International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; SP-9(3): 01-07

https://www.mathsjournal.com
Received: 04-02-2024
Accepted: 13-03-2024

Richa Dhurandher
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

B Devi Priyanka
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

AK Singh
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

ML Lakhera
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

AK Gauraha
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

Deepak Sharma
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

Corresponding Author:
AK Singh
Indira Gandhi Krishi Vishwavidyalaya (IGKV), Raipur, Chhattisgarh, India

Estimating yield and production of paddy and maize crops of Bastar district of Chhattisgarh using regression analysis

Richa Dhurandher, B Devi Priyanka, AK Singh, ML Lakhera, AK Gauraha and Deepak Sharma

Abstract
This study focused on estimating yield and production of paddy and maize crops of Bastar district of Chhattisgarh State. Since paddy and maize are the most important cereals among others in Bastar, these crops are selected for the study. 10 years data during the period of 2010-11 to 2019-20 are collected for the variables area, yield, production, proportionate gross irrigated area, proportionate total npk consumption, rainfall, relative humidity, maximum temperature and minimum temperature. ARIMA (p, d, q) parameters for the residuals of their linear model are calculated. From the ARIMA (p, d, q) parameters, it is understood that there is no auto-correlation for the data in paddy crop, while auto-correlation is observed in the data for the variables of maize crop. As a result, for the paddy crop, linear regression model was finalized whereas for the maize crop, generalized least square modelling was fitted. Then forecasting for the yield upto 2025 was done for the three selected crops of Bastar.

Keywords: ARIMA (p, d, q), linear models, generalised linear models, multicollinearity

Introduction
Regression analysis is an average measure of linear relationship between two or more variables. The regression model with one dependent variable and many independent variables is called multiple regression analysis (Draper and Smith 1985) [13]. Wei and Molin (2020) [42] estimated yield of soyaben using linear regression approach. Nazir *et al*., (2021) studied on estimating and forecasting rice yield using phenology based algorithm and linear regression model. Sellam and Poovammal (2016) [27] studied on prediction of crop yield using regression analysis. Shastry (2017) [28] studied on prediction of crop using regression analysis.
If the residuals of a time series model might exhibit serial correlation or heteroscedasticity, violating the assumptions of classical regression models like ordinary least square. When this happens, one approach to address these issues is by using GLS. Otto *et al*., (1987) [43] explained GLS approach to maximize likelihood estimation of regression models with ARIMA errors. Kadanali *et al*., (2019) [44] studied ARIMA model for forecasting wheat production in wheat. Dritsakis *et al*., (2019) [45] explained time series analysis using ARIMA models. Kwasi *et al*., compared the forecasting power of multivariate VAR and univariate ARIMA models.
Since there is multicollinearity among the dependent variables, high multi-collinear variables are eliminated from the model using the correlation matrix to increase its accuracy. How to identify multi-collinearity in regression analysis was described by Shrestha (2020). Haitovsky (1969) [30] and Daoud (2017) [10] conducted research on multi - collinearity in regression analysis. Multi-collinearity and misleading outcomes were clarified by Kim (2019) [18]. Schroeder *et al*. conducted research on multi-collinearity diagnosis and treatment. Regression and ARIMA models can be connected to take use of each method's advantages and increase forecasting accuracy, particularly in situations where the data show both linear correlations and temporal dependencies. Shumway *et al*. (2017) [31] provided an explanation of time series analysis and explored ARIMA models. The usefulness of time series modeling (ARIMA) in stock price forecasting was investigated by Mondal *et al*. in 2014 [46]. Benvenuto *et al*.'s (2020) [9] research examined the use of the ARIMA model with the COVID -19 pandemic dataset.

Material and Methods Study area

This study has been conducted in Bastar district of Chhattisgarh for the crops paddy and maize. It is located on the latitude of 19° 12' 0" N and longitude of 81° 56' 0" E. It is bounded on the northwest by Narayanpur District, on the north by Kondagaon district, on the east by Nabarangpur and Koraput Districts of Odisha State, on the south and southwest by Dantewada and Sukma.

Data Description

Secondary data provide the sole basis of this investigation. In 2000, the State of Chhattisgarh was divided into two halves from Madhya Pradesh. Since then, Chhattisgarh has formed a large number of new districts by the bifurcation or trifurcation of the large districts three times: in 2007–2008, in 2011–12, and in 2018–19. Bastar is considered as a combined district in this study.

The variables that included are area (x1), gross irrigation area (x2), total npk consumption (x3), maximum temperature (x4), minimum temperature (x5), relative humidity (x6) and rainfall (x7). Yield (y), production (p).

