Regression Analysis for Identifying Factors Responsible for Total COVID-19 Cases

Anita Sahu[1], *Lokesh Parashar[2], Hina Arya[3], S. L. Vig[4], Jagdish Prasad[5], Shailendra Sahu[6]

[1]*Research Scholar, Department of Statistics, Amity School of Applied Sciences, Amity University, Jaipur, Rajasthan, India*
[2]*Research Scholar, Department of Statistics, Amity School of Applied Sciences, Amity University, Jaipur, Rajasthan, India*
[3]*Statistician cum Tutor, Department of Community Medicine, FMHS, SGT University, Gurugram, Haryana, India*
[4]*Assistant Professor, Department of Community Medicine, Employees' State Insurance Corporation Medical College and Hospital, Faridabad, Haryana, India*
[5] *Professor & Head, Department of Statistics, Amity School of Applied Sciences, Amity University, Jaipur, Rajasthan, India.*
[6]*Research Scholar, School of Information & Computer Sciences, University of Hyderabad, Hyderabad, India*

* **Corresponding author**
Mr. Lokesh Parashar
* Research Scholar,
Amity School of Applied Sciences,
Amity University,
Jaipur 303002, Rajasthan, India
Email: lokeshparashar3889@gmail.com

KEYWORDS
Model, regression analysis, Covid-19, Spread.

ABSTRACT

Purpose: The research conducts a comprehensive analysis using multiple linear regression to investigate the factors influencing the spread of COVID-19 in India.

Methods: Data related to COVID-19 such as total cases, total deaths, movements of cases(transportation), and population of all Indian States and Union Territories, along with the number of hospitals in these were collected. Rigorous data preprocessing, normalization, and addressing multicollinearity were performed.

Results: The multiple regression model identifies "Aircraft Movements" and "Public Hospitals" as significant predictors of "Total Cases", highlighting the impact of air travel and the number of public hospitals on transmission. "Outgoing Trains from States" positively associated with "Total Cases," indicating the role of transportation networks in the virus's spread. However, "Incoming Trains to States" and "Population" exhibited weak and insignificant relationships with "Total Cases."

Conclusions: The study emphasizes the importance of infrastructure, population density, and transportation in shaping the COVID-19 pandemic in India. Policymakers can use these findings to devise targeted measures, allocate resources, and bolster healthcare infrastructure in vulnerable regions. While the model shows moderate explanatory power, it acknowledges the influence of unmeasured factors, necessitating further research and advanced modeling techniques for enhanced accuracy and addressing limitations.

Implications: This research contributes to understanding of COVID-19's impact in India, enabling stakeholders to make informed decisions to mitigate the pandemic's effects and safeguard the population's well-being. Continued data-driven approaches and ongoing research remain vital in our efforts to combat COVID-19. The insights gained can aid in adapting public health strategies to address the unique challenges posed by the pandemic in the Indian context.

1. Introduction

Covid-19 was the biggest epidemic in several decades that the world suffered. Since the starting of the pandemic, researchers have analyzed several factors to analyze the spread pattern of COVID-19. Different studies conclude that there is no standard pattern of COVID-19 across the world, even no similarity in the States within the counties. In this, the research analyzed the factors that may play an important role in the spread of COVID-19 so as to help the policymakers to devise a strategy to check the spread of the pandemic.

2. Literature Review

The literature surrounding the COVID-19 pandemic has witnessed a proliferation of research endeavors aimed at comprehending the virus's dynamics, predicting its progression, and formulating effective mitigation strategies. The study by Kishor Kulkarni et al. [1] revolves around modeling the total cases of COVID-19 in India and assessing its socioeconomic ramifications. Employing a regression model, the authors establish robust correlations by considering variables such as daily new cases, total deaths, and daily new deaths. Besides, the study also delves into strategies for enhancing India's GDP amidst the pandemic, emphasizing fiscal mechanisms and sector-specific self-sufficiency. However, the study acknowledges its limitations, chiefly the reliance on linear relationships and its applicability confined to the Indian context. In a complementary study by Smita Rath et al. [2], the focus shifts to predicting the trajectory of active COVID-19 cases in Odisha and India through multiple linear regression models. The models exhibit a high prediction capability, with R2 scores approaching perfection, indicating their accuracy in forecasting forthcoming active cases. This study highlights the importance of employing multiple linear regression models to decipher the intricate relationships between dependent and independent variables while shedding light on the limitations inherent in linear regression models for COVID-19 prediction.

