ILLINOIS STATE UNIVERSITY

United States Of America

MAT 490 Project Title:

Linear Regression analysis of COVID-19 Confirmed Cases and Mortality in the USA -State by State analysis

Submitted by

Celdrick Ndze K (kcndze)

To

Pr.Krzysztof Ostaszewski

Department of Mathematics

November 25, 2020

.

Contents

1 Introduction **1**
1.1 Expectations 2

2 Data and Methodology **2**
2.1 Data 2
2.2 Methodology 6

3 Results and Discussions **9**

4 Conclusion **14**
Appendix A 15
Appendix B 16
Appendix C 18
Appendix D 20

References **21**

List of Figures

1 COVID-19 deaths distribution by states 6
2 Histogram of the transformed confirmed cases 9
3 Histogram of the transformed Mortality rates 9
4 Linearity tests 15
5 Normality tests 15
6 Pearson moment coefficient of correlation 16
7 Confirmed Cases 20
8 Mortality rate 20

List of Tables

1 Explanatory Variables and publicly available data sources used in the analysis 4
2 Characteristics of the study Cohort Up to and Including July 22nd, 2020 5
3 Model 1-Determinants for number of Confirmed cases 11
4 Model 2-Determinants of mortality rate 13

5 Pearson Correlations 17
6 Anova table for lm.fit1 and Model1 CNCS 19
7 Anova table for lm.fit2 and Model2 MRAT 19

Abstract

This article examines the determinants of COVID-19 mortality rate, number of confirmed/positive cases,a state by state comparison of 50 states in the United States of America accumulated from February 1,2020 to June 22,2020. To study the changes of the epidemic and to make appropriate decisions in order to help flattened the epidemic curve and consequently prevent further deaths.

Multiple regression analysis was used to establish a linear relationship between each dependent variable and all other independent variables.Cumulative number of confirmed cases and the Mortality rate are the dependent variables. There was a linear correlation between number of confirm cases and population density, number of homeless individuals,average precipitation,cigarette smokers and those currently hospitalize and the multiple regression model was statistically significant. We also find that there was a linear relationship between Mortality rate and percentage of medical coverage,life expectancy by age,age dependency ratio and average annual temperature.The multiple regression model was also statistically significant.

During this time,7485 tests has been performed for the first quarter with a mean number of confirmed cases of 36457 and a maximum number of deaths and confirmed cases of 375133 in the state of New York alone (the state with the highest number of confirmed cases). The states with the highest mortality rates and accumulated cases of COVID-19 should implement high-level control measures that can effectively control the spread of COVID-19.

keywords: multiple linear regression,novel coronavirus; COVID-19; statistical map; epidemic

1 Introduction

The Covid-19 (SARS-CoV-2) pandemic is a major global health threat. The Novel Covid-19 has been reported as the most detrimental respiratory virus since 1918 H1N1 influenza pandemic. According to the World Health Organization [1] as of June 6, 2020, a total of 6,800,604 confirmed cases and 396,590 deaths have been reported across the world. Global spread has been rapid, with approximately 170 countries now having reported at least one case. The coronavirus disease 2019 (Covid-19) is an infectious disease caused by a novel coronavirus called severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). Corona virus belongs to a family of viruses which is responsible for illness ranging from common cold to deadly diseases as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) [6] which were first discovered in China (2002) and Saudi Arabia (2012). Despite the statewide stay-at-home order that was imposed, the suspension of non-essential businesses, all public transport, flights (by some states) and trains on March 2020 there was still rise in the number of Covid-19 cases due to increase laboratory testing, community spread and reporting across the country however there have been slow rate of increase in the spread of the virus in the USA. The 2019-novel Coronavirus (Covid-19) reported in Wuhan, China for the very first time on December 31^{st} 2019. According to Jiang et al [2] ,the fatality rate for this virus has been has been estimated to be 4.5% but for the age group 70-79 this has gone up to 8.0% while for those > 80 it has been noted to be 14.8%. This has led to elderly persons above the age of 50 with underlying diseases like diabetes, Parkinson's disease and cardiovascular disease to be considered at the highest risk. Symptoms for this disease can take 2-14 days to appear and can range from fever, cough, shortness of breath to pneumonia, kidney failure and even death [1]. The virus that causes Covid-19 is thought to spread mainly from person to person, mainly through respiratory droplets produced when an infected person coughs or sneezes. These droplets can land in the mouths or noses of people who are nearby or possibly be inhaled into the lungs. Spread is more likely when people are in close contact with one another (within about 6 feet) but the virus is not considered airborne [12]

Machine learning algorithms have proven to give efficient predictions in healthcare for instance research papers based on deterministic mass action models, regression models, SEIR, ARIMA forecasting models etc. Furthermore, during a pandemic, getting timely and accurate research insights is essential for taking effective countermeasures and reducing economic losses. With limited availability of data most studies on this virus are mostly exploratory. With no effective and well tested vac-

cine for Covid-19 the key part in managing the pandemic has been to decrease the epidemic peak or flattening the epidemic curve.