Outliers were detected and removed in order to maintain the accuracy of the model. The parameters of the auto-regressive integrated moving average, or ARIMA (p, d, q), were calculated in order to ascertain whether serial correlation existed or not for the time series data of the paddy, maize and jowar crop of Bastar. Using a correlation matrix, the predictor variables that had a strong connection with the dependent variable were chosen in descending order of magnitude. Linear model and generalised linear models are used wherever applied.

Linear Regression

Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable y and one or more independent variables $X1$, $X2$, .., Xn. It assumes that this relationship is linear, meaning that a change in the independent variable(s) is associated with a constant change in the dependent variable.

The general form of a linear regression model with *n* predictors is

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 \ldots \beta_n \times X_n + \in$$

Where Y is the dependent variable, $X1$, $X2$, .., Xn are the independent variables, $\beta0$ is the intercept, $\beta1$, $\beta2$, .., $\beta n$are the coefficients (slopes) of the independent variables, ∈ is the error term, which represents the difference between the observed and predicted values of Y.

Generalised least square model

Generalised least square linear model is a technique for estimating the unknown parameters in a linear regression model when the assumption for the residuals of the linear regression model fails, either in terms of zero mean, normality or equality of variances apart from the independence of different residuals.

GLS tackles these problems by introducing a weighting matrix. This matrix assigns different weights to different data points based on their estimated variance. Here's how it works:

Estimate the error covariance matrix: This matrix captures how the errors are related to each other. It might show higher variances for certain groups of data points.

Apply the weights: Each data point in the regression equation is multiplied by a weight based on the error covariance matrix. This gives more influence to data points with lower error variance and less influence to those with higher variance.

Minimize the weighted squared residuals: Similar to OLS, GLS minimizes the sum of squared residuals, but these residuals are now weighted based on the error covariance matrix.

Multi-collinearity

When two or more independent variables (predictor variables) in a regression analysis have a strong correlation with one another, this is referred to as multicollinearity. When analyzing the regression model's findings, this correlation may cause issues. It is challenging to discern each variable's unique impact on the dependent variable when there is a significant degree of correlation between the independent variables. This makes it difficult to determine the actual influence of each element on the result. Consequently, when the model struggles t o discern the underlying impacts of the associated variables, its predictions may become less accurate.

Auto Regressive Integrated Moving Average (ARIMA)

ARIMA stands for Auto Regressive Integrated Moving Average. It is a widely used statistical method for time series forecasting and analysis. ARIMA models are capable of capturing a wide range of temporal patterns in data, making them useful for various applications, including economics, finance, epidemiology, and weather forecasting.

The general ARIMA (p, d, q) model can be expressed with the following equation

$$Y_t = \mu + \varphi_1 Y_{t_{-1}} + \varphi_2 Y_{t_{-2}} + \cdots + \varphi_p Y_{t_{-p}} + \epsilon t - \theta_1 \varepsilon_{t_{-1}} - \theta_2 \varepsilon_{t_{-2}} - \cdots - \theta_q \varepsilon_{t_{-q}}$$

Where, Yt is the actual value at time t, μ is the constant mean (optional, may not be present in all models), $\varphi1$ (phi) are the autoregressive parameters ($\varphi1$ to φp). These represent the coefficients of the past p values of Yt, $Yt-1$ to $Yt-p$ are the lagged values of Yt, influencing the current value (Yt), ϵt is the white noise error term at time t (represents unpredictable random shocks), θ (theta) are the moving average parameters ($\theta1$ to θq). These represent the coefficients of the past q forecast errors ($\varepsilon t-1$ to $\varepsilon t-q$).

Model evaluation

After putting all of the estimate methods through diagnostic plots and goodness of fit metrics like R2, Adj R2, and their P-values, the process is concluded.