Ekta Gambhir et al. [3] pivot towards the integration of machine learning algorithms, specifically Polynomial Regression and Support Vector Machine (SVM), in their research. Notably, the Polynomial Regression model demonstrated an impressive accuracy rate of approximately 93% in forecasting the rise in COVID-19 cases. The study underscores the value of these models in comprehending the pandemic's dynamics. However, it does not delve into the detailed limitations or potential biases within the utilized datasets. In [4], Shruti Sharma et al. focus on building a reliable predictive model for COVID-19 cases, by assessing various models' performance, particularly highlighting the Adaptive gradient LSTM (AGLSTM) model.

A distinct approach is taken in the study by Deepmala et al. [5], which applies the Bayesian model to analyze and predict the spread of COVID-19 in Uttar Pradesh, India. This research considers confirmed, deceased, and recovered cases and employs logistic and Gompertz non-linear regression models within a Bayesian setup. The prior information is derived from advanced-stage outbreak States in India. The study aims to predict the evolution of confirmed, deceased, and recovered cases to enhance healthcare system preparedness. Notably, it does not account for changes in government control measures. Though the study acknowledges the significance of incorporating external information but does not elaborate on the specific limitations of the regression models utilized.

Raji and Deeba Lakxsmi [6] present a study that employs regression analysis techniques,

specifically linear and polynomial regression, to analyze the impact of COVID-19 in various States of India. Using a dataset from Kaggle, this research not only identifies the most affected States but also delves into forecasting the future outcomes of the pandemic, aiding in preparedness and resource allocation. Additionally, the paper emphasizes the role of contact tracing and responsible citizen behavior in controlling the spread of the virus. Ekta Rahangdale et al. [7] analyzed COVID-19 data in India using linear and polynomial regression algorithms. Focusing on data from March 1, 2021, to May 8, 2021, the study predicts the future outcomes of the pandemic in India and its top 4 States with the highest number of confirmed cases. This research underscores the potential of AI algorithms in understanding and combating the COVID-19 pandemic. Poonam Chauhan et al. [8] in its study applied linear and polynomial regression models to analyze the COVID-19 outbreak in India and its States. The study estimated case fatality rates (CFR) and recovery rates (RR) and utilized these metrics to identify the most affected States. Moreover, it employed polynomial regression to forecast COVID-19 cases, highlighting the importance of State-wise analysis and acknowledging the potential impact of government interventions and testing strategies. Sweeti Sah et al. [9] focused on analyzing the COVID-19 epidemic using linear and polynomial regression, considering geographic distribution and real-time datasets. The study aimed to predict the future reachability of COVID-19 across nations. This research highlighted the significance of the everyday exponential behavior of COVID-19 and emphasized the importance of accurate forecasting for policy and resource allocation.

Suganya et al. [10] proposed a COVID-19 forecasting model using Multivariate Linear Regression, aiming to predict confirmed and death cases for one and two weeks in top affected regions worldwide. The research achieved a high level of accuracy and highlighted the importance of accurate forecasting for preparedness and resource allocation.

P. Panda et al., in [11], examine the trend and pattern of COVID-19 deaths, aiming to discern the contributing factors to COVID-19 cases and fatalities. The study employs regres- sion analysis incorporating socio-economic variables. The findings reveal that the highest fatality burden resides in Southern Asia, with Western Asia bearing the greatest fatality burden concerning population density. Moreover, the study concludes that urbanization significantly correlates with the death burden, indicating that countries in the Asian region with larger urban populations tend to exhibit higher daily new death cases.

