BMS 331: Numerical Analysis Project Report

COVID-19: Epidemic Spread Prediction Models

Done By:

Mayar Mansour 201700072

Muhammad ElSadany 201700857

Supervised By:

Dr. Abdallah Aboutahoun

Abstract:

In December 2019, a novel coronavirus was found in a seafood wholesale market in Wuhan, China. The World Health Organization (WHO) officially named this coronavirus as COVID-19. Since the first patient was hospitalized on December 12th, 2019, China has reported 78824 confirmed COVID-19 cases and 2788 deaths as of February 28th, 2020. Wuhan's cumulative confirmed cases and deaths accounted for 61.1% and 76.5% of the whole China mainland, making it the priority center for epidemic prevention and control. Meanwhile, 51 countries and regions outside China have reported 79251 confirmed cases and 2835 deaths as of February 28th, 2020. The COVID-19 epidemic does great harm to people's daily life and the country's economic development. Finding the best spread prediction model is a priority now to help countries prepare for it, especially after WHO announced that COVID 19 is a pandemic disease as of March 12th, 2020. We'll mainly get the number of infected cases from different countries through a defined time interval and try to predict the model of disease spreading by 3 different methods for curve fitting: Linear regression, spline interpolation, and interpolation. The countries we will use for their statistical data are China, United States of America and Italy. After getting the three models, we'll compare them and choose the best model of fitting that will help in predicting new cases that will be infected by time. The comparison of these 3 models will be mainly based on comparing the error of each case. We will use the best fitting model of these 3 for predicting the spreading mode and statistical data for Egypt.

Introduction

On 31 December 2019, Chinese specialists cautioned the World Health Organization (WHO) that a flare-up of pneumonia of obscure etiology in Wuhan City, Hubei Province, China. A new strain of coronavirus (2019-nCoV) was in this manner segregated from a patient on 7 January 2020 [1]. Contaminations inside family groups and in medical workers affirm the occurrence of human-to-human transmission, however, the degree of this method of transmission is unclear. On 21 January 2020, the WHO recommended there was a potential continued human-to-human transmission. Side effects of detailed cases incorporate fever, hack, and brevity of breath. Pneumonia, serious intense respiratory disorder and kidney disappointment have been accounted for in extreme cases. Current clinical and epidemiological information is deficient to comprehend the full degree of the transmission capability of the pestilence. As of May, nearly all the countries worldwide have had reported cases of the virus and issued quarantine protocols as a way of limiting the progression of the virus. When the decision of enforcing the quarantine protocol was taken too late as in Italy, the casualties were severe and the country started to prioritize treating young people and children over the elderly.

Infectious disease transmission could be a complicated diffusion process occurring within the crowd. Models can be established for this process to investigate and study the transmission process of infectious diseases theoretically, so we can accurately predict the longer-term development trend of infectious diseases. Therefore, to regulate or reduce the harm of infectious diseases, the research and analysis of infectious disease prediction models became a hot research topic.

In this project, we will depend on the regression approach and curve-fitting of the data to make our prediction model for the spread of COVID-19 depending on the daily updates from the WHO regarding the number of new cases and deaths for around one month in 3 countries: Italy, China, and the USA. We will use least-square regression and interpolation and find the best fitting model to describe the data using the residuals. All models will be built using MATLAB.

Methods

All of the statistics and data for the 3 countries: China, US, and Italy, were retrieved from the official site for WHO [2, 3, 4 respectively]. The excel sheet attached contains the daily reported new cases and death, and the total of cases and deaths in each of the 3 countries. This data was used for building all the fitting modes for each country. To be more accurate, the models built were all for the same time interval which is a one-month interval. The chosen interval differs for each country as we're mainly checking the start time and peaks of disease spreading. For example, in **China**, the chosen 1-month interval was from **Jan 23rd to Feb 22nd**. On the other hand, the chosen interval for the **US** was from **Mar 12th to Apr 11th** and **Italy's** interval started from **Feb 25th to Mar 26th**.

The x values were considered as the day number through the 1-month interval starting from 1 till 31 while the y values were varied for different models from being the total cases to being the total deaths. We applied the main 2 models: least squares regression and interpolation on all the data of the 3 countries we have. We mainly applied 3 models in the least-squares regression: linear regression, quadratic polynomial, and cubic polynomial. On the other hand, we only applied two methods for the interpolation model: Newton and Lagrange. By applying the least-squares model on the 3 countries, we compared all the residuals norms for the 3 countries to get the best fitting model of the 3. The same was applied for the interpolation model; the residuals were compared for

both Newton and Lagrange. After that, the most fitting 2 chosen models are used for data curve fitting of Egypt.

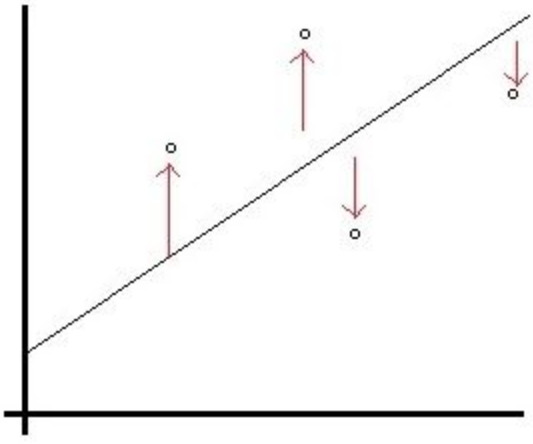
1. Least Squares Regression:

1.1. Linear Regression

It represents the simplest example of a least-squares approximation that works by fitting a straight line to a set of paired observations. The mathematical expression of that line is: $Y = a + bx + e$

Where a is a coefficient representing the intercept and b represents the slope of the line. While e represents the error or residual between the real values of Y and the approximated ones by the model. It can be calculated through the equation of:

$$e = Y - (a+ bx)$$



In this model, all needed is computing the values of both a and b to create a line as close as possible to the given data points. The best strategy for applying this model is to

minimize the sum of the squares of the residuals between the measured y and the y calculated with the linear model.

$$S_r = \sum_{i=1}^{n} e_i^2 = \{E\}^T\{E\} = \|E\|^2$$

$$S_r = (\{Y\} - [X]\{a\})^T(\{Y\} - [X]\{a\})$$

$$S_r = \{Y\}^T\{Y\} - \{Y\}^T[X]\{a\} - \{a\}^T[X]^T\{Y\} + \{a\}^T[X]^T[X]\{a\}$$

$$S_r = \{Y\}^T\{Y\} - 2\{a\}^T[X]^T\{Y\} + \{a\}^T[X]^T[X]\{a\}.$$

So, simply we need to reduce the Sr as much as we can and this can be done by getting the {a} that minimizes it. For that, we differentiate the Sr with respect to the vector {a} and equate it to 0:

$$\frac{dS_r}{d\{a\}} = -2[X]^T\{Y\} + 2[X]^T[X]\{a\} = 0$$

Then {a} can be given by this formula:

$$\{a\} = \left([X]^T[X]\right)^{-1}\left([X]^T\{Y\}\right)$$

This is the general least squares solution. And in our case, the linear regression, the {a} vector will be only of 2 rows for a and b the coefficients.

