COVID vaccination efficacy. A study of the United States[*]

Isaac Ng

April 27, 2022

Abstract

The most recent pandemic which started in 2019 has had many controversial events related to its vaccines. Since the vaccine rollout occurred quickly after the pandemic took shape, many had reason to doubt its efficacy. This paper uses data collected by the CDC to compare the percent change in vaccinated population with the the ratio of known COVID-related deathsand cases. The resulting model fitting the data shows a logarithmic relationship between these two rates. Thus, it is imperitive to continue encouraging everythong to become vaccinated as it not only leads to reduced COVID-19 cases, but it reduces the severe outcomes of death from COVID.

Keywords: COVID-19, United States, CDC, vaccines, death

Contents

[*] Code and data are available at: LINK.

1 Introduction

1.1 Inspiration

Throughout history, many of nature's deadliest diseases have taken out hordes of Earth's population. The most notable would include TB, the Black Plague, and most recently, COVID-19. In ancient times, various forms of medicine arose and many claimed to have cures for these diseases. While some have been scientifically proven to be effective, other methods and practices are more questionable. One which is heavily relied upon and has been deemed effective is that of vaccines. According to the WHO, we can define vaccination as a simple, safe, and effective way of protecting yourself against harmful diseases, before you come into contact with them. The vaccines use your body's natural defenses to build resistance to specific infections in order to make your immune system stronger. Vaccines help your body's natural defenses build up protection by recognizing the germ, producing antibodies to fight them, and remembering the disease and how to fight it (WHO August 2021). While vaccines have been proven time and again to be effective in tackling these diseases, many who are stubborn refuse to believe so, citing unreliable sources from the internet.

This paper focuses on COVID-19 and the relationship between the percentage of the population with their first dose of the vaccine and the ratio between deaths and cases. Since a total number would not be able to take into account of population size, taking deaths and cases known to be affected by COVID-19 and comparing the two will suffice. Rather than involving the science behind vaccinations and their various sources, this paper finds a logarithmic relationship between the percentage of the population vaccinated with one dose and the ratio between new deaths and cases. Since the CDC is a reliable source of information, the data collected there can be relied upon for further analysis. Furthermore, with the United States being a neighbouring country to Canada, the relevance of this paper is further magnified.

1.2 Overview

In the following sections, I first discuss the Data 2 used in this paper. I will start by discussing the various resources used in this paper. Then I discuss where the data was gathered from and how the data was collected. Then the variables used in the rest of the paper are stated along with why I chose them. Finally, I will show tables and plots of the raw data to lead into how I chose my model. In Model 3. I talk about the model used and how I came up with it. In Results 4, I use plots and tables to verify that the model works with the data. In the Discussion 5 section, I discuss the paper, giving a summary of the data collected, any weaknesses and next steps involved with my study, and concluding remarks discussion the implications of this paper.

2 Data

2.1 Datasource

For this paper, data is analyzed using R, a programming language for statistical computing and graphics (R Core Team 2022). The package tidyverse is used to help manipulate the data (Wickham et al. 2019). Since we use R projects to manage our data, we use here (Müller 2020) to reference file locations. bookdown (Xie 2021) is used for the formatting of the project. In order to keep tables and figures in place, package float (**float?**) is used. To include a nicely formatted table, kable and kableExtra (Zhu 2021) is used. To help manipulate the data further, the package dplyr was used (Wickham et al. 2022). For tables and models, I used modelsummary (Arel-Bundock 2022), lmtest (Zeileis and Hothorn 2002), broom (in the helper function of the script file) (Robinson, Hayes, and Couch 2022), and ggplot2 (Wickham 2016). Forcats (Wickham 2021) is used for reordering the x axis for a bar graph. The data used in this paper was gathered by the CDC; the data for cases and deaths can be found here (Disease Control and Prevention April 27, 2022) and the data for vaccination rates can be found here (COVID-19 Vaccination Trends in the United States and Jurisdictional April 26, 2022).

