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

E-ISSN: 2278-4136
P-ISSN: 2349-8234
JPP 2017; 6(5): 2495-2499
Received: 24-07-2017
Accepted: 25-08-2017

A Dash
College of Agriculture, Chiplima (OUAT), Odisha, India

DS Dhakre
Institute of Agriculture, PSB, Visva-bharati, Santiniketan, West Bengal, India

D Bhattacharya
Institute of Agriculture, PSB, Visva-bharati, Santiniketan, West Bengal, India

Correspondence
A Dash
College of Agriculture, Chiplima (OUAT), Odisha, India

Fitting of appropriate statistical model for study of growth and instability in cereal production of Odisha

A Dash, DS Dhakre and D Bhattacharya

Abstract
Cereals are the most important kharif season crops of odisha. Agriculture of Odisha is affected by the economic reforms in the 1991-92, following which our study period (1970-71 to 2013-14) has been divided into pre-reform (1970-71 to 1991-92) and post reform (1992-93 to 2013-14) period. Attempts have been taken to make a comparative study of the growth rate and instability of area, production and yield of rice in the two periods. For studying the growth rate and instability, appropriate model which could best possibly describe the behavior of the phenomenon are fitted by applying spline regression technique with kink or knot placed at the year of transition from pre–reform period to post- reform period which is considered to be 1991-92. Then by using the best fit model, identified with the help of residual diagnostics and model fit statistics, the average growth rates of area, production and yield are found. Coefficient of variation is used as a measure of instability. The difference in growth rates from the pre-reform to post–reform period is found to be negative but non-significant in case of area, negative and highly significant in case of production and significantly positive in case of yield. The difference in coefficient of variation between the two periods is highly significant and positive only for yield whereas, for both area and production, it is negative, but significant only in case of area.

Keywords: Spline regression, kink, economic reforms, growth rate, instability, coefficient of variation

1. Introduction
Agriculture is the backbone of rural economy and livelihood of Odisha. Cereals are the most important food grain crops of odisha. Rice is the most important cereal crop of Odisha. Other cereal crops grown in odisha are maize, ragi, wheat and small millets. They are mainly the kharif season crops of Odisha. Cereals share 74 per cent and 13 per cent of the total cropped area in the state, in kharif and rabi season respectively They share nearly 86 per cent and 19 percent of total area under food grains in kharif and rabi respectively. (Odisha Agricultural Statistics).
The economic reformation in 1991-92, is expected to have substantial effects in Indian agriculture which in turn affect the state agriculture. As economic reform is said to have brought about a clear shift in the focus on growth strategy, it may be useful to analyze the scenario of agriculture in the state by comparing the growth rate and instability in area, yield and production of cereals in pre-reform and post-reform period.
To describe the behavior of data and to take care of any significant jumps in the data over a long period of time, there is need for fitting of spline regression models. Keeping these perspectives in mind, the study has been made with objective of exploring appropriate model that best fits the area, production and yield of cereals, in Odisha, and making a comparative study of growth rate and instability in pre-reform and post reform period.

2. Materials and Methods
2.1 Period of Study
The study pertains to area, production and yield of cereals in the state of Odisha from the year 1970-71 to 2013-14. The entire study period is divided into two periods – Pre-reform period (1970-71 to 1991-92) and Post-reform period (1992-93 to 2013-14). Pre- reform period is referred to as Period I and Post-reform period is referred to as Period II.

2.2 Sources of Data
The analysis is based on secondary source data relating to the area, production and yield of cerealsin Odisha for the period from 1970-71 to 2013-14. The data are obtained from various volumes of Odisha Agricultural Statistics published by the Directorate of Economics and Statistics, Government of Odisha. The area, production and yield are expressed in '000 ha, '000 MT and kg ha^{-1} respectively. (1 ha = 10000 m^2, 1 MT = 1000 kg).

2.3 Analytical Techniques

Fitting of appropriate model to the data

A model is an equation or a set of equations which represents the behavior of the system. Models that are used to describe the behavior of the variables that vary with respect to time are called the growth models.

Fitting a single regression line for a long period to estimate the growth, may be misleading due to significant change (i.e. jumps or breaks) in direction of the growth path, which would make the estimate biased. To avoid this situation, the models are fitted by spline regression technique. A spline regression model avoids the inappropriate "jump" (i.e., break) in the data for a long period of time by placing a kink in the line at the point of jump without allowing a break in the line.