1.1 Expectations

We take into account number of factors that can put pressure on the number of confirmed cases and mortality rate of Covid-19.
Confirmed Cases: The number of individuals in a population that where tested positive for Covid-19.
Mortality rate: Covid-19 mortality rate was defined as the number of deaths per 100 Covid-19 cases and is therefore a measure of severity among detected cases.These factors are economic factors (which are a category generally recognized by economists as having a major influence on the rate of Covid-19), or a combination of economic and demographic factors, or economic, demographic and institutional factors.
The records entries of our data includes these variables otherwise called *predictors*. The two main question for which this project is seeking to answer are the following questions:

1. *What are the determinants for the number of confirmed Covid-19 cases and*
2. *What factors play a huge role in the mortality rate?*

By answering the first question, will lead to a more effective understanding of each factor and their contributions to the damage or remedy in the pandemic.The answer to the second question will lead to a more effective understanding of each factor and their contributions to the damage or remedy in the pandemic. It is therefore important for scientists to integrate the related data and technology to better understand the virus and its attributes/characteristics, which can help in taking right decisions and concrete plan of actions in developing vaccines and appropriate inferences in possibly eradicating future and similar outbreaks.

2 Data and Methodology

2.1 Data

The original data file was extracted from the Center for Disease Control (CDC) and delivered in the format of Microsoft excel spreadsheet. Initial data browsing was done for variable selection through multiple public data sources. due to quality issues of the data, a request for updated versions of the original data file from CDC

was placed which was later obtained, sorted, prepared and used alongside with data obtained from variety of public data sources for this project. The data set consist of 55 States (observations); fifty states and five major territories of the United States of America and 17 variables. The values for the entries represent the cumulative of the explanatory variables from $2014 - 2020$ and the dependent variable given up to the end of the week of July 22 , 2020. In order to avoid confounding data for Covid-19 with states population sizes , we downloaded data on confirmed cases per million people and indexed the mortality rate through total number of deaths due to Covid-19 divided by the total number of confirmed cases. The data was collected into three main categories

Demographic parameters: The parameters used in this analysis are population density per square miles, population size by states, age dependency ratio , life expectancy at birth.
Environmental and Urban parameters: Percentage of individuals under the federal poverty line, state income for first quarter in 2020, number of homeless individuals by states, percentage humidity, unemployment rates, Medicaid coverage, annual temperature, annual precipitation, percentage of adult smokers and percentage of Obese adults.
Institutional parameters: Because the pandemic is ongoing and lack of sufficient data, the only institutional factors in this analysis that have been considered are number of individuals currently hospitalized and those in intensive care units. For daily analysis of the possible trend of confirmed cases of Covid-19 and confirmed number of deaths, the data of each state was used.

2.1.1 Quality Control

One concern regarding data quality comes from the high percentage of missing (blank) values across the file. As an example the observation for state Puerto Rico and Wyoming has null cells. Most of these cases are due to the fact that data with low frequency (< 5) are suppressed. Suppression include states with low frequency counts and uncommon combinations of demographic characteristics (sex,age groups, race/ethnicity). Another concern is that, outcomes are not yet known at the time of reporting. Suppressed values are re-coded to the NA answer option. Explanation for the variables used in this analysis, data sources and their provenance, including links where the raw data can be extracted directly is shown in Table 1.

Table 1: Explanatory Variables and publicly available data sources used in the analysis

CODE	DESCRIPTION	DATA SOURCES
CNCS	Confirmed Cases	
HOSC	Currently Hospitalized	https://covidtracking.com/data
ICU	Intensive Care Unit	
CDHS	Covid-19 Deaths	https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku/data
PLUF	%Living under federal poverty line	https://data.ers.usda.gov/reports.aspx?ID=17826
SINC	State income for the first quarter of 2020	
POPD	Population density per square miles	https://www.bea.gov/system/files/2020-06/spi0620_0_0.pdf
POPS	Number of homeless	usich.gov
HUMI	Average humidity(%)	http://www.usa.com/rank/us–average-humidity–state-rank.htm
UNEM	Unemployment Rates	https://www.bls.gov/web/laus/lauhsthl.htm
MEDA	Medical Coverage	https://www.kff.org/interactive/medicaid-state-fact-sheets/
LEXP	Life Expectancy by age	https://worldpopulationreview.com/state-rankings/life-expectancy-by-state
ADEP	Age Dependency Ratio	https://worldpopulationreview.com/state-rankings/age-dependency-ratio-by-state
ATEM	Annual Temperature (^{o}F)	https://www.currentresults.com/Weather/US/average-state-temperatures-in-summer.php
APRE	Annual precipitation	https://www.currentresults.com/Weather/US/average-annual-state-precipitation.php
CIGA	Adult smokers (%)	https://www.cdc.gov/statesystem/cigaretteuseadult.html
OBEC	Obesity of adults (%)	https://www.cdc.gov/obesity/data/prevalence-maps.html#states

[1] [2] Table 1 presents detailed definitions of these variables. Table 2 [3] summarizes