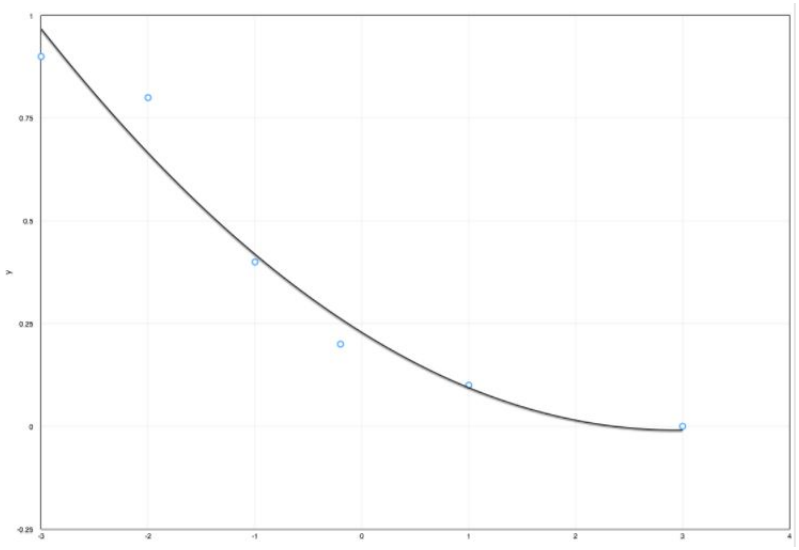
1.2. Quadratic Polynomial Regression

The least-squares procedure can be readily extended to fit the data to a higher-order polynomial. The governing matrix equation is still given by

$$\{Y\} = [X]\{a\} + \{E\}$$

The regression model equation for a quadratic polynomial can be given by:

$$y = a_0 + a_1 x + a_2 x^2 + e$$



And in this case the {a} vector will have 3 rows representing the 3 coefficients in the model equation. And in this case, the residual can be calculated using this form:

$$\begin{Bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{Bmatrix} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix} - \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \end{Bmatrix}$$

1.3. Cubic Polynomial Regression

The same case of quadratic polynomial regression is applied here with a difference in the polynomial degree. So that the regression model can be given by this form:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + e$$

Therefore, the residual errors can be calculated by this form:

$$\begin{Bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{Bmatrix} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{Bmatrix} - \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{Bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{Bmatrix}$$

2. Interpolation

2.1. Newton's Interpolation

Interpolation is the estimation of a value within two known values in a sequence of values.

Derivation of Newton's Interpolation:

Suppose a set of k+1 data points (x_0, y_0) , (x_1, y_1), (x_2, y_2), . … . .(x_k, y_k).

where, every x_j is unique.

Then the given interpolation of polynomial in Newton's form can be expressed in linear combination of Newton basis polynomial as follows:

Where $n_j(x)$ can be defined as $$N(x) := \sum_{j=0}^{k} a_j n_j(x)$$

$$n_j(x) := \prod_{i=0}^{j-1} (x - x_i)$$

Similarly, the coefficients are defined as:

$a_j := [y_0, y_1, y_2, \ldots\ldots y_j]$ where, $[y_0, y_1, y_2, \ldots\ldots y_j]$ is the notation of divided difference which can be expressed as

$$[y_0] = y_0$$

$$[y_0, y_1] = \frac{y_1 - y_0}{x_1 - x_0}$$

$$[y_0, y_1, y_2] = \frac{[y_1, y_2] - [y_0, y_1]}{x_2 - x_0} = \frac{\frac{y_2 - y_1}{x_2 - x_1} - \frac{y_1 - y_0}{x_1 - x_0}}{x_2 - x_0} = \frac{y_2 - y_1}{(x_2 - x_1)(x_2 - x_0)} - \frac{y_1 - y_0}{(x_1 - x_0)(x_2 - x_0)}$$

$$[y_0, y_1, y_2, y_3] = \frac{[y_1, y_2, y_3] - [y_0, y_1, y_2]}{x_3 - x_0}$$

Now, Newton's interpolation or polynomial can be expressed as:

$N(x) = [y_0] + [y_0, y_1] (x - x_0) + \ldots + [y_0, \ldots y_k](x - x_0) (x - x_1) \ldots (x - x_{k-1})$

2.2. Lagrange Interpolation

In a set of distinct points and numbers x_j and y_j respectively, this method is the polynomial of the least degree at each x_j by assuming corresponding value at y_j.

Consider a given set of k+1 points, (x0, y0) , (x1, y1), (x2, y2)….. (xk, yk) where each point is distinct.