Spline regression models are fitted with the help of dummy variables. Since the whole period of study (1970-71 to 2013-14) is divided into two periods – Pre-reform period (190-71 to 1991-92) and post-reform period (1992-93 to 2013-14) basing on the economic liberalization in 1991-92, the spline or kinked models have been fitted with one break point (i.e. kink) which is considered at the year 1991-92. The time variable for the year 1970-71 to 2013-14 is referred to as t = 1,2,3,…..22,23,24……43,44 i.e., t =1 for the year 1970-71, t = 2 for the year 1971-72, ……., t = 22 for the year 1991-92,……, t = 44 for the year 2013-14.

The time variable for the year in which kink is placed (i.e. 1991-92) is referred as k. Thus k = 22.

The kinked models are obtained with the help of dummy variables (Paltasingh, R.K., 2013) [3].

The kinked linear model, power model, compound model and quadratic model are obtained as:

Linear model: $Y_t = \beta_0 + \beta_1.\ t.\ I_{(1 \le t \le 22)} + \{\beta_1.\ t + A_1\ (t-k)\}.\ I_{(23 \le t \le 44)} + \varepsilon_t$

Power model: $Y_t = \beta_0.\ t^{\beta_1}.\ I_{(1 \le t \le 22)}\ \{t^{\beta_1}.\ (t-k)^{A_1}\}.\ I_{(23 \le t \le 44)}.\ \text{Exp}\ (\varepsilon_t)$

The power model obtained is transformed to linear model by natural log transformation as,

$\ln Y_t = \ln \beta_0 + \beta_1.\ \ln t.\ I_{(1 \le t \le 22)} + \{\beta_1.\ln t + A_1 \ln (t-k)\}.\ I_{(23 \le t \le 44)} + \varepsilon_t$

Compound model: $Y_t = \beta_0.\ \beta_1^{\ t}.\ I_{(1 \le t \le 22)}.\ \{\beta_1^{\ t}.\ A_1^{\ (t-k)}\}.\ I_{(23 \le t \le 44)}.\ \exp\ (\varepsilon_t)$

The compound model obtained is transformed to linear model by natural log transformation as,

$\ln Y_t = \ln \beta_0 + t.\ \ln \beta_1.\ I_{(1 \le t \le 22)} + \{t.\ \ln \beta_1 + (t-k) \ln A_1\} I_{(23 \le t \le 44)} + \varepsilon_t,$

Quadratic model: $Y_t = \beta_0 + \{\beta_1.t + \beta_2.t^2\}\ I_{(1 \le t \le 22)} + \{\beta_1.t + A_1\ (t-k) + \beta_2.t^2 + A_2\ (t-k)^2\}\ I_{(23 \le t \le 44)} + \varepsilon_t$

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

Using Ordinary Least Square technique, the estimated values of the coefficients β_0, β_1 and A_1 are found out. The estimated values of β_0, β_1 β_2, A_1 and A_2 are written as b_0, b_1, b_2 and a_1, a_2 respectively.

The significance of the estimated coefficient is tested by applying t test statistic

$t = \frac{\beta_i}{S.E.(\beta_i)}$, which follows't' distribution with n – p degrees of freedom, n is the number of observations.

The overall significance of the model is tested by applying F statistic

$F = \frac{MSM}{MSE}$, which follows F distribution with (p-1, n-p) degrees of freedom

MSM is the mean square of the model, MSE is the error mean square;

$MSM = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{p-1}$, $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}$; n is the number of observations and p is the number of parameters involved in the model.

Assumptions in the model are:

(iv) Errors should be independent

(v) Error should have zero mean.

(vi) Errors should have constant variance i.e. errors should be homoscedastic.

(vii) Errors must follow normal distribution

The assumptions regarding the errors are tested by using.

(i) Durbin-Watson test for testing independence of residuals

(ii) Park's test for testing homoscedasticity of residuals.

(iii) 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) [2].

Durbin-Watson test statistic i.e., D-W statistic, $d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}\varepsilon_t^2}$, where, $\varepsilon_t = y_t - \hat{y}_t$, y_t and $\hat{y}_t$ are respectively the actual and estimated values of the response variable in time t.

The value of 'd' ranges from 0 to 4. Upper and lower critical values, dU and dL have been tabulated for different values of k (no. of explanatory variables) and n (no. of observations) for corresponding level of significance (α) in the Durbin – Watson statistical table.