R. Anandan et al., in [12], investigate the COVID-19 pandemic's impact globally, emphasizing the lack of definitive treatment and the surge in fatalities. It explores the significance
of lockdown measures in India and the economic challenges the outbreak poses. Using publicly available datasets, the study utilizes visualization techniques and regression analysis to highlight vulnerable regions and forecast infection trends till March 27, 2020. The research employs Jupyter Notebook, Matplotlib for visualizations, and Time Series Analysis to predict COVID-19-related deaths in various states, primarily focusing on the 2020 mortality rates per 1000 individuals.

In [12], Rezaul Karim and Nazmin Akter examine the impact of climate factors on COVID-19 transmission using Bayesian regression analysis. It indicates that high temperatures potentially mitigate the virus's spread, while low temperatures contribute to increased cases and fatalities. This study suggests that warm and wet climates might play a role in reducing COVID-19 transmission, although further investigation into additional climate variables is recommended for a comprehensive understanding. In [13], the research conducted between January and September 2020 aimed to develop scalable analysis methods for observing and estimating COVID-19 spread worldwide. The study employs various mathematical models to identify outbreak levels across 187 countries and predicted epidemic trends, estimating an outbreak peak in 9–12 months. These findings offer crucial decision-making support for public health management in monitoring and promptly responding to high-risk regions.

In conclusion, the literature survey presents a diverse array of research contributions that collectively

enrich our understanding of COVID-19. These studies encompass modeling, prediction, and analysis efforts, shedding light on various facets of the pandemic and its repercussions. The findings serve as a valuable foundation for further investigations and potential policy interventions in the ongoing fight against COVID-19.

3. **Objective:** The objective of the study is to analyze the factors influencing the spread of Covid 19 in India using multiple linear regression and developing a model.
4. **Method**

4.1 Data Collection, Preparation and Analysis

A comprehensive data of total COVID-19 cases, total deaths, the count of public hospitals, incoming and outgoing train movements, aircraft movements, and population figures for all States and Union Territories in India has been sourced from https://data.gov.in/. and the population data is taken from the census 2011 (https://censusindia.gov.in/census.website). In the case of States and UTs that did not exist during that period, an aggregate of the populations of the constituent districts is taken to obtain accurate population estimates.

4.2 Data Normalization

Since all the variables considered are of different ranges, the data normalization by the Python sci-kit-learn library is done to ensure that all variables were scaled appropriately to enhance model performance and achieve more reliable results. [14].

4.3 Correlation

The analysis by Pearson's correlation using the IBM SPSS Statistics 25 tool was done to understand the relationships between the variables.

Table 1 shows the correlation between the variables.

Features	Total Cases	Total Deaths	Public Hospitals	Incoming Trains to States	Outgoing Trains from States	Aircraft Movements	Population
Total Cases	1.000	0.928	0.684	0.113	0.585	0.702	0.616
Total Deaths	0.928	1.000	0.560	0.014	0.640	0.684	0.513
Public Hospitals	0.684	0.560	1.000	0.515	0.595	0.502	0.923
Incoming Trains to States	0.113	0.014	0.515	1.000	0.181	0.013	0.725
Outgoing Trains from States	0.585	0.640	0.595	0.181	1.000	0.544	0.545
Aircraft Movements	0.702	0.684	0.502	0.013	0.544	1.000	0.426
Population	0.616	0.513	0.923	0.725	0.545	0.426	1.000

Table 1: Correlation Table

4.4 Multicollinearity

'Total Cases' were taken as the dependent variable while the independent variables considered were total deaths, the count of public hospitals, incoming and outgoing train movements, aircraft movements, and population. IBM SPSS was used to assess multicollinearity and ensure the robustness of the results. Table 2 shows the tolerance and Variance Inflation Factor (VIF) values. The tolerance and Variance Inflation Factor (VIF) values in Table 2 provide insights into the degree of multicollinearity among the independent variables. Multicollinearity occurs when two or more independent variables exhibit high correlation, which can lead to unreliable regression coefficients and hinder the interpretability of the model [15].

Based on the findings, it is observed that certain independent variables demonstrate high multicollinearity. Specifically, 'Public Hospitals' and 'Population' exhibit Tolerance values of 0.093 and 0.050, respectively, which are close to zero. Correspondingly, their VIF values of 10.734 and 20.114, respectively, exceed significantly the commonly accepted threshold of 5. These high VIF values suggest substantial multicollinearity and potential issues in the regression analysis.