Results and Discussions

Paddy crop of Bastar

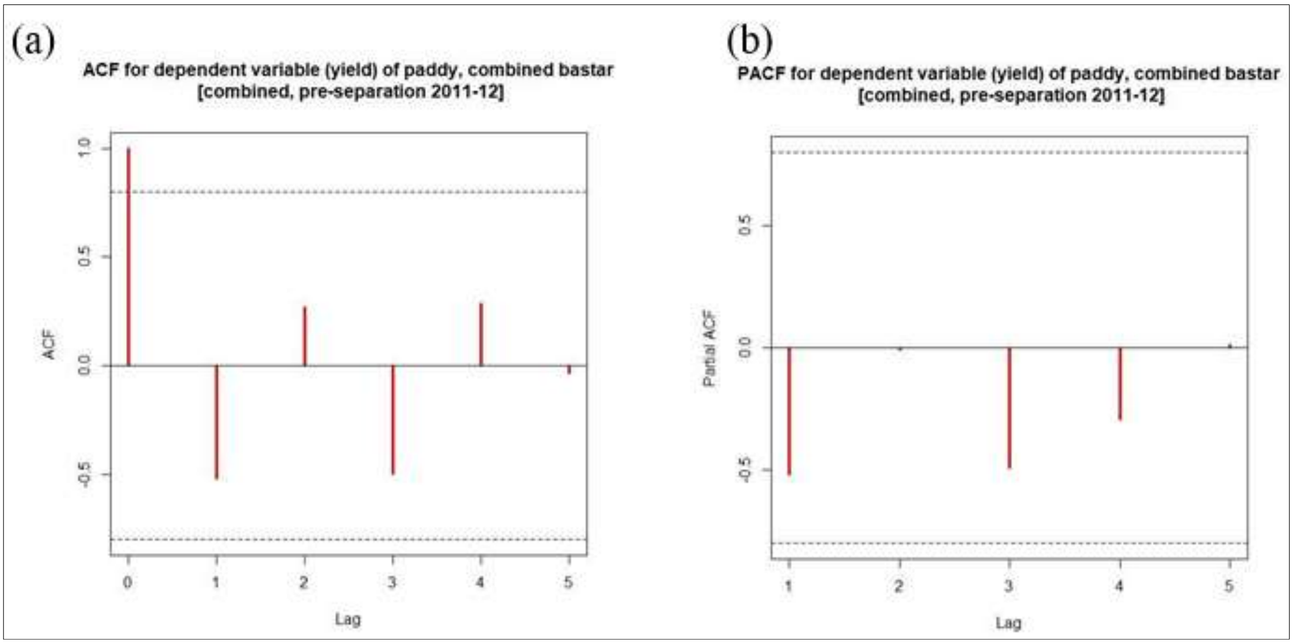

Fig 1: ACF and PACF for dependent variable (yield) of paddy, Bastar

For the time series data of Bastar's paddy yield, the ARIMA (0, 2, 0) model was produced in order to determine and eliminate any auto-correlation (Figure 1). As a result, the linear statistical model was employed for the estimate and prediction of paddy crop yield and production rather than the generalized least square model. The order of the correlations of magnitudes for the variables, x3, x4, x5, and x6, was used to choose which variables to include in the linear statistical model based on the correlation matrix.

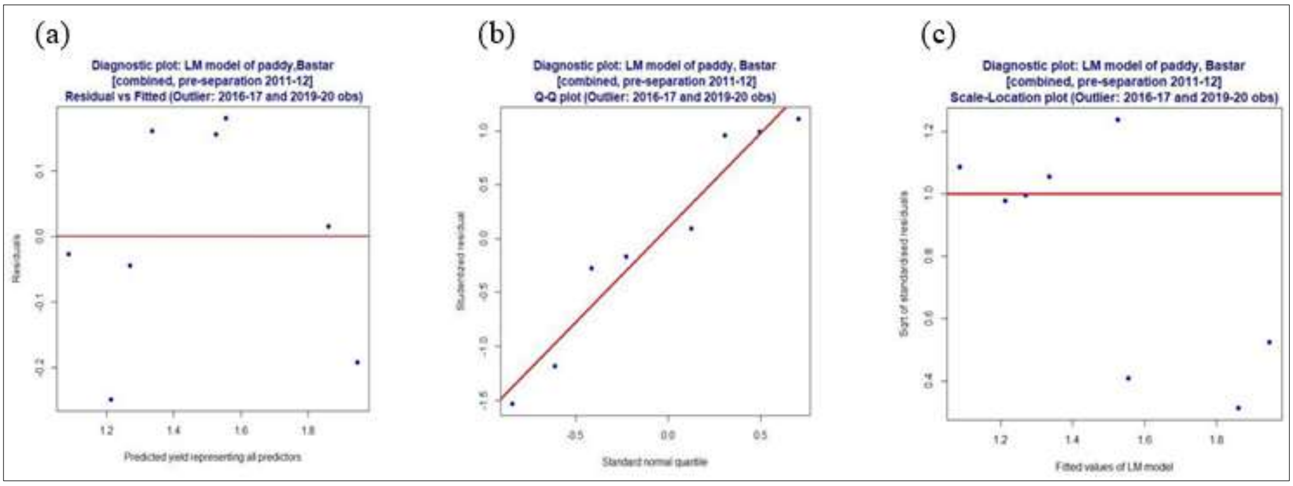

Fig 2: Three major diagnostic plots for the linear model of paddy, Bastar

The diagnostic plots indicating a good fitting model with R2 = 0.7823 and Adj R2 = 0.4921 and the finalized linear statistical model is given below.

$$y = 56 + 0.00136 \times x_3 - 1.74 \times x_4 + 1.33 \times x_5 - 0.423 \times x_6$$

Further, the yield, predicted yield, standard error and confidence interval of the finalized model are given in Table 1.

