Which Factors Explain the Difference in Severity of COVID-19 in Countries, and in What Ways?

John Lee

Note: Minor error after doing most of the work, which is that any statistic labeled “tests per million people” should be “tests per person”

Project Statement:

It has been over a year since the COVID-19 pandemic started. Most people in the world have felt the effects of the pandemic in their everyday lives. However, people in different countries have experienced the pandemic quite differently. Some countries have done a decent job in controlling the outbreak, while others struggle more. After a year, now we may start to see trends and commonalities in countries that have done better, and/or countries that have struggled more. The purpose of this project is to identify which qualities of a country or its population are correlated with the success (or a lack thereof) of dealing with COVID-19, and in what ways they are correlated.

Data Description

The first step is to define how to measure “success” of a country. I have decided to focus on two measures: deaths due to COVID-19 per million people, and cases of COVID-19 per million people. It might be interesting to see if conclusions are different between deaths and cases, and adjustment for population is obviously needed.

The next step is to brainstorm, explore, and collect data for variables that might be related to the success of a country with COVID-19. I tried to gather a comprehensive set of data that describes a country’s demographics, healthcare, and COVID-19-related policies, and after browsing through https://ourworldindata.org/charts, I decided on 15 variables.

- Demographics and Economy

    - Population density (people per sq km of land area)
    - Population (number of people)
    - GDP per capita (constant international US $)
    - Median age (years)
    - Urbanization (share of population living in urban areas)
    - Life expectancy
- Health
    - Respiratory disease death rate (deaths per 100,000 individuals)
    - Physicians per 1000 people
    - Health expenditure per capita (current international US $)
- COVID-19 response policy
    - Contact tracing (portion of days since January 1st, 2020 in which "comprehensive tracing" was done in a country)
    - International Travel Controls (portion of days since January 1st, 2020 in which a country implemented "Ban on high-risk regions" or "Total border closure")
    - Stay-at-home requirements (portion of days since January 1st, 2020 in which staying at home was "required (except essentials)" or "required (few exceptions)
    - Public gathering restrictions (portion of days since January 1st, 2020 in which gatherings of over 10 people were restricted)
    - Internal movement restrictions (portion of days since January 1st, 2020 in which a country "restricted movement")
    - Tests per million people (Total tests / population in millions)

Ourworldindata.org is itself an organizer of data, and the original sources (such as the World Bank for GDP) are specified in the data descriptions in the ourworldindata.org pages (hyperlinks above). All data from the links above are available in the form of csv files. However, most of the datasets are designed to show change in time, and have multiple data points for long time periods. Some preprocessing is needed to select most recent data points and create a new csv that contains merged information. This is done in preprocessing.py, which is included in the submission.

For all types of variables, I tried my best to get the most updated data possible, but for some variables, slightly older data could have seeped through during the Python processing. This was a compensation that was made. If I had to ignore countries that didn't have data for physicians per 1000 people in, for example, 2018, then I would have had to exclude many countries from analysis, or not use the variable altogether in order to include them.

By analyzing a diverse set of data like this, it could be possible to identify which variables are more important and correlated to the spread of a pandemic, and which variables are less relevant.

Exploratory data analysis

After I collected and merged all the variables, a temporary dataset I created was "data.csv". However, it can be seen that some countries do not have available data for many of the variables. A true global study would Ideally consider all countries, but since too much data is simply not available for some countries, I decided to exclude those countries from the dataset. A list of excluded countries can be found in preprocessing.py in the excludeCountries list global variable.

After excluding these countries, there are two final datasets that I must introduce. The csv file by the name of "preprocessed data.csv" features all countries that had data available for all variables with the exception of "COVID-19 tests per million people" (or Tests_per_mill). This dataset is imported into the R code with the name "dat1", and contains data for 164 countries. This dataset excludes only the countries specified in the previous paragraph. The csv file by the name of "preprocessed data with testing.csv" is a subset of "preprocessed data.csv", but excludes countries that did not have COVID-19 testing data available. This dataset is imported into the R code with the name “dat2”, and contains data for 104 countries.

In the R code “Exploratory Data Analysis” section, I have plotted all the dependent variables with cases_per_mill as the independent variable, for both dat1 and dat2. If we look closely, the points that are in dat1 but not in dat2 are located at the bottom of the plots for the most part, meaning that COVID-19 cases per million people are very low, no matter what the dependent variable is. This means that countries with no testing data available have significantly lower cases reported.

Normally, these trends could possibly be just incorporated in the model, and the regression could be left to decide what meaning it has in the big picture. However, after some more data analysis, (see this link) it turns out that almost a third of the countries without testing data have less than 1000 cases per million people, while the average for countries with testing data is nearly 40,000 cases per million people. The relationship between testing and confirmed cases is unique in that without robust testing, “true” data for confirmed cases is impossible to be reliable.

Therefore, I made the decision to eliminate the 60 countries without testing data. The cost of this decision is that the sample size decreased significantly (about 30% of all countries in the world) and thus represents less of the world’s nations, but the possible benefits are noise reduction, and increased robustness of the remaining dataset.

