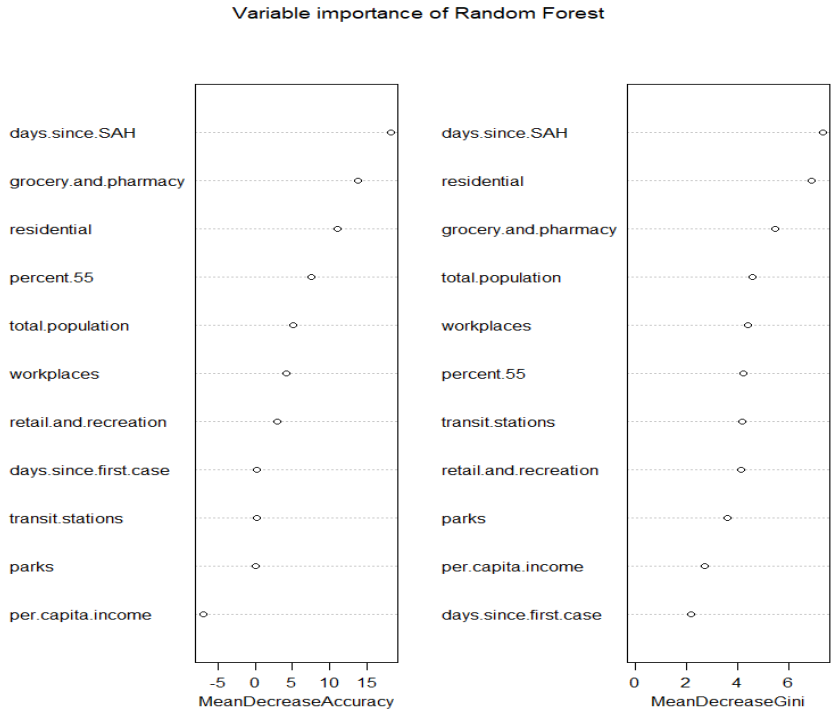

Figure 2: The importance of variables determined by the random forest with variables that are more important being higher up the graph.

Additional Analysis

Classification Tree (CT) Model

Using the CT model, we also see in our individual tree model that “Days Since SAH” is once again the primary variable of deciding whether there is a large decrease in the ADGR, as well as movement (“Grocery and Pharmacy”). Although the CT model is less robust with changes in the data, it consistently demonstrates that the SAHO is the primary factor in determining change in the ADGR, and that a higher number of days since SAHO is associated with a greater decrease in the ADGR (**Figure 3**).

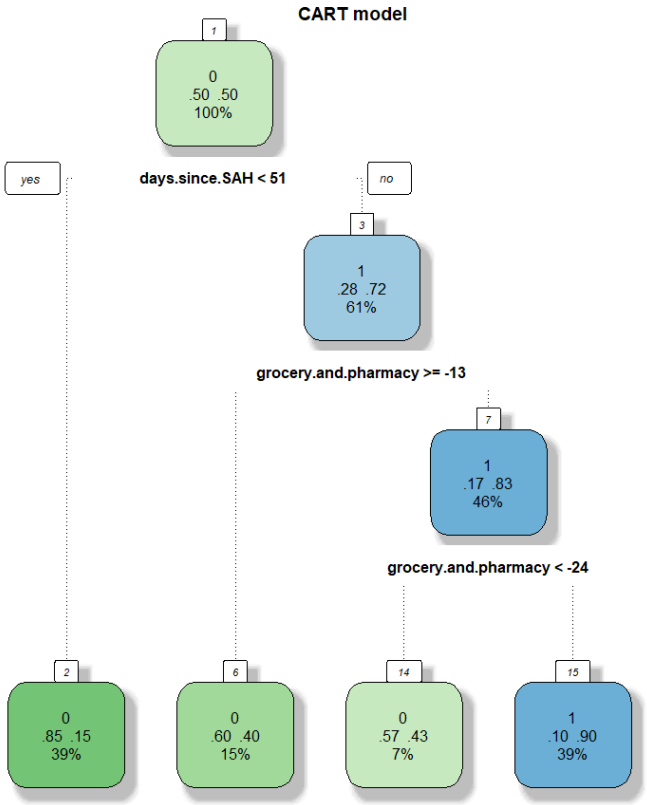

Figure 3: Outcome of the classification tree model (1 as high decrease and 0 as low decrease) showing “Days Since SAH” and “Grocery and Pharmacy” as significant.

