Study on Growth Models: A Critical Analysis with Reference to Andhra Pradesh and India

Pradip Kumar Nandi[*] and G Krishna Kanth

Acharya N G Ranga Agricultural University, Rajendranagar, Hyderabad-500030

ABSTRACT

Different linear and nonlinear growth models are studied for the purpose of estimating the growth rate and fitting the best model, which will help in better future prediction. The use of R^2, Adj. R^2 or RMS as a measure of goodness of fit and, therefore, as a criterion for choosing the best model, is not sufficient. The criteria of randomness and normality of time-series data should be satisfied. This particular aspect of fitting trend and growth rate estimation is rarely given due importance by different researchers, as a result various misleading inferences are drawn in earlier studies. Keeping in this mind we have estimated the parameters of the model and could able to infer correctly. One most important thing we have observed here is that the area under the total foodgrain crop is declining day by day and this is due to the replacement of area by different cash crops. Projected production of total foodgrain is expected to be 13.87 million tones and 232.79 million tones for Andhra Pradesh and India respectively.

Key Words: Linear and Nonlinear Growth model, Adj.R^2, RMS, Run test, Shapiro-Wilk test.

INTRODUCTION

In India the pace of agricultural growth significantly governs the pace of economic development, as a lion's share of the Gross Domestic Product (GDP) of the country comes from agriculture and allied sector measuring around 29 per cent. On the contrary, there is no consistent growth of agricultural commodities in our country, which is oscillatory in nature. The present study has been undertaken to evaluate the development in area, production and productivity of important crops in Andhra Pradesh in comparison with India using various linear, non-linear and linearisable growth models with a view to provide analytical approach of fitting appropriate growth model. The best-fitted model is taken for future projection of area, production and productivity of foodgrain crops of

---

[*] Present Address: Indian Agricultural Statistics Research Institute (IASRI), New Delhi
E-mail: nandi_stat@yahoo.co.in

Andhra Pradesh and India upto 2020 AD. Different researcher has made some isolated attempts. Borthakur and Bhattacharya (1998) analyzed trend of area, production and productivity of potato in Assam for the period of 1951 to 1993 using three different functional forms (linear, quadratic and exponential). Chattopadhyay and Das (2000) analyzed growth rates of agricultural production in West Bengal using log-quadratic, exponential and parabolic trend equations. To choose the best model, which explains the underlying phenomenon, Adj. R^2 and Residual Mean Square (RMS) were calculated. Some important assumptions regarding randomness (using one sample run test), test for normality (Shapiro-Wilk test) was also checked. In the present study the effort has been made in such a way that parameters of the best-fitted model have appropriate biological interpretation. The utility of such modeling efforts is that as different agricultural mechanism follows distinct patterns, an overview of the selection criteria of model will provide an insight into the underlying mechanism. Moreover the knowledge of estimated future production would help agriculture researcher to redirect their investigation towards the goal of sustainable development.

METHODOLOGY

The data pertaining to area, production and productivity of selected crops of Andhra Pradesh and India were collected for the period of 28 years (1972-73 to1999-2000) from the publication of Centre For Monitoring Indian Economy (CMIE'2001), Agriculture. To analyze trend of area, production and productivity following linear and nonlinear models were selected.

Linear	$Y_t = a + bt + e$
Quadratic	$Y_t = a + bt + ct^2 + e$
Logarithmic	$Y_t = a + b \ln(t) + e$
Exponential	$\ln Y_t = \ln a + t \ln b + e$
Inverse	$Y_t = a + b/t + e$
Power	$\ln Y_t = \ln a + b \ln t + e$
Logistic	$Y_t = 1 / [1/K + a\, b^t] + e$
Gompertz	$Y_t = Exp - (Exp - (a + bt)) + e$
Modified exponential	$Y_t = K + Exp(a + bt) + e$

Parameter Estimation: For the purpose of estimation of parameters let us consider a general linear model, $y = X\beta + \in$

