CAPSTONE PROJECT

UDACITY MACHINE LEARNING ENGINEER NANODEGREE

1. Project Overview

1.1. Problem Statement

On December 31st, 2019, the outbreak of "pneumonia of unknown cause" started in Wuhan City, Hubei Province, China. The virus was named novel Coronavirus, or 2019-nCoV in short. At January 23rd, there had been over 800 cases of 2019-nCoV confirmed globally, and it is still raising despite many effort from governments.

Amid the outbreak of COVID-19, it is a critical need to analyze and forecast the progress of the pandemic and the effective of current policy. An important task is to predict the number of confirmed cases and fatalities globally, and in each country in order to have appropriate preparation and policy. Many data scientist are trying to use AI to forecast these numbers by studying mass social interaction, analyzing patients travel records, and clustering high risk areas.

The goal of this project is to use time series record by days to forecast the number of cases (confirmed and deceased) in the next 14 days.

The method I use for modelling is inspired from this Kaggle notebook:

https://www.kaggle.com/saga21/covid-global-forecast-sir-model-ml-regressions

1.2. Metrics

In the context of this project, because of limited resource and data, I will only conduct Exploratory Data Analysis and use basic log-linear and logistics functions model to predict the number of cases based on time series data.

The metrics that I will use to evaluate the model is Root mean squared error (RMSE)

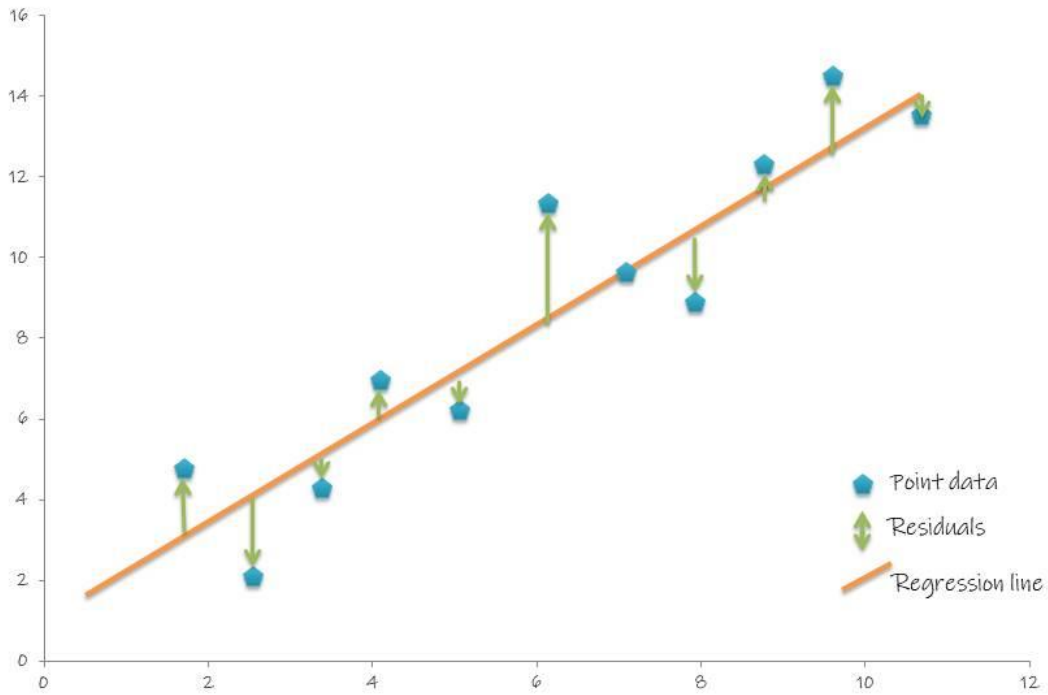

Figure 1: Root mean square error. Source: https://www.hatarilabs.com

To calculate the RMSE, this equation is used:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2}$$

With:

- n: number of samples
- p: predicted value
- o: observed value

Unlike other classification problem where we can use metrics like accuracy, precision, recall, because this is a regression problem, the RMSE should be effective enough to evaluate the performance of the model.

