International Journal of Current Microbiology and Applied Sciences
ISSN: 2319-7706 **Volume 9 Number 3 (2020)**
Journal homepage: http://www.ijcmas.com

Original Research Article https://doi.org/10.20546/ijcmas.2020.903.051

An Appropriate Model to Fit the Production of Rice and Wheat Data for India

Bhola Nath, D. S. Dhakre, K. A. Sarkar and D. Bhattacharya*

Visva-Bharati, Santiniketan, India

**Corresponding author*

Keywords

Assumptions, exponential fitting, MAPE, nonparametric regression, normal distribution, parametric regression.

Article Info

Accepted:
05 February 2020
Available Online:
10 March 2020

ABSTRACT

Fitting of an appropriate model to an observed time series data for the purpose of predicting the future values efficiently is always a challenging task. The practitioners of statistics in their first attempt always try to fit parametric regression model to the data. For all parametric models to be fitted, it is assumed that the model errors follow independent normal distributions. If that assumption on error distribution is not satisfied, then we should search for an alternative procedure. Here, we propose the nonparametric regression procedure as the alternative procedure and study its performance. In the present investigation the secondary data on production of rice crop forthe *Kharif* season and production of wheat for *Rabi* season for India as a whole for 51 years (1962-63 to 2012-13) have been used. It has been observed that the variable, production of rice,does not satisfy the assumption of normal distribution of errors but the variable, production of wheat satisfies the assumption of normality of error distribution. Here we have applied Parametric and nonparametric regression approaches to both the data sets. It has been found that there is a great reduction in the value of Mean Absolute Percentage Error (MAPE) of prediction for the dependent variable production of rice when nonparametric regression is used. It is concluded that the nonparametric regression works well for the data set for which the normality assumption of the error distribution does not hold and gives better prediction than the usual parametric regression.

Introduction

For growing population of India, it is an interesting problem to examine the growth rate and instability in the production of different crops, say, for example, paddy. If population growth rate is much higher than the growth rate of paddy, then we should look for technologies that can increase the yield of paddy. If the growth rate of paddy is more than the population growth rate, then we can earn some foreign currency by exporting the excess production of rice to the foreign countries.

Growth rate of a certain variable is defined as the percentage change of that variable within a specific time period. In the field of agriculture, the study of growth rates has enough importance and widely used in planning as these have important policy implications. The casual statements about these growth rates as falling or rising or constant may lead to some wrong decisions. The decision taken on the nature of ups and downs in growth rates should be based on fitting of models and by examining the real situation. The compound growth rates can be computed by fitting the exponential function as below:

$$y_t = y_0(1+r)^t, \qquad (1)$$

where y_t= dependent variable like area, production, productivity for the year '*t*'; y_0= the value of the variable *y* at the beginning of the time period; *t* = time element, *t* =1, 2, …, *n* and *r*= compound growth rate. Most commonly used model for computing growth rate in agriculture is given in the equation (1). Estimates of the parameters are obtained using the method of least squares. After logarithmic transformation, equation (1) becomes:

$$\log y_t = \log y_0 + \log(1+r) \times t.$$

Thus, the compound growth rate (*r*) is estimated by

$$\log(1+\hat{r}) = \hat{b}$$

or

$$\hat{r} = \exp(\hat{b}) - 1, \qquad (2)$$

where $\hat{b}$ is the least square estimate *b* in the linearized model, $y_t^* = a + bt$, where $y_t^* = \log y_t$, $a = \log y_0$, $b = \log(1+r)$.

The instability in the variable under study can be measured by the co-efficient of variation (C.V.) of that variable:

$$C.V. = \left(\frac{standard\ deviation}{Mean}\right) \times 100 \qquad (3)$$

Several authors have worked in the area of study of instability and growth of a particular variable like area, production or productivity of a crop over a period of time. Mention may be made of Dash *et al.*, (2017(a)). The said paper has widely studied the growth and instability in pulse production of Odisha state. They used the secondary data associated with area, production and productivity of the pulses in the state of Odisha over the period of (1970-2014) and divided the whole time period into two periods in reference to some economic reforms viz., pre-reform period and post-reform period. The work focuses on the comparison of the efficiency of different models fitted to the data such as linear model, compound model, quadratic model.

