Nonlinear and Non-parametric Regression Models and their Applications

Shri C.G. Joshy
Scientist, ICAR-CIFT, Cochin
cgjoshy@gmail.com

Introduction

The traditional approach to modeling the relationship between the explanatory/independent variables and the response is regression analysis which assumes that the underlying functional form to express the relationship can be expressed parametrically (Myers, 1990). If the true relationships among the variables known exactly, the researcher would be in a position to understand, predict and control the response. The true relationship among the studied variables, however, will rarely be known, and one must rely on empirical evidence to develop approximations. Here one assumes that the structure of the regression function is known and depends only on finitely many parameters, and one uses the data to estimate the (unknown) values of these parameters. If a parametric estimate cannot approximate the regression function better than the best function which has the assumed parametric structure, the parametric estimate will lead to bad estimate.

This inflexibility concerning the structure of the regression function is avoided by using nonparametric regression analysis which does not assume any function form of relationship between the explanatory variables and response variables (Vining and Bohn, 1998 and Anderson-Cook and Prewitt, 2005). Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates. . If the user misspecifies the parametric model, the estimates may be highly biased, whereas the nonparametric fits may be highly variable, especially in small sample settings.

Linear and Nonlinear Models

A linear model is the one in which all the parameters appear linearly whereas, in a nonlinear model, at least one of the parameters appears nonlinearly. More formally, in a nonlinear model, at least one derivative with respect to a parameter should be a function of that parameter.
The standard functional form of a nonlinear model can be given as

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\theta}) + \varepsilon_i, i = 1, 2, \cdots, N, \qquad 1$$

where Y_i is the dependent variable, $\mathbf{X}_i$ is independent variable and $\boldsymbol{\theta}$ is the vector of parameters and ε_i is the error term. Here, the function 'f' can have any nonlinear functional form.
A nonlinear model, which can be made linear by any suitable transformation, is called "intrinsically linear" (Draper and Smith, 2003). Monomolecular, Logistic and Gompertz are some of the important nonlinear growth models used in fisheries (Bard, 1974).

Some Important Nonlinear Growth Models

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 (Bates *et al*, 1988). In general growth models are mechanistic in nature rather than empirical. 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.

Malthus Model

The standard functional form of Malthus model for computing growth rate over a period of time 't' is

$$Y_t = Y_0(1+r)^t, t = 1,2,\cdots,N, \qquad 2$$

where Y_t is the observation at time t, Y_0 is the value of Y when t = 0 and r is the compound growth rate. A well-known limitation of the above model is that the response variable Y_t tends to infinity as $t \to \infty$, can't happen in reality. We can linearize the Equation (2) by assuming a multiplicative error term on the right side of the equation as exponential of error terms using logarithmic transformation and the parameters of the linearized model are estimated by method of ordinary least squares.

$$Y_t^* = Y_0^* + Bt, \text{where } Y_t^* = \log(Y_t), Y_0^* = \log(Y_0) \text{ and } B = \log(1+r) \qquad (3)$$

Thus, the compound growth rate (r) of the linearized model is computed as $\hat{r} = \exp(\hat{B}) - 1$. The merits and demerits of the log linearized model were discussed in detail by Prajneshu (2005). The above models assume a constant rate of growth over the entire period of time.

Monomolecular Model

The 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.,

$$dY/dt = r(K - Y),$$
$$\frac{dY}{(K-Y)} = r\,dt, \qquad (4)$$

for some r > 0. Integrating Equation (4), we get

$$Y_t = K - (K - Y_0)\exp(-rt), \qquad (5)$$

where t > 0 and r > 0.

Logistic Model

This model postulates that growth takes place at an exponential rate as in Malthusian model. However, subsequently a "deterrent force" comes into play because of crowding effect and does not let system to grow beyond the limits. This model is represented by the differential equation

$$\begin{aligned} dY/dt &= rY(1-Y/K), \\ \frac{K}{Y(K-Y)}dY &= r\,dt, \\ \left(\frac{1}{Y}+\frac{1}{(K-Y)}\right)dY &= r\,dt. \end{aligned} \tag{6}$$

Integrating Equation (6) and applying initial condition $Y(0) = Y_0$, we get

$$Y_t = K/[1+(K/Y_0-1)\exp(-rt)]. \tag{7}$$

The graph of Y_t versus 't' is an elongated S-shaped and the curve is symmetrical about its point of inflexion.