Let's assume a function $L(x_j)$ such that $L(x_j) = y_j$, j = 0, 1 , 3 , 3 .. k

Observing the following points

- $L_j(x)$ contains k factors in product and each factor has x

$$\ell_j(x_i) = \prod_{m=0,\, m\neq j}^{k} \frac{x_i - x_m}{x_j - x_m}$$

Now, consider what happens when this product is expanded. Since the product skips $m = j$, when $i = j$ then all terms are $[x_j - x_m] / [x_j - x_m] = 1$

Also, when $i \neq j$, $m \neq j$ does not produce it and one term in the product will be for $m = I$, that is, $[x_i - x_i]/[x_j - x_i] = 0$

Zeroing the entire product,

$$\ell_j(x_i) = \delta_{ji} = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{if } j \neq i \end{cases}$$

So, it can be written that:

$$L(x_i) = \sum_{j=0}^{k} y_j \ell_j(x_i) = \sum_{j=0}^{k} y_j \delta_{ji} = y_i$$

Therefore, the considered function $L(x)$ is a polynomial with degree at most k and where $L(x_j) = y_j$

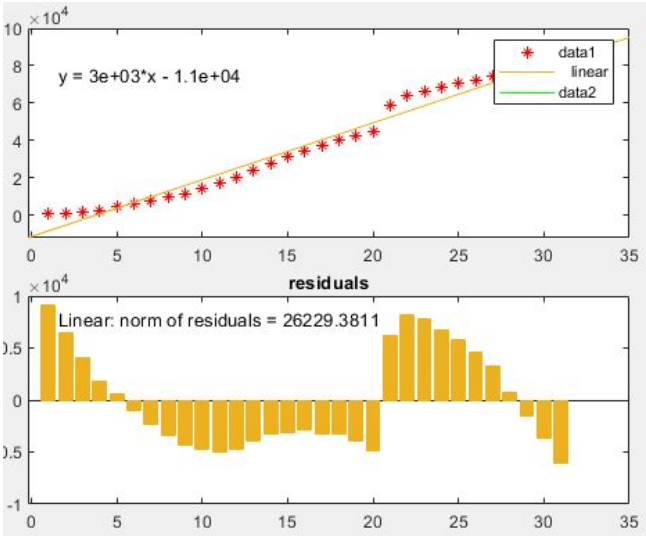
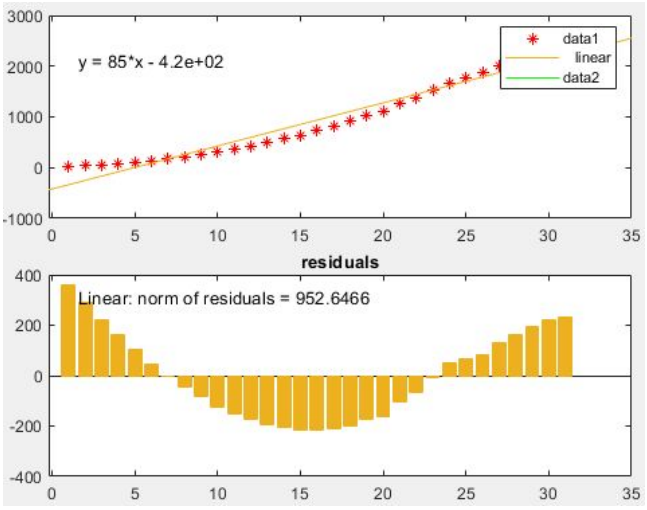
So, for all $i \neq j$, $l_j(x)$ includes the term $(x - x_i)$ in the numerator, therefore the entire product will be found to be zero at $x = x_j$

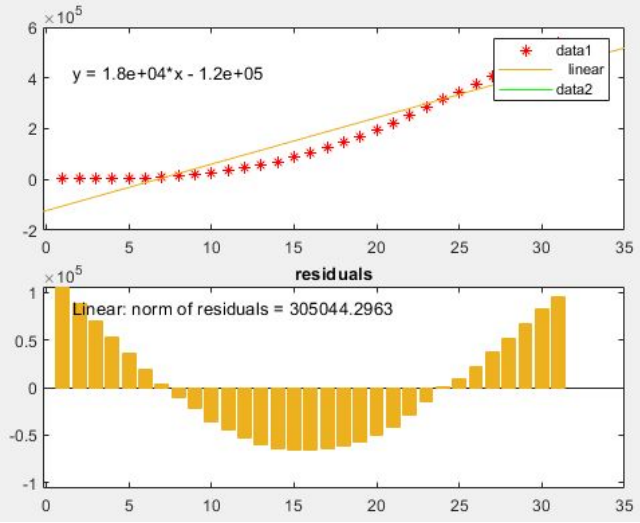
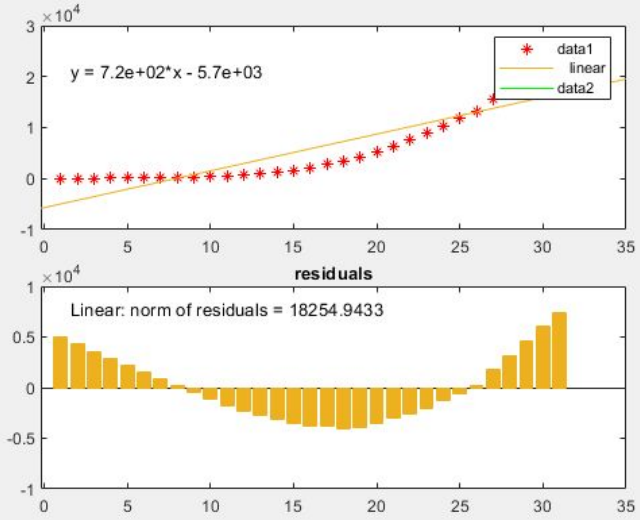
$$\ell_{j\neq i}(x_i) = \prod_{m\neq j} \frac{x_i - x_m}{x_j - x_m} = \frac{(x_i - x_0)}{(x_j - x_0)} \cdots \frac{(x_i - x_i)}{(x_j - x_i)} \cdots \frac{(x_i - x_k)}{(x_j - x_k)} = 0$$

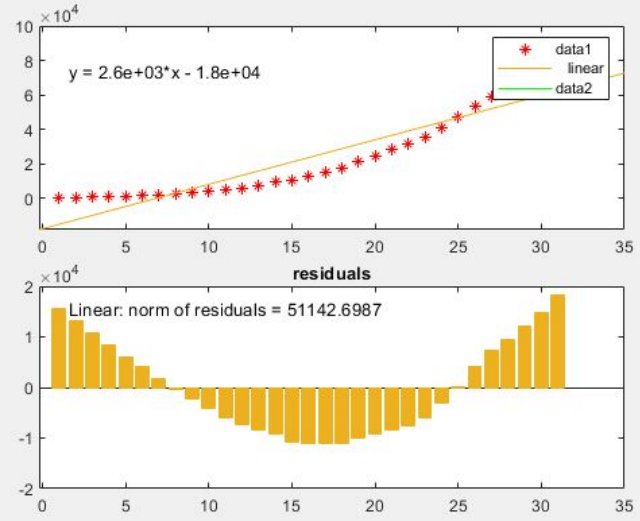
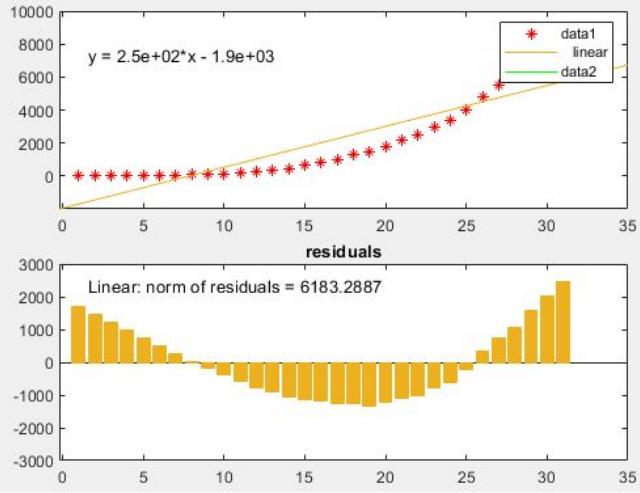
Results