Table 2: Collinearity between variables

Collinearity Statistics		
Variables	Tolerance	VIF
Total Deaths	0.327	3.060
Public Hospitals	0.093	10.734
Incoming Trains to the States	0.207	4.822
Outgoing Trains from the States	0.500	1.999
Aircraft Movements	0.492	2.032
Population	0.050	20.114

4.5. Variable Transformation

In response to the multicollinearity among certain independent variables, namely 'Public Hospitals' and 'Population,' variable transformation as a viable strategy was opted to address this issue thereby aiming to create new derived features that retain the valuable information embedded in these variables while reducing their correlation.

For the independent variable, 'Population,' a logarithmic transformation (the natural logarithm) was used to normalize the data and alleviate potential skewness [16, 17]. Additionally, to address potential issues with negative values, a condition was applied to ensure that only non-negative population values transform.

Similarly, for the other independent variable 'Public Hospitals,' a square root transformation was used to control the variable's scale and mitigate potential issues with negative values [17] with a condition statement to ensure that it only applies to non-negative 'Public Hospitals' values.

Further to ensure that the transformed variables are on a consistent scale, Min-Max scaling was used. This scaling technique brings the transformed variables within a common range of values, specifically between 0 and 1, thus facilitating a more robust analysis.

Variable transformation in this manner was employed with the aim of reducing the impact of multicollinearity on the regression analysis while preserving the essential aspects of the original data. This approach aligns with the research goal of maximizing the insights derived from the data while ensuring the reliability and interpretability of the model. Through these efforts, an effort was made to make a meaningful contribution to understanding the relationship between 'Total Cases' and the independent variables, bolstering the research

outcomes and their implications in our field of study.

Table 3 shows the collinearity statistics after variable transformation.

4.6 Regression

After dealing with the multicollinearity, the multiple linear regression analysis was performed to examine the relationship between 'Total Cases' and other independent variables, including 'Total Deaths,' 'Incoming Trains to States,' 'Outgoing Trains from States,' 'Aircraft Movements,' 'Population*,' and 'Public Hospitals*.' The regression model's performance.

Collinearity Statistics

Variables	Tolerance	VIF
Total Deaths	0.327	3.060
Public Hospitals*	0.254	3.937
Incoming Trains to the States	0.207	4.822
Outgoing Trains from the States	0.500	1.999
Aircraft Movements	0.492	2.032
Population*	0.355	2.814

Table 3: Collinearity between variables after variable transformation. *Variables after transformation are evaluated using various statistical metrics to assess their goodness-of-fit and explanatory power.

4.7 Model Summary

Table 4 shows the performance of the multiple linear regression model. The multiple linear regression model yielded promising results, as evidenced by an impressive R-squared value of 0.919. This indicates that the selected independent variables can explain approximately 91.9% of the variability in the 'Total Cases'. The Adjusted R-squared value of 0.902 reinforces the model's robustness, accounting for the number of independent variables and providing a more conservative estimate of the variance explained.

The model's coefficient of determination (R-squared) suggests a high level of fit, meaning that the chosen independent variables collectively contribute substantially to predicting the number of COVID-19 cases in the States and Union Territories of India. Additionally, the relatively low Standard Error of the Estimate (0.0626) indicates that the model's predictions will likely be close to the actual observed values.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959	.919	.902	.062642911492167

Table 4: Model Summary. Predictors: (Constant), Public Hospitals*, Incoming Trains to States, Total Deaths, Outgoing Trains from States, Aircraft Movements, Population*. Dependent Variable: Total Cases.

Table 5 presents the unstandardized coefficients, standardized coefficients (Beta), t- t-values, and significance levels for each independent variable in the model.

Among the independent variables, 'Total Deaths' shows a significant positive relationship with 'Total Cases,' as indicated by its Beta value of 0.831 ($p < 0.001$). The positive Beta value suggests that increased COVID-19 deaths are associated with a proportional increase in confirmed cases.