Gompertz Model

This is another model having sigmoid type of behaviour, however, gompertz model is not symmetric about its point of inflexion. The differential equation for this model is

$$\begin{aligned} dY/dt &= rY\log(K/Y), \\ \frac{d}{dt}(Y/K) &= -r(Y/K)\log(Y/K), \\ \frac{dz}{dt} &= -rz\log(z), \text{where } z = Y/K, \\ \frac{dz}{z\log(z)} &= -r\,dt. \end{aligned} \tag{8}$$

A general solution to Equation (8) is obtained by integrating it and applying initial condition $Y(0) = Y_0$. That is,

$$Y_t = K\exp[\log(Y_0/K)\exp(-rt)]. \tag{9}$$

For all the three models, Y_0 is the value of Y_t at t = 0, K is the carrying capacity of the system and r is the intrinsic growth rate. The above three models have been proposed deterministically by adding an error term e_t on the right side of the equations. The error terms are assumed to be independently and identically distributed with constant variance.

Fitting of Nonlinear Statistical Models

The parameters (r, K, Y_0) of the Malthus, monomolecular, logistic and gompertz models appear in nonlinear fashion. The parameters are estimated by minimizing the residual sum of squares using an iterative process that commences with user-supplied starting values and attempts to continually improve on the parameter estimates. The best iterative method to estimate the parameters of nonlinear model is Levenberg-Marquardt algorithm (Seber and Wild, (2003)). We now explain this method in some detail for the model:

$$Y_t = f(x_t, \boldsymbol{\theta}) + e_t, t = 1, 2, \cdots, N,$$

where Y_t is observation at time t, $f(x_t, \theta)$ is any parametric function relating Y_t and x_t, θ is the parameter vector and e_t is the error term. A random error is added to the growth relationship by deterministic approach, which could be reasonable with cross-sectional data in which single measurement is taken on each time period. The residual sums of squares is obtained as

$$S(\theta) = \sum_{t=1}^{N} \left[Y_t - f(x_t, \theta) \right]^2 . \quad (10)$$

Let $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \cdots, \theta_{p0})'$ be the vector of initial parameter values. Then the algorithm for obtaining successive estimates is essentially given by

$$(\mathbf{H} + \tau \mathbf{D})(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) = \mathbf{g}, \quad (11)$$

where

$$\mathbf{g} = \partial S(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \text{ at } \boldsymbol{\theta} = \boldsymbol{\theta}_0 ,$$

$$\mathbf{H} = \partial^2 S(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}' \text{ at } \boldsymbol{\theta} = \boldsymbol{\theta}_0 ,$$

τ is a conditioning/damping factor adjusted at each iteration factor and $\mathbf{D}$ is the diagonal matrix with entries equal to the diagonal elements of $\mathbf{H}$. The Levenberg-Marquardt algorithm is an iterative procedure which minimizes the residual sum of squares $S(\boldsymbol{\theta})$ (Levenberg (1944); Marquardt (1963)). The approximating sum of squares function will have a stationary point when its gradient is zero, that is, when

$$(\boldsymbol{\theta}_0 - \boldsymbol{\theta}) = (\mathbf{H} + \tau \mathbf{D})^{-1} \mathbf{g},$$

$$\delta(\tau) = (\mathbf{H} + \tau \mathbf{D})^{-1} \mathbf{g},$$

and this stationary point will be a minimum if $\mathbf{H}$ is positive definite (all its eigen values positive). This method of estimation requires initial parameter values which were either selected by linearizing the model by ignoring the error term or by graphical method. The parameters of linearized model were estimated using ordinary least square (OLS) method and used as initial estimates.

Nonparametric Regression Analysis

Situations may arise in which the relationship between the explanatory variables and the response is not adequately modeled parametrically. In these situations, any degree of model misspecification may result in serious bias of the estimated response. Furthermore, the optimal regressor settings may be miscalculated. For such situations, nonparametric methods have recently been suggested as they can capture structure in the data that a misspecifed parametric model cannot. Nonparametric smoothing techniques use curves to describe the relationship between the explanatory variables and the response without any parameters. Thus, the dependent variable Y is modeled as:

$$Y_i = m(X_{1i}, X_{2i}, \cdots, X_{pi}) + \varepsilon_i, i = 1, 2, \cdots, N, \quad (12)$$

where m is a function that represents the intrinsic behavior of data, may be of reasonably smooth form; X's are explanatory/independent variables and ε_i is error term assumed to be independently distributed with constant variance σ^2.

Similar to parametric regression, the estimator is a linear combination of the response values; however, the weighting schemes in some nonparametric regression methods assign more weight to observations closest to the point of prediction. The nonparametric fit is more flexible than the parametric fit as it is not

confined to the user's specified form. Myers (1999) suggested the use of nonparametric regression uses in the following scenarios:

i. The researcher is less interested in an interpretive function (i.e., interpreting the estimated regression coefficients) and more interested predicting the value of response variable.
ii. The functional form of the relationship between the explanatory variables and the response is not well behaved.

