COVID-19:
A Mathematical Approach

Rishikesh Sarma

March 18, 2020

0.1 Acknowledgement

This report was made possible due to data taken from **worldometer.com** and **Wikipedia**. All computations were done in **MATLAB R2019b**. All formatting were done in **LaTeX**.
[1]

[1] All the data presented in this report is true to March 08, 2020 18:34 GMT.

0.2 Introduction

0.2.1 A history of COVID-19

Coronaviruses are a group of viruses that cause diseases in mammals and birds. In humans, coronaviruses cause respiratory tract infections that are typically mild, such as some cases of the common cold (among other possible causes, predominantly rhinoviruses), though rarer forms such as SARS, MERS, and SARS-CoV-2 can be lethal. Symptoms vary in other species: in chickens, they cause an upper respiratory tract disease, while in cows and pigs they cause diarrhea. There are yet to be vaccines or antiviral drugs to prevent or treat human coronavirus infections.

In December 2019, a pneumonia outbreak was reported in Wuhan, China. On 31 December 2019, the outbreak was traced to a novel strain of coronavirus, which was given the interim name 2019-nCoV by the World Health Organization (WHO), later renamed SARS-CoV-2 by the International Committee on Taxonomy of Viruses. Some researchers have suggested that the Huanan Seafood Market may not be the original source of viral transmission to humans.

As of 8 March 2020, there have been at least 3,646 confirmed deaths and more than 107,351 confirmed cases in the coronavirus pneumonia outbreak. The Wuhan strain has been identified as a new strain of Betacoronavirus from group 2B with approximately 70% genetic similarity to the SARS-CoV. The virus has a 96% similarity to a bat coronavirus, so it is widely suspected to originate from bats as well.

0.2.2 The aim of this report

This is a basic report of the outbreak of the COVID-19 aka coronavirus that originated in Wuhan, China. The aim of this report is to provide a brief analysis of the coronavirus outbreak so far and what we can expect to encounter in the coming days, using mathematical tools such as linear regression, logistics curves, etc.

