Effects of Social Distancing and Stay at Home Orders on the COVID-19 Growth Rate in the United States

By Guangpeng Ren and Grant Gambetta

Introduction

The Coronavirus (COVID-19) pandemic originated in Wuhan, China in December of 2019, and since then, the virus has spread to the majority of countries around the world. A large number of countries have experienced pandemics that have been much more severe than China's was, and unfortunately, the pandemic in the US has been the greatest of them all. As of May 5, the US had roughly 1.23 million cases and 71,532 deaths as a result COVID-19, making it the country with the largest number of cases and deaths worldwide. Once the number of cases began to increase exponentially in the US, social distancing and stay at home orders were implemented in an effort to contain the virus and reduce the amount of exposure people have with it. These two measures went into effect around early to mid March, so we analyzed the data from March 14 to March 29 (days 15 to 30) and the data from April 8 to April 25 (days 40 to 57). The ultimate goal of this project was to determine whether or not the social distancing and stay at home orders implemented in the US have had a significant effect on the growth rate of COVID-19. With that being said, the null hypothesis is that the growth rates for days 15 to 30 and 40 onward are not significantly different, or in other words, the social distancing and stay at home orders have not caused a change in the COVID-19 growth rate in the US. To determine an answer to this hypothesis, a general linear model was created to analyze the relationship between the number of deaths in the US and the time periods from above (days 15 to 30 and 40 to 57). This model will be referred to as model 1 throughout this paper. We also created a second general linear model that consists of days 15 to 30 and days 40 to 64 (April 8 to May 3), in order to determine if there is a difference in the results from model 1, as well as to observe if the confidence intervals of the two models have any overlap. This model will be referred to as model 2 throughout this paper.

Materials and Methods

From the time since the first case of COVID-19 was discovered in Wuhan in December 2019, the Johns Hopkins University (JHU) Center for Systems Science and Engineering has been conducting extensive collection of COVID-19 data. JHU acquires the data from a variety of trustworthy sources, such as the World Health Organization and the US Centers for Disease Control and Prevention (CDC) and then combines all the data from these sources into CSV files which are updated daily. The dataset used for this project is a time series dataset which consists of the number of deaths each day since 1/22, in every country that has a confirmed death from COVID-19. For the sake of this project, the confirmed deaths dataset was more optimal to analyze than the number of confirmed cases dataset because the confirmed cases data depends heavily on the availability of testing in each country. Since countries have had varying levels of testing capabilities, there is a vast uncertainty in the actual number of COVID-19 cases due to the inability of most countries to perform consistent and randomized testing.

The software used to conduct the statistical analysis was the R programming language. Before carrying out the statistical analysis, a substantial amount of data manipulation and preprocessing was required, in order to obtain only the relevant data needed for creating the general linear model. A subset of the dataset was created that consisted of only US data, and all days where there were no confirmed deaths were dropped. To do the filtering and subsetting, the dplyr package was utilized, specifically the select and filter functions. Once the relevant US data was selected, a log transformation was applied to the number of confirmed deaths using the log function, in order to make the data more linear. After the US data was filtered and on log scale, it then was observed for the most linear parts within the time periods before and after the social distancing and stay at home measures were implemented. Once the linear parts of the data were identified, the R plotting package, ggplot2, was used to visualize the linear data points using the ggplot and geom_point functions. Also, in order to create the general linear model, a grouping factor variable was added to the filtered dataset to distinguish between the data before and after the social distancing and stay at home orders were implemented. This column is called “group” and 0 denotes the data from March and 1 denotes the data from April and early May. Upon completion of all the preprocessing and filtering, the final dataset consisted of three columns: days, number of deaths in the US on a log scale, and group (0 or 1). Finally, for the analysis, the two general linear models were created consisting of days (x) versus the logarithmic number of deaths in the US (y), with group as a factor variable, and were created using the lm function.

