
# 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**

**October 23, 2020**

.

# Contents



# List of Figures

**Abstract**

# 1 INTRODUCTION

## 1.1 Project Overview

### 1.1.1 Backgroud

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 on 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. Coronavirus disease 2019 (COVID-19) is an infectious disease caused by 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) 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 [8]

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

### 1.1.2 Expectations

We will take into account number of factors that can put pressure on the confirmed and mortality rate of COVID-19 , these factors can either be 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 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 key factors that influence 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.

## 1.2 Data Description

The original data file was extracted from the Center for Disease Control (CDC) data base and delivered in the format of Microsoft excel spreadsheet. Initial data browsing was then 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 then 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 observations (States) and 18 variables. The file includes fifty states and five major territories of the United States of America. 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. (After this date the data will be updated and future recollection of the data was done until the project was presented). 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 the virus divided by the total number of confirmed cases.

### 1.2.1 Model Specification

A Multiple linear regression model Is appropriate for modeling responses of numeric type with one of the underlying assumptions being that the response 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:

$$y_i = \beta_0 + \sum_{i=1}^{n}\sum_{j=1}^{p} \beta_j \textit{(Explanatory Variables)}_{i,j} + \epsilon_i$$

where:
$y_i$ is the observed responses for the response variables.$\beta_0$ is the intercept and $\beta_j$ is the 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 utilize this method to predict the amount of contribution for each of the the known contributors.

# 2 MATERIALS AND METHOD

*A two-stage modelling approach was used in the analysis*

For the first stage, the goal was to estimate the key factors that influence the number of confirm cases and Covid-19 mortality 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.

*Condition for Analysis*

**Demographic parameters** in the analysis included population density per square miles, population size by states, age dependency ratio , life expectancy at birth.
**Environmental and Urban parameters** in this analysis included percentage of individual 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 the 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.

*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.

*Software package and data*

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.
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. A descriptive statistics of the variables mean,standard deviation and number of observations available for each predictors are on Table 2

Table 1: **Explanatory Variables and publicly available data sources use 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://worldpopulation review.com/state-rankings/age- dependency-ratio-by-state |
| ATEM | Annual Temperature ($^o$F) | https://www.currentre sults.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 |

Note:1. Variables collected are from 2018-2020 with all data from CDC most recent
2. All Variables collected for each state.

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

| | N | Mean | Std.Dev |
|---|---|---|---|
| COVID-19 Deaths | 54 | 2036.83 | 2922.16 |
| Currently hospitalized | 54 | 1104.22 | 2233.21 |
| In intensive care units | 54 | 193.65 | 543.76 |
| % Living under Federal Poverty line | 51 | 12.86 | 2.78 |
| State income for first quarter 2020 | 51 | 371629.22 | 469020.25 |
| Population density | 50 | 203.9 | 267.41 |
| Population Size | 50 | 6611969.86 | 7480029.48 |
| # of homeless | 51 | 11113.26 | 24398.29 |
| % humidity | 51 | 77.57 | 2.39 |
| Unemployment rate | 51 | 9.83 | 3.09 |
| Medicaid coverage (%) | 52 | 20.15 | 6.20 |
| Life Expectancy by age | 50 | 78.69 | 1.69 |
| Age dependency ratio | 50 | 62.32 | 3.37 |
| Annual temperature ($^o$F) | 50 | 51.94 | 8.71 |
| Annual precipitation | 50 | 36.98 | 15.13 |
| % of Adult smokers | 51 | 16.47 | 3.26 |
| Obesity in Adults (%) | 51 | 31.29 | 3.83 |
| Confirmed COVID-19 cases | 54 | 73081.26 | 100868.70 |

# 3 DATA PREPARATION

## 3.1 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, the models was run with the empty cells and another model ran without the empty cells to check if the variable significance will be change since our observation points where 55. It turned out that more variables where significant without the empty cells and so these 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 which was calculated using the following expression.

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

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 22nd,2020 all the territory and states which did not report any deaths where suppressed to zero.