Table 1: Comparison of yield observed with the predicted one along with confidence interval for paddy, Bastar

Year	Yield	Predicted yield	Standard error	Confidence interval (95%)	
				Lower	Upper
2010-11	1.68	1.52	0.25	0.85	2.20
2011-12	1.06	1.08	0.25	0.41	1.76
2012-13	1.75	2.00	0.25	1.27	2.62
2013-14	1.73	1.55	0.25	0.87	2.23
2014-15	1.87	1.86	0.25	1.18	2.53
2015-16	0.96	1.21	0.25	0.54	1.90
2017-18	1.22	1.27	0.25	0.60	1.95
2018-19	1.50	1.33	0.25	0.66	2.01

Moreover, goodness of fit plot between predicted yield and yield is further done and depicted in Figure 3.

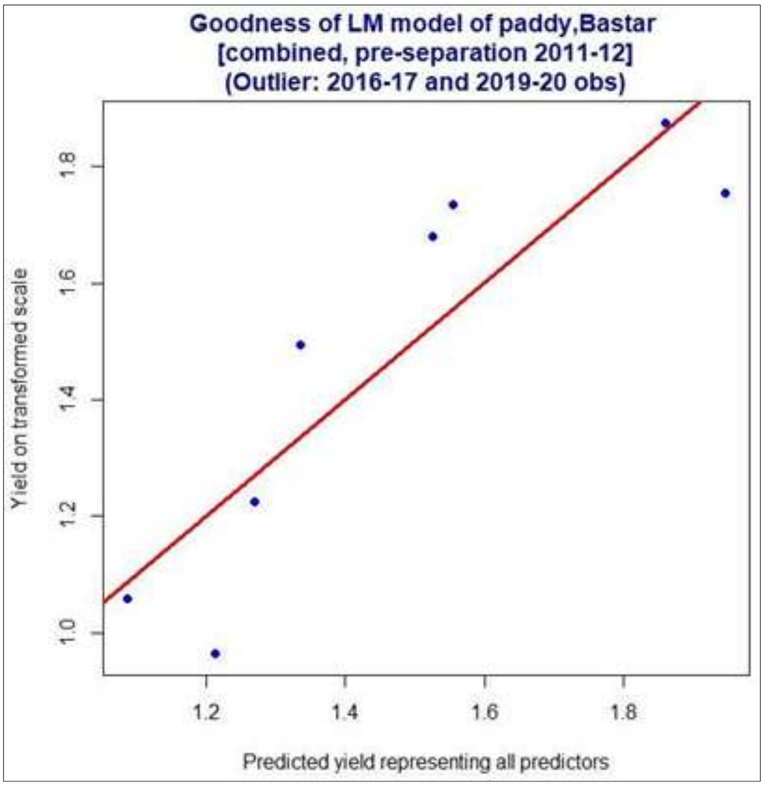

Fig 3: Goodness of the prediction model in terms of predicted vs observed yield for paddy, Bastar

After fitting and finalizing the model, a forecast for the yield and production has been made upto 2025.

Table 2: Forecast of yield and production for 2020-21 to 2024-25 based on the projected area under paddy, Bastar

Year	Projected area (hectare)	Predicted yield (tonnes per hectare)	Predicted production (tonnes)
2020-21	235984	2.03	480331
2021-22	235916	2.16	496843
2022-23	235848	2.18	513346
2023-24	235780	2.25	529840
2024-25	235711	2.32	546323

Maize crop of Bastar

From the time series data of the yield of the maize crop in Bastar, autoregressive integrated moving average, ARIMA (2, 2, 0) are determined to confirm the presence or absence of auto- correlation.