Results

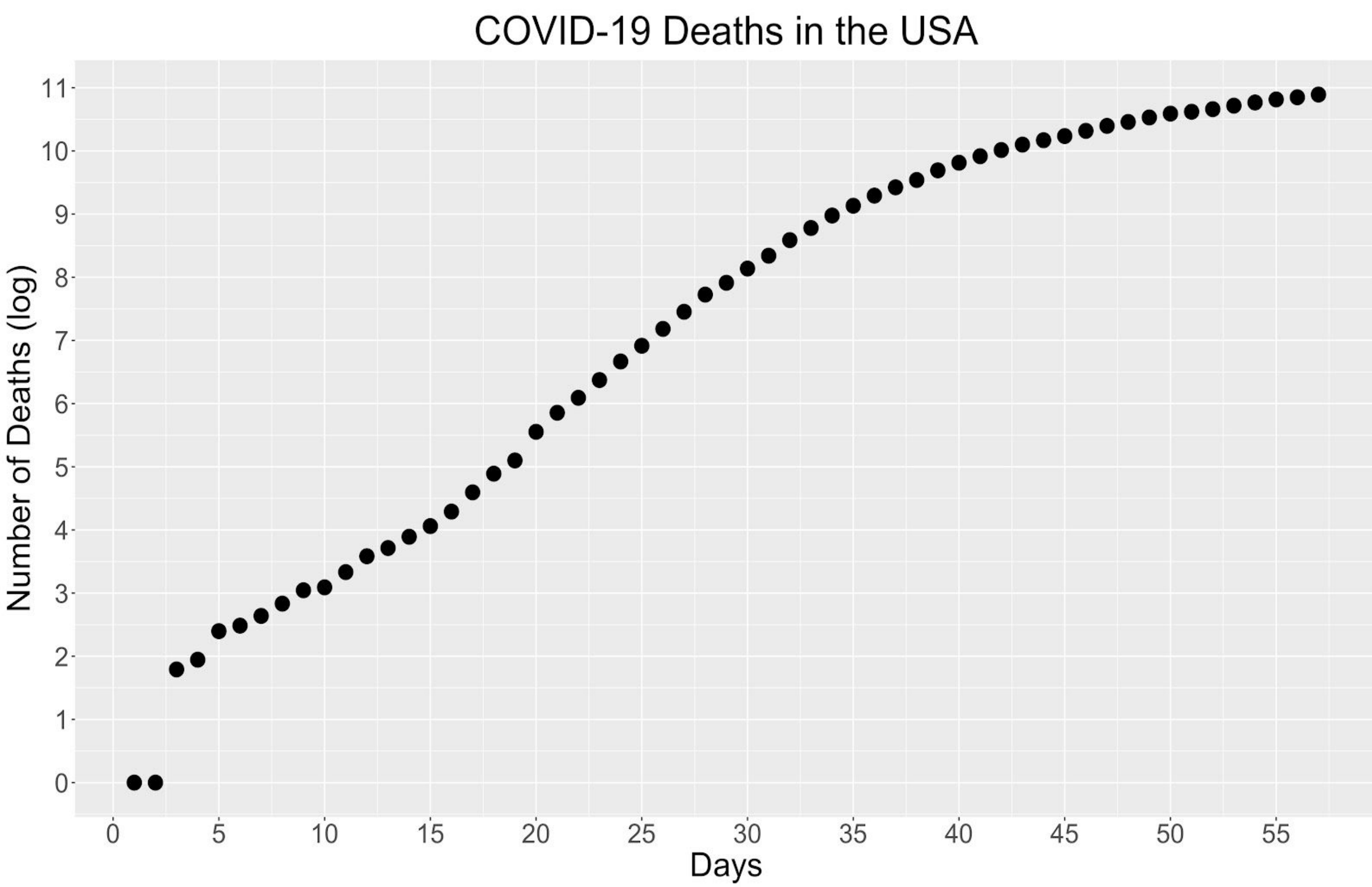

Figure 1. This figure depicts the number of COVID-19 deaths in the United States from 1/22 to 4/25 on a log scale.

