NON-LINEAR REGRESSION MODELS AND THEIR APPLICATIONS

PRAJNESHU
Indian Agricultural Statistics Research Institute
Library Avenue, New Delhi-110 012
prajnesh@iasri.res.in

1. Linear Model

A mathematical model is an equation or a set of equations which represents the behaviour of a system (France and Thornley, 1984). It can be either 'linear' or 'nonlinear'. A linear model is one in which all the parameters appear linearly. Some examples of linear model are:

1.1 Multiple Linear Regression

$$Y = a_0 + a_1X_1 + ... + a_pX_p + \varepsilon ,$$

where Y is the dependent (or response) variable, X_i are independent (or predictor) variables and ε is the error term.

1.2 Polynomial Models with One Predictor Variable

$$Y = a + b X + \varepsilon \qquad \text{(First-order model)}$$
$$Y = a + b X + c X^2 + \varepsilon \quad \text{(Second-order or Quadratic or Curvilinear model)}$$

Above models are very widely used in Agriculture, Industry, Education, Medicine, etc. 'Method of least squares' is generally employed for estimation of parameters. However, if polynomial models of a certain order are fitted to data by applying this procedure and later it is decided to add an extra term of a higher order, then estimates of all the parameters in the model have to be computed afresh.

2. Nonlinear Model

It is well recognized that any type of statistical inquiry in which principles from some body of knowledge enter seriously into the analysis is likely to lead to a 'Nonlinear model' (see e.g. Seber and Wild,1989). Such models play a very important role in understanding the complex inter-relationships among variables. A **nonlinear model** is one in which at least one of the parameters appears nonlinearly. More formally, in a nonlinear model, at least one derivative with respect to a parameter should involve that parameter. Examples of a nonlinear model are:

$$Y(t) = \exp(at+bt^2) \qquad \qquad ...(2.1)$$

$$Y(t) = at + \exp(-bt). \qquad \qquad ...(2.2)$$

Note. Some authors use the term 'intrinsically linear' to indicate a nonlinear model which can be transformed to a linear model by means of some transformation. For example, the model given in (2.1) is 'intrinsically linear' in view of the transformation $X(t) = \log_e Y(t)$.

3. Some Important Nonlinear Growth Models

Those models which describe the growth behaviour over time, are applied in many fields. In the area of population biology, growth occurs in plants, animals, organisms, etc. The type of model needed in a specific situation depends on the type of growth that occurs. In general, growth models are mechanistic in nature, rather than empirical. In the former, the parameters have meaningful biological interpretation; the latter is just like a 'black-box' where some input is given and some output is taken out. A mechanistic model usually arises as a result of making assumptions about the type of growth, writing down differential or difference equations that represent these assumptions, and then solving these equations to obtain a growth model. The utility of such models is that, on one hand, they help us to gain insight into the underlying mechanism of the system and on the other hand, they are of immense help in efficient management. We now discuss briefly some well-known nonlinear growth models:

3.1 Malthus Model

If $N(t)$ denotes the population size or biomass at time t and r is the intrinsic growth rate, then the rate of growth of population size is given by

$$dN/dt = rN. \qquad \qquad ...(3.1)$$

Integrating, we get

$$N(t) = N_o \exp(rt), \qquad \qquad ...(3.2)$$

where N_o denotes the population size at $t=0$. Thus this law entails an exponential increase for $r>0$. Furthermore, $N(t) \to \infty$ as $t \to \infty$, which cannot happen in reality.

Note. The parameter r is assumed to be positive in all models.

3.2 Monomolecular Model

This model describes the progress of a growth situation in which it is believed that the rate of growth at any time is proportional to the resources yet to be achieved, i.e.

$$dN/dt = r(K-N), \qquad \qquad ...(3.3)$$

where K is the carrying size of the system. Integrating (3.3), we get

$$N(t) = K- (K-N_o) \exp(-rt). \qquad \qquad ...(3.4)$$

3.3 Logistic Model

This model is represented by the differential equation

$$dN/dt = rN (1-N/K). \qquad \qquad ...(3.5)$$

Integrating, we get

$$N(t) = K / [1+(K/N_o-1) \exp(-rt)]. \qquad \qquad ...(3.6)$$

The graph of $N(t)$ versus t is elongated S-shaped and the curve is symmetrical about its point of inflexion.

