ECO 322 Project 1 : COVID Data Analysis
Group Members: Jesse Freitag, Charlie Clark, Kailey Ali, Gavin Vergara, Qiushi Yin
Stony Brook University - ECO 322: Data Science and Machine Learning in Economics
Professor Mark Montgomery
March 11, 2022

Slide 1: Title Page (COVID-19 Analysis: The Effectiveness of Restrictions in the United States)

COVID-19 data was analyzed using R statistical programming in order to determine the effectiveness of lockdowns and restrictions implemented during the pandemic in the United States.

Slide 2: Introduction

Each group member had specific roles which contributed to the progress of the project. Jesse was the manager of the project and an R-coder. Charlie was the head R-coder and quality control. Kailey was a note taker and quality control. Gavin was also a note taker and quality control, and Qiushi was an R-coder and note taker. We wanted to research this topic because there is much controversy about whether or not the COVID-19 lockdown and its restrictions were effective enough to control the spread of COVID-19. So, we explored this topic through the different characteristics that R's COVID-19 database had to offer.

Slide 3: Academic Studies on Lockdown Significance

We came across two academic studies: one which confirmed that lockdowns were effective in controlling COVID-19 cases, and the other argued that it was not effective in controlling the spread. We hypothesized that, in fact, there is a statistically significant difference in cases during stricter lockdown periods and increased restriction levels, meaning stricter lockdown and restriction policies are associated with lower COVID-19 transmission.

Slide 4: COVID-19 Package in R

The main package we used for this research project was the COVID-19 database available in R. The package contains regional, day-by-day data on COVID cases throughout the world, on the country, state, and local levels, from March 2020 to the present day.

Includes 47 relevant variables such as cumulative cases, deaths, hospitalizations, number of people fully vaccinated, updated continuously as new data is found.

> Policy measures are included in the data, which are specific factors of the stringency index. They are measured from zero, no measures taken, up to a maximum 5, the most stringent (see next slide).

Our major independent variable was the stringency index, which is the overall strictness of lockdown measures in a given area. How this index is calculated is explained in the next slide. Our major dependent variable was the number of cases.

Slide 5: Stringency Index

The stringency index is a measure of how strict a given location's lockdown regulations are at a specific point in time *t*. It was created by a team of researchers at Oxford University who diligently collect policy data for a large number of different countries, provinces/states, and even counties when possible. Because of how technical the calculation is, we decided not to cover it in our slides or presentation, however it does seem appropriate to explain the process here.

Let *t* represent a specific point in time. Then,

$$S_t$$

is the stringency index value at time *t*. Based on the documentation provided on the Oxford project's GitHub page, the value of each index at any point in time is simply the arithmetic mean of the values of a set of sub-indices, each of which is based on the value of a specific indicator variable at the given time. In other words,

$$X_t = \frac{1}{k}\sum_{j \in J} I_{j,t}$$

where *t* is an arbitrary point in time, X_t is the value of an arbitrary index at time *t*, *k* is the number of sub-indices used to calculate index *X*, *j* is a an arbitrary indicator variable inside the set *J*, which is the set of all indicator variables upon which index *X* is based on, and $I_{j,t}$ is the value of the sub-index associated with indicator variable *j* at time *t*. This same formula is used to calculate the stringency index. There are 9 sub-indices which are used to determine the stringency index at any given time; they themselves are calculated based on the following indicator variables (note: these indicator variables were mentioned in the slides and in our presentation, however they are restated here solely to ensure the reader has all the pertinent information in one place and therefore does not have to refer back to our slides):

$J = \{C1,\ C2,\ C3,\ C4,\ C5,\ C6,\ C7,\ C8,\ H1\}$, where *J* is the set of indicator variables that the stringency index is based on