Table 2: Characteristics of the study Cohort Up to and Including July 22nd, 2020

	N	Mean	Median	Std.Dev
COVID-19 Deaths	54	2036.83	809.5	2922.16
Currently hospitalized	54	1104.22	403.5	2233.21
In intensive care units	54	193.65	13	543.76
% Living under Federal Poverty line	51	12.86	12.8	2.78
State income for first quarter 2020	51	371629.22	226135	469020.25
Population density	50	203.9	107.7835	267.41
Population Size	50	6611969.86	4572435	7480029.48
# of homeless	51	11113.26	4355	24398.29
% humidity	51	77.57	77.14	2.39
Unemployment rate	51	9.83	8.7	3.09
Medicaid coverage (%)	52	20.15	19	6.20
Life Expectancy by age	50	78.69	78.9	1.69
Age dependency ratio	50	62.32	62.2	3.37
Annual temperature (^{o}F)	50	51.94	51.2	8.71
Annual precipitation	50	36.98	41.75	15.13
% of Adult smokers	51	16.47	16.1	3.26
Obesity in Adults (%)	51	31.29	30.9	3.83
Confirmed COVID-19 cases	54	73081.26	39225	100868.70
Mortality rate	54	2.84	1.98	1.99

Covid-19 mortality rates,Confirmed cases and regression covariates. For the 54 studied states, the average Covid-19 mortality rate was 2.84% (we expect a possible decrease in future as more measures are taking to curb down the pandemic). The mean confirmed number of cases was 73081.26. An exploratory analysis of the number of

[1]Variables collected are from 2018-2020 with all data from CDC most recent

[2]All Variables collected are for each state respectively

[3]The number of observations is based on the sample with missing and non-missing values of all the variables specified in the table. Codes for the variables are provided in Table 1

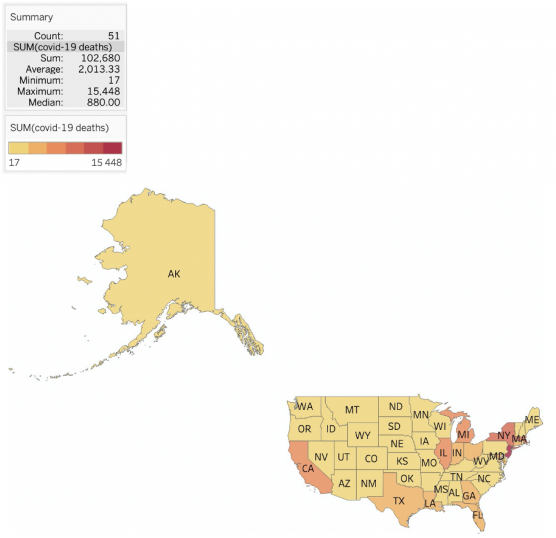


Figure 1: COVID-19 deaths distribution by states

Covid-19 deaths shows high increase in east coast states with highest recorded number of deaths, some parts in the west coast particularly California which are the most affected (Figure 1) [4]. 7485 tests performed for the first quarter with a mean number of confirmed cases of 36457 and a maximum number of positive test of 375133 in the state of New York.High risks states are New york,California,Illinois,Michigan,Florida ,Texas,Georgia with highest number of confirmed deaths in that order .

2.2 Methodology

A Multiple linear regression model Is appropriate for modeling responses of numeric type with one of the underlying assumptions being that the response are continuous and comes from a normal distribution [5]. For a multiple linear regression model with p distinct predictors, and n observations (with Y and X explanatory variables), the model equation is of the form:

$$y_i = \alpha + \sum_{i=1}^{n}\sum_{j=1}^{p} \beta_j (Explanatory\ \ Variables)_{i,j} + \epsilon_i \tag{1}$$

where:
y_i is the observed responses for the response variables.α is the intercept and β_j is the

[4]Accumulate Covid-19 deaths for each states color shows details about states.

coefficient for the j^{th} predictor in X_{ij} (i=1,2...,n) and *i.i.d.* $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
We shall adopt multiple regression analyses to examine the determinants to predict the amount of contribution for each of the known contributors.We perform separate regressions for two dependent variables corresponding to different contributors.

A two-stage modelling approach was used in the analysis

For the first stage, the goal we estimate the determinants for the number of confirmed cases and Covid-19 mortality rate and to assess the ability of this estimation in predicting the variables which are contributors. A multiple linear regression model was chosen to model the relation between predictor variables and mortality rate (MRAT). The goal of the second stage was to locate the factors that have a statistically significant impact . Note that the response variables are two , confirmed cases (CNCS) and Covid-19 mortality rate (MRAT) whose values are on a continuous scale and thus a multi linear regression model was a natural choice.