3.4 Gompertz Model

This is another model having a sigmoid type of behaviour and is found to be quite useful in biological work. However, unlike the logistic model, this is not symmetric about its point of inflexion. The differential equation for this model is

$$dN/dt = rN \log_e (K/N). \qquad ...(3.7)$$

Integration of this equation yields

$$N(t) = K \exp[\log_e (N_o/K) \exp(-rt)]. \qquad ...(3.8)$$

3.5 Richards Model

This model is given by

$$dN/dt = rN(K^m - N^m)/(mK^m), \qquad ...(3.9)$$

which, on integration, gives

$$N(t) = K\, N_o\, /[N_o + (K^m - N_o^m) \exp(-rt)]^{1/m}. \qquad ...(3.10)$$

Evidently, the last three models are particular cases of this model when m = −1,1,0 respectively. However, unlike the earlier models, this model has four parameters.

4. A Nonlinear Model for Aphid Population Growth

Aphids are recognized as serious pests of cereals, oilseeds, pulses and vegetable crops in our country. It is highly desirable to investigate optimal control policies for controlling this pest. To this end, as a first step, the following model is developed by Prajneshu (1998) for describing the dynamics of aphid population growth. The model is expressed in terms of the integro-differential equation:

$$\frac{dN}{dt} = rN - \frac{N}{c} \int_0^t N(s)\, ds, \qquad ...(4.1)$$

and it is solved to obtain

$$N(t) = ae^{bt} (1 + ce^{bt})^{-2}. \qquad ...(4.2)$$

As an application of this model, optimum time for insecticidal spray is determined.

5. Fitting of Nonlinear Models

The above models have been posed deterministically. Obviously this is unrealistic and so we replace these deterministic models by statistical models by adding an error term on the right hand side and making appropriate assumptions about them. This results in a 'Nonlinear statistical model'. As in linear regression, in non-linear case also, parameter estimates can be obtained by the 'Method of least squares'. However, minimization of residual sum of squares yield normal equations which are nonlinear in the parameters. Since it is not possible to solve nonlinear equations exactly, the next alternative is to obtain approximate analytic solutions by employing iterative procedures. Three main methods of this kind are:

(i)   Linearization (or Taylor Series) Method
(ii) Steepest Descent Method
(iii) Levenberg-Marquardt's Method

The details of theses methods along with their merits and demerits are given in Draper and Smith (1998). The linearization method uses the results of linear least square theory in a succession of stages. However, neither this method nor the Steepest descent method, is ideal. The latter method is able to converge on true parameter values even though initial trial values are far from the true parameter values, but this convergence tends to be very slow at the later stages of the iterative process. On the other hand, the linearization method will converge very rapidly provided the vicinity of the true parameter values has been reached, but if initial trial values are too far removed, convergence may not occur at all.

The most widely used method of computing nonlinear least squares estimators is the Levenberg-Marquardt's method. This method represents a compromise between the other two methods and combines successfully the best features of both and avoids their serious disadvantages. It is good in the sense that it almost always converges and does not 'slow down' at the latter part of the iterative process. We now discuss this method in some detail.

Let us consider the model

$$y_i = f(x_i, \underset{\sim}{\theta}) + \varepsilon_i \ , i = 1,2,...,n, \qquad\qquad ...(5.1)$$

where y_i is the ith observation of the dependent variable, x_i is i^{th} independent variable, $\boldsymbol{\theta} = (\theta_1 \ \theta_2 \ ... \ \theta_p)\,'$ are parameters, ε_i, the error terms are independent and follow N(0, σ^2). The residual sum of squares is

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n}[y_i - f(x_i, \boldsymbol{\theta})]^2 . \qquad\qquad ...(5.2)$$

Let $\boldsymbol{\theta}_0 = (\theta_{10} \ \theta_{20} \ ... \ \theta_{po})\,'$ be the vector of initial parameter values. Then the algorithm for obtaining successive estimates is essentially given by

$$(\mathbf{H} + \tau\,\mathbf{I})\,(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_1) = \mathbf{g}, \qquad\qquad ...(5.3)$$

where

$$\mathbf{g} = \left.\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \quad H = \left.\frac{\partial^2 S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0},$$

I is the identity matrix and τ is a suitable multiplier.

Note
(i)   Marquardt scaled the various quantities appearing in (5.3) by the standard deviations of the derivatives with respect to the parameter values.
(ii) Now a days most of the standard statistical packages contain computer programmes to fit nonlinear statistical models based on Levenberg-Marquardt algorithm. For

example, SPSS has NLR option, SAS has NLIN option, IMSL has RNSSQ option to accomplish the task. In SAS package, one more procedure for nonlinear estimation viz. Does not use derivatives (DUD) procedure is also available.

