Analyzing the COVID-19 Mortality Rate with respect to Various Variables

Robin Chenxu Mao, 1008267475, robin.mao@mail.utoronto.ca

Apr. 04, 2023

Introduction

To date, the COVID-19 epidemic remains as a major global public health concern. With recent data showing the COVID-19 confirmation count reaching 671M with total death toll of 6.83M [(1)], with the daily new case count over 2000.

Throughout online literature, many focus on the spreading pattern prediction and aggregate mortality rate severity in specific countries and regions,

Analysis and forecast of COVID-19 spreading in China, Italy and France [(2)], in which the paper focuses on modelling the COVID-19 spread in China, Italy, France, analyze the spreading pattern and create empirical mathematical models such as SIR model;

Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States [(3)], in which the paper focuses on the COVID-19 mortality rate prediction in India using advanced statistical methods and neuro-network;

Risk factors prediction, clinical outcomes, and mortality in COVID-19 patients [(4)] , in which the paper analyzes the risk factors of the COVID-19 and focus on the prediction of clinical outcomes and medical factors.

However, none analyzes the COVID-19 severity distribution from a global perspective nor none focuses on the resulting economic factors of severity. As the global mortality rate variation has intensified the epidemic severity in many regions, this report will focus on the global COVID-19 severity variation from an economic and mathematical perspective with the core message "what are the important economic factors resulting in the COVID-19 global mortality rate variation and how", analyzes on the potential resulting factors within and utilize statistical to create model. Used data sets provided by John Hopkins University, World Bank and more data sources, cited in the references section.

Before entering the detailed mathematical and economic analysis, the beginning section will first analyze the global COVID-19 mortality rate distribution to enhance readers' understanding.

Section 1. An Introduction to Global COVID-19 Mortality Rate Distribution

The formal definitions are given as below,

- **The COVID-19 Mortality Rate:** the estimated probability (sampled from the population) that a COVID-19-infected individual will result in disease-related mortality in a country.
- **Confirmation Count:** the population count of clinical diagnosed infected individuals. Can be referred to as "Case Count", "Confirm Count", etc.
- **Omicron Variant Proportion:** the estimated proportion of Omicron COVID-19 case out of the total COVID-19 confirmation count. Can be referred to as "omicron proportion", etc.

As the main y-variable, the mortality rate distribution will be demonstrated first.

Note that the regression analysis section will start when capable. This paper will strictly follow the course instruction with enhancement in order to further understand the research topic.

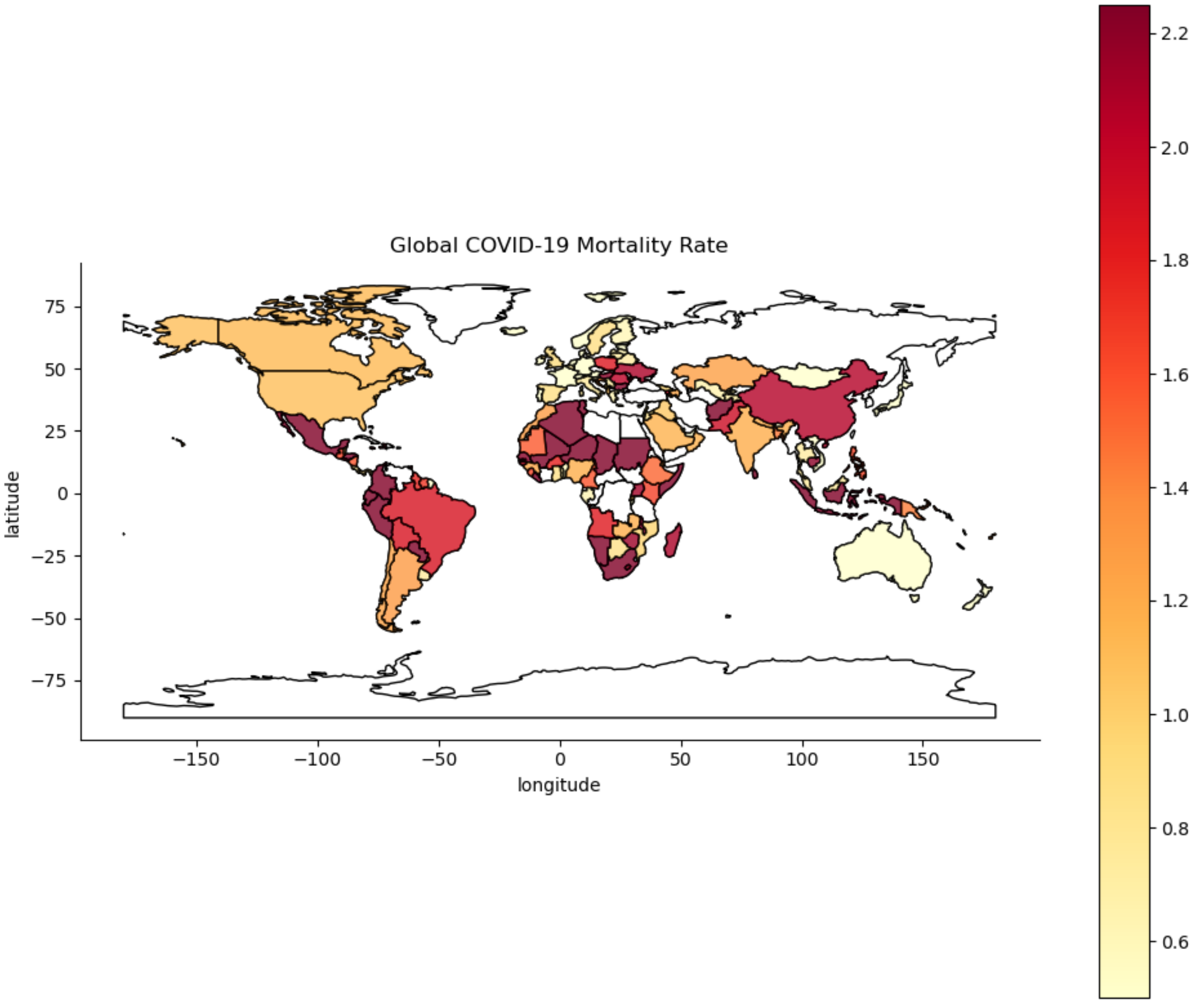
1.1 Data

The visualization dataset is queried based on the John Hopkins University COVID-19 statistics dataset $^{Data(1)}$, which contains the time horizon for the global COVID-19 confirmation count and fatal count, and country code dataset from UN Statistics $^{Data(7)}$ which contains the detailed latitude and longitude data.

Section 1.2 continued in the next page for better visual experience.

