International Journal of Multidisciplinary Research and Development
Online ISSN: 2349-4182, Print ISSN: 2349-5979; Impact Factor: RJIF 5.72
Received: 23-10-2019; Accepted: 25-11-2019
www.allsubjectjournal.com
Volume 6; Issue 12; December 2019; Page No. 107-116

Use of regression model and ARIMA model for forecasting kharif food grain production of Odisha: A comparative study

A Dash[1], D Bhattacharya[2], DS Dhakre[3]
[1] Assistant Professor (Agril. Statistics), College of Agriculture, Chiplima (OUAT), Odisha, India
[2] Professor (Agril. Statistics), PSB, Visva, Bharati, Odisha, India
[3] Assistant Professor (Agril. Statistics), PSB, Visva, Bharati, Odisha, India

Abstract
Present work discusses the issue related to the model selection for efficiently forecasting the area, yield and hence production of food grains grown in Odisha. Several models have been tried on the observed data on area and yield for the period from 1992-93 to 2010-11and the best model have been selected by comparing the model fit statistics after testing the model diagnostics criteria. The models tried are ordinary regression models, spline regression models and ARIMA models. The model diagnostic criteria used are Shapiro-Wilk's Statistic and Durbin-Watson statistic. The model fit statistics used are R^2, adjusted R^2 and Root Mean Square Error (RMSE). The selected models are also cross validated by using the known values for the year from 2011-12 to 2015-16. The cross validation of the selected model test the efficiency of the model in forecasting. Lastly, the best selected model has been used for forecasting area and yield of kharif food grains. Using the forecast values of area and yield the forecasts are obtained for production of kharif food grains.

Keywords: model fit statistics, adjusted R^2, root mean square error, forecast

Introduction

Role of agriculture in economy of the state is important for its contribution to the state income. Odisha is an agriculture dependent state Agriculture not only provides food to its population but also provides employment opportunities to about 60% of the total workforce of the State. About 70% of our population is directly or indirectly engaged in agriculture. The state has a total geographical area of 15,571 thousand hectares of which total cultivated land are about 6,180 thousand hectares. The net area sown is about 5,331 thousand hectares which is 34% of the state geographical area.

The share of' agriculture sector in the GDP of the state has been declining over the years. It declined from 23.5% in 2004-05 to about 16 % in 2015-16. The shares of other two sectors (i.e. industrial and service sector tend to dominate over the agriculture sector in their contribution to GDP of the state. The service sector being the leading contributor and it contributes to a tune of 51% in the year 2014-15. With the decrease in contribution of agriculture sector to the GDP of the state, the relative income of farmers is going down compared to those working in other two sectors. Besides this, per capita land holding is also declining gradually. All these events combined have adversely affected the financial condition of the farmers of Odisha.

There is no doubt about it that for ensuring food, livelihood and nutritional security of the growing population, agricultural production and yield have to be increased. So, the need is felt to use better agricultural technology, higher public and private investments, more awareness in different programmes related to agriculture, effective implementation of ongoing programmes in agriculture and allied sectors for increasing yield per unit land area and increasing the cropping intensity.

Recognizing the challenges faced by the agriculture sector and the fact that higher priority to agriculture would help achieve the goals of faster reduction of poverty as well as malnutrition and make the growth process more inclusive, a separate agriculture budget for the growth of agriculture and its allied activities like Horticulture, Fisheries, Animal Husbandry, Irrigation, Co-operative Credit, etc. has been set in force in the state budget of Odisha since 2013-14. Odisha is among the few states in the country to present a separate agriculture budget.

Out of 15,571 thousand hectares total geographical area of the State, 5813 thousand hectares is forest area, 342 thousand hectares of miscellaneous trees and groves, 454 thousand hectares of permanent pastures, 375 thousand hectares of culturable waste land and 840 thousand hectares of barren and unculturable land. The total cultivated land of the State is 6180 thousand hectares out of which 2914 thousand hectares (47%) is high land, 1755 thousand hectares (28%) is medium land and 1511 thousand hectares (25%) is low land (Odisha Agricultural Statistics, 2013-14).

In Odisha nearly 84% of the population of the state lives in rural areas. Kharif is the main cropping season in Odisha and rice is the principal crop which occupies 67% of the cultivated land in the state. Cropping during rabi season is confined to the irrigated tracts and land with available residual moisture in the soil, which mostly depends on the occurrence of rainfall during the last part of September.

The food grains produced in Odisha include cereals and pulses. Rice, maize, ragi, wheat, jowar, bajra and small millets are the cereal crops grown in Odisha. The pulse crops grown in the state include arhar, mung, biri, kulthi, cowpea, field pea, gram and lentil. The crops like wheat, bajra, jowar, small millets are grown to lesser extent. The major pulse crops are arhar, mung, biri and kulthi. Pulses

are grown mainly in uplands during kharif season predominantly in inland districts and in rice fallows during rabi season, mostly in coastal districts under available moisture. If there is a good rainfall during last part of October, then the area of coverage and production of pulse crops are higher. Above all, rice is the most important cereal and also the most important crop of Odisha. The agricultural scenario of the state can be best reflected from the status of food grains

The state of Odisha ranks l0th, 12th and 23rd position with respect to area, production and yield of food grains taken as average over the years 2010-11 to 2015-16, at all India level (Odisha Economic Survey, 2016-17.). The rank position shows that Odisha lags much behind the other states of India with respect to the yield of food grains. Because of vulnerability of the State to natural calamities, the food grain production generally fluctuates from year to year. The yield of food grains of the state is usually above 80% of that of India. Food grains shares almost 86% of the total cropped area in the state in kharif season.

