Journal of Pharmacognosy and Phytochemistry
Available online at www.phytojournal.com
J P P Journal of Pharmacognosy and Phytochemistry

E-ISSN: 2278-4136
P-ISSN: 2349-8234
www.phytojournal.com
JPP 2020; 9(2): 2301-2304
Received: 16-01-2020
Accepted: 20-02-2020

Abhiram Dash
Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

Pragati Panigrahi
Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

Corresponding Author:
Abhiram Dash
Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

On search of suitable regression model to forecast production of kharif pulse in Odisha: A statistical approach

Abhiram Dash and Pragati Panigrahi

Abstract
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. In this study the forecast values of area, yield and hence production of kharif 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 kharif pulse and power model for yield of kharif pulse. It is found that though there is increase in future areas, the decrease in future yield causes a slow increase in production of kharif pulse.

Keywords: Suitable regression, kharif pulse, statistical approach

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) [1] developed appropriate ARIMA models for the time series data on production of food grains in Odisha. Vijay *et al*. (2018) [4] 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 yearsie. 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 kharif 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 kharif pulses are obtained for the years from 2016-17 to 2023-24.

Based on the scatter plot of data on area and yield of kharif 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 non linear 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 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 = $\sum_{t=1}^{n}\left(\hat{y}_t - \bar{y}\right)^2$, SSE = $\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2$,

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

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

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

$$\text{MAPE}=\left(\sum_{i=1}^{n}\frac{|P_i - O_i|}{O_i}\text{X}100\right)/n$$

, where P_i and O_i are respectively the predicted and observed values for ith 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:

1. Durbin-Watson test for testing independence of residuals (Montgomery *et al.* (2001) [3].
2. Park's test for testing homoscedasticity of residuals (Basic Econometrics by Gujarati (2004) [2].
3. Shapiro-Wilk's test for testing normality of residuals (Lee *et al.* (2014)

4. 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.
5. 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 kharif pulses in Odisha. The study of the table shows that the cubic model does not have significant coefficients. So it 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 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 kharif pulses in Odisha. The study of the table shows that the quadratic and logarithmic models have all significant coefficients. So they can only be considered for selection. The study of table 4 shows that out of the two qualified models from table 3, only logarithmic model satisfy all the three assumptions of errors. Quadratic model does not satisfy the assumption of independency of errors. So logarithmic model is considered to the best among the selected models. Logarithmic 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 kharif 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 5.959%. The absolute percentage error for the selected power model for yield of kharif pulses is found to be below 14% for all the years included in the testing data and thus a low value of MAPE is obtained which is 10.249%. Thus from the table 5 it is found that both the selected models i.e. logarithmic model for data on area under kharif pulses and logarithmic model for data on yield of kharif pulses are successfully cross-validated.

Table 6 shows the forecast values of area and yield of kharif pulses of Odisha for the year from 2016-17 to 2023-24. The forecast values of production of kharif 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 kharif pulse is expected to increase, whereas, the future yield of kharif pulse is expected to decrease. This result in a slow increase in future production of kharif 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 Kharif pulses.

Model	b_0	b_1	b_2	b_3
Linear Model	201.209** (0.00)	14.277** (0.00)		
Quadratic Model	25.9479 (0.562)	40.5660** (0.00)	-0.6741** (0.00)	
Cubic Model	12.7214 (0.8409)	44.3896** (0.0028)	-0.9160 (0.2693)	0.0041 (0.7655)
Power Model	101.0457** (0.00)	0.5288** (0.00)		
Compound Model	199.5949** (0.00)	1.0391**(0.00)		
Logarithmic Model	-34.81*(0.049)	189.84** (0.00)		

Table 2: Model fit statistics of the linear and non-linear models fitted to the training set data on the area of Kharif 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	110.019	99.786	24.923	0.669	0.659	72.8** (0.00)	0.180** (.001)	1.64	0.2 (0.437)
Quadratic Model	82.91	74.488	18.934	0.812	0.802	75.75** (0.00)	0.150** (0.009)	1.48	0.394 (0.174)
Cubic Model	82.80	74.488	19.274	0.813	0.796	49.21** (0.00)	0.142** (.007)	1.42	0.347 (0.259)
Power Model	101.44	87.01	20.88	0.789	0.784	135.2** (0.00)	0.152** (0.007)	1.62	0.286 (0.505)
Compound Model	144.37	125.08	26.69	0.683	0.674	77.44** (0.00)	0.187** (.001)	1.45	0.956 (0.001)
Logarithmic Model	100.49	89.728	25.814	0.724	0.717	94.58** (0.00)	0.130 (.070)	1.88	-0.114 (0.664)

