E-ISSN: 2618-0618
P-ISSN: 2618-060X

www.agronomyjournals.com
2024; SP-7(8): 315-321
Received: 08-05-2024
Accepted: 13-06-2024

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

Linear statistical forecast model based on agronomic-meteorological variables for productivity and production of paddy crop in Raipur district of Chhattisgarh

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

DOI: https://doi.org/10.33545/2618060X.2024.v7.i8Se.1264

Abstract
Estimation of linear statistical forecast model for productivity and production of paddy crop for Raipur district of Chhattisgarh plains, based on major agronomic-meteorological predictors was carried out in this study. Statistical summaries of the variables along with some important graphical explorations were conducted based on 10 years data points from 2010-11 to 2019-20. An upward slope has been observed in area, yield, production, proportionate gross irrigated area, maximum temperature and minimum temperature, whereas proportionate total NPK consumption, proportionate rainfall and relative humidity were found to be decreasing gradually for paddy crop of Raipur district. Based on the ARIMA (p, d, q) parameters, estimation regarding detection of serial correlation, linear statistical model was obtained for the paddy after box-cox transformation, to ensure normality. The analysis was carried out using "R" software (R core team 2023).

Keywords: Kharif paddy, ARIMA (p, d, q), linear statistical model, box-cox transformation, multi-collinearity

1. Introduction
Chhattisgarh, known as rice bowl of India, is the 26th state of the India (Gregory 2013) [11]. It came into existence on November 1, 2000. As a result of its diverse soil types, varied climatic conditions, mountains, plateaus, rivers, native plants, and forests, this State has a great potential for agriculture. In this State, paddy occupies nearly 36 lakh acres of land during the *Kharif* season, making up roughly 77% of the total area sown.

Since paddy is the major crop in Raipur, yield forecast is necessary and useful for future predictions. 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) [7]. When there is a specific level of correlation between the regression model's residuals, the generalised least squares (GLS) method is employed to estimate the unknown parameters. Ordinary least squares and weighted least squares may need to be more statistically effective in these situations to avoid producing false conclusions.

Several attempts have been undertaken by various researchers to estimate and predict the crop yield and production at the national and state levels. Sellem *et al*., (2016) [25] studied on estimating crop yield through linear regression analysis. Belov (2018) [2] explored a mathematical-statistical method for estimating a linear multiple regression model's least-square parameter. Ekanayake *et al*. (2021) [8] used regression techniques for estimating yield of maize in the north-west parts of Sri Lanka. Matthew and Chris Chatfield explained on exploratory data analysis individually. Rajan provided the use of regression models for area, production and productivity growth trends of cotton crop in India. Rajarathinam studied on estimating models for area, production and productivity trends of tobacco for Anand region of Gujarat State.

When the dependent variable is not normally distributed, Box-cox transformation is used to stabilize the variance and/or make the data more normally distributed. Osborne (2010) [18] applied box-cox transformations to ensure normality in dependent variable. Studied on box-cox transformations and reviewed its applications. Kim and Hill (1993) [15] explained the box-cox transformations in regression analysis. Sakia (1992) [22] demonstrated the box-cox techniques.

Since multi-collinearity exists between the dependent variables, high multi-collinear variables are removed using correlation matrix in order to make the model more accurate. Shrestha (2020) [26] explained how to detect multi-collinearity in Regression analysis. Daoud (2017) [5] studied on multi-collinearity in regression analysis. Kim (2019) [14] explained multi-collinearity and misleading results. Schroeder *et al*., studied on diagnosing and dealing with multi-collinearity.

By connecting regression and ARIMA models, we can leverage the strengths of each approach to improve forecasting accuracy, especially when the data exhibit both linear relationships and temporal dependencies. Shumway *et al*., (2017) [27] studied on ARIMA models and explained time series analysis. Mondal *et al*., (2014) [16] studied on effectiveness of time series modeling (ARIMA) in forecasting stock prices. Benvenuto *et al*., (2020) studied on application of ARIMA model on the covid-2019 epidemic dataset.

As no study has been done on estimating and predicting the models of productivity and production of paddy crop in Raipur district of Chhattisgarh through linear regression model, this study has been undertaken.

2. Material and Methods

2.1 Study area

The study has been conducted in paddy crop for Raipur district of Chhattisgarh. District Raipur extends from latitude 21° 23″ to longitude 81° 65″. Paddy, Soybean, Urd and Arhar are the major Kharif Crops while Rabi season is mainly led by Chickpea and Lathyrus.

