Analyzing Relationships Between COVID-19 Cases, Political Parties, and Stay-at-Home Orders with Multiple Linear Regression

By

EMILY BEASLEY
JERRI SCHORR
NICHOLAS VASQUEZ

CAPSTONE PROJECT

Submitted in partial satisfaction of the requirements for the degree of

BACHELOR OF SCIENCE

in

STATISTICS

in the

COLLEGE OF SCIENCE

at

CALIFORNIA STATE UNIVERSITY,
MONTEREY BAY

Approved:

(Advisor)

Dr. Steven Kim

Department of Mathematics and Statistics

Spring 2021

Acknowledgements

Thank you to Professor Judith Canner, Professor Steven Kim, and Professor Alana Unfried, for contributing to the creation of the Bachelor of Science in Statistics program and providing an excellent educational experience to CSUMB students.

Abstract

Due to the global effect of the Coronavirus 2019 (COVID-19) Pandemic, most states in the United States implemented their own strategies to protect public health. Different levels of "Stay-at-Home" orders, ranging from least strict (1) to most strict (4), were applied and varied by each state. All orders had different durations due to the differences in severity of each state's COVID-19 situation. We created new variables to summarize the average weekly percent change of cases before and after the pivotal date of May 2^{nd}, 2020. We address the following questions: Does the decision of stay-at-home order depend on the cases before May 2^{nd}, 2020 and political party affiliation? Additionally, do the cases after May 2^{nd} 2020 depend on the level of stay-at-home orders and the cases before May 2^{nd}, 2020? Through data preparation and multiple linear regression techniques, our results suggest that there is not enough statistical evidence ($p > 0.05$) to claim an association between the choice of stay-at-home order levels and the amount of COVID-19 cases near that time. We found that states with Republican voter majority have less strict stay-at-home orders by one level in comparison to Democratic majority states under similar COVID-19 situations. After the implementation of stay-at-home orders, the average weekly percent change of COVID-19 cases shows a decrease for both Democratic and Republican states, with the former showing a higher drop. Overall, comparisons are made between the average weekly percent change of COVID-19 cases before and after May 2^{nd} 2020 and we show the association of these percentages with the stay-at-home order levels and political party affiliation.

Contents

Acknowledgements ii

Abstract iii

Background 1

Introduction 3

Methods 5
- Data Collection Methods 5
- Data Cleaning 6
- Statistical Methods 9

Results 11

Discussions and Conclusions 16
- Discussion 16
- Conclusion 17

Bibliography 19

Appendix A: R Code 21

Appendix B: Graphs 22

Background

Coronavirus 2019 (COVID-19) has changed the daily lives of people in almost every country. In the United States, most businesses and educational institutions were either closed or operating at less than full capacity. To respond to this public health crisis, most states developed their own official stay-at-home orders, or "Reopening plans." One of the reasons for this was to slow the exponential growth of new COVID-19 cases.

The original stay-at-home orders in response to COVID-19 were implemented between the months of March 2020 to May 2020 and labeled as: "No Statewide Stay-at-Home Order," "Partial Reopening Planned," "Partial Reopening Underway," and "Stay-At-Home Order Intact." By May 2^{nd} 2020, each state had been labeled as one of these orders.

As of May 2^{nd} 2020, American news channel MSNBC reported that"Texas Starts to Reopen as Coronavirus Deaths Hit Single-Day High," and shared the following map of the United States. To supplement the findings of this map, NBC News provided a web link for the reopening plans for each state in much more detail than just the map, including the exact time ranges for each reopening plan [**1**]. The map shown in Figure 1 consists of four colors that represent the reopening plans of each state.

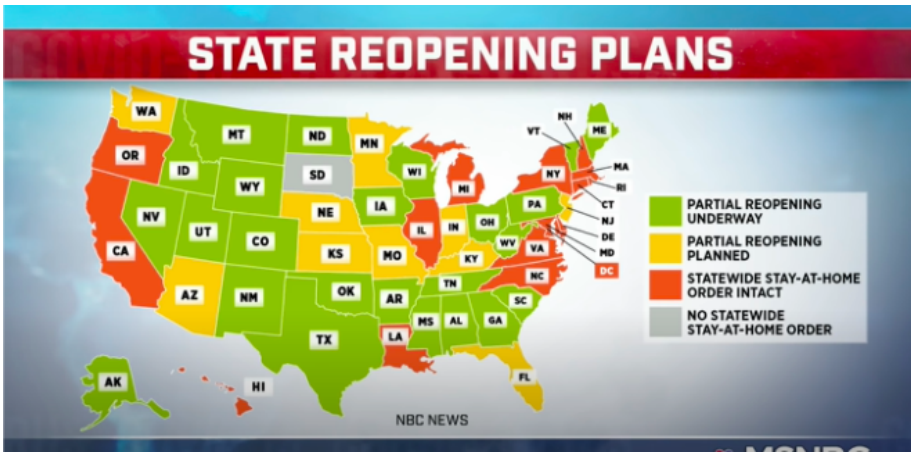

FIGURE 1. MSNBC Map of the United States' State Reopening Plans: 23 states had partial reopening underway (46%), 16 had stay-at-home orders intact (32%), 10 states had partial reopening planned (20%), and 1 state had no statewide stay-at-home order in place (2%).

Each stay-at-home order is leveled from 1 to 4, going from least strict to most strict and is defined as follows:

- 1: No Statewide Stay-At-Home Order
- 2: Partial Reopening Underway
- 3: Partial Reopening Planned
- 4: Statewide Stay-At-Home Order Intact

Partisanship plays a significant role in what information the partisan collects, what information they choose to pay attention to and how they may or may not respond to certain information [**2**]. Our work investigates the possibility that partisan affiliation shows an association with how strict the stay-at-home orders are. We hope our work can help other researchers determine whether or not state officials used their own personal beliefs when deciding which stay-at-home order to choose.

According to the public policy organization named The Brookings Institute, a state's political preferences and affairs contribute to their public health policies as well as the duration of stay-at-home orders. We would think that the pandemic would have brought the country together, yet it has only made the political divide worse [**3**]. Additionally, according to the New York Times, federal scientist Rick Bright warned that consequences to public health would happen due to the inability of the nation to coordinate a response that was based on science rather than politics. Lastly, federal relief funds were disproportionately given to hospitals associated with affluent populations, which should be observed due to the chance that politics, not data, was the reason behind these outcomes [**4**].

Contrarily, other studies deem political affairs as a very important component to understanding the connection between the compliance of each state's residents toward stay-at-home orders and their compliance to politician's recommended next steps to address COVID-19 [**5**]. Publicists, researchers and scholars have a different approach with their reporting and findings, yet each result seems to lean toward politics having some biases in their decision making regarding the public.

Because of this, we grew curious about what we may be able to find by conducting our own analysis and forming our own research questions. We want to study the potential association(s) between stay-at-home orders, political parties and COVID-19 case quantity. We want to quantify the influence that politics may have had in the decisions of stay-at-home order level throughout the states. When researching this topic of public health and safety, we respect the differences between political parties, and we hope that our data analysis delivers reproducible results that positively contribute to scientific research on COVID-19.