Hence, forecasting the area, production and yield of food grains have become very important in formulating strategies to feed the growing population of the state. We know that the production depends on area under cultivation and yield per unit area (production = area x yield) In the present study the forecast of area and yield of food grains have been obtained for kharif season and using these forecasts the forecast of production is obtained. Appropriate forecasting model based on different forecasting methods such as regression methods (ordinary and spline) needs to be fitted to data on area and yield of food grains for forecasting purpose. Different forecasting methods are compared to select the best forecasting model. The method providing the best forecast for a particular variable is used for obtaining the forecast value of that variable.

Review of earlier works related to the present study are done which would guide us in the adopting right approach for the conducting the research study.

Nelson (1972) [7] compared econometric method (regression model) with time-series method (ARMA model) for a longer time horizon. He concluded that the simple ARMA models are relatively more accurate with respect to post sample (future) predictions than the complex econometric models.

Bajpai and Venugopalan (1996) [1] forecast the all-India sugarcane production by applying regression analysis technique and autoregressive integrated moving average (ARIMA) time series modeling. It was found that ARIMA models produced higher R^2 value, lower Root Mean Square Error and Mean Absolute Error values compared to regression models. ARIMA (0, 1, 1) model has been used to obtain forecast values of sugarcane production in India for subsequent years. The forecast values obtained through ARIMA were found close to the observed values.

Prabakaran and Sivapragasam (2014) [10] used ARIMA model to forecast area and yield of rice in India. Standard statistical techniques were used to test the validity of the model. ARIMA (1, 1, 1) model was used to forecast both area and production in India for four leading years. The results also showed that the forecast of area for the year 2015 was about 44.75 thousand hectares with upper and lower limits 47.53 and 41.97 thousand hectares, respectively. The model also showed that the forecast for the rice production for the year 2015 was about 104.37 thousand tonnes with upper and lower limits 115.26 and 93.48 thousand tonnes respectively.

Borkar and Tayade (2016) [2] made an empirical study on forecasting time series data of cotton production in India by using Box-Jenkins ARIMA methodology. The diagnostic checking reveals that ARIMA (2, 1, 1) is found to be the appropriate model for forecasting cotton production in India. Based on the selected model, forecasts are obtained for the period 2015-16 to 2020-21.

Methodology

The study relates to forecasting of area, production and yield of food grains in the state of Odisha in kharif season for the year 2016-17 to 2018-19. The data on area and yield of kharif food grains in Odisha are collected for the period 1992-93 to 2015-16 from various volumes of Odisha Agricultural Statistics published by the Directorate of Economics and Statistics, Government of Odisha.

Here three different approaches have been used for fitting the observed data for the purpose of forecasting viz. ordinary regression, spline regression and ARIMA. Different forecasting models have been fitted to the data for the years 1992-93 to 2010-11. The data for the years 2011-12 to 2015-16 are kept for cross validation purpose of the selected model under each approach. The possible models that could fit well to the data under different approaches are linear, compound, logarithmic, power and quadratic. These models are also fitted by using spline regression technique. Spline regression technique involves in joining two or more separate regression lines at a point known as spline knot(s) while their slopes are allowed to be different at that point. From scatter of the data on area and yield of kharif food grains in Odisha from the year 1992-93 to 2010-11, it is seen that there is a change point in the data in the year 2002-03. So, the knot of spline regression is placed at the year 2002-03 (t = 11). Thus the whole period (1992-93 to 2010-11) is divided into two periods – 1992-93 to 2002-03 (Period I) and 2003-04 to 2010-11 (Period II). The best forecasting model is selected from each approach and are then cross validated by using the data from the period from 2011-12 to 2015-16. The model yielding the lowest MAPE during cross – validation is used for forecasting of the variable for the period from 2016-17 to 2018-19.

For clarity a brief descriptions of different regression models tried here are given below. In all the models Y_t is the value of the variable at time t, β_0 and β_1 are the parameters of the model used in the study and ε_t is the random error component

1. **Linear model:** Linear model is of the form $Y_t = \beta_0 + \beta_1.t + \varepsilon_t$
2. **Power model:** Power model is of the form: $Y_t = \beta_{0.\ t^{\beta_1}}.$ Exp (ε_t).
   The form of power model after logarithmic transformation is: $\ln (Y_t) = \ln (\beta_0) + \beta_1. \text{Ln } (t) + \varepsilon_t$
3. **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 logarithmic transformation is
   $\text{Ln } (Y_t) = \ln (\beta_0) + \ln (\beta_1). t + \varepsilon_t$
4. **Logarithmic model:** Logarithmic model is of the form, $Y_t = \beta_0 + \beta_1. \ln(t) + \varepsilon_t$
5. **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 observed data.
The various spline regression models fitted are:
The linear spline model is made continuous for the whole period (i.e., 1992-93 to 2010-11) with a single knot at k = 11, and it is written in the form of

$$Y_t = \beta_0 + \beta_1. t. I_{(1\le t\le 22)} + \{\beta_1. t + A_1 (t-k)\}. I_{(23\le t\le 46)} + \varepsilon_t,$$

Where $I_{(P)}$ is the indicator function which is 1 if P holds and 0 otherwise.

Power spline model

$$Y_t = \beta_0. t^{\beta_1} . I_{(1\le t\le 22)} \{ t^{\beta_1} . (t-k)^{A_1} \}. I_{(23\le t\le 46).} \text{Exp}(\varepsilon_t)$$

The power spline model is transformed to linear form by natural log transformation as,
$\text{Ln}(Y_t) = \ln \beta_0 + \beta_1 . \text{Ln}(t). I_{(1\le t\le 22)} + \{\beta_1 . \text{Ln}(t) + A_1 \ln(t-k)\}. I_{(23\le t\le 46)} + \varepsilon_t,$
Where $I_{(P)}$ is as defined earlier.