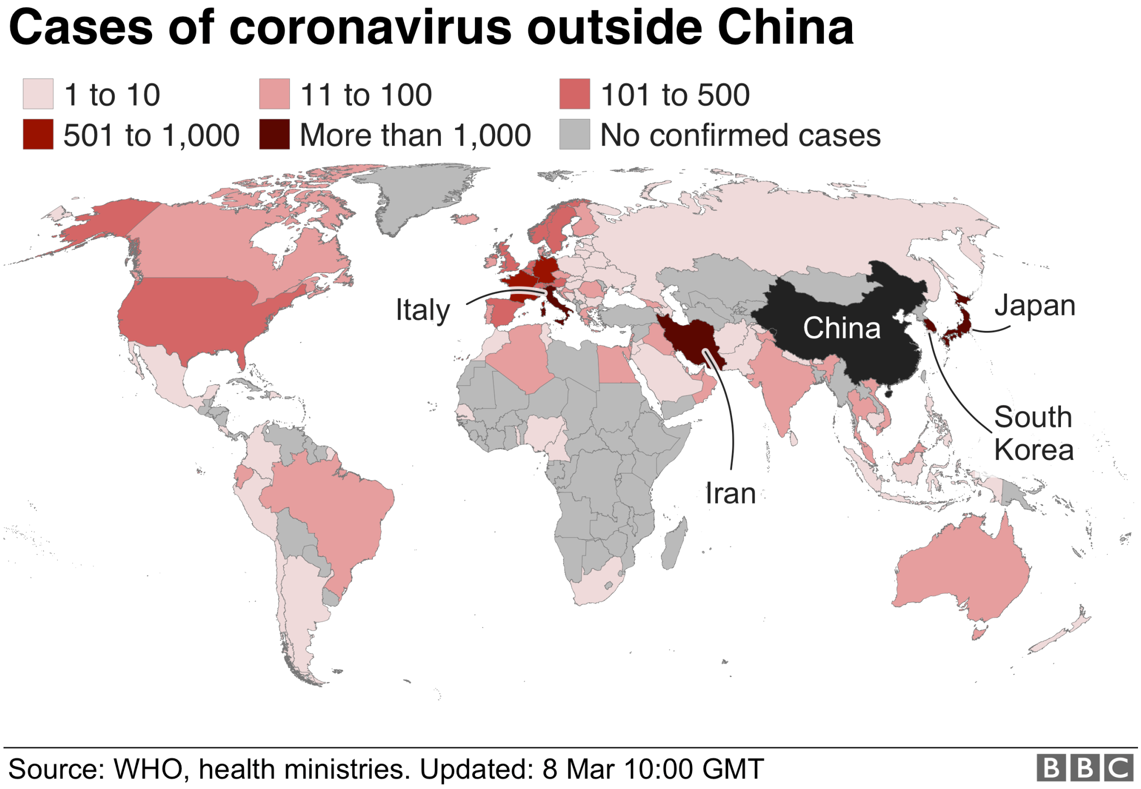
Contents

Chapter 1

the reach so far

1.1 Overview

There are currently 109,680 confirmed cases and 3,802 deaths from the coronavirus COVID-19 outbreak as of March 08, 2020, 18:34 GMT. The coronavirus COVID-19 is affecting 104 countries and territories around the world and 1 international conveyance (the Diamond Princess cruise ship harbored in Yokohama, Japan).



Cases of coronavirus outside China

Source: WHO, health ministries. Updated: 8 Mar 10:00 GMT

1.2 Worldwide

There are 44,933 active cases, of which 38,808 (86%) are in Mild condition and 6,125 (14%) are in Serious or Critical condition. The graph below shows the total number of recorded cases with each passing day. The spike observed on Feb. 12 is the result, for the most part, of a change in diagnosis classification for which 13,332 clinically (rather than laboratory) confirmed cases were all reported as new cases on Feb. 12, even though they were diagnosed in the preceding days and weeks.

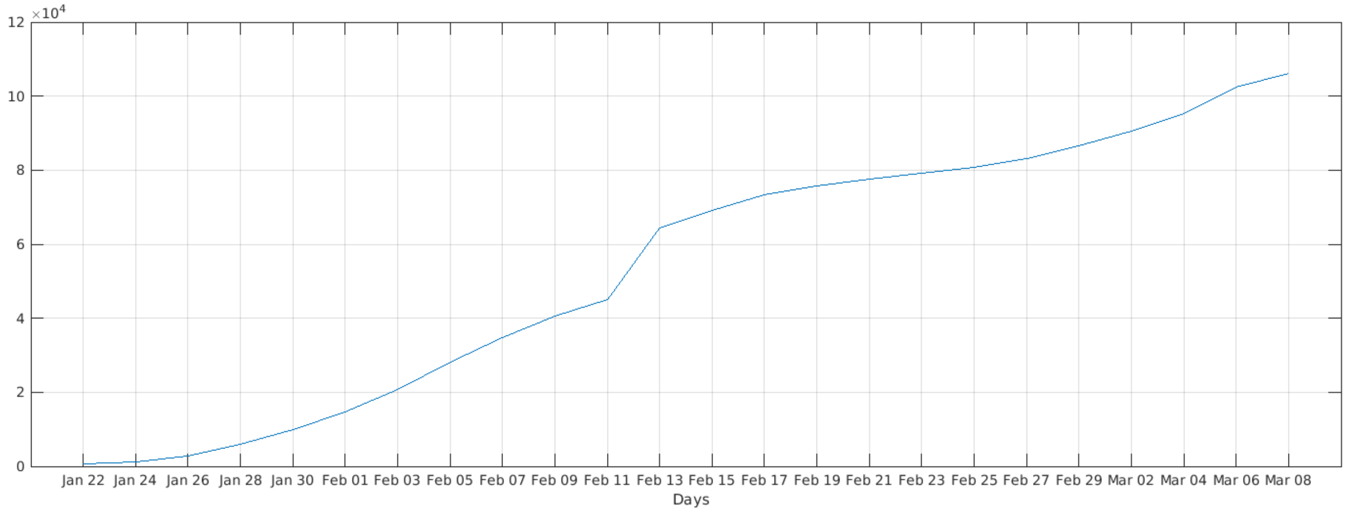


Figure 1.1: Total number of people infected with each passing day

1.3 China vs World

The pie-chart (Figure 1.1) shows the distribution of the number of recorded COVID-19 cases in China and the rest of the world.