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

Model	b_0	b_1	b_2	b_3
Linear Model	529.339** (0.00)	-1.541 (0.211)		
Quadratic Model	447.9018** (0.00)	10.6745* (0.0251)	-0.3132** (0.009)	
Cubic Model	456.1612** (0.00)	8.2868 (0.494)	-0.1621 (0.821)	-0.0025 (0.831)
Power Model	515.3706** (0.00)	-0.02406 (0.457)		
Compound Model	514.0877** (0.00)	0.9969 (0.207)		
Logarithmic Model	529.44** (0.00)	-11.13* (0.043)		

Table 4: Model fit statistics of the linear and non-linear models fitted to the training set data on yield of Kharif 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	79.57	67.936	14.086	0.043	0.0165	1.623 (0.210)	0.130 (.377)	1.46	0.53 (0.852)
Quadratic Model	72.11	59.422	12.477	0.214	0.1694	4.773* (0.014)	0.096 (0.897)	1.58	0.14 (0.652)
Cubic Model	72.06	59.35	12.47	0.215	0.1461	3.1118* (0.039)	0.095 (.788)	1.62	0.297 (0.346)
Power Model	83.29	67.61	13.85	0.015	-0.0119	0.5651 (0.457)	0.121 (0.426)	1.52	0.076 (0.908)
Compound Model	82.14	67.37	13.79	0.044	0.0173	1.655 (0.206)	0.133 (.371)	1.54	0.157 (0.595)
Logarithmic Model	80.79	68.588	14.23	0.013	-0.0136	0.5028 (0.4828)	0.121 (.424)	1.92	0.282 (0.393)

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

Year	Area			Yield		
	Actual values	**Forecast Values**	**APE**	**Actual values**	**Forecast Values**	**APE**
2008-09	730.00	665.49	8.837	529	488.38	7.678
2009-10	720.13	670.17	6.937	539	488.11	9.442
2010-11	729.40	674.75	7.493	540	487.84	9.659
2011-12	722.89	679.22	6.042	564	487.58	13.55
2012-13	685.6	683.58	0.295	559	487.32	12.823
2013-14	712.62	687.85	3.476	556	487.07	12.397
2014-15	721.69	692.02	4.111	536	486.83	9.174
2015-16	738.01	660.68	10.478	527	488.67	7.274
MAPE			5.959	MAPE		10.249

Table 6: Forecast values of area, yield and production of kharif pulse in Odisha

Year	Area	Yield	Production
2016-17	696.1	1207.79	1312.4
2017-18	700.1	1208.87	1312.57
2018-19	704.01	1209.93	1312.73
2019-20	707.85	1210.96	1312.9
2020-21	711.61	1211.97	1313.05
2021-22	715.29	1212.95	1313.21
2022-23	718.91	1213.91	1313.36
2023-24	722.46	1214.84	1313.5

Conclusion

The regression model used for forecasting of area and yield of kharif 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 also logarithmic 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 kharif 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 kharif pulse in Odisha which might be the result of shifting of non-food grain crops to pulse crops in kharif season. But adequate measures must be taken to enhance yield of kharif crops so as to have a sufficient increase in production of kharif pulse in Odisha in the future period which could ensure the nutritional security of the growing population.

References

1. Dash A, Dhakre DS, Bhatacharya D. Study of Growth and Instability in Food Grain Production of Odisha: A Statistical Modelling Approach, Environment and Ecology. 2017; 35(4D):3341-3351.
2. Gujarati DN. Basic Econometrics, Fourth Edition, McGraw-HiII Publication, lrwin, 2004, 403-404
3. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis, 3rd Edition, New York, John Wiley & Sons, USA, 2001.
4. Vijay N, Mishra GC. 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, 2018, 7(12).