Nonlinear regression

Gordon K. Smyth

Volume 3, pp 1405–1411

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Nonlinear regression

The basic idea of nonlinear regression is the same as that of linear regression, namely to relate a response Y to a vector of predictor variables $\mathbf{x} = (x_1, \ldots, x_k)^T$ (*see* **Linear models**). Nonlinear regression is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters. Whereas linear regression is often used for building a purely empirical model, nonlinear regression usually arises when there are physical reasons for believing that the relationship between the response and the predictors follows a particular functional form. A nonlinear regression model has the form

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where the Y_i are responses, f is a known function of the covariate vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$ and the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$, and ε_i are random errors. The ε_i are usually assumed to be uncorrelated with mean zero and constant variance.

Example 1: Retention of Pollutants in Lakes Consider the problem of build-up of a pollutant in a **lake**. If the process causing retention occurs in the water column, it is reasonable to assume that the amount retained per unit time is proportional to the concentration of the substance and the volume of the lake. If we can also assume that the lake is completely mixed, then an equilibrium equation leads to

$$Y_i = 1 - \frac{1}{1 + \theta x_i} + \varepsilon_i \tag{2}$$

where Y_i is retention of the ith lake, x_i is the ratio of lake volume to water discharge (the hydraulic residence time of the lake) and θ is an unknown parameter [5].

The unknown parameter vector $\boldsymbol{\theta}$ in the nonlinear regression model is estimated from the data by minimizing a suitable goodness-of-fit expression with respect to $\boldsymbol{\theta}$. The most popular criterion is the sum of squared residuals

$$\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})]^2$$

and estimation based on this criterion is known as nonlinear **least squares**. If the errors ε_i follow a normal distribution, then the least squares estimator for $\boldsymbol{\theta}$ is also the **maximum likelihood estimator**. Except in a few isolated cases, nonlinear regression estimates must be computed by iteration using optimization methods to minimize the goodness-of-fit expression.

The definition of nonlinearity relates to the unknown parameters and not to the relationship between the covariates and the response. For example the quadratic regression model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \tag{3}$$

is considered to be linear rather than nonlinear because the regression function is linear in the parameters β_j and the model can be estimated by using classical linear regression methods.

Practical introductions to nonlinear regression including many data examples are given by Ratkowsky [8] and by Bates and Watts [3]. A more extensive treatment of nonlinear regression methodology is given by Seber and Wild [9]. See also Section 15.5 [7]. Most major statistical software programs include functions to perform nonlinear regression.

Common Models

One of the most common nonlinear models is the exponential decay or exponential growth model

$$f(x, \boldsymbol{\theta}) = \theta_1 \exp(-\theta_2 x) \tag{4}$$

(*see* **Logarithmic regression**). This model can be characterized by the fact that the function f satisfies the first-order differential equation

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial x} = c f(x, \boldsymbol{\theta}) \tag{5}$$

for some constant c. Physical processes can often be modelled by higher-order differential equations, leading to higher-order exponential function models, of the form

$$f(x, \boldsymbol{\theta}) = \theta_1 + \sum_{j=1}^{k} \theta_{2j} \exp(-\theta_{2j+1} x) \tag{6}$$

where k is the order of the differential equation. See, for example, [3, Chapter 6] or [9, Chapter 8].

Another common form of model is the rational function:

$$f(x, \boldsymbol{\theta}) = \frac{\sum_{j=1}^{k} \theta_j x^{j-1}}{1 + \sum_{j=1}^{m} \theta_{k+j} x^j} \qquad (7)$$

Rational functions are very flexible in form and can be used to approximate a wide variety of functional shapes.

In many applications the systematic part of the response is known to be monotonic increasing in x, where x might represent time or dosage. Nonlinear regression models with this property are called growth models (*see* **Age–growth modeling**). The simplest growth model is the exponential growth model (4), but pure exponential growth is usually short-lived. A more generally useful growth curve is the logistic curve

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1}{1 + \theta_2 \exp(-\theta_3 x)} \qquad (8)$$

This produces a symmetric growth curve which asymptotes to θ_1 as $x \to \infty$ and to zero as $x \to -\infty$. Of the two other parameters, θ_2 determines horizontal position or 'take-off point', and θ_3 controls steepness. The Gompertz curve produces an asymmetric growth curve

$$f(x, \boldsymbol{\theta}) = \theta_1 \exp[-\theta_2 \exp(-\theta_3 x)] \qquad (9)$$

As with the logistic curve, θ_1 sets the asymptotic upper limit, θ_2 determines horizontal position, and θ_3 controls steepness. Despite these interpretations, it can often be difficult in practice to isolate the interpretations of individual parameters in a nonlinear regression model because of high correlations between the parameter estimators. More examples of growth models are given in [9, Chapter 7].