Software package

The statistical computing package RStudio and Tableau for data visualization was used throughout this project. The choice was partially due to the extensive availability of documentations and technical support and programming flexibility for the RStudio software and data visualization capabilities of the Tableau software. The version for R software is R 3.6.1.
A descriptive statistics of the variables mean,standard deviation and number of observations available for each predictors are on Table 2

Quality Control and data cleaning

Quality control and data cleaning started with the detection of variable with empty data cells. In order to solve this problem, empty cells where deleted. Three addition columns where added onto the data set,the first two columns are for the state codes and respective regions into which the states where all classified, this will provide a clearer picture for our data visualization and the second column, the mortality rate column calculated using the formula[5].

$$MRAT = \frac{CDHS}{CNCS} * 100 \tag{2}$$

[5]According to the *Dictionary of Epidemiology [8]*, the mortality rate is an "estimate of the portion of a population that dies during a specified period"

For the purpose of coding and new variables creation, we have used a four letter abbreviation scheme for our variables which will be loaded onto our statistical software for easy comprehension of our models. For data visualization on the number of Covid-19 deaths accumulated till the period of June 22,2020 all the territory and states which did not report any deaths where suppressed to zero.

correlation

To carry out pearsian correlation analysis amongst the response variables and the demographic, institutional and socio economic factors, a set of data from 50 states excluding missing values from the various U.S territories. The pearsian moment coefficient of correlation shows no strong correlation exist amongst variables except for the positive correlation. The red squares indicates positive correlation amongst variables while the blue squares in Figure 6 indicate negative correlation.

The number of confirmed cases is positively correlated with number of individuals currently hospitalized, individuals in intensive care unit, percentage living under federal poverty line, state income, average population density, population size, percentage of homeless individuals, annual average temperature, and annual average precipitation. Also, mortality rate is positively correlated (Table 5) with population density,unemployment rates, medicaid and live expectancy by age.

Model Construction

Given the structure of our data set a Fixed-Effect-Model was employed to test between two key dependent variables, Covid-19 mortality and number of confirmed cases (positive tests) against the various demographic, economic and institutional factors.A multiple linear regression model was constructed using OLS to estimate the specification of the model. The dependent variables where checked for normality and afterwards, the explanatory variables were also checked for multicollinearity [3] We then construct the model by splitting our data into 70% training data set and 30% testing data set.The 70% data set was used to train the model while the remaining 30% was used for prediction. Next we perform the model validations and checking the model assumptions.

2.2.1 Box-Cox transformation

The responses where highly skewed, so we chose a Box-Cox transformation [5] (See Appendix C).
Figure 2 and Figure 3 shows a histogram of the transformed response with a fitted normal curve.

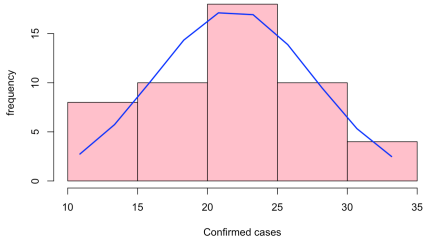


Figure 2: Histogram of the transformed confirmed cases

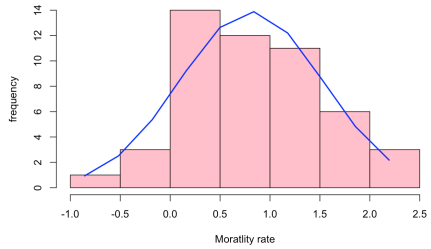


Figure 3: Histogram of the transformed Mortality rates

3 Results and Discussions

Table 3 to Table 7 and Figure 4 to Figure 8 presents the summary result of this study. A histogram was used to check for normality assumption of the dependent variables and then Box-Cox transformation was employed to transform the dependent variables since they were not normally distributed[6]. Variance inflation factors (VIFs) analysis was conducted to check for multicollinearity amongst the independent variables. Using the cut-off point of 10 as suggested by [7] we exclude all variables with VIF greater than 10. These variables include population size, percentage living under federal poverty line and state income for the first quarter of 2020[7].The data was

[6]The response variables where both positively skewed Figure 7 and Figure 8
[7]The maximum VIFs for selected variables is 4.87

split into two parts for training and testing our model [8]. In the next step of the analysis, we identify firm characteristics that affect the number of confirmed cases of Covid-19.We run the following regression model.

$$BT\_CNCS_i = \alpha + \beta X_i + \epsilon_i \tag{3}$$

Where the dependent variable BT_CNCS_i is the box-cox transformation of the direct number of confirmed cases with a $\lambda = 0.15$ (Table 3 summarizes this model). X_i is a vector of firm characteristic variables, including a subset of variables from the general regression equation (1) .
A second regressions to examine the relationship between firm characteristics and COVID-19 mortality rate. This regression model given by:

$$BT\_MRAT_i = \alpha + \beta X_i + \epsilon_i \tag{4}$$

Where the dependent variable BT_MRAT_i is a special case of box-cox[9] of the direct number of mortality rate with a $\lambda \approx 0.0$ (Table 4 summarizes this model). X_i is a vector of firm characteristic variables, including a subset of variables from the general regression equation (1) .
For a robustness check , we fitted several regression models using each of the models in equation (3) and equation (4) with slightly different groups of candidate predictors and significance levels before and after step-wise variable selection [13] were tried and the two models in APPENDIX C ended up being the best two.However to get a parsimonious model [10], an analysis of variance (APPENDIXB) [4] comparison was carried out to check if there models where statistically the same or not.From the analysis of variance (Table 6 and Table 7) ,we conclude that these pair of models are not statistically different, and hence the model with less predictors and smallest BIC after stepwise regression in Table 3 and Table 4 are the optimal models. [10]