Several fitting techniques have been proposed in the nonparametric regression literature such as kernel regression (Nadaraya,1964, Watson,1964, Priestley and Chao,1972, and Gasser and MÄuller,1984), local polynomial models (Fan and Gijbels, 1996 and Fan and Gijbels, 2000), spline-based smoothers, and series-based smoothers (Ruppert, Wand, and Carrol, 2003). Details of two popular methods, kernel regression and local polynomial regression, are presented in the next sections.

Kernel Regression

As previously mentioned, nonparametric methods estimate the regression function using a weighted average of the data. For example, the kernel regression estimate of the response at the point of interest x_0 is a weighted average of the responses:

$$\hat{Y}_0^{(KER)} = \frac{\sum_{i=1}^{n} h_{0i}^{(KER)} y_i}{\sum_{i=1}^{n} h_{0i}^{(KER)}},$$

where $h_{0i}^{(KER)}$ represents the weights. A common weighting scheme proposed by Nadaraya (1964) and Watson (1964) in which the weight associated with the i^{th} response at prediction point x_0 is given by:

$$h_{0i}^{(KER)} = \frac{K(\frac{x_0 - x_i}{b})}{\sum_{i=1}^{n} K(\frac{x_0 - x_i}{b})}$$

where K is an univariate kernel function and b is the bandwidth.

The kernel function is taken to be some appropriately chosen decreasing function in $|x_0 - x_i|$ such that observations close to x_0 receive more weight than observations far from x_0. Kernel functions are often chosen to be nonnegative and symmetric about zero. Some common kernel functions include the Gaussian, uniform, and Epanechnikov kernels.

For the multiple regressor case, the linear prediction at $\tilde{x} = (x_1, x_2 \cdots \cdots \cdots x_n)$ is given by

$$\hat{Y}^{KER} = \boldsymbol{L^{(KER)}Y}$$

Where $\hat{Y}^{KER}$ is the kernel hat or smoother matrix defined as

$$L^{(KER)} = \begin{bmatrix} h_1^{(KER)^I} \\ h_2^{(KER)^I} \\ \vdots \\ h_n^{(KER)^I} \end{bmatrix}$$

And $h_i^{KER^I} = \left(h_{i1}^{KER^I}, h_{i2}^{KER^I}, \cdots h_{1in}^{KER^I}\right)$ and $h_{ij}^{(KER)} = \frac{K(\tilde{x}_i, \tilde{x}_j)}{\sum_{j=1}^{n} K(\tilde{x}_i, \tilde{x}_j)}$, more details on multivariate kernel regression is given by Scott (1992) and Simono (1996).

The smoothness of the estimated function is controlled by the bandwidth, b. A larger bandwidth value results in a smoother function, whereas a smaller value results in a less smooth function. However, if the

bandwidth is chosen too large, the estimated function may be too smooth, resulting in estimates with low variance but high bias. On the other hand, a bandwidth that is too small may result in a rougher fit with low bias but high variance. Thus, a bandwidth should be chosen that offers a balance between bias and variance.

The choice of bandwidth is critical, and the literature is rich with bandwidth selection methods (HÄardle, 1990 and HÄardle et al., 2004). The trade-off between bias and variance naturally leads to the minimization of optimality criteria such as mean squared error. The selection of optimum smoothing parameter was done based on the values of generalized cross validation (GCV) (Takezawa, (2006)) and it is given by

$$GCV = \frac{N\hat{\sigma}^2}{(N - Trace(L))^2}, \tag{13}$$

where $\hat{\sigma}^2$ is the estimated mean square error. The optimum smoothing parameter is obtained for lowest values of GCV.

Local Polynomial Regression

A local polynomial equation m(x, x*) of degree 'p', when the value of the predictor (x) is close to (x*) , is given by

$$m(x, x^*) = a_0(x^*) + \sum_{j=1}^{p} a_j(x^*)(x - x^*)^j. \tag{14}$$

The coefficients $a_j(x^*), j = 0, 1, \cdots, p$ are derived by minimizing

$$E_{local}(x^*) = \sum_{i=1}^{N}\left[w\left(\frac{X_i - x^*}{h}\right)\left(Y_i - m(X_i, x^*)\right)^2 \right]$$

$$= \sum_{i=1}^{N}\left[w\left(\frac{X_i - x^*}{h}\right)\left(Y_i - a_0(x^*) - \sum_{j=1}^{p} a_j(x^*)(X_i - x^*)^j \right)^2 \right], \tag{15}$$

where $w\left(\frac{X_i - x^*}{h}\right)$ is a kernel function and h is the bandwidth. We used tri-cube weight function (Takezawa, (2006)) defined by

$$w\left(\frac{X_i - x^*}{h}\right) = \begin{cases} \left[1 - \left(\frac{|X_i - x^*|}{h}\right)^3\right] & \text{if } \left(\frac{|X_i - x^*|}{h}\right)^3 \le 1 \\ 0 & \text{if } \left(\frac{|X_i - x^*|}{h}\right)^3 > 1 \end{cases}$$