1.2 Visualization: General Global Mortality Rate Distribution

The distribution is demonstrated as the heat map below,



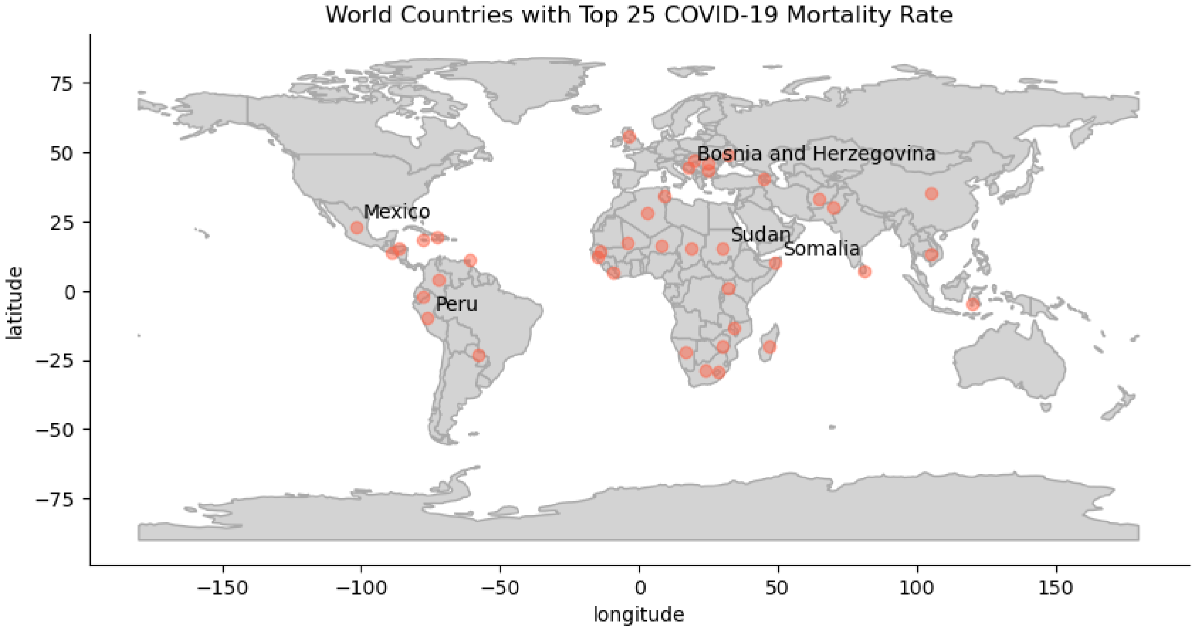
1.3 Analysis: General Global Mortality Rate Distribution

The global variation is demonstrated as above. Countries such as Sudan (7.856259%), Somalia (4.98096%) and Peru (4.891534%) have intuitively significantly greater COVID-19 mortality rates than countries like Bhutan (0.033536%), Burundi (0.070855%) and New Zealand (0.114354 %). With this huge difference times the large current global population bases, millions of viral-related fatal outcomes can be avoided if the resulting factors are determined.

Countries with the top 25 mortality rates are visualized as below,

1.4 Visualization: Top 25 COVID-19 Mortality Rate Countries

Note that the dataset usage are identical to section 1.1 with a different visualization method.



1.5 Analysis: Top 25 COVID-19 Mortality Rate Countries

Countries with the top 25 COVID-19 mortality rate are mostly distributed in South America, Africa and East Europe. There are also significant number of top GDP countries in Southeast Asia. However, countries with the top 5 mortality rate (marked with names) are evenly distributed in the Americas, Africa and East Europe. With this intuitive trend, further introductory intuitive analysis on the potential factors will begin in the next section.

Section 2 continues on the next page.

Section 2. The Introductory Intuitive Factors

Section 1 introduces the intuitive mortality rate difference among the globe and analyzed the relating geographical traits. In this section, introductory casual analyses on intuitive factors begin. Note that more detailed mathematical regressions are provided in the latter section in detailed.

2.1 Time Horizon: Confirmation Count, Death Toll w.r.t Mortality Rate

The x-variables here are confirmation count and death toll. This section will visualize and provide implications on the hypothetical distribution similarities among the variables due to the clinical nature of which one individual has to be infected to be resulted from a viral-related fatality. Thus, the hypothesis here is that the data will show a somewhat significant relation between the confirmation count and mortality rate, as well as the death toll and the latter. The relation is hypothesized to be the mortality rate will increase overtime as the confirmation count rises, vice versa and the mortality rate will increase overtime as the death toll rises, vice versa.

2.1.1 Data

This section uses the confirmation count and fatal count dataset from John Hopkins University $^{Data(1)}$ and queried the mortality rate based on. The dataset is introduced before in section 1.1.

2.1.2 Summary Statistics

The data is divided into monthly interval for better visual experience.

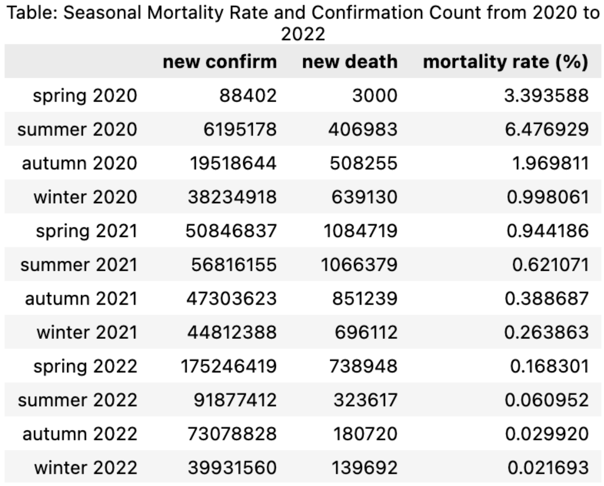


Table: Seasonal Mortality Rate and Confirmation Count from 2020 to 2022

	new confirm	new death	mortality rate (%)
spring 2020	88402	3000	3.393588
summer 2020	6195178	406983	6.476929
autumn 2020	19518644	508255	1.969811
winter 2020	38234918	639130	0.998061
spring 2021	50846837	1084719	0.944186
summer 2021	56816155	1066379	0.621071
autumn 2021	47303623	851239	0.388687
winter 2021	44812388	696112	0.263863
spring 2022	175246419	738948	0.168301
summer 2022	91877412	323617	0.060952
autumn 2022	73078828	180720	0.029920
winter 2022	39931560	139692	0.021693

Note that the COVID-19 mortality rate refers to the ratio of new death toll over cumulative confirmation count as below,

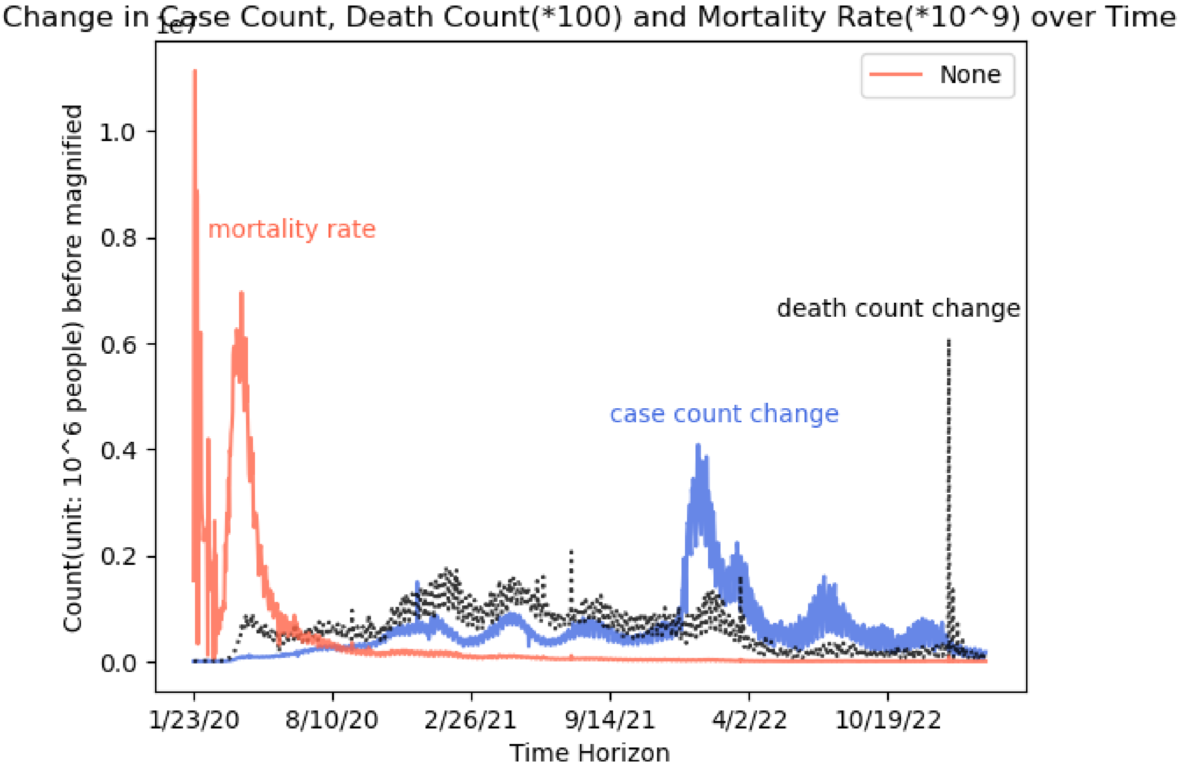
$$MR_i(\%) = \frac{DeathToll_i}{ConfirmationCount_i} * 100 \tag{1}$$

Where MR_i is the mortality rate for country i.