Transformably Linear Models

Some simple nonlinear models can be converted into linear models by transforming one or both of the responses and the covariates. For example, the exponential decay model

$$Y = \theta_1 \exp(-\theta_2 x) \qquad (10)$$

can, if $Y > 0$, be transformed into

$$\ln Y = \ln \theta_1 - \theta_2 x \qquad (11)$$

If all the observed responses are positive and variation in $\ln Y$ is more symmetric than variation in Y, then it is likely to be more appropriate to estimate the parameters θ_1 and θ_2 by linear regression of $\ln Y$ on x rather than by nonlinear least squares. The same considerations apply to the simple rational function model

$$Y = \frac{1}{1 + \theta x} \qquad (12)$$

which can, if $Y \neq 0$, be linearized as

$$\frac{1}{Y} - 1 = \theta x \qquad (13)$$

One can estimate θ by proportional linear regression (without an intercept) of $(1/Y) - 1$ on x. Transformably linear models have some advantages in that the linearized form can be used to obtain starting values for iterative computation of the parameter estimates. Transformation to linearity is only a possibility for the simplest of nonlinear models, however.

Iterative Techniques

If the function f is continuously differentiable in $\boldsymbol{\theta}$, then it can be linearized locally as

$$f(\boldsymbol{\theta}, \mathbf{x}) = f(\boldsymbol{\theta}_0, \mathbf{x}) + \mathbf{X}_0(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \qquad (14)$$

where $\mathbf{X}_0$ is the $n \times p$ gradient matrix with elements $\partial f(\mathbf{x}_i, \boldsymbol{\theta}_0)/\partial \theta_j$. This leads to the Gauss–Newton algorithm for estimating $\boldsymbol{\theta}$,

$$\theta_1 = \boldsymbol{\theta}_0 + (\mathbf{X}_0^{\mathrm{T}} \mathbf{X}_0)^{-1} \mathbf{X}_0^{\mathrm{T}} \mathbf{e} \qquad (15)$$

where $\mathbf{e}$ is the vector of working **residuals** $y_i - f(\mathbf{x}_i, \boldsymbol{\theta}_0)$. The Gauss–Newton algorithm increments the working estimate $\boldsymbol{\theta}$ at each iteration by an amount equal to the coefficients from the linear regression of the current residuals $\mathbf{e}$ on the current gradient matrix $\mathbf{X}$. If the errors ε_i are independent and normally distributed, then the Gauss–Newton algorithm is an application of Fisher's method of scoring for obtaining maximum likelihood estimators.

If $\mathbf{X}$ is of full column rank in a neighborhood of the least squares solution, then it can be shown that the Gauss–Newton algorithm will converge to the solution from a sufficiently good starting value [6]. There is no guarantee, though, that the algorithm will converge from values further from the solution. In practice, it is usually necessary to modify the Gauss–Newton algorithm in order to secure convergence.

Example 2: PCB in Lake Trout Data on the concentration of Polychlorinated biphenyl (PCB) residues in a series of lake trout from Cayuga Lake, New York, were reported in Bache et al [1]. The ages of the fish were accurately known because the fish were annually stocked as yearlings and distinctly marked as to year class. Each whole fish was mechanically chopped and ground and a 5-g sample taken. The samples were treated and PCB residues in parts per million (ppm) were estimated by means of column chromatography. A combination of empirical and theoretical considerations leads us to consider the model

$$\ln[\text{PCB}] = \theta_1 + \theta_2 x^{\theta_3} + \varepsilon \tag{16}$$

in which [PCB] is the concentration of PCB, and Λ is age. The natural logarithm of [PCB] is modeled as a constant plus an exponential growth model in terms of $\ln(x)$. This model was estimated from the data with using least squares, and the sequence of working estimates resulting from the Gauss–Newton algorithm are given in Table 1. The iteration was started at $\theta_3 = 0.5$ whereas the least squares solution occurs at $\theta_3 = 0.19$. The other two parameters were initialized at their optimal values given $\theta_3 = 0.5$. It can be seen that the iteration initially overshoots markedly before oscillating about the solution and then finally converging. The algorithm fails to converge from starting values too far from the least squares estimates. For example the algorithm diverges from $\theta_3 = 1.0$. The data and fitted values are plotted in Figure 1.