2.2 Data Collection

The dataset for cases and deaths is collected by the Centers for Disease Control and Prevention, COVID-19 Response. The dataset is structured to include daily numbers of confirmed and probable case and deaths reported to CDC by states over time. The states include the 50 states as well as other territories such as Washington DC, Guam, Palau, and Puerto Rico. Because these provisional counts are subject to change, including updates to data reported previously, adjustments can occur. These adjustments can result in fewer total numbers of cases and deaths compared with the previous data, which means that new numbers of cases or deaths can include negative values that reflect such adjustments. (Disease Control and Prevention April 27, 2022)

The dataset for vaccinations is also collected by Centers for Disease Control and Prevention, COVID-19 Response. The dataset is structured to include daily numbers of vaccinated people by states over time. The dataset is updated daily after review and verification between 1:30pm and 8:00pm EST. The data is received by the CDC at 6:00am every morning and comes from all 50 states as well as territories (including Puerto Rico, Washington DC, Palau, etc.). (COVID-19 Vaccination Trends in the United States and Jurisdictional April 26, 2022)

2.3 Variables

In this paper, the Administered first dose by percentage population is used. With data including the first, second, and third (booster) doses, the first dose percentage would be the most evident indicator. Comparing deaths and cases to the percentage of people with one vaccine would provide insight on how effective the vaccines are. As a first step, this paper introduces the first dose reports to find a relationship between being vaccinated and the deaths/cases ratio. Since the paper is centered around being vaccinated, first dose is the attribute I am looking for since it encompasses people with their second and third doses. With the case and death data, new_case and new_death columns were selected as they indicate the number of COVID-19 related cases and deaths at a given date.

As stated above in the Data Collection section, some of these values can be negative in order to adjust for updates in the data. As this is the case, aggregating the values over months and seasons would be the most effective as the updates would balance themselves out by summing a larger value with a negative value.

2.4 Plots

In the first plot 1, we see the total number of people vaccinated and the total number of cases for each state. The second plot 2 compares the total number of people vaccinated with the total number of cases for each state. Note that the relationships seems to have some sort of exponential here. Now let us look at the various vaccination rates as compared to their date/time in plot 3. Here there are two colours, one for the total accumulated percentage of the population being vaccinated. The other is for the change in percentage over each month. The next plot 4 is from raw data which looks at the number of new cases and deaths over time.