Table above gives the seasonal global new confirm count and death toll between 2020 and 2022. As the seasonal confirmation change had a significant growth from spring 2020 to spring 2022, the global mortality rate had a sudden increase in spring 2020, reaching its peak of over 6.74% around summer 2020, and is relatively stabilized since. Their increment trend does not seem to match.

2.1.3 Visualization

Note that the visualization uses the continuous data without the monthly interval from section 2.1.2.



Note that to analyze the trend, the death count is magnified by 100, and the mortality rate is magnified by 1×10^9. This is reasonable as we are multiplying the values by a constant coefficient, and if there exists a relation (which will be unchanged after the magnification due to the nature of constant multiplication), it will be easier to tell.

2.1.4 Analysis

The mortality rate had huge fluctuations in the very beginning, then was stabilized around 8/10/20. which is before any significant fluctuation in the death change and the case change starting from the midpoint of 8/10/20 and 2/26/21. As well as the peak of the mortality rate lies between 1/23/20 and 8/10/20, while the case count peak lies between 9/14/21 and 4/2/22, and there still exist significant fluctuations after the peak.

Due to the significant distribution difference between the two variables, conclusion for research question 2.1, Confirmation Count, Death Toll w.r.t Mortality Rate, is drawn as **data shows no significant direct relation between case count and the mortality rate, nor between fatal count and the mortality rate.**

This can be reasoned as many countries enforced lockdowns when the mortality rate is high, resulting in a low case count, and they will end the quarantine if and only if the mortality rate is stabilized due to exogenous factors such as mass vaccinations or herd immunity. For example, the Ontario government announced the end of lockdown on Nov.20, 2020 (5), when the mortality rate ended its fluctuation. It would be reasonable for the confirmation count to rise after people gain back rights to outside activities as there exist more social interactions, resulting in more individuals and more ways to spread the virus.

However, when the mortality rate rises between 1/23/20 and 8/10/20, the death count also has a relatively sudden increase. As the mortality rate goes down, the death count also has a downfall before 8/10/20. Thus, another conclusion drawn for section 2.1 is that a rise in the death toll rate will lead to an increase in the COVID-19 mortality rate; however, there exist other factors that can result in a more significant impact (on the mortality rate).

2.2 The Role of Governmental Mode towards the Mortality Rate

Countries shown in the section 1.4 graph have a common trait, which is a dictatorial government. This section will analyze the role of governmental mode towards the mortality rate, with specifying countries into high population and low population, where countries with a population higher than 14049687 is considered high population, which is the mean for the data set population without the 10% edge values (no top 10% population & no bottom 10% population).

The governmental modes are categorized into the following subgroups, Civilian Dictatorship, Military Dictatorship, Parliamentary Democracy, Presidential Democracy, Royal Dictatorship, Semi-Presidential Democracy.

A government with a civilian dictatorship are ruled by dictators who do not derive their powers from the military. For example, Afghanistan, Angola, etc.

A government with a military dictatorship are ruled by dictators who holds the military powers. For example, Botswana, Burkina Faso, etc.

A parliamentary democracy system is a system of democratic governance of a state where the executive derives its democratic legitimacy from its ability to command the support of the legislature, typically a parliament, to which it is accountable. For example, Germany, India, etc.

A presidential democracy** system is defined by the separation of the executive branch from other aspects of government. The head of government is elected to work alongside, but not as a part of, the legislature. There are several types of powers that are traditionally delegated to the president. For example, the Indonesia, Argentina, etc.

A royal dictatorship is a system ruled by ”royal” members. For example, Bahrain, Cambodia, etc.

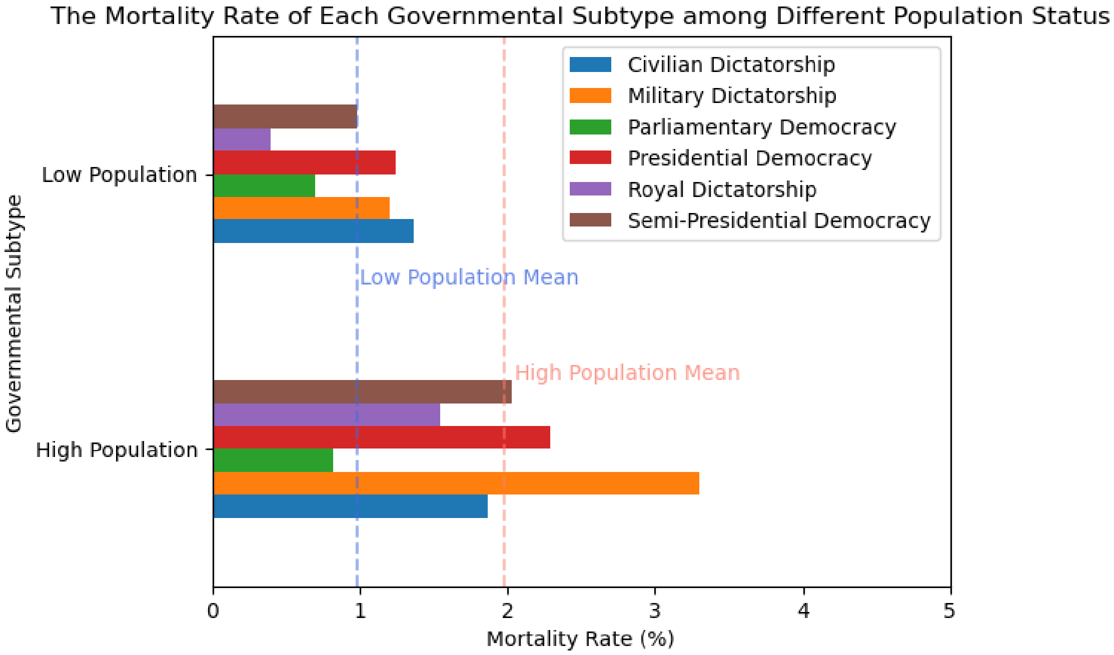
A semi-presidential democracy is a system in which a president exists alongside a prime minister and a cabinet, with the latter two being responsible to the legislature of the state. For example, Ireland, Kyrgyzstan, etc.

2.2.1 Data

The data used in section 2.2 are the John Hopkins COVID-19 Info $^{Data(1)}$ and Dictatorship Index from Kaggle $^{Data(6)}$. All other data are queries manually. The former was introduced before while the latter contains only the dictatorship index info (and country name).

2.2.2 Visualization

Graph below demonstrates the mortality rate differences among all governmental modes.



In countries with **low population**, **civilian dictatorial** governments bring the highest average mortality rate while **royal dictatorial** governments bring the least.

In countries with **high population**, **military dictatorial** governments bring the highest average mortality rate while **parliamentary democratic** governments bring the least. High population countries also have a higher average mortality rate no matter of the governmental mode.

Section 3. Simple Mathematical Relations among the Variables

Simple mathematical regression begins in this section to further identify the relations among the variables resulting in the COVID-19 mortality rate.

