

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2026): 4.49
© 2026 Stats & Maths
Maths 2026; SP-11(9): 09-16
www.mathsjournal.com
Received: 24-07-2026
Accepted: 28-08-2026

Dr. Anil Kumar Singh
Professor & Dean (Academics)
Jagran Institute of Management,
Kanpur, Uttar Pradesh, India

A nonlinear mathematical production model for crop yield estimation and resource optimization in Indian agriculture

Anil Kumar Singh

DOI: <https://www.doi.org/10.22271/math.2026.v11.i9Sa.2694>

Abstract

A nonlinear production model is developed in this study to forecast the yield of crops based on 19,689 data points of Indian agriculture. Factors analyzed include the area cultivated, production, annual rainfall, fertilizer, pesticide, and temperature. The variable 'yield' exhibited high skewness (12.79), with mean and median values being 79.95 and 1.03 respectively. 'Production' had the highest correlation with the yield ($r = 0.5708$, $p < 0.001$), while 'rainfall' and 'temperature' showed weak associations with 'yield'. Multicollinearity was severe between the variables of 'area', 'fertilizer', and 'pesticide'. Thus, a reduced semi-logarithmic OLS regression model is formulated with $\ln(\text{Yield}+1)$ as the dependent variable and $\ln(\text{Area})$, $\ln(\text{Production})$, $\ln(\text{Annual "Rainfall"})$, and Avg. Temperature as the independent variables. This model could explain 89.50% of the variations in the yield ($R^2 = 0.8950$; adjusted $R^2 = 0.8950$) and was highly significant ($F = 41$, 965.98 , $p < 0.001$). All the four independent variables were statistically significant. Diagnostic tests found that there was heteroscedasticity in residuals and residuals did not follow normal distribution, while 632 influential observations, which included some observations with zero yield, were identified with the help of Cook's distance.

Keywords: Crop yield, production model, OLS regression, agricultural modelling, sustainable agriculture

Introduction

Agriculture is significant in ensuring food security, rural livelihoods, and economic growth. Estimation of crop yield is vital in agricultural planning, resource management, food supply, and risk assessment. Crop yield is influenced by various interacting factors such as land under cultivation, production factors, rainfall, temperature, and crop-management practices ^[1, 2]. Production function methods offer a framework that enables an understanding of the association between agricultural inputs and outputs. Methods such as Cobb-Douglas framework and subsequent statistical production functions have extensively been used in modeling the associations between input and output and assessing the contributions of production factors ^[1, 3]. Nevertheless, agricultural data can be highly variable, skewed, exhibit multicollinearity, and nonlinearities in relationships that can influence the statistical estimation process.

The availability of vast data on agriculture and weather has promoted the adoption of machine learning and deep learning methods for crop yield prediction. Such methods can model the interactions between agricultural, climatic, and environmental variables in yield predictions and have shown tremendous promise in yield prediction ^[4]. In spite of the fact that modern sophisticated prediction methods have been invented, the development of easily understandable, statistically proven, and mathematically formulated models which will help measure the dependency between agricultural production and climatic parameters is required. That is why the purpose of this research is the construction of a logarithmic OLS production model which will include the area of cultivation, production volume, annual precipitation, and average temperature. Moreover, it is necessary to check the presence of multicollinearity, heteroscedasticity, residual behavior, and influential observations.

Corresponding Author:
Dr. Anil Kumar Singh
Professor & Dean (Academics)
Jagran Institute of Management,
Kanpur, Uttar Pradesh, India

Literature Survey

The problem of agricultural production has been examined using the statistical, econometric, process-based, remote sensing, and machine learning methods. Previous works in the area of agricultural production have found that the relationship between agricultural output and major factors of production like land, labor, capital, and others is vital. The framework of the Cobb-Douglas production function helped to establish the input-output relationship and production elasticities [5]. Another important factor in agricultural productivity is the climatic factor. Lobell and Field [6] have found that there are considerable relations between climate and global crop yields. They emphasized the role of temperature and precipitation in agricultural productivity. Later researches included climatic and environment variables in their models of agricultural production and yield estimation.