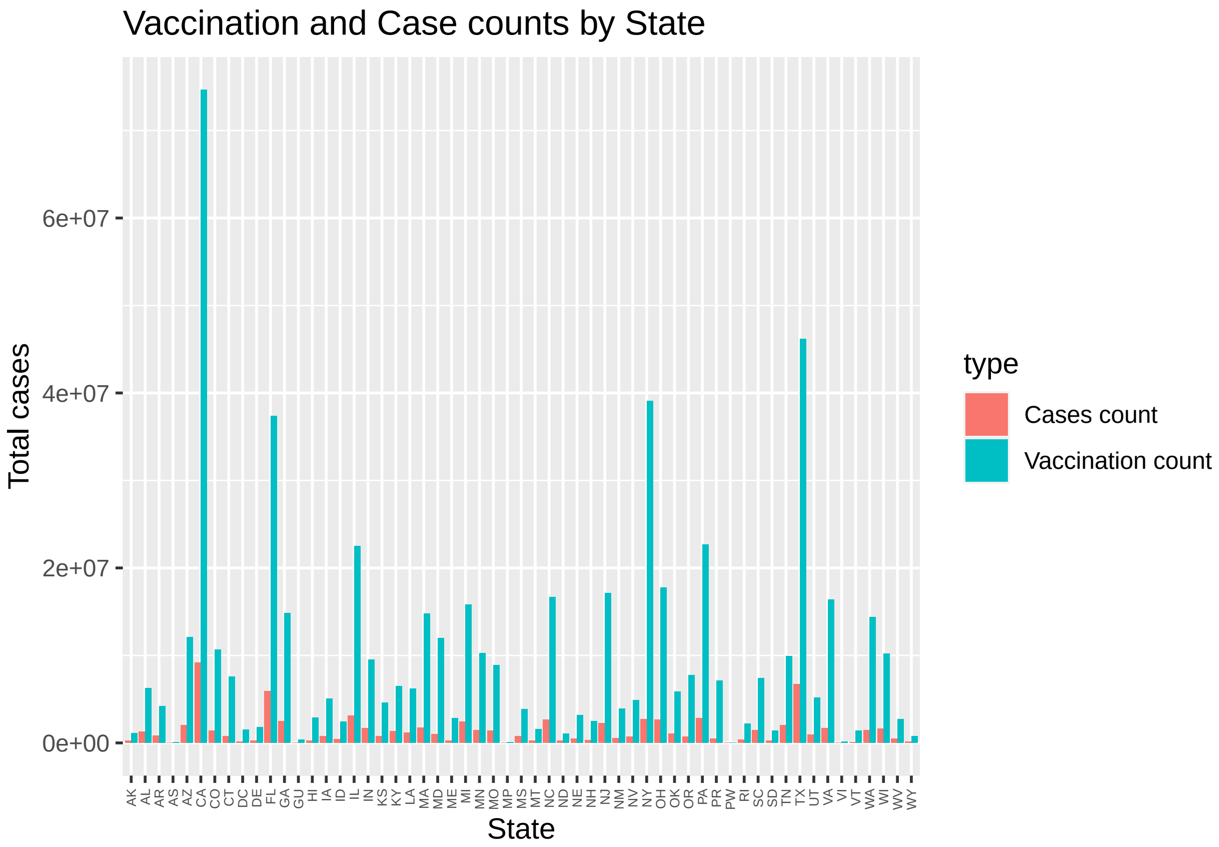

Figure 1: Vaccination and Case counts by State

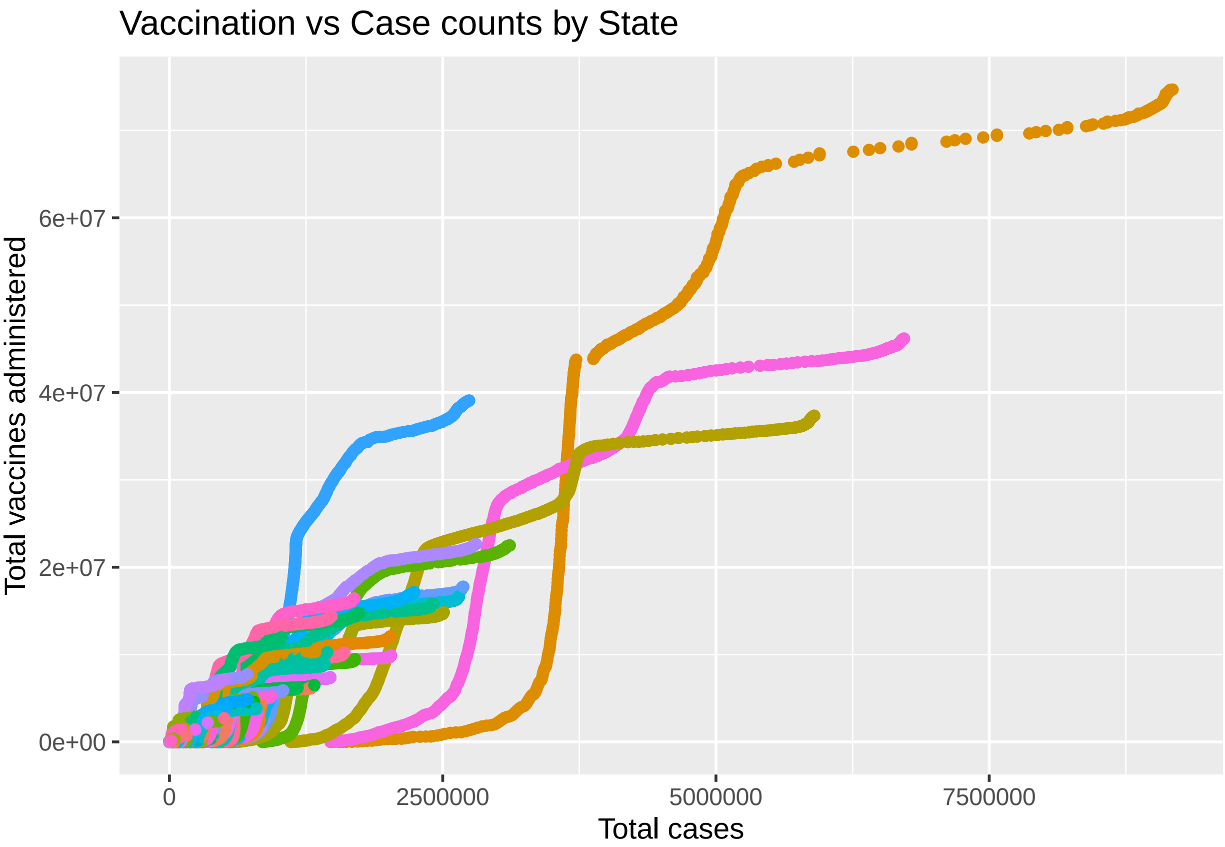

Figure 2: Vaccination vs Case counts by State

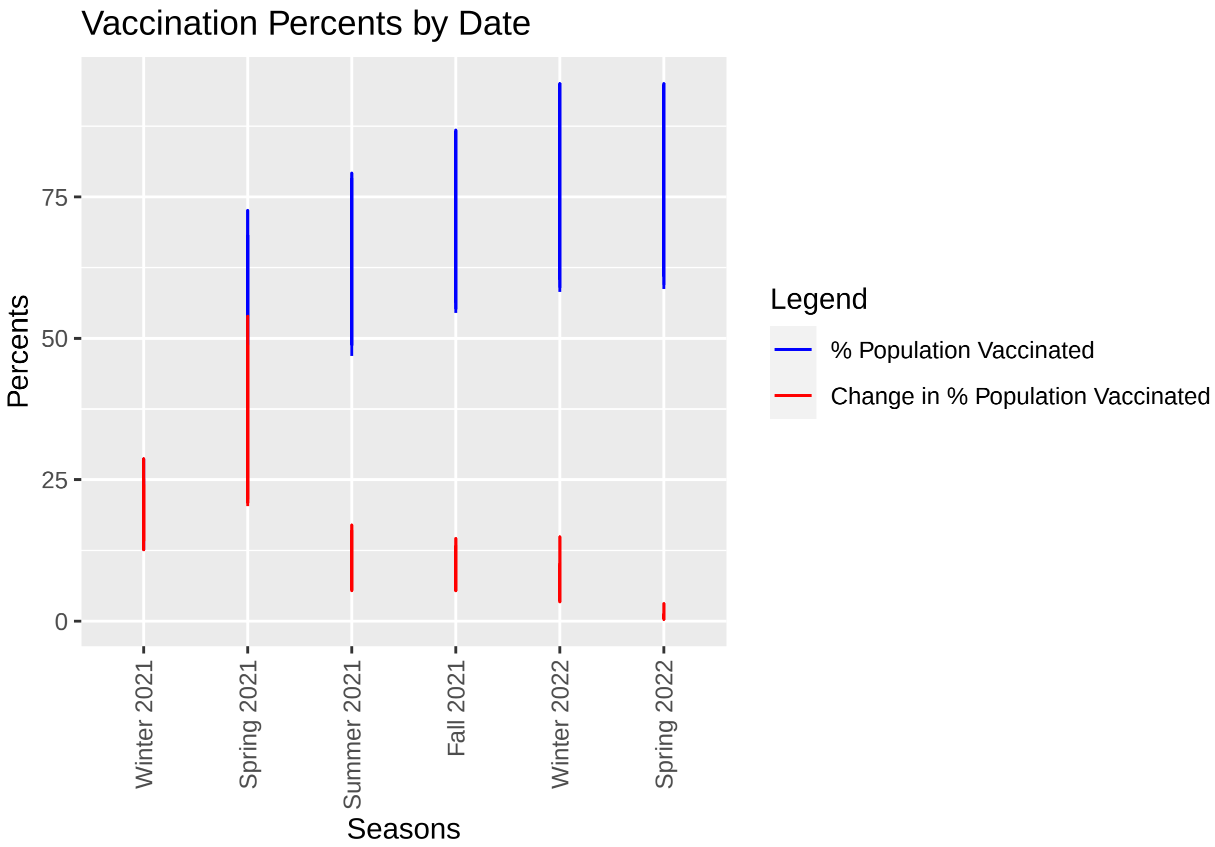

Figure 3: Vaccination Percents by Date