3.1 GDP vs Mortality Rate

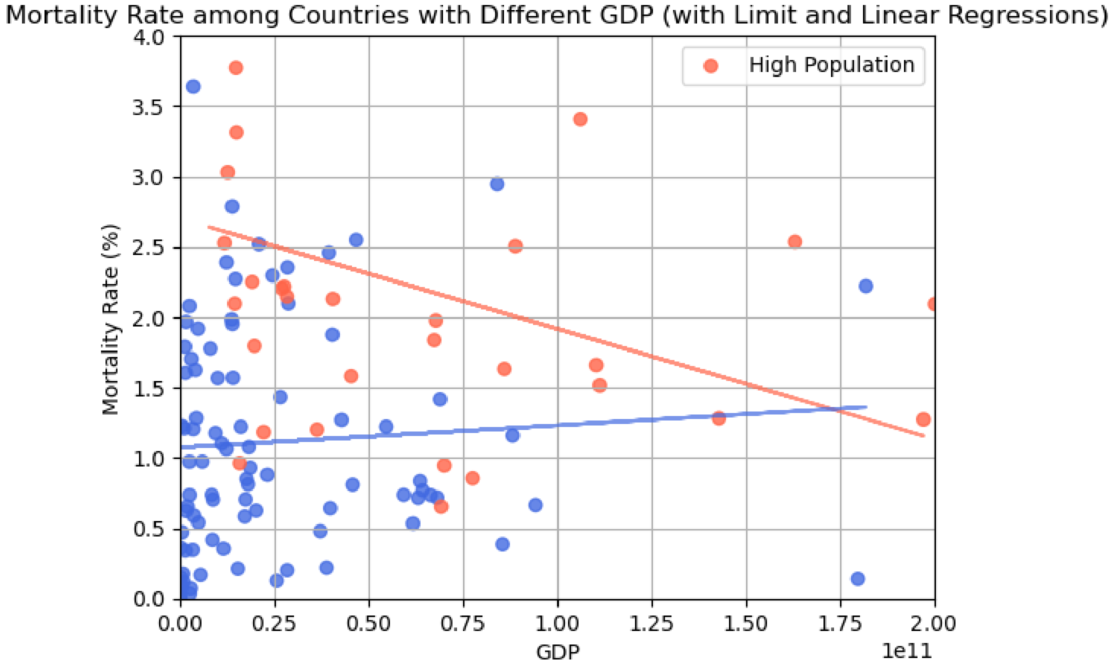
Gross Domestic Product (referred to as GDP), is a measurement of a nation's aggregate output. Theoretically, an higher GDP indicates a better national development and should result in a better medical environment, decrease the mortality rate.

3.1.1 Data

The data used section 3.1 are John Hopkins Data $^{Data(1)}$ introduced in section 1.1 and World Bank Global GDP dataset $^{Data(5)}$, which contains countries and their corresponding GDP time horizon.

3.1.2 Visualization

Note that the outliners such as China and Sudan were removed from the plot, and were not a part of the regression simulation.



For low-population countries, the relation between GDP and mortality rate is positive and elastic. For countries with high population, the relation is negative and relatively more inelastic.

3.1.3 Regression Result

Their mathematical relations are shown as below. For countries with low population,

$$MR_{low} = 1.07305102 + 1.57916441 * 10^{-12} Pop \tag{2}$$

For countries with high population,

$$MR_{high} = 2.70057998 - 7.83312433 * 10^{-12} Pop \tag{3}$$

The models are simple uni-variate linear regression functions.

Section 3.1 concludes that for countries with low population, an increase in GDP increases the aggregate mortality rate. Vice versa, an GDP increase in high-population countries decreases the mortality rate, in a more inelastic way. Moreover, more populated countries also tend to have more outliner in GDP and mortality rate

This can be reasoned as low-population countries having smaller elasticity between the population density and GDP, and countries with high-population and high GDP such as the United States and China tend to have the most advanced medical resources and better epidemic prevention policies, leading to a mortality rate decrease.

3.2 Literacy Rate vs Mortality Rate

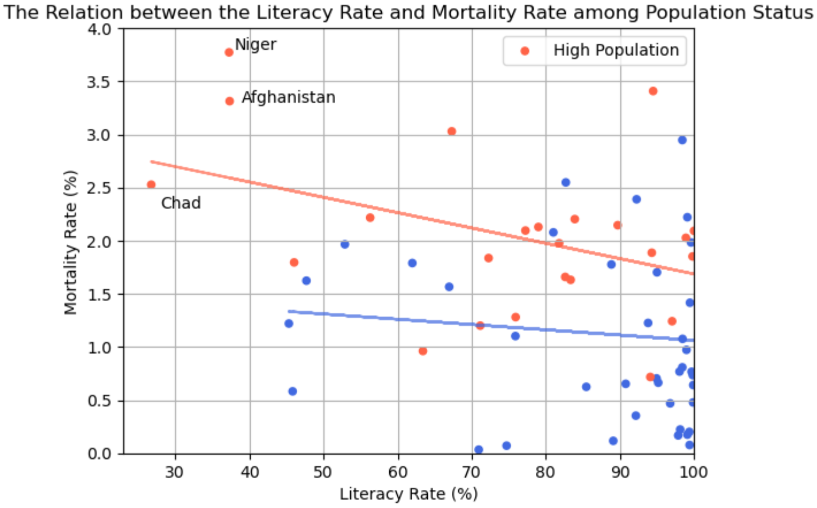
Literacy rate is another essential economic factor. Ideally, as literacy rate increases, citizens will have more basic medical knowledge resulting in a decrease in the mortality rate.

3.2.1 Data

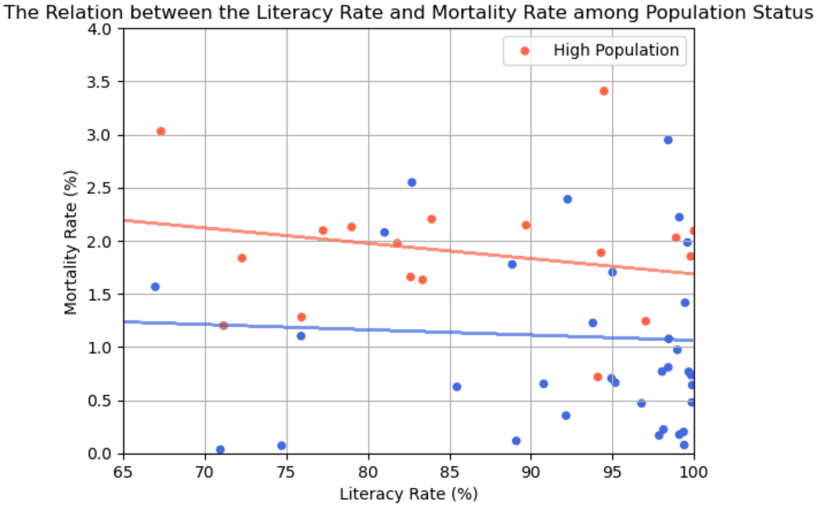
Research question 3.2 uses the John Hopkins COVID-19 Info $^{Data(1)}$ introduced before and Global Literacy Rate Dataset offered by World Bank $^{Data(8)}$, which contains countries and their most recent literacy rate.

3.2.2 Visualization

The first visualization below demonstrates the visualization with outliners included.



It shows more inelastic relaation for both high-population and low-population countries. Below is the graph without the outliners,



The distribution hotspot shows a more elastic relation between the literacy rate and mortality rate, which implies an increase in the literacy rate decreases less mortality rate in the second plot.

3.2.3 Regression Result

After removing the outliners, the regression for countries with low population is,

$$MR_{low} = 1.55856683 - 0.00496083 * LR \tag{4}$$

Where MR is the mortality rate and LR is the literacy rate. For countries with high population,

$$MR_{high} = 3.12982315 - 0.01442991 * LR \tag{5}$$

The elasticity for two population are similar, which indicates an increase in literacy rate leads to similar decrease amount in mortality rate for both low-population and high-population countries. This proves that education, which is a major independent variable regarding the literacy rate, is another important role towards the COVID-19 severity.