2.2 Data description

This study is entirely based on secondary data. The State of Chhattisgarh was bifurcated from Madhya Pradesh in 2000. Since then many new districts were created in Chhattisgarh by bifurcating or trifurcating the big districts, three times in 2007-08, 2011-12 and in 2018-19. Raipur is considered as a combined district at the time 2010-11, where after the present districts of Balodabazar and Gariyaband were separated from it in 2011-12. The data set for a period of 10 years are taken for the period 2010-11 to 2019-20 from the websites of Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Govt. of Chhattisgarh, Department of Agriculture, Government of Chhattisgarh, Indirawati Bhavan, New Raipur, Chhattisgarh etc.

The independent variables that included are area (x_1), gross irrigation area (x_2), total npk consumption (x_3), maximum temperature (x_4), minimum temperature (x_5), relative humidity (x_6) and rainfall (x_7). Yield (y) is the dependent variable. Production is represented as “p”. For the variables under study, mean and standard errors are calculated.

An outlier of 2016-17 observation is found and hence removed. To ensure presence or absence of serial correlation for the time series data of yield of paddy crop in Raipur, auto-regressive integrated moving average parameters, ARIMA (p, d, q) were determined. The predictor variables that were highly correlated with the dependent variable were selected in descending order of magnitude using correlation matrix. As there was no serial correlation, a linear statistical model was used instead of generalised least square model. Box-cox transformation was used to ensure normality.

2.3 Linear Regression

Linear 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) [7]. The fitted multiple regression model is given below:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \in \quad \ldots(1)$$

Where, $E(\in) = 0, V(\in) = V\sigma^2$ and $\in \sim N(0, V\sigma^2)$.

The error-variance-covariance matrix structure can be given as

$$V\sigma^2 = \sigma^2 \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \sigma^2 I \quad \ldots(2)$$

2.4 Box-cox transformation

The Box-Cox transformation is a technique used in statistics to transform a non-normal dependent variable (usually the target variable in a regression analysis) into a more normal distribution. Many statistical tests, like linear regression, rely on the assumption that the errors are normally distributed. This normality assumption allows us to make inferences about the population from the sample data. If the data isn't normal, the results of these tests might be unreliable.

The Box-Cox transformation applies a power function to the data, with a parameter lambda (λ) that can take on different values. By choosing the right lambda, the transformation can significantly reduce skewness and make the distribution of the data more closely resemble a normal bell curve. The formula for the Box-Cox transformation involves taking the natural logarithm of the variable and then raising it to the power of lambda. The lambda value is chosen through a process called maximum likelihood estimation, which basically finds the value that best fits the data to a normal distribution.

$$V = \begin{cases} \dfrac{(Y^{\lambda} - 1)}{\lambda \dot{Y}^{\lambda - 1}}, & for\ \lambda \neq 0 \\ \dot{Y} \ln Y, & for\ \lambda = 0 \end{cases} \quad \ldots(3)$$

Where, $\dot{Y}$ is geometric mean of Y.

2.5 Multi-collinearity

Multicollinearity refers to a situation in regression analysis where two or more independent variables (predictor variables) are highly correlated with each other. This correlation can create problems when interpreting the results of the regression model. When independent variables are highly correlated, it becomes difficult to isolate the individual effect of each variable on the dependent variable. This makes it challenging to understand how each factor truly influences the outcome. As a result, the model's predictions may become less accurate, as the model struggles to distinguish the true effects of the correlated variables.

Examining the correlation coefficients between all independent variables. High correlations (close to 1 or -1) suggest potential collinearity. If a variable can be reasonably explained by other variables in the model, consider removing it.

2.6 Auto Regressive Integrated Moving Average (ARIMA)

ARIMA, which stands for Autoregressive Integrated Moving Average, is a powerful statistical model used for analyzing and forecasting time series data. It's a popular technique in various fields like finance, economics, and supply chain management. ARIMA is a combination of three models:

1. Autoregressive (AR): This component predicts future values based on a specific number of past values (lags) in the time series.
2. Integrated (I): Sometimes, time series data might exhibit trends or seasonality that make it non-stationary (meaning the statistical properties change over time). The integration step involves differencing the data (taking the difference between consecutive observations) to remove trends and stationaries the data.
3. Moving Average (MA): The moving average component considers the average of past forecast errors (the difference between predicted and actual values) to account for randomness and improve the accuracy of the model.