1.4 India vs World

There have been a total of 45 cases of COVID-19 in India, out of which 41 are active and 4 have recovered. Globally, India constitutes for 0.04% of the total registered cases and 0.09% of the active cases (Figure 1.2).

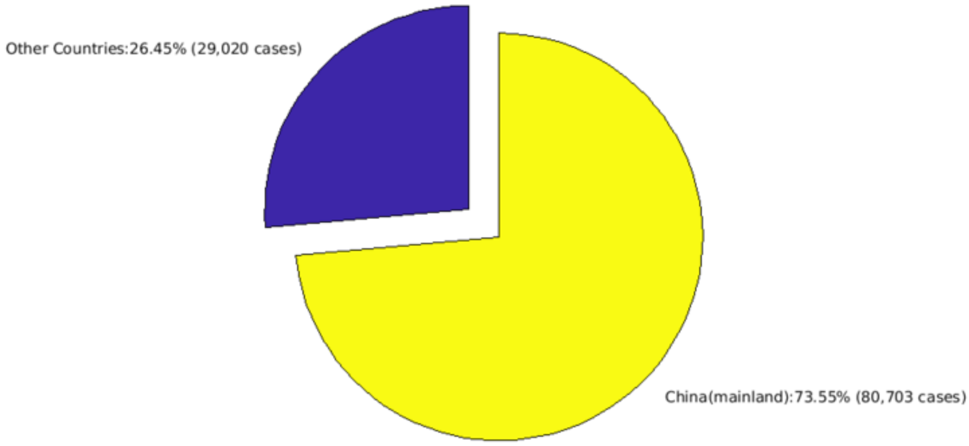


Figure 1.2: Pie comparing the outbreak in China and the rest of the world

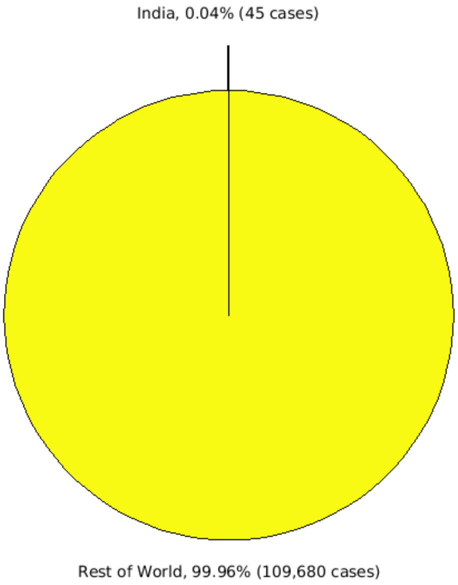


Figure 1.3: A minuscule fraction

Chapter 2

the math (so far)

2.1 The math behind the behind the outbreak

Though viral outbreaks are quite hazardous, the math behind them is fairly simple. This is because all viral outbreaks have a common anomaly: all of them spread through individuals who are already infected. This means that the total number of cases on any given day is proportional to the number of active cases on the previous day. It also means the the number of new cases of any given day is also proportional to the same quantity.

Let the number of active cases of on any given day be N,
Average number of people someone infected is exposed to be E
Probability of someone exposed to COVID-19 being infected be p
Then,
The number of ***new cases*** , ΔN the next day is given by,

$$\Delta N = N \times E \times p \tag{2.1}$$

Of course, while deriving this equation, we have made some assumptions. First, we have assumed that new cases only emerge after one day, but that may not be necessarily the case. Next, we have also not taken into account the people who have acquired the virus but get no longer spread it, e.g. they have recovered (or died). Although keep in mind that people who have gotten infected would have acquired the virus recently. Also let's not get into the more complex factors involving the reach of someone infected with the virus, travel patterns, etc.

