International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.8, Issue.5, May 2020 E-ISSN: 2347-2693

A Mathematical Model to Forecast and Compare COVID-19 Outbreak in Male and Female using Polynomial Regression Analysis

A. Bhattacharya[1], P. Dutta[2], S. Halder[3*], P. Banerjee[4], S.K. Bandyopadhyay[5]

[1,2,3,4]Department of Computer Science, The Bhawanipur Education Society College, Kolkata, India
[5]The Bhawanipur Education Society College, Kolkata, India

[*]*Corresponding Author: sanjib.halder@thebges.edu.in, Mob.: +91-9830351321*

DOI: https://doi.org/10.26438/ijcse/v8i5.139143 | Available online at: www.ijcseonline.org

Received: 18/May/2020, Accepted: 23/May/2020, Published: 31/May/2020

Abstract— COVID-19, a fast-spreading infectious disease, has taken away the life of a large number of people miserably with its dreadful consequences globally. The situation demands immediate controlling measures to stop the quick spread of disease. Governments and other legislative bodies rely on insights from different predictions to suggest and enforce relevant steps. Forecasting, however, requires sufficient authentic historical data. Due to deficiency of important authentic data, standard models have shown comparatively lower accuracy for prediction. Among the standard models for COVID-19 global pandemic prediction, statistical models have received more attention by authorities. So, far the predictions from all the standard models gave insights that are irrespective of gender. In this study we proposed a statistical model that uses regression analysis techniques to predict the rate of growth of Covid-19 infections in males and females based on their population density in each country. Finally, we used a dataset of COVID-19 patients to evaluate the accuracy of the proposed model.

Keywords— Covid-19, Corona Virus, Curve fitting, Regression analysis, Pandemic

I. INTRODUCTION

The novel Coronavirus disease 2019 (COVID-19), caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus2) started spreading at Wuhan in China in the late 2019 and started its uncertain and precarious journey in January 2020 globally [1,2]. As a result, the World Health Organization (WHO) declared the outbreak as a Public Health Emergency of International Concern on 30th January and later in March 2020 they declared it as COVID-19 Pandemic [1,2]. This highly infectious virus primarily spreads between people during close contact, mainly via small droplets produced by coughing, sneezing, and talking. Its symptoms are fever, cough, shortness of breath, loss of smell or it might show none. Its usual onset time is 2-14 days from the day of infection [3]. Most of the COVID-19 effected countries in the world have taken the policy of nation-wide lock-down and isolation to minimise the spreading of the disease [4]. As of 10th May, 2020, there have been more than 4 million confirmed cases and reported more than 279300 deaths globally. USA has the most number of cases of deaths. The recent global scenario of COVID-19 pandemic has exhibited complex and nonlinear in nature. In this paper a mathematical model is proposed to predict the future COVID-19 outbreak separately for males and females. The model was applied on a dataset collected from Kaggle.com.

The rest of the paper is organized as follows: Section II introduces the proposed research method, which has two subsections. The first subsection describes the pre-processing of data, the second subsection discusses the implementation strategy, Section III contains data analysis and prediction. Finally, the paper is concluded in Section IV.

II. METHODOLOGY

A. Data Pre-Processing

In this research work the dataset from kaggle.com is pre-processed to get the most informative data values considering top 22 severely affected countries. To obtain effective data, the time frame considered to be 25th March to 7th May, 2020. The severity of COVID-19 has been in full swing during this time. After the collection of data, the parameters like latitude and longitude, total population and population densities for the selected countries are obtained. The aims of the analysis and predictions are based on COVID-19 cases affecting males and females based on population densities. So, the population densities of each country are used as parameters for regression analysis and then perform necessary calculations.