The ARIMA Model Notation

ARIMA models are represented by ARIMA (p, d, q), where, 'p' represents the number of autoregressive terms (past values included in the model), 'd' represents the degree of differencing needed to stationaries the data and 'q' represents the number of moving average terms (past errors included in the model)

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

$$Y_t = \mu + \varphi_1 Y_{t_{-1}} + \varphi_2 Y_{t_{-2}} + \cdots + \varphi_p Y_{t_{-p}} + \epsilon t - \theta_1 \varepsilon_{t_{-1}} - \theta_2 \varepsilon_{t_{-2}} - \cdots - \theta_q \varepsilon_{t_{-q}} \quad \ldots(4)$$

Where, Y_t is the actual value at time t, μ is the constant mean (optional, may not be present in all models), φ_1 (phi) are the autoregressive parameters (φ_1 to φ_p). These represent the coefficients of the past p values of Y_t, $Y_{t_{-1}}$ to $Y_{t_{-p}}$ are the lagged values of Y_t, influencing the current value (Y_t), ϵt is the white noise error term at time t (represents unpredictable random shocks), θ (theta) are the moving average parameters (θ_1 to θ_q). These represent the coefficients of the past q forecast errors ($\varepsilon_{t_{-1}}$ to $\varepsilon_{t_{-q}}$).

The ARIMA model can be used to predict future values in the time series after establishing the proper values for p, d, and q (often by a procedure known as model fitting). The model makes predictions based on historical data, trends, seasonality (where applicable), and random mistakes.

2.7 Model Evaluation

The entire estimation processes is finalized after passing them through diagnostic plots and goodness of fit measures like R^2, Adj R^2 and their P-values.

2.7.1 Coefficient of Determination

The coefficient of determination, also known as R-squared (denoted by R^2 or r^2), is a statistical measure used in regression analysis to assess how well a regression model fits the data. It essentially tells how much of the variation in the dependent variable can be explained by the independent variable(s) in the model. The formula for coefficient of determination is

$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} \quad \ldots(6)$$

Where, Σ represents the sum of squares (often calculated using summation notation), y is the actual value of the dependent variable, ŷ is the predicted value of the dependent variable from the model, ȳ is the mean of the dependent variable (y).

2.7.2 Adjusted R^2

Adjusted R-squared penalizes the addition of unnecessary predictors by taking into account the number of predictors in the model and the sample size. It provides a more accurate assessment of the goodness-of-fit of a regression model, particularly when comparing models with different numbers of predictors. The formula for adjusted R-squared is

$$Adjusted\ R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{n-k-1}\right) \quad \ldots(7)$$

Where, R^2 is the R-squared value, *n* is the sample size, *k* is the number of predictors in the model.

After the model was finalized, yield and production were forecasted for the next five years i.e. from 2020-21 to 2024-25 using projected areas for these years.

3. Results and Discussion

The statistical summaries like arithmetic mean and standard error for all the variables are calculated (Table 1). Similarly, graphical representations for distributions of different variables were represented through barplots (Figure 1). The results of the barplots of all the variables revealed that there were increasing trends in x1, y, p, x2, x4 and x5, whereas x3, x7 and x6 were decreasing gradually during this study period 2010-11 to 2019-20.

Table 1: Statistical summaries for all the variables of the paddy crop, Raipur

S. No.	Variable	Arithmetic mean	Standard error
1	Area (hectare)	543543	3038
2	Yield (tonnes per hectare)	1.7	0.131
3	Production (tonnes)	918742	68909
4	Proportionate gross Irrigated area (hectare)	62434	171
5	Proportionate total npk consumption (metric tonnes)	84721	14428
6	Proportionate rainfall (millimetre per year)	1444	65.5
7	Relative humidiy (percent)	83.28	0.55
8	Maximum temperature (degree Celsius)	30.35	0.43
9	Minimum temperature (degree Celsius)	23.41	0.3