Section 4. Multivariate OLS among the Variables and Accuracy Examination

This section will introduce more complicated mathematical model to explain the relation among different variables and examine their accuracy.

4.1 Omicron Proportion, Death Toll and Mortality Rate

As a more deadly variant of the COVID-19, Omicron Proportion is hypothesized to have an influential role towards the mortality rate change. Combine with death toll, this multivariate OLS model will explain the variation in detail.

4.1.1 Data

The datasets used in section 4.1 are John Hopkins Global COVID-19 $^{Data(1)}$ Info as introduced before, Global Population Dataset by World Bank $^{Data(2)}$ which contains countries and their current population and Omicron Time Horizon Dataset from Kaggle $^{Data(3)}$ which recorded each Omicron case, their country, date, etc.

4.1.2 Regression Result

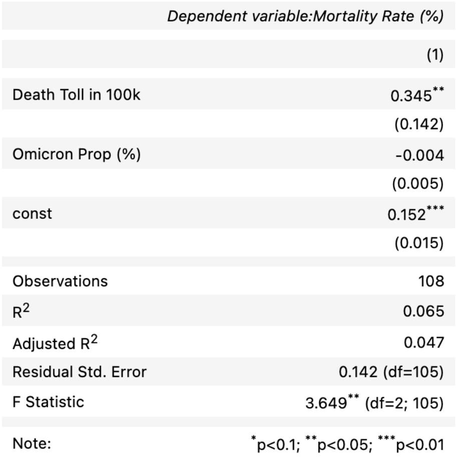
The regression model is provided as below,

$$MortalityRate(\%) = 0.0152 - 0.004 OmicronProp + 0.345 DeathToll \tag{6}$$

Where death toll is measured in $1 * 10^6$. This equation implies that for each percentage increase in the Omicron proportion, the resulting mortality rate will fall by 0.004; for each $1*10^6$ increase in the death toll, the mortality rate increases by 0.345%. The relation is consistent partially with the null hypothesis as the mortality rate shows a positive relation with respect to the death toll; however, the negative correlation between it and the Omicron proportion is unexpected. This can be reasoned by Omicron is a variant that appeared in the later part of the time horizon, where the infection treatment is already greatly improved (compared to the leading part of the time horizon), leading to a lower mortality rate.

4.1.3 Accuracy and Model Statistics

The detailed model statistics is provided as below,



	Dependent variable:Mortality Rate (%)
	(1)
Death Toll in 100k	0.345**
	(0.142)
Omicron Prop (%)	-0.004
	(0.005)
const	0.152***
	(0.015)
Observations	108
R^2	0.065
Adjusted R^2	0.047
Residual Std. Error	0.142 (df=105)
F Statistic	3.649** (df=2; 105)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficient of determinant, R^2, calculated from below,

$$R^2 = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{7}$$

Where y_i is the actual value, $\hat{y}_i$ is the model predicted value and $\bar{y}_i$ is the actual mean. The 6.5% R^2 of the model implies that the model can explain 6.5% of the total y-variation, which is significantly higher than the uni-variate OLS model.

4.2 Temperature, Population and Mortality Rate

Throughout the human epidemic history, environmental factors such as temperature, population density are also critical towards the overall disease severity. Section 4.2 will first use a simple OLS model to analyze the relation among the variables, then use a 2SLS model to further examine other instrumental factors, using the high-population and low-population categorization.

4.2.1 Data

The data used in this sections are John Hopkins COVID-19 Info $^{Data(1)}$ as introduced before, Global Population Dataset by World Bank $^{Data(2)}$, Country Code Dataset from UN Statistics $^{Data(7)}$ as introduced before and Temperature data scraped from Wikipedia $^{Data(10)}$ which contains country and their monthly average temperature time horizon.

4.2.2 Regression Result

For countries with low population, the regression result is calculated as below,

$$MortalityRate(\%) = 0.295 + 0.025Temp + 0.117P \tag{8}$$

Where P is population measured in millions. For countries with high population,

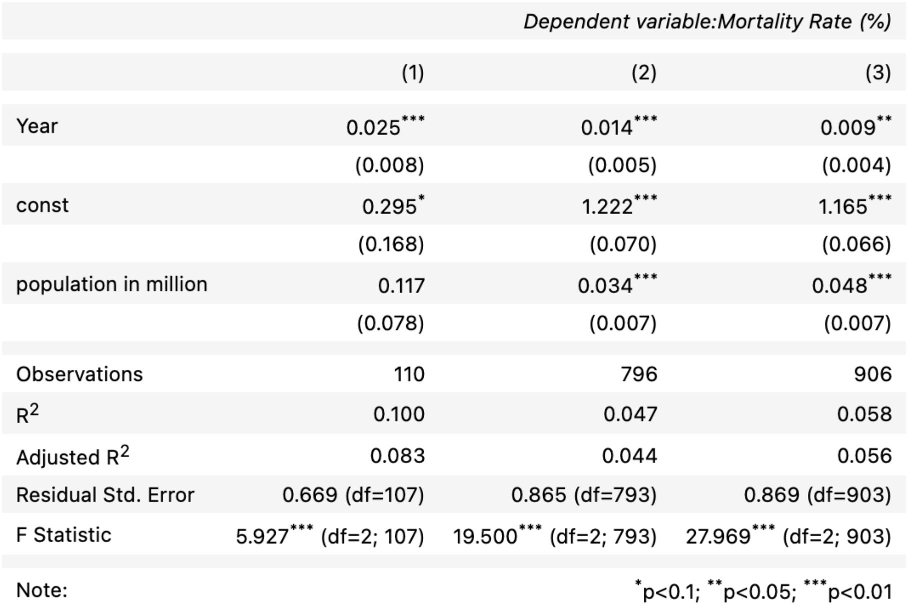
$$MortalityRate(\%) = 1.222 + 0.014Temp + 0.034P \tag{9}$$

For low-population countries, each Celsius degree increase in temperature raise the mortality rate by 0.025, and each unit rise in population increases the mortality rate by 0.117; for countries with high population, each temperature increase raises the mortality rate by 0.014 with each population unit rise increases the mortality rate by 0.034.

Moreover, the mortality rate for low-population countries show a more inelastic relation towards the temperature while being more inelastic towards the population in high-population countries. This can be reasoned by high-population countries having a high population density which results in a situation that the increase in population will have more chance resulting in fatal outcome.

4.2.3 Accuracy and Model Statistics

The model statistics are provided as below,



	Dependent variable:Mortality Rate (%)		
	(1)	(2)	(3)
Year	0.025^{***}	0.014^{***}	0.009^{**}
	(0.008)	(0.005)	(0.004)
const	0.295^{*}	1.222^{***}	1.165^{***}
	(0.168)	(0.070)	(0.066)
population in million	0.117	0.034^{***}	0.048^{***}
	(0.078)	(0.007)	(0.007)
Observations	110	796	906
R^2	0.100	0.047	0.058
Adjusted R^2	0.083	0.044	0.056
Residual Std. Error	0.669 (df=107)	0.865 (df=793)	0.869 (df=903)
F Statistic	5.927^{***} (df=2; 107)	19.500^{***} (df=2; 793)	27.969^{***} (df=2; 903)
Note:			*p<0.1; **p<0.05; ***p<0.01

Where (1) is for low-population countries, (2) for high, (3) for the overall result without categorization.

For countries with low population, the 0.100 R^2 implies that the model explains 10% of the total Y-variation. For countries with high population, the model explains only 4.7% of the total Y-variation. Compare to the aggregate regression without specifying the population status, linear regressions that specifying population status brings a more accurate model for low-population countries but a less accurate model for high-population countries. However, the R^2 is still relatively low, let us further increase the accuracy by dividing time horizon and find relations for each monthly time interval.