Introduction

Research shows that political affiliation changes how people see the world. Some political science researchers theorize that a person's own identity is strongly affected by the group they identify with, and in turn that person often wants to defend their group and promote their group in a positive light [**6**]. When the time comes to make major decisions that concern the general public (e.g, public health, public transit, public education, etc.), we want to see how strong the influence of political party affiliation has on these major decisions. The current pandemic required state officials to make choices on what the response should be to best serve the state's citizens. Naturally, a majority of the state officials have a political party they prefer. We are trying to understand if these public health orders were made with more than just data being considered. We also want to see the trend of COVID-19 cases after most of the stay-at-home order levels were implemented. Let us consider two objectives for this project that we refer to as Research Question 1 and Research Question 2.

Research Question 1: Does the number of recent new cases and political party affiliation explain the level of stay-at-home orders?

Research Question 2: Do the near-future cases depend on the decisions of the stay-at-home orders?

We are taking a deeper look into the stay-at-home orders and how each level (1-4) is impacted by factors like political party affiliation and the amount of COVID-19 cases observed during that time, which is before May 2^{nd} 2020. Additionally, we research the trend of COVID-19 cases after May 2^{nd} 2020 and how it was impacted by the stay-at-home order levels and COVID-19 cases leading up to that point.

The average week-to-week percent change of COVID-19 cases are a helpful calculation for analysis and hypothesis testing. Both research questions have been addressed with multiple linear regression methods using new variables calculated by us. We use a programming language called R to address our research questions. RStudio is a user friendly version of R that provided the programming environment we needed to manage the entire data science process, from data preparation to regression analysis.

Regression analysis is a statistical processing method that allows researchers to estimate the relationships between a response variable and an explanatory variable. Response and explanatory variables are more commonly known as "dependent" and "independent" variables, respectively. For our variables, *order* is the dependent variable for Research Question 1 while *after* is the dependent variable for Research Question 2.

It is important to address the first research question because we can provide insight for the stay-at-home orders under the consideration of average weekly percent change of COVID-19 cases and political party. Comparing the stay-at-home order levels, cases, and political

parties (between Democrat or Republican states) can show how the role of politics might intertwine in a public health scenario.

On a similar note, the second research question can inform the public of how the percentage of average weekly COVID-19 cases might change depending on the level of strictness of stay-at-home orders.

Throughout this entire research project, we have kept the replication crisis in science on our minds and have made every effort to make our research reproducible. We hope this project inspires future researchers, data science students, and aspiring statisticians to use their knowledge to make a positive impact on scientific research and society.

Methods

Data Collection Methods

Primarily, we recorded the daily number of new COVID-19 cases for each state in the United States. The data was collected from a time-series visualization made by Google [**7**], which combined several sources together to report the count of new cases from the start of 2020 through 2021 and beyond. Figure 2 depicts a time-series visualization for California made by Google that was utilized for the original data collection process.

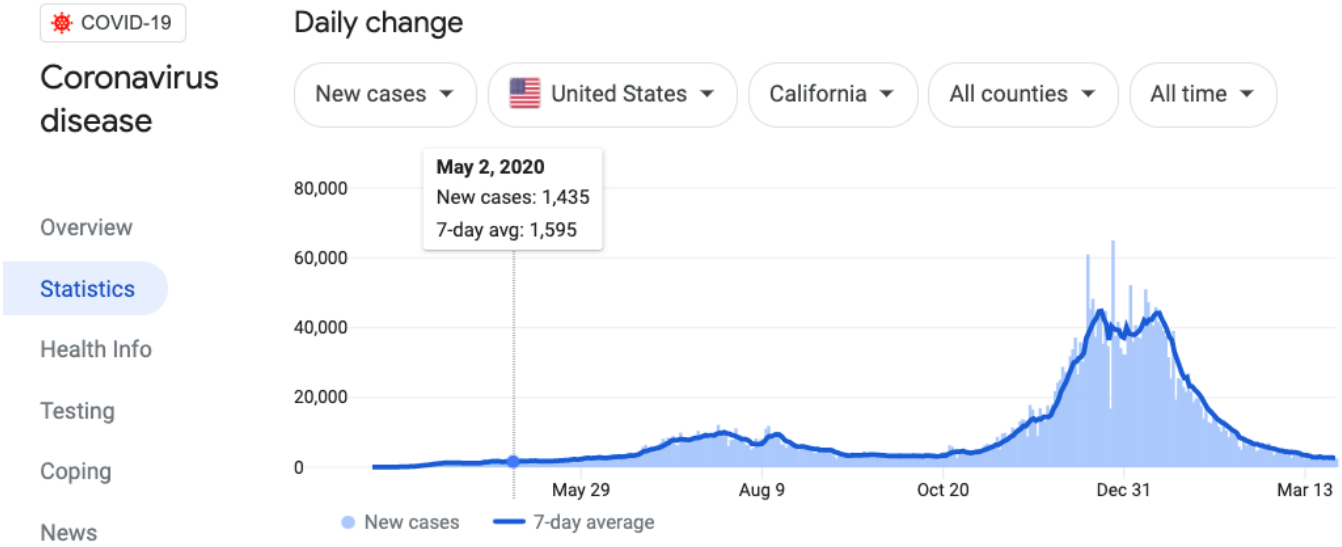

FIGURE 2. Original Data Source [**7**].

Google acquired all of its sources from around the world which are constantly updated such as: The New York Times, Government Health Ministries, Wikipedia and many other authoritative sources [**7**]. The data structure that we created from the time-series visualization consists of States as columns and days as rows as shown in Figure 3 .

	A	B	C	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK
1	Y	M	D	Maine	Maryland	Massachusetts	Michigan	Minnesota	Mississippi	Missouri	Montana	Nebraska	Nevada	New Hampshire	New Jersey	New Mexico	New York
59	2020	5	2	29	1001	1952	851	498	229	319	0	488	112	119	2527	219	3438
60	2020	5	3	33	989	1824	547	435	109	232	1	333	68	89	3027	118	2538
61	2020	5	4	20	946	1000	196	571	327	368	0	424	103	70	1525	181	2239
62	2020	5	5	21	709	1184	447	617	330	162	0	355	69	48	2324	107	2786
63	2020	5	6	28	1046	1754	657	728	217	186	0	333	103	104	1297	153	3491
64	2020	5	7	76	1211	1696	592	786	262	239	2	419	118	103	1745	202	2938
65	2020	5	8	44	1111	1612	680	723	404	148	0	641	84	104	1819	180	2715
66	2020	5	9	34	1049	1410	430	587	288	177	0	403	104	64	1631	105	2273
67	2020	5	10	28	1053	1050	382	239	123	178	1	81	4873	60	1447	85	1660
68	2020	5	11	26	786	669	414	150	173	74	2	257	159	89	1413	206	1430
69	2020	5	12	15	688	870	469	332	234	88	1	120	83	79	798	143	2176
70	2020	5	13	38	751	1165	370	69	182	136	0	341	105	60	817	152	2390
71	2020	5	14	50	1091	1685	1191	7	393	175	4	383	115	83	1144	139	2762
72	2020	5	15	38	1083	1239	497	805	318	139	2	356	95	82	1201	159	2419
73	2020	5	16	45	982	1512	425	729	322	219	0	448	148	92	1184	185	1889
74	2020	5	17	39	836	1077	638	699	173	114	2	128	49	40	1245	91	1250
75	2020	5	18	26	958	1041	773	704	136	156	1	277	140	56	1705	158	1474
76	2020	5	19	28	1784	1094	435	506	535	135	7	221	120	69	974	66	1525
77	2020	5	20	78	777	1046	659	79	255	152	1	276	89	147	1386	155	2088
78	2020	5	21	58	1208	1013	501	8	420	108	0		146	67	1073	155	1696
79	2020	5	22	71	893	888	403	805	363	218	0	237	295	79	1247	153	1772
80	2020	5	23	65	1071	398	452	840	247	194	0	327	74	75	385	170	1589
81	2020	5	24	42	818	280	314	728	206	236	0	145	109	60	1050	148	1249
82	2020	5	25	19	839	205	202	742	223	179	0	221	118	48	938	83	1072
83	2020	5	26	35	535	900	223	645	313	124	2	264	116	34	672	104	1129