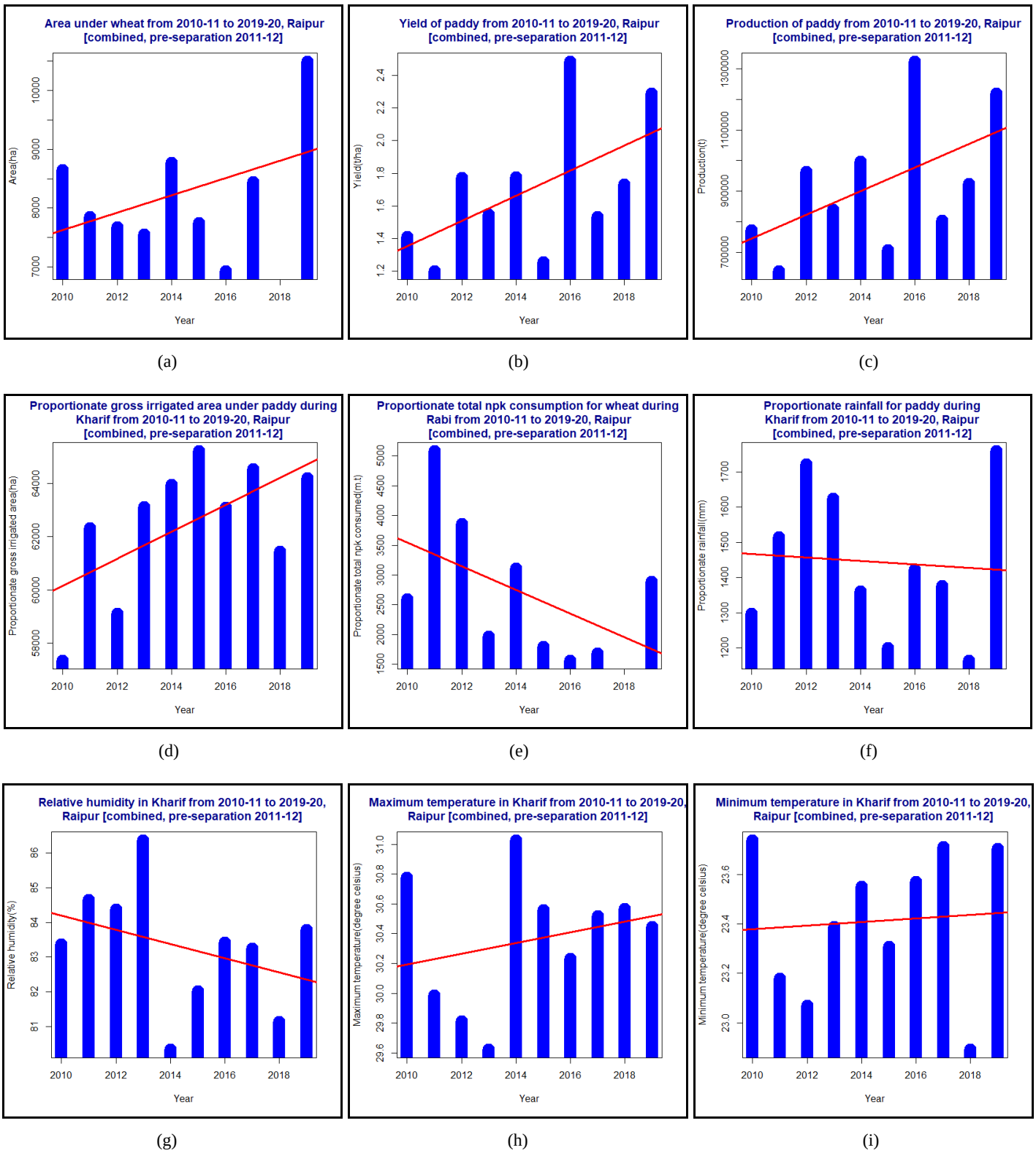
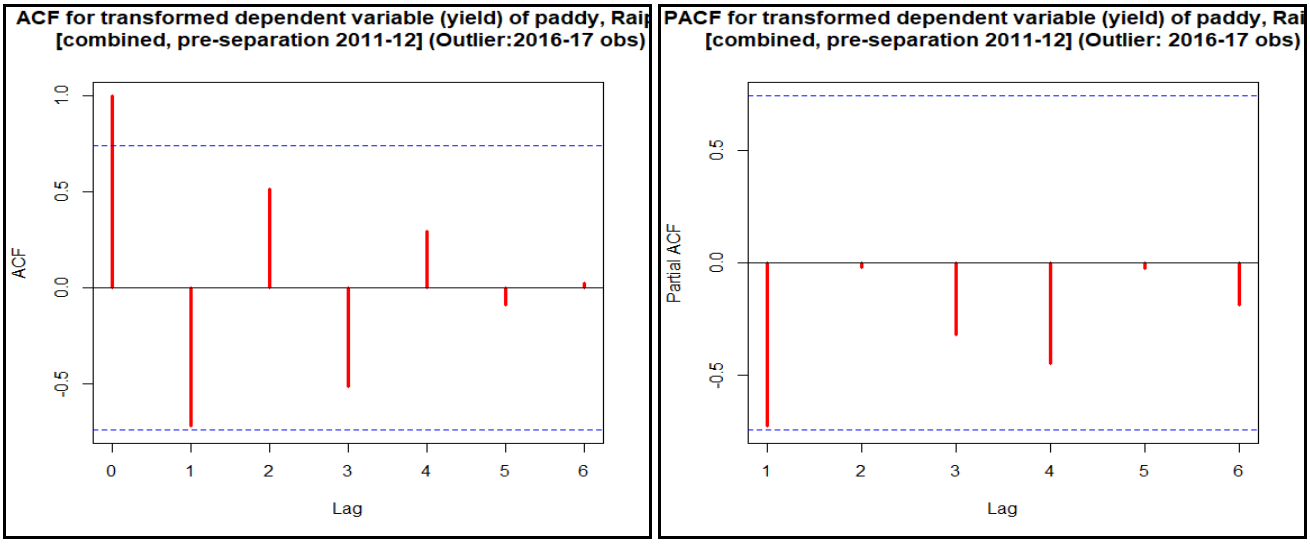