On the other hand, 'Outgoing Trains from States' and 'Aircraft Movements' exhibit negative Beta values of -0.144 and 0.067, respectively, albeit with varying significance levels. These variables may have a modest impact on 'Total Cases,' with 'Outgoing Trains from States' indicating a negative association.

Additionally, 'Population*' and 'Public Hospitals*' demonstrate positive Beta values of 0.027 and 0.258, respectively. The significance level of 'Public Hospitals*' at p = 0.020 suggests that the count of public hospitals, after the square root transformation, plays a role in influencing the number of COVID-19 cases.

Independent Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	-.028	.029		-.952	.349
Total Deaths	.957	.095	.831	10.083	.000
Incoming Trains to States	.002	.055	.002	.036	.972
Outgoing Trains from States	-.126	.066	-.144	-1.922	.064
Aircraft Movements	.057	.066	.067	.872	.390
Public Hospitals*	.185	.075	.258	2.458	.020
Population*	.021	.067	.027	.308	.760

Table 5: Coefficient Summary. Dependent variable: Total Cases

4.8 Rationale for Excluding 'Total Deaths' as an Independent Variable

Despite the high R-squared value and the significant coefficient of 'Total Deaths' in the multiple linear regression model, we made a deliberate decision to exclude 'Total Deaths' as an independent variable from our analysis. This subsection elucidates the reasoning behind this crucial step.

Endogeneity Concerns:. The primary concern that led to the exclusion of 'Total Deaths' as an independent variable is the potential issue of endogeneity. 'Total Deaths' can be considered an outcome of 'Total Cases,' as the number of deaths resulting from COVID-19 is naturally influenced by the number of confirmed cases. This inherent relationship creates a feedback loop between the two variables.

Reverse Causality. Introducing 'Total Deaths' as a predictor in the regression model can lead to reverse causality, where causality flows in the opposite direction than intended. In this context, while 'Total Cases' may influence the number of deaths, the reverse may also hold true – an increase in reported deaths may influence an increase in reported cases. Factors such as increased testing, improved healthcare interventions, or public health measures in response to higher death rates can contribute to changes in the number of reported cases.

Unreliable Coefficient Estimates. Due to endogeneity and reverse causality, the coefficient of 'Total Deaths' may be subject to bias and unreliability. The correlation between 'Total Deaths' and the error term in the regression equation can confound the coefficient estimate, making it difficult to discern the true impact of 'Total Deaths' on 'Total Cases.'

Focus on Predictive Power. While the coefficient of 'Total Deaths' may indeed be significant and positively associated with 'Total Cases,' our primary objective is predictive modeling rather than establishing a causal relationship. In predictive modeling, we aim to identify and leverage independent variables that are not influenced by the outcome variable to create a robust and reliable model for forecasting future 'Total Cases.'

Enhancing Model Interpretability. By removing 'Total Deaths' from the analysis, we focus on other independent variables that are not prone to endogeneity concerns. This enhances the interpretability of the model and allows us to gain more actionable insights into the various factors that impact the spread of COVID-19.

4.9 Regression Analysis Results after Removing 'Total Deaths'

An attempt to see the revised outcomes of the multiple linear regression anal- ysis, following the exclusion of 'Total Deaths' as an independent variable was made . The model's performance, ANOVA results, and coefficients for the remaining independent variables are assessed, shedding light on the factors that significantly impact the total number of COVID- 19 cases in India.

4.9.1 Model Summary

The revised multiple regression model demonstrates a moderate level of goodness-of- fit, as indicated by the R-squared value of 0.634. Approximately 63.4% of the variabil- ity in 'Total Cases' can be explained by the selected independent variables, which include 'Public Hospitals*,' 'Incoming Trains to States,' 'Aircraft Movements,' 'Outgoing Trains from States,' and 'Population*.'. Table 6 presents the model's performance after removing 'Total Death' variable.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.796	0.634	0.573	.130738

Table 6: Model Summary. Predictors: (Constant), Public Hospitals*, Incoming Trains to States, Outgoing Trains from States, Aircraft Movements, Population*. Dependent Variable: Total Cases.