Dhakre & Bhattacharya (2013) analyzed the growth and instability in the production of vegetables in the state of West Bengal by fitting an exponential model for variables like area, production and productivity. They have also estimated the parameters using ordinary least square techniques and estimated the growth rate and tested its significance using appropriate test statistic.

Bhattacharya & Roychowdhury (2017) discussed the necessity of the nonparametric regression model when the errors in the linear regression model do not satisfy the necessary assumptions required by the linear regression to be satisfied. In such cases if we use linear regression, then we may get a very poor result. For those cases nonparametric regression works pretty well. The work also discussed about testing the significance of the regression parameters by Spearman's rank correlation.

Dash *et al.*, (2017 (b)) proposed the appropriate model for studying the growth rate and instability of mango production in India. In this research they have used spline model and discussed about the appropriateness of the spline model with the help of different evaluation criteria, such as, R^2, Adjusted-R^2 and root mean square error (RMSE). They have found that compound model with spline, compound model without spline and linear model with spline best fits the observed data on production, area and productivity of Odisha, respectively. Dash *et al.*, (2017(b)) also studied the growth and instability in the food grain production of Odisha by using the time series model fitted over the period of 1970-2014. Coefficient of variation was used for the instability in area, production and productivity for the total food grain production.

Estimation and test of significance for growth rate and instability in production

The relative rate of change (r_t) in variable y between periods ($t-1$) and t is defined as:

$$\frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1 = r_t, \quad t = 1,2,\dots,n.$$

The average of relative growth rate (*AGR*), $\bar{r}_t$, can be obtained by taking the arithmetic mean of the relative rate of changes.

Let, y_0 denote the value of y at the beginning of the time period and $y_t$denote the value of y at the end of period t, $t = 1,2,\dots,n$. The growth relatives are defined as follows:

Now the ratios:

$\frac{y_t}{y_{t-1}} = r_t, t = 1,2,\dots,n$, are called growth relatives.

To find the average of growth relatives we should not use arithmetic mean but use geometric mean of the growth relatives so that the correct picture of average growth rate is captured.

Thus, average of growth relatives of *n* time period $= \left(\frac{y_1}{y_0} \times \frac{y_2}{y_1} \dots \times \frac{y_n}{y_{n-1}}\right)^{\frac{1}{n}} = \left(\frac{y_n}{y_0}\right)^{\frac{1}{n}}$

Next, the estimate of *r* in equation (1) can be used as: $r_t = \left(\frac{y_t}{y_0}\right)^{\frac{1}{t}} - 1$.Next we discuss how to predict the average growth rate.

Using the fitted regression model, the estimated value($\hat{y}_t$) of the dependent variable (production) are obtained. Next, using the predicted values of production, the estimated annual growth rates can be obtained in the following ways:

Estimated Annual Growth Rate for the year *t*, $\widehat{AGR}_t = \left(\frac{\hat{y}_t - \hat{y}_{t-1}}{\hat{y}_{t-1}}\right) \times 100$, $t = 1,2,\dots,n$,

where $\hat{y}_t$ and$\hat{y}_{t-1}$ are the predicted values of the variable *y* at time *t* and (*t*-1), respectively.

Estimated average growth rates for the whole period of study is obtained by taking arithmetic mean of the annual growth rates of the respective periods.

Thus, estimate of the 'Average Annual Growth Rate' is obtained as, $\widehat{AAGR} = \frac{1}{n}\sum_{t=1}^{n} \widehat{AGR}_t$.

The significance of the average annual growth rate in the population is tested by using student's *t*- statistic.

The null hypothesis is taken as H_0: population *AAGR* = 0, which is tested against the alternative hypothesis H_1: population $AAGR \neq 0$ at 1% level of significance.

The test statistic used is,

$t = \frac{\widehat{AAGR}}{Se(\widehat{AAGR})}$,where, $Se(\widehat{AAGR}) = \frac{Sd(\widehat{AAGR})}{\sqrt{n}}$.

Here the statistic *t* follows a *t* distribution with (*n* – 1) *d.f.*, where *n* is the number of observations. (Dash *et al.*, (2017) (c))

Measuring instability in production

The coefficient of variation (*C.V.*) is used to measure the instability in production. To eliminate the effect of trend in the calculated *C.V.*, it is estimated from the detrended values. For linear trend, where the effects of different components are assumed to be additive in nature, the detrended values are obtained by subtracting the predicted values from the actual values obtained from the best fitted model.

