Regression Data Analysis Approach On COVID-19 Prediction

Dr. S. Sathish[1]*, Dr. A. Jeeva[2]

[1]*School of Mathematics and Statistics, Assistant Professor, MIT World Peace University Pune-38, India-411038 Email: sathish.s@mitwpu.edu.in*
[2]*Department of Mathematics, Assistant Professor, Vel Tech Rangarajan Dr. Sagunthala R & D Institute of Science and Technology Chennai, India. Email:-drjeevaa@veltech.edu.in*

Corresponding Author: *Dr. S. Sathish*
School of Mathematics and Statistics, Assistant Professor, MIT World Peace University Pune-38, India-411038 Email: sathish.s@mitwpu.edu.in

Article History	Abstract
Received: *Revised:* *Accepted* **CC License** CC-BY-NC-SA 4.0	The Proposed method is to develop the regression models for the observed frequency distribution process and generate expected frequency distribution. This study analyzed the daily COVID 19 cases site, Regression models they are used to estimate daily confirmed, Death and New cases data of per day. The error estimates RMSE, MAE of forecasts from the above models is compared to identify the most suitable approaches for forecasting trend analysis. ***Keywords— Regression analysis , Forecast, COVID-19***

I. INTRODUCTION

The year 2020 has been a disastrous year for humankind. We humans, all around the globe have come across the Coronavirus. India witnessed an outbreak of COVID-19, during the last week of January 2020 when a few Indian students traveled to Kerala from Wuhan located in China. In 2020, from January to November to date, we have not been able to get rid of the virus. As per the World Health Organization (WHO), numerous potential COVID-19 antibodies are being examined, and many voluminous clinical trials may report their results later at the near end of 2020 or the very beginning of 2021 [1]. WHO is working with partners around the world to help coordinate with the key steps in this process. Companies such as Pfizer and BioNTech have concluded a phase 3 study of the COVID-19 vaccine and claim to be 95% efficient against the virus [2]. How the epidemic in India will top or decrease is foremost concerning the issue. Therefore, it is pivotal to predict the trends of the pandemic, nationwide. With this view of helping the Government, we undertook this research to aid them in making informed decisions about the spread of coronavirus thereby taking precautionary measures. For this, we have analyzed India's COVID-19 dataset using regression models with supporting empirical evidence including error analysis and accuracy juxtapositions. Also, we have forecasted the trend of coronavirus cases using the Time Series Forecasting approach of Tableau. These methods are applied for four different types of cases: Confirmed, Active, Cured, and Death Cases as available in [3].

II. METHODOLOGY

This section discusses the different methods applied to India's dataset for COVID-19 for the analysis, prediction, and forecasting of different cases. Fig. 1 shows the flowchart of our methodology which includes

data collection, followed by data preprocessing, data visualization, implementation of regression models, time series forecasting approach, and their results.

(a) Data Collection
The data for the ongoing Covid-19 outbreak in India is collected from [3]. The columns of this dataset include the Total number of Confirmed, Active, Cured, and Death cases of Covid-19 patients accumulating all the states, on a day-to-day basis from 12th March 2020 to 31st October 2020.

(b) Data Pre-processing
In the section of data preprocessing, redundant or null values were removed by data cleaning. Further, we have set the column attributes as - "Confirmed, Active, Cured and Deaths cases" for which the dependent variable is Y and "Months" as the independent variable X. To achieve this target, the data was then split into 75% for training purpose and 25% for testing purpose. Standardization of the variables pertinent to training and testing was done using the Standardizable() function and the fit_transform() function, the object was fit into data to transform these values into standard form.

