Int.J.Curr.Microbiol.App.Sci (2020) *9*(5): 829-836

International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 9 Number 5 (2020)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2020.905.092

Exploring Appropriate Regression Model to Forecast Production of Rabi Pulse in Odisha, India

Abhiram Dash* and Pragati Panigrahi

Odisha University of Agriculture and Technology, Bhubaneswar, India

**Corresponding author*

Keywords

Agricultural sector, Crop yield, Logarithmic model

Article Info

Accepted:
05 April 2020
Available Online:
10 May 2020

A B S T R A C T

Forecasting of area/yield/production of crops is one of the important aspects in agricultural sector. Crop yield forecasts are extremely useful in formulation of policies regarding stock, distribution and supply of agricultural produce to different areas in the country. In this study the forecast values of area, yield and hence production of rabi pulses are found. ARIMA method should not be used for finding the forecasted values for the testing period as this would increase the uncertainty with the end period of testing data. The uncertainty will further increase for the next future periods for which we want to obtain the forecast values. So, in the present study, the regression models are tried for the purpose of forecasting as these models have no such limitation. The regression models used for the study are Linear, Quadratic, Cubic, Power, Compound and Logarithmic. The parametric co-efficients are tested for significance, the error assumptions are also tested and the model fit statistics obtained for different models are compared. Logarithmic model is found to be the best model for area under rabi pulse and power model for yield of rabi pulse. It is found that though there is increase in future areas, the decrease in future yield causes a slow increase in production of rabi pulse.

Introduction

Pulses are an important commodity group of crops that provide high quality protein complementing cereal proteins for pre-dominantly substantial vegetarian population of the country. Pulses have long been considered as poor man's only source of protein. At present, pulses are grown in 18.7 lakh ha with production of 9.4 lakh tonnes and productivity of 502 kg/ha, in Odisha. The most important pulses grown in Odisha are gram, tur, arhar. According to the classification of pulses of Odisha can be broadly divided into kharif and rabi crops.

The Mahanadi delta, the Rushikulya plains and the Hirakud and the Badimula regions are favorable to the cultivation of pulses. Production of pulses is basically concentrated in districts like Cuttack, Puri, Kalahandi, Dhenkanal, Bolangir and Sambalpur.

The Rushikulya plain is the most important agricultural region in Odisha and is dominated by pulses. Odisha covers nearly about 9% area and 8% production of pulses as compare to the total area & production of pulses in India respectively.

Forecasting of area/yield/production of crops is one of the important aspect in agricultural sector. Crop yield forecasts are extremely useful in formulation of policies regarding stock, distribution and supply of agricultural produce to different areas in the country. Statistical forecasting techniques employed should be able to provide objective crop forecast with reasonable precisions well in advance for taking timely decisions. Various approaches have been used for forecasting time series data. Dash *et al.,* (2017) developed appropriate ARIMA models for the time series data on production of food grains in Odisha. Vijay *et al.,* (2018) have studied time series prediction is a vital problem in many applications in nature sciences, agriculture, engineering and economics.

ARIMA technique is most widely used for forecasting time series data. But, in ARIMA, it is not advisable to obtain forecast for future period which is too far from the last period of training data set. This is because the standard error associated with the forecast increases with increase in the length of the forecast period. The increase in standard error of forecast will increase the uncertainty of forecast made for periods which are quite far in future time (Sarika *et al.,* 2011). Since the testing set data in our study comprises of 8 years i.e the end period of the testing data is 8 years far from the end period of the training data, ARIMA method should not be used for finding the forecasted values for the testing period as this would increase the uncertainty with the end period of testing data. The uncertainty will further increase for the next future periods for which we want to obtain the forecast values. So, in the present study, the regression models are tried for the purpose of forecasting as these models have no such limitation.

Materials and Methods

The secondary data pertaining to the area, yield and production of rabi pulses in Odisha are collected for the period from 1970-71 to 2015-16 from various issues of Odisha Agricultural Statistics published by the Directorate Agriculture and Food Production, Government of Odisha. The area, yield and production are expressed in '000 ha, kg/ha and ‘000 tonnes respectively. The data on area and yield of pulses for the year from 1970-71 to 2007-08 are used for model building and hence known as training set data, and for the year from 2008-09 to 2015-16 are not used for model building and kept for cross-validation of the selected model and hence known as testing set data. The forecast values of area and yield and hence production of rabi pulses are obtained for the years from 2016-17 to 2023-24.

Based on the scatter plot of data on area and yield of rabi season in Odisha, the following models are used for the study:

(i) linear model (ii) power model (iii) compound model (iv) logarithmic model and (v) quadratic model (polynomial model of degree two) (vi) cubic model (polynomial model of degree three).