Statistical and econometric methods remain very useful since they allow establishing interpretative relations between inputs and outputs of agricultural production. The production function models were extensively used in estimation of influence of land, water, fertilizers, climatic, and other factors on agricultural production [3-5]. However, nonlinearities, multicollinearity, heteroscedasticity, and influential observations can be present in agricultural systems. The development of data-driven approaches has greatly widened the scope of yield prediction studies. The application of neural networks and data mining approaches in wheat, sugarcane, and apple showed that nonlinear computational models have great capabilities to model the interrelations between agricultural and environment-related features [7-10]. Then Random Forest, Support Vector Machines, Decision Trees, Gradient Boosting, and ensemble-learning were extensively applied [11-15]. It is possible to use many predictors without the need for a specific functional form of relations. Remote sensing has made a significant contribution to yield estimation by means of satellite images, vegetation indices, weather data, and crop properties [16-18]. Thus, there are opportunities for making a spatially distributed evaluation of crop production and to get information which cannot be collected using standard field data sets. A number of systematic reviews show that machine learning and deep learning have been rapidly developing in agriculture. For example, Van Klompenburg *et al.* [19] analyzed machine learning and deep learning studies and found that the most popular features are temperature, rainfall, and soil properties. Other recent investigations have made use of CNN, LSTM, DNN, and recurrent models, using remote sensing and environmental data to increase the level of accuracy [20-22]. Javed and Murad [23] also noted the popularity of Random Forest, ANN, and SVM methods, with CNN, LSTM, and DNN being used in deep-learning models. Furthermore, Yadav *et al.* [24] have pointed out the rising popularity of machine learning, deep learning, and hybrid models from 2017 through 2024 to enhance the accuracy of crop yield and resource planning.

Moreover, comparative analysis and the combination of process-based and data-driven models are being increasingly carried out. Such research indicates that the combination of process-based and data-driven modeling approaches can provide more accurate prediction while retaining the information on crop-growth and environmental processes [24]. In spite of these improvements, however, there is still a need for interpretable mathematically formulated models that estimate the contribution and elasticity of important agricultural and climatic factors. The majority of the previous research has paid more attention to the accuracy of prediction without much concern about multicollinearity, heteroscedasticity, non-normality of residuals, and influential observations. This study aims at filling this gap by proposing a log-transformed OLS production model that includes area,

production, rainfall, and temperature factors while assessing multicollinearity, heteroscedasticity, residuals, and influential observations.

Materials and Methods

Data Set

The research will employ the crop-yield data set consisting of 19, 689 observations. This data set encompasses various crops, seasons, states, and years in relation to agricultural production and climatic variables.

Table 1: Variables Employed for the Research

Variable	Description	Role
Crop	Crop Type	Categorical
Crop_Year	Year of cultivation	Temporal
Season	Agricultural season	Categorical
State	State of India	Categorical
Area	Cultivation	Predictor
Production	Crop production	Predictor
Annual_Rainfall	Annual rainfall	Predictor
Fertilizer	Fertilizer application	Initial predictor
Pesticide	Pesticide application	Initial predictor
Yield	Crop yield	Dependent variable
Avg_Temperature	Average temperature	Predictor
Max_Temperature	Maximum temperature	Descriptive/climatic
Min_Temperature	Minimum temperature	Descriptive/climatic

Exploratory Statistical Analysis

Descriptive statistics were computed for the dependent variable Yield. Measures like mean, median, standard deviation, minimum, maximum, skewness, and kurtosis were computed for this purpose.

Table 2: Descriptive Statistics of Crop Yield

Statistic	Value
Observations	19, 689
Mean	79.954
Median	1.030
Standard deviation	878.306
Minimum	0.000
Maximum	21, 105.000
Skewness	12.785
Kurtosis	175.577

The significant gap between the mean and median as well as the very high skewness value suggest that Yield is significantly right-skewed. Thus, logarithmic transformation was performed to minimize the effect of outliers.

Yield Summary

As per the descriptive analysis, the mean and median of Yield are 79.95 and 1.03 respectively, which suggests a large gap between average and outlier values. Its standard deviation is 878.31, suggesting very high variability. The values for Yield range from 0 to 21, 105. As seen from high positive skewness (12.79), there is a strong right skewness in its distribution whereas the high kurtosis (175.58) reflects the presence of outliers. Hence, transformation of Yield through $\ln(\text{Yield}+1)$ will be done prior to regression analysis.

Heatmap Correlation Analysis

From the above heatmap, one can see the strength and the direction of the linear relationship between crop production variables. Positive correlations were found between the Area, Fertilizer, and Pesticide as well as between different temperature variables, indicating the existence of multicollinearity. While there is a moderate positive correlation ($r=0.57$) between Production and Yield, weak correlations are present with other variables.