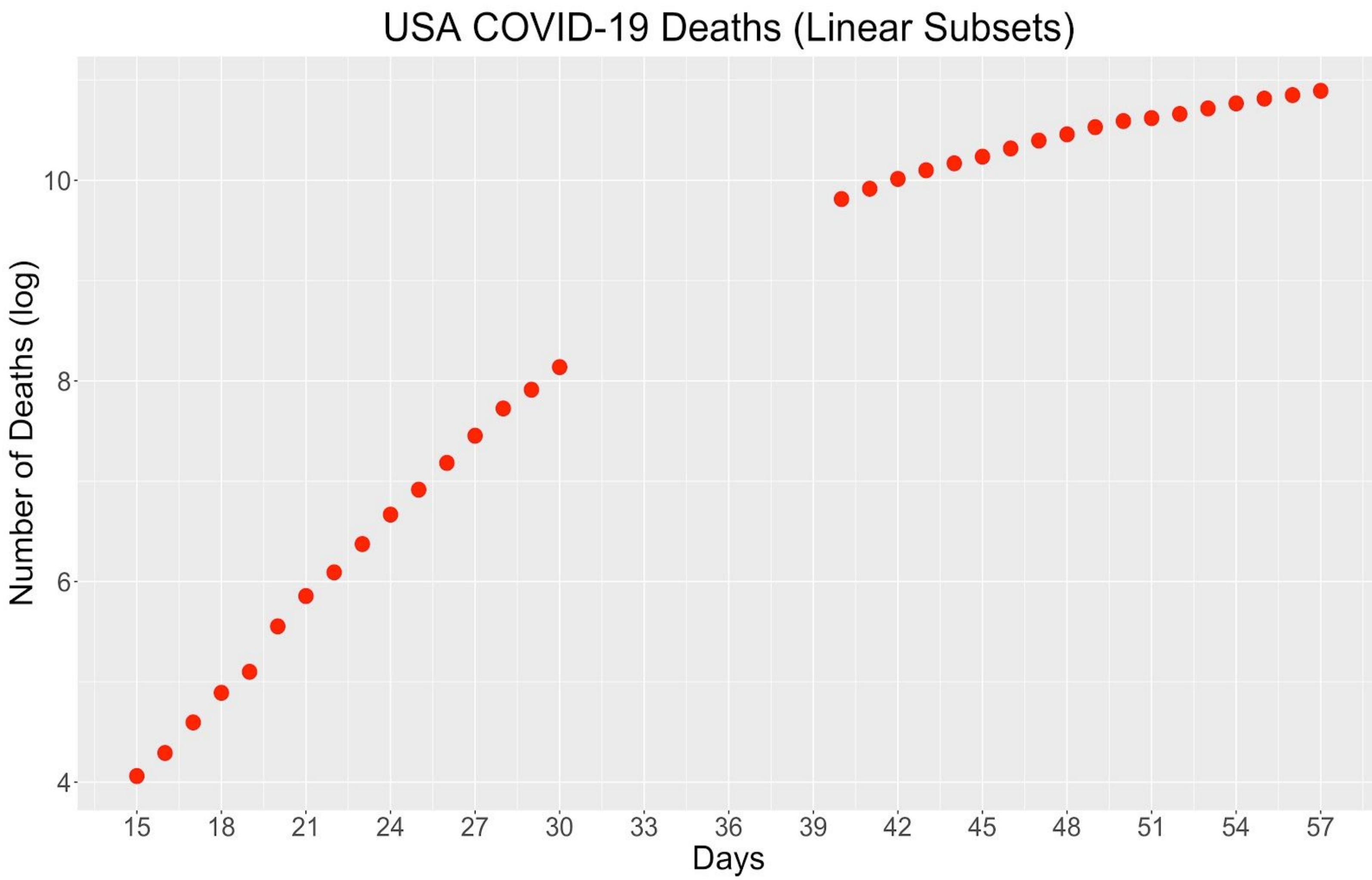

Figure 2. This figure depicts the number of COVID-19 deaths in the United States from days 15 to 30 (3/14 to 3/29) and days 40 to 57 (4/8 to 4/25) on a log scale. These linear subsets were used for the first general linear model (Table 1).