2. Exploratory Data Analysis

2.1. Dataset

There are many data sources available online, providing many insight and data in various aspect of the diseases. I will use the time series dataset collected and updated daily by Center for Systems Science and Engineering, John Hopkins University. The dataset can be downloaded from this github repository: https://github.com/CSSEGISandData/COVID-19

The time series dataset is recorded with daily number of cases so we do not need to do a lot of data processing. It contains 4 region information fields: "Province/State", "Country/Region", "Lat", "Long"

- "Province/State": for China, the numbers are measured in scale of province
- "Country/Region": country like US, Spain, Vietnam
- "Lat" and "Long" is the latitude and longitude of the region
- The other fields are number of confirmed or death cases, counting from 01/22/2020

2.2. Exploratory Data Analysis

I will conduct some basic data aggregations and visualization to have a sense of the distribution of data and trend of the number, and to see what model can be applied to the data

The plots below shows the number of confirmed and decease cases globally and in some other countries.

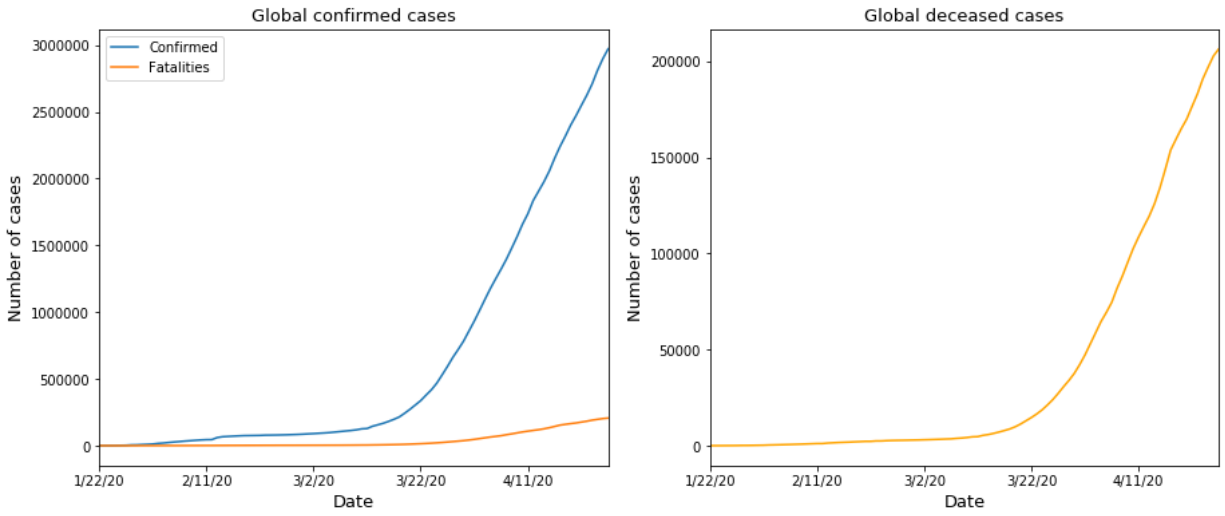

Figure 2: Global confirmed and deceased cases

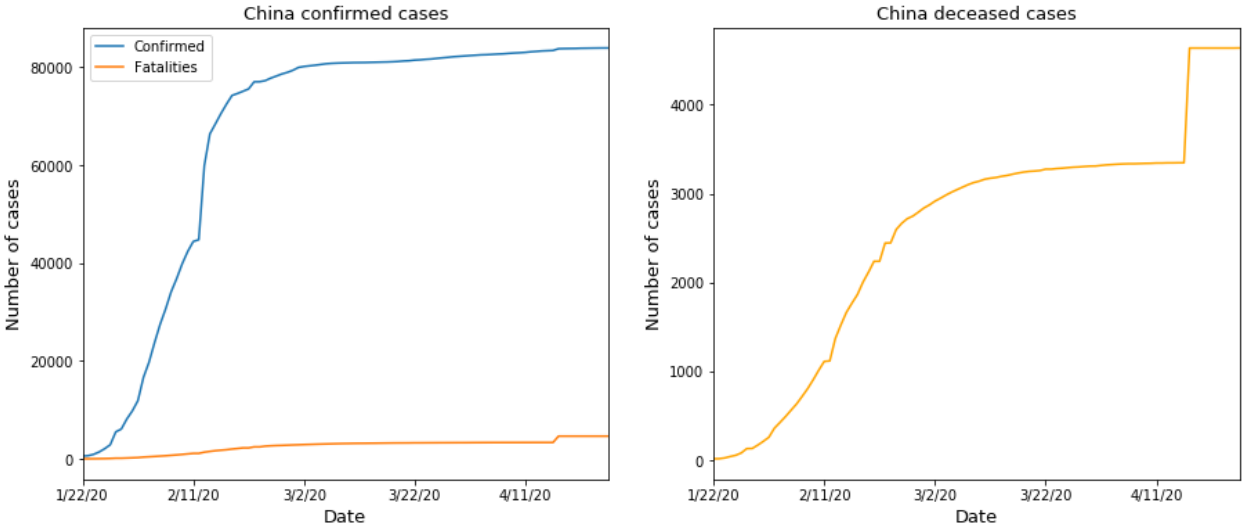

Figure 3: Confirmed and deceased cases in China

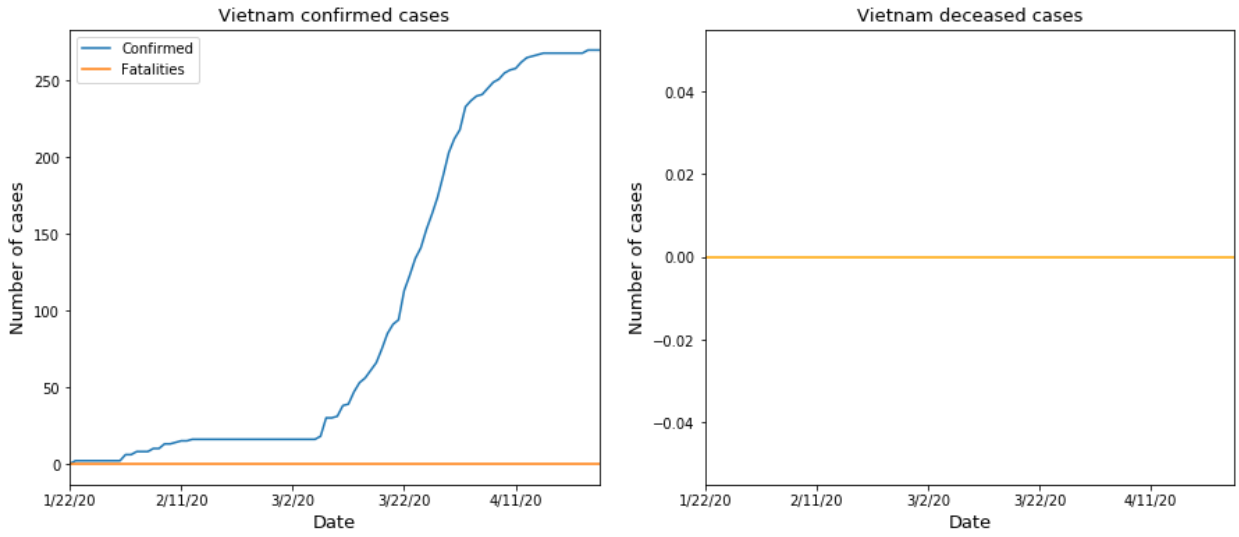

Figure 4: Confirmed and deceased cases in Viet Nam

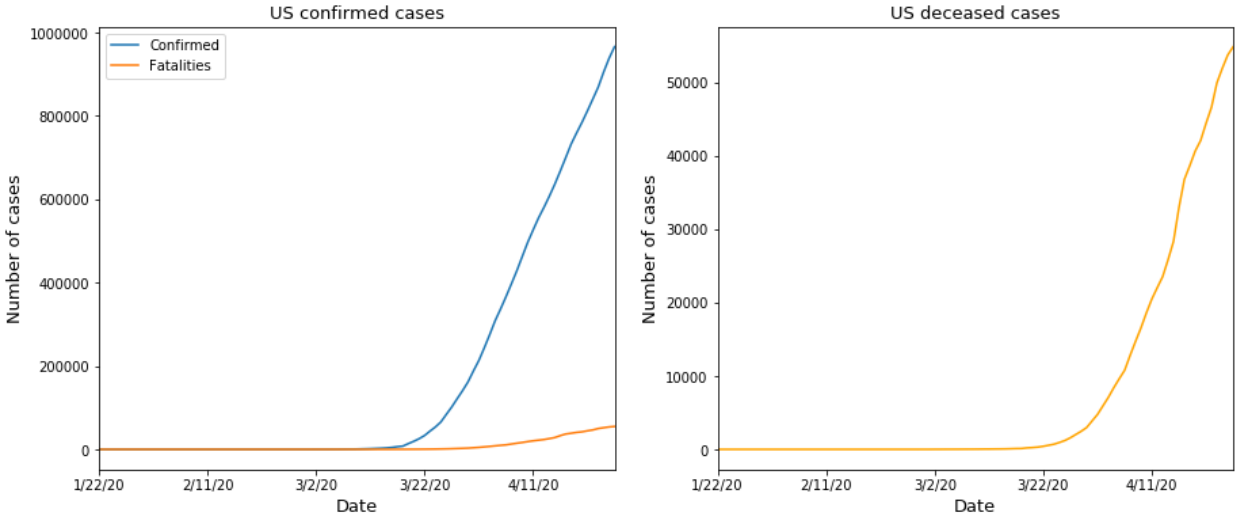

Figure 5: Confirmed and deceased cases in US

Looking at the above plots, we can see the growing number of confirmed cases, globally and in each country. From there we can also spot some abnormal points. For example, in China, there is a straight up rise after April 11th in the deceased cases by about nearly 2000. That was when they released a new number due to wrong statistics method. Confirmed and deceased cases in "Global" and US are rather "smooth", but have shown no sign of slowing down. We now use some models to forecast the progress and see when the curve is flatten.

3. Methodology

The method I use for modelling is inspired from this Kaggle notebook:

https://www.kaggle.com/saga21/covid-global-forecast-sir-model-ml-regressions

I will use 3 models to forecast the number of cases

3.1. SIR model