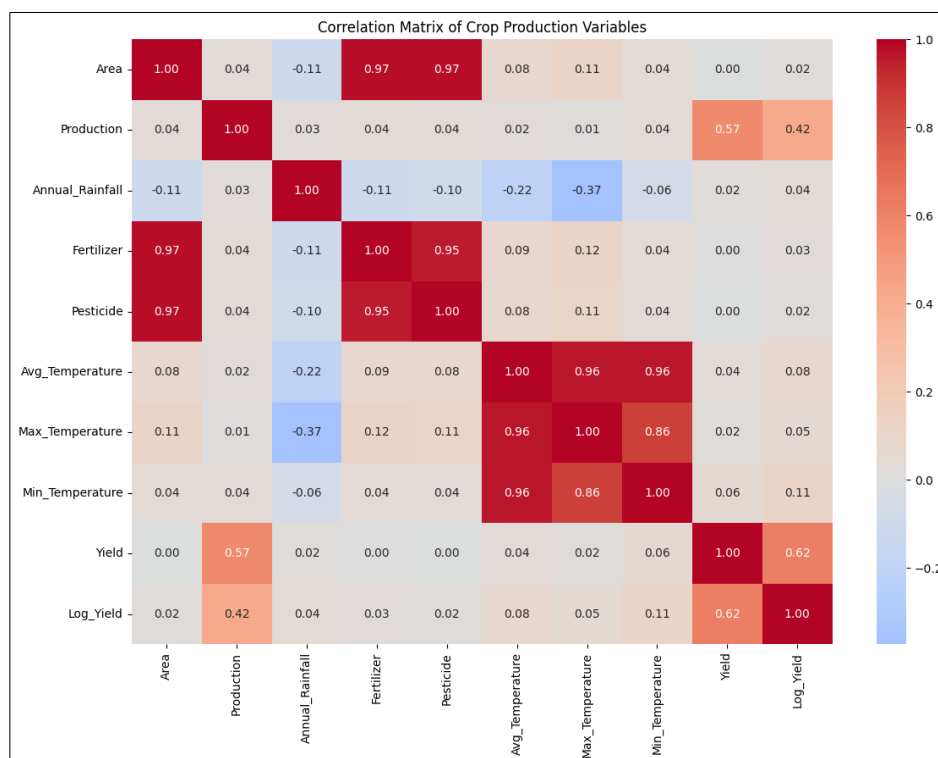


Fig 1: Correlation Matrix

Pearson Correlation Significance

Based on Pearson correlation, Production has a significant positive relationship with Yield ($r = 0.571$, $p < 0.001$). Though Annual Rainfall, Avg. Temperature, Max. Temperature, and Min. Temperature are significant, they have

a very weak correlation ($|r| < 0.06$). The factors Area, Fertilizer, and Pesticide are not significant ($p > 0.05$). Production is the most significant linear predictor of Yield, whereas the rest of the variables have weak relationships with Yield.

Table 3: Pearson Correlation Significance with Yield

Variable	Pearson r	p-value	Significance
Area	0.0019	0.7943	NS
Production	0.5708	<0.001	***
Annual Rainfall	0.0208	0.0036	**
Fertilizer	0.0029	0.6880	NS
Pesticide	0.0018	0.8026	NS
Avg. Temperature	0.0393	<0.001	***
Max. Temperature	0.0187	0.0085	**
Min. Temperature	0.0576	<0.001	***

Note: *** $p < 0.001$, ** $p < 0.01$, NS = Not Significant ($p \geq 0.05$)

OLS Regression Result

The OLS regression model analysis was done with Log_Yield as the dependent variable. The regression model can explain 89.5% of the variability in crop yield ($R^2 = 0.895$; adjusted R^2

= 0.895), and is statistically significant overall ($F = 28, 020$; $p < 0.001$). All the six independent variables were statistically significant at the 1% level or even more apart from Pesticide that was statistically significant at the 1% level.

Table 4: OLS Regression Result

Variables	Coefficients (β)	Standard Error	t-value	p-value	Significance Level
Intercept	1.2762	0.078	16.324	<0.001	***
Log_Area	-0.6238	0.014	-44.850	<0.001	***
Log_Production	0.6904	0.002	406.310	<0.001	***
Log_Annual_Rainfall	-0.0348	0.005	-6.394	<0.001	***
Log_Fertilizer	-0.0541	0.013	-4.255	<0.001	***
Log_Pesticide	-0.0213	0.008	-2.646	0.008	**
Avg_Temperature	0.0049	0.000	10.414	<0.001	***

Statistics for the model: $N = 19, 689$; $R^2 = 0.895$; Adjusted $R^2 = 0.895$; F-statistic = 28, 020 ($p < 0.001$); AIC = 16, 490; BIC = 16, 550.

Log_Production is highly positively correlated with Log_Yield ($\beta = 0.6904$), whereas Log_Area has a large negative coefficient ($\beta = -0.6238$). Temperature contributes only slightly to the production, positively. Rainfall, fertilizer,

and pesticide have negative coefficients. Thus, the model shows high explanatory power, though the significant test results for the diagnostic statistics (Jarque-Bera $p < 0.001$ and Durbin-Watson = 1.