If $d < d_L$, it is significant. If $d > d_U$, then it is insignificant and the residuals are 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) Park's test: In this test, natural logarithm of the residual (ε_t) is regressed with natural logarithm of the independent variable (which is time, t) by fitting linear regression, i.e, $\ln (\varepsilon_t) = a + b \ln (t)$. If the slope of the regression coefficient, b is found to be non-significant, then it is concluded that residuals are homoscedastic (i.e. constant error variance), otherwise, residuals are heteroschedastic (error variance not remaining constant). (Gujarati, D.N., 2003) [1].

(iii) Shapiro-Wilk's test: Shapiro-Wilk's test statistic i.e., S-W test staitistic, $w = s^2 / b$

where, $s^2 = \Sigma\ a(k)\ \{x_{(n+1-k)} - x_{(k)}\}$; $b = \sum_{t=1}^{n}(y_t - \bar{y})^2$ (Thode Jr., H.C., 2012) [5].

The parameter k takes the values 1,2,….,n/2, when n is even and 1,2,.....,(n-1)/2, when n is odd, n is the number of observations. $x_{(k)}$ is the k^{th} order statistic of the set of residuals.

The values of coefficients a(k) for different values of n and k are obtained from the table of Shapiro-Wilk. If w is non-significant, then the residuals are normally distributed.

The model fit statistics, viz, R^2, adjusted R^2 and RMSE (Root mean Square Error) are also computed. Among the models fitted for the dependent variable, which satisfy the error assumptions and show overall significance and significant parameter estimates, the one having highest adjusted R^2 and lowest RMSE is considered to be the best fit model for that variable.

$R^2 = \frac{SSM}{SSE}$, where, SSM is the sum of square due to model; SSE is the sum of square due to error.

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

Adjusted $R^2 = 1 - (1 - R^2) \times \frac{n-1}{n-p}$; RMSE (Root Mean Square Error) $= \left\{\frac{\sum_{i=1}^{n}(y_t - \hat{y}_t)^2}{n-p}\right\}^{1/2}$

where 'n' is the number of observations ; p is the no. of parameters involved in the model.

Estimation of growth rates

Using this best fit model, the estimated/predicted values ($\hat{Y}_t$) of the dependent variable (Area/Production/Yield) are found for the period I, period II and the whole period. By using the predicted values, the annual growth rates are found.

Annual Growth Rate for the year t, $AGR_t = (\frac{\hat{Y}_t - \hat{Y}_{t-1}}{\hat{Y}_{t-1}})$ x 100

Average Growth rate for the period I (1970-71 to 1991-92), period II (1992-93 to 2013-14) and the whole period is obtained by taking arithmetic mean of the annual growth rates of the respective periods. (Prajneshu, 2005) [4].

Also the difference in average growth rates of period I and period II is obtained as,

$\Delta GR = GR_2 - GR_1$.

Study of instability in area/production/yield

Coefficient of variation is used as a measure of instability. The simple c.v. often contains the trend component and thus overestimates the level of instability in time series data characterized by long term trends. So to eliminate the effect of trend, C.V. is estimated from the detrended values (i.e., trend eliminated values).

For linear and quadratic model, where the effects are assumed to be additive in nature, the detrended values are obtained by subtracting the predicted values (obtained from the model) from the actual values.

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

where, y_t is the actual value of the variable in time t.

$\hat{y}_t$ is the predicted value obtained from the best fit model.

For power, compound and logarithmic model, where the effects are assumed to be multiplicative in nature, the detrended values are obtained by dividing the actual values with predicted values (obtained from the model).

Thus, detrended value, $y_D = \frac{y_t}{\hat{y}_t}$ (assuming multiplicative model)

where, y_t is the actual value of the variable in time t.

$\hat{y}_t$ is the predicted value obtained from the best fit model.

The detrended values are then centered by adding the mean of the actual values ($\bar{y}_t$)

The C.V. is found from these detrended and centered values.

$CV = \frac{\sigma_{yD}}{\bar{y}_D} \times 100$ C.V. is found for period I, period II and for the whole period. Also the difference in CV (ΔCV) between period I and period II is found.

3. Result and discussion

Table 1 shows the parameter estimates, residual diagnostics and model fit statistics of the fitted models for area under cereals. From the table it is found that among all the models fiited for area under cereals,quadratic model is the only model in which all the parameter estimates are significant, all the assumpptions regarding the error are satisfied and has the highest adjusted R^2 and lowest RMSE than other fitted models. So, quadratic model is found to be the best fit model for area under cereals in Odisha.