Fig 1: Barplot for variables for paddy crop of Raipur

To ascertain and remove auto-correlation, if any, the ARIMA (0, 2, 0) model was obtained for the time series data of yield of paddy of Raipur (Figure 2).



(a) (b)

Fig 2: ACF and PACF for transformed dependent variable (yield) of paddy, Raipur

As a result, instead of generalised least square model, the linear statistical model was used for estimation and prediction of yield and production of paddy crop.

From the correlation matrix, the variables to enter into the linear statistical model were determined based on the order of the correlations of magnitudes for variables was x3, x1, x7 and x5 as given in Table 2.

Table 2: Correlation matrix among the yield contributing variables for the paddy crop of Raipur [combined, pre-separation 2011-12]

	x_1	**y**	**x_3**	**x_2**	**x_4**	**x_5**	**x_7**	**x_6**
x_1	1	0.369	-0.013	0.093	0.608	0.219	-0.174	-0.604
y	0.369	1	-0.382	0.132	0.02	0.261	0.352	-0.066
x_3	-0.01	-0.38	1	-0.2	-0.33	0.074	0.043	0.46
x_2	0.093	0.132	-0.202	1	0.102	0.179	-0.042	-0.179
x_4	0.608	0.02	-0.33	0.102	1	0.383	-0.631	-0.862
x_5	0.219	0.261	0.074	0.179	0.383	1	0.116	0.057
x_7	-0.17	0.352	0.043	-0.04	-0.63	0.116	1	0.69
x_6	-0.6	-0.07	0.46	-0.18	-0.86	0.057	0.69	1

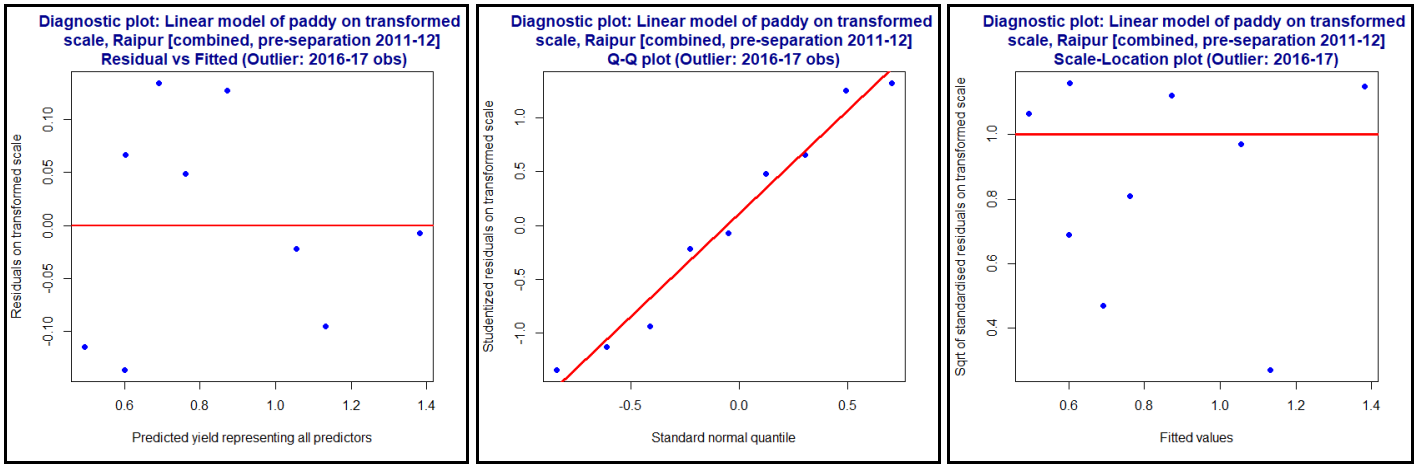
The estimation of linear model was done by using R. The diagnostics plots are depicted in Figure 3. The linear statistical model that was finalized is given below:

$$y = -9.3 - 1.1e^{-5} \times x3 + 1.95e^{-5} \times x1 + 8.3e^{-4} \times x7 - 2.4e^{-4} \times x5 \quad \ldots (3)$$

All the diagnostic plots, as in Fig. 3, indicate that the linear statistical model, as in Equation (3) above, is a good fitting model, which is further verified from the goodness of fit measures like R^2= 0.8933 (P-value= 0.0001206), Adj R^2 = 0.7866 and goodness of fit plot (Fig. 4). Therefore, based on this model the predicted values along with the confidence intervals for the model have been computed and are given in Table 3.

Table 3: Comparison of yield observed with the predicted one along with confidence interval for paddy, Raipur [Combined, pre-separation 2011-12]

Year	**Yield**	**Predicted yield**	**Standard error**	**Confidence interval (95%)**	
				Lower	**Upper**
2010-11	1.41	1.165	0.143	-0.51	2.84
2011-12	1.203	1.131	0.143	-0.544	2.805
2012-13	1.773	1.342	0.143	-0.333	3.016
2013-14	1.547	1.195	0.143	-0.48	2.869
2014-15	1.777	1.378	0.143	-0.296	3.053
2015-16	1.257	1.164	0.143	-0.511	2.839
2017-18	1.533	1.22	0.143	-0.455	2.895
2018-19	1.733	1.263	0.143	-0.412	2.938
2019-20	2.287	1.517	0.143	-0.158	3.191



(a) (b) (c)

Fig 3: Three major diagnostic plots for the linear model of paddy, Raipur

After fitting the final model, a future forecast of 5 years has been made, ahead of the last data points used for estimating the model, i.e. from year 2020 to 2024. This forecast is given below in Table 4 along with the forecast of production corresponding to the projected areas under the crop for same years, because we did not have the observed data for them.

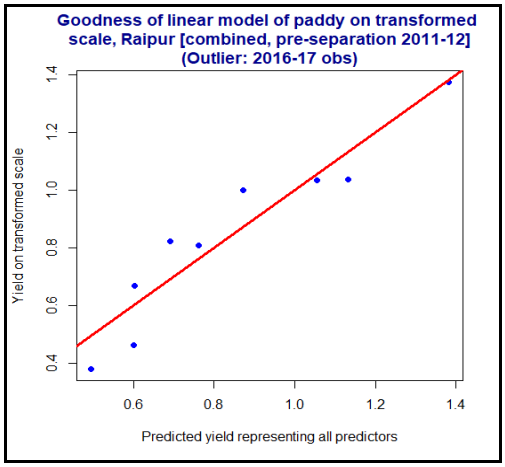


Fig 4: Goodness of the prediction model in terms of predicted vs observed yield for paddy on transformed scale

Table 4: Forecast of yield and production for 2020-21 to 2024-25 based on the projected area under paddy, Raipur [combined, pre-separation 2011-12]

Year	Projected area (hectare)	Predicted yield (tonnes per hectare)	Predicted production (tonnes)
2020-21	547433.7	1.391813	761925.3
2021-22	548118.8	1.418111	777293.3
2022-23	548804	1.445642	793374.1
2023-24	549489.1	1.474499	810221.1
2024-25	550174.3	1.504788	827895.7

4. Conclusions

Paddy is a major crop in Raipur district of Chhattisgarh. A study was conducted to know behaviour of yield during the period of 2010-2011 to 2019-20. There was an increasing trend in area, yield, production, proportionate gross irrigated area, maximum temperature and minimum temperature. After ensuring stationarity in the yield of paddy, the parameters ARIMA (0, 2, 0) were estimated. For ensuring the normality, box-cox transformation was carried out. In view of above, linear model was estimated. Diagnostic plots were examined. R^2 = 0.8933 (P-value = 0.03172) and Adj R^2 = 0.7826. Forecasting has been made for the next five years upto 2025 and observed an increased trend in the yield for paddy crop of Raipur.

