Jour. Ind. Soc. Ag. Statistics
56(1), 2003 : 1-6

Nonlinear Regression: A Realistic Modeling Approach in Horticultural Crops Research

R.Venugopalan and K.S. Shamasundaran
Indian Institute of Horticultural Research, Bangalore-560089
(Received : October, 2001)

SUMMARY

Nonlinear statistical models play a very important role in almost all branches of agricultural/horticultural and biological sciences due to the existence of wide population fluctuations and complex non-linear inter-relationship among variables of interest. This article brings out the realistic nature of modeling such variables using non-linear regression approach. Four different methods of non linear regression are discussed and measures of goodness of fit are presented with a view to develop a suitable non-linear model for describing data pertaining to the period 1960-61 to 1976-77, on average fruit yield of coorg mandarin trees. Using the Gompertz model selected, it was inferred that 94 percent of the carrying capacity (maximum sustainable yield) had already been achieved by the year 1977 and hence there is little scope for its survival in Coorg region.

Key words : Statistical modeling, Nonlinear regression, Goodness of fit.

1. Introduction

Statistical modeling essentially consists in constructing a model, represented by a set of equations to describe the input-output relationship among the variables of interest. From a realistic point of view, such a relationship among variables in agricultural/horticultural and biological sciences are 'non linear' in nature. In such models, a unit increase in the value of independent variable(s) may not result in an equivalent unit increase in the dependent variable. For example, the relationship between yield of a crop and the spacing of density of plants, the dose-response relationship between yield of various crops and zone, models of *aphid* population growth are usually non-linear in nature.

However, due to complex nature of these models, approximate transformed versions of these models are utilized in drawing conclusions, ignoring the possible disturbance in error structure. Furthermore, as a measure of goodness of fit R^2 is usually used in the literature. But, while dealing with such a transformed model, R^2 measures only the variability in the new model;

and not of the original nonlinear model. Problems of this nature are enormous while dealing with linearized version of nonlinear model. Hence, it is the purpose of this present article to highlight these and brings out the realistic way of statistical modeling through nonlinear regression approach. Four such methods viz. (i) Gauss-Newton method, (ii) Steepest Descant method, (iii) Lavenberg-Marquardt technique and (iv) Do not use derivatives (DUD) are discussed with their relative merits and demerits. As a measure of goodness of fit of such models, importance of carrying out residual analysis is also highlighted. The above said theory is used to develop a suitable non-linear model for describing data on average fruit yield of orange trees.

2. *Materials and Methods*

Average fruit yield (in terms of number of fruits/tree) of coorg mandarin trees (based on 30 samples) during the period 1960-61 to 1976-77, observed at experimental station of IIHR at Gonnikopal/Chethalli is considered with a view to assess its survival in Coorg region and also to workout the maximum yield that could be attained on sustainable basis. Standard cultural practices for cultivation was followed and the spacing used was 6m × 6m. The above objective was materialized by the following procedures as delineated below.

(i) Common Procedure Followed for Parameter Estimation in Nonlinear Models and the Drawbacks

A non-linear regression model is one in which at least one of the parameters appear non-linearly. Mathematically, in nonlinear models at least one of derivatives of the expectation function with respect to at least one parameter is a function of parameter(s). For example

$$Y_t = a X_t^{\,b} \qquad (1)$$

is a nonlinear regression model as the derivatives of Y_t with respect to a and b are both functions of a and /or b. Details about family of nonlinear models are mentioned in Ratkowsky [2]. Like in linear regression, parameters in a nonlinear model can also be estimated by the method of least squares. However, due to the difficulty in the procedure of computation, the common practice is to work with the log transformed model.

The above transformation is valid only when error term 'e' in equation (1) is multiplicative in nature. Thereafter, method of least square is used to estimate the unknown parameters. Furthermore, R^2 value is calculated to measure the goodness of fit of the model.

This above-mentioned procedure suffers from some important drawbacks.

(a) Original structure of the error term got disturbed once we use suitable transformation.

(b) R^2 values computed, assess the goodness of fit of the transformed model and not of the original nonlinear model.

(c) Proceeding further to carryout residual analysis for the residuals generated by the transformed model, will result in erroneous conclusions.

