# COVID-19 Data Analysis

Rishav Bhagat

April 25, 2020

# Contents

# 1 Dataset

I used the dataset that was collected and used by John Hopkins University for this project. This dataset was collected from many international health organizations and all compiled into onto github repository. It contains information on daily reports (including new cases, deaths, and recoveries), and global trends.

Here is the link to the datasets I used:

https://github.com/CSSEGISandData/COVID-19

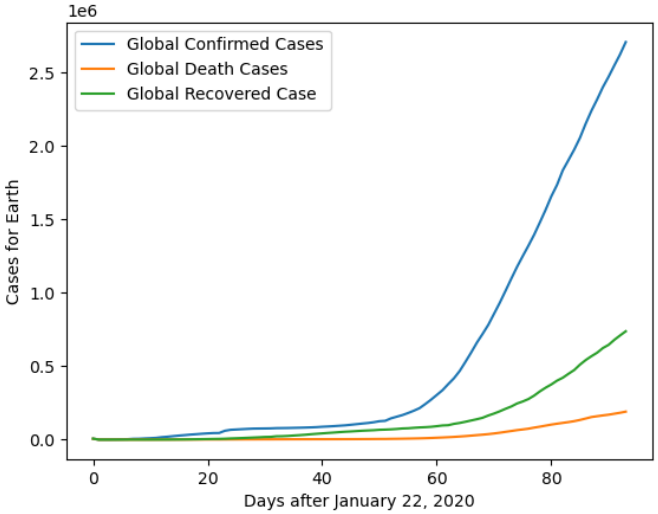
## 1.1 Downloading the Data

I created a batch script to download all this data using git, delete files that are not needed for my analysis, and move the files into more convienient locations.

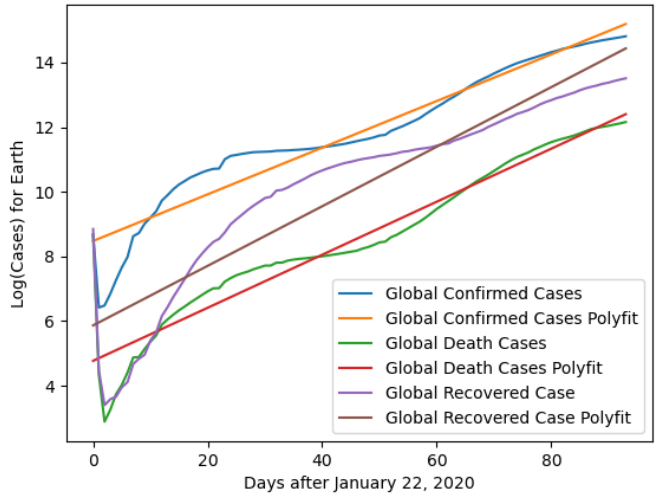
```
@echo off
rmdir data /s /q
git clone https://github.com/CSSEGISandData/COVID-19.git
rename COVID-19 data
cd data
rmdir .git /s /q
rmdir archived_data /s /q
rmdir who_covid_19_situation_reports /s /q
del README.md
mv csse_covid_19_data/* .
rmdir csse_covid_19_data /s /q
cd ..
git add data
git commit -m "updated data from John Hopkins github repo"
git push
```

## 1.2 Initial Glance

First, I plotted the global coronavirus data (cases vs days) in a similar fashion to the way it was plotting on the John Hopkins project. This is to reveal any obvious trends.




At a first glance, this looks like exponential growth, so I then plotting a logarithmic graph with a line of best-fit.




A linear relationship in the logarithmic graph corresponds to an exponential relationship in the original graph. As we can see, the logarithmic graph is approximately linear, which confirms that so far, the coronavirus is exponentially growing. But we also see that the curve is beginning to flatten out, corresponding to the true logistic growth of the original graph. The beginning of all logarithmic graphs appear to be exponential, which explains why it apears as it

does.

Now, we know that logistic growth will eventually level out, and it is in our interest to see when and where it will do this globally. We can do this by tracking the growth rate of the graph, as we shall see in the next section.