Table 1 Gauss–Newton iteration, starting from $\theta_3 = 0.5$, for PCB in lake trout data

Iteration	θ_1	θ_2	θ_3
1	0.0315	0.2591	0.5000
2	−2.7855	2.6155	−1.0671
3	3.7658	−3.8699	−2.8970
4	2.2455	−2.3663	1.9177
5	−0.1145	0.1698	1.9426
6	−0.0999	0.1619	1.6102
7	−0.3286	0.3045	0.9876
8	−1.1189	0.9774	0.1312
9	−2.9861	2.8261	0.6888
10	−1.9183	1.7527	0.5595
11	−2.4668	2.2974	0.3316
12	−3.9985	3.8307	0.1827
13	−4.7879	4.6254	0.2033
14	−4.8760	4.7126	0.1964
15	−4.8654	4.7023	0.1968

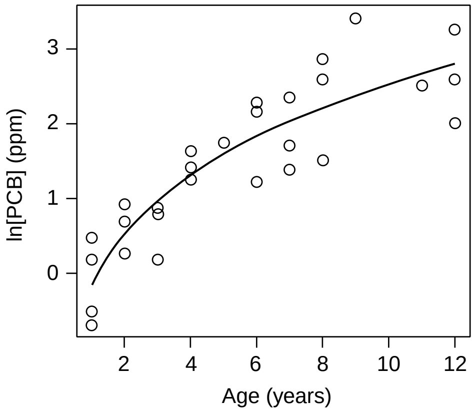

Figure 1 Concentration of PCB as a function of age for the lake trout data

The basic idea behind modified Gauss–Newton algorithms is that care is taken at each iteration to ensure that the update does not overshoot the desired solution. A smaller step is taken if the proposed update would lead to an increase in the residual sum of squares. There are two ways in which a smaller step can be taken: line-search methods and Levenberg–Marquart damping. The key to line-search algorithms is the fact that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is a positive definite matrix. This guarantees that $(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{e}$ is a descent direction; that is, the sum of squared residuals will be reduced by the step from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_0 + \alpha(\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0)^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{e}$, $\alpha > 0$, where α is sufficiently small. The line-search method consists of using one-dimensional **optimization** techniques to minimize the sum of squares with respect to α at each iteration. This method reduces the sum of squares at every iteration and is therefore guaranteed to converge unless rounding error intervenes.

Even easier to implement than line searches and similarly effective is Levenberg–Marquardt damping. Given any positive definite matrix $\mathbf{D}$, the sum of squares will be reduced by the step from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_0 + (\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0 + \lambda\mathbf{D})^{-1}\mathbf{X}_0^{\mathrm{T}}\mathbf{e}$, if λ is sufficiently large. The matrix $\mathbf{D}$ is usually chosen to be either the diagonal part of $\mathbf{X}_0^{\mathrm{T}}\mathbf{X}_0$ or the identity matrix. In practice, λ is increased as necessary to ensure a reduction in the sum of squares at each iteration, and is otherwise decreased as the algorithm converges to the solution.

Although the Gauss–Newton algorithm and its modifications are the most popular algorithms for nonlinear least squares, it is sometimes convenient to use derivative-free methods to minimize the sum of squares and in so doing to avoid computing the gradient matrix **X**. Possible algorithms include the Nelder–Mead simplex algorithm and the pseudo-Newton–Raphson method with numerical derivatives (*see* **Optimization**).

Inference

Suppose that the ε_i are uncorrelated with mean zero and variance σ^2. Then the least squares estimators $\hat{\boldsymbol{\theta}}$ are asymptotically normal with mean $\boldsymbol{\theta}$ and covariance matrix $\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ [6]. The variance σ^2 is usually estimated by

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})]^2 \tag{17}$$

Standard errors and confidence intervals for the parameters can be obtained from the estimated covariance matrix $s^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ with **X** evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. In practice, the linear approximations that the standard errors and confidence intervals are based on can be quite poor [3]. The approximations tend to be more reliable when the sample size n is large or when the error variance σ^2 is small.