Compound spline model

$$Y_t = \beta_0. B1^{t.} I_{(1\le t\le 22).} \{\beta_1^{t} . A_1^{(t-k)}\}. I_{(23\le t\le 46).} \exp(\varepsilon_t)$$

The compound spline model can be transformed to linear form by a natural log transformation and written as,
$\text{Ln}(Y_t) = \ln \beta_0 + t. \ln(\beta_1). I_{(1\le t\le 22)} + \{t. \ln(\beta_1) + (t-k). \ln(A_1)\} I_{(23\le t\le 46)} + \varepsilon_t$
Where $I_{(P)}$ is as defined earlier.

Logarithmic spline model

$$Y_t = \beta_0 + \beta_1. \text{Ln}(t). I_{(1\le t\le 22)} + \{\beta_1. \text{Ln}(t) + A_1. \text{Ln}(t-k)\}. I_{(23\le t\le 46)} + \varepsilon_t$$

Where $I_{(P)}$ is as defined earlier.

Quadratic spline model

$$Y_t = \beta_0 + \{\beta_{1.} t + \beta_{2.} t^2\} I_{(1\le t\le 22)} + \{\beta_{1.} t + A_{1.} (t-K) + \beta_2 . t^2 + A_2 . (t-K)^2\} I_{(23\le t\le 46)} + \varepsilon_t$$

Where $I_{(P)}$ is as defined earlier.
In case of both ordinary and spline regression models, for testing the significance of the coefficient β_0 the null hypothesis is taken as H_0: $\beta_0 = 0$ and the alternative hypothesis as, H_1: $\beta_0 \neq 0$, for linear, logarithmic and quadratic models, whereas, for power model and compound model the null hypothesis is taken as H_0: $\beta_0 = 1$ and the alternative hypothesis as,
H_1: $\beta_0 \neq 1$.
For testing the significance of the coefficient β_1 in case of linear model, power model, logarithmic model and quadratic model the null hypothesis is taken as
H_0: $\beta_1 = 0$ and the alternate hypothesis as, H_1: $\beta_1 \neq 0$, whereas, for compound model the null hypothesis is taken as H_0: $\beta_1 = 1$ and the alternative hypothesis as, H_1: $\beta_1 \neq 1$.
For testing the significance of the coefficient β_2 in case of quadratic model the null hypothesis is taken as H_0: $\beta_2 = 0$ and the alternate hypothesis as, $H_{1:}$ $\beta_2 \neq 0$,
The appropriate test statistic is $t = \frac{b_j}{SE(b_j)}$ 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, $SE(b_j)$ is the standard error of b_j, j =0,1,2.

$$SE(b_j) = \left(\frac{\sum_{t=1}^{n} e_t^2}{(n-p)} \times X_{kk} \right)^{\frac{1}{2}},$$

Where e_t is the residual at time t; k = 1 for j=0, k = 2 for j = 1 and k = 3 for j = 2.
X_{kk} is the element in the k^{th} row and k^{th} column of the inverse of the matrix X, Where

$$X = \begin{bmatrix} n & \sum t \\ \sum t & \sum t^2 \end{bmatrix}$$ for linear spline model and compound spline model;

$$X = \begin{bmatrix} n & \sum \ln(t) \\ \sum \ln(t) & \sum [\ln(t)]^2 \end{bmatrix}$$ For power spline model and logarithmic spline model;

$$X = \begin{bmatrix} n & \sum t & \sum t^2 \\ \sum t & \sum t^2 & \sum t^3 \\ \sum t^2 & \sum t^3 & \sum t^4 \end{bmatrix}$$, for quadratic spline model.

Let t_o and $t_{\alpha,n-p}$ be the observed and tabulated values of t_{n-p} respectively, such that $P(t_{n-p} > t_{\alpha,n-p}) = \alpha$.
Now, if $t_o > t_{\alpha,n-p}$, then H_0 is rejected at α% level of significance and the parameter is considered to be significantly different from zero or one, as the case may be.
The influence of the parameter β_0 in the model is considered to be significant if β_0 is significantly different from zero (in case of linear model, logarithmic model and quadratic model) or significantly different from one (in case of power model and compound model), otherwise the influence of the parameter β_0 in the model is considered to be non-significant.
The influence of the parameter β_1 in the model is considered to be significant if β_1 is significantly different from zero (in case of linear model, power model, logarithmic model and quadratic model) or significantly different from one (in case of compound model) otherwise the influence of the parameter β_1 in the model is considered to be non-significant.
The influence of the parameter β_2 in the quadratic model is considered to be significant if β_2 is significantly different from zero, otherwise the influence of the parameter β_2 in the model is considered to be non-significant.
To test the overall significance of the model, F test is used.

$$\text{The appropriate test statistic is } F = \frac{\text{mean square due to the model (MSM)}}{\text{mean square due to the error (MSE)}},$$

$$\text{Where MSM} = \frac{\sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2}{p-1}; \text{ MSE} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n-p};$$

y_t And $\hat{y}_t$ are, respectively, the actual and estimated values of the response variable y at time t and $\bar{y}$ is the mean of y_t.
Let F_o and $F_{\alpha,(p-1,n-p)}$ be the observed and the tabulated values

of $F_{(p-1,n-p)}$, respectively, such that $P(F_{p-1,n-p} > F_{\alpha; p-1,n-p}) = \alpha$
If $F_o > F_{\alpha, (p-1,n-p),}$ then H_0 is rejected at α% level of significance and the model is considered to be significant.
To test the significance of the coefficients A_i of the model, where i=1 (for linear model, power model, compound model and logarithmic model) and i = 1, 2 (for quadratic model), t-test has also been used.
In case of linear model, power model, logarithmic model and quadratic model the null hypothesis is taken as, $H_{0:}\ A_i = 0$ and the alternative hypothesis as, $H_{1:}\ A_i \neq 0$.
In case of compound spline model the null hypothesis is taken as, $H_{0:}\ A_i = 1$ and the alternative hypothesis as, $H_{1:}\ A_i \neq 1$.
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. and $SE(a_i)$ is the standard error of a_i.
In order to carry out the above tests we have to assume that errors follow normal distribution and are independently distributed.
The following statistical tests are considered for testing the assumptions regarding errors in the model:

1. Durbin-Watson test for testing independence of residuals.
2. Shapiro-Wilk's test for testing normality of residuals.

i) Durbin-Watson test: This test considers the first order autocorrelation among the residuals. (Montgomery *et al*. (2001)).
Null hypothesis is taken as, H_0: the errors are independent.
And the alternative hypothesis as, H_1: the errors are not independent.

Durbin-Watson test statistic (D-W statistic) $d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$,

Where, $e_t = y_t - \hat{y}_t$, y_t and $\hat{y}_t$ are respectively the actual and estimated values of the response variable at time t and *n* is the no. of observations
The value of'd' ranges from 0 to 4. Upper and lower critical values, d_U and d_L have been tabulated for different values of k (the number of explanatory variables) and *n* (the number of observations) for corresponding level of significance (α) in the Durbin – Watson statistical table of critical values.
If $d < d_L$, dependence of errors is significant. If $d > d_U$, then the dependence of errors is non-significant and the residuals are considered to be independent.
If $d_L < d < d_U$, test is inconclusive.
For testing negative autocorrelation, the statistic 4 - d is used to compare with d_U and d_L.

ii) Shapiro-Wilk's test: This test is used for testing normality of the residuals.
Null hypothesis here is H_0: the errors follow normal distribution. The null hypothesis is tested against the alternative hypothesis, H_1: The errors do not follow normal distribution. To carry out the test, the data pertaining to errors are arranged in ascending order so that $e_{(1)} \leq e_{(2)} \leq \ldots \leq e_{(n)}$

Next we calculate the Shapiro-Wilk's (S-W) test statistic as given by

$W = \frac{s^2}{b}$,

Where, $s^2 = \sum_{k=1}^{m} a(k)\{e_{(n+1-k)} - e_{(k)}\}^2$; $b = \sum_{t=1}^{n}(e_t - \bar{e})^2$ (Lee *et al*. (2014)

If *n* is even, then $m = \frac{n}{2}$. If n is odd, then $m = \frac{n-1}{2}$
The parameter k takes the values 1, 2, m.
n is the number of observations, $e_{(k)}$ is the k^{th} order statistic in the set of residuals,
e_t is the residual at time 't' and $\bar{e}$ is the mean of e_t.
The values of coefficients a (k) for different values of *k* and particular values of *n* are obtained from the table of Shapiro-Wilk. (Hanusz and Tarasinska (2011), Table no. 2)
For a given value of n, the value of p that is closest to 'W' can be obtained from Shapiro-Wilk's table. If the p value exceeds 0.05, then the null hypothesis cannot be rejected. If it lies below 0.05 but above 0.01, then the null hypothesis is rejected at 5% level. If the *p* value is below 0.01, then the null hypothesis is rejected at 1% level. (Hanusz and Tarasinska (2011), Table no. 3)

Model selection
The model to be considered for selection must satisfy the assumptions regarding the errors, should have overall significance, and the parameters are to be significantly different from zero.
Next the model fit statistics, viz., R^2, adjusted R^2 and RMSE 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 is considered to be the best fit model for that dependent 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.
The expressions for SSM and SSE are, respectively,

$$SSM = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2,\ SSE = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2,$$

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

Adjusted R^2 is defined as Adjusted $R^2 = 1 - (1-R^2) \times \frac{(n-1)}{(n-p)}$

It is known that the adjusted R^2 penalizes the model for adding some independent variables which are not necessary to fit the data and thus adjusted R^2 will not necessarily increase with the increase in the number of independent variables included in the model.

Again, Root Mean Square Error is defined as RMSE =

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

Fitting of ARIMA Model

By looking at the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the differenced series, the orders of AR and MA terms that are needed can be tentatively identified. The lag beyond which the PACF cuts off is the indicated number of AR terms to be retained in the model. The lag beyond which the ACF cuts off is the indicated number of MA terms to be retained in the model. The number of AR terms is denoted by 'p' and the number of MA terms is denoted by 'q'

Let Y_t be the value of the time series at time t, where, t = 1, 2, 3, n

If the order of differencing, d=1, then $y_t = Y_t - Y_{t-1}$.

If the order of differencing, d=2, then $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$.

The forecasting equation of ARIMA model with 'p' number of AR terms 'q' number of MA terms and order of differencing 'd' is expressed as:

$Y_t = \mu + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \theta_p Y_{t-p} - \Phi_1\varepsilon_{t-1} - \Phi_2\varepsilon_{t-2} - \ldots - \Phi_q\varepsilon_{t-q}$,

where, Y_t, Y_{t-1}, Y_{t-2},… are the stationarized values of time series for time points t, t-1, t-2,… which may be the original values of the series or the values obtained after first or second order differencing, μ is the constant term; θ_1, θ_2,… are the AR coefficients; Φ_1, Φ_2,… are the MA coefficients and ε_{t-1}, ε_{t-2},…, ε_{t-q} are the error terms at lags 1, 2, …, q respectively.