Where y is the (n x 1) vector of dependent variable, X is the (n x p) matrix of observation, β is the (p x 1) vector of parameters included in the model, $\in$ is the (n x 1) residual vector, and includes the repressors not included in the model. For first five (*i to vi*) models that have linear or linearisable form, so parameters can be estimated using Ordinary Least Square (OLS) technique. The assumptions about the linear model are,

$$E(\in) = 0, \; E(\in \in') = \sigma^2 I_n \text{ and X is non-stochastic.}$$

Then the estimate of the parameter β is given by,

$$\hat{\beta} = (X'X)^{-1}X'y$$
$$\text{and} \quad v(\hat{\beta}) = (X'X)^{-1}\sigma^2$$

It may be noted that the last three models are 'nonlinear', as each one of these involves at least one parameter in a nonlinear manner. Like linear regression, parameter estimates in nonlinear case also can be obtained by minimizing the residual sum of squares.

Let us consider a non-linear model,

$$y = f(\gamma, x)) \ + \ \in$$

where y is the nx1 response vector, x is the regressor variable and γ is the vector of parameters, $\in$ is the residual vector.

The normal equations can be obtained by minimizing the residual sum of squares,

$$Q = \sum_{i=1}^{n} \in_i^2 = \sum_{i=1}^{n} [y - f(\gamma, x)]^2$$

However, because of nonlinearity, the resulting normal equations are nonlinear in parameters and so cannot be solved exactly. Accordingly, a number of iterative procedures have been developed to obtain approximate solutions. A good account of procedures has been included in the book by Seber and Wild (1989). Four main methods of this kind are:

      i)   Linearization method

      ii)   Gradient method

      iii)   Levenberg-Marquardt (LM) method, and

      iv)   Does not use Derivatives (DUD) method

The linearization method uses the results of linear least square theory in a succession of stages. However, the drawback of this procedure is that it may converge very slowly or oscillates widely or may not converge at all. In short, neither this method nor gradient method is recommended in practice. The LM procedure represents a compromise between these methods and combines successfully the best features of both as well as avoids their serious disadvantages. The DUD procedure, as the name suggests, is a derivative-free method. Now a days most of the standard statistical software packages like SPSS and SAS contain programs for fitting nonlinear models by LM and DUD

procedures using PROC NLIN in SAS and NLR option in SPSS. However, convergence to biologically meaningful values cannot always be guaranteed by any procedure and the success rate is quite low. In the present study SPSS package using LM method has been used to estimate the constants of non-linear models. Each of these trend equations imposes certain restrictions upon the character of the growth process.

So in fitting trend equation the choice of trend equation from amongst the available alternatives is very crucial.   The most commonly used statistical criterion is the coefficient of determination, R^2 or adjusted R^2 ($\overline{R}^2$).

$$R^2 \quad = \quad \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \quad 1 - \frac{e'e}{y'y - (1'y)^2/n}$$

As R^2 is not an adequate measure for choice of nonlinear models, because the reduced linear model obtained through Linearization method or LM method are not having intercept term. In case of model without intercept term, sum of residuals is not zero.
We know,

$$e_i = y - X\hat{\beta} = (I - H)y \qquad \text{where} \quad H = X(X'X)^{-1}X'$$
$$1'e = 1'(I - H)y = \left(p'X' - p'X'X(X'X)^{-1}X'\right)y = (p'X' - p'X')y = 0$$

where,   1 is the (n x 1) unit vector .

$e = y - X\hat{\beta}$, estimated residual vector.

It is possible only in case of intercept model as unit vector belongs to the column space of X. Keeping this in mind adjusted R^2 and Residual Mean Square are considered here for the purpose of choosing the best model. Adjusted R^2 is calculated as,

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

where   p is the number of constants in the equation.

                  n is the total number of observation.