SIR model is widely used to analyze the progress of transmitted diseases, including COVID-19. There are many versions of this model, some of which pay more attention to demography and external factors like lock down policy. I can only implement a simple model that monitor a population in 3 states:

- **Susceptible (S)**: The individual has not contracted the disease, but might be infected due to transmission from infected people. Initially, all of the population will be in S, and for each time step, some will be infected and is counted to I
- **Infected (I)**: The person has contracted the disease. The group initializes with 0, and is increased by the number of infected people and decreased by recovered people
- **Recovered/Deceased (R)**: The person who had been infected but has either recovered or deceased.

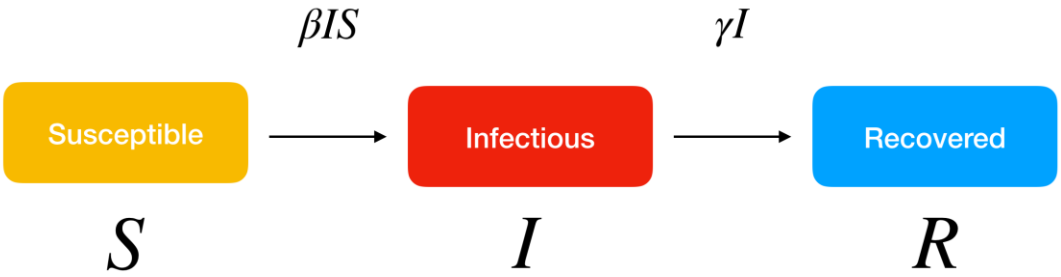

Figure 6: SIR model. Source: lewuathe.com

This model can be expressed by these differential equations:

$$\frac{dS}{dT} = -\frac{\beta IS}{N}$$

$$\frac{dI}{dT} = \frac{\beta IS}{N} - \gamma I$$

$$\frac{dR}{dT} = -\gamma I$$

With β the transmission rate and γ the recovery rate and N is the total size of the population.