Table 3 presents the regression results for number of confirmed cases(CNCS). A total of 50 states were included in the regression analysis.We found that an increase of only 1sq miles in population density is associated with a statistically significant 1.2% increase in the number of confirmed cases.The number of individuals currently hospitalized and number of homeless are important predictors for the number of

[8]The data was split into 70% for training the model and 30% for predictions

[9]When $\lambda = 0$ the box-cox transformation is the natural logarithm

[10]Heteroscedasticity-consistent standard errors (allowing for clustering at the group level) are in parentheses. Definitions of the dependent and independent variables are provided in Table 3. *** p<0.01, ** p<0.05, * p<0.1 denote significance levels of 1%, 5% and 10%, respectively

Table 3: Model 1-Determinants for number of Confirmed cases

Predictors	Estimate	P-value (std.Error)	95% Conf.Interval
HOSC	0.0033	0.0179* (0.0013)	(0.0006 0.0060)
ICU	-0.0072	0.0972 (0.0042)	(-0.0159 0.0014)
POPD	0.0120	0.0013** (0.0033)	(0.0052 0.0189)
HOML	0.0002	0.0001*** (0.00005)	(0.0001 0.0003)
ATEM	0.2302	0.0597 (0.1172)	(-0.0101 0.4707)
APRE	-0.158	0.0430* (0.0746)	(-0.3116 -0.0053)
CIGA	0.6604	0.02326 (0.274)	(0.09690 1.22380)

confirmed cases.We also found that a 1 inch increase in annual precipitation is associated to a statistically significant 1.5% decrease in the number of confirmed cases[11]

[11]Model 1, the R-squared value was 0.72, adjusted R-squared value was 0.65. Model 1 explains approximately 65% of Covid-19 confirmed cases. The difference between the root mean square error (RMSE) for the training and testing data set is 0.11, which justify better fit for Model1.

(annual precipitation was negatively and significantly correlated with Covid-19 confirmed cases).

Table 4 presents the regression results for mortality rate (MRAT). We find that annual temperature was positively and significantly correlated with mortality rate. An increase 1^0F in annual temperature is associated with a statistically significant 7% increase in mortality rate.We also found that the age dependency ratio, life expectancy, population density and individuals currently hospitalized are important predictors for the mortality rate.However during the time of this study the maximum average annual temperature statewide was $51.1^0F(\approx 10.6^0C)$ and maximum annual temperature of $70.1^0F(\approx 21^0C)$ for the state of Florida,the study could not evince a negative effect on temperatures above $70.1^0F(\approx 21^0C)$ .A likely reason may be the lack of quantitative data to explore, or perhaps that Covid-19 could, in fact, fit these higher temperatures. Further studies need to be conducted to discover new findings and determinants.Our results are consistent with previously reported findings that shows the impact of temperature, dry weather and precipitation on human west Nile virus infections [9] [11].A variety of arguments can be given for the positive relationship between temperature and new cases. One could be a hypothesis that people are more prone to break lock-down 'stay-home'rules when the sun is shining outside, so eventually become exposed to the virus. In the contrary,the negative relationship between precipitation and new cases is the reverse: whereby people avoid coming out if it is rainy[12].The Mortality rate estimates from this model monotonically increased as annual temperature and life expectancy increases,supporting the assumption of a linear relationship between Covid-19 mortality rate and annual temperature and a rather surprisingly increase in the mortality rate by 32% with a very small increase (decline) in life expectancy.[13][14] Although predictive capability was the principal feature of interest in these models, residual plots were evaluated to check the usual assumptions of normality and heteroscedasticity and appropriateness of fit [5] A histogram plot is given in Figure 4. No visible clear pattern in the residual plot indicates linearity. For normality Figure 5 indicates no significant deviation from the 45° angle line [15] were detected for both models hence normality.

[12]The linear Model2 predicted R-square was a reasonable indicating that the model explains approximately 62% of the mortality rate.

[13]With all the seven variables that where selected from stepwise regression analysis , each time the models where run in a robust check, at least four variables where statistically significant.

[14]In Table 3 four variables showed strong linear correlation with the number of confirm unlike Table 4 where 5 variables showed strong linear relationship with the mortality rate with Individuals currently hospitalized and population density being statistically significant in both models within the 5% -10% level.