Thus, the total number of cases,N_T on any particular day can be given by,

$$N_T = N + \Delta N \tag{2.2}$$

$$\implies N_T = N + N \times E \times p$$

$$\implies N_T = N \times (1 + E \times p)$$

Hence, it is clear from the above equation (and the graph below) that the nature of the outbreak is exponential, as all viral outbreaks tend to be initially. By studying the graph below, it was found that the factor $(1 + E \times p)$ lies roughly between $1.15 - 1.25$.

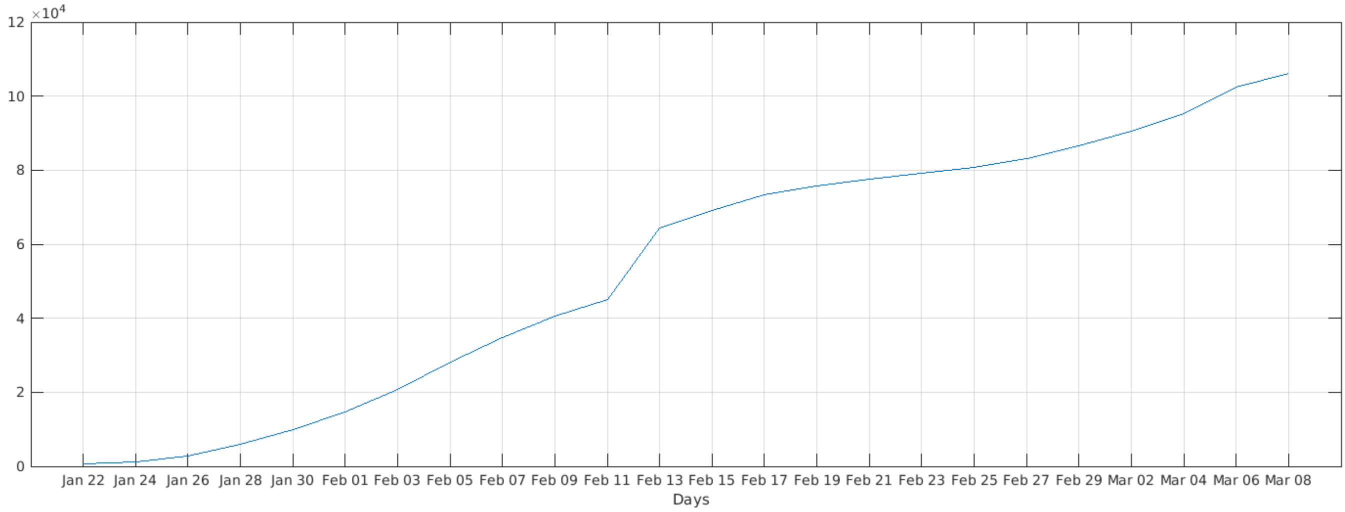


Figure 2.1: No. of people infected vs Time

2.2 Ananlysis through Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable), in this case the successive days, and one or more explanatory variables (or independent variables), which in this case is the number of recorded cases. The case of one explanatory variable is called simple linear regression. To put it simply, Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set we have, with the belief that those outputs would fall on the line.

As we have already seen, the graph between the total number of recorded cases with each passing day, as a plot of scattered points can be shown as below.