The Equation (2.57) can be written as

$$E_{local}(x^*) = (\mathbf{Xa} - \mathbf{Y})' \mathbf{W} (\mathbf{Xa} - \mathbf{Y}), \tag{16}$$

where **W** is a $N \times N$ diagonal matrix containing the kernel weights associated with **X**,

$$\mathbf{W} = \begin{pmatrix} w\left(\frac{X_1 - x^*}{h}\right) & 0 & 0 & \cdots & 0 \\ 0 & w\left(\frac{X_2 - x^*}{h}\right) & 0 & \cdots & 0 \\ 0 & 0 & w\left(\frac{X_3 - x^*}{h}\right) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w\left(\frac{X_N - x^*}{h}\right) \end{pmatrix}.$$

Differentiating Equation (2.58) with respect to $a_j(x^*), j = 0, 1, \cdots, p$ and set to zero, then we get

$$\hat{\mathbf{a}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}. \tag{17}$$

The local polynomial regression fit is given by

$$\hat{\mathbf{Y}}^{\mathbf{LPR}} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} = \mathbf{L}\mathbf{Y}, \tag{18}$$

where **L** is the local linear HAT or smoother matrix, which transforms the observed $\mathbf{Y}$ values into the $\hat{\mathbf{Y}}$ values, can be defined as

$$\mathbf{L}^{(\mathrm{LPR})} = \begin{bmatrix} \mathbf{L}_1^{(\mathrm{LPR})'} \\ \mathbf{L}_2^{(\mathrm{LPR})'} \\ \vdots \\ \mathbf{L}_N^{(\mathrm{LPR})'} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{x}}_1'(\mathbf{X}'\mathbf{W}_1\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_1 \\ \tilde{\mathbf{x}}_2'(\mathbf{X}'\mathbf{W}_2\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_2 \\ \vdots \\ \tilde{\mathbf{x}}_N'(\mathbf{X}'\mathbf{W}_N\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_N \end{bmatrix},$$

$\tilde{\mathbf{x}}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})$ and $\mathbf{W}_i$ is a $N \times N$ diagonal matrix containing the weights associated with $\tilde{\mathbf{x}}_i$.

The variance of $\hat{\mathbf{Y}}^{\mathbf{LPR}}$ is given by:

$$\begin{aligned} \mathrm{Var}(\hat{\mathbf{Y}}^{\mathbf{LPR}}) &= \mathrm{Var}(\mathbf{L}\mathbf{Y}) \\ &= \mathbf{L}\mathrm{Var}(\mathbf{Y})\mathbf{L}' \\ &= \sigma^2\mathbf{L}\mathbf{L}'. \end{aligned} \tag{19}$$

The local linear regression fit at the prediction point $\mathbf{x}_0' = (x_{10}, x_{20}, \cdots, x_{p0})$ to predict $\mathbf{Y}$ is given by

$$\hat{Y}_0^{LPR} = \mathbf{x}_0' \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{Y} = \mathbf{L}_0^{(LPR)'}\mathbf{Y}. \quad (20)$$

Since the LPR estimates are dependent on the kernel weights, bandwidth selection remains important. For more details on local polynomial regression, see for example Fan and Gijbels (1996) and Fan and Gijbels (2000). The selection of optimum smoothing parameter was done based on the values of generalized cross validation (GCV) (Takezawa, (2006)).

References

1. Bard, Y. (1974). *Nonlinear Parameter Estimation*. Academic Press, New York.
2. Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Application*. John Wiley, New York.
3. Cleveland, W.S., (1979). Robust locally weighted regression and smoothing scatterplots." *Journal of the American Statistical Association*, 74, 829-836.
4. France, J. and Thornley, J. H. M. (2006). *Mathematical Models in Agriculture*. Butterworths, London.
5. Gallant, A. R. (1987). *Nonlinear Statistical Models*. John Wiley, New York.
6. HÄardle, W., (1990). *Applied Non-Parametric Regression*. Cambridge University Press, Cambridge.
7. Hastie, T.J. and Tibshirani, R.J., (1990). *Generalized Additive Models*. Chapman & Hall, London.
8. Kvalseth, T. O. (1985). Cautionary notes about R^2. *The American Statistician*, **39(4)**, 279-285.
9. Marquardt, D. W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of Society of Industrial and Applied Mathematics*, **11**, 431-441.
10. Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Its Applications* 9, 141-142.
11. Seber, G. A. F. and Wild, C. J. (2003). *Nonlinear Regression*. John Wiley and Sons, New York.
12. Watson, G., (1964). Smoothing regression analysis." *Sankhya Series A* 26, 359-372.