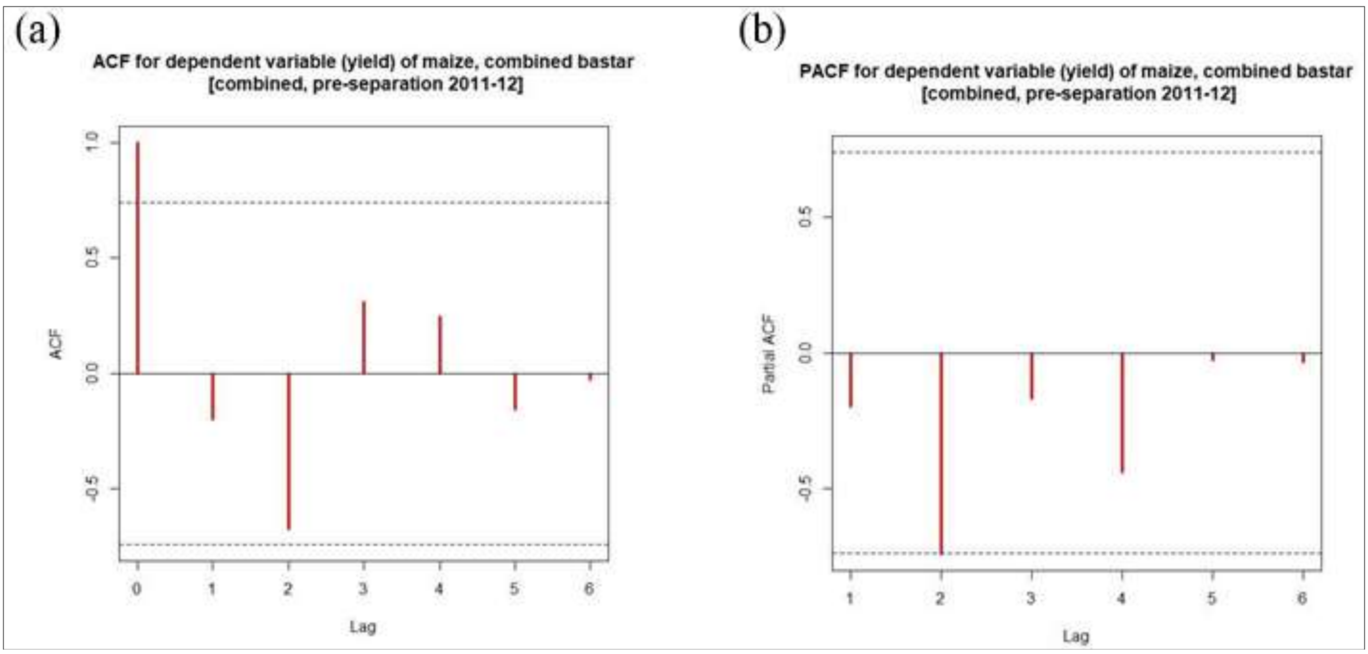

Fig 4: ACF and PACF for dependent variable (yield) of maize, Bastar

The values of p=2 and q=0 from ARIMA indicating the presence of serial correlation. As a result, generalised least square is used instead of linear model. Further, to know the order of predictor variables for entering into the generalised least square model, correlation matrix was used. From the correlation matrix, the variables having highest correlation with dependent variable were selected in descending order. The order of the variables is as follows: x1, x3, x2, x4, x5 and x7.

The generalized linear model for the maize crop is given below.

$$y = 2.01 + 0.236 \times x1 - 0.0034 \times x3 + 1.27 \times x2 - 0.84 \times x4 + 1.09 \times x5 + 2.40 \times x7$$