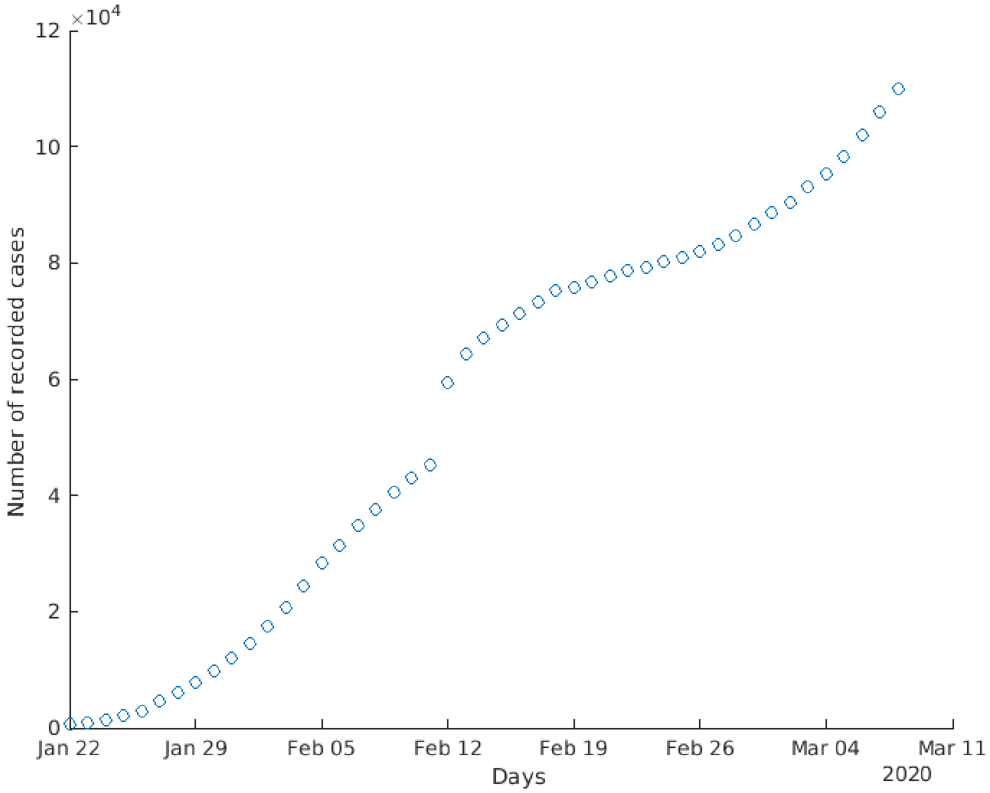


Figure 2.2: No. of total recorded cases of COVID-19 as scattered points

Using linear regression on the above data, we find the line that best fits the data points. We obtain the line shown below.

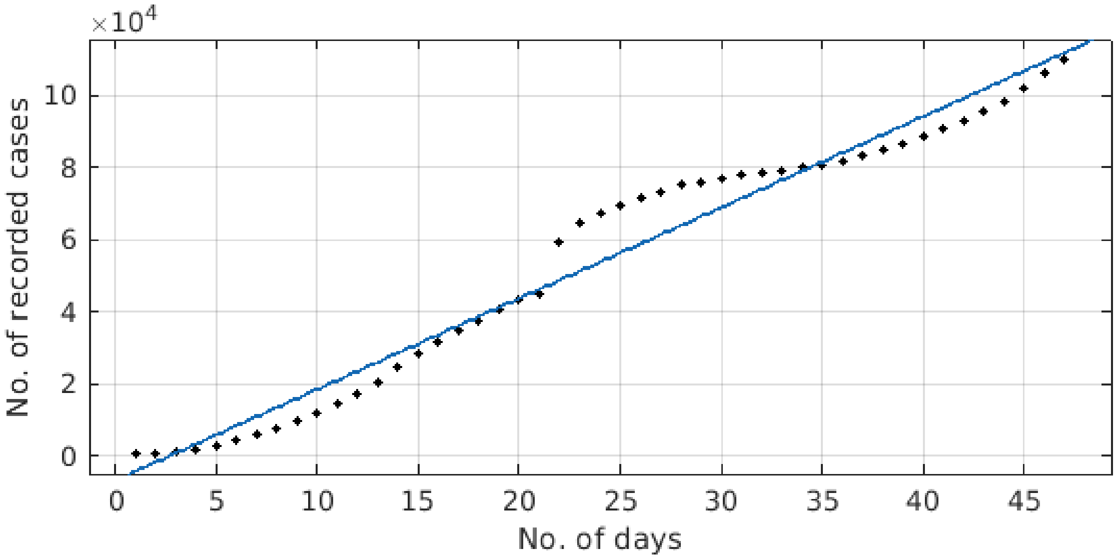


Figure 2.3: The best fit line

In this case, we find $\boldsymbol{R^2}$ to be **0.9668**.