Table 5: Regression Coefficient Analysis

Variables	Coef. (β)	Std. Err	t-stat	p-value	95% Confidence Interval Lower	95% Confidence Interval Upper	Significant?
Intercept	1.2762	0.0782	16.324	<0.001	1.1229	1.4294	Yes
Log_Area	-0.6238	0.0139	-44.850	<0.001	-0.6510	-0.5965	Yes
Log_Production	0.6904	0.0017	406.310	<0.001	0.6871	0.6937	Yes
Log_Annual_Rainfall	-0.0348	0.0054	-6.394	<0.001	-0.0455	-0.0241	Yes
Log_Fertilizer	-0.0541	0.0127	-4.255	<0.001	-0.0791	-0.0292	Yes
Log_Pesticide	-0.0213	0.0081	-2.646	0.008	-0.0371	-0.0055	Yes
Avg_Temperature	0.0049	0.0005	10.414	<0.001	0.0040	0.0058	Yes

Table 6: Model Summary

Statistic	Value
R-Squared (R^2)	0.8952
Adjusted R-Squared	0.8952
F-statistic	28, 021.43
F-test p-value	<0.001
AIC	16, 492.75
BIC	16, 547.97
Observations	19, 689

The regression model has an R squared value of about 89.52%. The extremely large F-statistic and p-value < 0.001 suggest that the whole regression model is statistically significant.

Assessment of Multicollinearity

The Variance Inflation Factor (VIF) test was conducted to detect multicollinearity in the explanatory variables.

Table 7: Variance Inflation Factor (VIF) Test

Variables	VIF	Explanation
Constant	889.701	
Log_Area	246.435	Very high multicollinearity
Log_Production	5.078	Moderate multicollinearity
Log_Annual_Rainfall	1.098	Little or no significant multicollinearity
Log_Fertilizer	208.224	Very high multicollinearity
Log_Pesticide	80.405	Very high multicollinearity
Avg_Temperature	1.109	Little or no significant multicollinearity

The values from the VIF indicate that there is very high multicollinearity between Log_Area, Log_Fertilizer, and Log_Pesticide. There is moderate multicollinearity between Log_Production (VIF=5.08), but there is no multicollinearity between rainfall and temperature. Consequently, dimensionality reduction should be done prior to regression.

Table 8: Model Summary

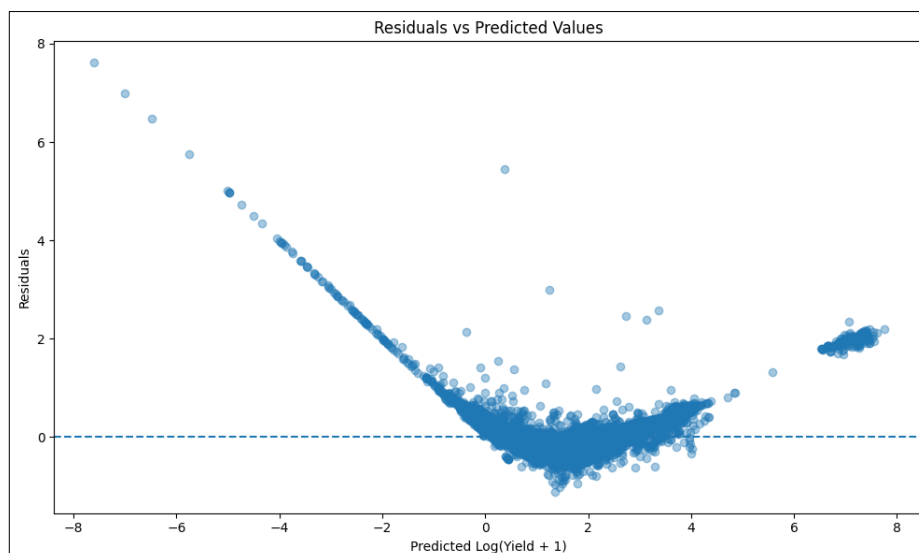
Statistic	Value
R-Squared (R^2)	0.895
Adjusted R-Squared	0.895
F-statistic	41, 970
F-test p-value	<0.001
AIC	16, 508
BIC	16, 560
Observations	19, 689
Durbin-Watson	1.782

The reduced OLS regression model accounts for 89.5% of the variations in Log_Yield and is significant ($F = 41, 970$, $p < 0.001$). All four variables in the regression are significant. The relationship between Log_Production and yield is strongly positive ($\beta = 0.6899$), whereas the relationship between Log_Area and yield is strongly negative ($\beta = -0.6984$). Elimination of Fertilizer and Pesticide solves the problem of multicollinearity without affecting the explanatory power of the regression model.