SIR model can be implemented in many ways. I will simply run a numerical method Runge-Kutta to solve the differential equations system and fit the real data with the theory. Based on available research in the field, I choose β = 0.78735 and γ = 0.154. Global population is 7.8 billion, we have SIR model from start to end below

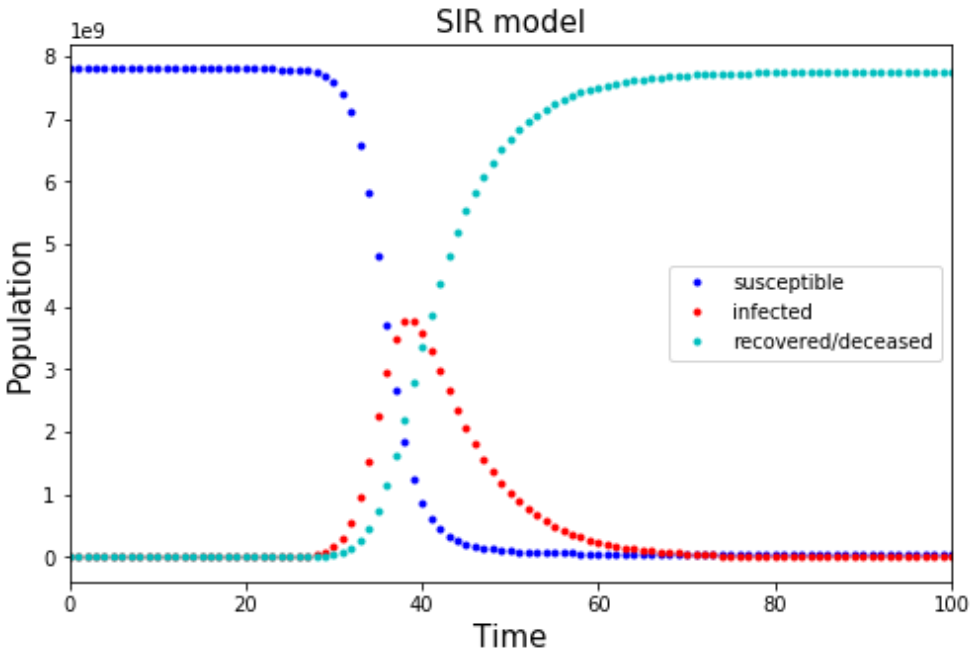

Figure 7: SIR model applied for world population

3.2. Log – Linear Regression

One thing we can observe when doing the EDA is that the transmission rate of the disease is exponential at the early stage (for about 10 - 15 days starting from the outbreak). I will try plotting

the number of cases for that period and try applying the log function to see if we can use log-linear model to observe the spreading of this virus at the early stage.