B. Implementation Technique

This research is conducted to develop a regression model to predict the increase or decrease in the number of COVID-19 cases around the world based on gender. Here polynomial regression and curve fitting methods are used to get a visual detail as well as numerical results [5]. The entire process is shown in Figure1. In recent times, prediction analysis using polynomial regression is very effective as it provides best curve fit and more accurate results than linear regression, time series, and random forest methods. Polynomial regression uses N number of

test cases to make prediction and perform curve fitting where the minimum value of N is 2 and this N is added with the original linear regression value to get the final result of the prediction analysis. As per the dataset the male cases share a negative linear regression and female cases share a positive linear regression with the population distribution of the males and females respectively. Square test and cubic test methods are used to obtain p-values and f-values of male and female test data cases where p-values are used for probability analysis and f-values are used to verify the statistical validity of data. Square test is capable of making prediction in a small number of intervals whereas the cubic test uses a large number of intervals. The formulas for both are as follows [5].

$$XSQ = x^2 \quad (1)$$

$$XCUB = x^3 \quad (2)$$

Now, the dataset is categorised into two sub-parts based on male and female instead of using the individual data values. Next it will divide the population data in certain intervals (like 0-10,10-20,….) with the help of Simpson's 1/3rd rule and then make predictions using test case equations. Thus, we obtain approximated values for curve fitting. The formula of Simpson's 1/3rd rule is as follows:

$$\int_a^b f(x)dx = h/3[(y0 + yn) + 4(y1 + y3 + \cdots + yn - 1) + 2(y2 + y4 + \cdots + yn - 2)] \quad (3)$$

Where, b = upper limit, a = lower limit and f(x) represents the regression function for each interval.