(c) Data Visualization
Figures 2 and 3 represent Heat Maps of data and information to supply an open way to see and get trends, exceptions, and patterns in information. A Heat Map visualization could be a combination of colored rectangles, each representing a quality component that permits clients to rapidly get a handle on the state and effect of an expansive number of factors at one time. For example, Maharashtra state has the highest number of cases which is shown by the high intensity of red coloration, whereas for Lakshadweep state, color intensity is the least.
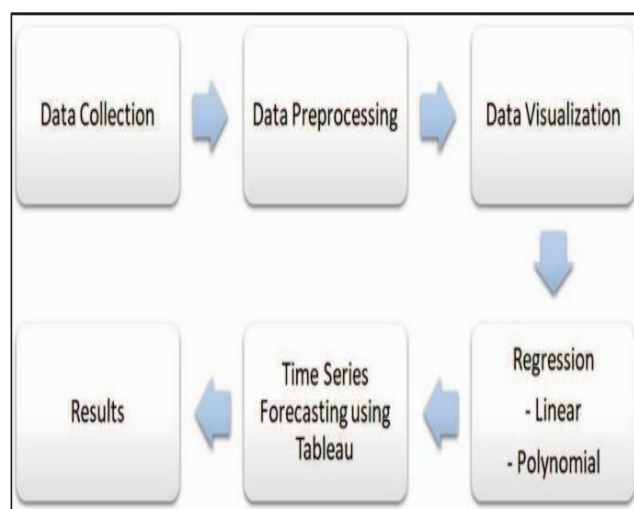
(d) Error Analysis
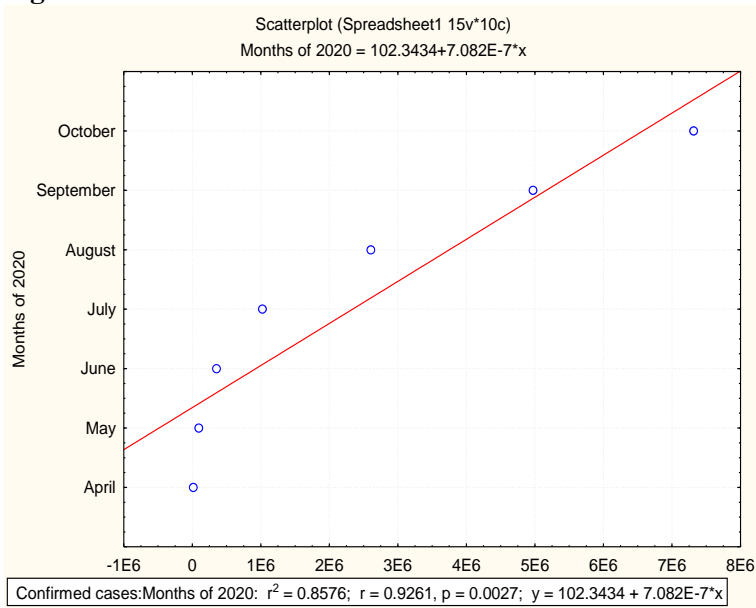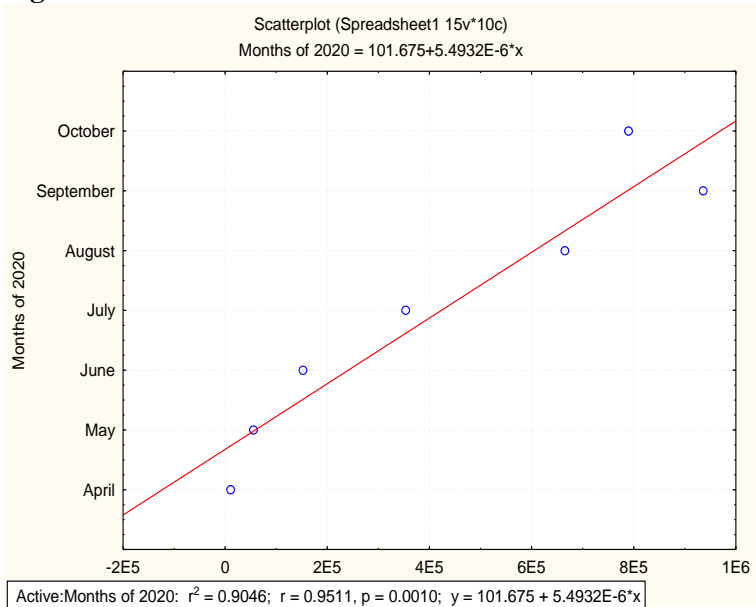Mean, Standard deviation, MAE, RMSE, MAPE and R^2

III. RESULTS AND DISCUSSIONS

(Whatever I collected from WHO, ICMR)

Months of 2020	Types of Cases			
	Confirmed	*Active*	*Cured*	*Death*
April	13788	10902	2451	435
May	94822	55396	36529	2896
June	351945	152324	188964	10656
July	1023435	353318	644520	25597
August	2604826	665008	1889706	50111
September	4970459	935915	3953097	81445
October	7315171	789945	6413690	1536

Fig. 1

Scatterplot (Spreadsheet1 15v*10c)
Months of 2020 = 102.3434+7.082E-7*x

Confirmed cases:Months of 2020: r^2 = 0.8576; r = 0.9261, p = 0.0027; y = 102.3434 + 7.082E-7*x

Fig 2: Observed and Predicted of confirmed cases

Scatterplot (Spreadsheet1 15v*10c)
Months of 2020 = 101.675+5.4932E-6*x

Active:Months of 2020: r^2 = 0.9046; r = 0.9511, p = 0.0010; y = 101.675 + 5.4932E-6*x