# 2 Data Analysis and Implications

First let's define some variables:

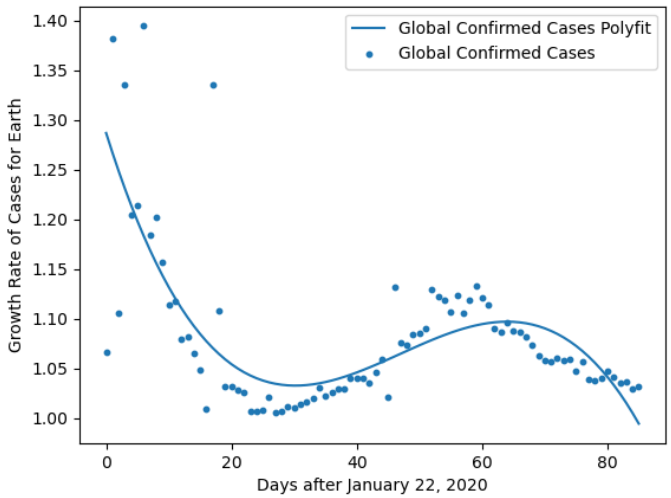
- Let us refer to the region we are talking about as $X$. For now $X$ refers to the entire globe.
- Let $N^X$ be the current number of people who have COVID-19 within $X$. I may sometimes just use $N$, which will refer to the region in question at the time.
- Let $N_i^X$ be the current number of people who have COVID-19 within $X$ on the $i$-th day after January 22nd (which is when John Hopkins began collecting data).
- Let $\Delta N_i^X = N_i^X - N_{i-1}^X$ be the number of new cases on a given day, defined by $i$.

## 2.1 Growth Rate

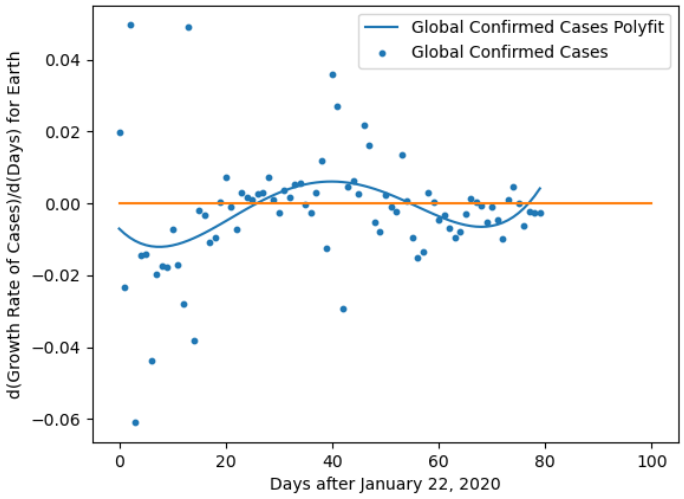
Now we can define our growth rate as

$$G_i^X = \frac{\Delta N_i^X}{\Delta N_{i-1}^X} \tag{1}$$

We are interested when this value crosses from $G > 1$ to $G < 1$, which corresponds to a point of infection. It is when the numbers of new cases each day begins to decrease rather than increase. Or in terms of calculus, when the second derivative becomes negative. The reason we do not just use the second derivative is that this data is noisy and approximating a high-order derivative would not work too well. Now plotting the growth rate:




If the graph continues to follow the polyfit curve drawn, then we expect the inflection point to occur soon, but we do not know if it will go back up as it did at around $i = 25$. In fact, we can expect it to go up, which is evident by plotting the approximated derivative [1] of the growth function.




[1] When I refer to the derivative I am using the central difference approximation give by:

$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

with $h = 1$ since we have a discrete input space (each day).

The derivative of the growth rate at the current time ($i \approx 80$) is following the same pattern as at $i = 25$ and it has gone above the zero-line, implying that the growth rate will increase again.

This is a dissapointing result since finding the true point of inflection will let us estimate where the number of cases will max out. If the point of inflection occurs at $i$, then it is reasonable to guess that the number of cases will max at $2 * N_i^X$ by the way logistic growth works.

### 2.1.1 Modelling with the Growth Rate

The curve generated by `np.polyfit` fits the growth rate pretty well, so it may be possible to create a model based on the growth rate. To do so, we just have to follow the definitions of the growth rate and the solve for $N_i$:

$$G_i^X = \frac{\Delta N_i^X}{\Delta N_{i-1}^X} = \frac{N_i^X - N_{i-1}^X}{N_{i-1}^X - N_{i-2}^X}$$
$$G_i^X (N_{i-1}^X - N_{i-2}^X) = N_i^X - N_{i-1}^X$$

and finally

$$N_i^X = N_{i-1}^X + G_i^X (N_{i-1}^X - N_{i-2}^X) \tag{2}$$

But there are a lot of problems with a model like this. For one it is based on previous estimates, so any error in early estimates will propagate through to the later estimates, making the model extremely inaccurate for anything too far out of the domain of the data.

Also, as we showed before, the growth rate will likely not follow the polyfit curve outside the domain of the data.