Based on the GitHub documentation provided by the Oxford researchers, the value of an arbitrary sub-index at time *t* is calculated as follows:

$$I_{j,t} = 100\left\{\frac{v_{j,t} - \frac{1}{2}\left(F_j - f_{j,t}\right)}{N_j}\right\}$$

where $I_{j,t}$ is the value of the sub-index associated with indicator variable *j* at time *t*, *j* is the indicator variable associated with the sub-index being calculated, *t* is an arbitrary point in time, N_j is the maximum value of indicator variable *j*, F_j is a flag variable which specifies whether or not indicator variable *j* has a Boolean flag component, $f_{j,t}$ is the value of indicator variable *j*'s Boolean flag component at time *t,* and $v_{j,t}$ is the recorded policy value for indicator variable *j* at time *t*, based on indicator variable *j*'s ordinal scale. F_j is valued at 0 if indicator variable *j* does not have a Boolean flag component, and 1 if *j* does have a Boolean flag component. A final condition for the calculation of sub-index $I_{j,t}$ is that $v_{j,t} \neq 0$. More specifically,

$$v_{j,t} = 0 \Rightarrow I_{j,t} = 0.$$

With all this said, the value of the stringency index at time *t* is calculated as follows:

$$S_t = \frac{1}{9}\sum_{j\in J} 100\left\{\frac{v_{j,t} - \frac{1}{2}\left(F_j - f_{j,t}\right)}{N_j}\right\} = \frac{100}{9}\sum_{j\in J}\frac{1}{N_j}\left\{v_{j,t} - \frac{1}{2}\left(F_j - f_{j,t}\right)\right\}$$

Slide 6: R Libraries Used

For both parts of the analysis of our project, we used several packages. Obviously, our main library we used was the COVID-19 package, which contained all the necessary components and data of our analysis.

To filter and clean the data, we found it useful to use the data.table package. Using the setDT() function, we can convert a data.frame object into a data.table object which allows for efficiency when modifying column data without making copies of the original data.table.

The packages, usmap, map view and tigris, were very helpful when it came to GIS plotting and were used in our analysis with the state case concentrations and provided a nice visualization of the case concentration levels across the United States.

Slide 7: Problems Encountered/Solutions

Originally, we seeked to explore the differences in cases of lockdown periods and non-lockdown periods on a city to city basis. We quickly realized, unfortunately, that this would lead us into some trouble because cities inherit restriction levels from state levels. In this case, whatever policy is enforced on a state-wide level overwrites the city restriction level. This causes issues when a city restriction may be more strict/lenient than the aggregate statewide level.

Regarding the negative outliers in daily cases and daily deaths, we came to the conclusion that they were caused by adjustments made to the cumulative case numbers, most likely in response to clerical errors made when the data was initially entered in. Our first thought was that the data might be more accurate if we included them as they were, however we eventually started to reason that corrections made in response to clerical errors might not be entirely accurate (i.e. the -40,000 outlier in Florida’s data). Furthermore, if they were responses to errors made outside the scope of our analysis(i.e. Before the beginning of 2021), then the errors being corrected wouldn’t be present in our data to cancel out the effect of the negative outliers, thereby skewing our results. We decided to replace all negative outliers with a value of zero in both the daily cases and daily deaths features because it seemed likely to us that minimizing the effects of large outliers to a reasonable degree would reduce the inaccuracy of our analysis as well as make our results more credible (research is partially judged on the soundness of the diagrams created during presentation, and we figured our audience would probably question a diagram that showed negative daily cases or deaths). We feel this reasoning is justified, as our results would likely be skewed to some degree regardless of what actions we took to correct the issue and we made a logical judgment on what goals should be prioritized (we chose accuracy and credibility).

Slide 8: Assumptions

We assumed that all COVID-19 cases in the periods studied were unique (someone is unlikely to test positive twice within 12 months). This was an especially important consideration when we analyzed the proportion of the population affected by COVID for each state. We seeked to not overcount or inflate the percentage of population affected.

We also assumed the population did not change from 2020 to 2022. Obviously, people are constantly moving in and out of any given state, but for consistency purposes we kept the population the same– which is what the database provided us with.

We assumed the stringency index at any given moment accurately reflected statewide restrictions, and there was not any lag (i.e. when *x* restriction was lifted on *y* date, the data reflects this on *y* date.)