After identifying the values of p (the order of AR terms) and q (the order of MA terms), the parameters of the autoregressive and moving average terms are estimated using simple least square techniques. In the present study, the parameters of the AR and MA terms are obtained with the help of the forecasting tool available in the statistical package SPSS 20.0.

Next, after determining the values of p, q and d the parameters associated with AR and MA terms are estimated. Later the constant and the coefficients of the AR and MA terms are tested for their significance. If the constant term appears not to be significantly different from zero, then ARIMA model without constant is fitted. After testing the significance of the model parameters, the diagnostic test for the residuals of the selected model is done. The residual ACF and PACF plots were obtained by using the forecasting tool of SPSS 20.0. If none of the residual ACFs and PACFs are significant, then the model can be considered to be adequate.

A formal goodness of fit test of the selected model to the observed time series data is also done by using Box-Ljung statistic (Ljung and Box (1978)). Independence of the error terms is also checked in the following manner:

Here the null hypothesis is set as H_0: the errors are distributed randomly. And the alternative hypothesis H_1: the errors are non-random.

The Box-Ljung test statistic, $Q=n(n+2)\sum_{k=1}^{m}\frac{r_k^2}{(n-k)}$,

Where n is the number of observations, r_k is the estimated autocorrelation of the series at lag k = 1, 2… m, m is the number of lags being considered and the statistic Q follows a chi-square distribution with (n-p-q) degrees of freedom.

The test rule is to reject the null hypothesis i.e., the errors are not random if $Q \geq \chi^2_{1-\alpha,\,h}$, otherwise we fail to reject the null hypothesis i.e., errors are considered to be random or independent, if $Q < \chi^2_{1-\alpha,h}$, where, $\chi^2_{1-\alpha,h}$ is the value of the chi-square variable with 'h' degrees of freedom such that P ($\chi^2_h > \chi^2_{1-\alpha,\,h=}$) = 1- α, α being the level of significance.

Here the degrees of freedom, h = (m – p – q); p and q are the numbers of AR and MA terms, respectively.

The normality of the residuals is tested by Shapiro-Wilk's test. After the diagnostic checking of the model and its parameters, the evaluation of the model is made. Among the models satisfying the tests for residual diagnostics, the best fit model is chosen using any one of the criteria like R^2, adjusted R^2, RMSE and mean absolute percentage error (MAPE). The model having lowest value of any of these measures is considered to be the best fit ARIMA model for the given data.

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

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

After exploring the best fit ARIMA model, cross validation is done by obtaining the forecast values of the dependent variable from the fitted model for the time period for which the observations were left out for the validation purpose and not considered for developing the model. From the actual and forecast values of the dependent variable for the time period left out for validation, the absolute percentage error (APE) value is obtained for each observations in the left out 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 the i^{th} year, i= 1, 2, …, 9. Low value of APE ensures the appropriateness of the selected model for forecasting.

After successful cross validation of the selected model, it is used for the purpose of forecasting. The forecast values of area and yield are obtained for three future years along with the 95% confidence intervals using the forecasting tool available in SPSS 20. The forecast value of production of food grains are obtained from the forecast values of area and yield in the following manner:

Forecast value of production (in '000 t) = Forecast value of area ('000 ha) x Forecast value of yield (in kg/ha) /1000

Results and Discussion

The scatter of data on area and yield of kharif food grains in Odisha as shown in Figure 1 and Figure 2, respectively, shows that there is a jump in area and yield after the year 2002-03. So the year 2002-03 is taken as the year of jump or change point and the corresponding value of time variable for this year is t = 11.

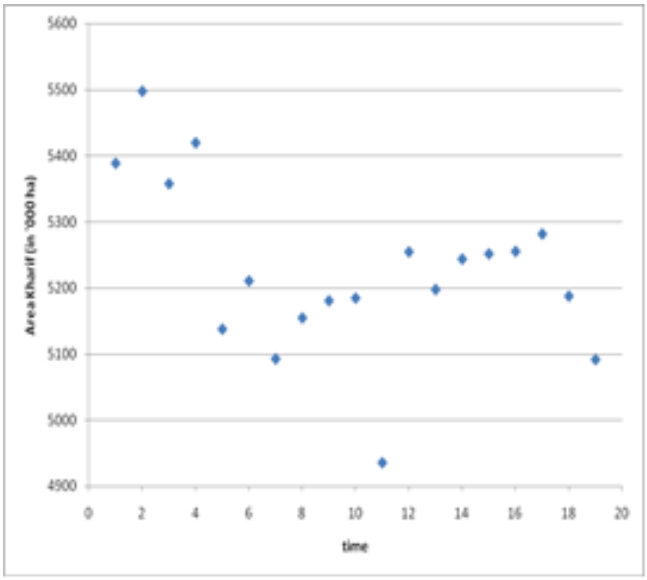

Fig 1: Scatter of area under kharif food grains in Odisha from 1992-93 to 2010-11