As a remedy to these pitfalls, nonlinear regression procedures are already developed in literature which necessitates computer intensive tools to find solution for the parameters. Four such kinds of methods are discussed briefly in the following section.

(ii) Nonlinear Regression Methods

Four main methods are available in literature (Seber and Wild [4]) to obtain estimates of the unknown parameters of a nonlinear regression model. These are: (i) Gauss-Newton method (ii) Steepest-Descent method (iii) Levenberg-Marquardt technique and (iv) Do not use derivative (DUD) method. However, in all these methods the following steps are carried out.

Step (i) Starting with a good initial guess of the unknown parameters, a sequence of θ's which hopefully converge to θ is computed.

Step (ii) Error sum of squares or objective function expressed as $S(\theta)=\sum_{t=1}^{n}[Y_t - F_i(\theta)]^2$ is minimized with respect to the current value of θ. The new estimates are obtained.

Step (iii) By feeding the recently obtained estimates as the initial guess for the next iteration, objective function $S(\theta)$ is minimized again to obtain fresh estimates. This procedure is continued till the successive iteration yielded parameter estimate values are close to each other.

Marquardt method is widely used for computing non-linear least squares. The reason for this is almost all the standard statistical packages have built in programs for estimating nonlinear parameter estimates. For example, statistical analysis systems (SAS) have NLR option and statistical package for social sciences (SPSS) has NLR option to achieve the above task. To sum up, all these methods do not necessitate transforming the original data points and the original structure is preserved which otherwise is also important while dealing with horticultural data.

(iii) Measures of Goodness of fit

The following measures of goodness of fit statistics are generally used to judge the adequacy of the model developed (Agostid'no and Stephens[1])

Root mean squared error (RMSE)

$$RMSE = \sqrt{\left[\Sigma_t\left(Y_t - \hat{Y}_t\right)^2 / n\right]}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \left[\Sigma\left(Y_t - \hat{Y}\right)^2 / \Sigma(Y_t - \overline{Y})^2 \right]$$

However, while fitting regression models to the data considered, it may be noted that even if we add one more independent variable to the model, R^2 value will also get increased. Further more, while dealing with time-series data it may be possible that successive observations may be auto correlated among themselves. To overcome all these problems, performing residual analysis is strongly advised. Randomness assumption of the residuals need to be tested before taking any final decision about the adequacy of the model developed. To carry out the above analysis "Run test" procedure is developed in the literature (Ratkowsky [3]). Further, to test for the presence or absence of autocorrelation in the data set Durbin-Watson test procedure (Lewis-Beck [2]) was utilized. To demonstrate the forgoing ideas we present the following illustration.

3. Illustration

Data on average fruit yield of coorg mandarin trees (based on 30 samples) during the period 1960-61 to 1976-77 collected at experimental station of IIHR at Gonnikopal/Chethalli is considered with a view to develop a suitable model. The observation corresponding to the period 1972-73 was not included while developing models, as the trees were affected and resulted in low yield which lead to an outlier data for the regression analysis. The observed yield data over years (see Fig. 1), indicated about the appropriateness of using nonlinear models, due to its S-shaped pattern. The following nonlinear growth models (Seber and Wild [3]) were tried to explain the data set.

Logistic model

$Y_t = c/(1 + b \exp(-at)) + e,\ b = c/Y(0) - 1$

Gompertz model

$Y_t = c \exp(-b \exp(-at)) + e,\ b = \ln(c/Y(0))$

Richards model

$Y_t = c[1 + b \exp(-at)]^{(-1/d)} + e,\ b = c^d/Y^d(0) - 1$

Monomolecular model

$Y_t = c - (c - a) \exp(-bt) + e$

Morgan–Mercer–Flodin (MMF) model

$Y_t = (bc + at^d)/(c + t^d) + e$

Where Y_t is the orange yield observed during the time t; a,b,c,d, are the parameters, and e is the error term. The parameter a is the intrinsic growth rate and the parameter c represents the carrying capacity for each model. Symbol b represents different functions of the initial value Y(0) and d is the added

parameter in Richards model. Levenberg-Marquardt algorithm, which is widely used, is utilized for fitting all the three models. Different sets of initial parameter values were tried so as to ensure global convergence. The iterative procedure was stopped whenever the successive iterations parameter estimates values were negligibly low.