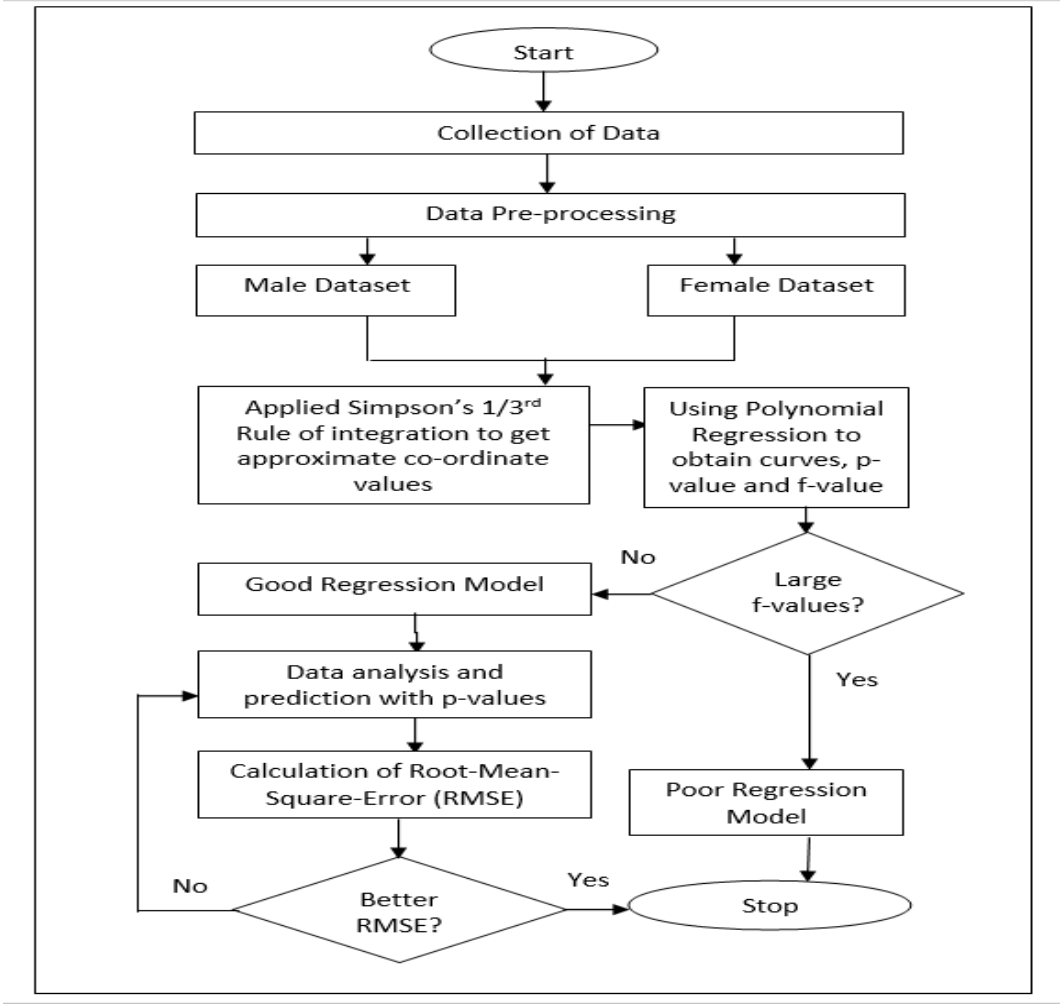

Figure.1: Workflow diagram

III. RESULTS AND DISCUSSION

Data Analysis and Prediction

To predict the gender wise confirmed cases, the cubic and the square tests are used. It was also checked that there is no overlapping of the training data and the test data. Next it is calculated the Root-Mean-Square-Error (RMSE) value to check the validity and accuracy of the data and regression model. The estimated f-values obtained for the male dataset are 2.4851 and 3.4457 by square and cubic tests respectively. Since, both the values are low so we can consider that the errors in our calculations and data analysis are minimal and hence our regression model to be a perfect one. The following graph is obtained.

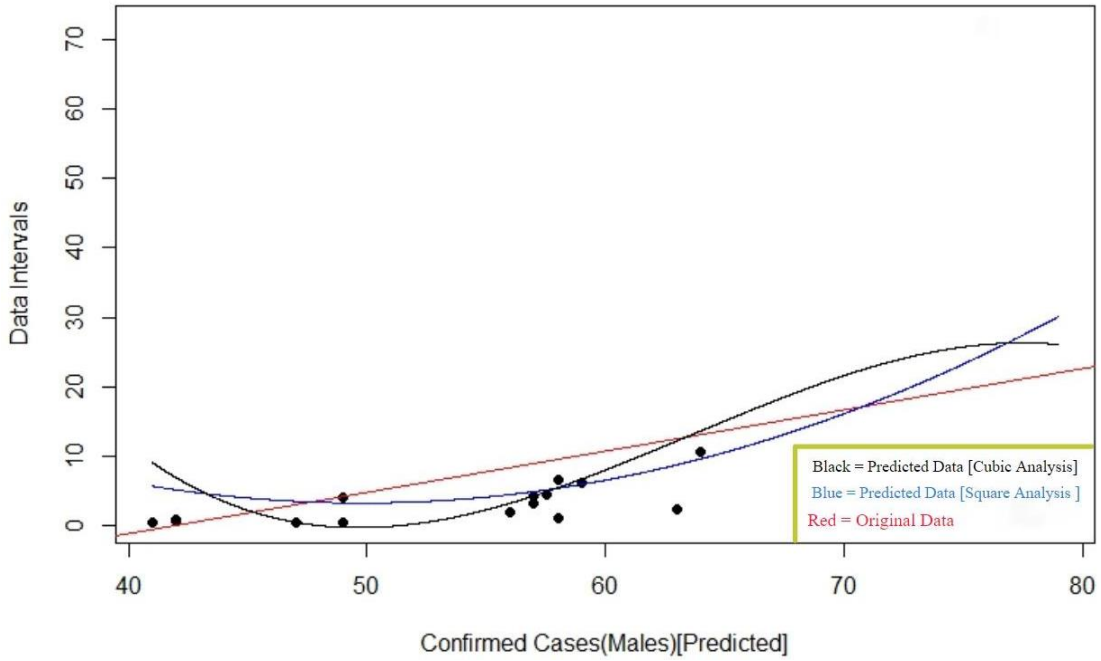

Figure.2: Population density vs Confirmed Covid-19 cases of male

The data is divided into certain intervals which in-turn consists of the population density of males and females and is represented by y-axis. The intervals are divided equally with a common difference of 10 for all in order to maintain a decent test result like, for each of the intervals created we would consider 10% of the population density every time when it is implementing a prediction test. With the help of this the coordinates are obtained by selecting the approximate values that are generated by using Simpson's $1/3^{rd}$ rule from the confirmed cases represented by x-axis.
In Figure2, the red line represents the original data while the blue curve is for square test and black curve is for cubic test. From figure 2 the cubic curve gives a more accurate result as it covers the max number of intervals as compared to square test. The analysis gives an idea that the number of confirmed cases will initially increase and then will start to fall. For the above data we get a negative regression relationship.