[15]Shapiro-wilk normality test for Model1 with p-value =0.98 and Model2 with p-value=0.27

Table 4: Model 2-Determinants of mortality rate

Predictors	Estimate	P-value (std.Error)	95% Conf.Interval
HOSC	-0.0002	0.001500** (0.000050)	(-0.00031 -0.00008)
POPD	0.0008	0.0371000* (0.0042)	(0.000050 0.001630)
HOML	-0.0000007	0.214570 (0.000005)	(0.00001 0.0000050)
LEXP	0.32080	0.026670* (0.13650)	(0.0401371 0.601384)
ADEP	-0.09687	0.003470** (0.03013)	(-0.158804 -0.034935)
ATEM	0.07079	0,002470** (0.02112)	(0.027300 0.1142000)
CIGA	0.8542	0.105190 (0.05089)	(-0.019174 0.190024)

No heteroscedasticity present therefore equal variance assumption holds for Model 1 and Model 2. [16] However autocorrelation fails to hold for Model 2 [17]. Our re-

[16]studentized Breusch-Pagan test with p-value of 0.7 and 0.6 for Model 1 and 2 respectively

[17]DurbinWatson test for autocorrelation with p-value= 0.004 for Model 1 and p-value=0.87 for Model 2

sults were adjusted for a large set of institutional, demographic and environmental parameters.

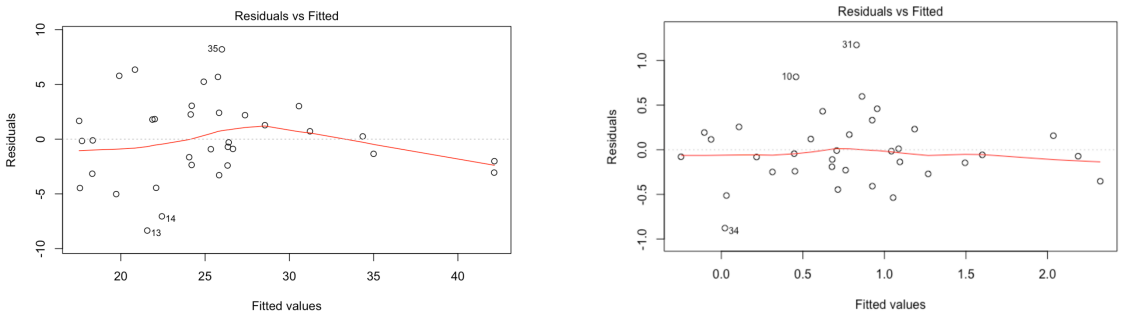
4 Conclusion

This report examines the transmission dynamics and mortality of the novel coronavirus 2019 (Covid-19) considering state by state transmissions.We use a machine learning method (stepwise regression) to select instrumental variables with strong predictive power for the endogenous variables.We find significant and expected associations between most demographic ,socioeconomic factors and the Covid-19 confirmed cases and mortality rates.
We find that Covid-19 mortality is associated with population density,less number of patients currently hospitalized (this is possible due to patients currently undergoing treatment and better healthcare supervisions), life expectancy ,annual temperature and higher age dependency ratio.
We also found that,number of confirm Covid-19 cases increase with population density (this increase is very slow, only 1.2% one reason could be that dense areas have better access to health care facilities and greater implementation of social distancing policies and practices), number of homeless being the most significant predictor and with decrease in average precipitation.

explain result of tableau here.

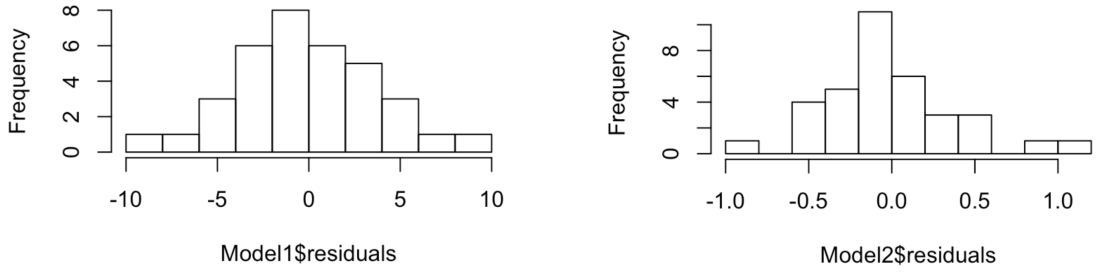
It is important to acknowledge that, this study has several limitations. First this study is based on Covid-19 cases reported by states and therefore less observations with inaccurate reporting and increases in number of cases may have influenced the predictive power of our models.Secondly the Covid-19 related factors used in this study are from state-level data not patient-level data. If patient-level data is made available for analyses, the prediction accuracy will further improve.Thirdly, we selected a limited number of predictors for this study that potentially determine the state's confirmed cases and mortality.Future study may exploit other state related factors to improve the prediction accuracy.However these results can contribute in future pandemic policymaking at state levels.