4. Results and Discussion

The observed sample values are utilized to develop non-linear models and the results are presented in Table 1. The Goodness of fit statistics, viz. RMSE and R^2 are also presented. Run-test statistic (|Z|) value is also presented. Perusal of Table 1 indicates, among many other things, that logistic, Gompertz and Mono-molecular models provide satisfactory results. R^2 values being almost equal in all three models, very low RMSE value in case of Gompertz model viz., (65.37) , Gompertz model is selected to explain the data set. However, Richards and MMF models failed to produce convergence of parameter estimates for the data set considered. Further, the Durbin-Watson test statistic values computed under both the models, being nearer to zero, reflect about the absence of autocorrelation in the time series data. Before drawing final conclusion about the adequacy of the selected model, randomness assumption of the residual was carried out. Value of test statistic viz., 0.934 well below the critical value 1.96 of normal distribution at 5% level, also ensures the suitability of the selected model. A graphical representation of the adequacy of the fitted models is presented in Fig 1.

Further more, by working out the ratio of yield during the last year in the data set to the carrying capacity value of 895 fruits, it may be seen that 94% of the carrying capacity is already achieved and hence there is little scope for further increase. This may possibly be one of the reasons for the extinction of coorg mandarin variety in Coorg region of Karnataka. This study, in view of the recent ongoing efforts to look into the aspect of revival of coorg mandarin, would go a long way and may be used as a baseline for revival of the crop.

Table 1. Summary statistics of model fit

Parameter	Logistic	Gompertz	Mono-Molecular
A	0.57	0.38	-203.8
B	14.73	4.01	0.18
C	870.45	895.02	988.54
R square value	0.951	0.957	0.953
RMSE	70.08	65.37	68.27
Run test |Z| value	1.25	0.934	1.10
Durbin-Watson statistic	1.04	1.053	1.5
Carrying capacity achieved at present	92.5%	94%	98%

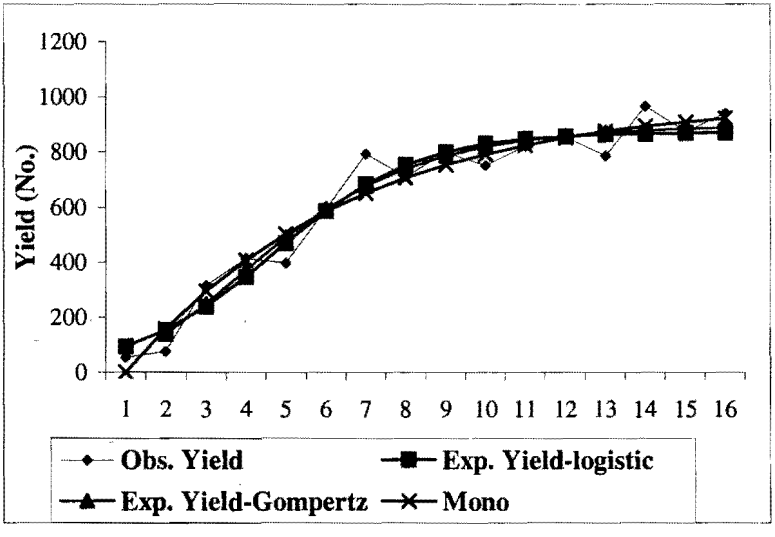

Fig.1 . Graphical display of fitted models along with sample values

ACKNOWLEDGEMENTS

The authors are thankful to the referees for their valuable suggestions that led to substantial improvement in the quality of the paper.

REFERENCES

[1] Agostid'no, R.B. and Stephens, M.A. (1986). *Goodness of Fit Techniques.* Marcel Dekker, New York.

[2] Lewis-Beck, S.M. (1993). *Regression Analysis.* Sage Publication, New York.

[3] Ratkowsky, D.A. (1990). *Handbook of Non-linear Regression Models.* Marcel Dekker, New York.

[4] Seber, G.A.F. and Wild, C.J. (1989). *Nonlinear Regression.* John Wiley and Sons, New York.