R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. In simple words, it is a measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. If for e.g., the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

In this case, with $R^2 = 0.9668$, we can conclude that the regression line fits our data points nearly perfectly. **By studying the slope of the regression line, we found that the number of recorded cases tends to increase by a factor of ten fold every 16 days, on average**.

Extrapolating the regression line, we can make predictions on the number of new cases. We found that the number of recorded cases would hit 1 million in 30 days (April 5, 2020), 10 million in 47 days (April 22, 2020), 100 million in 64 days (May 9, 2020) and 1 billion in 81 days (May 26, 2020). However, there is a grave mistake in this approach.

This approach would suggest that the number of infected people would keep increasing till the entire human population is infected. This represents a highly idealised and unrealistic case of a viral outbreak. Even the most deadliest outbreak in history, the Spanish flu, managed to infect about one-third of the world population and claimed about 50 million lives (2.5% of the world population).

Although it may be seem like viral outbreaks follow an exponential function, they do so only initially. In reality, all viral outbreaks follow a logistic function or a logistic curve. A logistic curve is a common 'S' shaped curve that are used to model functions that increase gradually at first, more rapidly in the middle growth period, and slowly at the end, leveling off at a maximum value after some period of time. Thus, logistic curves are the perfect tool to depict the nature of any viral outbreak as all of them follows the above mentioned trend. The maximum value of the logistic curve is the positive saturation point, and in this case, it denotes the total number of people COVID-19 has left infected after the neutralisation of the outbreak.

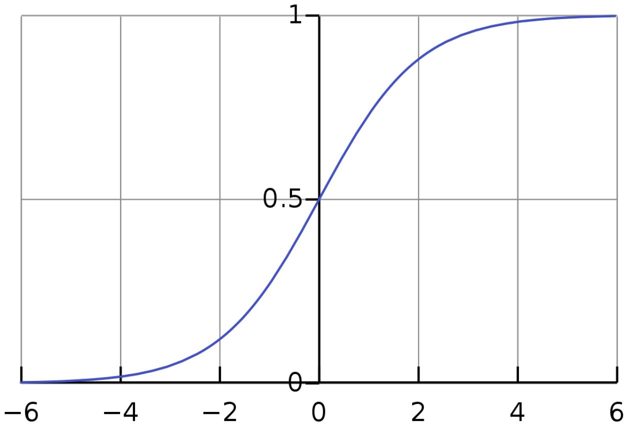


Figure 2.4: A simple logistic curve, where $L = 1$

2.3 Why the logistic approach makes sense

All viral outbreaks tend to follow a exponential function only in the initial stages of the outbreak. In actuality, they follow a logistic function (which we will see later) and as with all things, they come to an end one day. This is because of very basic fact that all viral outbreaks are dependent on the number of existing active cases to spread new cases. In the worst case scenario of a perfectly pernicious model of a viral outbreak, the outbreak infects almost the entire population. But even then, since the entire population is already infected, the virus does not get new uninfected hosts to spread its' disease. We have already seen the equation (2.2) relating the total number of recorded cases to the number of new cases.

$$N_T = N \times (1 + E \times p) \tag{2.3}$$

From the above equation, we can see that there is a valid reason to expect an exponential here. If the number of recorded cases on any given day is proportional to the number of existing cases, it necessarily means each day we multiply some constant, which in our case turns out to be $(1 + E \times p)$ and lies between 1.15 to 1.25. So moving forward d days is the same as multiplying the constant d times, i.e.

$$N_d = N \times (1 + E \times p)^d \tag{2.4}$$

The only way N_d goes down is if the factor $(1 + E \times p)$ decreases, or strictly speaking the factor $(E \times p)$ decreases. It is inevitable that this will eventually happen. It may happen if the standards of quarantine efforts are improved, people stop gathering and travelling, or as simple as if more people start washing their hands.