Exploratory Data Analysis (EDA) Section

Slides 10 and 11: Methodology - Retrieving the Data

We are fortunate to have access to US statewide data in the COVID database whereas other countries like China only have country-wide data. This allowed for a more granular analysis. To obtain our data, we used administration level 2, the 2 stands for statewide data.

We quickly realized we needed to convert our cumulative cases for each state into daily cases. To do this, the shift() function was utilized to “lag” the data by 1. A new variable was created in the data.table named “daily cases” and the computation was *confirmed - previous.* We then dropped the previous column from our data.table since it is no longer needed.

Using a for loop and the assign() function, we were able to create fifty (50) different data.table objects with ease. About 2 lines of code, completed the work of what would be 50 lines of code (1 line for each state).

As mentioned earlier, utilizing the usmap library allowed us to create our GIS plot of the United States after lining the case concentrations up with the corresponding FIPS code for each state.

Slide 12: Percentage of Population Vaccinated in Suffolk, Nassau, NYC

This plot shows the percentage of population vaccinated throughout the pandemic in Suffolk, Nassau County, and New York City. Leading the numbers is New York City– with around 77% of the total population fully vaccinated as of February 20th, 2021. This is followed by approximately 71%-72% of the population currently fully vaccinated in Suffolk and Nassau County.

Slide 13: US Heat map and histogram of COVID-19 concentration by state in 2021.

Concentration refers to the total number of cases a given state recorded in 2021 as a proportion of the state's population. The darker blue on the map shows areas of high concentration and the lighter blue represents lower concentrations.

Surprisingly, the histogram of the state case concentrations appears to follow an approximate normal distribution, with n = 50.

Slide 14: Summary

We found it very interesting that the two non-contiguous states, Alaska and Hawaii, were on the extremes of the case concentration spectrum. Alaska had the highest concentration with around 15% of the population affected, and Hawaii had the lowest concentration of about 6.2% of population affected. The average state case concentration was 10.6% of the population, with a standard deviation of 1.8%

Slide 15: EDA Conclusions

See above explanations that elaborate on the past few slides. This concludes our EDA.

Main Experimentation Section

Slide 17 - 21: Methodology Comparing Data from Periods of Strict and Lenient Regulations

Walk the audience through some of the steps we took, along with code snippets used in our research.

Slide 22: Average Stringency Index by State (2021)

Includes a heat map for the average stringency index of each state during the year 2021. Decided to include this after our presentation because it seemed relevant for comparison purposes.

Slide 23: Average Daily Cases and Deaths by State (2021)

Includes heat maps for average daily cases and average daily deaths for each state in 2021. Decided to include this after our presentation because they are mentioned in methodology to a degree, so it didn't make sense not to include it in our final version.

Slide 24: Statistical Significance of Lenient-Strict Mean Daily Cases Difference for each state

Includes heat maps for the 3 sets of 50 T-tests performed on daily cases by state. Followed the advice provided by the audience during our presentation and tried to make the descriptions under each heat map more understandable instead of offering a brief and crude explanation of the nature of the T-tests used.

Slide 25: Box Plots of Daily Cases in Strict and Lenient Periods

Used New York, California, and Florida.

Slide 26: Violin Plots of Daily Cases in Strict and Lenient Periods

Used New York, California, and Florida. Violin plots are like box plots, but instead of a box, there is a rotated kernel density plot on each side.

Slide 27: Time series plots of Daily Cases, Daily Deaths, and Stringency Index

We thought that it would be interesting to select New York and two different states, other than California and Florida. We decided to select Louisiana because it (surprisingly) had a very high average stringency index during the year 2021, so we thought it could be a good idea to see if the time series plot shed some light on why this was the case. We also decided to pick Hawaii because Jesse's work seemed to indicate that it had the lowest concentration of cases among all the states in the US.

Slide 28: Statistical Significance of Lenient-Strict Mean Daily Cases Difference: Summary and Conclusion

Wraps up the Main Experimentation section by providing some raw numbers and discussing our conclusions. Based on 150 separate T-tests (3 for each state), we determined that stricter lockdown regulations did seem to have the desired effect of reducing COVID-19 transmission among a large fraction of US states over the course of the year 2021.