*Exploratory Analysis*
An exploratory analysis of the number of COVID-19 deaths shows high increase in east coast states with highest re-coded number of deaths, some parts in the west coast particularly California which

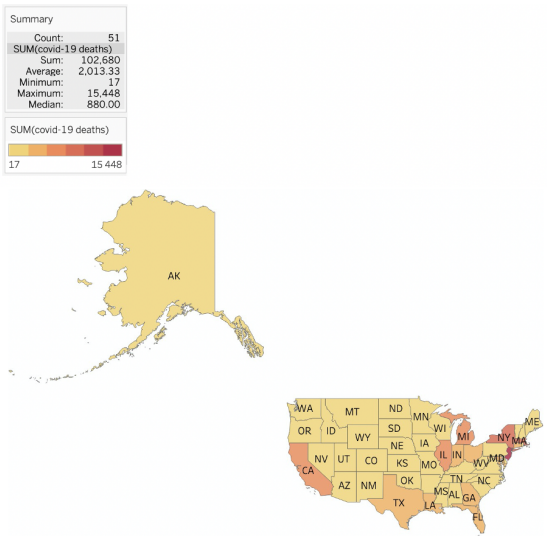



Figure 1: Covid-19 deaths distribution by states

are the most affected (figure 1) [1]. 7485 tests performed for the first quarter with a mean number of confirmed cases of 36457 and a Max number of positive test of 375133 in the state of New York.

*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 exluding missing values from the various U.S territories. The pearson 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 5 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 (Figure 5) with population density,unemployment rates, medicaid and live expectancy by age. (table goes here for the correlation)

## 3.2 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 COVID-19 cases (positive test) 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

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

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 is then used to train the model while the remaining 30% was used for prediction. Next we perform the model validations and checking the model assumptions.

### 3.3 Box-Cox transformation

The responses where highly skewed, so we chose a Box-Cox transformation [5](See Appendix C for more information),which turned out to be a natural log.
Figure 5 shows a histogram of the transformed response with a fitted normal curve.

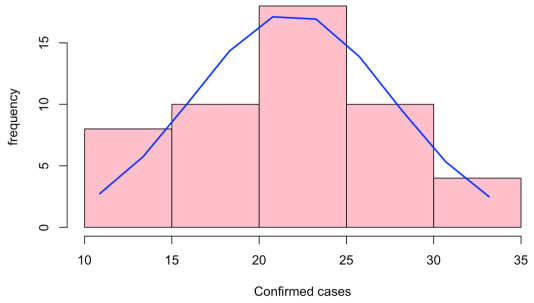



Figure 2: Histogram of the transformed confirmed cases

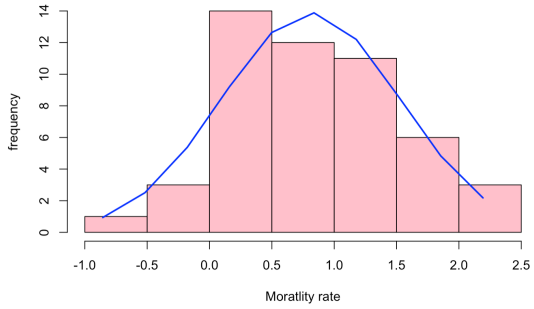



Figure 3: Histogram of the transformed Mortality rates

# 4 RESULTS AND DISCUSSIONS

Table 2-Table 5 and Figure 5-Figure 9 present the summary result of this study .In both models, A probability plot (Q-Q plot) 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. Variance inflation factor analysis was conducted to check for multicollinearity amongst the independent variables. Using the cut-off point of 10 as suggested by [6] 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.Our data was then split into two parts for training and testing our model.

For the first stage, we fitted Several linear regression models using our first response for Confirmed cases with slightly different groups of candidate predictors and significance levels for step-wise variable selection [9] were tried and the two models in Table (See APPENDIX B) ended up being the best two.However to get a parsimonious model [7], an analysis of variance (ANOVA,APPENDIX B) [4] comparison was carried out to see if there where any significant difference between our two models and we ended up with conclusion that the two models where not statistically different, and hence the model with less predictors after subset selection in table 3 is the optimal model.