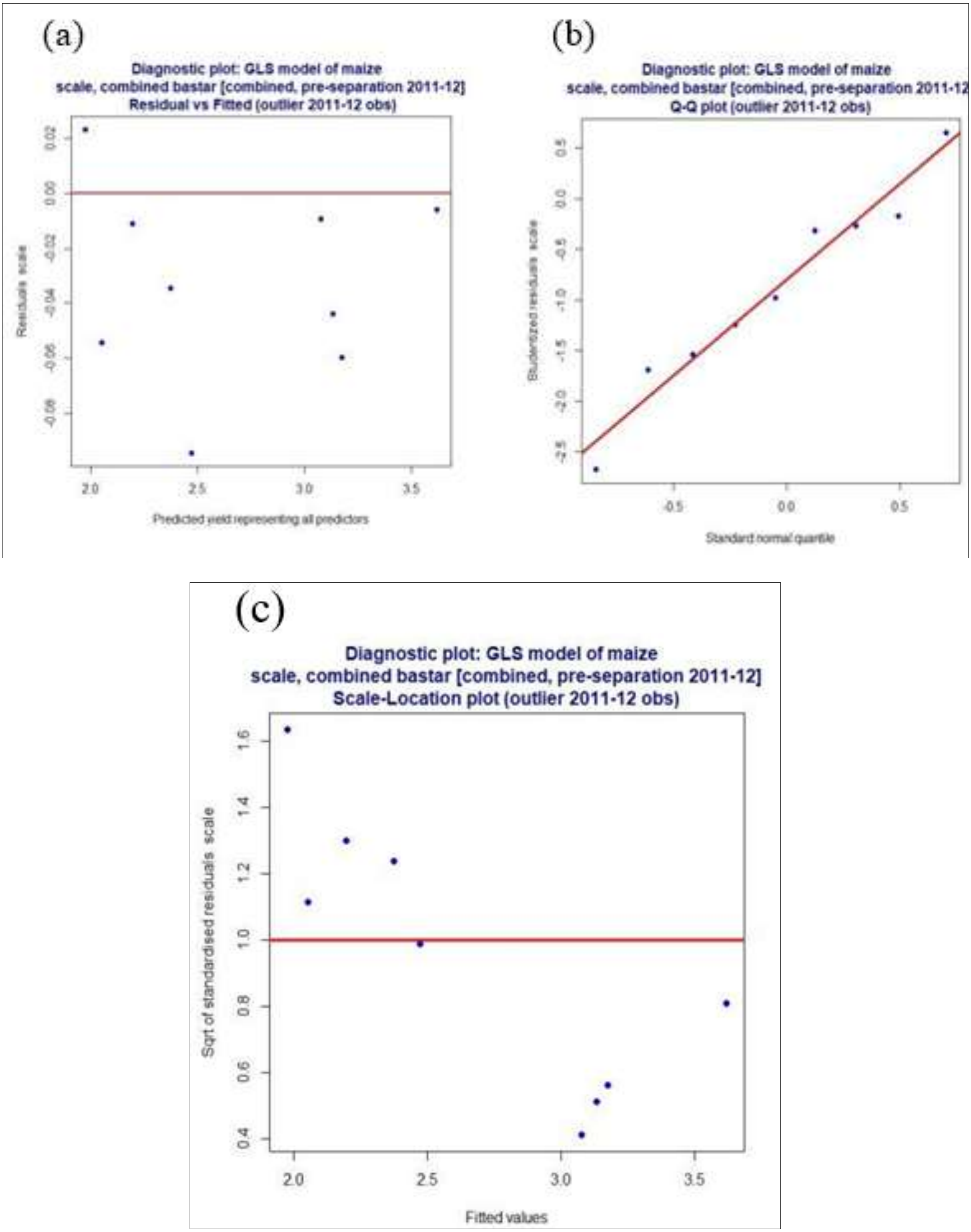

Fig 5: Three major diagnostic plots for the linear model of maize, Bastar

Since the diagnostic plots were seemed to fit well, the generalised least square model was finalized. Further, the yield, predicted yield, standard error and confidence interval of the finalized model are given in Table 3.

Table 3: Comparison of yield observed with the predicted one along with confidence interval for maize, Bastar

Year	Yield	Predicted yield	Standard error	Confidence interval (95%)	
				Lower	Upper
2010-11	2.00	1.97	0.17	1.28	2.67
2012-13	2.18	2.20	0.17	1.50	2.90
2013-14	2.34	2.38	0.17	1.68	0.36
2014-15	2.00	2.05	0.17	1.26	3.74
2015-16	2.38	2.47	0.17	2.78	1.36
2016-17	3.11	3.17	0.17	2.40	3.80
2017-18	3.09	3.13	0.17	2.40	3.80
2018-19	3.07	3.07	0.17	2.78	3.77
2019-20	3.61	3.62	0.17	0.66	2.01

Moreover, goodness of fit plot between predicted yield and yield is further done and depicted in Fig. 6. It is very clear that the data points are close to the fitted line indicating a good fit with R2 = 0.9964 (P-value: $8.285e-10$) and Adj R2 = 0.9959