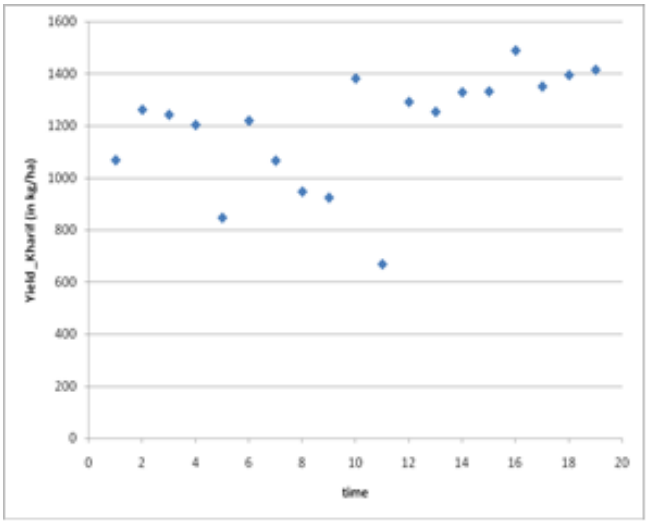

Fig 2: Scatter of yield of kharif food grains in Odisha from 1992-93 to 2010-11

The study of Table 1 shows that among the fitted ordinary regression models, quadratic model has significant F-value (which shows the model is adequate) has all significant estimates of the coefficient parameters, has independent and normally distributed errors (which is evident from insignificant D-W statistic and S-W statistic, respectively), low values of RMSE and MAPE and high value of R^2. Thus, among the ordinary regression model, quadratic model is the best fit for the data on area under kharif food grains.

Table 1: Estimated model parameters, model diagnostics and model selection criteria of ordinary regression models fitted to data on area under kharif food grains of Odisha

	linear	**logarithmic**	**compound**	**power**	**quadratic**
b_0	5327.91**	5418.88**	5325.74**	5418.31**	5483.32**
SE(b_0)	57.11	70.17	58.41	73.19	80.057
b_1	-9.99	-92.22**	0.998**	-0.017**	-54.31**
SE(b_1)	5.009	31.68	0.001	0.006	18.43
b_2					2.21*
SE(b_2)					0.901
D-W Statistic	1.33	1.65	1.33	1.65	1.9
S-W Statistic	0.959	0.947	0.961	0.948	0.924
Coefficient of ln(t)	-0.97	-0.01	-0.75	-0.01	0.24
SE	1.02	0.44	0.67	0.45	0.51
R^2	0.619	0.703	0.684	0.624	0.729
RMSE	119.59	108.54	119.46	108.4	104.29
MAPE	1.72	1.56	1.72	1.56	1.47
F- Value	3.984	8.473**	3.831	8.15**	5.762**

*Significant at 5 % level of significance; ** Significant at 1 % level of significance

The study of Table 2 shows that among the fitted spline regression models, logarithmic spline model has all significant estimates of the coefficient parameter, independent and normally distributed errors (which is evident from insignificant D-W statistic and S-W statistic respectively), low values of RMSE and MAPE and high value of R^2. Quadratic model though havethe highest value of R^2 is not considered to be the best as the estimate of b_1 is not found significant. Thus, among the ordinary regression model, logarithmic spline model is the best fit for the data on area under kharif food grains.

Table 2: Estimated model parameters, model diagnostics and model selection criteria of spline regression models fitted to data on area under kharif food grains of Odisha

	Linear spline	**Logarithmic spline**	**Compound spline**	**Power spline**	**Quadratic spline**
b_0	5485.25**	5515.45**	5486.52**	5089.67**	5509.34*
SE(b_0)	60.691	77.455	60.613	92.682	108.811
b_1	-42.03**	-177.46**	0.99**	-0.03**	-53.15
SE(b_1)	8.948383	44.4499923	0.0017295	0.000215	41.67563
a_1	86.01**	276.22**	1.02**	0.05**	185.26**
SE(a_1)	26.30868	76.9506882	0.00507337	0.000465	58.94824
b_2					0.93
SE(b_2)					3.38
a_2					-17.02**
SE(a_2)					5.647
D-W Statistic	1.321*	1.971	1.332*	1.962	2.534*
S-W Statistic	0.922	0.911	0.923	0.912	0.932
coefficient of ln(t)	1.919	1.094	1.847	1.129	1.778
SE	1.263	0.986	1.271	0.927	1.263
R^2	0.648	0.771	0.749	0.736	0.838
RMSE	117.202	103.103	117.172	103.441	76.369
MAPE	1.681	1.507	1.684	1.526	1.167
F	2.784*	8.272**	2.793*	8.114**	28.241**

*Significant at 5 % level of significance; ** Significant at 1 % level of significance

The study of Table 3 shows that among the fitted ARIMA models, the constant of fitted ARIMA (1, 1, 0) model is insignificant. So, ARIMA (1, 1, 0) model is fiited without constant. The coefficient of AR (1) is significant. The non-significance of D-W Statistic and S-W Statistic indicates that the errors of the fitted ARIMA (1, 1, 0) model without constant are independent and follow normal distribution. The values of RMSE and MAPE for the fitted ARIMA (1, 1, 0) model without constant are low which indicated the apprpriateness of the model in fitting the data on area under kharif food grains. Thus, among the ARIMA group of models, ARIMA (1, 1, 0) model without constant is the best fit for the data on area under kharif food grains.

Table 3: Estimated model parameters, model diagnostics and model selection criteria of selected ARIMA models fitted to data on area under kharif food grains of Odisha

	ARIMA (1,1,0) (with constant)	**ARIMA (1,1,0) (without constant)**
Constant	-40.79	
SE(Constant)	42.22	
a_1	a_1 = -0.584*	a_1 = -0.539*
SE(a_1)	0.201	0.209
D-W Statistic	1.976	1.981
S-W Statistic	0.956	0.964
R^2	0.722	0.729
RMSE	110.92	109.68
MAPE	1.671	1.670

*Significant at 5 % level of significance

The cross-validation of the selected model from each of the three groups of models shown in Table 4 shows that all the three selected models from each group have low values of MAPE. This shows that all are almost efficient for forecasting the area under kharif food grains in Odisha. Since among these three models, the ARIMA (1, 1, 0) model without constant provides the lowest MAPE (2.33) so this model is selected to best model among all the three groups of models fitted to the data on area under kharif food grains. So ARIMA (1, 1, 0) model without constant is used for forecasting of area under kharif food grains in Odisha for the future years 2016-17 to 2018-19.