For the female dataset, the estimated f-values obtained for both the cubic and square test are 3.892 and 2.434 respectively. Similar like male cases, both the f-values of the female dataset are low so the regression model to be a perfect one.

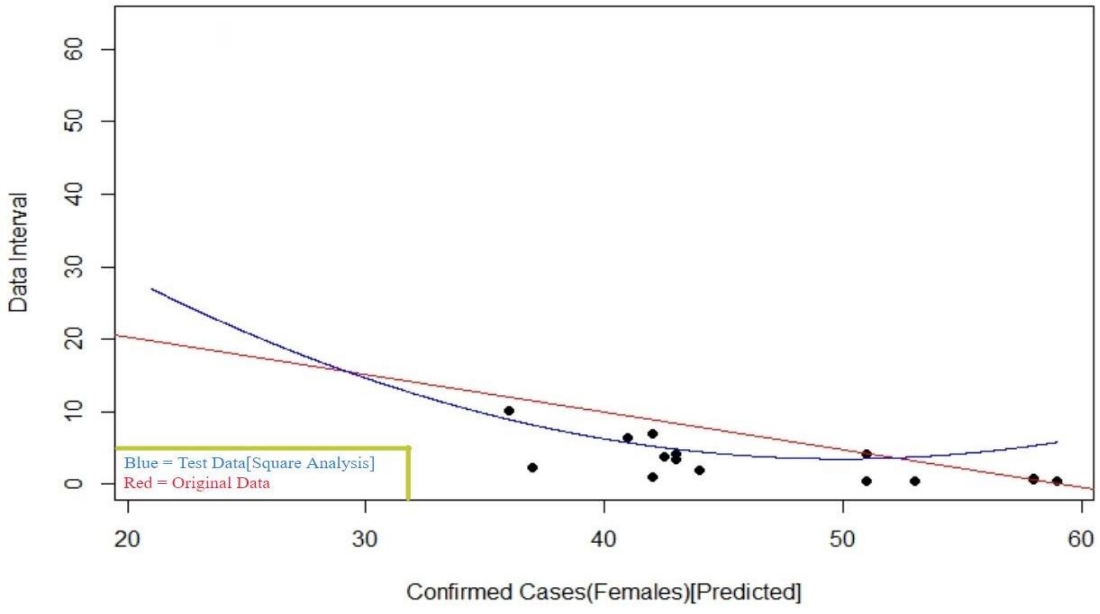

Figure.3: Population density vs Confirmed Covid-19 cases of female

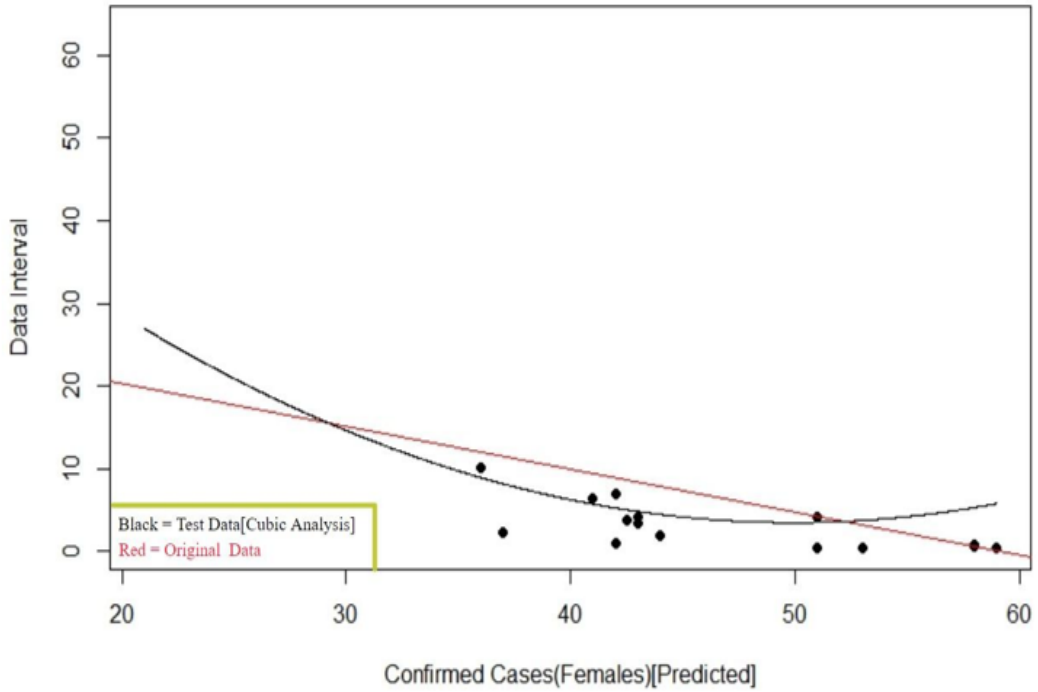

Figure.4: Population density vs Confirmed Covid-19 cases of female

The results of both tests are equal so one curve is generated. In Figure3, the blue curve represents the predicted value using the square test and the red line represents the original data and in Figure4, black curve represents the predicted value using the cubic test and the red line represents the original data. Analysing the above two graphs, an assumption can be made that the confirmed cases of females tend to fall in both the tests and then makes a small rise in the end of our graph. The confirmed cases of females and female population density share a positive regression relationship.

Based on the p-value we can make our prediction using the formula-

$(p(sq)/p(o))\times 100=$ The percentage form of prediction

Where p(sq) and p(o) is the p-value of square analysis and the p-value of the original dataset respectively. If the p-value is low, then the predictions will be more accurate. So the necessary calculations for male dataset are-

$$p(sq)=0.995$$
$$p(o)=0.22$$

As the value of *p(sq)* is greater than *p(o)*, the square test provides a prediction that the cases of infection will increase by 45.22%approximately in each data interval with a frequency of 10 for males.

Similarly, for cubic test the same formula is by replacing p(sq) with p(cubic). The p-values for original dataset and cubic analysis are 0.3 and 1.265 respectively. Cubic test for male dataset shows that the cases of infection will increase by 42.1%approximately in each data interval with a frequency of 10.