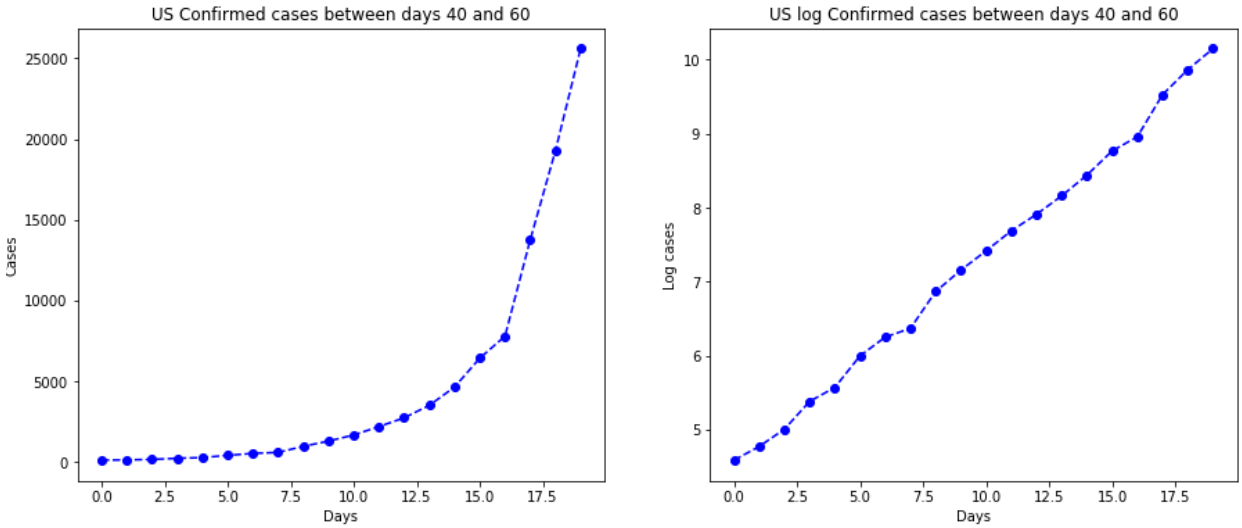

Figure 8: Confirmed case in US, raw data and apply log between days 40 and 60

We can see that from day 40 to 60, the virus started spreading out very fast in exponential rate and we can model the number with a log-linear function. This only happens at the early stage of the outbreak as the population had the first cases. Before or after that, the assumption no longer holds.

3.3. Logistics function

In reality, the number of cases should look similar to a logistic function, with lower rate at the beginning and late stages of the outbreak, while higher at the middle when the disease starts to transmit.

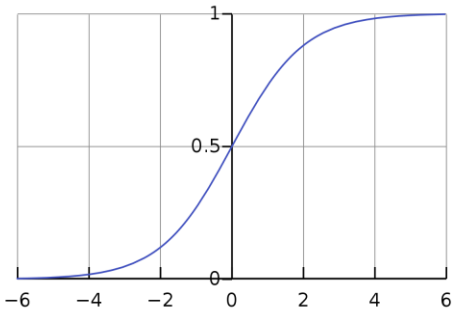

Figure 9: Logistic function. Source: [Wikipedia]

Logistic function can be expressed in the below equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

4. Result

4.1. SIR model result

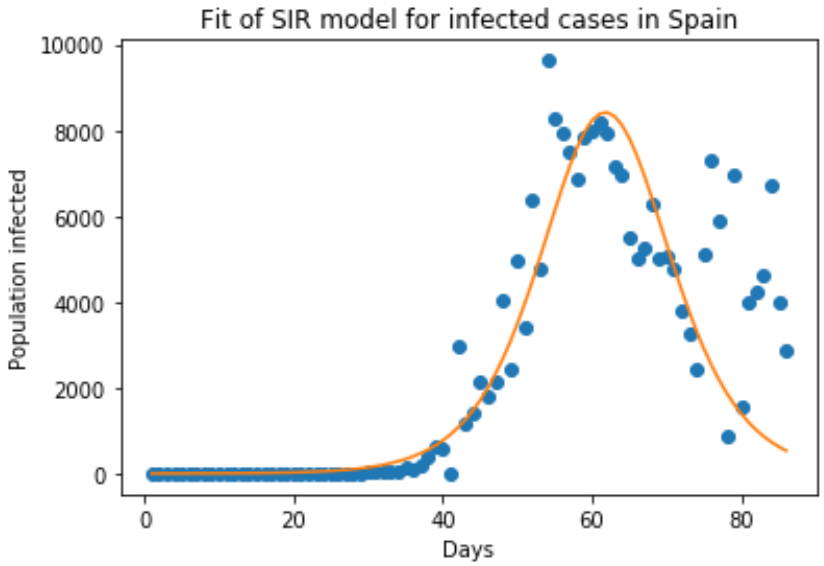

Figure 10: SIR model applied for Spain

We can see that the SIR model is not fitted very well with the real data. The formula that I used in this project is the most basic representation of SIR model. There are many other versions which also consider other factors like demographic, or the intermediate states, which I cannot implement accurately and effectively.

4.2. Log – Linear function forecast

```
Root mean squared error for confirmed case:  0.11262179191614262
Root mean squared error for fatal case:  0.2012122020962682
```