Brief descriptions of different models are given below. In all the models Y_t is the value of the variable in time t, β_0 and β_1 are the parameters of the models used in the study and ε_t is the random error component.

Linear model

Linear model is of the form $Y_t = \beta_0 + \beta_1.t + \varepsilon_t$

Power model

It is of the form: $Y_t = \beta_0 . t^{\beta_1} . \exp(\varepsilon_t)$.
The form of power model after logarithmic transformation is

$$\ln(Y_t) = \ln(\beta_0) + \beta_1 . \ln(t) + \varepsilon_t$$

Compound model

The compound model is a nonlinear model of the form, $Y_t = \beta_0 . \beta_1^t . \exp(\varepsilon_t)$

The form of the compound model after logarithm transformation is

$$\ln(Y_t) = \ln(\beta_0) + \ln(\beta_1) . t + \varepsilon_t$$

Logarithmic model

Logarithmic model is of the form, $Y_t = \beta_0 + \beta_1 . \ln(t) + \varepsilon_t$

Quadratic model

Quadratic model is a second degree polynomial model of the form,

$Y_t = \beta_0 + \beta_1 . t + \beta_2 . t^2 + \varepsilon_t$,

where β_2 is the parameter of the model.

In all the cases the parameters of the model are estimated optimally using the data.

Cubic model

Cubic model is a third degree polynomial model of the form,

$Y_t = \beta_0 + \beta_1 . t + \beta_2 . t^2 + \beta_3 . t^3 + \varepsilon_t$,

where β_3 is the parameter of the model.

In all the cases the parameters of the model are estimated optimally using the data.

The test of overall significance of the model is tested by applying an F test. (Dash *et al.*,)

The significance of the coefficients of the fitted models are tested by using t test (Dash *et al.*,)

The appropriate test statistic is $t = \frac{a_i}{SE(a_i)}$ which follows a 't' distribution with (n – p) degrees of freedom, where 'n' is the number of observations and 'p' is the number of parameters involved in the model. a_i is the estimated value of A_i. $SE(a_i)$ is the standard error of a_i.

Next the model fit statistics, viz., R^2, adjusted R^2 and RMSE, MAPE and MAE are computed for the purpose of model selection. Among the models fitted for the dependent variable, the model which has highest R^2, highest adjusted R^2 and lowest RMSE, MAPE and MAE is considered to be the best fit model for that variable.

Note that, $R^2 = \frac{SSM}{SSE}$,

where, SSM is the sum of square due to model; SSE is the sum of square due to error.

$$SSM = \frac{\sum_{t=1}^{n}\left(\hat{y}_t - \bar{y}\right)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}$$

$$SSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2}$$

where y_t and $\hat{y}_t$ are respectively the actual and estimated values of the response variable at time t, $\bar{y}$ is the mean of y_t.

Adjusted R^2 is defined as

$$\text{Adjusted } R^2 = 1 - (1-R^2) \text{ X } \frac{(n-1)}{(n-p)}$$

To know that the adjusted R^2 penalizes the model for adding independent variables those are not necessary to fit the data and thus adjusted R^2 will not necessarily increase with the increase in number of independent variables in the model.

Again, Root Mean Square Error is defined as

RMSE = $\left\{\frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{(n-p)}\right\}^{1/2}$,

For the sake of clarity we define Mean Absolute Percentage Error (MAPE) here.

$$MAPE=(\sum_{i=1}^{n}\frac{|P_i-O_i|}{O_i}X100)/n$$

where P_i and O_i are respectively the predicted and observed values for i^{th} year, i= 1, 2, …, n.

Absolute Error, = $\sum_{i=1}^{n}|P_i-O_i|$;

Mean Absolute Error. MAE = $\frac{AbsoluteError}{n}$

The residuals diagnostics tests must also be done to render a model fit for selection. The test checks whether or not the errors follow normal distribution with constant variance and are independently distributed.

Here we have considered the following statistical tests for testing the assumptions regarding errors in the model:

(i) Durbin-Watson test for testing independence of residuals (Montgomery *et al.,* (2001)).

(ii) Park's test for testing homoscedasticity of residuals (Basic Econometrics by Gujarati (2004)).

(iii) Shapiro-Wilk's test for testing normality of residuals (Lee *et al.,* (2014))

(iv) After exploring the best fit model, cross validation is done by obtaining the forecast values of the variable from the model for the time period left out for the validation purpose and not considered for developing the model. From the actual and forecast values of the variable for the time period left out for validation, the Absolute Percentage Error (APE) value is obtained for each observation in the validation period. The APE for the i^{th} year of validation period is obtained as,

$$APE_i=\frac{|P_i-O_i|}{O_i}\times 100$$

where P_i and O_i are respectively the predicted and observed values for i^{th} year, i= 1, 2, …, 9. Low value of APE ensures the appropriateness of the selected model for forecasting.