FIGURE 3. Original Data Structure

In addition to this data, we used a data-set (stored on Github, a programming collaboration website) from the time-series visualization above. It has the same data we collected but presented as daily new cumulative case count instead of daily new case count.

Data Cleaning

There were several columns and rows that were unnecessary for our work in the data-set found via Github. So, our data preparation consisted of removing rows, removing columns, creating variables, adding columns and creating data frames. To clean our data in this manner, below are a few key functions we used to help clean and create the data-set in Figure 3:

- `data[-c(1, 2, 3), ]`
- `merge(data, data2, by = 'columnName')`
- `cbind(data, columnName)`

which are provided with more detail and comments in Appendix A: R Code.

Note that the final average weekly percent change of COVID-19 cases for each state, named *before* and *after* (before and after May 2^{nd} 2020), was from a data frame of $18,000$ rows. The formula and methods we used to bring these $18,000$ rows of cumulative case data to one summarized percentage per state is mentioned below.

Filter

	state	before	after	party	order	mask
1	Alabama	98.15	17.00	R	2	Y
2	Alaska	141.13	17.37	R	2	N
3	Arizona	126.92	20.14	R	3	N
4	Arkansas	96.05	17.71	R	2	Y
5	California	109.58	15.54	D	4	Y
6	Colorado	119.66	7.38	D	2	Y
7	Connecticut	115.25	3.36	D	4	Y
8	Delaware	104.47	7.28	D	4	Y
9	Florida	123.54	17.86	R	3	N
10	Georgia	125.90	13.28	R	2	N
11	Hawaii	100.51	16.45	D	4	Y
12	Idaho	91.47	17.05	R	2	N

FIGURE 4. Data Structure Needed to Address Both Research Questions
Each of these variables were created with the help of the data-set found via GitHub.

- **State**: The names of the states where data were collected in the United States.

To address the first question, the following variables were needed:

- **Order**: Level of reopening status (1: least strict, 4: most strict)
- **Before**: Average percent change of weekly cases before May 2^{nd} 2020 (1/4/20 - 5/1/20).
- **Party**: The political party per state between Republican and Democrat.

Order is the response variable, while **Before** and **Party** are the explanatory variables. We aim to focus on whether **Order** depends on the amount of COVID-19 cases before May 2^{nd} 2020 and/or the political party of each state. Between **Before** and **Party**, there is a chance that one or both do not have any association with **Order**. So, the results will include how each stay-at-home order has an association with the COVID-19 cases before May 2^{nd} 2020 along with the strength of that association. This will be done by using methods of multiple linear regression.

To address the second question, we needed the following variables:

- **After**: Average percent change on May 2^{nd} 2020 and after (05/02/20 - 8/29/20).
- **Order**: Level of reopening status (1- least strict, 4- most strict)
- **Before**: Average percent change of cases before before May 2^{nd} 2020 (1/4/20 - 5/2/20).

After is the response variable while **Order** and **Before** are the explanatory variables. Our focus here is whether the strictness of **Order** plays a role in the amount of COVID-19 cases after May 2^{nd} 2020. It is likely that **Before** will play no role and could have no kind of relationship with **After**. So, the results will include how the strictness of stay-at-home orders plays a role with the change in COVID-19 cases after May 2^{nd} 2020 and the strength of the

association. This will be done by using multiple linear regression methods.

Percent Change. Taking population proportions into account was pertinent to answering our research questions. For example, the population size of California is about 39 million people, while the population size of Tennessee is 9 million people. The cumulative new cases were reported and collected on a daily basis and varied from state to state based on their population size. So, California's ratio of COVID-19 cases to its population substantially outweigh the proportion of COVID-19 cases to the population in Tennessee. For instance, some states reported as little as no cases a day, while other states reported up to hundreds of cases, daily. In statistics, we do not want to run analysis, descriptive statistics or any hypothesis testing without checking these details. Therefore, we decided that finding a percentage change of the cumulative cases is more appropriate than the cumulative number of cases.

For both research questions, we noticed that most states vary in time ranges for stay-at-home orders, making it difficult to report long term results. The proposed solution to this dilemma was finding a common date where stay-at-home orders took place for each state in our data-set. As previously mentioned, MSNBC's report on May 2^{nd} of 2020 is the common date for our research purposes [**9**]. Using R, we were able to transform our date format $(Y - m - d)$ in the data-set to numeric form of day labeled as 'd.' From this quick analysis, we were able to use the cumulative case numbers from each Saturday before and after May 2^{nd} 2020 to calculate our average weekly percent change. We found that from January 1^{st}, 2020 to May 2^{nd}, 2020, there are eighteen Saturdays, and since we want the weeks before and after the common date to match, eighteen Saturdays from May 2^{nd}, 2020 to August 29^{th}, 2020 were used. The formula used to calculate the weekly percent change for each state is:

$$Weekly\ \%\ Change = \frac{X_n - X_{n-1}}{X_{n-1}}$$

where X_n represents the n^{th} week in the sequence of Saturdays from January 1^{st}, 2020 to May 2^{nd}, 2020 and from May 2^{nd} 2020 to August 29^{th}, 2020. To explain further, if the weekly % change is equal to 0%, then there was no change. If the weekly % change is equal to 100%, then the cumulative cases that week have doubled. If weekly % change is negative then there is a decrease in cumulative cases.

Average Percent Change. To complete proper analysis and start hypothesis testing for our research questions, we calculated the average weekly percent change per state for before and after May 2^{nd} 2020. To do so, we used the definition of the sample average and took the weekly % change values calculated previously. The formula we used to calculate this average is as follows. Let

$$Y_i = Weekly\ \%\ Change$$

for the i^{th} Saturday in the sequence from 01/01/2020 to 05/02/2020. Therefore we have that

$$Average\ \%\ Change = \frac{1}{m} \cdot \sum_{i=1}^{m} Y_i$$

where $m = 18$ Saturdays before May 2^{nd} 2020, which now gives us the "Before" variable. The same formula was used for $m = 18$ Saturdays after May 2^{nd} 2020 to calculate the "After" variable. Having the average % change of "Before" and "After" May 2^{nd} 2020 is useful for comparison purposes and is relevant to answering our research questions.