Lasso Regression Model and Results

The lasso regression is fitted to our data to identify significant predictors using the original ADGR as the responding variable. Prior to fitting the data, each predictor was standardized, in order to compare the magnitudes of the selected variables. The model estimation and selection are simultaneously performed using the cv.glmnet() function in the glmnet R package. To select the penalty parameter, the one that minimizes the mean squared error from 10-fold cross validation is calculated.

The resulting model identifies two predictors, namely, “Days Since SAH” (with a coefficient of 0.007572) and “Residential” (with a coefficient of 0.000407) as significant. The former variable appears much more significant than the latter by comparing the magnitudes of their coefficients directly. In fact, when the same fitting procedure is repeated with multiple seeds, the only predictor not eliminated consistently by the lasso regression is “Days Since SAH”. The corresponding coefficient consistently shows a positive value, demonstrating that the longer the stay-at-home order is active, the larger the decrease in the ADGR is.

The residuals of the model are displayed through a normal probability plot and residual plots (**Figure 4).** These plots show that the residuals closely follow a normal distribution with a constant variance. The normality of the data is checked again by using the Anderson-Darling tested (p-value = 0.2571), showing that the distribution of the residuals is consistent with a normal distribution.

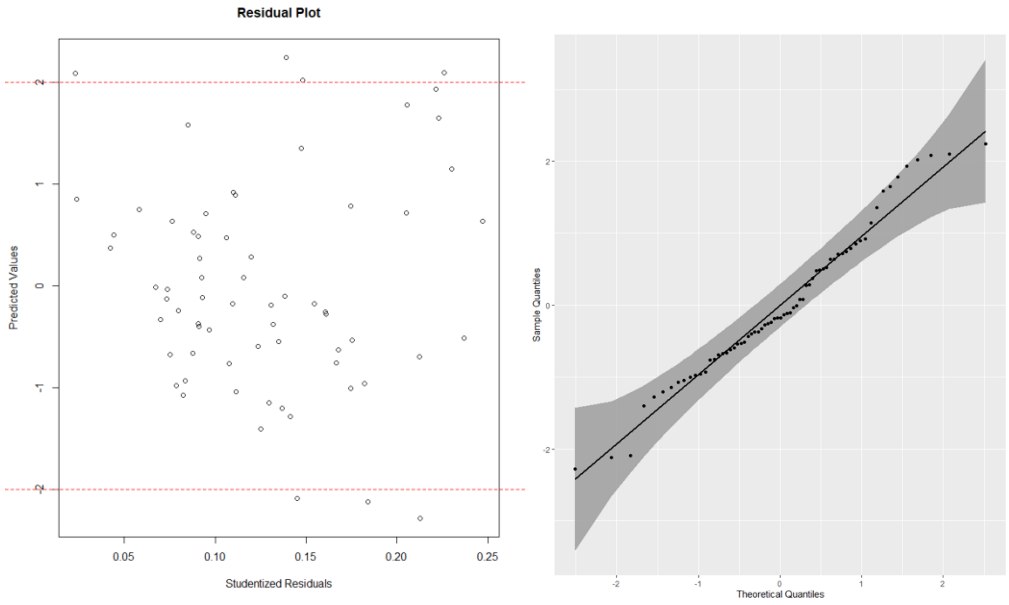

Figure 4: Residual plot with lines at 2 standard deviations **(left panel)** and the normal probability plot with confidence band **(right panel)**.