Table 9: Comparison of Original and Reduced OLS Models

Model	R^2	Adjusted R^2	AIC	BIC	p-value of F-test
Initial Model	0.895203	0.895171	16492.75	16547.97	<0.001
Reduced Model	0.895045	0.895024	16518.31	16557.75	<0.001

The two models have near-identical explanatory capabilities ($R^2 \approx 0.895$). However, the Initial Model has higher explanatory power, AIC, and BIC than the Reduced Model, but the latter has fewer predictors. Therefore, the Reduced Model gives a simpler model without significant reduction in explanatory capability. Hence, the Reduced Model is preferred when simplification of the model is important.

**Fig. 2:** Residuals versus Predicted Values

Residuals are almost randomly distributed around the zero line, however, the distinct curvilinear/funnel shape along with several extreme residuals suggest non-linearity and existence

of outliers. Widening of residuals in certain areas also raises questions regarding heteroscedasticity. Thus, the model might be violating assumptions of homoscedasticity and linearity.

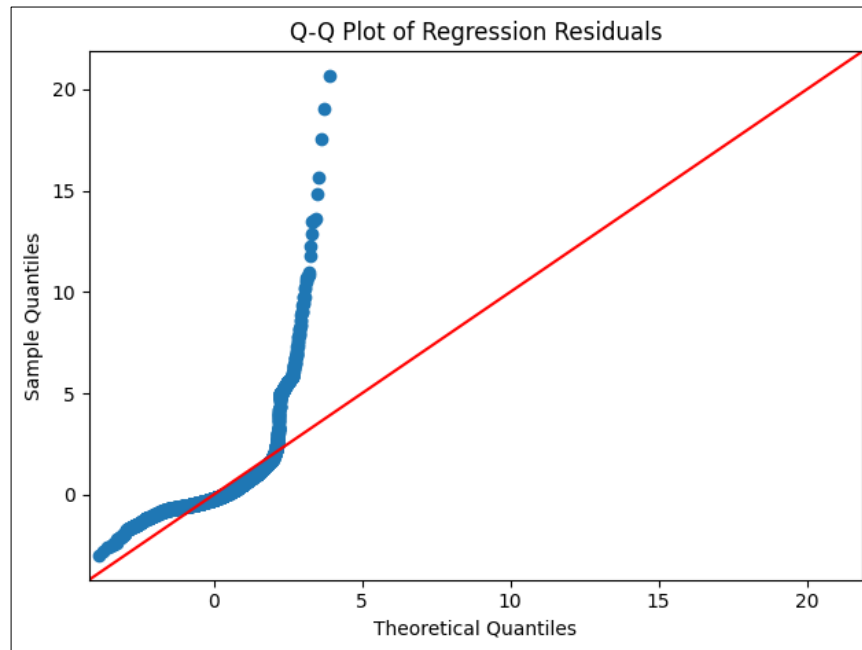


Fig 3: Q-Q Plt of Regression Residuals

The Q-Q plot clearly demonstrates a deviation of the residuals from the reference line, especially in the upper tail. It means that the residuals are not normally distributed and are skewed to the right. Thus, the normality assumption regarding the regression residuals is violated, which means that there are outliers in the data set.

Table 10: Normality of the Residuals Test using Jarque-Bera test statistic

Statistic	Value
Jarque-Bera test statistic	2, 849, 661.964
p-value	<0.001
Residual skewness	5.627
Residual kurtosis	60.853
Conclusion	Residuals are not normally distributed

Jarque-Bera is statistically very significant (p -value < 0.001)

implying a violation of normality of OLS residuals. The positively skewed and kurtosis values show that the residuals are right-skewed and have fat tails.

Table 11: Breusch-Pagan Test for Heteroscedasticity

Statistic	Value
LM statistic	102.058
LM p-value	<0.001
F statistic	25.641
F p-value	<0.001
Conclusion	Heteroscedasticity is present

The LM test and F-test statistics are significant ($p < 0.001$), and hence the variance of the error term is not constant. The data therefore exhibit heteroscedasticity, and hence it would be prudent to consider the use of robust standard errors.

Table 12: OLS Regression with HC3 Robust Standard Errors

Variable	Coefficient (β)	Robust Std. Error	t-value	p-value	95% CI Lower	95% CI Upper	Significance
Intercept	1.0361	0.048	21.810	<0.001	0.943	1.129	***
Log_Area	-0.6984	0.007	-99.056	<0.001	-0.712	-0.685	***
Log_Production	0.6899	0.007	98.373	<0.001	0.676	0.704	***
Log_Annual_Rainfall	-0.0348	0.006	-5.827	<0.001	-0.047	-0.023	***
Avg_Temperature	0.0049	0.000	12.779	<0.001	0.004	0.006	***

Table 13: Model Summary

Statistic	Value
R-squared	0.895
Adjusted R-Squared	0.895
F-statistic (robust)	3, 404
p-value (F-test)	<0.001
AIC	16, 518
BIC	16, 558
Observations	19, 689
Covariance	HC3 robust
Durbin-Watson	1.782
Jarque-Bera p-value	<0.001

Once corrected for heteroscedasticity through HC3 robust standard errors, all predictors retain their statistical significance ($p < 0.001$). The model still explains a very large proportion of the data ($R^2 = 0.895$), meaning that the key regression relations are robust even in the face of heteroscedasticity. Log_Production is the most strongly correlated predictor with Log_Yield ($\beta = 0.6899$), whereas Log_Area is negatively correlated ($\beta = -0.6984$).