Statistical Methods

To address Research Question 1, hypothesis testing was formulated based on the multiple linear regression model

$$Orders = \beta_0 + \beta_1 \cdot (\textbf{Before}) + \beta_2 \cdot (\textbf{Party}) + \epsilon$$

where β_1 tells us that for every 1 percentage increase in average weekly cases before May 2^{nd} 2020, there will be a shift in order level, when comparing two states with the same political party. While β_2 is telling us that when we compare a Republican state to a Democratic state of the same average weekly percent change in cases before May 2^{nd} 2020, we expect to see how much of a difference there was between stay-at-home order levels.

Before represents the *Average % Change* in cumulative cases before May 2^{nd} 2020. **Party** represents the political party that each state is affiliated with. If a state is a Democratic State it is represented by a 0, and Republican states are represented by a 1. We removed two states that had split party affiliation, which were Maine and Nebraska. We also removed the District of Columbia, which had "NA" as its political party. They were removed since we are focusing on Democrat or Republican states, and we do not want to assign a party affiliation to these places.

In statistics, a hypothesis is a statement about the population, or in our case, a statement about the model parameter. To test for the association between 'Order' and 'Before', the null hypothesis is stated as $H_0 : \beta_1 = 0$, and the alternative hypothesis is stated as $H_1 : \beta_1 \neq 0$. In addition, to test if the political party explains the level of "Reopening status"/Stay-at-home orders (given the same rate of new cases until May), we have $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 \neq 0$. We conduct a two-tailed hypothesis test for both parameters β_1 and β_2 because, if there is a relationship between 'Order,' 'Before,' and 'Party,' we want to know if it is a positive or negative relationship. During hypothesis testing for both β_1 and β_2, we assume to have a significance level $\alpha = 0.05$ where we'd have a 5% chance in concluding the alternative hypothesis if the null hypothesis is true.

Overall, the adjusted model given above will be used to study the association between the new COVID-19 cases and stay-at-home orders, while adjusting for state political parties.

For Research Question 2, we also use multiple linear regression. The proposed model for Research Question 2 has the same mathematical form as the model for Research Question 1, and is written as

$$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Before}) + \epsilon$$

where β_1 tells us that for every 1 level increase in order level, there is a shift in the percent change of average weekly cases after May 2^{nd} 2020 when comparing two states with the same **Before** value. **After** is the *Average % Change* of cumulative cases after May 2^{nd}

2020, **Order** is a numerical representation of each stay-at-home order, and **Before** is the *Average % Change* of cumulative cases before May 2^{nd} 2020.

However, after running analysis, we found that we needed to alter our model for Research Question 2 by adding a squared term such that

$$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Order.sq}) + \beta_3 \cdot (\textbf{Before}) + \epsilon$$

where $\textbf{Order.sq} = \textbf{Order}^2$, the Order variable term squared. This quadratic model provided plots which addressed our research questions more explicitly than our former linear model. The **Before** variable did not provide any substance in the plot (seen in Figure 6). In the results section, you can see some quadratic pattern in residual vs. fitted plots for each model to be able to better visualize the process. A quadratic pattern indicates that our assumptions may not be reasonable. By adding the squared term, we are able to see the quadratic pattern become more linear. We continue to explain this in the results section.

During hypothesis testing for β_2, we assume to have a significance level $\alpha = 0.05$ where we'd have a 5% chance in concluding the alternative hypothesis if the null hypothesis is true.

With that being said, if there is an association between **After** and **Order.sq**, we expect to see a negative relationship, because as the order level increases in strictness, we assume the average weekly cumulative COVID-19 cases would decrease. Hence we conduct a hypothesis test such that $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 < 0$. Since the **Before** variable does not add anything of notice for either model, we will not be conducting a hypothesis test for β_3.

Results

For the first linear model ($Orders = \beta_0 + \beta_1 \cdot (\textbf{Before}) + \beta_2 \cdot (\textbf{Party}) + \epsilon$) the program R, by default, has Democrats set as 0 and Republicans set as 1 to compare both political parties.

The estimate for β_1 is 0.001761. When we compare two arbitrary states of the same political party, one state should have a higher weekly percent (%) change by 568% in order to expect one higher level of stay-at-home order. The estimate for β_2 is -0.9162 which is approximately -1. When we compare two arbitrary states of the same weekly change before May 2^{nd} 2020, a Republican state would have about one level lower than a Democratic state.
And according to the p-value for political party, $partyR$ ($p = 0.000132$) , there is sufficient evidence to claim that there is a relationship strong enough between political party affiliation and stay-at-home orders to be detected by hypothesis testing.

According to the p-value for **Before** (0.2123), there is insufficient evidence to claim that there is a relationship between the **Order** and **Before** variables and hence, average weekly percent increase before May 2^{nd} 2020 is not a significant predictor. In other words, across the states, the level of stay-at-home orders are not explained by the intensity of the spread of the virus, but by the political party affiliation.

For the second linear model, recall we used the model $After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Before}) + \epsilon$, where $\beta_1 = -1.833155$ and $\beta_2 = -0.007806$. We found that when comparing the level of stay-at-home orders and the average weekly percentage *after* May 2^{nd}, **Order** was a significant predictor. As the stay-at-home order became more strict (going from 1 to 4), the average percent increase in weekly cumulative COVID-19 cases decreased. However, the plot visualizing this result is unable to portray this thoroughly (Figure 5).

Below, Figure 5 graphically assesses whether the mean of **After** is well captured by its multiple linear regression model. In that figure (5), a quadratic pattern is observed, and it motivated us to include the squared term of **Order** so we can attempt to flatten the quadratic curve. So, the updated model we conducted analysis on for our second research question is $\textbf{After} = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Order.Sq}) + \beta_3 \cdot (\textbf{Before}) + \epsilon$. We had adjusted our multiple linear regression model a few times before committing to the updated model in order to present the progress of the quadratic pattern flattening. From our first adjustment of said model, we were given the following plot in Figure 6 such that it progressively flattened compared to the plot in Figure 5.

After our final configuration of our multiple linear regression model, we were able to find a suitable plot which not only helped in addressing our second research question, but allowed us to explicitly interpret the results as well.

We considered removing the **Before** variable when we drew the curve in Figure 7 because it is not statistically significant and it's coefficient is close to zero. Also, **Before** would not have enough sufficient information that could have been detected with hypothesis testing.

After including the squared term and excluding the **Before** variable, the red curve in our plot is fairly straight (Figure 7) which we believe is more appropriate. When building any model in statistics, there will always be some assumptions involved that we must address. If our assumptions are reasonable, then the red curve in a residual vs. fitted plot should be approximately flat. Figure 7 addresses our first two assumptions: (1) that we captured the average weekly percent change of COVID-19 cases after May 2^{nd} reasonably well and (2) that the constant variance assumption is considered.

To put briefly, the constant variance assumption is where the variance of error predicted value is constant. On a plot, it would show the dispersion of data points such that they are similar to a residual vs. fitted model.

Since our curve is flat, then we can conclude that our assumptions are reasonable when adding a squared **Order** term.