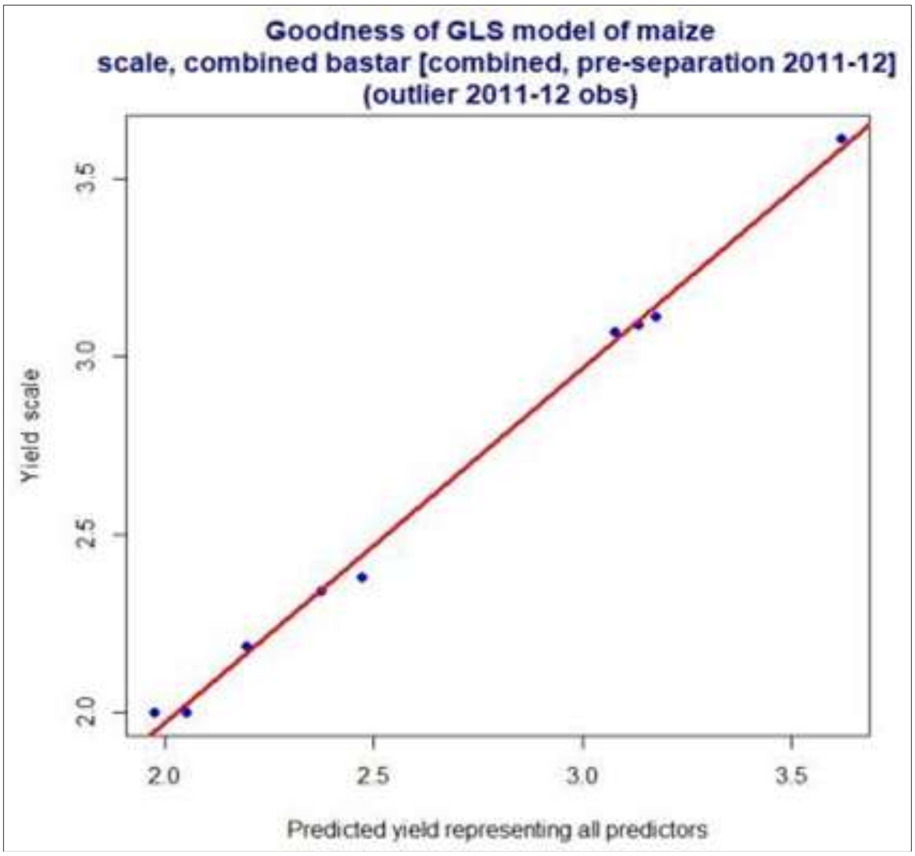

Fig 6: Goodness of the prediction model in terms of predicted vs observed yield for maize, Bastar

A forecast upto 2025 has been made for yield and production after finalizing the model.

Table 4: Forecast of yield and production for 2020-21 to 2024-25 based on the projected area under maize, Bastar

Year	Projected area (hectare)	Predicted yield (tonnes per hectare)	Predicted production (tonnes)
2020-21	27886.8	3.627899	101170.5
2021-22	28619.05	3.816518	109225.1
2022-23	29351.31	4.005137	117556
2023-24	30083.56	4.193756	126163.1
2024-25	30815.82	4.382375	135046.5

Conclusion

In this study, predicted yield is very much close to the actual yield indicating that the finalized model is a good fit. Through forecasting of yield of paddy and maize upto 2025, there is showing an increasing yield in future which is a good index.

References

1. Anonymous. List of Districts of Chhattisgarh [Internet]. Wikipedia. Available from: https://en.wikipedia.org/wiki/List_of_districts_of_Chhattisgarh.