(v) After successful cross validation of the selected model, it is used for the purpose of forecasting.

Results and Discussion

Table 1 shows the parametric coefficients of different regression models fitted to data on area under rabi pulses in Odisha. The study of the table shows that the linear and compound model does not have significant coefficients. So they cannot be considered for selection.

The study of table 2 shows that out of the remaining models, only logarithmic model satisfy all the three assumptions of errors. So logarithmic model is considered to be the best among the selected models. Logarithmic model also has low value of RMSE, MAPE and MAE and high value of adjusted R^2.

Table 3 shows the parametric coefficients of different regression models fitted to data on yield rabi pulses in Odisha. The study of the table shows that the linear, quadratic and cubic models do not have all significant coefficients. So they cannot be considered for selection.

The study of table 4 shows that out of the remaining models, only power model satisfy all the three assumptions of errors. Logarithmic model does not satisfy the assumption of homoscedasticity of errors and compound model do not satisfy the assumption of independency of errors. So power model is considered to the best among the selected models. Power model also has low value of RMSE, MAPE and MAE and high value of adjusted R^2.

In table 5, the result of cross validation of the selected models have been presented. The absolute percentage error for the selected logarithmic model for area under rabi pulses is found to be below 6% for all the years included in the testing data except for 2014-15 for which it is around 17% and thus a low value of MAPE is obtained which is 4.676%. The absolute percentage error for the selected power model for yield of rabi pulses is found to be below 11% for all the years included in the testing data and thus a low value of MAPE is obtained which is 8.079%.

Thus from the table 5 it is found that both the selected models i.e. logarithmic model for data on area under rabi pulses and power model for data on yield of rabi pulses are successfully cross-validated.

Table 6 shows the forecast values of area and yield of rabi pulses of Odisha for the year from 2016-17 to 2023-24. The forecast values of production of rabi pulse in Odisha are obtained from the forecast values of area and yield. The forecast value of area shows that the future values of area under pulse is expected to increase, whereas, the future yield of rabi pulse is expected to decrease. This result in a slow increase in future production of rabi pulses in Odisha which is due to increase in area.

Table.1 Parametric coefficient of the different linear and non-linear models fitted to the training set data on area of Rabi pulses

Model	**b_0**	**b_1**	**b_2**	**b_3**
Linear Model	1060.993** (0.00)	5.712 (0.139)		
Quadratic Model	637.986** (0.00)	69.163** (0.00)	-1.627** (0.00)	
Cubic Model	332.672** (0.0052)	157.425** (0.00)	-7.2119** (0.00)	0.0954** (0.0004)
Power Model	735.6812** (0.00)	0.1546** (0.0002)		
Compound Model	1003.244** (0.00)	1.0065 (0.0669)		
Logarithmic Model	770.88** (0.00)	148.18** (0.0015)		

Table.2 Model fit statistics of the linear and non-linear models fitted to the training set data on the area of Rabi pulses

Model Fit Statistics							Residual Diagnostics		
Model	**RMSE**	**MAE**	**MAPE**	**R^2**	**Adj. R^2**	**F Statistic**	**S-W Statistic**	**D-W Statistic**	**Coefficient of ln(t)**
Linear Model	248.55	219.83	20.72	0.057	0.037	2.287 (0.139)	0.916** (.008)	1.64	-0.669* (.045)
Quadratic Model	176.69	143.46	13.06	0.525	0.4977	19.33** (0.00)	0.991 (0.976)	1.48	0.21 (.628)
Cubic Model	146.67	118.56	11.15	0.673	0.6437	23.28** (0.00)	0.929* (.021)	1.42	-.03 (0.942)
Power Model	239.24	205.38	2.48	0.310	0.291	16.18** (0.0002)	0.948 (0.084)	1.62	0.28 (.296)
Compound Model	150.64	117.06	2.86	0.086	0.065	3.571 (0.0668)	0.921* (.012)	1.45	-0.531 (.096)
Logarithmic Model	222.57	197.12	17.83	0.246	0.2251	11.75** (0.0015)	0.945 (.065)	1.88	0.09 (.787)

Table.3 Parametric coefficient of the different linear and non-linear models fitted to the training set data on yield of Rabi pulses

Model	**b_0**	**b_1**	**b_2**	**b_3**
Linear Model	527.828** (0.001)	-3.261 (0.001)	-	-
Quadratic Model	463.931** (0.00)	6.323 (0.066)	-0.246** (0.0055)	
Cubic Model	438.183** (0.00)	13.767 (0.122)	-0.717 (0.172)	0.008 (0.359)
Power Model	552.163** (0.00)	-0.068 * (0.02)		
Compound Model	520.119** (0.00)	0.993** (0.00)		
Logarithmic Model	544.76** (0.00)	-29.72* (0.0224)		

