Proceedings of GREAT Day

Volume 2020 Article 1

2021

Curve Fitting Techniques for Predicting Corona Virus Cases

Frank Addeo
SUNY Geneseo

Curve Fitting Techniques for Predicting Corona Virus Cases

Erratum

Sponsored by Ahmad Almomani

Predicting Coronavirus Cases Using Curve-Fitting Techniques

Frank Addeo

sponsored by Ahmad Almomani

Abstract

With coronavirus becoming an increasing global concern each day, it is obvious that we must understand what is truly happening. Using curve fitting techniques such as Fourier series, cubic splines and least squares approximation, an accurate model can be made to fit the data. Analyzing and comparing the three approaches, the best method to fit these data can be seen. We will also compare the fitting curve methods with the solution of the logistic model solution. *Editor's note: This paper was presented in April 2020, relying only on data and understandings about COVID-19 available then.*

The COVID-19 pandemic of 2020 has been one of the most significant recent events worldwide. It has impacted each and every American in one way or another. As millions of people contracted the lethal virus, it seemed the world as we know it took a turn for the worst. Intensive care units filled rapidly because of the need for ventilators. Essential items flew off of supermarket shelves, students of all ages were required to take classes online, working-class people were urged to work remotely, and sports venues, concert halls, and restaurants were all temporarily shut down. Consequently, the stock market crashed as panic settled in. Our lives were derailed and forced into a period of almost complete quarantine as thousands of Americans perished at the hands of the novel Coronavirus.

Due to the severity of the virus, it is important to analyze how the virus is spreading. For example, predicting how many Coronavirus cases there would be at a particular time in the United States could offer plenty of insight. Anticipating the number of sick people can enable hospitals to reserve ample intensive care units and ventilators for those whose lives depend on them. So, how could the number of cases be predicted? Using various curve-fitting techniques, the number of future cases can be accurately estimated. For the sake of this study, three methods have been used to create an estimate: Least Squares Approximation, Cubic Spline Approximation, and the Logistic Model. As a point of reference, April 22, 2020, SUNY Geneseo's GREAT Day, would be the date to predict how many Coronavirus cases there would be using each method.

In order to predict coronavirus cases over time using curve-fitting, previous data regarding the total number of cases must be examined. *Figure 1* is a representation created using MATLAB depicting coronavirus cases over time beginning on February 15, 2020. The total number of cases over time was reported from Worldometer. On the x-axis, days since February 15 are incremented. Points along the x-axis can be referenced simply by a number. For example, April 4, 2020, would mark the 50th day since the initial number of cases was recorded on February 15. On the other hand, the y-axis measures the total Coronavirus cases. Hence, as shown in the graph, there were 311,357 total cases on April 4. All data was recorded up until April 12 which is reference point 58 (the 58th day since February 15), at which final predictions for GREAT Day were calculated.