Data Analysis: Methods and progression

Unfortunately, a “normal” MLR did not seem to be the best model for this particular dataset (dat2). At first glance, not many independent variables were strongly correlated linearly with the dependent variable (in the plots described above). Even if we extended the model to include polynomials of degree larger than 2, the plots do not give us any simple intuition about the modeling options for the independent variables (cubic, quadratic, etc).

For the sake of trying, I attempted to fit the data using a normal MLR for cases per million people, to see what happens. (Section 3.1.1. of R code) We see some significant variables, but adjusted R-squared is only 0.4747, and the VIF for two variables are over 5. I decided to remove life expectancy, as life expectancy itself is a variable predicted by other information. (life expectancy will be dropped from now on) (Section 3.1.2.) The VIFs decreased, adjusted R-squared increased, but there was still some unfavorable behavior in residuals vs Fitted and Scale-Location plots, as the left side and right side looked different from each other, and close to 10 countries had negative fitted values. (Section 3.1.3) A similar attempt with deaths per million people as the dependent variable had similar results. Adjust R-squared was even lower at 0.424, and about the same number of countries had negative fitted values.

In an attempt to find an improved model, I turned to the GAM function. Given the not so obvious nature of the relationship between variables and apparent nonlinearity, the GAM model seemed like an appropriate model to use. (Section 3.2.1.) I first fit an s-curve to all the variables, and then after looking at the plots, changed some variables to a linear curve if the graph for that variable indicated a linear association. The results can be found in the output of section 3.2.1 in the R code. Two of the linear coefficients were found to be significant, and four of the smooth terms were found to be significant.

However, in population and population density, one outlier each could be found. India, with its enormous population, could be dragging the curve and influencing it too much. The region between India and the next most populous country is larger than the region between the next most populous country and zero, and India's population alone makes the country unique in the world. Singapore was the outlier with population density, and Singapore is also a unique country in that the entire country is basically one city. Its population density being four times the next dense country in the dataset could be influencing the curve heavily, and the amount of leverage it has made it justifiable to remove it from the dataset for the sake of finding a more accurate curve, in my opinion (or at least worth a shot). (section 3.2.2.) After taking the two outliers out, the population and population density curves did change, and the adjusted R-squared also increased by 0.022.

(section 3.2.3.) I used the GAM function again, but using deaths per million people instead of cases per million as the dependent variable. Interestingly, more variables turned out to be nonlinear.

Data Analysis: Comparing Models and Variables with a Table

Note: p-values less than 0.05 were considered “significant”, but variables with p-values < 0.1 were also included in the table for reference. All plots for the GAM functions can be found in section 3.2.2 and 3.2.3 of the R code

	MLR: Cases	MLR: Deaths	GAM: Cases	GAM: Deaths
Adj R-squared	0.4747	0.424	0.819	0.677
No. of significant variables (linear)	4	2	1	0
No. of significant variables (nonlinear)	N/A	N/A	6	6
Population			. P-val: 0.072 Slope: slightly negative, almost linear	
Population density			* P-val: 0.0206 Slope: Many ups and downs	* P-Val: 0.0437 Slope: Decreasing, then increasing
GDP per capita				. P-Val: 0.068443 Slope: Decreasing, concave up
Respiratory disease death rate				
Median age	* P-val: 0.04770 Slope: positive Coef: 1371	*** P-val: 2e-06 Slope: positive Coef: 55.71	** P-val: 0.008148 Slope: Monotone increasing, concave up	*** P-Val: 0.000110 Slope: Slightly increasing, then suddenly spikes

Contact tracing				
International travel controls	* P-val: 0.00379 Slope: negative Coef: -36520	. P-val: 0.08 Slope: negative Coef: -456.8	** P-val: 0.002931 Slope: Decreasing, stable, then decreasing again	* P-Val: 0.027333 Slope: Almost constant, then sudden decrease
Physicians per 1000 people			*** P-val: 0.000574 Slope: Slightly increasing until 3, then decreasing	*** P-Val: 0.000135 Slope: Slightly increasing until 3, then decreasing
Percent urbanization	. P-val: 0.08217 Slope: positive Coef: 371.3		. P-val: 0.0724 Slope: Mostly increasing, accelerates towards end	
Healthcare expenditure per capita			*** P-val: 0.000256 Slope: Increasing, decreasing, and increasing again	* P-Val: 0.0289 Slope: Increasing, constant, then increasing again
Stay home requirements	* P-val: 0.0257 Slope: positive Coef: 27470	*** P-val: 0.000175 Slope: Positive Coef: 995.2	*** P-val: 0.000867 Slope: positive Coef: 29398	*** P-Val: 8.14e-05 Slope: Almost linearly increasing, slightly concave down
Gathering restrictions				
Internal movement restrictions	. P-val: 0.0745 Slope: negative Coef: -21970			
Tests per million people	* P-val: 0.01366 Slope: positive Coef: 6640		*** P-val: 0.000118 Slope: Increasing until about 2.5, then decreasing	

Data Analysis: Interpretations and Insights

Overall, the GAM models were more informative and insightful than the normal MLR models, and had higher adjusted R-squared values, as expected. To answer the question of which variables are correlated with cases and/or deaths, and how, I will attempt to create brief interpretations and insights of each variable.