Hypotheses about the parameters can also be tested using F-statistics obtained from differences in sums of squares. Suppose for example that

$$f(x, \boldsymbol{\theta}) = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x) \tag{18}$$

and we wish to determine whether it is necessary to retain the second exponential term in the model. Let SS_1 be the residual sum of squares with one exponential term in the model (i.e. with $\theta_3 = 0$) and SS_2 the residual sum of squares with two exponential terms. Then

$$F = \frac{\mathrm{SS}_1 - \mathrm{SS}_2}{2s^2} \tag{19}$$

follows approximately an F-distribution on 2 and $n-4$ degrees of freedom. This is closely analogous to the corresponding F-distribution result for linear regression. Tests and confidence regions based on the residual sum of squares are generally more reliable than tests or confidence intervals based on standard errors. If the ε_i follow a normal distribution then tests based on F-statistics are equivalent to tests based on likelihood ratios (*see* **Likelihood ratio tests**).

Example 3: PCB in Lake Trout The least squares estimates and associated standard errors for the model represented by (16) are given in Table 2. Notice the large standard errors for the parameters, which are caused largely by high correlations between the parameters. The correlations obtained from the off-diagonal elements of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ are in fact 0.997 between $\hat{\theta}_1$ and $\hat{\theta}_3$, -0.998 between $\hat{\theta}_2$ and $\hat{\theta}_3$, and -0.9998 between $\hat{\theta}_1$ and $\hat{\theta}_2$. A 95% confidence interval for θ_3 based on its standard error would suggest that θ_3 is not significantly different from zero. However, if θ_3 is set to zero then $f(x, \boldsymbol{\theta})$ no longer depends on age, and the sum of squared residuals increases from 6.33 to 31.12. For these data, $s^2 = 6.33/25 = 0.253$, and the F-test statistic for testing $\theta_3 = 0$ is $F = (31.12 - 6.33)/(2s^2) = 48.95$, for 2 and 25 degrees of freedom. The null hypothesis of θ_3 is soundly rejected.

Separable Least Squares

Even in nonlinear regression models, many of the parameters enter the model linearly. For example, consider the exponential growth model

$$f(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x) \tag{20}$$

For any given value for θ_3, values for θ_1 and θ_2 can be obtained from linear regression of Y on $\exp(\theta_3 x)$. The parameters θ_1 and θ_2 are said to be 'conditionally linear' [3]. For any given value of θ_3 the least squares estimators of θ_1 and θ_2 are available in a closed-form expression involving θ_3. In this sense, θ_3 is the only nonlinear parameter in the model. Estimation of the parameters can be greatly simplified if the expression for the conditionally linear parameters is substituted

Table 2 Least squares estimates and standard errors for PCB in lake trout data

Parameter	Value	Standard error
θ_1	−4.8647	8.4243
θ_2	4.7016	8.2721
θ_3	0.1969	0.2739

into the estimation process. Regression models with conditionally linear parameters are called 'separable' because the linear parameters can be separated out of the least squares problem in this way.

Any separable model can be written in the form

$$f(\mathbf{x}, \theta) = \mathbf{X}(\boldsymbol{\alpha})\boldsymbol{\beta} \tag{21}$$

where $\boldsymbol{\alpha}$ is a vector of nonlinear parameters, $\mathbf{X}$ is a matrix of full-column rank depending on $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ is a vector of conditionally linear parameters. The conditional least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\alpha}) = [\mathbf{X}(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{X}(\boldsymbol{\alpha})]^{-1}\mathbf{X}(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{Y} \tag{22}$$

The nonlinear parameters can be estimated by minimizing the reduced sum of squares

$$[\mathbf{Y} - X(\boldsymbol{\alpha})\hat{\boldsymbol{\beta}}]^{\mathrm{T}}[\mathbf{Y} - X(\boldsymbol{\alpha})\hat{\boldsymbol{\beta}}] = \mathbf{Y}^{\mathrm{T}}\mathbf{P}\mathbf{Y} \tag{23}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{X}(\boldsymbol{\alpha})[\mathbf{X}(\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{X}(\boldsymbol{\alpha})]^{-1}\mathbf{X}(\boldsymbol{\alpha})^{\mathrm{T}}$ is the orthogonal projection onto the space orthogonal to the column space of $\mathbf{X}(\boldsymbol{\alpha})$. Several algorithms have been suggested to minimize the reduced sum of squares [9, Section 14.7], the simplest of which is probably the nested Gauss–Newton algorithm [10]:

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^{k} + (\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{P}\mathbf{y} \tag{24}$$

This iteration is equivalent to performing an iteration of the full Gauss–Newton algorithm and then resetting the conditionally linear parameters to their conditional least squares values. Separable algorithms can be modified by using line searches or Levenberg–Marquardt damping in the same way as the full Gauss–Newton algorithm.