Table 4: MAPE values for the selected best fit models for area under kharif among the ordinary regression models, spline regression models and ARIMA models

		Forecast values			Absolute Percentage Error		
Year	**Actual value**	**Ordinary regression (quadratic model)**	**Spline regression (logairthmic model)**	**ARIMA (1,1,0 without Constant Model)**	**Ordinary regression (quadratic model)**	**Spline regression (logairthmic model)**	**ARIMA (1,1,0 without constant Model)**
2011-12	4939	5281.12	5306.92	5114.64	6.93	7.45	3.56
2012-13	4930	5317.42	5317.33	5071.82	7.86	7.86	2.88
2013-14	5017	5358.14	5326.74	5062.82	6.8	6.17	0.91
2014-15	5012	5403.28	5335.34	5034.1	7.81	6.45	0.44
2015-16	4829	5452.84	5343.24	5014.53	12.92	10.65	3.84
Mean Absolute Percentage Error					8.46	7.72	2.33

The study of Table 5 shows that among the fitted ordinary regression models, linear model has significant F-value (which shows that the model is adequate), has all significant estimates of parametric coefficients, has independent and normally distributed errors (which is evident from insignificant D-W statistic and S-W statistic respectively), low values of RMSE and MAPE and high value of R^2. Thus among the ordinary regression model, linear model is the best fit for the data on yield of kharif food grains.

Table 5: Estimated model parameters, model diagnostics and model selection criteria of ordinary regression models fitted to data on yield under kharif food grains of Odisha

	Linear	**logairthmic**	**compound**	**Power**	**quadratic**
b_0	1009.38**	1010.99**	1008.03**	1017.86**	1221.88**
SE(b_0)	94.12	136.89	93.14	134.57	140.39
b_1	18.41*	88.14	1.015**	0.068**	-42.31
SE(b_1)	8.26	61.81	0.008	0.06	32.33
b_2	-	-	-	-	3.036
SE(b_2)	-	-	-	-	1.57
D-W Statistic	2.26	1.95	2.27	1.94	2.8*
S-W Statistic	0.943	0.939	0.935	0.942	0.94
coefficient of ln(t)	-0.61	0.07	-0.24	0.33	-0.26
SE	0.633	0.485	0.447	0.376	0.514
R^2	0.726	0.707	0.668	0.671	0.773
RMSE	187.07	211.76	196.45	212.23	182.9
MAPE	13.11	14.82	13.84	15.66	12.72
F	4.974**	2.033	3.443	1.295	4.761**

*Significant at 5 % level of significance; ** Significant at 1 % level of significance

The study of Table 6 shows that among the fitted spline regression models, power spline model has significant F-value (which shows that the model is adequate), has all significant estimates of the parameters, has independent and normally distributed errors (which is evident from insignificant D -W statistic and S -W statistic respectively), low values of RMSE and MAPE and high value of R^2. Quadratic spline model though have the highest value of R^2 and lowest value of R^2, it is not considered to be the best as the parametric estimate b_1 is not significant. Thus among the ordinary regression models, logarithmic spline model is the best fit for the data on yield of kharif food grains.

Table 6: Estimated model parameters model diagnostics and model selection criteria of spline regression models fitted to data on yield of kharif food grains of Odisha

	Linear spline	**Logarithmic spline**	**Compound spline**	**Power spline**	**Quadratic spline**
b_0	1225.84**	1220.39**	1249.35**	1223.13**	1148.08**
SE(b_0)	132.8823	156.5162124	134.245617	156.2206	236.92746
b_1	-25.11	-91.27	0.972**	-0.102**	10.78
SE(b_1)	19.592	89.822	0.019	0.027	90.745
a_1	115.44**	358.75**	1.12**	0.35**	223.82
SE(a_1)	35.431	123.293	0.245	0.033	125.088
b_2					-2.99
SE(b_2)					7.365
a_2					-20.71
SE(a_2)					11.893
D-W Statistic	2.359	2.831*	1.776	2.452	2.825*
S-W Statistic	0.957	0.955	0.955	0.952	0.917
coefficient of ln(t)	1.934	1.524	2.166	1.644	0.969

SE	0.834	0.823	0.778	0.809	1.203
R^2	0.784	0.742	0.352	0.751	0.852
RMSE	187.81	180.09	220.28	194.72	164.54
MAPE	13.211	12.921	14.613	13.732	10.881
F	6.352**	8.309**	0.248	4.794*	13.122**

*Significant at 5 % level of significance; ** Significant at 1 % level of significance

The study of Table 7 shows that among the fitted ARIMA models, the constant of fitted ARIMA (0, 1, 2) model is insignificant. So ARIMA (0, 1, 2) model without constant is fiited. The coefficient of MA (1) and MA (2) are both significant. The non-significance of D-W Statistic and S-W Statistic indicates that the errors of the fitted ARIMA (0, 1, 2) model without constant are independent and follow normal distribution. The values of RMSE and MAPE for the fitted ARIMA (0, 1, 2) model without constant are low, which indicates the apprpriateness of the model in fitting the observed data. Thus among the ARIMA group of models, ARIMA (0, 1, 2) without constant model is the best fit for the data on yield of kharif food grains.