Thus, detrended value is $y_d = y_t - \hat{y}_t$(assuming an additive model),

Where y_t is the actual value of the variable at time *t* and $\hat{y}_t$ is the predicted value obtained from the fitted model.

Centering of detrended values are done by adding the mean of the actual values$\bar{y}_t$and the detrended values (y_d's) are obtained accordingly by using the additive model. The *C.V.* is found from these detrended and centered values, as sample *C.V.* is defined as, $CV = \frac{s_{y_d}}{\bar{y}_d} \times 100$,

wheres_{y_d} is the standard deviation of the detrended values (y_d) of *y* and$\bar{y}_d$is the mean of$y_d$values.

The significance of the coefficient of variation is tested by using student's *t*-statistic. The null hypothesis for the test is taken as H_o: population *CV* = 0 which is tested against the alternative hypothesis H_1:population *CV*> 0.

Here the test statistic, $t = \frac{CV}{Se(CV)}$, which follows a*t* distribution with (*n*-1) degrees of freedom, where *n* is the number of observations and $s_e(CV)$ is the standard error of the coefficient of variation which is given by, $s_e(CV) = \frac{CV}{\sqrt{2n}}$ (Koopmans *et al.*, (1964)), $\widehat{CV}$ is the estimated *CV*obtained from the sample data.

Table.1 Test of significance of population average annual growth rate (*AAGR*) and co-efficient of variation (*CV*)

Table: 1Test of significance of population average annual growth rate (*AAGR*) and coefficient of variation (*CV*)				
Dependent variable	$\widehat{AAGR}$ **(%)**	***t*-value**	***CV* (%)**	***t*-value**
Production of rice (*Kharif*)	2.52	18.77** (2.68)	8.27	10.10** (2.68)
Production of wheat (*Rabi*)	4.78	9.24** (2.68)	7.54	10.10** (2.68)

Note: Figures in the parentheses are the tabulated values of '*t*' and ** denotes that the value is significantly different from 0 at 1% level of significance.

Regression Approach to model fitting

Parametric Regression

It is an approach to modeling the relationship (linear or nonlinear) between one dependent variable and one or more independent variables using a functional relationship. Linear regression model in one variable is given by:

$$y = \alpha + \beta x + \epsilon$$

where, y is the dependent variable, α is the intercept, β is the regression coefficient, x is the independent variable and ϵ is the random error. The estimates of α and β can be obtained by the following formulae:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The formula to calculate R^2 and adjusted R^2 are given as:

$$R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}$$

and

$$adjR^2 = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

where SS_{reg}= regression sum of squares, SS_{total}=total sum of squares, $\hat{y}_i$ =estimated value of the dependent variable y, y_i=ith observed value of y, n= number of observations, k= number of regressors. Instead of fitting exponential model as given in (1) other parametric regression models can also be tried. Further, it is to be noted that the errors associated with the parametric regression model should satisfy some assumptions. If those model assumptions are not satisfied, then nonparametric regression approach is adopted for the growth rate studies.

Nonparametric regression

The model for nonparametric regression fitting is given by,

$$y_t = m_t + \varepsilon_t, \quad i = 1,2,\dots,n, \qquad (4)$$

where y_t is the observation value at tth time point, m_t is the trend function which is assumed to be smooth and ε_t'sare random error with mean zero and finite variance. Since there is no assumption of parametric form of function m_t, this approach is flexible and robust to deviations from any particular form of the assumed model.

Generally, the parametric approach uses transformations like logarithmic or so in order to stabilize variance or linearize the relationship but in nonparametric approach there is no need of any such transformation.

Theil's method for estimating slope and intercept in nonparametric regression

Without loss of generality, let us assume that for the data set $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

$$X_1 < X_2 < \dots < X_n$$

We compute S_{ij} for all $\binom{n}{2}$ possible combinations of (i, j), $i = 1, 2, \dots, n$, $j = 1, 2, \dots, n$, $i<j$, as follows:

$$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i} (i < j)$$

The estimates of β and α are:

$$\hat{\beta} = median\{S_{ij}\}$$

$$\hat{\alpha} = median\{Y_i\} - \hat{\beta} \times median\{X_i\}$$

Analysis of data

Data on two dependent variables viz., production of rice (*Kharif*) and production of wheat (*Rabi*);(all are in thousand tonne) for the period of 1962-63 to 2012-13have been used for analysis. Two separate regressions considering production of rice and production of wheat as respective dependent variables and time as the independent variable have been carried out. It has been found by Q-Q plot and Shapiro- Wilk's test of errors for the regression that the dependent variable production of rice does not follow normality assumption but the dependent variable production of wheat follows the normality assumption of the error. The data set used is given in the Table-2.