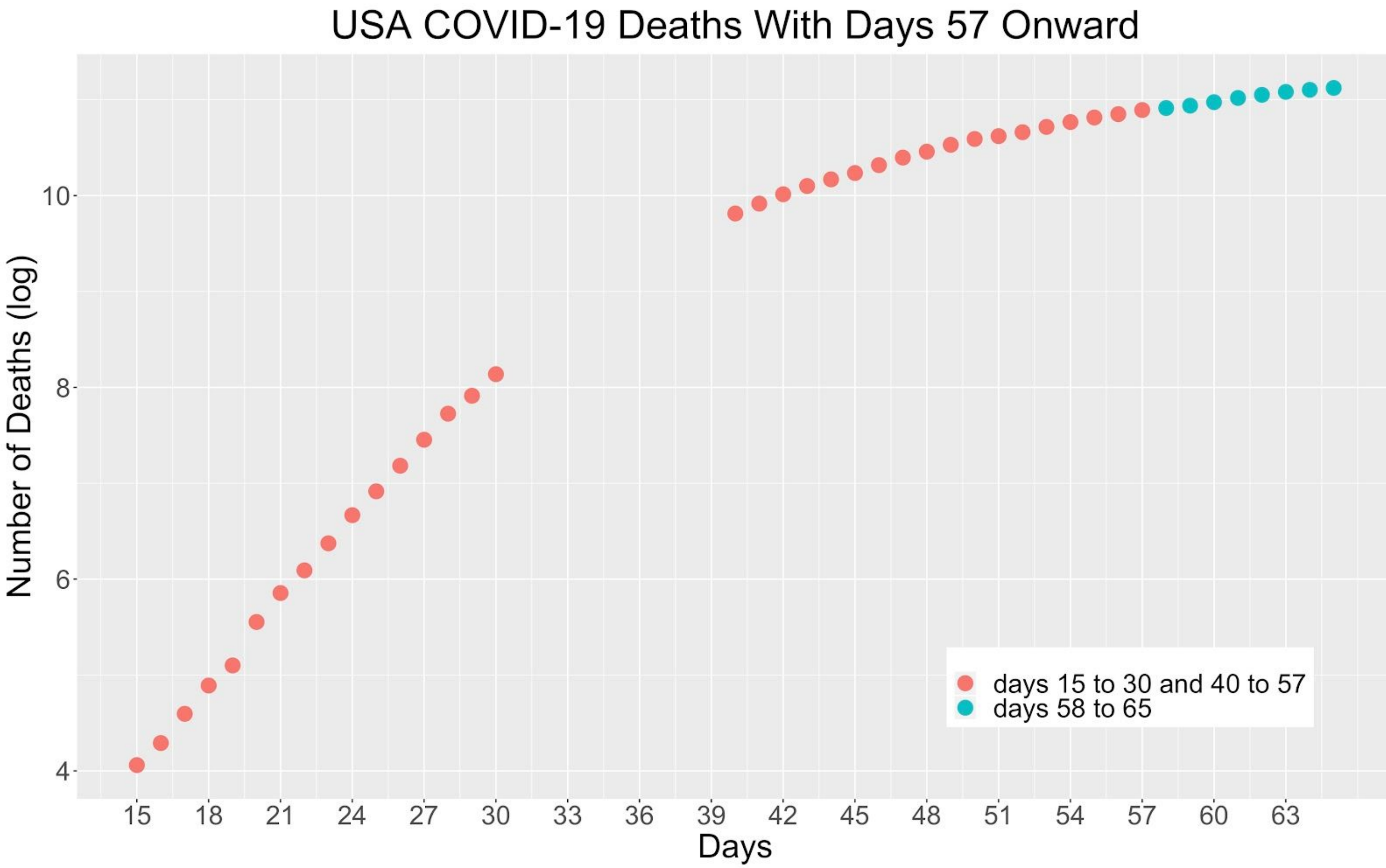

Figure 3. This figure depicts the number of COVID-19 deaths in the United States from days 15 to 30 (3/14 to 3/29) and days 40 to 65 (4/8 to 5/4) on a log scale. These linear subsets were used for the second general linear model (Table 2).