2. Arunachalam R, Balakrishnan V. Statistical modeling for wheat (*Triticum aestivum*) crop production. Int. J Stat Appl. 2012;2(4):40-46.
3. Bahrami M, Shabani A, Mahmoudi MR, Didari S. Determination of effective weather parameters on rainfed wheat yield using backward multiple linear regressions based on relative importance metrics. Complexity; c2020. p. 1-10.
4. Barnett V, Lewis T. Outliers in statistical data. New York: Wiley; c1994.
5. Bartlett MS. The use of transformations. Biometrics. 1947;3(1):39-52.
6. Belov AG. A mathematical-statistics approach to the least squares method. Comput. Math Model. 2018;29:30-41.
7. Chatfield C. Exploratory data analysis. Eur. J Oper. Res. 1986;23(1):5-13.
8. Chaudhary JL. Rainfall-rice yield relationship-a case study for Chhattisgarh region of Madhya Pradesh in central India.
9. Ciotti M, Angeletti S, Minieri M, Giovannetti M, Benvenuto D, Pascarella S, *et al*. COVID-19 outbreak: An overview. Chemotherapy. 2020;64(5-6):215-223.
10. Daoud JI. Multicollinearity and regression analysis. J Phys Conf Ser. 2017;949(1):012009.
11. Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Govt. of Chhattisgarh. Area, Production and Productivity of Various Districts of Chhattisgarh for the Years 2010-11 to 2019-20. National Informatics Centre. Website: http://aps.dac.gov.in/APY/Public_Report1.aspx. 2023.
12. Diwan UK, Puranik HV, Das GK, Chaudhary JL. Yield prediction of wheat at preharvest stage using regression based statistical model for 8 district of Chhattisgarh, India. Int. J Curr. Microbiol. Appl. Sci. 2018;7(1):2180-2183.
13. Draper NR, Smith H. Applied regression analysis. John Wiley & Sons; c1998.
14. Ekanayake EMP, Wickramasinghe LCD, Weliwatta RT. Use of regression techniques for rice yield estimation in the North-Western province of Sri Lanka. Ceylon J Sci. 2021;50(4):439-447.
15. Ferdushi KF, Hossain MK, Kamil AA. Production Risk with Feasible Generalized Least Square. J Phys. Conf. Ser. 2020;1641(1):012109.
16. Ferguson IS, Leech JW. Generalized least squares estimation of yield functions. For Sci. 1978;24(1):27-42.
17. Haitovsky Y. Multicollinearity in regression analysis: Comment. Rev Econ Stat; c1969. p. 486-489.
18. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, *et al*. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102-D1109.
19. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. John Wiley & Sons; c2021.
20. Nasa Prediction of Worldwide Energy Resources. The Power Project. Website: https://power.larc.nasa.gov/. 2023.
21. Nazir A, Ullah S, Saqib ZA, Abbas A, Ali A, Iqbal MS, *et al*. Estimation and forecasting of rice yield using phenology-based algorithm and linear regression model on sentinel-ii satellite data. Agriculture. 2021;11(10):1026.
22. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; c2023.
23. Rajarathinam A, Parmar RS, Vaishnav PR. Estimating Models for Area, Production and Productivity Trends of Tobacco (*Nicotiana tabacum*) Crop for Anand Region of Gujarat State. Indian J Appl Sci. 2010;10:2419-2425.
24. Ran Y, Chen H, Ruan D, Liu H, Wang S, Tang S, *et al*. Identification of Factors Affecting Paddy Yield Gap in Southwest China: An experimental study. Plos One, 2018, 13(11).
25. Rossi R, Murari A, Gaudio P, Gelfusa M. Upgrading Model Selection Criteria with Goodness of Fit Tests for Practical Applications. Entropy (Basel). 2020;22(4):447.
26. Sahu G, Nishad D, Lakhera ML, Mishra P, Joshi RP. Influence of Area and Yield on the Production of Maize in Chhattisgarh Plain. J Pharmacogn. Phytochem. 2018;1:71-75.
27. Sellam V, Poovammal E. Prediction of crop yield using regression analysis. Indian J Sci. Technol; c2016.
28. Shastry A, Sanjay HA, Bhanusree E. Prediction of crop yield using regression techniques. Int. J Soft Comput. 2017;12(2):96-102.
29. Shirazi SM, Yusop Z, Zardari NH, Ismail Z. Effect of Irrigation Regimes and Nitrogen Levels on the Growth and Yield of Wheat. Hindawi. Publ. Corp Adv. Agric.; c2014.
30. Shrestha N. Detecting multicollinearity in regression analysis. Am J Appl. Math Stat. 2020;8(2):39-42.
31. Shumway R, Stoffer D. Time series: A data analysis approach using R. Chapman and Hall/CRC; c2019.
32. Singh DP, Kumar D, Paikra MS, Kusro PS. Developing Statistical Models to Study the Growth Rates of Paddy Crops in Different Districts of Chhattisgarh. Am Int. J Res Formal Appl. Nat Sci. 2014;5(1):102-104.
33. Singh R, Brahme R, Singh VB. Growth in Area, Production, and Productivity of Kharif Paddy in Chhattisgarh. Int. J Agric. Sci. Res. 2018;8(4):65-72.
39. Usman A, Tukur K, Suleiman A, Abdulkadir A, Ibrahim H. The Use of the Weighted Least Squares Method When the Error Variance is Heteroscedastic. Benin J Stat. 2019;2:85-93.
40. Uyanık GK, Güler N. A study on multiple linear regression analysis. Procedia. Soc. Behav. Sci. 2013;106:234-240.
41. Wasnik S, Pandey S, Patel P, Choudary V. Trends in Area, Production and Yield of Paddy, Wheat and Gram in Chhattisgarh State: A Critical Analysis. Pharma Innov. J. 2022;11(6):1397-1402.
42. Wei MCF, Molin JP. Soybean yield estimation and its components: A linear regression approach. Agriculture. 2020;10(8):348.
43. Maroni G, Wise J, Young JE, Otto E. Metallothionein gene duplications and metal tolerance in natural populations of Drosophila melanogaster. Genetics. 1987 Dec 1;117(4):739-744.
44. Kadanali E, Yazgan Ş, Yalçinkaya Ö. Arima Model for Forecasting Wheat Production in Turkey. In 4th International Conference on Advances in Natural & Applied Sciences; c2019 Jun 19. p. 743.
45. Stamatiou P, Dritsakis N. Causality among CO_2 emissions, energy consumption and economic growth in Italy. International Journal of Computational Economics and Econometrics. 2019;9(4):268-286.
46. Mondal SP, Yamage M. A retrospective study on the epidemiology of anthrax, foot and mouth disease, haemorrhagic septicaemia, peste des petits ruminants and rabies in Bangladesh, 2010-2012. PloS one. 2014 Aug 7;9(8):e104435.