In table 3 approximately 65% of the variation has been explain by the model. The difference

Table 3: **Factors influencing Confirmed cases**

| | Estimated Coefficients | Standard Error | 95%CI | P-value |
|---|---|---|---|---|
| Currently Hospitalized | 0.0033 | 0.0013 | (0.0006 0.0060) | 0.0179 |
| In intensive care units | -0.0072 | 0.0042 | (-0.0159 0.0014) | 0.0972 |
| Population density | 0.0120 | 0.0033 | (0.0052 0.0189) | 0.0013 |
| # of homeless | 0.0002 | 0.0000 | (0.0001 0.0003) | 0.0001 |
| Annual temperature | 0.2302 | 0.1172 | (-0.0101 0.4707) | 0.0597 |
| Annual Precipitation | -0.158 | 0.0746 | (-0.3116 -0.0053) | 0.0430 |
| % of Adult Smokers | 0.6604 | 0.2746 | (0.0969 1.2238) | 0.0232 |
| Adjusted R-Squared | | | | 64.77% |
| Model p-value | | | | 0.000004374 |
| RMSE.Training data | | | | 0.72 |
| RMSE.Testing data | | | | 0.61 |

between the root mean square error(RMSE) for the training and testing data set is 0.11, which justify better fit of the models.Although predictive capability was the principal feature of interest in these models, residual plots were evaluated to check the usual assumptions of normality and homoscedasticity and appropriateness of fit [5] The normal probability plot is given in Figure 3.5 as an example. No substantial deviations from these assumptions were detected.