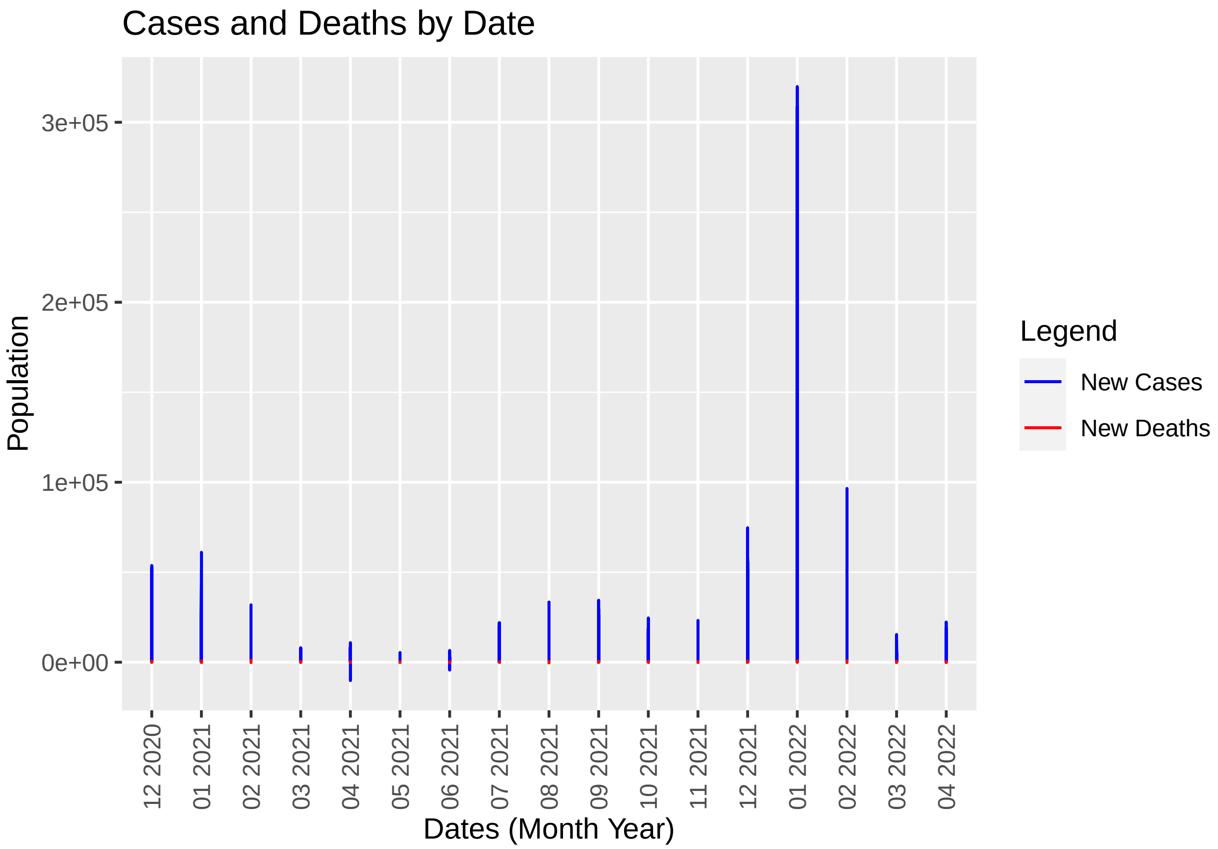

Figure 4: Cases and Deaths by Date

3 Model

In order to find some relationship between the data, we can use the pairs function to see the correlations between all the data we have gathered 5. In this case, we can visually see that there is some sort of logarithmic relationship between change_in_perc_first and new_case as well as new_death. This is can be more clearly seen in the plot between dtc and change_in_first