This criterion, though a very powerful one as a test of goodness of fit, has several deficiencies under certain circumstances. For example, a power model may first be liberalized by using a logarithmic transformation and then fitted to empirical data by using ordinary least square method. The R^2 value is then often calculated using the data points (log $\hat{Y}_i$, log Y_i). This R^2 is generally interpreted as a measure of goodness of fit of even the original non-linear model which is incorrect. So the best criterion to choose a model is the residual mean square (RMS), which will also ensure accurate forecasting.

$$\text{Residual Mean Square (RMS)} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\text{residual degrees of freedom}}$$

Before choosing a model one should be certain that the disturbance term satisfies all the conditions of randomness, non-autocorrelation, homoscedasticity and normality. The property of homoscedasticity is hardly checked perhaps on the assumptions that it is not a problem of time series data. In the present study an attempt has been made to verify two most important assumptions of normality and randomness of residuals.

Test of Randomness: One Sample Run Test

Randomness of residuals can be tested by using non-parametric one sample run test (Siegel and Castellan, 1965). A run is defined as a succession of identical symbols in which the individual scores or observations originally were obtained. For example, suppose a series of binary events occurred in this order:

＋＋ - - - ＋ - - - - ＋＋ - - ＋

If very few runs occur, a time trend or some bunching owing to lack of independence is suggested. If a great many runs occur, systematic short-period cyclical fluctuations seem to be influencing the scores.

Let 'm' be the number of elements of one kind, and 'n' be the number of elements of the other kind in a sequence of N = m + n binary events. If both m and n are less than or equal to 20, then the number of runs, r if falls between the critical values, we cannot reject null hypothesis. Where null hypothesis is,

$$H_0 : \text{The sequence is random.}$$

For large samples if either 'm' or 'n' is large than 20, a good approximation to the sampling distribution of 'r' is the normal distribution with.

$$\text{Mean} = \mu_r = \frac{2mn}{N} + 1$$

$$\text{And standard deviation} = \mu_r = \sqrt{\frac{2mn(2mn - N)}{N^2(N-1)}}$$

$$\text{Then } H_0 \text{ may be tested by } Z = \frac{r - \mu_r}{\sigma_r}$$

The significance of any observed value of Z computed by using the equation may be determined from a normal distribution table.

Test of Normality: Shapiro-Wilk Test

It is important to note that for regression the normality test should be applied to the residuals rather than the raw scores. However, this assumption is not so stringent while selecting non linear models because residuals of non linear models may not follow normal distribution. But it is must for linear models and linearised models. That is why in the present study normality check has been included. There is not a general agreement of the best way to test normality. Three some popular tests for normality namely, the

Shapiro-Wilk (Shapiro and Wilk, 1965), the Kolmogorov – Smirnov test, Cramer Von

Misses test and Anderson-darling test. According to the SAS manual, if the sample size

is less than 2000, the Shapiro-Wilk test is better. The null hypothesis is

H_0: there is no significant departure from normality.

Shapiro-Wilk statistic W (Pearson and Hartley, 1972) is given as

$$W = \frac{\left[\sum_{i=1}^{n} a_i X_{(i)}\right]^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

where $i = 1, \ldots, n$

$X_{(i)}$ = ordered sample values

a_i = the constants generated from mean, variances and covariances of the order statistics

of a sample of size 'n' from a normal distribution.

If the p-value is smaller than the critical value, normally 0.05, H_0 is rejected and we

conclude that the population is not normal. Shapiro-Wilk statistic is W; distribution of W-

Statistic ranges in between 0 and 1, highly skewed to the right.

RESULTS AND DISCUSSION

Results obtained after fitting the growth models to the total foodgrain production in

Andhra Pradesh are shown in table 1. Area under total foodgrain crops had no systematic

trend; it was randomly distributed over the years. Power function was chosen for the

purpose of projection of total foodgrain production because minimum RMS was

exhibited by power function and also satisfied other assumptions, that is the number of

runs was found to be insignificant and Shapiro-Wilk statistic was also insignificant at

0.05 probability level. So the assumptions of randomness of residual and normal

distribution of residual are satisfied. Projection of total foodgrain production in Andhra

Pradesh was made on the basis of power function and it was found to be 13.87 million tones by 2020 AD.