Table.2 Data set on production of rice for *Kharif* and *Rabi* seasons and production of wheat for the period of 1962-63 to 2012-13

Year	Production of rice (*Kharif*)	Production of wheat (*Rabi*)
1962-1963	32340	10776
1963-1964	36175	9853
1964-1965	38388	12257
1965-1966	29429	10394
1966-1967	28622	11393
1967-1968	35313	16540
1968-1969	37127	18651
1969-1970	37591	20093
1970-1971	39559	23832
1971-1972	39992	26410
1972-1973	36324	24735
1973-1974	40904	21778
1974-1975	35926	24104
1975-1976	44745	28846
1976-1977	39266	29010
1977-1978	48947	31749
1978-1979	49337	35508
1979-1980	38486	31830
1980-1981	50089	36313
1981-1982	49245	37452
1982-1983	43164	42794
1983-1984	55052	45476
1984-1985	53782	44069
1985-1986	59392	47052
1986-1987	53561	44323
1987-1988	48763	45096
1988-1989	63376	54110

1989-1990	65878	49850
1990-1991	66317	55135
1991-1992	66368	55690
1992-1993	65243	57210
1993-1994	70723	59840
1994-1995	72603	65767
1995-1996	67879	62097
1996-1997	71323	69350
1997-1998	71571	66345
1998-1999	71092	71288
1999-2000	77480	76369
2000-2001	72778	69681
2001-2002	80522	72766
2002-2003	63084	65096
2003-2004	78619	72156
2004-2005	72230	68637
2005-2006	78272	69355
2006-2007	80171	75807
2007-2008	82703	78570
2008-2009	84951	80679
2009-2010	75959	80804
2010-2011	80607	86874
2011-2012	92738	94882
2012-2013	92368	93506

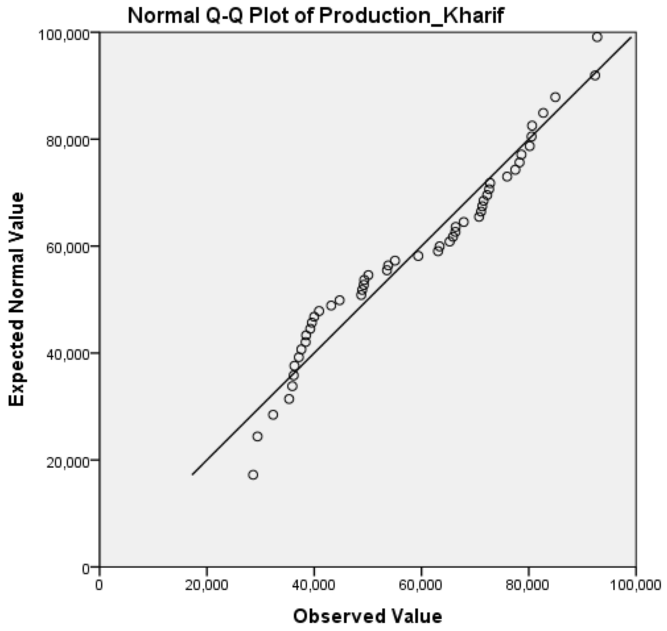

Figure.1 Q-Q plot for the dependent variable production of rice (*Kharif*)