4.3 Model with More Variables

To increase the model accuracy, the temperature is divided into 12 monthly interval. The data usage is identical to section 4.2.1 with a different calcualtion method.

4.3.1 Regression Result

For countries with low population

$$\begin{aligned} MR(\%)_{low} = 0.089 - 0.230T_1 - 0.049T_2 - 0.105T_3 + 0.269T_4 - 0.118T_5 - 0.056T_6 - 0.183T_7 \\ +0.301T_8 - 0.132T_9 + 0.130T_{10} - 0.087T_{11} + 0.295T_{12} + 0.024P_Y \end{aligned}$$

Where T_i is the temperature of the i^{th} month.

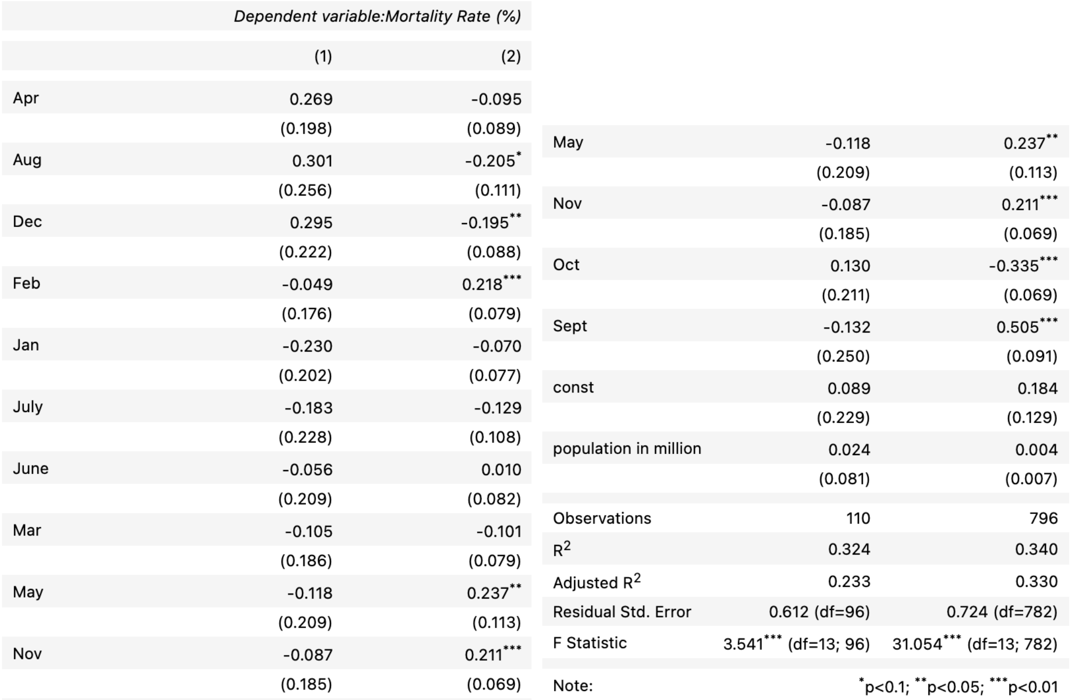
For high-population countries,

$$\begin{aligned} MortalityRate(\%) = 0.184 - 0.070T_1 + 0.218T_2 - 0.101T_3 - 0.095T_4 + 0.237T_5 + 0.010T_6 - 0.129T_7 \\ -0.205T_8 + 0.505T_9 - 0.335T_{10} + 0.211T_{11} - 0.195T_{12} + 0.004P_Y \end{aligned}$$

The new model uses more variables to explain a more mathematical complicated equation, showing a more inelastic relation towards the population in low-population countries, which is consistent with the first model (hypothesis reasoned in the previous section). In low-population countries, mortality rate shows positive relations with the temperatures in Apr, Aug, Oct and Oct. In low-population countries, mortality rate shows positive relations with the temperatures in Feb, May, June, Sept and Nov. This could be hypothesized to relate to seasonal events that are related to tourism leading to seasonal population density change, etc.

4.3.2 Accuracy and Model Statistics

The model statistics are given as below,



	Dependent variable:Mortality Rate (%)	
	(1)	(2)
Apr	0.269	-0.095
	(0.198)	(0.089)
Aug	0.301	-0.205*
	(0.256)	(0.111)
Dec	0.295	-0.195**
	(0.222)	(0.088)
Feb	-0.049	0.218***
	(0.176)	(0.079)
Jan	-0.230	-0.070
	(0.202)	(0.077)
July	-0.183	-0.129
	(0.228)	(0.108)
June	-0.056	0.010
	(0.209)	(0.082)
Mar	-0.105	-0.101
	(0.186)	(0.079)
May	-0.118	0.237**
	(0.209)	(0.113)
Nov	-0.087	0.211***
	(0.185)	(0.069)
Oct	0.130	-0.335***
	(0.211)	(0.069)
Sept	-0.132	0.505***
	(0.250)	(0.091)
const	0.089	0.184
	(0.229)	(0.129)
population in million	0.024	0.004
	(0.081)	(0.007)
Observations	110	796
R^2	0.324	0.340
Adjusted R^2	0.233	0.330
Residual Std. Error	0.612 (df=96)	0.724 (df=782)
F Statistic	3.541*** (df=13; 96)	31.054*** (df=13; 782)
Note:		*p<0.1; **p<0.05; ***p<0.01

The new specification provides detailed mathematical relations, the 0.324 R^2 of low population implies that the model now explains 32.4% of the y-variance. The 0.340 R^2 of high population implies that the model explains 34.0% of the total Y-variance. This detailed model provides with 240% accuracy increase in mortality rate prediction in low-population country and 623.4042253% increase in high-population countries.

4.4 2SLS Model

The initial simple linear model between the Temperature, Population and the COVID-19 Mortality Rate has a low R^2, which indicates that the model is highly likely to be suffered from endogeneity. Thus, this section conducts a 2SLS (two stage least squares regression) on the variables. Our null hypothesis is that latitude and longitude will impact the temperature (because latitude and longitude variations result in different weathers, which determine temperature), which will result in endogenous changes in the COVID-19 mortality rate, decreasing the total y-vairance the OLS model can explain. The 2SLS will have two stages, one finding the correlation between latitude, longitude and the temperature and another one using the stage one model prediction to obtain another regression, eliminate the endogeneity.

4.4.1 Data

The data usage is identical to section 4.2.1.

4.4.2 2SLS Stage One: Latitude, Longitude vs Temperature

The relation is obtained as below using OLS. For countries with low population,

$$Temp_{Low} = 21.939 - 0.144 * Lat - 0.065 * Long \tag{10}$$

For countries with high population,

$$Temp_{High} = 16.301 - 0.124 * Lat + 0.015 * Long \tag{11}$$

The linear models show a negative relation between the temperature and the latitude for both high-population and low-population countries. However, low-population temperatures show negative relations toward the longitude while high-population temperatures show positive.