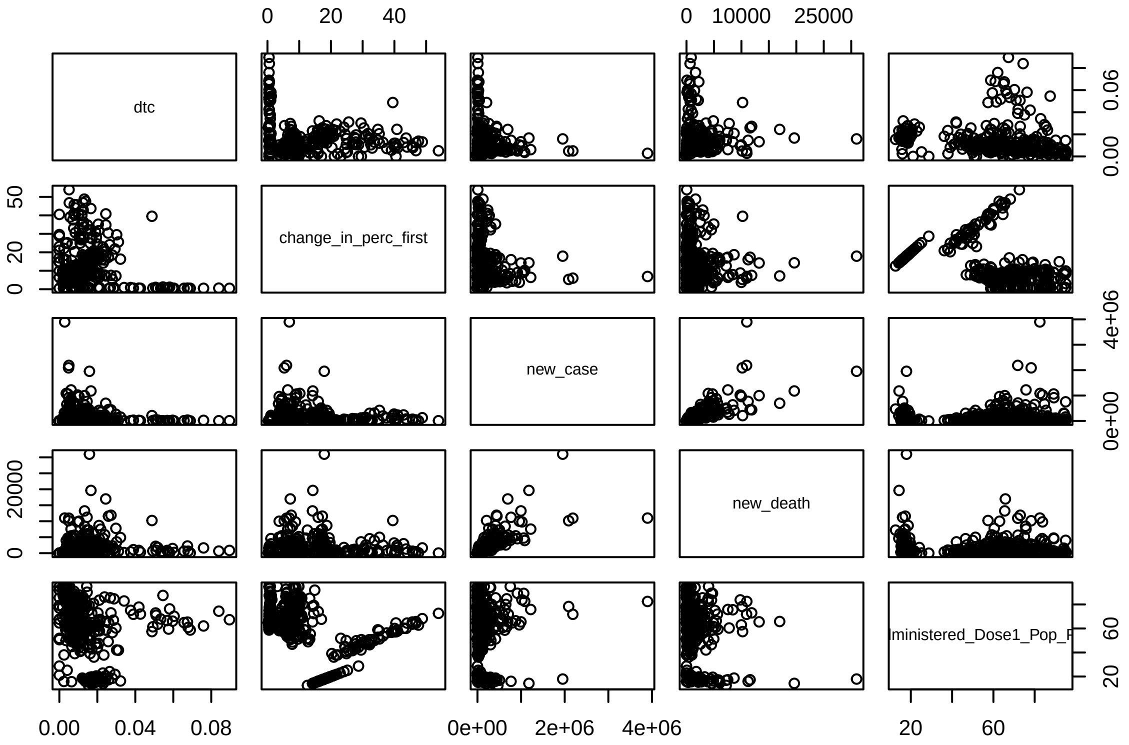

Figure 5: Scatterplots for all pairs

The model we want to use is a logarithmic model that also takes into account each state as a factor. X1 represents the seasonal change in percentage of each state's population being vaccinated. The Xi in the summation represent each state being factored into the equation. The equation would look as follows:

$$\text{dtc} = \beta_0 + \beta_1(log(X_1)) + \Sigma_{i \in State}\beta_i X_i + \epsilon$$

4 Results

The summary for this model is massive as there are more than 50 entries for State. Here is a nicely summarized table of the more important factor, change_in_perc_first alongside relevant values such as F-value, Multiple and Adjusted R-squared, and the Residual values ??. For a further detailed summary, take a look at the file 01-model.R in folder scripts. Here we see that the p value is very small, rejecting the null hypothesis. The residuals are also very small (less than 0.1) which indicates that the prediction is not far off from the actual values. The R squared and adjusted R squared are high but not as high as I would have expected. The F value is quite a large value. Altogether, we see that the chosen model is quite good. In the model script, I also found that there are 16 outliers. Since there are only 16 outliers our of 337 values, as well as having the maximum residual being so low, it is not necessary to remove them from the dataset.

p-value of log(change_in_perc_first)	f-statistic	R-squared	Adjusted R-squared	Max residual
3.97e-13	10.27272	0.6810765	0.614777	0.05296441

To further validate the model, let us look at the diagnostic plots for this model 6.From the Residuals vs Fitted graph, the line is relatively horizontal and the points seem scattered quite randomly, indicating a null plot and that the model is good. In the scale-location, the same null plot is seen, with no kind of trend in the graph, also indicating that the model is good. The Normal Q-Q plot does not concern me as there are over 300 points and most of them lie along the diagonal line. We cannot even see the Cook's distance lines, indicating that the outlier points are not influential.