Table 2 shows the parameter estimates, residual diagnostics and model fit statistics of the fitted models for production under cereals. From the table it is found that among all the models fiited for production under cereals, power model is the only model in which all the parameter estimates are significant, all the assumpptions regarding the error are satisfied and has the highest adjusted R^2 and lowest RMSE than other fitted models. So, power model is found to be the best fit model for area under cereals in Odisha.

Table 3 shows the parameter estimates, residual diagnostics and model fit statistics of the fitted models for yield under cereals. From the table it is found that among all the models fiited for yield under cereals, quadratic model is the only model in which all the parameter estimates are significant, all the assumpptions regarding the error are satisfied and has the highest adjusted R^2 and lowest RMSE than other fitted models. So, quadratic model is found to be the best fit model for area under cereals in Odisha.

Table 4 shows the average growth rates and coefficient of variation of area, producftion and yield of cereals in period I, period II and whole period. From deep perusal of table 4 regarding the average growth rates, it is found that the average growth rates of area uder cereals remained negative but non-signifiacnt in period I. In period II, the average growth rate becomes more negative and significant at 0.01 level of significance. Thus, in the whole period of study the average growth rate remained negative and significant at 0.01 level of significance. The difference in average growth rates of area from period I to period II is negative but non-significant. The average growth rate of production and yield of cereals remained positive and significant at 0.01 level of significance in both the periods and in the whole period. The difference in the average growth rate from period I to period II is negative and significant at 0.01 level of significance for production and yield of cereals.

From the scrutiny of table 4 regarding the coefficient of variation, it is found that the coefficient of variation is significant at 0.01 level of significance for all the variables under study i.e., area, production and yield in period I, period II and the whole period. The difference in coefficient of variation from period I to period II is positive and signifcant at 0.01 level of significance only in case of yield of cereals. For area, coefficient of variation is negative and significant at 0.05 level of significance, whereas, for production it is negative and non-significant.

Table 1: Parameter estimates, residual diagnostics and model fit statistics of the fitted models for area under cereals in Odisha

Fitted Models →		Linear	Power	Compound	Quadratic
Parameter estimates	b_0	4706.01** (76.71)	4615.91** (103.3)	4501.29**(76.6)	4856.57** (101.17)
	b_1	--11.46 (5.84)	-0.035*(0.015)	0.997**(0.013)	-75.01** (20.26)
	a_1	24.36* (9.79)	0.037* (0.016)	1.005**(0.02)	-25.73 (43.51)
	b_2	-	-	-	2.82** (0.86)
	a_2	-	-	-	-1.71** (0.57)
Residual diagnostics	D-W Statistic	0.66**	0.946*	0.932*	1.978
	Coefficient of ln(t) in Park's test	0.677	0.85**	0.66	0.93
	S-W Statistic	0.933*	0.707	0.695	-0.36
Model fit statistics	R^2	0.336	0.046	0.34	0.573
	Adjusted R^2	0.305	0.022	0.304	0.542
	RMSE	203.62	180.15	203.52	118.8
	F Value	9.59**	9.9**	9.79**	12.24**

(Figures in the parentheses indicates the standard errors)
* Significant at 0.05 level of significance ** Significant at 0.05 level of significance

Table 2: Parameter estimates, residual diagnostics and model fit statistics of the fitted models for production of cereals in Odisha

Fitted Models →		Linear	Power	Compound	Quadratic
Parameter estimates	b_0	3396.2**	3438.03**	3406.36** (352.38)	4334.28** (495.48)
	b_1	93.44	0.15** (0.056)	1.019** (0.007)	-167.08* (79.24)
	a_1	-7.94	0.023 (0.14)	0.985** (0.009)	-483.15* (223.4)
	b_2	-	-	-	11.28* (4.19)
	a_2	-	-	-	0.02 (7.67)
Residual diagnostics	D-W Statistic	2.34	1.96	2.37	2.13
	Coefficient of ln(t) in Park's test	-0.251	0.666	0.175	0.733
	S-W Statistic	0.944*	0.956	0.947*	0.916*
Model fit statistics	R^2	0.607	0.532	0.613	0.679
	Adjusted R^2	0.598	0.521	0.600	0.663
	RMSE	972.90	1061.38	965.80	879.22
	F Value	63.33**	47.79**	66.45**	20.1**