6. Choice of Initial Values

All the procedures for nonlinear estimation require initial values of the parameters and the choice of good initial values is very crucial. However, there is no standard procedure for getting initial estimates. The most obvious method for making initial guesses is by the use of prior information. Estimates calculated from previous experiments, known values for similar systems, values computed from theoretical considerations all these form ideal initial guesses. Some other methods are:

6.1 Linearization

After ignoring the error term, check the form of the model to see if it could be transformed into a linear form by means of some transformation. In such cases, linear regression can be used to obtain initial values.

6.2 Solving a system of equations

If there are p parameters, substitute for p sets of observations into the model ignoring the error. Solve these equations for the parameters, if possible. Widely separated x_i often work best.

6.3 Using properties of the model

Consider the behaviour of the response function as the x_i go to zero or infinity, and substitute in for observations that most nearly represent those conditions in the scale and context of the problem, solve, if possible, the resulting equations.

6.4 Graphical method.

Sometimes a visual estimate can be obtained by plotting the data.

7. Goodness of Fit of a Model

This is generally assessed by the coefficient of determination, R^2. However, as pointed out by Kvalseth(1985), eight different expressions for R^2 appear in the literature. One of the most frequent mistakes occurs when the fits of a linear and a nonlinear model are compared by using the same R^2 expression but different variables. Thus, for example, a power model or an exponential model may first be linearized by using a logarithmic transformation and then fitted to data by using ordinary least squares method. The R^2-value is then often calculated using the data points ($\log_e y_i$, $\log_e \hat{y}_i$). The R^2 is generally interpreted as a measure of goodness of fit of even the original nonlinear model, which is incorrect. Scott and Wild (1991) have given a real example where two models are identical for all practical purposes and yet have very different values of R^2 calculated on the transformed scales.

Kvalseth (1985) has emphasized that, although R_1^2 given by

$$R_1^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2 \qquad \qquad ...(7.1)$$

is quite appropriate even for nonlinear models, other summary statistics like

Mean Absolute Error (MAE) $= \Sigma |y_i - \hat{y}_i| / n,$

Mean Squared Error (MSE) $= \Sigma (y_i - \hat{y}_i)^2 / (n-p),$

should also be computed. Here n is the total number of observed values and p denotes the number of model parameters.

8. Examination of Residuals

Uncritical use and sole reliance on the above statistics may fail to reveal important data characteristics and model inadequacies. Additional detailed analysis of the residuals is strongly recommended to decide about the suitability of a model. Two important assumptions made in the model are:

(i) errors are independent

(ii) errors are normally distributed.

These assumptions can be verified by examining the residuals. If the fitted model is correct, the residuals should exhibit tendencies that tend to confirm or at least should not exhibit a denial of the assumptions. The principal ways of plotting the residuals are: (a) in time sequence, (b) against fitted values. We now discuss the tests for the assumptions (i) and (ii) above.

8.1 Test for independence of errors (Run test).

We test H_o: Errors are independent

Against H_1: Errors are not independent.

Replace a residual by '+' or '−' sign according as it is positive or negative. Let m be the number of pluses and n be the number of minuses in the series of residuals. The test is based on the number of runs (r), where a run is defined as a sequence of symbols of one kind separated by symbols of another kind (Siegel and Castellan, 1988). A good large sample approximation to the sampling distribution of the number of runs is the normal distribution with

Mean $(\mu) = 2mn/(m+n) + 1$

and

Variance $(\sigma^2) = 2mn(2mn - m - n)(m + n)^{-2}(m + n - 1)^{-1}$

Therefore, for large samples the required test statistic is

$Z = (r+h-\mu)/\sigma \sim N(0,1),$

where

$$h = \begin{cases} 0.5, & \text{if } r < \mu \\ -0.5, & \text{if } r > \mu. \end{cases}$$

H_o is rejected at level α if $Z < -Z_\alpha$, where

$Z_\alpha = P\{Z > Z_\alpha\} = \alpha$

8.2 Test for normality (Shapiro-Wilk test, n<50)

We test H_o: Errors are normally distributed

Against H_1: Errors are not normally distributed.

The required test statistic W is defined as
$$W = S^2/b,$$
where

$$S^2 = \Sigma \, a(k) \left\{ x_{(n+1-k)} - x_{(k)} \right\}, \; b = \Sigma (x_i - \overline{X})^2$$

In the above, the parameter k takes the values

$$k = \begin{cases} 1,2,...,n/2 \text{ when n is even} \\ 1,2,...,(n-1)/2 \text{ when n is odd} \end{cases}$$

and $x_{(k)}$ is the kth order statistic of the set of residuals. The values of coefficients a(k) for different values of n and k are given in Table 5 of Shapiro-Wilk (1965). H_o is rejected at level α is W is less than the tabulated value which is given in Table 6 of the above paper.