In our equation, it means including a factor like $(1 - \frac{N_d}{Pop.size})$ to account for the people who are already infected or can no longer get the infection due to some reasons (for e.g. they have recovered). Hence, the modified equation becomes,

$$\Delta N = N \times E \times p \times (1 - \frac{N}{Pop.size}) \tag{2.5}$$

and

$$N_d = N + \Delta N \tag{2.6}$$

$$\implies N_d = N + N \times E \times p \times (1 - \frac{N}{Pop.size})$$

$$\implies N_d = N \times (1 + E \times p \times (1 - \frac{N}{Pop.size}))$$

Differentiating the above equation with respect to time, i.e. $\frac{dN_d}{dT}$, we obtain a **logistic curve**.

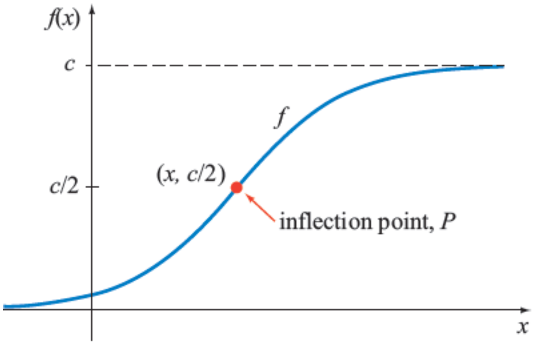


Figure 2.5: A logistic curve and the inflection point

The inflection point of the curve is defined as that point at which the curve goes from concave-up to concave-down (or vice-versa), and the derivative around this point remains roughly constant. Strictly speaking, the inflection point is defined as that point at which the second-order derivative equals zero, i.e. $\frac{d^2N_d}{dT^2} = 0$. The concept of inflection point will come in handy later.

2.4 The Growth Factor

Growth factor of an outbreak is defined as the ratio between the change in the number of recorded cases one day to the change in the previous day, i.e.,

$$Growthfactor = \frac{\Delta N_{d+1}}{\Delta N_d} \tag{2.7}$$

As long as the growth factor is greater than unity, it means that we're in the concave-up part of our logistic curve and the rate at which the recorded cases occur will keep on increasing. If the growth factor is less than unity, it means that we're in the concave-down part of our logistic curve and the rate of infection of the outbreak is slowly decreasing, and hopefully it will reach its' saturation point soon.

But if the growth factor is equal to unity (or round-about unity), it indicates that we have hit the inflection point in our logsitic curve, and the time it will take for the outbreak to reach its' saturation point is roughly double the time at this point. Using this fact, we can calculate the number of days it will require for the outbreak to reach its' saturation level and die out.

We've calculated the growth factor over a period of 45 days starting from January 22, 2020 to March 8, 2020 and the results obtained is plotted below.