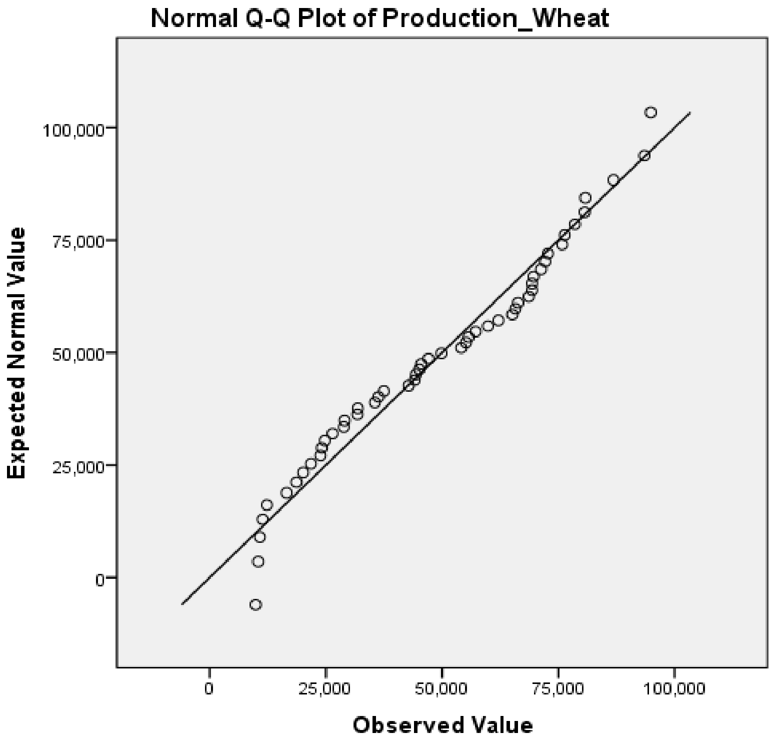

Figure.2 Q-Q plot for the dependent variable production of wheat (*Rabi*)

The graphical representation of the normality check has been done by Q-Q plot for the same data set and are given in the Figures 1 and 2; Figure:1 for the production of rice (*Kharif*) and Figure:2 for the production of wheat (*Rabi*).

Next, a formal test the Shapiro-Wilk's test is one of the most popular tests for normality assumption of normality is done. For the detail about Shapiro-Wilk test statistic and its derivations Shapiro and Wilk (1965) is referred to. The form of the test statistic is

$$W = \frac{\left(\sum a_i y_{(i)}\right)^2}{\sum (y_i - \bar{y})^2},$$

where $y_{(i)}$ is the i^{th} ordered statistic and a_i is the expected value of the i^{th} normalized order statistics. For independently and identically distributed observations, the values of a_i can be obtained from the table presented by Shapiro and Wilk (1965) for sample sizes up to 50. *W* can be expressed as a square of the correlation coefficient between a_i and $y_{(i)}$. So, *W* is location and scale invariant and is always less than or equal to 1. In the plot of $y_{(i)}$ against a_i an exact straight line would lead to *W* very close to 1.

So, if *W* is significantly less than 1, the hypothesis of normality will be rejected. Although the Shapiro-Wilk's test is very popular, it depends on availability of values of a_i, and for large sample cases their computation may be much more complicated.

Some minor modifications to the *W* test have been suggested by Shapiro and Francia (1972), Weisberg and Bingham (1975) and Royston (1982). An alternative test of the same nature for samples larger than 50 is designed by D'Agostino (1971).

Table.3 Test for normality by Shapiro-Wilk's test

Variable	Statistic	*df*	*p*-value
Production of rice (*Kharif*)	0.941	51	0.014
Production wheat (*Rabi*)	0.956	51	0.057

From the Table-3it is clear that the dependent variable production of rice (*Kharif*) does not satisfy the assumption of the normality but production of wheat (*Rabi*) does satisfy the normality assumption. This implies that the variable production of rice (*Kharif*) is non-normal and the variable production of wheat (*Rabi*) is normal. Next, we apply both the regression approaches (parametric and nonparametric) separately for the given two data sets to test which one fits the data better.

As we have the data set for two dependent variables under consideration and out of those two variables one dependent variable i.e., production of rice (*Kharif*) contains the non-normality of the error distributions and the other dependent variable i.e., production of wheat (*Rabi*) satisfies the normality of the error distribution. Here, we have applied both parametric, and nonparametric regression approaches to both the data sets and compared the MAPE's for prediction purpose. We have computed the model fit criteria like: R^2 and Adjusted $R^2$for comparison and the results are given inTable-4.

In the process of fitting the regression models we have used 41 observations to build up the model and remaining 10 observations were kept for cross validation of the fitted regression models. When we are concerned with the prediction of the dependent variables, it becomes a necessary task to evaluate the accuracy of the predicted results.