Since, for the female dataset both square and cubic test provide equal result so, the p-value obtained in both cases is 0.52 and p(o) is 0.22. It leads to a conclusion that the female cases of infection will increase by 23.63%approximately in each data interval with a frequency of 10.

Root Mean Square Error (RMSE) for both the male and female cases is 49.16788 and 50.95789 respectively. It is a good indicator as the lower the RMSE the better is the regression model. Apart from that it also worked on intervals for predictions instead of huge volumes of data. It is known that a linear regression model can be considered as a perfect one if the RMSE Value of the test data is almost equal to the RMSE Value of the original data. Results obtained 49.16788(for square test), 51.98735(for original data) and 50.95789(for cubic data) respectively. So, it is clear that the RMSE of the test cases of data is almost equal to the RMSE of our original data. Hence, the regression model is to be a more improved as well as more accurate model.

IV. CONCLUSION AND FUTURE SCOPE

The proposed method investigated the global pandemic of the SARS-CoV-2 worldwide in practical circumstances. A statistical model has been established which uses regression analysis to obtain our predictions. In this paper, the rate of growth of COVID-19 infections in males and females are based on their population density in each country. The f-values and the p-values obtained for both the cubic and square tests in polynomial regression analysis are used to determine the predictions. The percentage obtained from the p-values of both original and predicted data are used to obtain our observations. The result shows that man has higher tendency of getting infected with COVID-19 than woman.

