Regression Analysis of Statewide COVID-19 Data in the U.S.

Marqus Parker[1], Dr. Guoquing Tang[1]
North Carolina Agricultural and Technical State University[1]

Abstract

Pandemics can cause social, political, and economic turmoil that can interfere with the peoples' lives and everyday occupations. The COVID-19 pandemic is a virus spread from person to person through the release of respiratory substances generated by a cough or sneeze according to the National Institute of Health (NIH). The virus's mode of transmission has inspired the creation of global social distancing laws and transitions to a form of virtual proceedings for many professional and educational settings. Researchers have been studying the COVID-19 pandemic in order to create models that predicts the total number of cases and deaths that caused by the virus. In this study, a multiple linear regression and nonlinear regression model was derived to predict the total number of COVID-19 deaths since January 2020 in daily increments for each state in the United States. Multiple linear regression and Nonlinear regression models developed in this study in R and Python and the data used to plot daily U.S. state data have been generated from the Johns Hopkins University's Github Repository. The performance of the linear regression model features a significant p-value of < 2e-16 while the nonlinear regression holds a significant p-value of <0.001. This study will assist doctors and researchers in developing methods of mitigation to the spread of the COVID-19 pandemic. Based on the predictions received by the generated models, forecasting of COVID-19 deaths could be observed over various period of time.

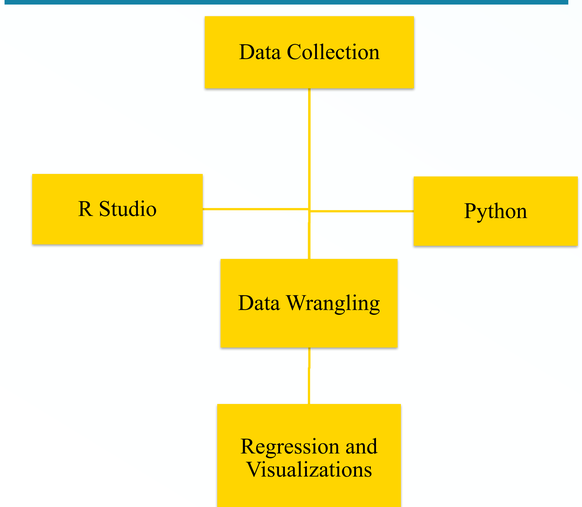
Introduction

- SARS-COV-2: Severe Acute Respiratory Syndrome
- December 2019 – COVID-19 emerged in Wuhan, China.
- Causes a respiratory sickness that transmits through substances released during a cough or sneeze.
- Symptoms usually show within 14 days of exposure.
- March 11,2020 – World Health Organization (WHO) declared COVID-19 as a global pandemic.
- The virus forced states in the U.S. to issue statewide social distancing orders.

Objectives

- The aim of this project is to test the significance of different factors contributing to the total deaths caused by COVID-19.
- To execute this study, a multiple linear regression and a nonlinear regression will be the chosen method of approach.

Methods



Data Description and Regression Models

A subset of the dataset retrieved from the Johns Hopkins's Github Repository was created containing the necessary variables to carry out the regression models with as shown in Figure 1.

- **Multiple Linear Regression Equation**

$$Y = \beta_0 += \beta_0 + \sum_{i=1}^{5} \beta_i x_i + \epsilon$$

- **Nonlinear Regression Equation**

$$Y = f(X_1, X_2, X_3, X_4, X_5, \beta)$$

Figure 1: On the left, are the equations used to perform each type of regression. On the right, are the dependent and independent variables used in the experiment.