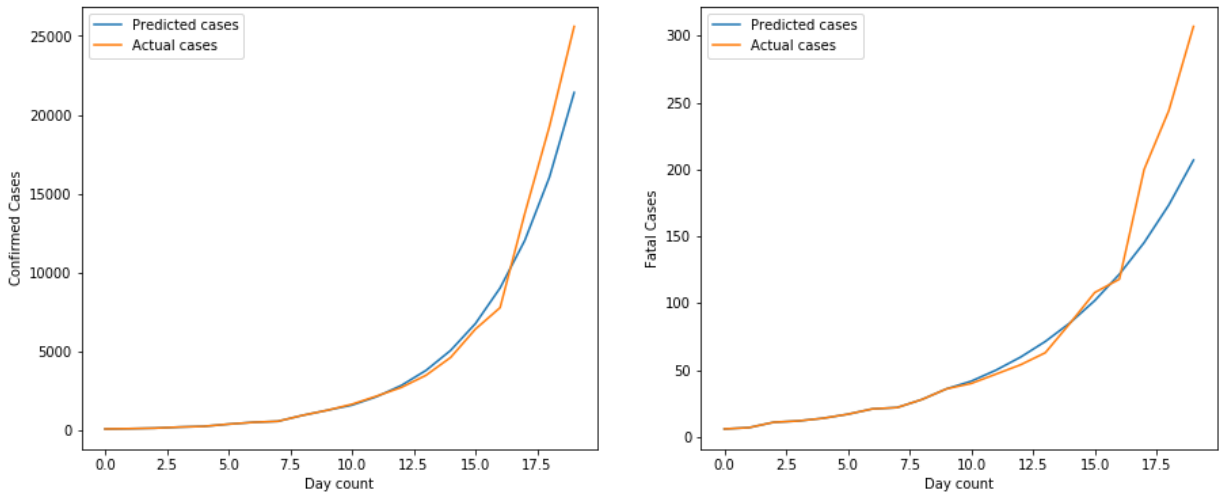

Figure 11: Log-linear function applied for US. RMSE: 0.1126

```
Root mean squared error for confirmed case:  0.2575176595976714
Root mean squared error for fatal case:  0.5264137269163458
```

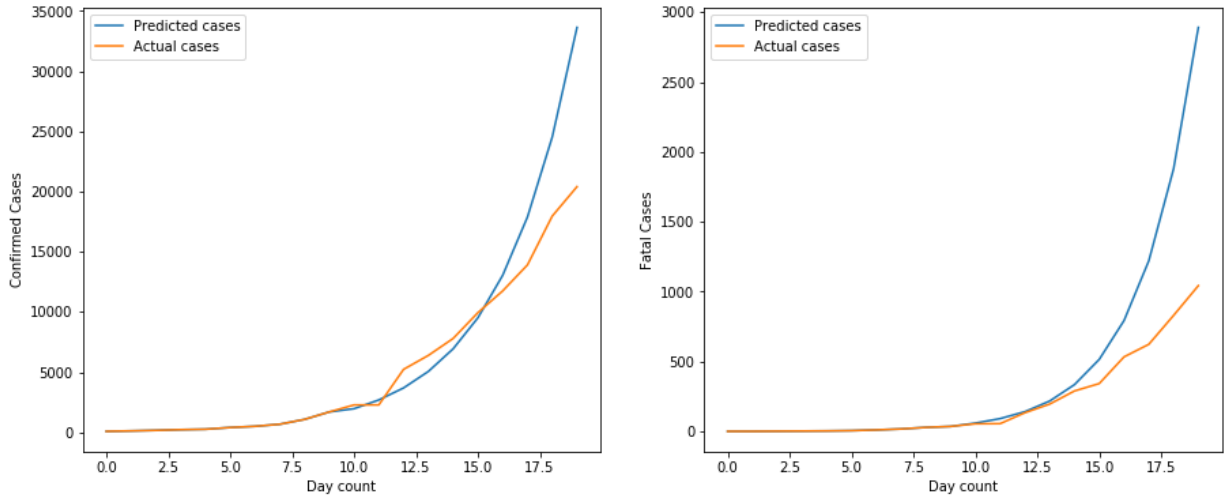

Figure 12: Log-linear function applied for Spain. RMSE: 0.2575

It can be seen that the log-linear model fit well only if we can specify a right data window that lie at exactly the early stage of the outbreak as for that period, the number of cases grow in an exponential form. After that period, with strict policy like social distancing and lock down, the

transmission rate will be lower, then the log-linear model is no longer accurate. We can conclude that this model is not very effective in forecasting the pandemic progress.