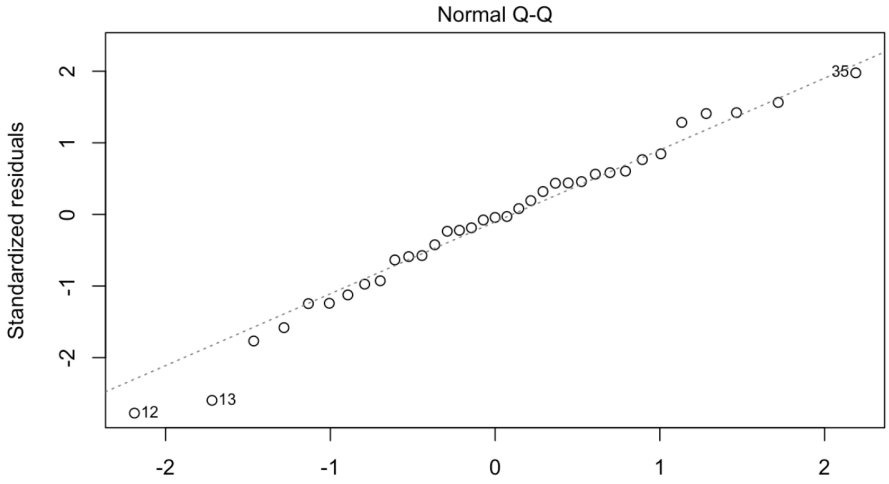



Figure 4: Residual diagnotics

# APPENDIX A

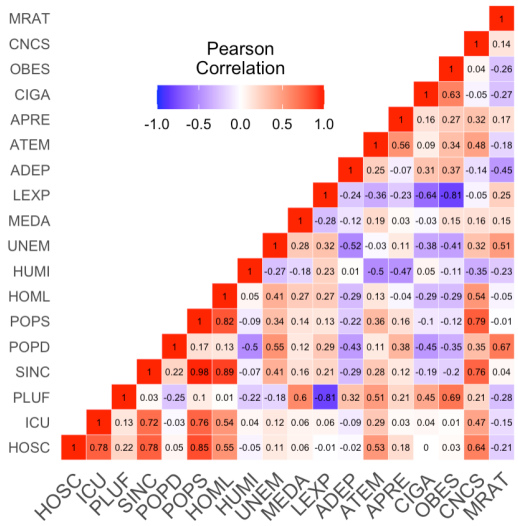



Figure 5: Pearson moment coefficient of correlation

# APPENDIX B

```
Call:
lm(formula = CNCS ~ HOSC + ICU + POPD + HOML + HUMI + UNEM +
    MEDA + LEXP + ADEP + ATEM + APRE + CIGA + OBES, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4505 -2.1008 -0.4406  1.2592  6.8516

Coefficients:
              Estimate   Std. Error t value Pr(>|t|)
(Intercept) -12.18021645  96.39786233  -0.126 0.900654
HOSC          0.00378307   0.00163127   2.319 0.030554 *
ICU          -0.00877732   0.00517698  -1.695 0.104767
POPD          0.01182435   0.00470896   2.511 0.020288 *
HOML          0.00024177   0.00006109   3.958 0.000719 ***
HUMI         -0.21353629   0.53100133  -0.402 0.691646
UNEM          0.31893260   0.35288466   0.904 0.376360
MEDA         -0.17802699   0.19165000  -0.929 0.363487
LEXP          0.13869756   0.96566288   0.144 0.887162
ADEP          0.25410655   0.35712355   0.712 0.484581
ATEM          0.14666632   0.16754931   0.875 0.391279
APRE         -0.17045604   0.08548180  -1.994 0.059291 .
CIGA          0.53792675   0.35133011   1.531 0.140667
OBES          0.36395804   0.36962373   0.985 0.335995
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.443 on 21 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.6114
F-statistic: 5.115 on 13 and 21 DF,  p-value: 0.0004908
```

(a) model before stepwise selection

```
Call:
lm(formula = CNCS ~ HOSC + ICU + POPD + HOML + ATEM + APRE +
    CIGA, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3508 -2.3831 -0.1628  2.2221  8.2031

Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  2.56475184  7.05013461   0.364 0.718850
HOSC         0.00331860  0.00131593   2.522 0.017879 *
ICU         -0.00728464  0.00423924  -1.718 0.097175 .
POPD         0.01202941  0.00334939   3.592 0.001290 **
HOML         0.00023112  0.00005078   4.551 0.000102 ***
ATEM         0.23032439  0.11719831   1.965 0.059750 .
APRE        -0.15848870  0.07463990  -2.123 0.043034 *
CIGA         0.66043119  0.27461020   2.405 0.023298 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.231 on 27 degrees of freedom
Multiple R-squared:  0.7202,    Adjusted R-squared:  0.6477
F-statistic: 9.928 on 7 and 27 DF,  p-value: 0.000004374
```

(b) after stepwise selection

Figure 6: Two main models

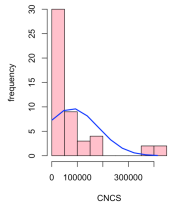
```
Analysis of Variance Table

Model 1: CNCS ~ HOSC + ICU + POPD + HOML + HUMI + UNEM + MEDA + LEXP +
    ADEP + ATEM + APRE + CIGA + OBES
Model 2: CNCS ~ HOSC + ICU + POPD + HOML + ATEM + APRE + CIGA
  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
```

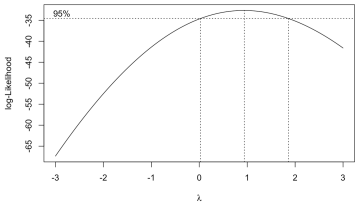
Figure 7: Analysis of variance

# APPENDIX C: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 8a 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 9b illustrate how the log-likelihood changes with the choice of different $\lambda$ , 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 5 shows the histogram along with a fitted normal curve of the transformed responses which presents a much more plausible shape for the confirmed cases.

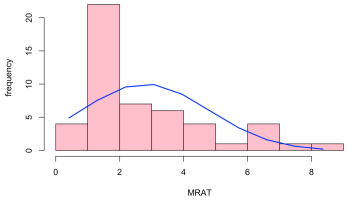



(a) histogram of the confirmed cases

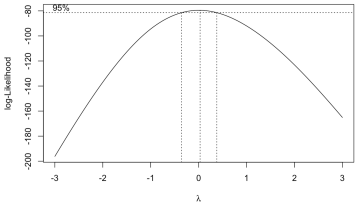



(b) plot of Box-cox result

Figure 8: Confirmed Cases




(a) histogram of Mortality rate




(b) plot of Box-cox result

Figure 9: 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] Robert M O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007.

[7] 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.

[8] 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.

[9] 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.