EXERCISES

Exercise 1: Wheat productivity data (in quintals/hectare) of the country during the years 1973-74 to 1996-97 is as follows:

11.72, 13.38, 14.10, 13.87, 14.80, 15.68, 14.36, 16.30, 16.91, 18.16, 18.43, 18.70, 20.46, 19.16, 20.02, 22.41, 21.21, 22.81, 23.97, 23.27, 23.80, 25.59, 24.93, 26.59.

Fit logistic and Gompertz growth models to this data using 'Levenberg-Marquardt' nonlinear estimation procedure. Compute the mean square error (MSE) in each case and forecast the values for the year 2010 by each model.

Exercise 2: The average weekly aphid count data on Yellow seeded mustard crop at I.A.R.I. farms during the year (1980-81) is as follows:

Week No. (t)	1	2	3	4	5	6	7	8	9	10	11
Average Aphid Population(N(t))	0	0	0	0	0	1.5	2.5	12.1	67.5	175.5	571.2

Week No.(t)	12	13	14	15	16	17	18	19
Average Aphid population (N(t))	926.2	1868.6	6483.8	9016.7	10057.1	8993.3	238.1	0

Fit the nonlinear statistical model corresponding to (5.1) and draw the graph of the fitted model along with data.

SPSS Commands

1. Click Start $\rightarrow$ Program $\rightarrow$ SPSS
2. Open $\rightarrow$ Data$\rightarrow$file $\rightarrow$ RPRA1.SAV(for pract.1)/RPRA2.SAV(for pract.2)
3. Analyze $\rightarrow$ Regression $\rightarrow$ Nonlinear
4. Define dependent variable $\rightarrow$ Nt
5. Write model Expression

Exercise 1: Logistic Model   c/(1+b*exp(-a*t))
        Gompertz Model   c*exp(-b*exp(-a*t))
Exercise 2:   (a*exp(b*t))/(1+c*exp(b*t))**2

6. Parameters→ Name → Starting Value → Then Add → Continue
           Exercise 1: a=0 b=1 c=30
           Exercise 2: a=0 b=0 c=1

7. Select save and tick the option Predicted values and Residuals→ Continue→ OK

Nonparametric Test
1. Click next → Data appears with predicted values and Residuals
2. Analyze →Nonparametric Tests →Runs
3. Resid (Test variable list) → Cut point→mean→ OK →Next

Normality Test
1. Analyze →Summarize →Explore
2. Resid (Dependent list)→ Display→Both→Plot→tick Normality plot with tests→continue→OK

Graph
1. Graph→Scatter→Overlay→ Define
2. Select pair (nt ,t)→Y-X Pairs→again select (pred,t)→Y-X Pairs→OK
3. Edit →select pred. points →option no. 7(line style)→Spline→Apply→Close
4. Select X-axis and double click →Axis title →Range →Minimum →Maximum →OK
5. Select Y-axis and double click →Axis title →Range →Minimum →Maximum →OK
6. Select border →Option no.2 →Select border → Apply→Close

RESULTS

Exercise 1.
        Logistic Model Estimates a = 0.0619   b = 2.2595   c = 40.3909
                          Run test   z = 1.0436
                   Shapiro-Wilks test   w = 0.9632
              Forecast for the year 2010   =32.8604
        Gompertz Model Estimates a=0.0295   b = 1.5271   c = 56.8202
                        Run test   z = 1.0436
                 Shapiro-Wilks test   w = 0.9682
            Forecast for the year 2010   = 34.0294

Exercise 2.   Estimates a = 0.0014   b = 1.1037   c = 3.1852E -08
                         Run test   z = -1.1572
                 Shapiro-Wilks test   w = 0.8117

References and Suggested Reading

Draper, N.R. and Smith, H.(1998). Applied Regression Analysis, 3rd ed. John Wiley

Kvalseth, T.O. (1985). Cautionary note about R^2, *Amer.Statistician*, **39**, 279-85

Prajneshu (1991). Cautionary note about nonlinear models in fisheries, *Ind. J. Fisheries*, **38**, 231-33

Prajneshu (1998). A nonlinear statistical model for aphid population growth, *Jour. Ind. Soc. Ag. Statistics*, **51**, 73-80

Ratkowsky, D.A. (1990). Handbook of nonlinear regression models, Marcel Dekker

Scott, A. And Wild, C.J.(1991). Transformations and R^2, *Amer. Statistician*, **45**, 127-28

Seber, G.A.F. and Wild, C.J. (1989). Nonlinear regression, John Wiley