- **Data**
  - JHU daily updated data
    - 18 columns by 58 observations
    - Consists of statewide cumulative COVID-19 data
    - Used for model creation
  - JHU time series data
    - 13 columns by 8925 observations
    - Consists of new and cumulative COVID-19 cases
    - Used for plot creation

- **Y** – Cumulative Deaths
- **X1** – Confirmed Cases
- **X2** – Recovered Cases
- **X3** – Active Cases
- **X4** – People Tested
- **X5** – People Hospitalized

Analytical Results and Visualization

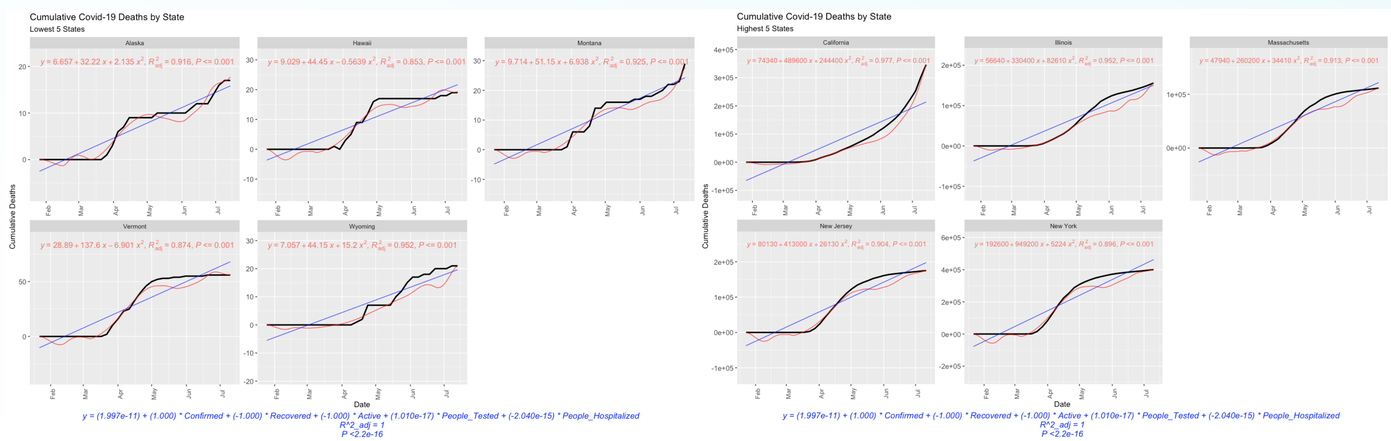


Figure 2: Visual plots of daily COVID-19 deaths in all states. Alaska has been enlarged to compare both models to actual data.

Statistical Analysis

- **Predictor Variable Significance**

Variable	P-value
Confirmed cases	< 2e-16
Recovered Cases	< 2e-16
Active Cases	< 2e-16
People Tested	0.177716
People Hospitalized	0.000939

Table 1: Tables containing coefficient estimates and their respective p-values generated for the model.

- A p-value of < 2e-16 generated from the multiple linear regression model signifies that there is a significance between Cumulative Deaths and the predictor variables.
- A p-value of _____ generated from the nonlinear regression model signifies that there is a significance between Cumulative Deaths and the predictor variables.

Future Discussion

- To implement this project further, the regression models could be extended to countywide data to predict cumulative deaths over each county.
- Demographics such as age, gender, and race could be an implementation in order to further describe how different groups of people are affected by COVID-19.

References

1. Sauer, L. (n.d.). What Is Coronavirus? Retrieved July 30, 2020, from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus
2. Ghosal, Samit et al. "Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020)." Diabetes & metabolic syndrome vol. 14,4 (2020): 311-315. doi:10.1016/j.dsx.2020.03.017
3. Pandey, Gaurav & Chaudhary, Poonam & Gupta, Rajan & Pal, Saibal. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. 10.1101/2020.04.01.20049825.

Acknowledgements

This work is supported by the National Science Foundation under Grant HRD-1719498. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the funding agencies. I would like to thank Dr. Guoquing Tang, Dr, Tamer Elbayoumi, Brett Hunter, Yang Xue, and Chandra Manivanan for their assistance with the project. I would also like to thank the Department of Mathematics and the HBCU-UP at North Carolina A&T. I would also like to acknowledge Dr. Kossi Edoh for providing me with the chance to conduct research.