Population: For the most part, population did not seem to be highly correlated with deaths or cases. The only general trend we see is that the slope is slightly negative for all four models.

Population density: The GAM models seemed to extract some trends with regards to population density, but the relationship between population density and COVID-19 cases per million seems to be very volatile, and an intuitive conclusion seems difficult to acquire. With regards to deaths, however, there is a clear negative correlation, meaning that the higher the density, the lower the number of deaths per million people. Perhaps this is because more “packed” countries have better infrastructure and better control.

GDP per capita only had a slightly significant p-value for deaths, but the negative trend could be noteworthy, as it indicates that richer countries have fewer deaths per capita.

Median age was very significant across all models. All had positive slopes, indicating older populations tend to suffer more from the pandemic, but a very interesting point is with the GAM death model, in which a rapid spike in deaths can be spotted when the median age exceeds 40 years.

International travel controls were also very significant across all models with an overall negative slope. One interesting trend is that in the GAM models, the slopes are more negative at the ends than the middle, indicating that the most contrast can be seen between countries that don’t control travel at all versus countries that control even just a little, and between countries that control about 80% of the time versus countries that control travel all the time. After accounting for other variables, 20% travel control and 80% travel control do not differ significantly.

The effect of Physicians per 1000 people is very similar between the two GAM models. Increasing cases/deaths until about 3 physicians per 1000 people, then decreasing after 3. It could be the case that the increase in density of physicians indicates more medical infrastructure that allows testing, which would increase cases of

COVID-19 and deaths attributed to it, but after a certain point, more doctors help reduce the damage done by the pandemic.

Percent urbanization had a slightly positive correlation only for COVID-19 cases and not for deaths. A reason could be that urbanization indicates more dense clusters of people, which increases interactions and virus spreading, but does not increase death due to better medical infrastructure in urban areas.

Healthcare expenditure per capita is significant in the GAM models, but is also the most mysterious. Perhaps a reason for the positive, negative, and positive correlation again might be similar to the “number of physicians effect”. An increase in cases and deaths due to more accurate diagnosis of the disease, followed by a decrease due to adequate medical infrastructure. However, the following increase again might be due to the fact that exorbitant amounts of money spent on healthcare is an indication of an inefficient and overpriced healthcare system.

Stay home requirements are significant in all models with a positive slope. This could be due to the fact that stay home requirements are usually not preventative measures, but reactionary measures against spreading that have already become hard to control. More days with stay home requirements mean more periods in which an already bad situation needs to be kept under control.

Tests per million people was significant for cases, but not for deaths. The explanation for the increase followed by a decrease could be similar to healthcare expenditure and no. of physicians; more testing leads to more cases discovered, but extremely thorough testing leads to a very strong control over the spread of the virus.

Respiratory disease death rate (from 2018 and before), contact tracing, internal movement restrictions, and gathering restrictions were not very significant, but this could be related to how I defined some of these variables (as portions of time), and a different approach (maybe categorical variables?) could produce different results.

Summary and Discussion:

In summary, population density, median age, international travel controls, physicians per 1000 people, healthcare expenditures per capita, tests per million people, and stay at home requirements were the most significant in explaining the differences between cases and deaths of COVID-19 in countries. Percent urbanization, GDP per capita and population were also worth analyzing, and past respiratory disease death rate, contact tracing, internal movement restrictions, and gathering restrictions were not very significant.

Given the spread between linear and nonlinear effects of the independent variables, a GAM model turned out to be a decent choice. The simple MLRs were worth comparing with, and the differences between the two models highlighted the significance of using an additive model for an application like this.

The limitations of this project come from the compromises that were made during data processing. About 30 countries couldn’t be considered because they didn’t have enough data for the independent variables I selected, 60 more countries were taken out because of a lack of testing data, and several outliers were subtracted too. The resulting dataset I eventually worked with was about 100 countries’ worth of data, which is about half of the number of countries in the world. If a lack of information indicates a lack of infrastructure and personnel required to acquire information in a country, the deletion of countries could have resulted in a biased dataset.

Future research might attempt to investigate changes over time instead of just using aggregate data. Also, variable selection could be refined to focus around a more specific theme or to test a narrower hypothesis than this project.

List of data and code:

I have included all original csv files, downloaded from the links in the **Data Description**, in “raw data.zip”. These csv files can be processed by preprocessing.py, but I have included the processed files in “processed data.zip”. All R code is in “Factors that correlate to the spread of COVID-19.rmd”, and “Factors that correlate to the spread of COVID-19.html” is the knit html file.