APPENDIX A : Model Assumptions



(a) Model 1 (b) Model 2

Figure 4: Linearity tests



(a) Model 1 (b) Model 2

Figure 5: Normality tests

APPENDIX B

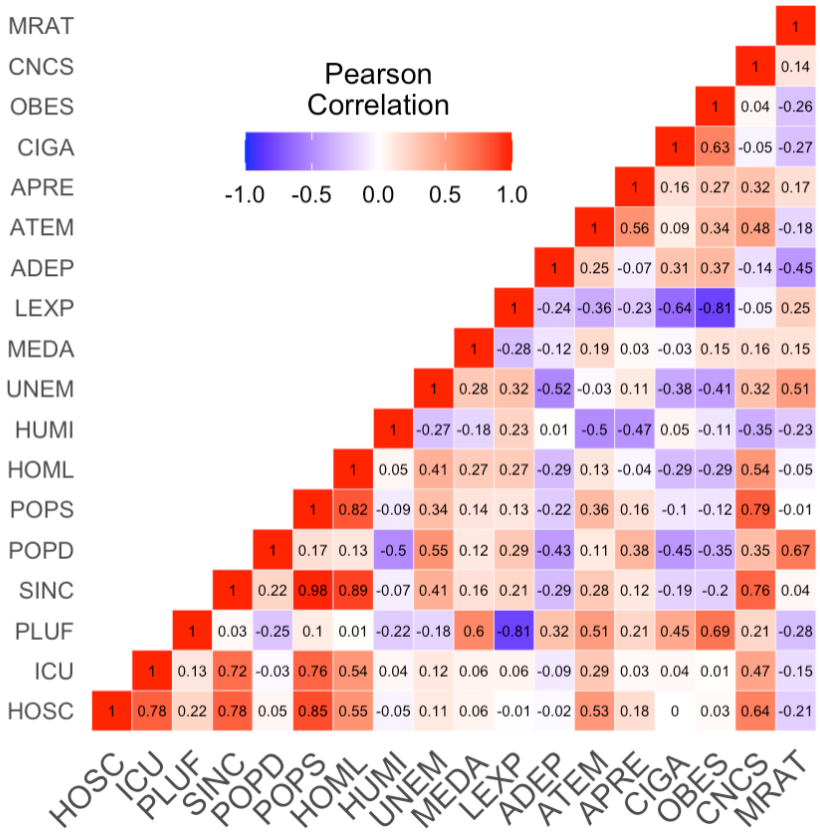


Figure 6: Pearson moment coefficient of correlation

Table 5: Pearson Correlations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
(1) MRAT	1																	
(2) CNCS	.15	1																
(3) OBES	.16	.04	1															
(4) CIGA	.25	−.05	.63	1														
(5) APRE	.18	.32	.27	.16	1													
(6) ATEM	−.09	.48	.34	.09	.56	1												
(7) ADEP	−.44	−.14	.37	.31	−.07	.25	1											
(8) LEXP	.19	−.05	−.81	−.64	−.23	−.36	−.24	1										
(9) MEDA	.21	.16	.15	−.03	.03	.19	−.12	−.28	1									
(10) UNEM	.45	.32	−.41	−.38	.11	−.03	−.52	.32	.28	1								
(11) HUMI	−.22	−.35	−.11	.05	−.47	−.5	.01	.23	−.18	−.27	1							
(12) HOML	−.01	.54	−.29	−.29	−.04	.13	−.29	.27	.27	.41	.05	1						
(13) POPS	.01	.79	−.12	−.1	.16	.36	−.22	.13	.14	.34	−.09	.82	1					
(14) POPD	.58	.35	−.35	−.45	.38	.11	−.43	.29	.12	.55	−.5	.13	.17	1				
(15) SINC	.06	.76	−.2	−.19	.12	.28	−.29	.21	.16	.41	−.07	.89	.98	.22	1			
(16) PLUF	−.18	.21	.69	.45	.21	.51	.32	−.81	.6	−.18	−.22	.01	.1	−.25	.03	1		
(17) ICU	−.14	.47	.01	.04	.03	.29	−.09	.06	.06	.12	.04	.54	.76	−.03	.72	.13	1	
(18) HOSC	−.2	.64	.03	0	.18	.53	−.02	−.01	.06	.11	−.05	.55	.85	.05	.78	.22	.78	1

APPENDIX C : First stage regressions for each dependent variable[18]

lm.fit 1 for CNCS			lm.fit2 for MRAT		
OLS	Estimate	P-value (std.Error)	OLS	Estimate	P-value (std.Error)
HOSC	0.00378	0.030554* (0.001631)	HOSC	-0.00022	0.009710** (0.000076)
ICU	-0.00877	0.104767 (0.005177)	ICU	0.000093	0.706370 (0.000244)
POPD	0.01182	0.020288* (0.004709)	POPD	0.001171	0.09584 (0.000672)
HOML	0.000241	0.000719*** (0.00006)	HOML	-0.000006	0.404060 (0.00006)
HUMI	-0.2135	0.691646 (0.5310)	HUMI	-0.2135	0.438620 (0.000007)
UNEM	0.3189000	0.376360 (0.191600)	UNEM	0.013370	0.740610 (0.039870)
MEDA	0.-0.1780	0.363487 (0.191600)	MEDA	0.0.037090	0.10585 (0.021950)
LEXP	0.138700	0.887162 (0.965700)	LEXP	0.261300	0.13522 (0.168200)
ADEP	0.254100	0.484581 (0.357100)	ADEP	-0.089920	0.01608* (0.034350)
ATEM	0.146700	0.391279 (0.167500)	ATEM	0.072000	0.01086* (0.025760)
APRE	-0.170500	0.059291 (0.085480)	APRE	-0.003015	0.75061 (0.009362)
CIGA	0.537900	0.140667 (0.351300)	CIGA	0.060420	0.38868 (0.068640)