That considered, the estimate for β_2 (**Order.Sq**) is -2.4236. When we compare two arbitrary states with the same **Before** value, the difference in the **After** value is calculated as $12.33 - 2.42 \cdot (2 \cdot \textbf{Order} + 1)$ percent per one order level increase. For example, if California and Alabama had the same **Before** value but California had order level 4 and Alabama had order level 3, then β_2 would be $12.33 - 2.42 \cdot (2 \cdot 4 + 1) = -9.45$ percent. In other words, for every one level increase in stay at home order, the average weekly percent change in cases after May 2^{nd} 2020 decreases by -9.45 percent.

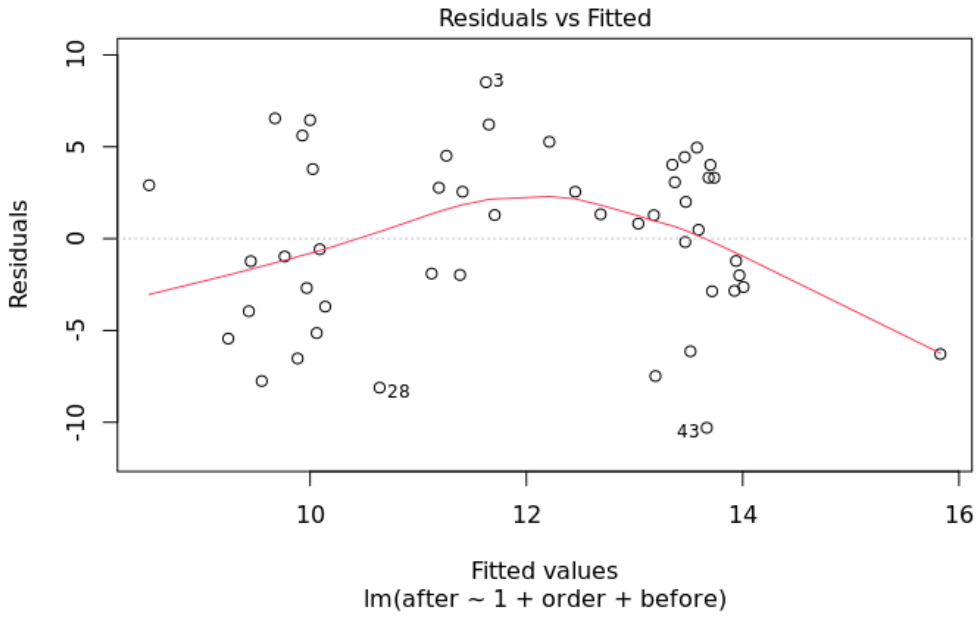

FIGURE 5. Residual Plot vs. Fitted Plot 1
This plot is from our first attempted model for Research Question 2
$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Before}) + \epsilon$.

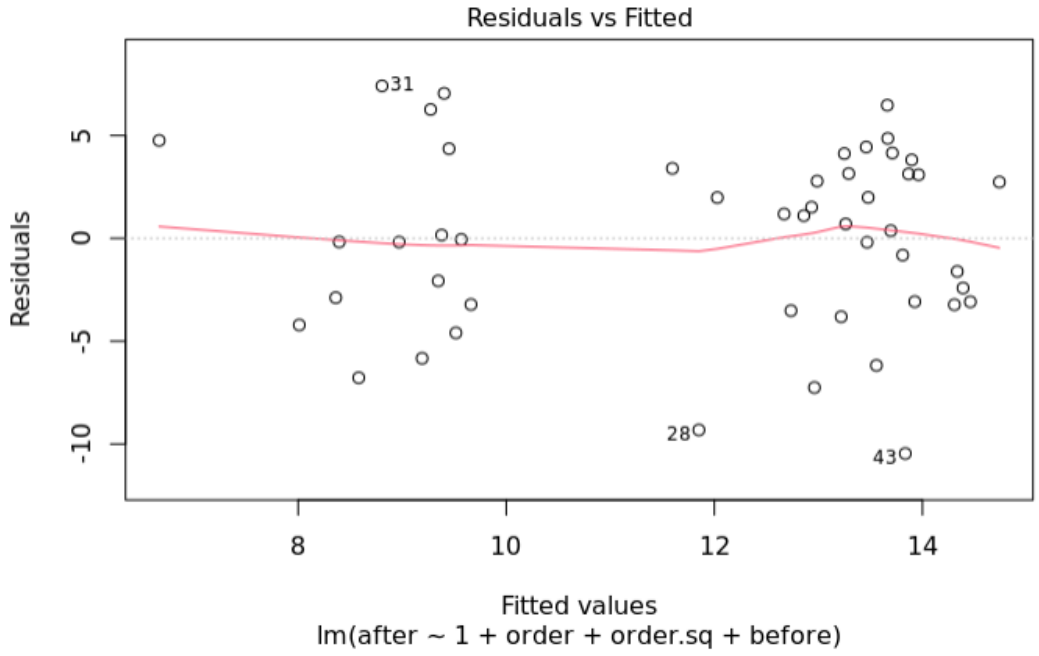

FIGURE 6. Residual Plot vs. Fitted Plot 2
This plot is from our second attempted model for Research Question 2
$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Order.Sq}) + \beta_3 \cdot (\textbf{Before}) + \epsilon$.

We were able to utilize the information we've assessed from our residual vs. fitted plots and create a graph that answers our second research question (Figure 8).

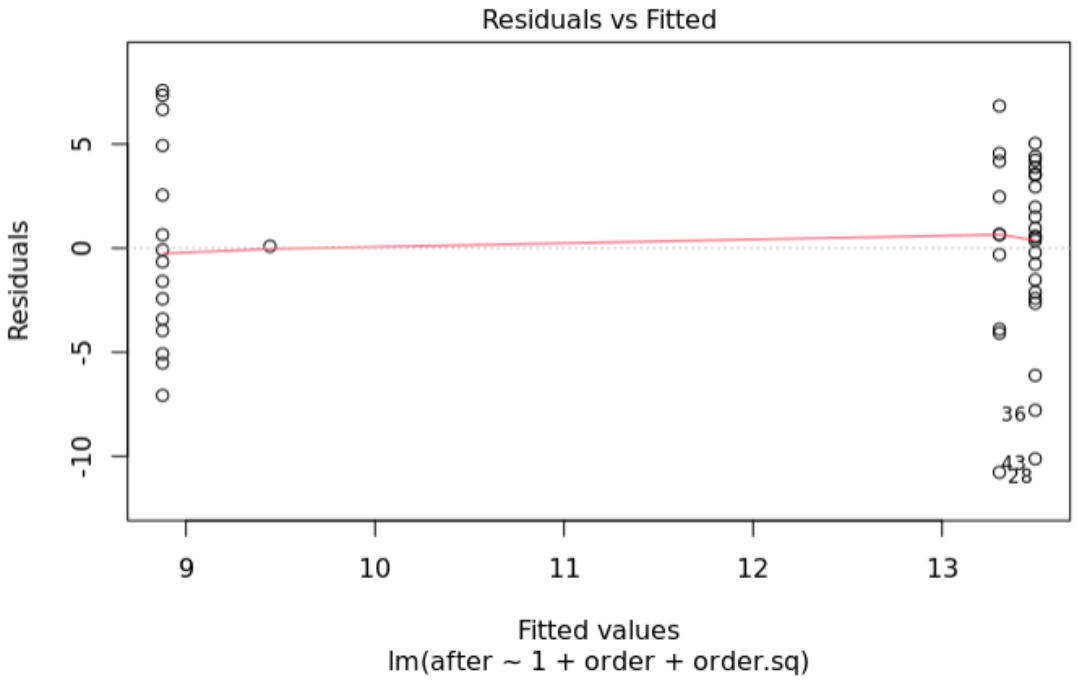

FIGURE 7. Residual Plot vs. Fitted Plot 3
This plot is from our final attempted model for Research Question 2
$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Order.Sq}) + \epsilon$.