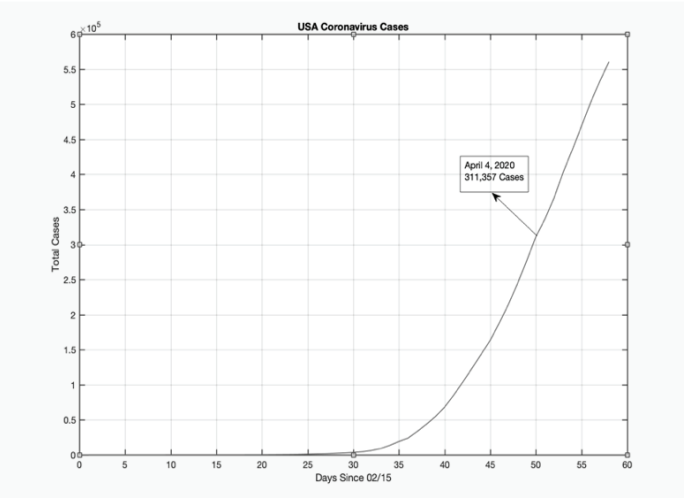

Figure 1: Coronavirus Cases in the USA from 02/15 to 04/12

The first curve-fitting method to be examined is the Least Squares Polynomial Approximation. The curve being observed is not linear. Thus, linear regression is not a suitable method to fit the data. Nonlinear regression will yield a higher-order polynomial and will be more accurate. The basic form of the nonlinear model is $\hat{Y}_i = f(\hat{X}_i:\hat{\beta}) + \epsilon$. The values $\hat{Y}_i$ refer to the predicted values along the *y*-axis. $f(\hat{X}_i:\hat{\beta}) + \epsilon$ is a function of x and beta where beta is the term which is the sum of squares error. Finally, epsilon is simply the error term for the model. To obtain this model, iterative procedures must be performed to minimize the sum of squares error, which is $\Sigma(Y_i - \hat{Y}_i)^2$. The model will be a simple function of and , and thus the only unknown parameters are the beta terms (Biran, 2019). To find the beta terms and find the general equation of the curve, independent least squares approximations are performed. Doing these iterations by hand would require a great deal of time and patience, so MATLAB code was created to quickly and efficiently estimate the beta parameters.

The method of Least Squares Regression does have strengths and weaknesses. For example, this method is appropriate for all kinds of datasets and should provide accurate

results each time. Furthermore, it is a rather efficient method and can produce a result quickly. The one potential downside to this method is the complexity of the iterative procedures (Biran, 2019). Displayed in *Figure 2* is the output of Least Squares Regression. In blue is the least squares approximation, and displayed in red are the actual reported values of Coronavirus cases. Running the code yields a simple polynomial that can be used to accurately predict the number of Coronavirus cases in the United States at a particular time. As mentioned before, GREAT Day is the 68th day since February 15. This means that 68 is plugged into the polynomial and the output is the total number of Coronavirus cases that we can expect. Using this model the prediction for GREAT Day was 825,570 cases, and the actual reported value was 854,385.

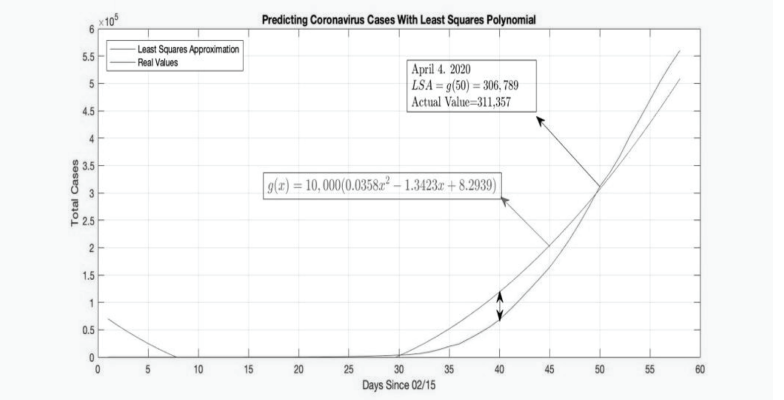

Figure 2: Output of Least Squares Regression

The next method is Natural Cubic Spline Approximation. The method of cubic splines constructs a piecewise defined function that fits many forms of data quite accurately. As previously discussed, this data is not linear and thus a smooth curve must be constructed to fit the data. Starting with points, intervals must be constructed to fit the data. For this data, there are 58 points (dates ranging from February 15 to April 12) and thus there will be 57 individually piecewise defined intervals. Using code in MATLAB, the coefficients , $s_{k,i}$, must be derived in order to create these intervals. The general form for a cubic spline is: $S(x)=s_k(x)+s_{k,0}+s_{k,1}(x-x_k)+s_{k,2}(x-x_k)^2+s_{k,3}(x-x_k)^3+\cdots$ *for* $x \in [x_k, X_{k+1}]$ *and* $K=0,1,\ldots,N-1$ (Matthews & Fink, 2005, p. 281).

In order to find these coefficients, there are a few properties that must be satisfied. First, the spline must interpolate the data points. Each node, x_k , represented in the spline matches the corresponding number of Coronavirus Cases at that given point. Second, the function must be smooth and continuous at each interval. This means that the second derivative exists implying that there are no corners or abrupt ridges in the curve (Newton, 2011).

Cubic Spline Approximation incorporates all datasets quite well. Essentially, this method creates a new approximation for each given interval. It modularizes the whole curve into smaller ones which create greater precision at points within the curve. Hence, this method is not greatly affected by irregularities and anomalies within a

dataset. This method however can be a bit difficult to use when trying to predict future values. Since each interval is individually defined, there is no "best" approximation for future values. Therefore, the last defined interval can be extended in order to make predictions. In other words, the polynomial given in the last interval can be evaluated at a given value of *x,* much like that of the least squares method. This may cause a greater error.