	Res.DF	RSS	DF	Sum of Sq	F	Pr($> F$)
1	21	414.56				
2	27	483.28	-6	-68.722	0.5802	0.742

Table 6: Anova table for lm.fit1 and Model1 CNCS

	Res.DF	RSS	DF	Sum of Sq	F	Pr($> F$)
1	21	5.0614				
2	26	5.3149	-5	-0.25344	0.2103	0.9544

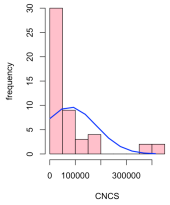
Table 7: Anova table for lm.fit2 and Model2 MRAT

[18]lm.fit1 and lm.fit2; The model for number of confirmed covid cases and mortality rate after VIFs selections was done respectively.

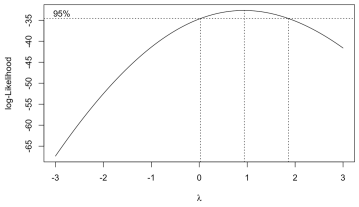
[18]Model1 and Model2; The models after stepwise regression analysis was perform for number of confirmed cases and mortality rate respectively.

APPENDIX D :Box-Cox transformation

The first phase of the analysis starts with an initial check for the necessity of transformation on the response variables (Confirm cases and the Mortality rates). Figure 7a shows the histogram of the response variable with a fitted normal curves.Clearly there is no way to believe it comes from a normal distribution. So a transformation is necessary here. The technique of Box-Cox transformation [5] is then utilized to optimally locate the choice of transformation. Figure 8b illustrate how the log-likelihood changes with the choice of different λ , the order of the transformation. Both the software printout and the line plot led to the choice of $\lambda = 0.202$ for confirmed cases and 0.0038 for Mortality rates which corresponds to a natural log transformation on the confirmed cases. Figure 6 shows the histogram along with a fitted normal curve of the transformed responses which presents a much more plausible shape for the confirmed cases.

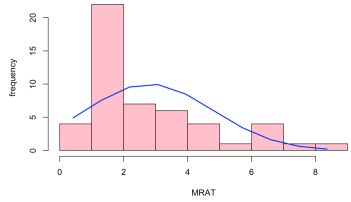


(a) Histogram:Confirmed cases

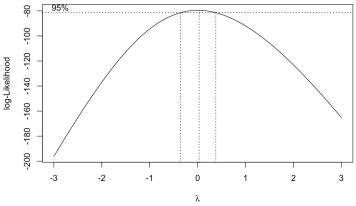


(b) plot of Box-cox result

Figure 7: Confirmed Cases



(a) Histogram:Mortality rate



(b) plot of Box-cox result

Figure 8: Mortality rate

References

[1] CDC Covid, CDC COVID, CDC COVID, Nancy Chow, Katherine Fleming-Dutra, Ryan Gierke, Aron Hall, Michelle Hughes, Tamara Pilishvili, Matthew Ritchey, et al. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—united states, february 12–march 28, 2020. *Morbidity and Mortality Weekly Report*, 69(13):382, 2020.

[2] Fang Jiang, Liehua Deng, Liangqing Zhang, Yin Cai, Chi Wai Cheung, and Zhengyuan Xia. Review of the clinical characteristics of coronavirus disease 2019 (covid-19). *Journal of general internal medicine*, pages 1–5, 2020.

[3] G Maddala. S.(1988), introduction to econometrics, 1988.

[4] Mary L McHugh. Multiple comparison analysis testing in anova. *Biochemia medica: Biochemia medica*, 21(3):203–209, 2011.

[5] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. Applied linear statistical models. 1996.

[6] World Health Organization et al. Middle east respiratory syndrome coronavirus (mers-cov), 2019.

[7] Robert M O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007.

[8] Miquel Porta. *A dictionary of epidemiology*. Oxford university press, 2014.

[9] Marilyn O Ruiz, Luis F Chaves, Gabriel L Hamer, Ting Sun, William M Brown, Edward D Walker, Linn Haramis, Tony L Goldberg, and Uriel D Kitron. Local impact of temperature and precipitation on west nile virus infection in culex species mosquitoes in northeast illinois, usa. *Parasites & vectors*, 3(1):19, 2010.

[10] Joachim Vandekerckhove, Dora Matzke, Eric-Jan Wagenmakers, et al. Model comparison and the principle of parsimony. *Oxford handbook of computational and mathematical psychology*, pages 300–319, 2015.

[11] Guiming Wang, Richard B Minnis, Jerrold L Belant, and Charles L Wax. Dry weather induces outbreaks of human west nile virus infections. *BMC infectious diseases*, 10(1):38, 2010.

[12] Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *Jama*, 323(13):1239–1242, 2020.

[13] Tian Yu, Guang Yu, Peng-Yu Li, and Liang Wang. Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, 101(2):1233–1252, 2014.