(Figures in the parentheses indicates the standard errors)
* Significant at 0.05 level of significance ** Significant at 0.05 level of significance

Table 3: Parameter estimates, residual diagnostics and model fit statistics of the fitted models for yield of cereals in Odisha

Fitted Models →		Linear	Power	Compound	Quadratic
Parameter estimates	b_0	7238.36** (69.43)	723.01** (113.87)	756.48*(68.01)	913.79** (99.65)
	b_1	25.38** (5.29)	0.136* (0.05)	1.02** (0.01)	-22.78*(9.96)
	a_1	-5.89 (10.72)	-0.023 (0.08)	0.989** (0.01)	-107.98* (52.75)
	b_2	-	-	-	1.98* (0.84)
	a_2	-	-	-	1.22(1.65)
Residual diagnostics	D-W Statistic	2.12	1.77	2.1	2.02
	Coefficient of ln(t) in Park's test	0.346	0.566	0.441	0.628
	S-W Statistic	0.906**	0.934**	0.901**	0.966
Model fit statistics	R^2	0.627	0.553	0.625	0.732
	Adjusted R^2	0.618	0.543	0.620	0.719
	RMSE	222.82	243.78	223.42	188.7
	F Value	68.86**	52.03**	69.94**	25.99**

(Figures in the parentheses indicates the standard errors)
* Significant at 0.05 level of significance ** Significant at 0.05 level of significance

Table 4: Average growth rates of coefficient of variation of area, production and yield of cerealsof Odisha in period I, period II and the whole period

(In per cent)

Variables	Average Growth Rate				Coefficient of Variation			
	Period I (GR_1)	Period II (GR_2)	Whole Period (GR)	Δ GR = GR_2 GR_1	Period I (CV_1)	Period II (CV_2)	Whole Period (CV)	Δ CV = $CV_2 - CV_1$
Area	-0.28 (0.16)	-0.43** (0.11)	-0.32** (0.1)	-0.18 (0.27)	3.16** (0.46)	2.14** (0.31)	2.69** (0.28)	-1.02* (0.39)
Production	1.97** (0.43)	1.67** (0.39)	1.83** (0.29)	-0.30** (0.1)	19.15** (2.99)	18.31** (2.83)	18.69** (1.92)	-0.84 (2.98)
Yield	2.03** (0.48)	1.95** (0.53)	1.88** (0.37)	-0.08** (0.03)	13.22** (2.03)	22.09** (3.51)	14.91** (1.55)	8.87** (2.82)

(Figures in the parentheses represents the standard errors)
* Significant at 5 % level; ** Significant at 1 % level

4. Conclusion

The above discussion hightlighted that though some models perform better in terms of model fit statistics, such as, high value of adjusted R^2 and low value of RMSE, they are not worthy of being selected as the best model if they do not satisfy the error assumptions. This is the case with the statistical modelling in case of production of cereals. Though the linear model, compound model and quadratic model fitted to the production of cereals data, have higher values of adjusted R^2 than the power model fiited to the same data, but these models show lack of normality of errors as indicated by their significant S-W statistic. The spline regression technique also takes care of any abrupt jumps in the data.

From the study of average growth rate found with the help of best fit model, it is found that area under cereals shows a declining trend and the differnce in the average growth rate from pre-reform period to post reform period is significantly negative. This decline in area may be assigned to the rapid growth of industrialisation and shift of agriculture from cereals to commercial crops in the post reform period. The average growth rate in case of production and yield shows an increasing trend in both the periods, but the rate of increase is less in the post reform period. This shows that attempts have been taken to improve the yield of cereals in both the periods which was less sucessful to some extent in post-reform period.

From instability point of view, there is increase in instability of area, production and yield of cereals in post-reform period as compared to pre-reform period which is significant only in case of yield. This shows that high growth rate is more prone to high instability due to the fact that the activites that leads to increase in yield, may not be carried out with equal potency in all the years throughout the period.

5. References

1. Gujarati DN. Basic Econometrics, Fourth Edition, Published by McGraw-HiII/lrwin, 2003.
2. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 3rd Edition. New York. John Wiley & Sons, 2001.
3. Paltasingh RK, Goyari P. Agricultural Economics Research Review. 2013; 26(Conference number):67-78.
4. Prajneshu, Chandan KP. Computation of compound growth rates in agriculture, Revisited. Agricultural Economics Research Review, 2005, 317-324.
5. Thode Jr HC. Testing for Normality. Marcel Dekker, Inc., New York, 2012.