Table 14: Splitting between Training and Testing Data

Dataset	No. of observations	Percentage
Training data set	15, 751	80%
Testing Data set	3, 938	20%
Total	19, 689	100%

Training data (80%, 15, 751 observations) and testing data (20%, 3, 938 observations) were separated to develop and test the model.

Table 15: Model Performance

Measure	Value
MAE	0.2044
RMSE	0.3473
R^2	0.9043

The regression model has a value of $R^2 = 0.9043$, which means that 90.43% of variability in the yield is explained by the model. The low values of MAE (0.2044)

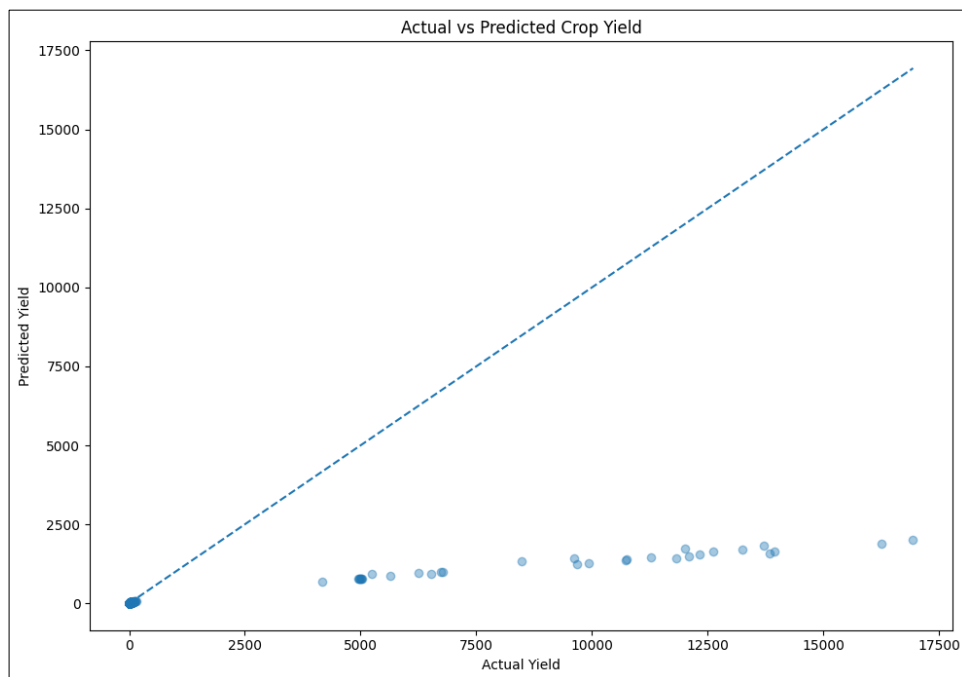


Fig 4: Actual vs. Predicted Crop Yield

As seen from the scatterplot between Actual and Predicted, the fit is quite accurate at lower levels of yields but shows large under prediction for extreme and high levels of yields. It

implies that the model is good enough to capture the overall level of the relationship between the variables but does not have much scope to reproduce the extreme values of yield.

Table 16: Actual vs. Predicted Yield

Observation	Experimental Yield	Theoretical Yield	Error	Difference Absolute Error
18238	3.820	5.1972	-1.3772	1.3772
6918	1.395	1.6756	-0.2806	0.2806
4894	6.523	7.1847	-0.6617	0.6617
10960	0.813	1.1207	-0.3077	0.3077
15615	0.508	0.5790	-0.0710	0.0710
17454	0.358	-0.1698	0.5278	0.5278
10480	0.749	0.8514	-0.1024	0.1024
16231	1.836	2.5862	-0.7502	0.7502
326	75.314	42.5739	32.7401	32.7401
2498	1.775	3.4940	-1.7190	1.7190
16272	0.771	0.9278	-0.1568	0.1568
19308	1.383	3.0607	-1.6777	1.6777
19219	0.940	1.1037	-0.1637	0.1637
1052	2.534	4.1498	-1.6158	1.6158
11750	1.008	1.2892	-0.2812	0.2812
13062	1.413	1.9030	-0.4900	0.4900
8763	0.540	0.4037	0.1363	0.1363
9983	0.953	1.2964	-0.3434	0.3434
14045	0.974	1.1107	-0.1367	0.1367
14202	1.347	1.7554	-0.4084	0.4084

The estimates are fairly accurate as compared to the actuals. Nevertheless, observation 326 has relatively high error

(32.74), which means that there is an outlier or challenging point.