4.3. Logistic function forecast

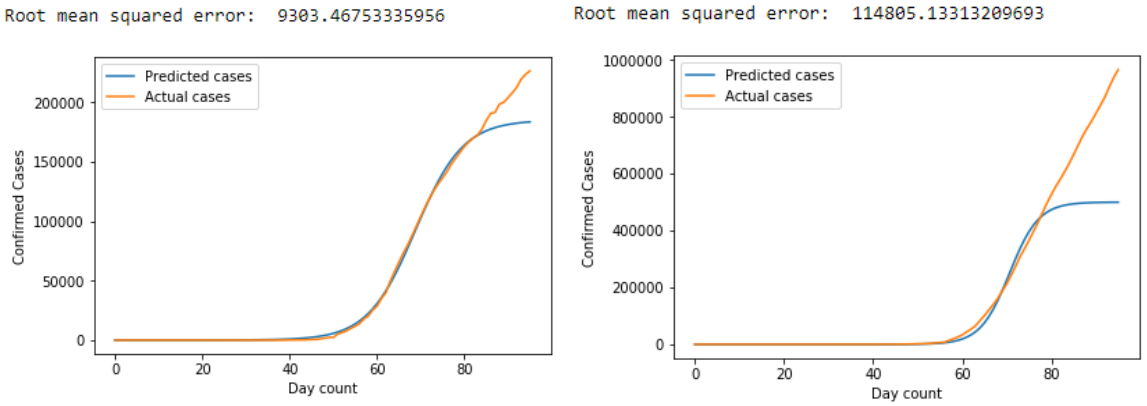

Figure 13: Logistic function applied for Spain and US

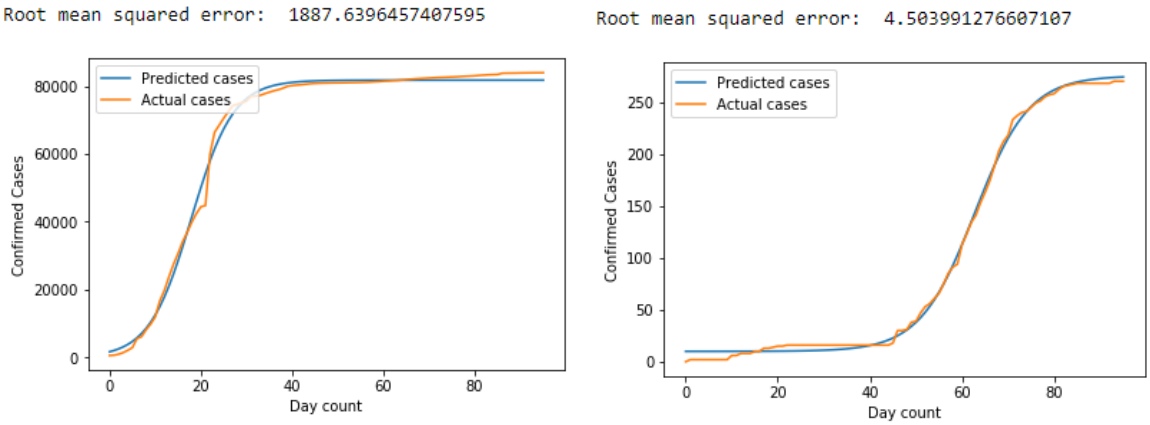

Figure 14: Logistic function applied for China and Vietnam

It looks like the logistic function fits quite well for countries like China and Viet Nam, where the virus is said to have been suppressed and the number of confirmed cases stop raising. For countries like Spain or the US, the confirmed cases data is still increasing so the logistic function is not fitting well.

5. Conclusion

In this project, I have used 3 different method to monitor and forecast the number of confirmed and deceased cases of COVID-19. However, those models are the most simple and need manual effort to fine tune and select which part of the data to apply and cannot fit the data well enough in general

5.1. Improvement

In order to improve the performance and flexibility of the forecasting model, we can try other methods, such as ARIMA (Autoregressive integrated moving average) or LSTM for time series prediction. We can also take into account other information: regions, mass social interaction, and lock down period in order to achieve more accurate result.

May 1st, 2020

Nguyen Manh Tuan

tuan.ngmanh@gmail.com