1. Least Squares Regression:

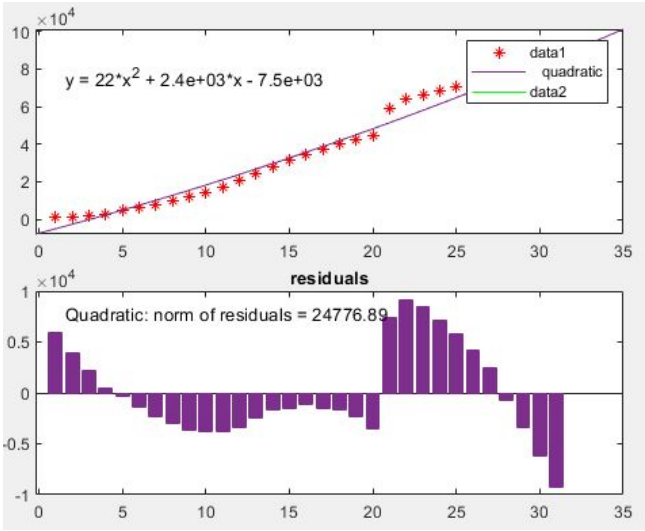
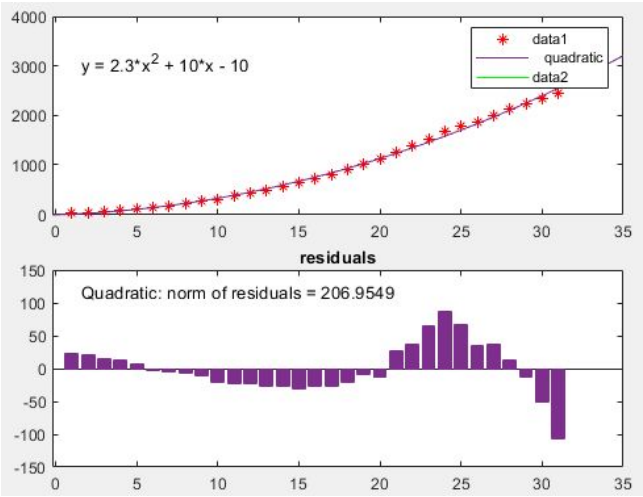
1.1. Linear Regression:

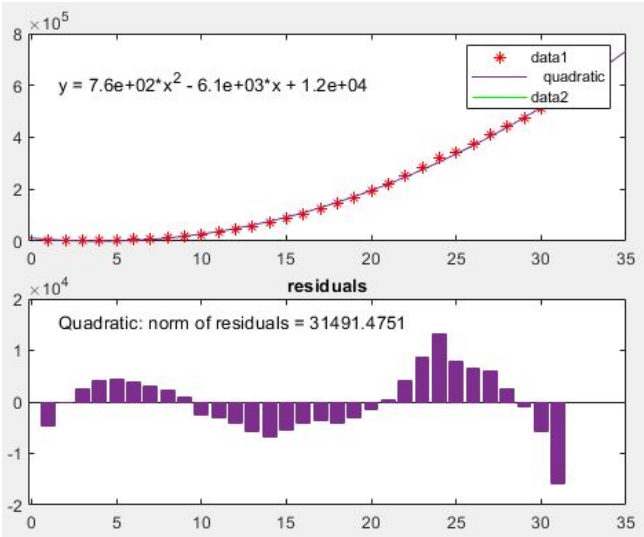
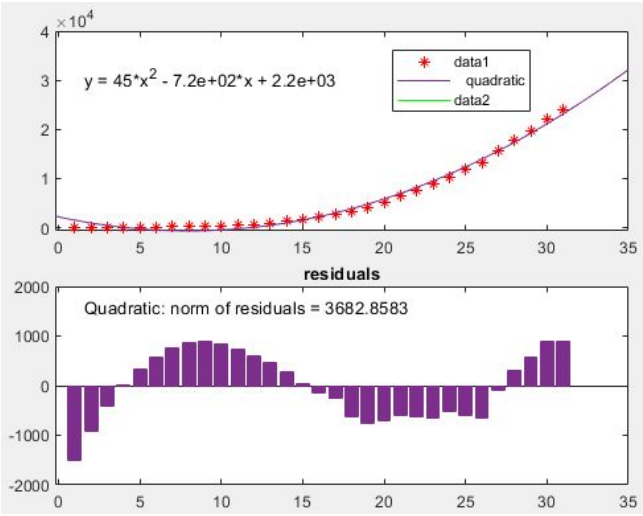
China: total cases	China: total deaths
	