5. Acknowledgement

The authors would like to acknowledge the Department of Agriculture, Indirawati Bhavan, New Raipur, Chhattisgarh for providing the data regarding gross irrigation area and fertilizers.

6. References

1. Atkinson AC, Riani M, Corbellini A. The Box–Cox transformation: Review and extensions. Statistical Science. 2021;36(2):239-255.
2. Belov AG. A mathematical-statistics approach to the least squares method. Computational Mathematics and Modeling. 2018;29(1):30-41.
3. Benvenuto D, Giovanetti M, Vassallo L, Angeletti S, Ciccozzi M. Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in Brief.

2020;29:105340.
4. Chatfield C. Exploratory data analysis. European Journal of Operational Research. 1986;23(1):5-13.
5. Daoud JI. Multicollinearity and regression analysis. In: Journal of Physics: Conference Series. IOP Publishing; 2017. p. 012009. Vol. 949, No. 1.
6. 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. Available from: http://aps.dac.gov.in/APY/Public_Report1.aspx. 2023.
7. Draper NR, Smith H. Applied Regression Analysis. 3rd ed. New York: John Wiley & Sons; 1998. (Wiley Series in Probability and Statistics); 1985.
8. Ekanayake EMP, Wickramasinghe LCD, Weliwatta RT. Use of regression techniques for rice yield estimation in the North-Western province of Sri Lanka. Ceylon Journal of Science. 2021;50(4):439-447.
9. Fattah J, Ezzine L, Aman Z, El Moussami H, Lachhab A. Forecasting of demand using ARIMA model. International Journal of Engineering Business Management. 2018;10:1847979018808673.
10. Gurka MJ, Edwards LJ, Muller KE, Kupper LL. Extending the Box–Cox transformation to the linear mixed model. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2006;169(2):273-288.
11. Gregory CA. Chhattisgarh: At the crossroads. In: The modern anthropology of India. London: Routledge; c2013. p. 46-65.
12. James A, Tripathi V. Time series data analysis and ARIMA modeling to forecast the short-term trajectory of the acceleration of fatalities in Brazil caused by the coronavirus (COVID-19). PeerJ. 2021;9
13. Khan S, Alghulaiakh H. ARIMA model for accurate time series stocks forecasting. International Journal of Advanced Computer Science and Applications. 2020;11(7):1-6.
14. Kim JH. Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology. 2019;72(6):558-569.
15. Kim M, Hill RC. The Box-Cox transformation-of-variables in regression. Empirical Economics. 1993;18(2):307-319.
16. Mondal P, Shit L, Goswami S. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications. 2014;4(2):13-29.
17. Nasa Prediction of Worldwide Energy Resources. The Power Project; c2023. Available from: https://power.larc.nasa.gov/.
18. Osborne J. Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation. 2010;15(1):1-9.
19. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; c2023.
20. Rajan MS, Palanivel M. Application of regression models for area, production and productivity growth trends of cotton crop in India. International Journal of Statistical Distributions and Applications. 2018;4(1):1-8.
21. 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, India. Agricultural Economics Research Review. 2010;23:79-85.
22. Sakia RM. The Box-Cox transformation technique: a review. Journal of the Royal Statistical Society: Series D (The Statistician). 1992;41(2):169-178.
23. Schaffer AL, Dobbins TA, Pearson SA. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. BMC Medical Research Methodology. 2021;21(1):1-12.
24. Schroeder MA, Lander J, Levine-Silverman S. Diagnosing and dealing with multicollinearity. Western Journal of Nursing Research. 1990;12(2):175-187.
25. Sellam V, Poovammal E. Prediction of crop yield using regression analysis. Indian Journal of Science and Technology. 2016;9(38):1-5.
26. Shrestha N. Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics. 2020;8(2):39-42.
27. Shumway RH, Stoffer DS. ARIMA models. In: Time Series Analysis and Its Applications: With R Examples. 4th ed. New York: Springer; c2017. p. 75-163.