Table 17: Final Regression Model Results

Measure	Values
Number of observations	19, 689
Number of predictors	4
R ²	0.8950
Adjusted R ²	0.8950
F-value	41, 965.98
p-value (F-test)	<0.001
AIC	16, 518.31
BIC	16, 557.75
Test MAE	66.174
Test RMSE	782.001
Test R ²	0.2368

The model is able to explain 89.50% of the variability in the training sample; however, the test R² falls to 0.237, showing a significant decrease in prediction capability. The high test RMSE shows that there may be large prediction errors or outliers present in the model.

Table 18: Regression Coefficients

Dependent Variables	Coefficient (β)	p-value	Statistical Significance
Intercept	1.0361	<0.001	***
Log_Area	-0.6984	<0.001	***
Log_Production	0.6899	<0.001	***
Log_Annual_Rainfall	-0.0348	<0.001	***
Avg_Temperature	0.0049	<0.001	***

Table 19: Cook's Distance - Influential Observations

Observation	Yield	Log_Yield	Cook's Distance
1883	0.0	0.0	0.313036
5367	0.0	0.0	0.224480
1058	0.0	0.0	0.168913
1273	0.0	0.0	0.108604
1441	0.0	0.0	0.074454
1135	0.0	0.0	0.074205
6733	0.0	0.0	0.072541
3873	0.0	0.0	0.063791
5810	0.0	0.0	0.053894
1134	0.0	0.0	0.048118
11580	0.0	0.0	0.040580
4303	0.0	0.0	0.032538
3772	0.0	0.0	0.030845
1651	0.0	0.0	0.030747
5811	0.0	0.0	0.029755
4014	0.0	0.0	0.028595
13246	0.0	0.0	0.028133
9864	0.0	0.0	0.026849
15822	0.0	0.0	0.026590
1678	0.0	0.0	0.023977

According to the Cook's Distance measure, there are 632 observations which are above the threshold, which means that there could be some influence on the regression. The highest value (0.3130) is much higher than the threshold and represents a very influential observation. The interesting thing about the highest influential observations is that they have Yield = 0 and Log_Yield = 0, meaning that the zero yield observations seem to have a very strong influence on the estimated model.

5. Results

There are 19, 689 observations in the analysis. Crop yield was highly variable and positively skewed (12.79), with mean = 79.95, median = 1.03, and std dev = 878.31. Production is correlated with the crop yield the most ($r=0.5708$, " " $p < 0.001$). Rainfall and temperature variables show small correlations.

There was a severe multicollinearity among Area, Fertilizer and Pesticide. That is why Fertilizer and Pesticide were

All four predictor variables are significant ($p < 0.001$). Log_Production is positively correlated with Log_Yield, whereas Log_Area and Log_Annual_Rainfall are negatively correlated. Avg_Temperature is weakly positively correlated with yield.

Significance: *** $p < 0.001$.

Estimated Regression Equation

$$\log(Yield + 1) = 1.0361 - 0.6984(\log Area) + 0.6899(\log Production) - 0.0348(\log Annual Rainfall) + 0.0049(Avg. Temperature)$$

The relationship from the equation shows that Production has a positive impact on yield, whereas Area and Annual Rainfall have negative coefficients. Average Temperature has a minor positive coefficient. All the coefficients are statistically significant at $p < 0.001$.

Cook's Distance - Influential Observations

The threshold value for Cook's Distance is 0.000203 and 632 observations have been identified as influential observations. Table 19 shows the top 20 influential observations as per the Cook's Distance.

removed from the model. The final semi-logarithmic OLS model is:

$$\ln(Yield + 1) = 1.0361 - 0.6984\ln(Area) + 0.6899\ln(Production) - 0.0348\ln(AnnualRainfall) + 0.00494AvgTemperature + \varepsilon$$

It is evident that the model was statistically significant ($F=41, 965.98$, " " $p < 0.001$), accounting for 89.50% of the variance in the transformed variable Yield ($R^2=0.8950$, adjusted $R^2=0.8950$). All four independent variables were statistically significant, where Production and Avg.Temp have positive influence, while Area and Annual Rainfall have negative conditional influence.