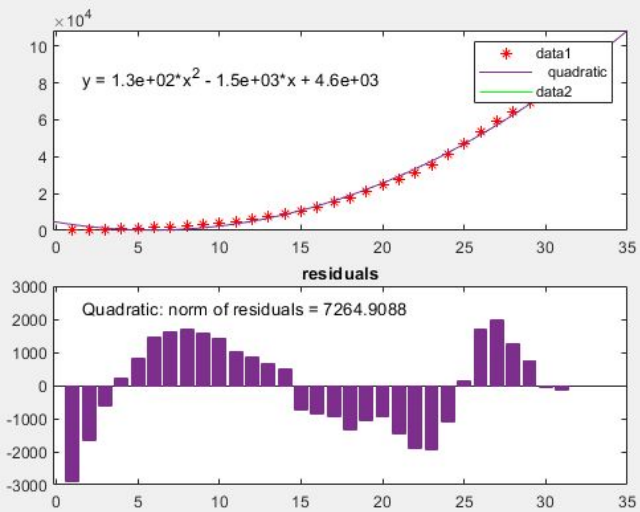
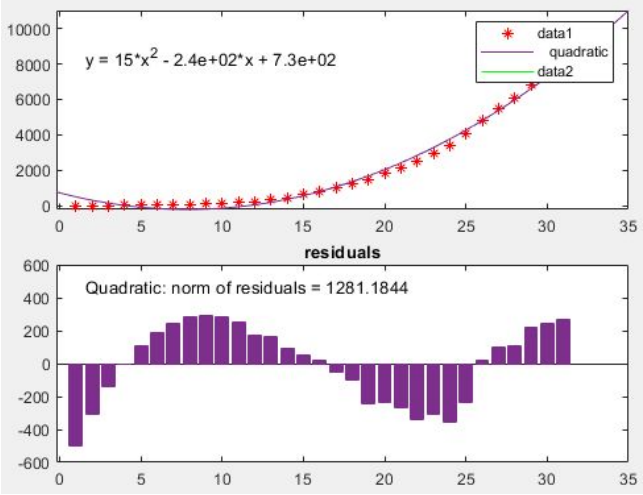
US: total cases	US: total deaths
	

Italy: total cases	Italy: total deaths
	

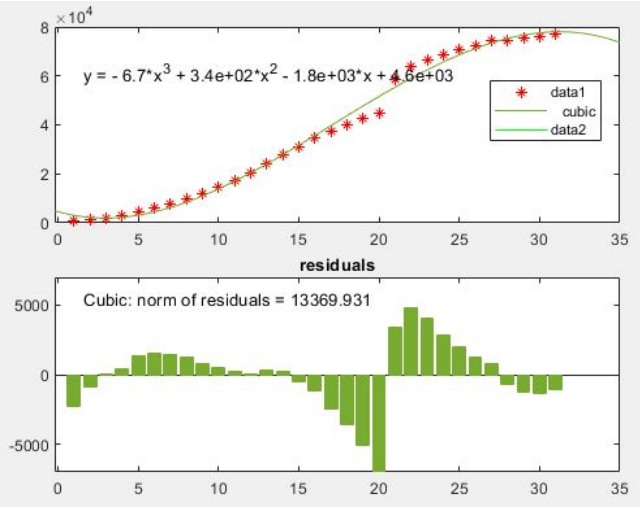
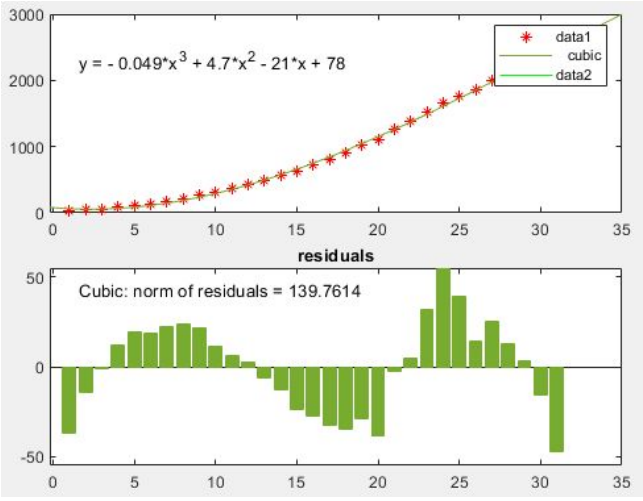
1.2. Quadratic polynomial

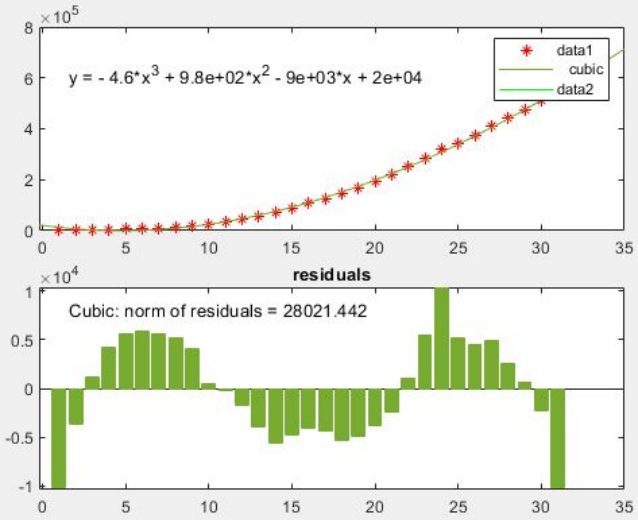
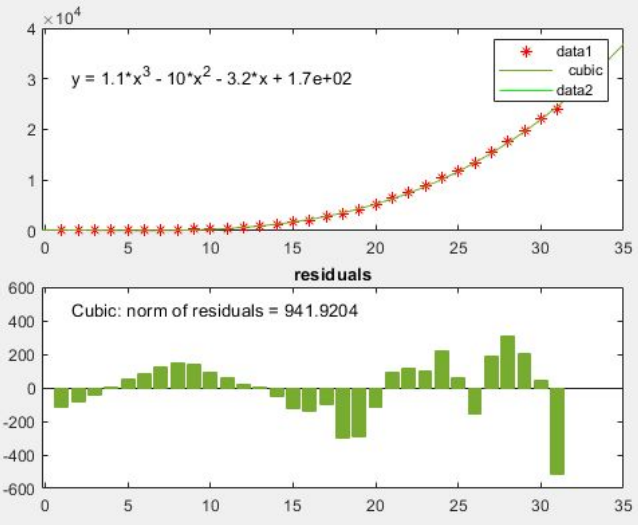
China: total cases	China: total deaths
	