In future, research can be carried out to find the impact of natural parameters like, Geographical position, change of atmospheric temperature due to seasonal change etc. Some other parameters like underlying health conditions, sanitation facilities, social distancing, average density of people in each city etc. can also be considered to perform analysis and hence obtain a clearer observation of COVID-19 infections in each city.

REFERENCES

[1] M.K. Arti, K. Bhatnagar, *"Modeling and Predictions for COVID 19 Spread in India"*, ResearchGate, DOI: 10.13140/RG.2.2.11427.81444, published on April 01, 2020.

[2] L. Li, Z. Yang, Z. Dang, C. Meng , J. Huang, H. Meng , D. Wang, G. Chen, J. Zhang, H. Peng , Y. Shao, *"Propagation analysis and prediction of the COVID-19"*, Contents lists available at ScienceDirect, Infectious Disease Modelling, https://doi.org/10.1016/j.idm.2020.03.002

[3] F.A.B. Hamzah, C. Lau, H. Nazri, D.V. Ligot, G. Lee, C.L. Tan, *"CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction"*, [Submitted]. Bull World Health Organ. E-pub: 19 March 2020. doi:http://dx.doi.org/10.2471/BLT.20.255695

[4] U.K. Tiwari, R. Khan, *"Role of Machine Learning to Predict the Outbreak of Covid-19 in India"*, Journal of Xi'an University of Architecture & Technology, Volume XII, Issue IV, 2020.

[5] D.R. Chatterjee, *"Log Book – Practical guide to Linear & Polynomial Regression in R"*.

[6] H. Bouhamed, *"Covid-19 Cases and Recovery Previsions with Deep Learning Nested Sequence Prediction Models with Long Short-Term Memory (LSTM) Architecture"*, International Journal of Scientific Research in Computer Science and Engineering, Vol.8, Issue.2, pp.10-15, April (2020).

[7] M.R. Bhatnagar, *"COVID-19: Mathematical Modeling and Predictions"*, submitted to ARXIV.

[8] A.V. Singhania, A. Bhattacharya, P. Banerjee, S. Halder, *"Statistical Inference of the Effectiveness of Lockdown for COVID 2019 in India", International* Research Journal of Engineering and Technology (IRJET), Volume: 07, Issue: 05 , May, 2020.

Authors Profile

Anuran Bhattacharya former student of South Point High School, Kolkata, pursuing B. Sc. Computer Science (Hons) course at The Bhawanipur Education Society College. His keen interests revolves around machine learning, robotics, network security, etc

Pratyush Dutta former student of Julien Day School, Kolkata, pursuing B. Sc. Computer Science (Hons.) course at The Bhawanipur Education Society College. His interest is in learning different programming language.

Sanjib Halder was awarded his 1st Master Degree (Master of Computer Application) from Indian Institute of Engineering Science and Technology, Shibpur, West Bengal, India and his Second Master Degree (M. E. in Software Engineering) from Jadavpur University, West Bengal, India. He worked in the Software industry for one year and then came to the academic world. He is teaching in the Computer Science Department of The Bhawanipur Education Society College for the last 19 years. His area of specialization is software Engineering and his research interest includes Software Engineering, mobile computing, System Security, etc.

Priyanka Banerjee was awarded B.Sc. degree in computer science(Hons) from Asutosh College and the B.Tech. and M. Tech. degrees from University of Calcutta. She is currently a Lecturer in the Department of Computer Science at The Bhawanipur Education Society College. Her research interests include machine learning, deep learning, computer vision, object detection, and semantic segmentation of image, etc.

Prof. (Dr.) **Samir Kumar Bandyo**pa**d**h**yay** is presently designated as Academic Advisor, The Bhawanipur Education Society College Kolkata. He is also Retired Senior Professor in the Department of Computer Science and Engineering of University of Calcutta. He published a number of papers in International Journals.