The Adjusted R-squared value of 0.573 accounts for the number of independent variables and offers a more conservative estimate of the variance explained. The model's Standard Error of the Estimate (0.131) suggests that the predictions of 'Total Cases' are expected to deviate by approximately 13.1% from the actual values.

4.9.2 ANOVA

The ANOVA table[Table 7] provides insights into the overall significance of the regression model. The Regression Sum of Squares is 0.888, with 5 degrees of freedom (df), leading to a Mean Square value of 0.178. The F-statistic of 10.391 is highly significant ($p < 0.001$), indicating that the model as a whole explains a significant amount of the variance in 'Total Cases.'

	Sum of Squares	df	Mean Square	F	Sig.
Regression	.888	5	.178	10.391	.0001
Residual	.513	30	.017		
Total	1.401	35			

Table 7: ANOVA

4.10. Coefficients

Table 8 presents the unstandardized coefficients, standardized coefficients (Beta), t- values, and significance levels for each independent variable in the model.

Independent Variables	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
(Constant)	B	Std. Error	Beta	-.742	.464
	-.045	.061			
Incoming Trains to States	-.087	.114	-.098	-.761	.453
Outgoing Trains from States	.127	.127	.144	1.000	.325
Aircraft Movements	.355	.123	.416	2.899	.007

Log Population	-.036	.139	-.047	-.255	.800
Sqrt Public Hospitals	.317	.154	.444	2.056	.049

Table 8: Coefficient Summary after removing 'Total Death'. Dependent variable: Total Cases

Among the remaining independent variables, 'Aircraft Movements' and 'Public Hospitals*' exhibit the most substantial standardized coefficients (Beta values) of 0.416 and 0.444, re-respectively. A higher Beta value indicates a stronger influence of these variables on 'Total Cases.'

'Outgoing Trains from States' also displays a positive standardized coefficient (Beta value of 0.144), albeit with a significance level of 0.325. This variable may have a modest impact on 'Total Cases.'

On the other hand, 'Incoming Trains to States' and 'Population*' demonstrate negative standardized coefficients (Beta values of -0.098 and -0.047, respectively), suggesting a weak inverse relationship with 'Total Cases.'

5. Conclusion

In this research, an in-depth analysis is conducted to investigate the factors influencing the spread of COVID-19 in India using multiple linear regression modeling. Comprehensive data was collected on 'Total Cases,' 'Total Deaths,' 'Incoming Trains to States,' 'Outgoing Trains from States,' 'Aircraft Movements,' 'Population,' and 'Public Hospitals' for all states and Union Territories in India.

The primary objective was to identify key predictors of 'Total Cases' and provide valuable insights to inform public health policies and interventions in managing the pandemic. Throughout the analysis, rigorous data preprocessing, and normalization was done as well as multicollinearity concerns were handled to ensure the robustness of the results. After addressing high multicollinearity concerns, the 'Population' and 'Public Hospitals' variables were transformed into 'Log Population' and 'Sqrt Public Hospitals' using variable transformation techniques. After considering potential endogeneity issues, a critical decision was made to remove 'Total Deaths' as an independent variable. Although 'Total Deaths' showed a significant association with 'Total Cases,' the risk of reverse causality was recognized, and it was focused on independent variables less influenced by the outcome.

The multiple regression model revealed 'Aircraft Movements' and 'Sqrt Public Hospitals' as substantial predictors of 'Total Cases,' implying that air travel and the number of public hospitals have a significant impact on the pandemic's transmission. Additionally,' Outgoing Trains from States' displayed a positive association with 'Total Cases,' suggesting the importance of transportation networks in the spread of the virus.

On the other hand, 'Incoming Trains to States' and 'Log Population' showed weak and in-significant relationships with 'Total Cases,' indicating limited predictive power in explaining case counts.

The findings of the present study underscore the significance of infrastructure, population density, and transportation in the dynamics of the COVID-19 pandemic in India. Policymakers can leverage these insights to strategize targeted measures, allocate resources, and strengthen healthcare infrastructure in regions more vulnerable to the virus's transmission.

Despite the model's moderate explanatory power, it is acknowledged that other unmeasured factors might influence the pandemic's spread. Further research and more sophisticated modeling techniques may enhance predictive accuracy and address these limitations.