Table.4 Model fit statistics of the linear and non-linear models fitted to the training set data on yield of Rabi pulses

Model Fit Statistics							Residual Diagnostics		
Model	**RMSE**	**MAE**	**MAPE**	**R^2**	**Adj. R^2**	**F Statistic**	**S-W Statistic**	**D-W Statistic**	**Coefficient of ln(t)**
Linear Model	60.64	50.23	11.43	0.268	0.238	13.214 (0.001)	0.112 (.149)	1.46	0.491 (0.123)
Quadratic Model	52.79	43.437	10.09	0.415	0.382	12.41** (0.00)	0.107 (0.111)	1.58	0.259 (0.444)
Cubic Model	52.13	42.15	9.74	0.429	0.379	8.528** (0.002)	0.132 (.095)	1.62	0.176 (0.636)
Power Model	66.47	56.51	12.69	0.141	0.117	5.91* (0.02)	0.125 (0.104)	1.92	0.335 (0.475)
Compound Model	61.44	51.22	11.52	0.271	0.251	13.41** (0.001)	0.105 (.174)	1.54	0.558 (0.078)
Logarithmic Model	64.13	55.94	12.70	0.137	0.113	5.691* (0.022)	0.121 (0.096)	1.52	0.73 (0.02)

Table.5 Cross validation of the selected best fit model for forecasting area and yield of rabi pulses in Odisha

Year	**Area**			**Yield**		
	Actual values	**Forecast Values**	**APE**	**Actual values**	**Forecast Values**	**APE**
2008-09	1300	1317.49	1.346	468	435.13	7.024
2009-10	1359.55	1321.16	2.824	450	434.39	3.468
2010-11	1274.17	1324.73	3.968	414	433.68	4.753
2011-12	1319.05	1328.22	0.695	477	432.98	9.229
2012-13	1402.69	1331.62	5.067	481	432.29	10.126
2013-14	1368.12	1334.95	2.424	483	431.63	10.636
2014-15	1143.37	1338.21	17.041	481	430.97	10.401
2015-16	1262.7	1313.75	4.043	479	435.88	9.002
MAPE			4.676	MAPE		8.079

Table.6 Forecast values of area, yield and production of rabi pulse in Odisha

Year	Area	Yield	Production
2016-17	1341.41	430.33	577.2
2017-18	1344.52	429.71	577.75
2018-19	1347.57	429.11	578.24
2019-20	1350.56	428.4947	578.71
2020-21	1353.50	427.91	579.17
2021-22	1356.38	427.33	579.62
2022-23	1359.20	426.77	580.06
2023-24	1361.97	426.21	580.48

The regression model used for forecasting of area and yield of rabi pulse in Odisha provides forecast values for much ahead future values. The best regression model for forecasting area is found to be logarithmic model and for yield it is found to be power model. These two models have all significant coefficients, satisfy all the error assumptions and have low value of RMSE, MAPE and MAE and high value of adjusted R^2. The forecast values of production of rabi pulses obtained from the forecast values of area and yield shows a slow increase despite of decrease in yield. This is only due to increase in area under rabi pulse in Odisha which might be the result of shifting of cereal crops to pulse crops in rabi season by enhancing and ensuring assured irrigation in rabi season. But adequate measures must be taken to enhance yield of rabi crops so as to have a sufficient increase in production of rabi pulse in Odisha in the future period which could ensure the nutritional security of the growing population.

References

Dash A, Dhakre DS and Bhatacharya D. (2017). Study of Growth and Instability in Food Grain Production of Odisha: A Statistical Modelling Approach, Environment and Ecology, 35(4D): 3341-3351.

Gujarati, D.N. (2004): *Basic Econometrics*, Fourth Edition, McGraw-HiII Publication, lrwin, 403-404

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*, 3rd Edition, New York, John Wiley & Sons, USA.

Vijay, N. and Mishra, GC. 2018. Time Series Forecasting Using ARIMA and ANN models for Production of Pearl Millet (BAJRA) Crop of Karnataka, India, *International Journal of Current Microbiology and Applied Sciences*, ISSN: 2319-7706 Volume 7 Number 12.

How to cite this article:

Abhiram Dash and Pragati Panigrahi. 2020. Exploring Appropriate Regression Model to Forecast Production of Rabi Pulse in Odisha, India. *Int.J.Curr.Microbiol.App.Sci.* 9(05): 829-836. doi: https://doi.org/10.20546/ijcmas.2020.905.092