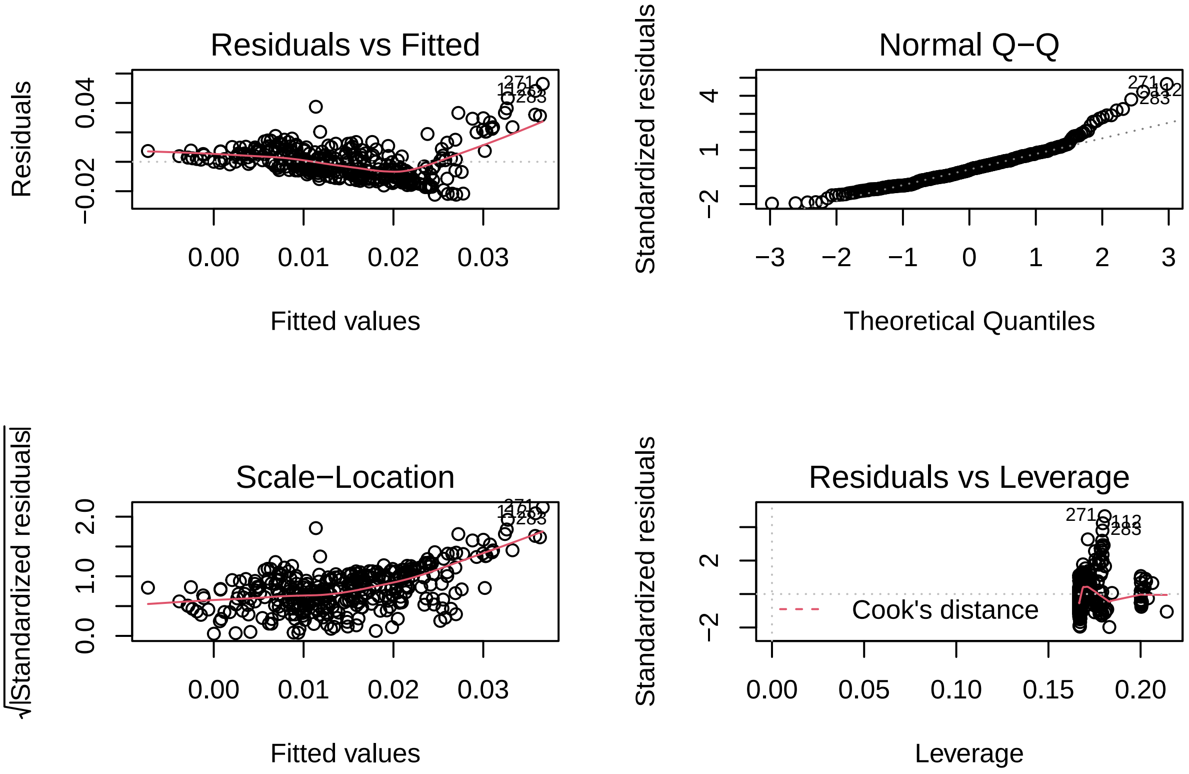

Figure 6: Diagnostic plots

5 Discussion

5.1 Summary

Using data taken from the CDC, available to the public, we showed that there is a logarithmic relationship between the death to cases ratio and the seasonal change in the population being vaccinated. After cleaning the data for NAs and organizing it in a way that is easy to manage, the data was fitted to a model and the model was proved to be valid using the summaries to said model. The logarithmic relationship indicates that as the death to cases ratio decreases quickly, the change in percentage of vaccinated population increased quickly. Conversely, when the death to cases ratio increases quickly, the change in percentage of vaccinated population decreased quickly. In other words, when more of the population becomes vaccinated, fewer and fewer people die from contracting covid and when there is a slower increase in the vaccinated population, more people are likely to die from COVID.

From the raw data, we see that the states with more vaccinations also have higher case counts. The states in question are California, Texas, New York and Florida, all of which have a larger population. When comparing the total number of people vaccinated with the total number of cases in a state, there is a logarithmic relationship, hinting towards which model should be used. When we plot each attribute against each other 5, we see that there is a linear trend between new_case and new_death, which is to be expected. There is also a linear trend between the change in percentage with the actual percentage of the population vaccinated, which is also to be expected. There is a logarithmic relationship between the change in percentage and new case as well as new deaths. However, when we compare the change in percentage with the death to cases ratio, the logarithmic relationship becomes more evident. The resulting summary tables of the model support the model being a good fit with the data. There is a low p value for all variables used, a high F statistic, relatively high R squared and adjusted R squared values, and low residuals.