In conclusion, this research contributes to the growing body of knowledge regarding the COVID-19 pandemic's impact in India. By understanding the interplay between various factors and case counts, it is aimed to aid stakeholders in making informed decisions to mitigate the pandemic's effects and ensure the well-being of the population.

6. Discussion

Understanding the spread of infection is very important so as to mitigate it from being epidemic or pandemic. Though the virulence of the agent, mode of transmission from host to other and environmental factors need to be assessed from an epidemiological point of view, the other factors particularly in COVID-19 like population movement and the robustness of public health system in terms of availability accessibility cannot be underestimated. This study has shown a strong association with movement for the spread. Any public health program should also emphasis these factors.

References

[1] K. Kulkarni, A. Kulkarni, N. S. Shaikh, S. Sayyed, Modeling of total cases due to covid-19 and its impact in india, Journal of The Institution of Engineers (India): Series B (2021) 1–12.

[2] S. Rath, A. Tripathy, 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 14 (5) (2020) 1467–1474.

[3] E. Gambhir, R. Jain, A. Gupta, U. Tomer, Regression analysis of covid-19 using machine learning algorithms, in: 2020 International conference on smart electronics and communication (ICOSEC), IEEE, 2020, pp. 65–71.

[4] S. Sharma, Y. K. Gupta, A. K. Mishra, Analysis and prediction of covid-19 multivariate data using deep ensemble learning methods, International Journal of Environmental Research and Public Health 20 (11) (2023) 5943.

[5] Deepmala, N. K. Srivastava, S. K. Singh, U. Singh, Analysis and prediction of covid-19 spreading through bayesian modelling with a case study of uttar pradesh, india, OPSEARCH (2022) 1–16.

[6] P. Raji, G. Deeba Lakshmi, Covid-19 pandemic analysis using regression, medRxiv (2020) 2020–10.

[7] Ekta, S. Rahangdale, S. More, G. Narnaware, S. Sahu, Gujar, Covid 19 data analysis in india using linear and polynomial regression algorithms, International Journal for Research in Applied Science & Engineering Technology (IJRASET) 10 (3) (2022).

[8] P. Chauhan, A. Kumar, P. Jamdagni, Regression analysis of covid-19 spread in india and its different states, medRxiv (2020) 2020–05.

[9] S. Sah, S. Kanmani, A. Kamerkar, B. Surendiran, R. Dhanalakshmi, Covid-19 epidemic analysis using linear and polynomial regression approach, Applications of Artificial Intelligence in COVID-19 (2021) 245–255.

[10] R. Suganya, R. Arunadevi, S. M. Buhari, Covid-19 forecasting using multivariate linear regression (2020).

[11] P. K. Panda, R. S. Varkey, P. Ranjan, A. K. Meher, S. Panda, Covid 19 fatalities burden in asian countries: An analysis of pattern and determinants, Social Sciences & Humanities Open 7 (1) (2023) 100378.

[12] R. Anandan, T. Nalini, S. Chiwhane, M. Shanmuganathan, P. Radhakrishnan, Covid-19 outbreak data analysis and prediction, Measurement: Sensors 25 (2023) 100585.

[13] X. Zhou, X. Ma, S. Gao, Y. Ma, J. Gao, H. Jiang, W. Zhu, N. Hong, Y. Long, L. Su, Measuring the worldwide spread of covid-19 using a comprehensive modeling method, BMC Medical Informatics and Decision Making 21 (Suppl 9) (2023) 384.

[14] O. Kramer, O. Kramer, Scikit-learn, Machine learning for evolution strategies (2016) 45–53.

[15] J. I. Daoud, Multicollinearity and regression analysis, in: Journal of Physics: Conference Series, Vol. 949, IOP Publishing, 2017, p. 012009.

[16] K. Benoit, Linear regression models with logarithmic transformations, London School of Economics, London 22 (1) (2011) 23–36.

[17] L. Fang, Y. Hong, Uncertain revised regression analysis with responses of logarithmic, square root and reciprocal transformations, Soft Computing 24 (2020) 2655–2670.