Forecasts of Total Deaths and Confirmed Cases

Pritam Shrestha Bellevue University

1000 Galvin Rd S, Bellevue, NE 68005

prishrestha@my365bellevue.edu

**Abstract**

The outbreak of coronavirus disease 2019 (COVID-19) has created a global health crisis and impacting the global economy and our daily lives. Due to the fast spread of the coronavirus, social distancing, use of hand sanitizer, frequent hand wash have come as major prevention activities. Coronavirus disease (COVID-19) is an inflammation disease from a new virus. Basically, it affects the respiratory system of the human and its known symptoms are fever, common cold, and cough. But it could make severe damage to the respiratory system and breathing problems. Most of the countries of the world are facing this virus because until now there is no known medicine to control it. Vaccines are being trialed worldwide and hopefully, we will get it very soon.

Here, I presented a model that could be useful to predict the spread of COVID-2019, estimated death, and recovery percentage. For that, I will perform a regression model based on the Kaggle datasets.

**Intro/Background of the problem**

Since the first known case of COVID-19 in Hubei china, 9451515 have tested positive, 482301 have died and 5111629 have recovered until today (6/24/2020). Gradually mortality rate is decreasing as of today due to the several actions have taken by countries. Every nation implemented different types of models to stop the spread of covid\_19 such as lockdown, social distancing, frequent hand washing, quarantine, and isolation, etc. After a few months, most of the countries are reopening based on the WHO, CDC guidelines but it is impacting to stop the spread of COVID-19. I just watched, governor Gavin Newsome reviled that after reopening California, positive test case has increased by 69% which is warning for future and human outside activities. So COVID-19 is impacting every sector of our life, such as the economy, health, social activities, sports, schools, and so on. Due to the loneliness, growing fear, bad economic status a lot of people is being depressed which could lead to the mental health crisis in the world.

Hence, it is clear that it is affecting everywhere so it is called a pandemic. Here, I am planning to use a couple of datasets related to COVID-19 and will extract some insights based on the available data. First of all, I will work on four datasets to predict death, recovery, and positive test case. More precisely, I will work on the different datasets and if needed, I will merge, join, and drop the variables. Besides that, I will apply the best fit model to answer my hypothetical questions and to solve the problems.

**Methods**

There are several methods to solve this problem but I am planning to use a regressing model right now. In the long run, if I found another best model for a prediction I will go through it. In statistics, Linear Regression (LR) is a direct way to deal with demonstrating the connection between a dependent variable and at least one independent variable. It is the main kind of regression analysis and used widely to solve the prediction problems discovering relation among the variables. Here one variable is viewed as an independent and the other is viewed as a dependent. An LR line has a condition of the structure:

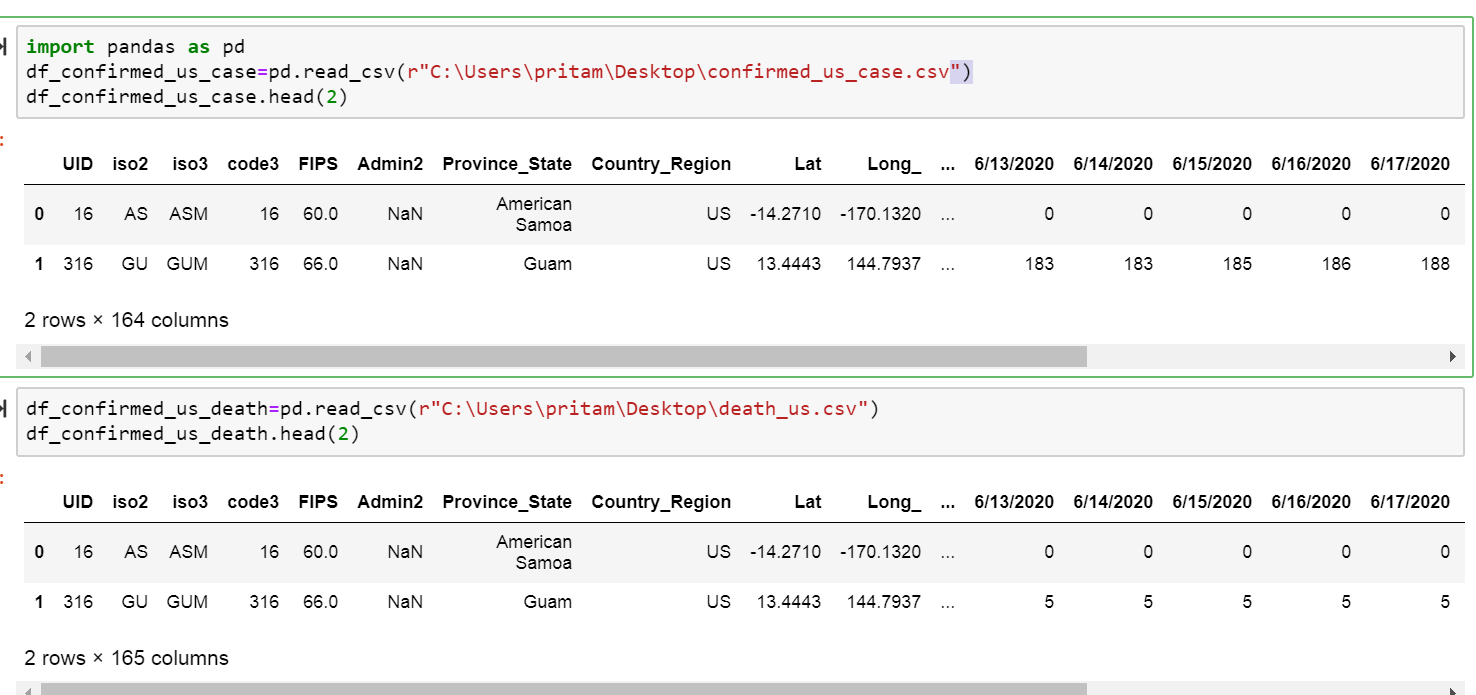
Y=mX+c

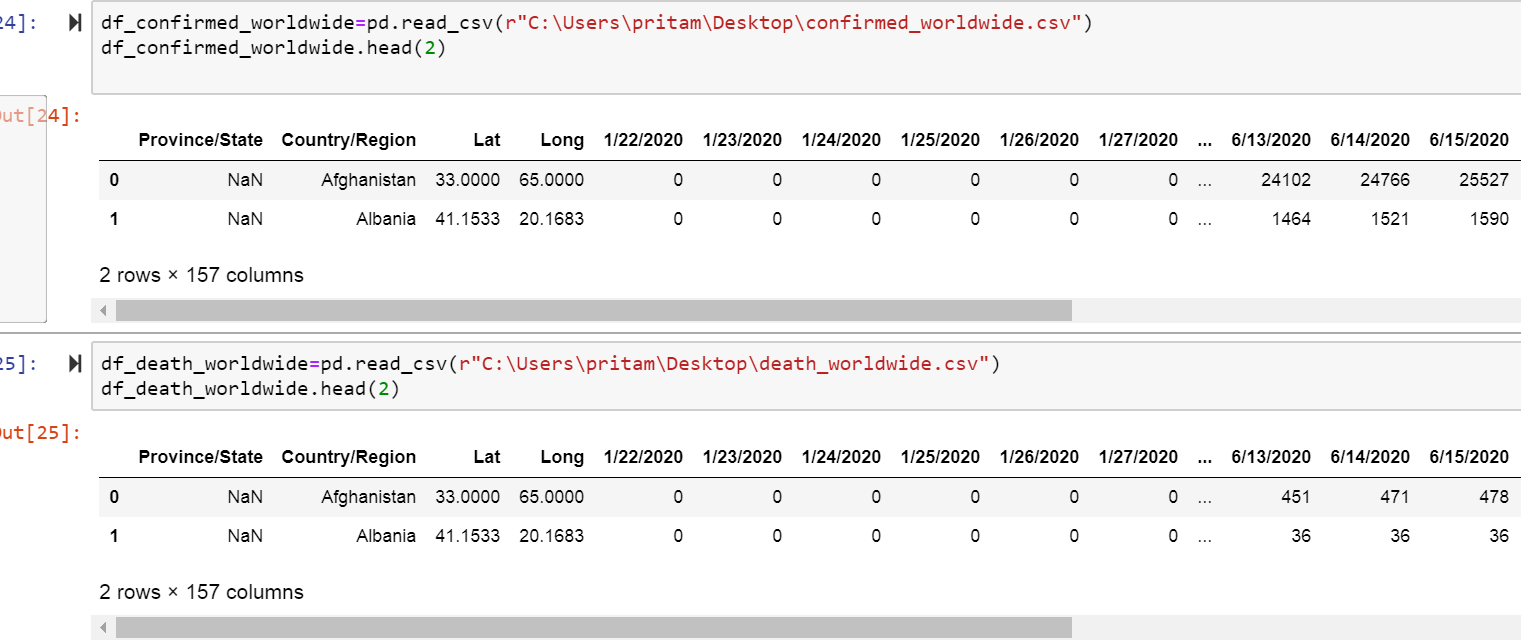
Above X is the independent and Y is the dependent variable. The slope of the line is m and c is the intercept (the value of y when x = 0).