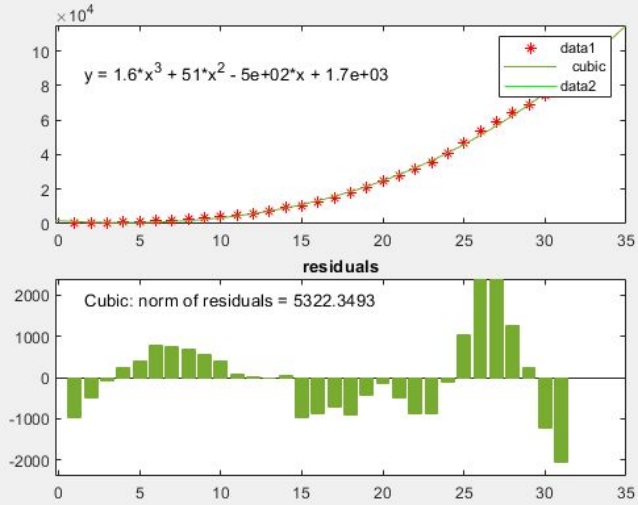
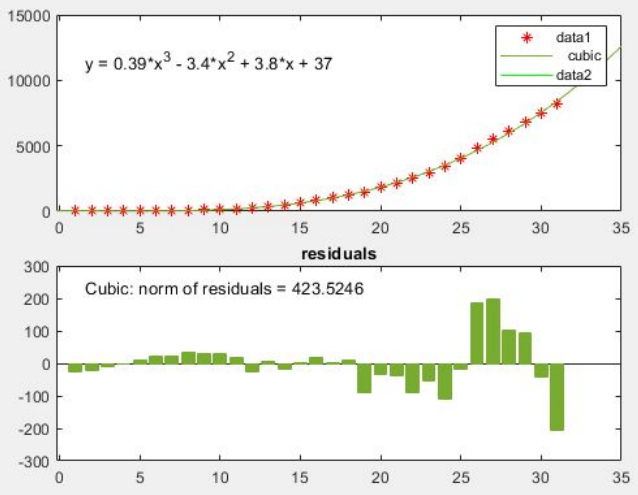
US: total cases	US: total deaths
	

Italy: total cases	Italy: total deaths
	

1.3. Cubic polynomial

China: total cases	China: total deaths
	

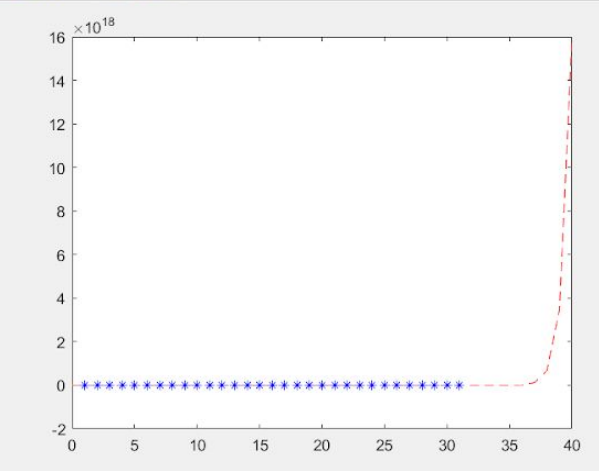
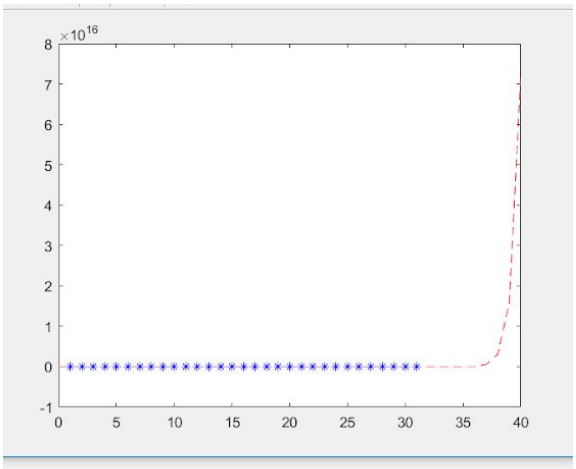
US: total cases	US: total deaths
	

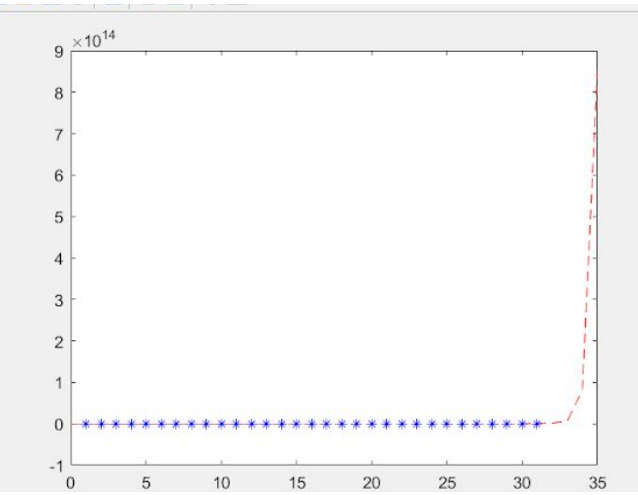
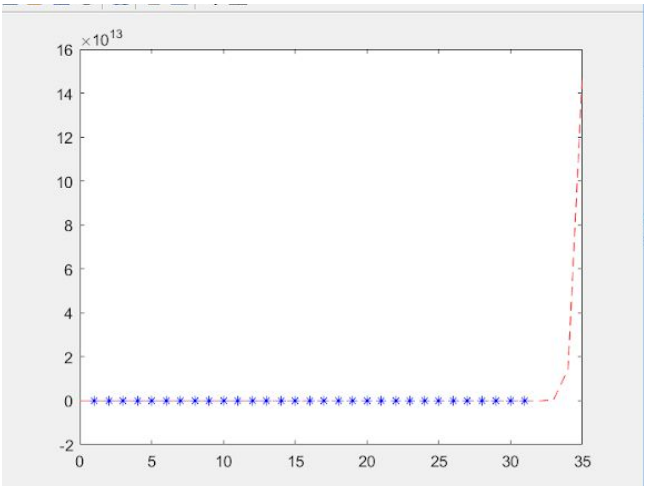
Italy: total cases	Italy: total deaths
	

Based on the previous 3 models results and predictions, the cubic polynomial regression was the best among them as it has the lowest norm of residuals.