Table 1. This table shows the results of the general linear model created with days 15 to 30 and days 40 to 57 as the independent variable, number of COVID-19 deaths in the US on a log scale as the dependent variable, and group as a factor variable. Group 0 represents the number of deaths for days 15 to 30 and Group 1 represents the number of deaths for days 40 to 57. The adjusted R squared value for this model was 0.9993.

Parameter	**Estimate**	**P-Value**	**95% CI Interval Lower Bound**	**95% CI Interval Upper Bound**
Group 0 (baseline)	-0.1066	0.166	-0.2599	0.0466
Group 0 * Days (baseline)	0.2792	< 2e-16	0.2725	0.2859
Group 1	7.5326	< 2e-16	7.2198	7.8454
Group 1 * Days	-0.2171	< 2e-16	-0.2258	-0.2084

Table 2. This table shows the results of the general linear model created with days 15 to 30 and days 40 to 65 as the independent variable, number of COVID-19 deaths in the US on a log scale as the dependent variable, and group as a factor variable. Group 0 represents the number of deaths for days 15 to 30 and Group 1 represents the number of deaths for days 40 to 65. The adjusted R squared value for this model was 0.999.

Parameter	**Estimate**	**P-Value**	**95% CI Interval Lower Bound**	**95% CI Interval Upper Bound**
Group 0 (baseline)	-0.1066	0.262	-0.2963	0.0831
Group 0 * Days (baseline)	0.2792	< 2e-16	0.2709	0.2875
Group 1	8.0762	< 2e-16	7.7923	8.3601
Group 1 * Days	-0.2286	< 2e-16	-0.2379	-0.2196

Discussion

Our null hypothesis for this project was that there is no significant difference in the COVID-19 growth rates for days 15 to 30 and days 40 onward in the US. Also, both R squared values are close to 1 which means that our polynomials fit the dataset well, and the variance can be mostly explained by our models. From the results, the baseline mortality rate was the slope of group 0 times days, so therefore, the growth rate of COVID-19 in the US for days 15 to 30 was 0.2792. From the general linear model not including days 58 to 65 (table 1), the p-values for the slopes of groups 0 and 1 were both less than 0.05, which means we can reject the null hypothesis and conclude there is a difference in growth rates before and after social distancing and stay at home measures were implemented in the US. To be exact, the estimate of group 1 times days was -0.2171 and if we add that value to the baseline slope of 0.2792, a slope of 0.0621 is obtained which is the growth rate of COVID-19 for days 40 to 57 in the US. Therefore, the growth rate of COVID-19 in the US has decreased from 0.2792 (days 15 to 30) to 0.0621 (days 40 to 57) which means that the social distancing and stay at home orders have helped to decrease the growth rate of the virus.

Next, regarding the general linear model that included days 58 to 65 (table 2), our goal was to compare the results to model 1 and determine if there was any overlap in the confidence intervals of the two models. As in model 1, the baseline for model 2 was also 0.2792, since we did not add any more data to the subset of data selected before social distancing and stay at home measures were implemented. On the other hand, the estimate for group 1 times days in model 2 was -0.2286, which was smaller than the estimate of this parameter in model 1 (-0.2171). If we add the estimate for group 1 times days of model 2 to the baseline slope of 0.2792, we obtain a resulting slope of 0.0506. This means that the growth rate of COVID-19 in the US after social distancing and stay at home orders were implemented was 0.0506, and this value is less than the growth rate of 0.0621 from this parameter in model 1. These results imply that the social distancing and stay at home measures are greatly helping in decreasing the growth rate of COVID-19 in the US. Even over days 58 to 65, which is a somewhat small time period of time, the growth rate decreased by about 0.0115, which is about a 20 percent decrease from the growth rate of model 1.

References

Johns Hopkins University, COVID-19 Deaths Time Series Dataset