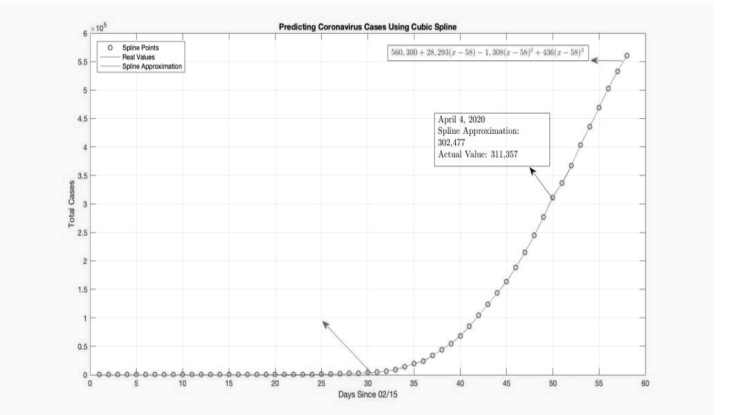

Figure 3: Cubic Spline Approximation of Coronavirus Cases from 02/15 to 04/12

Figure 3 contains the Cubic Spline Approximation fitted to the curve of Coronavirus cases over time. One may ask, where are the actual reported values? The Cubic Spline Approximation matches the data so precisely that it covers the actual reported values. Within the figure there is a small snapshot that shows the discrepancy between the spline and the actual curve between the 30th and the 31st day since February 15. Since the spline is so close to the actual reported values, the exact number of cases at a particular time during any given day can be estimated. In the right corner of the figure is the last interval of the spline which can be used to make future predictions. Thus, the value 68 can be plugged in for x to obtain the predicted future value for GREAT Day. Using this value of x, the expected number of Coronavirus cases is 1,149,230. This value is expected to be a bit high because only the last interval was used.

The third and final model used for this study is the Logistic Model. Coronavirus is extremely transmissible. Speculation has concluded that the virus could live on surfaces for days. Furthermore, scientists believe that the virus could travel a distance of six feet. For these reasons, the number of infected people was growing exponentially without bound through March and April of 2020. This rate of growth is not practical for the future however. It was expected that the curve would flatten out at some point. Due to government intervention requiring people to socially distance and wear masks, the rate of growth has decreased and the number of cases has approached an asymptotic value, or equilibrium solution. The logistic model accounts for this government intervention. Unlike the other curve-fitting methods, this method will

not grow without bounds and will at some point level out. Eventually, the number of cases will reach a carrying capacity (L) that cannot be sustained by the population, and thus will cause the curve to level out. In this circumstance, a suitable value for L would be the total number of hospital beds in the United States, which happens to be 924,147 (AHA, 2020). The only other unknown when constructing a Logistic Model is the rate of growth. In *Figure 4* the rate of growth at each day is graphed. After much consideration, the best approximation for a general rate of growth using this model is simply averaging out the rates over each day. Now that L, and R have been found, the values can be plugged into the general formula for the Logistic Model in order to create a realistic prediction for the future. The general equation for the Logistic Model is $y = \frac{Ly_0}{y_0+(L-y_0)e^{-rt}}$. So far L and R have been accounted for and is simply the number of cases that were initially present on February 15, 2020.

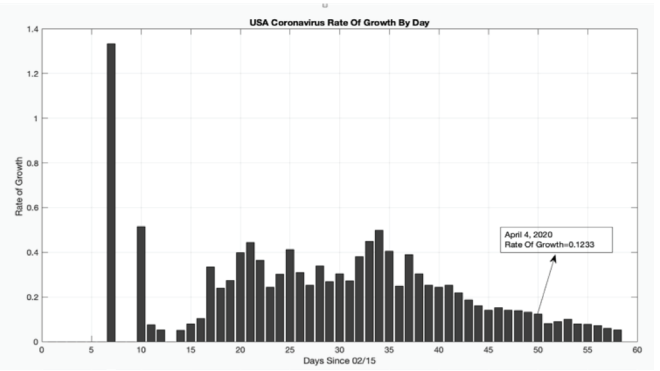

Figure 4: Rates of growth each day from 02/15 to 04/12

In *Figure 5*, the Logistic Model is represented by the values in blue and the real, reported values are in red. In order to make a prediction using this model, the value, t, represented in the general equation above is treated as x would be. Thus, 68 is inputted for t and the prediction is yielded. Using this method, the predicted number of cases for GREAT Day (April 22, 2020) is 897,490 which is relatively close to the actual value of 854,385.