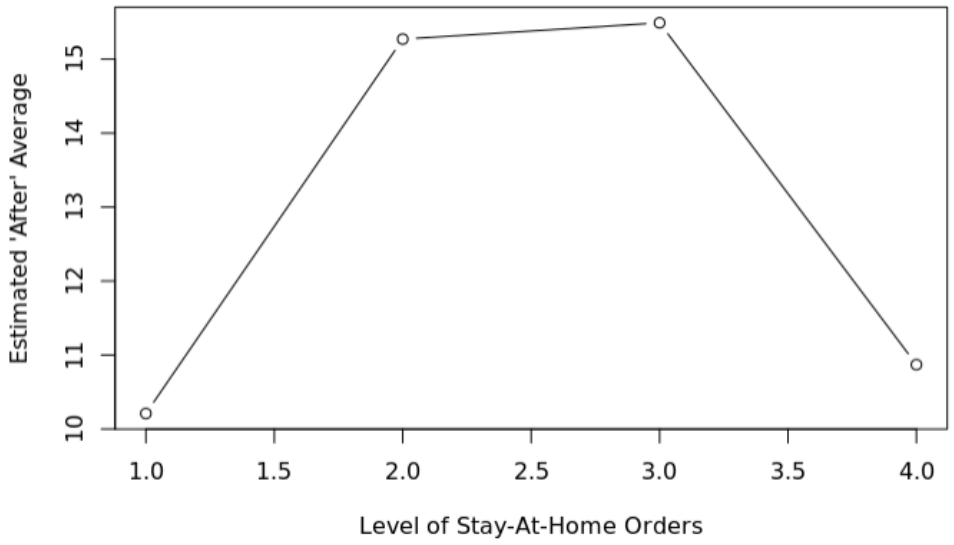

FIGURE 8. Estimated Average of 'After' Variable according to Order Level (1-4)
$After = \beta_0 + \beta_1 \cdot (\textbf{Order}) + \beta_2 \cdot (\textbf{Order.Sq}) + \beta_3 \cdot (\textbf{Before}) + \epsilon$.

Figure 8 tells us that under the order levels 1 or 4, states may have had a low average increase in COVID-19 cases because of two possible reasons:

(1) Level 1: There is only one state (South Dakota), therefore this order level was not as useful as the other order levels.

(2) Level 4: Since the estimated ‘after’ average is lower than level 2 and 3, we can say that having a strict stay at home order was a good decision made by the state.

Let us take into consideration order levels 2 and 3. States probably needed the stay-at-home order levels but did not take any formative action that could have been strong enough to assist their situation regarding the pandemic. So it is possible that this is why the estimated ’After’ average is higher for these order levels. If so, we would suggest that state officials (e.g, Governors) decide to enact order level 4. We cannot effectively compare order level 1 because South Dakota was the only state under this level, which is discussed further in the Limitations section.

Discussions and Conclusions

Discussion

From creating variables to merging data frames, we wanted to ensure we built the best performing models such that they would provide accurate and comprehensible results for both of our research questions. Fortunately, the models we utilized were able to predict a relationship between explanatory variables after performing multiple-linear regression with respect to each model. When performing linear and multiple linear regression, we have found that there is a negative relationship between political party and stay-at-home order levels for COVID-19 cases before most of the stay-at-home orders were set in place (May 2^{nd} 2020). Meaning, Republican States were less strict by one order level when compared to Democratic States of the same COVID-19 situation. From our findings, we can report that sources like *The Brookings Institute* [**3**], hold an accurate claim regarding political affairs and their contribution to public health policies. Lastly, the stay-at-home order levels show an association with lowering the average weekly percent change of cases.

Our intention is to utilize these findings as a pivotal research tool to expand further on the relationships between political party and stay-at-home order levels. It would be interesting to know more about an association between the average weekly percent change of COVID-19 cases *before* May 2^{nd} and stay-at-home order levels. From our estimated coefficient of the **Before** variable in our model for Research Question 1, we saw a very small positive relationship between **Before** and **Order**.

Furthermore, by discovering a negative relationship between **Order.sq** and **After** variables, it raised a question of why states are deciding on stay-at-home order levels 2 and 3. We briefly covered that it may have something to do with state political party affiliations, however, since the **Party** variable is not included in our model for Research Question 2, this is not a definite claim from us.

A variety of widely known public health centers [**14, 15, 16**] report that research has been done on whether certain racial or ethnic groups have been hit harder by COVID-19 [**14, 15**]. Their evidence suggested many reasons of why certain racial or ethnic groups were most impacted by COVID-19, such as residing in multi-generational homes or densely populated areas [**15**]. If this research project were continued, it might be interesting to know how our research would look like if data of racial and ethnic groups were included. Including, possibly, knowing more on how the pandemic may have contributed to the drop in student success rate within certain ethnic groups [**16**].

While a variety of research topics can be carried on from this study, the strengths of this research topic alone can be noted. Since the duration of stay-at-home orders were different for each state, we chose the MSNBC date of May 2^{nd} 2020 to create our new variables. Our strategy of choosing the week to week percentage change instead of using the data as

is was the most efficient way for us to proceed with our research to account for population proportion.

Limitations. Since the start of the spread of the virus and each day after that, states would report the amount of their infected citizens by end of day or end of week. During the data collection process, we noticed inconsistency in the reported COVID-19 case values. One example being that if one state would report that there were 30 new cases one day, then when checking this same day weeks later, this number would have changed to 50. We were not able to update each reported COVID-case value as they changed. Therefore, realizing that many states often change the new-case data, possibly to fix past errors, was the first limitation the we encountered. Because of this, we could not account for future changes to the 2020 COVID-19 new cases data. This turned out to be helpful because the dates from this data began in March of 2020 whereas the new data-set we used had a starting date of January 2020.

Another limitation was that we needed to take out Nebraska, Maine, and the District of Columbia from our data-set. This is because we decided to focus on the two political parties: Republican and Democratic. These three places had a political party titled as 'S' or 'NA'. We did not want to assign their political party affiliation ourselves.

The states that were omitted from our current data-set are able to be placed back easily, if necessary as we utilized a very simple algorithm which can be learned about in Appendix A.

When taking into account the low average increase in cumulative COVID-19 cases under order level 1 (Figure 8), it is important to note that there was only one state in our data-set that committed to order level 1 (South Dakota). That is why the claim for lower average cumulative COVID-19 cases cannot be compared effectively to the low amount of cases under order level 4.

When future research is performed, there may be more up-to-date data-sets that can be derived from reliable sources. We hope that future COVID-19 data-sets contain new-case numbers that are as accurate as possible.