In case of all India level it was observed that area under total foodgrain crop was increasing upto 1987-88 but after that declined drastically again starts increasing. For the country as a whole foodgrain production exhibited increasing trend over the study period. Gompertz curve was found to be the best fitted one as it has least RMS and also satisfied the assumptions normality and independence of residuals. Projected production of total foodgrain is shown in table 3 and by 2020 AD it will reach to the level of 232.79 million tones.

Productivity of total foodgrain crops in Andhra Pradesh was also exhibited increasing trend over the study period. Linear function was fitted well to the data pertaining to the productivity of total foodgrain crops. All other functions were also showed goodness of fit. But the least RMS was observed in case of linear function and the assumptions of normality and randomness of residuals were satisfied by this function. So linear function was chosen for the purpose of future prediction. Projected productivity of foodgrain crops by 2020 AD was found to be 2772.15 kg ha^{-1}. It can be said that though area under total foodgrain followed declining trend but the increase in productivity accounts for the increase in total foodgrain production. For the country as a whole, similar type of trend was observed. Linear function was fitted to the data and as it has fulfilled the criteria of best fit having least RMS, insignificant number of runs and insignificant Shapiro-Wilk statistic. Future projections were made based upon the above function. Projected productivity was 2319.96 kg ha^{-1} by 2020 AD.

CONCLUSION

From the above discussion it is obvious that one specific model is not enough to explain the production pattern in differing situations. That is why different growth models are fitting well separately for different crops, for different parameters (area, production and productivity) and also in diverse geographical situations. This is due to the fact that agriculture industry is getting diversified. Area under total foodgrain crops shifting towards the cash crops, which can provide more remuneration to the farming community. Again the productivity of foodgrain crops is increasing which is the sign of intensive agriculture and thereby the total production is increasing. But still there is the problem of underfeeding in different areas in India. So this study opens up new vistas for more intensive research to achieve the production level, which will lead us towards sustainable agriculture.

REFERENCES

Borthakur, S. and Bhattacharya, B.K. (1998), Trend Analysis of Area, Production and Productivity of Potato in Assam: 1951-1993, Economic Affairs, Vol. 43 No.4, December 1998.

Chattopadhyay, A.K. and Das, P.K. (2000), Estimation of Growth Rate: A Critical Analysis with Reference to West Bengal Agriculture, Indian Journal of Agricultural Economics, Vol. 55. No.2, April-June 2000

Chennarayudu, K.C, Narender, I. and Parthasarathy, P.B. (1993), Trends in Area, Production and Productivity of Foodgrain in Andhra Pradesh and India-Projection by 2000 AD: An Econometric Analysis, Journal of Indian Society of Agricultural Statistics, Vol. 45 No.1: 106.

Draper, N.R. and Smith, H. (1981), Applied Regression Analysis, Second Edition, John Wiley.

Gupta, S.C. and Kapoor, V.K. (2001), Fundamentals of Applied Statistics, Third Edition, Sultan Chand and Sons.

Montgomery, D.C. and Peck, E.A. (1982), Introduction to Linear Regression Analysis, John Wiley and Sons.

Seber, G.A.F. and Wild, C. (1989). Nonlinear Regression. John Wiley , New York.

Scot, A. and Wild, C. (1991), Transformations and R^2, The American Statistician, Vol. 45 No.2: 127-128.

Shapiro, S.S. and Wilk, M.B. (1965), An Analysis of Variance Test for Normality Biometrika, Vol.52: 229-240.

Siegel, S. and Centellen, N.J. (1988), Nonparametric Statistics for Behavioral Sciences, McGraw-Hill.

Srivastava, A.K. and Shobhit (2000), Comparison of R^2 and Adj. R^2 in Linear Regression Models Under Asymptotic Losses, Journal of Quantitative Economics, Vol. 16 No.2: 65-70.