4.4.3 2SLS Stage Two: New Regression

The new regression is summarized using equations from section 4.4.2 as below. For countries with low population,

$$MortalityRate(\%)_{low} = -0.1935 + 0.1204 * Pop + 0.0531 * Temp \tag{12}$$

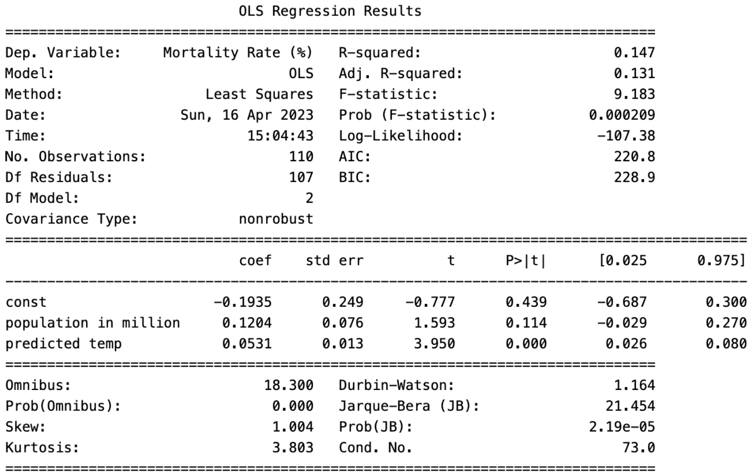
Where P is population measured in millions. For countries with high population,

$$MortalityRate(\%)_{high} = 1.7319 + 0.0395 * Pop - 0.0243 * Temp \tag{13}$$

Accuracy analysis is provided in the upcoming section.

4.4.4 2SLS Model Accuracy and Statistics

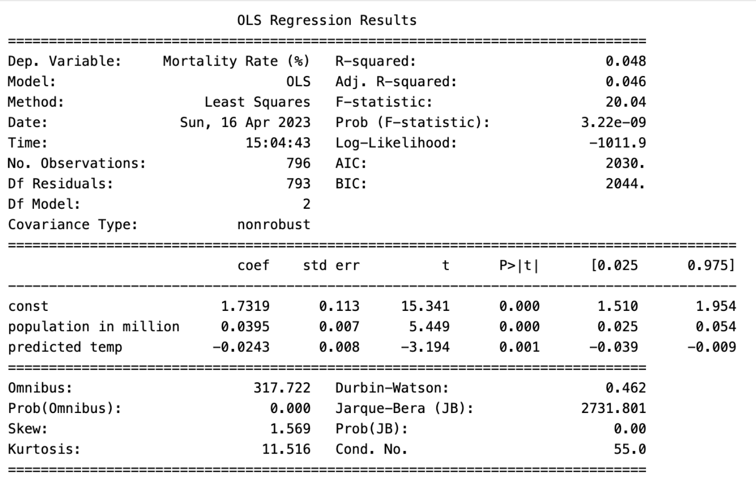
For countries with low population,



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Mortality Rate (%)   R-squared:                       0.147
Model:                            OLS   Adj. R-squared:                  0.131
Method:                 Least Squares   F-statistic:                     9.183
Date:                Sun, 16 Apr 2023   Prob (F-statistic):           0.000209
Time:                        15:04:43   Log-Likelihood:                -107.38
No. Observations:                 110   AIC:                             220.8
Df Residuals:                     107   BIC:                             228.9
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    -0.1935      0.249     -0.777      0.439      -0.687       0.300
population in million    0.1204      0.076      1.593      0.114      -0.029       0.270
predicted temp            0.0531      0.013      3.950      0.000       0.026       0.080
==============================================================================
Omnibus:                       18.300   Durbin-Watson:                   1.164
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               21.454
Skew:                           1.004   Prob(JB):                     2.19e-05
Kurtosis:                       3.803   Cond. No.                         73.0
==============================================================================
```

Without the instrument variable, the previous simple linear regression model has a R^2 of 0.100. Compare to the new model with a R^2 of **0.147**, the total Y-variation that the model can explain had an increase. This also proves our null hypothesis that latitude and longitude create endogenious effects on the temperature which changes the mortality rate prediction accuracy for countries with low population.

For countries with high population,



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Mortality Rate (%)   R-squared:                       0.048
Model:                            OLS   Adj. R-squared:                  0.046
Method:                 Least Squares   F-statistic:                     20.04
Date:                Sun, 16 Apr 2023   Prob (F-statistic):           3.22e-09
Time:                        15:04:43   Log-Likelihood:                -1011.9
No. Observations:                 796   AIC:                             2030.
Df Residuals:                     793   BIC:                             2044.
Df Model:                           2
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     1.7319      0.113     15.341      0.000       1.510       1.954
population in million    0.0395      0.007      5.449      0.000       0.025       0.054
predicted temp           -0.0243      0.008     -3.194      0.001      -0.039      -0.009
==============================================================================
Omnibus:                      317.722   Durbin-Watson:                   0.462
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2731.801
Skew:                           1.569   Prob(JB):                         0.00
Kurtosis:                      11.516   Cond. No.                         55.0
==============================================================================
```

The previous simple OLS model has a R^2 of 0.047, compare to the new mathematical relation obtained as above, the new model can explain 1% more variation than the previous model. Contradicting to the null hypothesis, this increase is relatively insignificant. Therefore, this section conclude that in countries with high population, there is not much endogenous effects on the COVID-19 mortality rate brought by the latitude and longitude. This can be reasoned by the COVID-19 mortality rate of high-populated countries mainly depend on the technology, medical resource availability and other economic factors instead of environmental factors.

Section 5. Regression Tree

Machine learning techniques can further improve the mathematical models obtained previously. This section will focus on analyzing the relation between Omicron Proportion and the COVID-19 Mortality Rate.

The objective function for this section as below, where R is a region that contains all MR values.

$$\min_{j,s}\left(\sum_{i:x_{i,j}\leq s,x_iR1}(MR_i-\hat{M}R_{R1})^2+\sum_{i:x_{i,j}>s,x_iR2}(MR_i-\hat{M}R_{R2})^2\right) \tag{14}$$

Where MR represents the COVID-19 mortality rate, $R1$ and $R2$ are the two regions defined by R. (i, j are indexes.)

The y_i represents the actual observed value and $\hat{y}_{R1}$ represents the predicted value, so $(MR_i-\hat{M}R_{R1})^2$ and $(MR_i-\hat{M}R_{R2})^2$ are the square residual (difference between the predicted values and the observed values, then square). The goal is to minimize the aggregate residual squares found by summing all the individual residual squares.

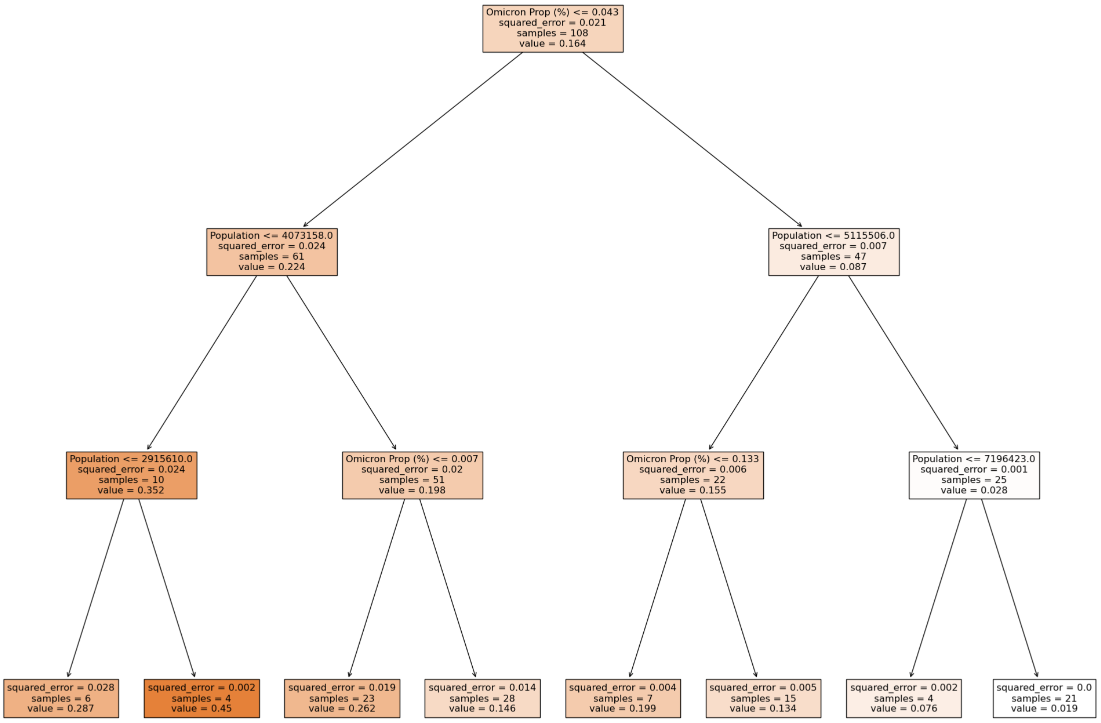
As repeating the simulation for each of the two smaller regions, we stop when $|R|$ = some chosen minimum size or when tree depth = chosen maximum.

$$\min_{tree\subset T}\sum(\hat{f}(x)-y)^2+\alpha|\text{terminal nodes in tree}| \tag{15}$$

Where α is defined as the **regularization parameter**. It can be chosen to be larger if we want a longer regression tree, vice versa.