Conclusion

Our box-plots in Appendix B show the difference before most of the stay at home orders were in place and after most of the stay at home orders were in place. Before May 2^{nd} 2020, we see that most states, either Democrat or Republican, approximately double (median is approximately 125 percent) their amount of cases. After May 2^{nd} 2020, most states, either Democrat or Republican, show below 20% for their average weekly percent change of cases. This box-plot shows a significant decrease in average weekly percent change of cases after most of the stay at home orders were implemented.

Our results from Research Question 1 show that Republican states had a less strict stay-at-home order by approximately 1 level on average when comparing Republican and Democrat states under the same COVID-19 situation *before* May 2^{nd} 2020. We found that there is an association between stay at home order level and political party affiliation. We did not find an association between stay at home order level and the COVID-19 cases observed before May 2^{nd} 2020.

For Research Question 2, we also have enough evidence to claim that *after* May 2^{nd} 2020, the average weekly COVID-19 cases seemed to have decreased as the stay-at-home order levels became more strict, but this is not accounting for order levels 2 and 3. In other words, a low value of "after" for order level 1 and 4, and a high value of "after" for order level 2 and 3, may be interpreted as:

It is not enough to place a stay-at-home order mandate with levels 2 and 3 to effectively protect the public from the virus. A strong unified action and order (e.g, order level 4) was needed to control the spread of the coronavirus.

Bibliography

[1] MSNBC: Reopening America. *All 50 states have begun to reopen. See what that means for your state.* 2020. MSNBC Map Resource.

[2] J. R. Zaller. *The Nature and Origins of Mass Opinion* 2012. Zaller Book.

[3] The Brookings Institute. *Politics is Wrecking America's Pandemic Response* 2020. Brookings.

[4] The New York Times. *On Politics: stay-at-home? States Can't Agree* 2020. New York Times.

[5] Proceedings of the National Academy of Sciences. *Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States* 2020. PNAS.org.

[6] American Psychological Association. *Politics is personal: Research by political psychologists helps to explain why we vote the way we do—and is informing ways to improve democratic elections* 2019. APA.

[7] Google Data Source. *Coronavirus Disease 2019.* 2019. Google.

[8] Jiachuan Wu et. al. *Reopening America.* 2020. NBC News.

[9] MSNBC News. *Texas Starts to Reopen As Coronavirus Deaths Hit Single-Day High* 2020. MSNBC Reopening Status Report.

[10] Harvard Public Health Review. *Political Affiliation and Human Mobility Under Stay-at-Home Orders: A Difference-in-Difference Analysis with County and Time Fixed Effects* 2020. Harvard PH Review.

[11] John Hopkins News. *Link found between state governors' political parties and COVID-19 case and death rates: Analysis finds that states with Republican governors had higher case and death rates beginning last summer through 2020* 2021. JHU Hub.

[12] National Institutes of Health. *Associations between governor political affiliation and COVID-19 cases, deaths, and testing in the United States* 2021. NIH Article.

[13] Taylor and Francis Online. *The importance of policy narrative: effective government responses to Covid-19* 2020. TANDF.

[14] Center for Disease Control and Prevention. *Health Equity Considerations and Racial and Ethnic Minority Groups* 2021. CDC.

[15] Mayo Clinic: Coronavirus infection by race: What's behind the health disparities? *Why are people of color more at risk of coronavirus complications?* 2020. Mayo Clinic.

[16] McKinsey & Company. *COVID-19 and learning loss—disparities grow and students need help* 2020. McKinsey Source.

[17] Office of Governor Gavin Newsom. *State Issues Limited stay-at-home Order to Slow Spread of COVID-19* 2020. CA.gov.

[18] American Association of Retired Persons. *Blacks, Hispanics Hit Harder by the Coronavirus, Early U.S. Data Show* 2020. AARP Source.

[19] Cleveland Clinic *Coronavirus, COVID-19: What Is It, Symptoms, Causes & More* 2020. Cleveland Clinic Source.

[20] National Public Radio *What Do Coronavirus Racial Disparities Look Like State By State?* 2020. NPR Source.

Appendix A: R Code

- `data[-c(1, 2, 3), ]`
  Simple function that can remove rows from data frame. Can be easily reversed if necessary.
- `merge(data, data2, by = 'columnName')`
  A function to combine two data frames by a specific column.
- `cbind(data, columnName)`
  Function that allows user to add specific column to a data frame.

For more information regarding R files used for this project, you may click here.

Appendix B: Graphs

In Appendix B, we show the graphs used for this study.

Below, we observe Figure 9, a box-plot that shows the average weekly percent increase of cases before May 2^{nd} 2020 for both political parties. We observe the average weekly percent increase of cases on and after May 2^{nd} 2020 for both political parties in Figure 6.

We also observe our Model Assumptions in Figure 11 and 12. To summarize, we want to see that the residuals of our model follow a normal distribution, and the Normal Q-Q plot can be used to determine that. Since the residuals (data points) in Figure 11 formed along the dotted line, our model for Research Question 2 met this assumption.

For Figure 12, we want to see that each vertical line is below 0.5. Since all data points have a Cook's Distance of below 0.5, there are no outliers that significantly affect our estimates, which is another assumption we must look out for when building models.

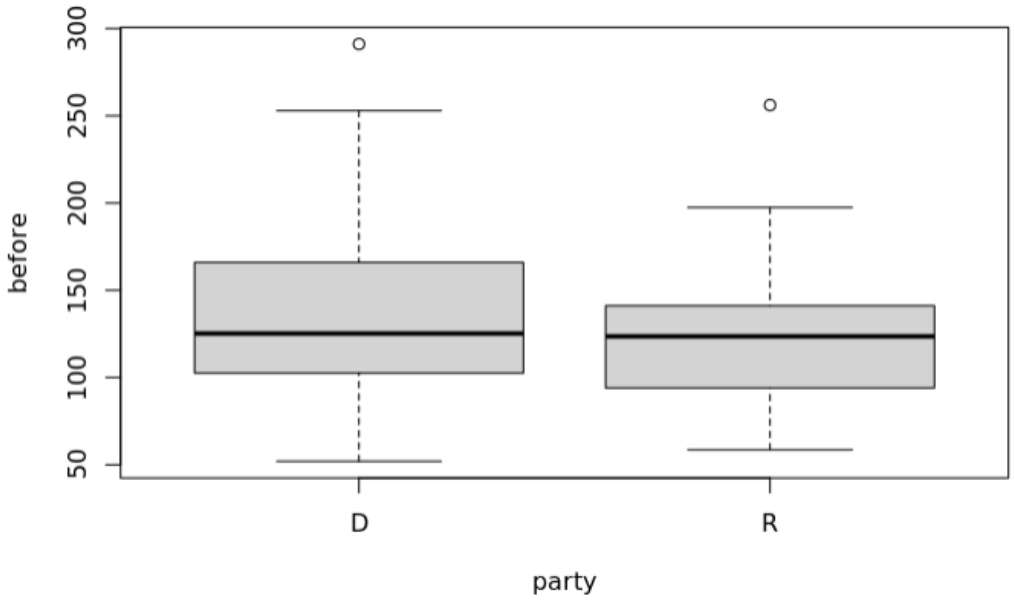

FIGURE 9. Average weekly percent increase of cases before May 2^{nd} 2020 for both political parties.