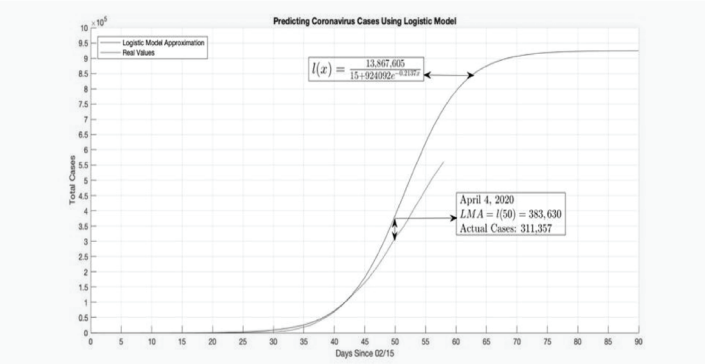

Figure 5: Logistic Model Approximation of Coronavirus Cases from 02/15 to 04/12

There are many factors contributing to error in this study. For example, approximately one of every four Coronavirus-positive patients is asymptomatic, and therefore may not be tested and thus will not be included in the total number of cases (Woodward, 2020a, 2020b, & 2020c). There has also been speculation of false reporting, which would indicate that these totals are wrong anyway. This is logical considering COVID-19 caused a time of crisis which may lead to faulty reporting. Additionally, in the United States, testing was quite hard to come by in the early months of the pandemic. As of March 8, only five people per million were tested for coronavirus which is significantly lower than per capita testing in South Korea, which had a rate of 3,692 tests per million people on the same date (Woodward, 2020a, 2020b, & 2020c). Obviously, people cannot be counted as having Coronavirus if they have not actually tested positive for the virus. Practically, this is the largest factor influencing the numbers in the United States.

As expected, the Logistic Model approximation and the Nonlinear Regression approach were the most accurate for predicting future Coronavirus cases. Generally, the Nonlinear Regression Approximation appears to be the most efficient method. It is quite flexible and can be easily changed to variance in the data. For example, making a prediction for October 2020 would only require incorporating data from April until late September. Using the Logistic Model to make the same prediction would be a bit more complicated because a new rate of growth must be established as well as a new carrying capacity. The Cubic Spline approximation was the least accurate result, but is not entirely useless because it fits the data best over the period that is already known.

References

American Hospital Association [AHA]. (2020). Fast Facts on U.S. Hospitals, from www.aha.org/statistics/fast-facts-us-hospitals.

Biran, A. (2018). Cubic Spline. In *Geometry for Naval Architects* (pp. 305-324). Oxford: Butterworth-Heinemann. doi: 10.1016/B978-0-08-100328-2.00018-3

Mathews, J. H., Fink K. D. (2005). *Numerical Methods Using MATLAB*. Pearson Prentice Hall.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Nonlinear Least Squares Regression. In *Introduction to linear regression analysis* (pp. 101-143). Hoboken, NJ: Wiley.

Newton, M. (2011). Cubic Splines on the HP 50g. Retrieved March 21, 2020, from https://resources.thiel.edu/mathproject/Spline/index.htm

Sharov, A. (1997, March 2). Logistic Model. Retrieved March 22, 2020, from https://web.ma.utexas.edu/users/davis/375/popecol/lec5/logist.html

Woodward, A. (2020a, March 9). One Chart Shows How Many Coronavirus Tests per Capita Have Been Completed in 8 Countries, The US Is Woefully behind. Business Insider. Retrieved April 10, 2020, from www.businessinsider.com/coronavirus-testing-covid-19-tests-per-capita-chart-us-behind-2020-3.

Woodward, A. (2020b, April 3). It's Estimated 1 in 4 Coronavirus Carriers Could Be Asymptomatic, Here's What We Know. *ScienceAlert*. Retrieved April 10, 2020, from www.sciencealert.com/here-s-what-we-know-so-far-about-those-who-can-pass-corona-without-symptoms.

Walton, A. G. (2020c, April 14). Small New Study May Provide Clues Into Asymptomatic Carriers Of Coronavirus/COVID-19. Retrieved April 17, 2020, from http://www.forbes.com/sites/alicegwalton/2020/04/14/small-new-study-in-pregnant-women-may-provide-clues-about-coronavirus-infection-rate/

Worldometer. United States Coronavirus Cases. (2020). Retrieved March 25, 2020, from www.worldometers.info/coronavirus/country/us/.