Table 7: Estimated model parameters, model diagnostics and model selection criteria of selected ARIMA models fitted to data on yield of kharif food grains of Odisha

	ARIMA (0,1,2) (with constant)	**ARIMA (0,1,2) (without constant)**
Constant	-57.08	
SE(Constant)	28.73	
b_1	b_1 =1.44	b_1 =1.031**
SE(b_1)	8.49	0.252
b_2	b_2 = -0.444	b_2 = -0.635*
SE(b_2)	3.632	0.314
DW	2.191	2.112
SW	0.977	0.986
R^2	0.843	0.891
RMSE	185.04	191.83
MAPE	13.74	13.75

The cross-validation of the selected model from each of the three groups of models is shown in Table 8. The table shows that all the three selected models from each group have low values of MAPE. This shows that all are almost efficient for forecasting the yield of kharif food grains in Odisha. Since among these three models, the ARIMA (0, 1, 2) model without constant provides the lowest MAPE (15.76) so this model is selected as the best model for kharif food grains. So ARIMA (0, 1, 2) without constant model is used for forecasting of yield of kharif food grains in Odisha for the future years 2016-17 to 2018-19.

Table 8: MAPE values for the selected best fit models for area under kharif among the ordinary regression models, spline regression models and ARIMA models

		Forecast values			**Absolute Percentage Error**		
Year	**Actual value**	**Ordinary regression (linear model)**	**Spline regression (power model)**	**ARIMA (0,1,2 without constant Model)**	**Ordinary regression (linear model)**	**Spline regression (power model)**	**ARIMA (0,1,2 without constant Model)**
2011-12	1264	1327.58	1684.91	1422.72	5.03	33.3	12.56
2012-13	1991	1345.99	1730.04	1394.42	32.4	13.11	29.96
2013-14	1564	1364.4	1771.91	1420.21	12.76	13.29	9.19
2014-15	1951	1382.81	1811.01	1447.17	29.12	7.18	25.82
2015-16	1457	1401.22	1847.75	1475.3	3.83	26.82	1.26
Mean Absolute Percentage Error					16.63	18.74	15.76

Table 9 shows the forecast values of area, yield and production of kharif food grains in odisha by using the best fit model which are ARIMA (1, 1, 0) model without constant in case of area under kharif food grains and ARIMA (0, 1, 2) model without constant in case of yield of kharif food grains. The forecast values of production of kharif food grains is found by using the forecast values of area and yield for the corresponding years. The forecast values are obtained for the years 2016-17, 2017-18 and 2018-19. The forecast vales for these three years show that there is increase in area, yield and thus production of kharif food grains in Odisha.

Table 9: Forecast values of area, yield and production of kharif food grains in Odisha using the best fit model

Year	**Area (in '000 ha)**	**Yield (in kg/ ha)**	**Production (in '000 tonnes)**
2016-17	4944.23	1504.61	7439.14
2017-18	4937.27	1535.09	7579.15
2018-19	4930.57	1566.74	7724.92

Summary and Conclusion

In case of area under kharif food grains, the quadratic model is found to be the best model among the group of ordinary regression models, logarithmic spline models is found to be the best fit model among the group of spline regression models and ARIMA (1,1,0) model without constant is found to be the best fit model among ARIMA models. On comparing these three models, though ARIMA (1,1,0) model without constant has slightly high RMSE and MAPE values than the other two models but the MAPE of the ARIMA (1, 1, 0) model without constant is lowest during cross validation of the models. So, ARIMA (1, 1, 0) model without constant is used for forecasting the area under kharif rice in Odisha for the future years from 2016-17 to 2018-19. But it is seen that spline regression model is an improvement over ordinary regression model for forecasting purposes.

In case of yield of kharif food grains, the linear model is found to be the best model among the group of ordinary regression models, power spline models is found to be the best fit model among the group of spline regression models and ARIMA (0, 1, 2) model without constant is found to be the best fit model among ARIMA models. On comparing these three models, though ARIMA (0, 1, 2) model without constant has slightly high RMSE and MAPE values than the other two models but the MAPE of the ARIMA (0, 1, 2) without constant model is lowest during cross validation of the models. So, ARIMA (0, 1, 2) without constant model is used for forecasting the yield of kharif rice in Odisha for the future years from 2016-17 to 2018-19. But it is also seen in case of yield of kharif food grains that spline regression model is an improvement over ordinary regression model for forecasting purposes.

The forecast values for production of kharif food grains obtained from the forecast of area and yield for the future years shows that production of kharif food grains is likely to increase during these future years which can be ascertained after the availability of the actual data for these years.

References

1. Bajpai PK, Venugopalan R. Forecasting sugarcane production by time series modeling, Indian Journal of Sugarcane Technology. 1996; 11(1):61-65.
2. Borkar P, Tayade PM. Forecasting of cotton production in India using ARIMA models, International Journal of Research in Economics and Social Sciences. 2016; 6(5):1-7.
3. Hanusz Z, Tarasinska J. Tables for Shapiro-Wilk W statistic according to royston approximation, Colloquium Biometricum. 2011; 41:211-219.
4. Lee R, Qian M, Shao Y. On Rotational Robustness of Shapiro-Wilk Type Tests for Multivariate Normality, Open Journal of Statistics. 2014; 4(11):964-969.
5. Ljung GM, Box GEP. On a Measure of a Lack of Fit in Time Series Models, *Biometrika*. 1978; 65(2):297-303.
6. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis, 3rd Edition, New York, John Wiley & Sons, USA, 2001.
7. Nelson C. The prediction performance of the FRB-MIT_PENN model of the US economy, The American Economic Review. 1972; 62(5):902-917.
8. Odisha Agricultural Statistics, 2013. Directorate of Agriculture and Food Production, Government of Odisha
9. Odisha Economic Survey. Planning and Convergence Department, Governnent of Odisha, 2016.
10. Prabhakaran K, Sivapragasam C. Forecasting area and production of rice in India using ARIMA model, International Journal of Farm Sciences. 2014; 4(1):99-106.