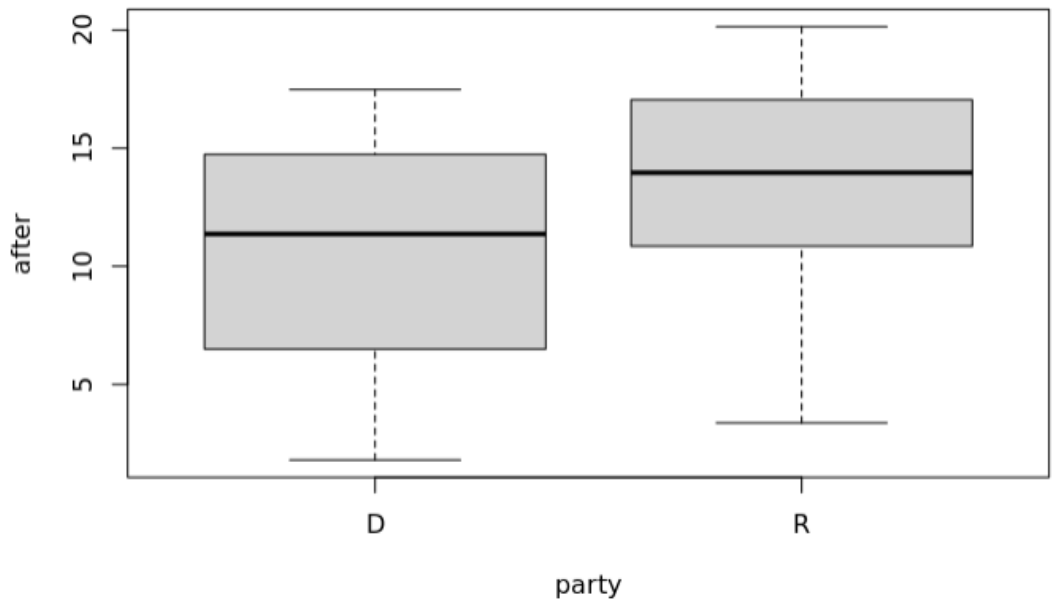

FIGURE 10. Average weekly percent increase of cases after May 2^{nd} 2020 for both political parties.

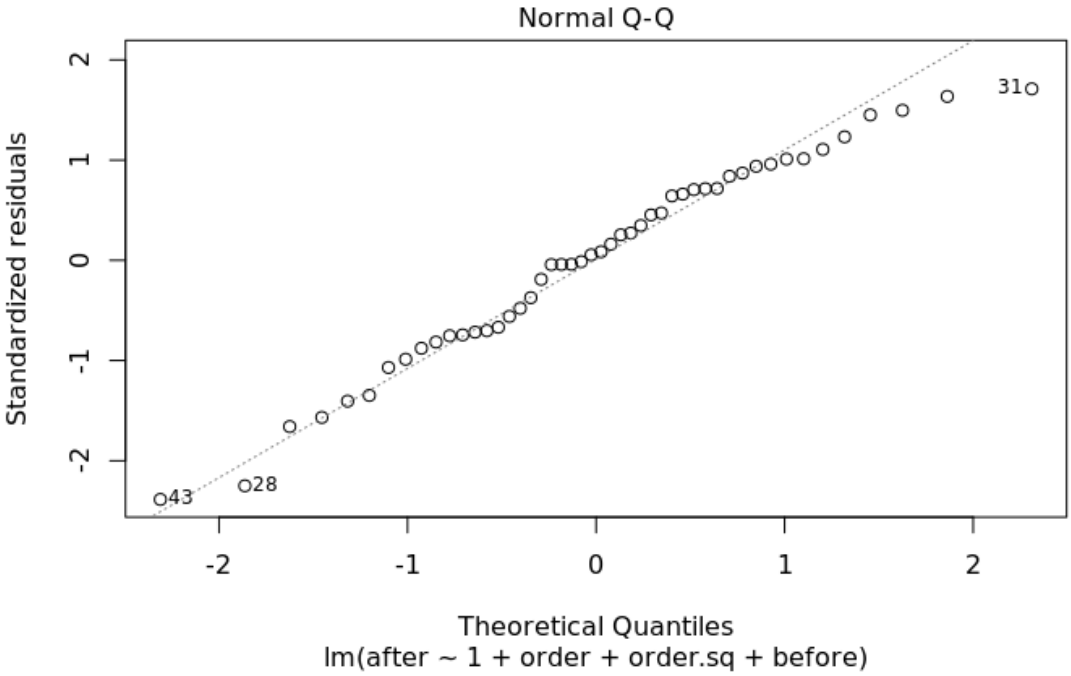

FIGURE 11. Model Assumption: Normal QQ Plot with respect to second model for Research Question 2.

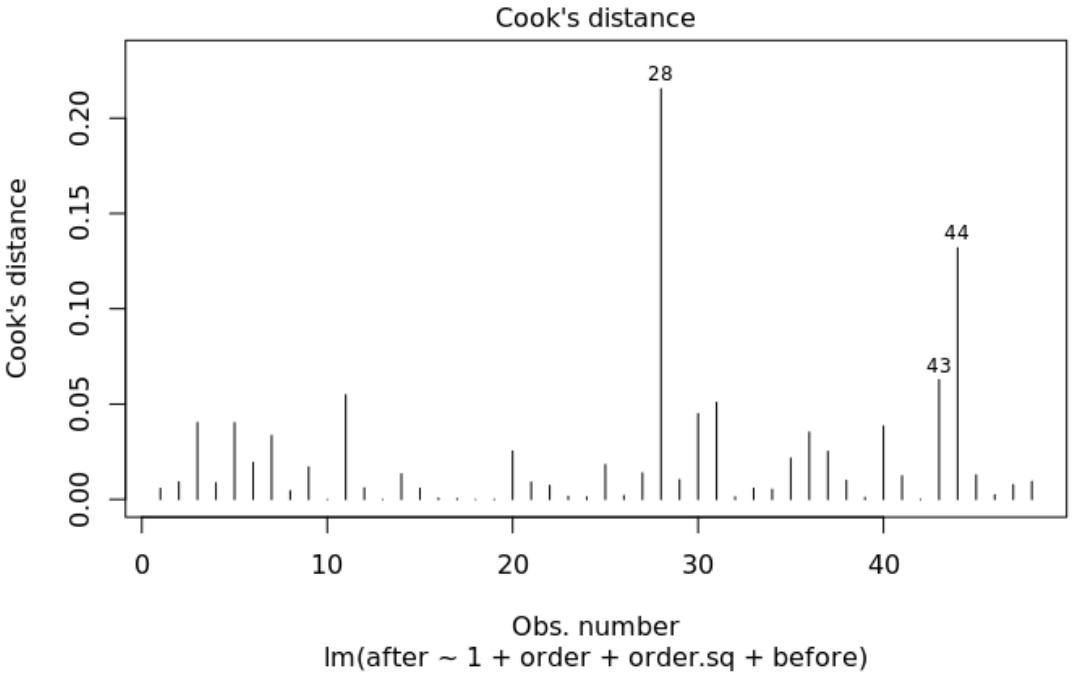

FIGURE 12. Model Assumption: Cook's Distance Graph with respect to second model for Research Question 2.