5.2 Implications

The unwillingness of the remaining unvaccinated population to become vaccinated has become evident through the media. Massive protests and riots have occurred all over the world and have become somewhat uncontrollable. In Canada, we saw the truckers protesting against the mandates set by the government and in America, many even doubt the existence of COVID-19. With more and more people dying from this disease every day, studying these numbers and comparing the increase in deaths to the increase in vaccinated persons will hopefully provide further incentive to become vaccinated. Through this paper, it has become evident that when more of the population becomes vaccinated, less of the population dies from COVID.

Concerns arising the vaccine have surfaced throughout the timeline of COVID. Some fear the reaction they face when being vaccinated, having prior incidents with other vaccines. Others have had milder reactions to the first or second dose but fear a stronger reaction in a future one. While these fears are rational and aim towards self preservation, Maslow has perfectly stated that the needs of the many outweigh the needs of the few. Having an increased population vaccinated will severely reduce the number of deaths that occur and the number of cases that take place in each country.

5.3 Weaknesses and next steps

One weakness of this paper comes from the dataset itself. Many of the cases and deaths data show 0 new cases or deaths in a given day. It appears to be non-zero about once every 7 days. Although this is an accurate number, having accurate daily counts would be much more beneficial as we would be able to tell when exactly there are no new deaths. Another weakness of the dataset is the lack of information regarding the new cases, new deaths and their respective pnew cases, and pnew deaths. One is the confirmed new cases/deaths while the other includes only the probable new case/deaths. In order to have a fuller picture, a definition of where the line between confirmed and probable needs to be defined.

As a next step, aggregating over month or even by day could provide better insight on the trends in the data. Having more locations, such as aggregating by country, could also greatly benefit this study into COVID-19. By introducing multiple geographic areas, vaccination efficacy can even be compared to population density or different climates. In order to see the efficacy of future/multiple vaccination shots, another next step could be to analyze the second and third shots. Those can be compared or further added into the model in hopes of finding a correlation between receiving more vaccination shots and case numbers.

Another possible step is to introduce a time lag into the equation. As it is unclear how much time should be lagged, this paper aggregated by season to aquire a fuller picture however, given more information and research, introducing a time lag could also benefit this study. The time lag would coincide with the amount of time it takes for a vaccine to become effective. Likewise, another time lag that could be introduced is the average time it takes for COVID to take a person's life. Given more information, these time lags could be introduced and a more detailed version of the data can be used to further examine relationships between the vaccination rate and new cases/deaths.

5.4 Conclusion

COVID-19 is a recent disease with plenty of unknowns that has spread across the globe. Not long after it surfaced, vaccines from Pfizer and Moderna were released to the public in order to combat the disease. Naturally, many became skeptical with these medicines for one reason or another. This paper has proven that there is a logarithmic relationship between the change in vaccinated population and the rate of deaths relevant to COVID cases. In other words, when a greater portion of the population becomes vaccinated, fewer people die from these cases. Not only do these vaccines aid in preventing the contraction of COVID, but it clearly reduces the disease's severity to those who do manage to get it. As a result, the general population should be reassured by the vaccines mandated by the government and is recommended to follow what has been instructed regarding vaccinations.

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://vincentarelbundock.github.io/modelsummary/.

COVID-19 Vaccination Trends in the United States, National, and Jurisdictional. April 26, 2022. *United States COVID-19 Cases and Deaths by State over Time.* https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2.

Disease Control, Centers for, and COVID-19 Response Prevention. April 27, 2022. *United States COVID-19 Cases and Deaths by State over Time.* https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.*

R Core Team. 2022. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles.*

WHO. August 2021. *Vaccines and Immunization: What Is Vaccination?* https://www.who.int/news-room/questions-and-answers/item/vaccines-and-immunization-what-is-vaccination.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Forcats: Tools for Working with Categorical Variables (Factors).*

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation.*

Xie, Yihui. 2021. *Bookdown: Authoring Books and Technical Documents with r Markdown.* https://github.com/rstudio/bookdown.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.*