The residual analysis reveals the presence of significant non-normality of residuals (Jarque-Bera = 2, 849, 661.96, $p < 0.001$) and heteroscedasticity (Breusch-Pagan $p < 0.001$). The

use of HC3 Robust Standard Errors is hence recommended, where all four predictors will still be statistically significant. There are 632 potentially influential observations according to Cook's Distance, including some

6. Conclusion

The research presented here developed a semi-logarithmic production mathematical model to estimate crop yield using 19, 689 data points on Indian agriculture. The model was able to explain 89.50% of the variance in transformed crop yield ($R^2 = 0.8950$), with production, area, rainfall, and temperature having statistically significant impacts. Nevertheless, issues like heteroscedasticity, non-normality of error terms, and presence of influential observations call for the use of robust estimation techniques and sound interpretation of the findings.

References

1. Ghoshal P, Goswami B. Cobb-Douglas production function for measuring efficiency in Indian agriculture: a region-wise analysis. *Agricultural Economics Research Review*. 2018;62(4):573-579. DOI:10.5958/0976-4666.2017.00069.9.
2. Xu X, Liu Y, Wang J. Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. *Ecological Indicators*. 2019;101:943-952. DOI:10.1016/j.ecolind.2019.01.059.
3. Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Computers and Electronics in Agriculture*. 2018;151:61-69. DOI:10.1016/j.compag.2018.05.012.
4. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*. 2020;177:105709. DOI:10.1016/j.compag.2020.105709.
5. Ghoshal P, Goswami B. Cobb-Douglas production function for measuring efficiency in Indian agriculture: a region-wise analysis. *Agricultural Economics Research Review*. 2018;62(4):573-579. DOI:10.5958/0976-4666.2017.00069.9.
6. Lobell DB, Thau D, Seifert C, Engle E, Little B. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*. 2015;164:324-333. DOI:10.1016/j.rse.2015.04.021.
7. Ahmad M, *et al.* Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. *Journal of the Indian Society of Remote Sensing*. 2018;46:1701-1712. DOI:10.1007/s12524-018-0825-8.
8. Wang X, Huang J, Feng Q, Yin D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sensing*. 2020;12(11):1744. DOI:10.3390/rs12111744.
9. Ali I, Cawkwell F, Dwyer N, Green S. Modeling managed grassland biomass estimation by using multitemporal remote sensing data: a machine learning approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2017;10(7):3254-3265. DOI:10.1109/JSTARS.2016.2561618.
10. Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Computers and Electronics in Agriculture*. 2018;151:61-69. DOI:10.1016/j.compag.2018.05.012.
11. Paudel DR, Boogaard H, *et al.* Machine learning for large-scale crop yield forecasting. *Agricultural Systems*. 2021;187:103016. DOI:10.1016/j.agry.2020.103016.
12. Ansarifar J, Wang L, Archontoulis SV. An interaction regression model for crop yield prediction. *Scientific Reports*. 2021;11:17754. DOI:10.1038/s41598-021-97221-7.
13. Cedric K, *et al.* Crops yield prediction based on machine learning models: case of West African countries. *Smart Agricultural Technology*. 2022;2:100049. DOI:10.1016/j.atech.2022.100049.
14. Iniyar S, *et al.* Crop yield prediction using machine learning techniques. *Advances in Engineering Software*. 2023;175:103326. DOI:10.1016/j.advengsoft.2022.103326.
15. Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*. 2023;218:406-417. DOI:10.1016/j.procs.2023.01.023.
16. Lobell DB, Thau D, Seifert C, Engle E, Little B. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*. 2015;164:324-333. DOI:10.1016/j.rse.2015.04.021.
17. Ahmad M, *et al.* Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. *Journal of the Indian Society of Remote Sensing*. 2018;46:1701-1712. DOI:10.1007/s12524-018-0825-8.
18. Zhang F, Liang S, *et al.* A review of crop yield estimation on pixel and field scales from remotely sensed data. *Science of Remote Sensing*. 2026;13:100342. DOI:10.1016/j.srs.2025.100342.
19. van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*. 2020;177:105709. DOI:10.1016/j.compag.2020.105709.
20. Paudel DR, Boogaard H, *et al.* Machine learning for large-scale crop yield forecasting. *Agricultural Systems*. 2021;187:103016. DOI:10.1016/j.agry.2020.103016.
21. Wang X, Huang J, Feng Q, Yin D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sensing*. 2020;12(11):1744. DOI:10.3390/rs12111744.
22. Javed MA, Murad MAA. Crop yield prediction in agriculture: a comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon*. 2024;10(24):e40836. DOI:10.1016/j.heliyon.2024.e40836.
23. Yadav R, Seth A, Dembla N. Optimizing crop yield prediction: data-driven analysis and machine learning modeling using USDA datasets. *Current Agriculture Research Journal*. 2024;12(1). DOI:10.12944/CARJ.12.1.22.
24. Wang G, *et al.* Machine learning crop yield models based on meteorological features and comparison with a process-based model. *Artificial Intelligence for the Earth Systems*. 2022;1(4). DOI:10.1175/AIES-D-22-0002.1.