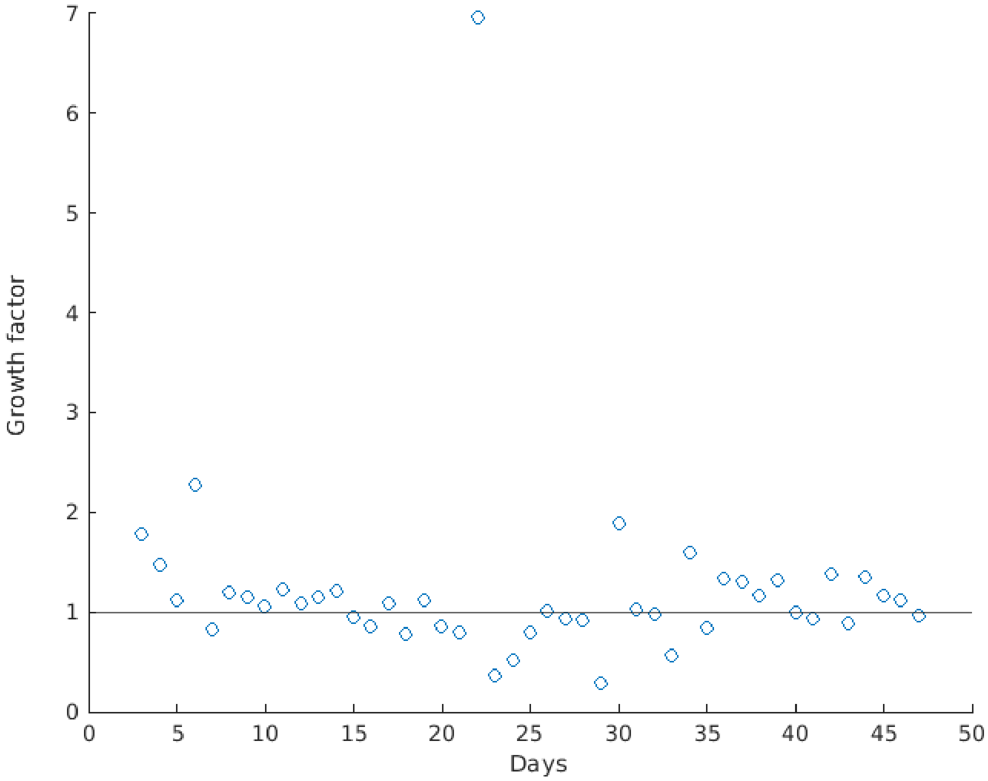


Figure 2.6: The growth factor. The horizontal lines is $y = 1$ and indicates how close the growth factor on a particular day is to 1.

As we can see, there is no clear pattern emerging from the plot. Sometimes, the growth factor is above unity and at other times, it lies below. Occasionally, it hits unity but since there is more than instance of this, we cannot say anything conclusively.

Even when we apply liner regression to the data, we find $R^2 = 0.007474$ which indicates the regression line is not a close fit at all.

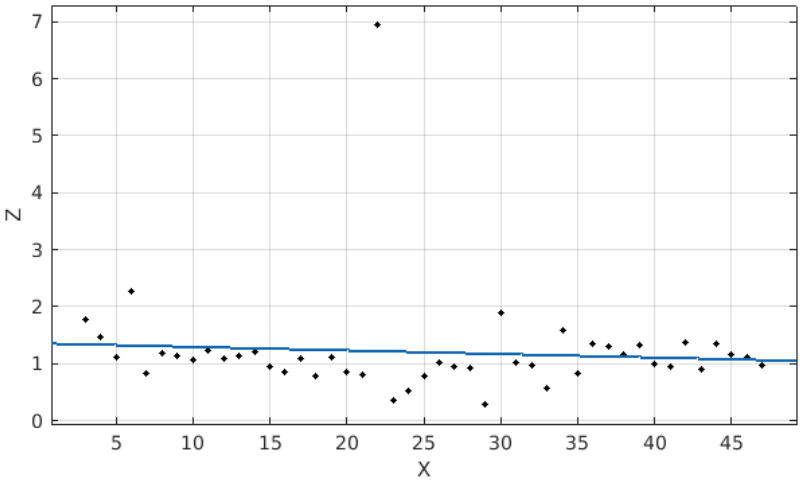


Figure 2.7: The regression line. Here $R^2 = 0.007474$

Chapter 3

conclusion (so far)

There is a need for more data. We cannot conclude when the outbreak will reach its' saturation with the data we've collected so far. But we can instill better practices. With better sanitation, health-care, quarantine and monitoring efforts as well as other factors like research and awareness, we can surely limit the saturation and contain the outbreak.