Fig 3: Observed and Predicted of Active cases

Scatterplot (Spreadsheet1 15v*10c)
Months of 2020 = 102.5079+7.9553E-7*x

Cured:Months of 2020: r^2 = 0.8173; r = 0.9041, p = 0.0052; y = 102.5079 + 7.9553E-7*x
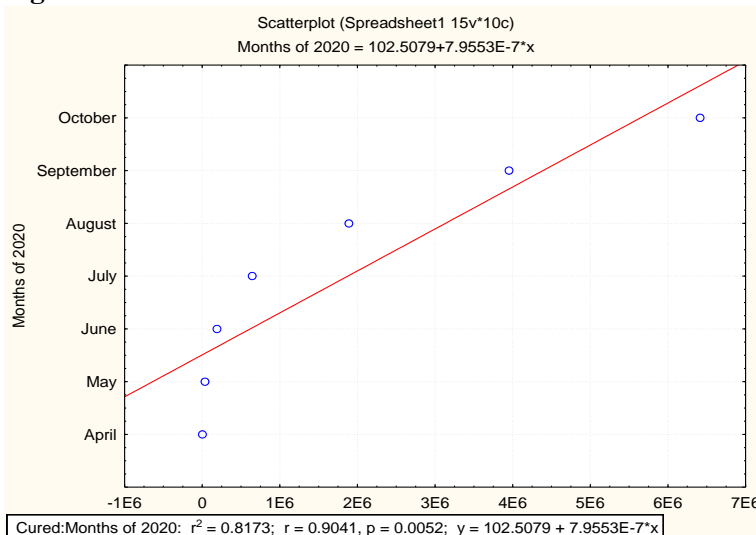
Fig 4: Observed and Predicted of Cured cases

IV. CONCLUSION

India's COVID-19 dataset is analyzed to mine the trends of the cases that can help the government and citizens to ensure safety by taking precautionary measures in the future. Analysis of dataset is done using linear and polynomial regressions which involved metrics like accuracy, R2 score, and MAPE. We can conclude that polynomial is better than linear regression. Forecasting is done using Tableau and the results are found to be satisfactory. However, the error rate in the future can be reduced by using a bigger dataset, betteralgorithms, and fine-tuning of the parameters.

REFERENCES

1. "World Health Organization," 7 April 1948. [Online]. Available: https://www.who.int. [Accessed 17 September 2020].
2. "Pfizer," Pfizer, 18 November 2020. [Online]. Available: https://www.pfizer.com/news/press-release/press-release-detail/pfizer- and-biontech-conclude-phase-3-study-covid-19-vaccine. [Accessed 19 November 2020].
3. PRS legislative research," 12 March 2020. [Online]. Available: https://prsindia.org/covid-19/cases. [Accessed 21 August 2020].
4. S. Tiwari, S. Kumar and K. Guleria, "Outbreak Trends of Coronavirus Disease–2019 in India: A Prediction," Disaster Medicine and Public Health Preparedness, vol. 14, no. 5, pp. 1-6, 2020.
5. P. Wang, X. Zheng, J. Li and B. Zhu, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," Chaos, Solitons and Fractals, vol. 139, no. 110058, pp. 1-7, 1 July 2020.
6. M. Yadav, M. Perumal and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," Chaos, Solitons and Fractals, vol. 139, no. 110050, pp. 1-12, 2020.
7. A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," Science of the Total Environment, vol. 728, no. 138762, pp. 1-6, 20 April 2020.
8. D. Parbat and M. Chakraborty, "A python based support vector regression model for prediction of COVID19 cases in India," Chaos, Solitons and Fractals, vol. 138, no. 109942, pp. 1-5, 2020.
9. S. Rath, A. Tripathy and A. R. Tripathy, "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 5, pp. 1467-1474, 2020.
10. R. Tamhane and S. Mulge, "Prediction of COVID-19 Outbreak using Machine Learning," International Research Journal of Engineering and Technology, vol. 7, no. 5, pp. 5699-5702, 2020.
11. A. Aggarwal, A. Rani, P. Sharma, M. Kumar, A. Shankar and M. Alazab, "Prediction of Landsliding using Univariate Forecasting Models," Internet Technology Letters, no. e209, pp. 1-6, 2020.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion and O. Grisel, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.2020. [Online]. Available: https://www.tableau.com/. [Accessed 10 September 2020].
13. P. Arora, H. Kumar and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," Chaos, Solitons and Fractals, vol. 139, no. 110017, pp. 1-9, 2020.