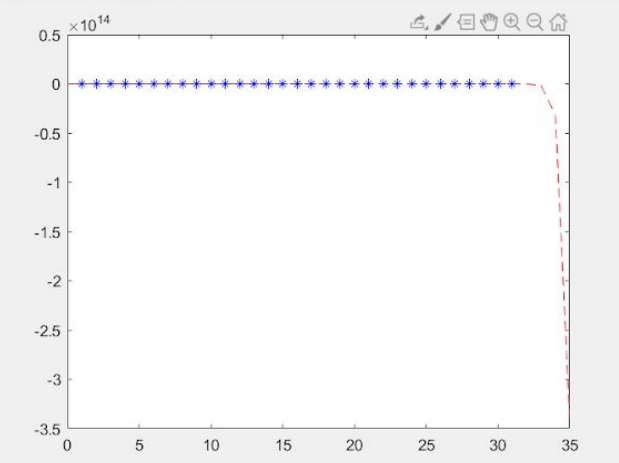
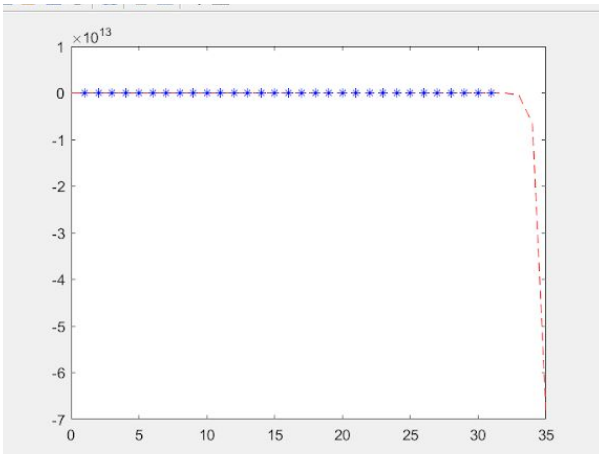
2. Interpolation

2.1. Newton's Interpolation

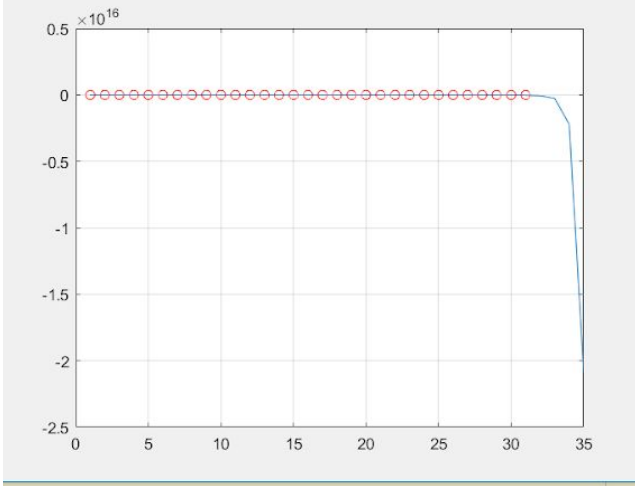
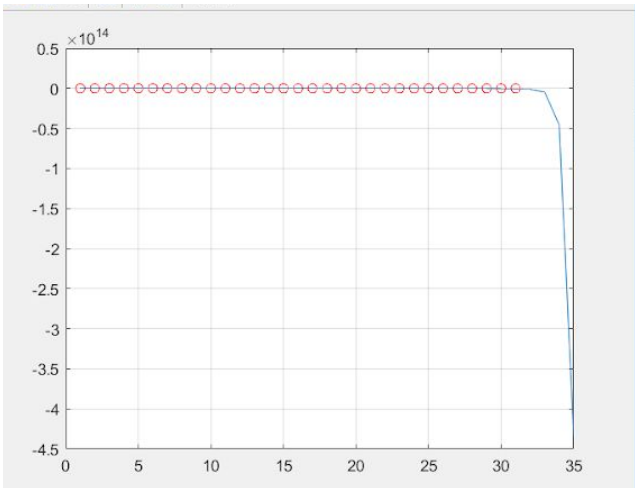
China: total cases	China: total deaths
 Residual L2 norm = 3.3728e+11	 Residual L2 norm = 1.5436e+09

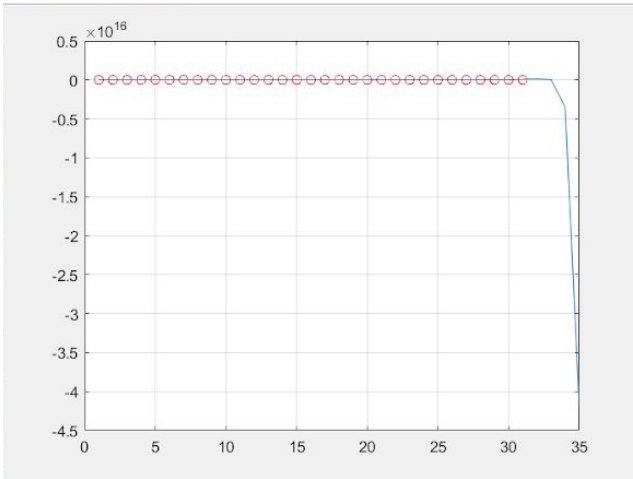
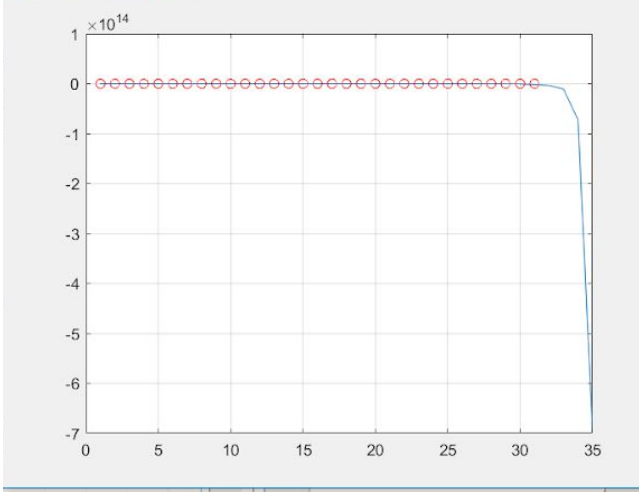
US: total cases	US: total deaths
	