To achieve my goal, I am planning to use four different types of datasets which are given below.

1. **df\_confirmed\_us\_case**=pd.read\_csv(r"C:\Users\pritam\Desktop\confirmed\_us\_case.csv")
2. **df\_confirmed\_us\_death**=pd.read\_csv(r"C:\Users\pritam\Desktop\death\_us.csv")
3. **df\_confirmed\_worldwide**=pd.read\_csv(r"C:\Users\pritam\Desktop\confirmed\_worldwide.csv")
4. **df\_death\_worldwide=**pd.read\_csv(r"C:\Users\pritam\Desktop\death\_worldwide.csv")

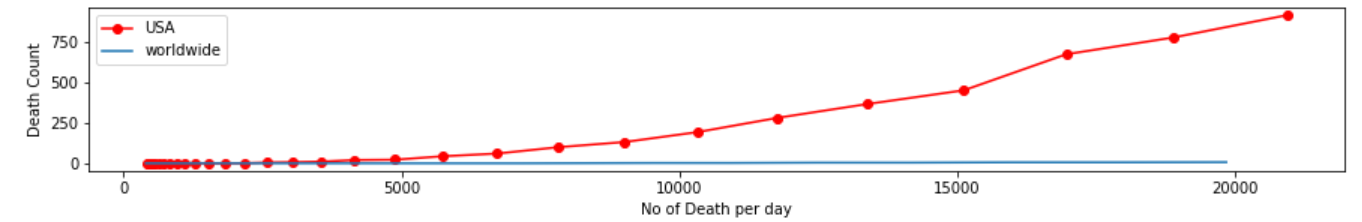
**Structure of data**

****



**Results**

The success and failure of the project depend on the achieved output or result. For this project, I will work on four different data types and will discover the relationship between them and find the output as a prediction. Here, I have taken four datasets among them two datasets are related to the USA and two are related to the entire world. When I predict the case and death I will compare the obtained result with a worldwide case like below.



**Source: Covid\_19\_DeathComparision\_by\_pritam\_shrestha**

Correlation plays a great role in finding the dependency among the features of the dataset. Our dataset revolves around the confirmed, recovered, and death of cases because of the COVID-19 outbreak over the time frame of around 6 months in the USA and worldwide. So I will implement Pearson’s correlation method to find the strong bound between variables. The main purpose of this project is finding or discovering the prediction of death and positive test case and based on the prediction finding the solution to minimize them and save human life.

**Discussion/Conclusion**

Coronavirus disease (COVID-19) is an inflammation disease from a new virus and spreading all over the world contacting its surrounding. It means it can spread through close contact of an infected person, food, and other various materials used by an infected person. So, until now there are no vaccines and medicine have invented. Due to this bitter truth, we have only one method to stop it is social distancing. There are so many biomedical companies are working rapidly to invent vaccines such as Gillard, A foster city-based company’s vaccine Remdesivir is being used for trial and several vaccines are also being used for the second phase of the trial. Besides that so, many data scientists are also working to discover the insights about it. Recently, I have seen my professor Fadi Alsaleem has also invented a new concept to minimize the coronavirus spread using kinsa thermometers.

To complete this project is very challenging because I am planning to use four different datasets and I have to do a lot of data cleaning, transformation, duplicate finding, data conversions, and best fit model implementation. Besides that, I have to use two different programming languages as minimum requirements.

So, I am planning to do all exploratory data analysis using python and will do modeling and plotting using R. I am sure that I will face a couple of problems while doing this project but I will give my 100% and work under the supervision of my professor Fadi Alsaleem. In fact, I am beginning of the project so I have no idea what model works better so if I found the best model other than what I am planning to implement will change with the permission of my professor.

**Acknowledgments**

I can’t express enough thanks to my professor Fadi Alsaleem for his continuous support and encouragement.This is just starting with the project and many more come in the future. Starting of this project could not have been accomplished without the support of my classmates and beautiful daughter Prisumsa.

Thanks to my wife Suchan as well. The countless times you kept the children during our hectic schedule will not be forgotten.

**References:**

<https://link.springer.com/article/10.1007/s00477-020-01827-8>

<https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv>

<https://www.frontiersin.org/research-topics/13638/coronavirus-disease-covid-19-the-impact-and-role-of-mass-media-during-the-pandemic>