5.1 Regression Tree

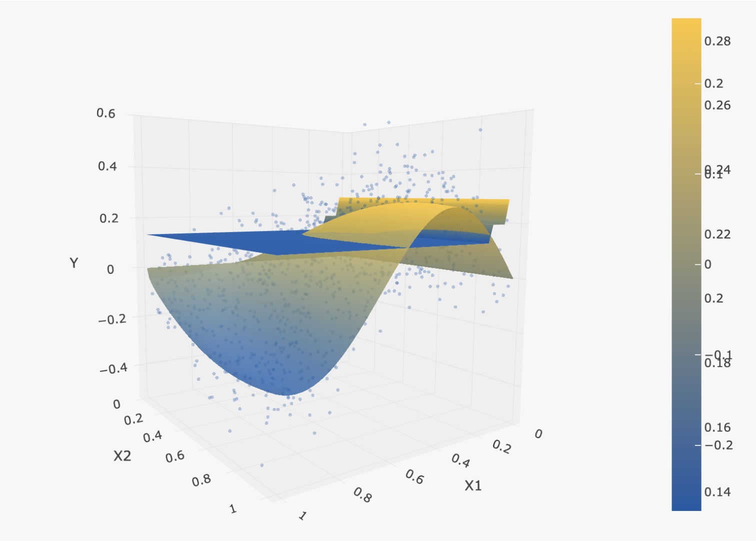
The regression tree has a depth chosen as 3 to maximize the accuracy. The tree is demonstrated as below,



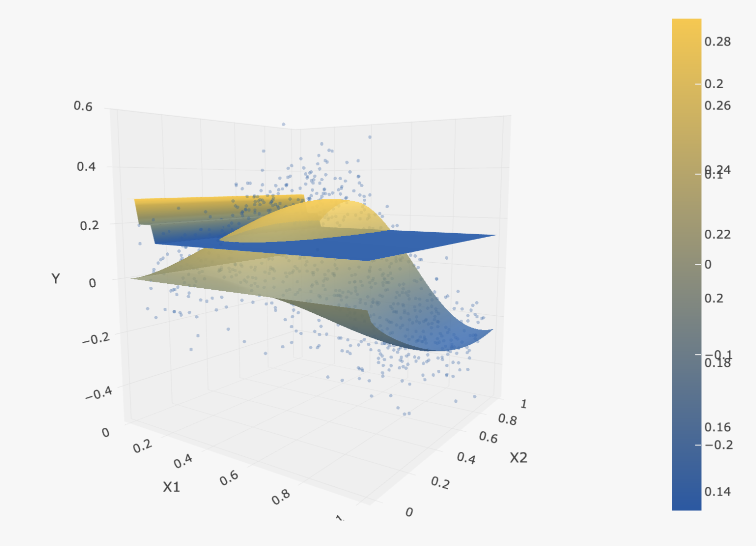
The depth is chosen to be three to minimize the MSE. The regression tree is demonstrated as above, with **MSE** of 0.010464898572977806, which means the aggregate residual square summation is 0.010464898572977806. This proves the accuracy increase compare to the previous OLS model which can only explain about 6% of the total Y-variation. A regression tree is a form of decision tree which improves the overall regression accuracy. From above, we can see the aggregate dataset is categorized into two nodes first (population $\leq$ 4073158 or population $\leq$ 5115506), and then further categorized into two other nodes, etc. The mathematical relation is further detailed visualized as below,

5.2 Mathematical Visualization

The mathematical visualization is shown as below,



On the other side,



Where X1 stands for Omicron Proportion in percentage, X2 for ‘Population‘ and Y for COVID-19 Mortality Rate. The regression tree predictions are piece-wise-constant on rectangle regions. This model explains the actual value trend intuitively, and minimize the square distance between the predicted values and the actual values.

Compare to the previous OLS, the regression tree provides more information and provide mathematical relation for each subgroup. The graph demonstrates that countries with low population and low Omicron proportion have the lowest COVID-19 mortality rate, and countries with low-to-mid Omicron proportion and mid-to-high population have the highest COVID-19 mortality rate. Whenever the Omicron proportion is too high or too low, the mortality rate decreases. These are observations contradicting the null hypothesis which that cannot be observed from a simple OLS.

Section 6. Conclusion

To answer the main question "what are the important economic factors resulting in the COVID-19 global mortality rate variation and how", different from other literature, this report analyzed numerous factors and their relations towards the COVID-19 mortality rate using various methods from shallow to deep, so that audiences with different mathematical background could all understand the paper. The results were constructed among important economic variables with both broad overview and niche review, which is different from other literature. For example, there is simple bar plot focusing on the governmental mode, and there is machine learning regression tree focusing on the detailed model.

However, this economic paper lacks more perspectives such as medical, cultural, which I personally hope future researchers could provide answer onto. If so, the model accuracy could be further improved with less limitations.

Reference Page

Data Source

1. CSSEGISandData. (2022). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. GitHub. https://github.com/CSSEGISandData/COVID-19

2. World Bank. (2021). Population, Total — Data. Worldbank.org. https://data.worldbank.org/indicator/SP.POP.TOTL

3. COVID-19 Variants Worldwide Evolution. (n.d.). Www.kaggle.com. Retrieved April 17, 2023, from https://www.kaggle.com/datasets/gpreda/covid19-variants

4. World Bank. (2021a). GDP per capita (current US$) — Data. Worldbank.org. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

5. World Bank. (2021b). GDP (current US$) — Data. The World Bank. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

6. Democracy-Dictatorship_Index. (n.d.). www.kaggle.com. Retrieved April 17, 2023,from https://www.kaggle.com/datasets/mathurinache/democracy-dictatorship-index

7. Duncalfe, L. (2022, February 4). lukes/ISO-3166-Countries-with-Regional-Codes. GitHub. https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes

8. The World Bank. (2020). Literacy rate, adult total (% of people ages 15 and above) —Data. Worldbank.org. https://data.worldbank.org/indicator/SE.ADT.LITR.ZS

9. US GDP by State 1997-2020. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/davidbroberts/us-gdp-by-state-19972020

10. Global Temperature (Scraping) https://en.wikipedia.org/wiki/List_of_cities_by_average_temperature.

Citations

1. Covid-19 data explorer. Our World in Data. (n.d.). Retrieved February 3, 2023.

2. Fanelli, D., & Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 134, 109761.

3. Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., ... & Georgescu, A. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences, 119(15), e2113561119.

4. Alizadehsani, R., Alizadeh Sani, Z., Behjati, M., Roshanzamir, Z., Hussain, S., Abedini, N., ... & Islam, S. M. S. (2021). Risk factors prediction, clinical outcomes, and mortality in COVID-19 patients. Journal of medical virology, 93(4), 2307-2320.

5. Ontario's COVID-19 response: A history of announced measures, 2020-2022. JD Supra. (n.d.). Retrieved February 3, 2023.