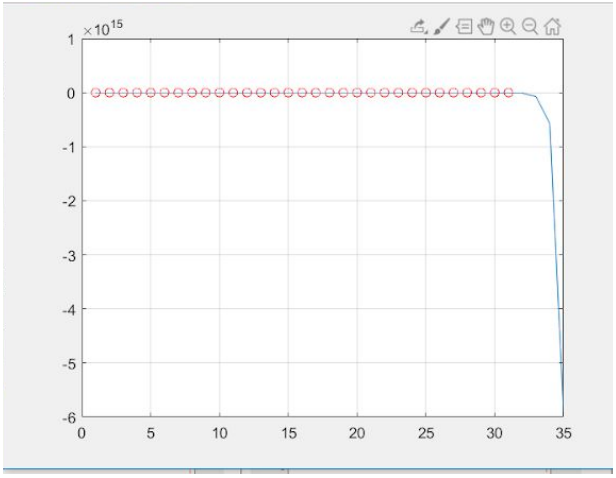
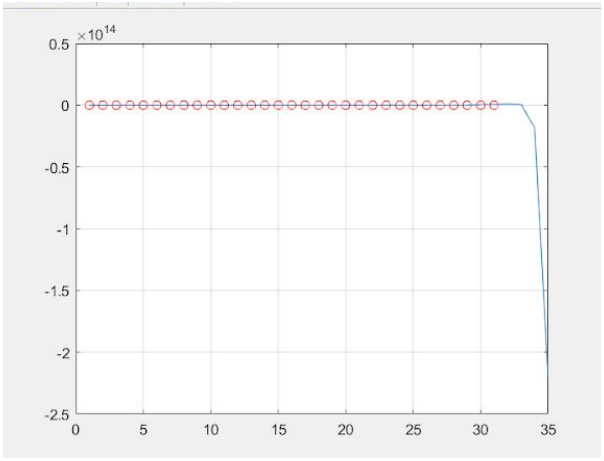
Residual L2 norm = 5.8004e+11	Residual L2 norm = 2.5928e+10

Italy: total cases	Italy: total deaths
	
Residual L2 norm = 1.5036e+11	Residual L2 norm = 1.6944e+10

2.2. Lagrange Interpolation

China: total cases	China: total deaths
	
Residual L2 norm = 1.0302e+15	Residual L2 norm = 1.2729e+13

US: total cases	US: total deaths
 Residual L2 norm =8.3070e+14	 Residual L2 norm =3.3496e+13

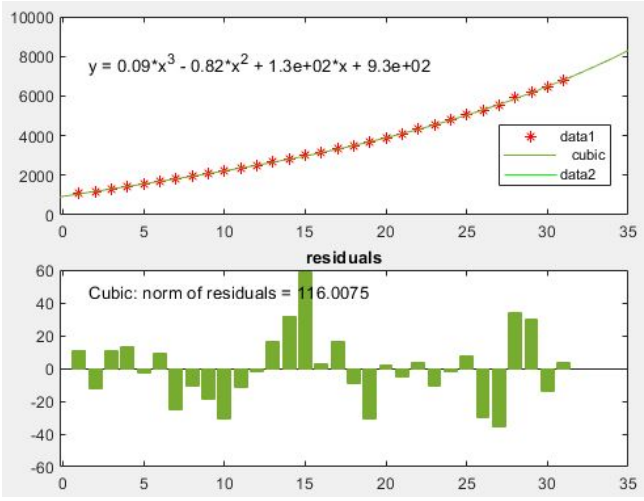
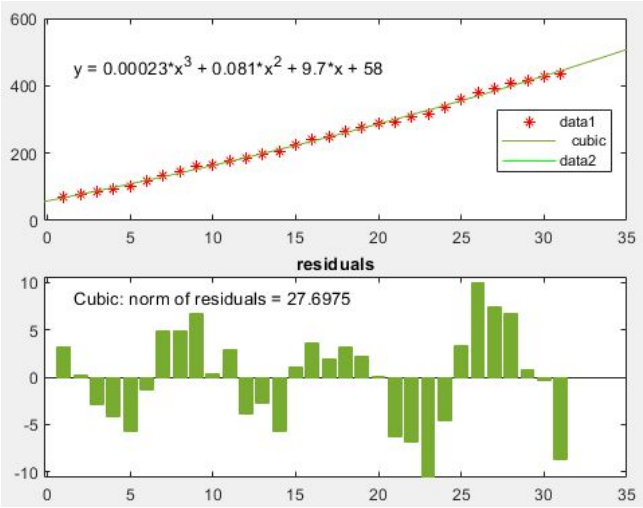
Italy: total cases	Italy: total deaths
 Residual L2 norm = 9.1146e+13	 Residual L2 norm = 9.4751e+12

Based on the previous 2 models results and predictions, the Newton interpolation regression was the best among them as it has the lowest norm of residuals.

3. Egypt model

When comparing the L2 norm of the residuals in our 2 best filt models: the cubic polynomial and Newton's interpolation, the cubic polynomial had the lower residual. That is why after testing the data collected from Egypt with the cubic polynomial least square regression. The data used in this model was also for a 1-month interval starting from April 4th to May 4th.

The results were as follows:

Egypt: total cases	Egypt: total deaths
	

Conclusion

After applying the cubic polynomial regression on the data from Egypt, it fit the data perfectly. This method can help us more in predicting the number of daily infected cases and deaths. Thus, that could help us in controlling the disease spreading by limiting all the factors that increase disease transmission. However, we can't take the results as a definite way in predicting the future of COVID-19 in Egypt as we chose the model based on data collected from 3 different countries who each dealt with the virus at first differently and continue dealing with it differently. Each country has a different Health care system. Moreover, this all depends on the citizens whether they will follow rules or not and follow social distancing. In conclusion, we cannot ignore the importance of these numerical methods helping us in predicting future data based on the current one.

References

[1] Wenjie Tan, Xiang Zhao, Xuejun Ma, Wenling Wang, Peihua Niu, Wenbo Xu, et al. A Novel Coronavirus Genome Identified in a Cluster of Pneumonia Cases — Wuhan, China 2019−2020[J]. China CDC Weekly, 2020, 2(4): 61-62. doi: 10.46234/ccdcw2020.017

[2] "China Coronavirus: 82,881 Cases and 4,633 Deaths - Worldometer", Worldometers.info, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/country/china/.

[3] "United States Coronavirus: 1,215,862 Cases and 70,147 Deaths - Worldometer", Worldometers.info, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/country/us/.

[4] "Italy Coronavirus: 211,938 Cases and 29,079 Deaths - Worldometer", Worldometers.info, 2020. [Online]. Available: https://www.worldometers.info/coronavirus/country/italy/.