Table.4 Comparison of the results for both the dependent variables

Sl. No.	Dependent Variable	Criteria	Parametric Regression	Nonparametric Regression
1.	**Production of rice (*Kharif*)** (non-normality of error distribution prevails)	**R^2**	0.9312	**0.9265**
		Adj. R^2	0.9298	**0.8865**
2.	**Production of wheat (*Rabi*)** (normality of errors distribution prevails)	**R^2**	**0.9785**	0.9739
		Adj. R^2	**0.9780**	0.9339

To meet this requirement, we have cross validated the predicted values using the data set which we have not used for the model building purpose. The comparison of the said values is given in the Table-5. It can clearly be observed from the Table-5 that the nonparametric regression can be preferred over the parametric regression approach in case where there is a violation in the normality assumption of the errors for the concerned variable.

Table.5 Summary of cross validation (MAPE)

Sl. No.	Dependent Variable	MAPE	
		Parametric	Nonparametric
1.	**Production of rice (*Kharif*)**	24.2219	**4.8351**
2.	**Production of wheat (*Rabi*)**	**6.6985**	8.2215

Here our main aim is prediction, so comparison of MAPE values is more important than that of comparing the values of other model fit criteria for the purpose of selecting the best approach and the best fitted model. It is observed that MAPE significantly reduces if nonparametric regression approach is used when the error distribution is non-normal.

As we have found earlier inTable-2that the variable production of rice (*Kharif*) does not satisfy the normality assumptions of the error distribution. The dependent variable for which the normality assumption of the error distribution holds i.e., production of wheat (*Rabi*) the parametric regression works well. Thus, it can be concluded that in cases where dependent variable encounters the issue of non-normally distributed errors, nonparametric regression is preferred over the parametric regression for better prediction result.

Findings in Table 4 reveal that in case of nonparametric regression though the value of R^2 is little less but MAPE value is better than that of the parametric regression for the variable production of rice (*Kharif*). Since in the present investigation we are concerned with the prediction of the variables under study, so here we compare the MAPE values in the cross validation of the concerned dependent variable. In case of production of rice (*Kharif*) which does not satisfy the normality assumption of error, the better prediction result can be achieved if nonparametric regression is used. On the other hand, the parametric regression may be of efficient for the production of wheat (*Rabi*), as reflected in R^2 and Adj.$R^2$and prediction (MAPE) values.

References

Bhattacharya, D. & Roychowdhury, S. (2017). *Nonparametric Statistical Methods*, Medtech: A Division of Scientific International.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large sample sizes, *Biometrika*, 58(August): 341-348.

Dash, A., Dhakre, D.S. & Bhattacharya, D. (2017(a)). Analysis of Growth and Instability in Pulse Production of Odisha during *Rabi* Session: A Statistical Modelling Approach, *International journal of current microbiology and applied sciences, 6* (8), 107-115.

Dash, A., Dhakre, D.S. & Bhattacharya, D. (2017(b)). Fitting of appropriate model to study growth rate and instability of mango production in India, *Agricultural Science Digest, 37* (3), 191-196.

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

Dhakre, D.S. & Bhattacharya, D. (2013). Growth and Instability Analysis of Vegetables in West Bengal, India, *International journal of Bio-resource and Stress Management, 4* (3), 456-459.

Koopmans, L. H., Owen, D. B. & Rosenblatt,

J. I. (1964). Confidence intervals for the coefficient of variation for the normal and lognormal distributions, *Biometrika*, *51*, 25-32.

Royston, J. P. (1982). An extension of Shapiro-Wilk's W test for non-normality to large samples, *Applied Statistics*,31: 115-124.

Shapiro, S. S. & Francia, R. S. (1972). An approximate analysis of variance test for normality, *Journal of the American Statistical Association*,67(337): 215-216.

Weisberg, S., & Bingham, C. (1975). An approximate analysis of variance test for non-normality suitable for machine calculation, *Technometrics*,17(1): 133-134.

How to cite this article:

Bhola Nath, D. S. Dhakre, K. A. Sarkar and Bhattacharya. D. 2020. An Appropriate Model to Fit the Production of Rice and Wheat Data for India. *Int.J.Curr.Microbiol.App.Sci.* 9(03): 439-449. doi: https://doi.org/10.20546/ijcmas.2020.903.051