Example 4: PCB in Lake Trout Table 3 gives the results of the nested Gauss–Newton iteration from the same starting values as previously used for the full Gauss–Newton algorithm. The nested Gauss–Newton algorithm converges far more rapidly. It also converges from a much wider range of starting values, for example from $\theta_3 = 1$.

Table 3 Conditionally linear Gauss–Newton iteration, starting from $\theta_3 = 0.5$, for PCB in lake trout data

Iteration	θ_1	θ_2	θ_3
1	−1.1948	1.1986	0.5000
2	−6.1993	6.0181	0.1612
3	−4.7777	4.6161	0.1997
4	−4.8740	4.7107	0.1966
5	−4.8657	4.7026	0.1968

Measures of Nonlinearity

Since most asymptotic inference for nonlinear regression models is based on analogy with linear models, and since this inference is only approximate insofar as the actual model differs from a linear model, various measures of nonlinearity have been proposed as a guide for understanding how good linear approximations are likely to be.

One class of measures focuses on curvature of the function f and is based on the sizes of the second derivatives $\partial^2 f/\partial\theta_j\partial\theta_k$. If the second derivatives are small in absolute value, then the model is approximately linear. Let $\mathbf{u} = (u_1, \dots, u_n)^{\mathrm{T}}$ be a given vector representing a direction of interest and write

$$A_{jk} = \sum_{i=1}^{n} u_i \frac{\partial^2 f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial\theta_j\partial\theta_k} \tag{25}$$

The matrix $\mathbf{A} = \{A_{jk}\}$ represents the size of the second derivatives in the direction $\mathbf{u}$. To obtain curvature measures it is necessary to standardize the derivatives by dividing by the size of the gradient matrix. Let $\mathbf{X} = \mathbf{QR}$ be the QR decomposition of the gradient matrix $\mathbf{X} = \partial f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})/\partial\theta_j$, and write

$$\mathbf{C} = (\mathbf{R}^{-1})^{\mathrm{T}}\mathbf{A}\mathbf{R}^{-1} \tag{26}$$

Bates and Watts [2] use $\mathbf{C}$ to define relative curvature measures of nonlinearity. If $\mathbf{u}$ belongs to the range space of $\mathbf{X}$ ($\mathbf{u} = \mathbf{X}\boldsymbol{\beta}$ for some $\boldsymbol{\beta}$), then $\mathbf{C}$ defines parameter effect curvatures. Bates and Watts [3] consider an orthonormal basis of vectors $\mathbf{u}$ for the range space of $\mathbf{X}$ and interpret the resulting curvatures in geometric terms. If $\mathbf{X}^{\mathrm{T}}\mathbf{u} = \mathbf{0}$ then $\mathbf{C}$ defines intrinsic curvatures. It can be shown that intrinsic curvatures depend only on the shape of the expectation surface $f(\mathbf{x}, \boldsymbol{\theta})$ as $\boldsymbol{\theta}$ varies and are invariant under reparameterizations of the nonlinear model; see [3, Chapter 7] or [9, Chapter 4].

Intrinsic curvature in the direction defined by the residuals, with $u_i = y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}})$, is particularly important. In this case, Bates and Watts [3] call $\mathbf{C}$ the 'effective residual curvature matrix'. The eigenvalues of this matrix can be used to obtain improved confidence regions for the parameters [4]. The largest eigenvalue in absolute size determines the limiting rate of convergence of the Gauss–Newton algorithm [10].

Another method of assessing nonlinearity is to vary one component of $\boldsymbol{\theta}$ and to observe how the profile sum of squares and conditional estimators vary. Let $\hat{\boldsymbol{\theta}}_{(j)}(\theta_{j0})$ be the least squares estimator of $\boldsymbol{\theta}$ subject to the constraint $\theta_j = \theta_{j0}$. Bates and Watts [3] study nonlinearity by observing the rate of change in $\hat{\boldsymbol{\theta}}_{(j)}(\theta_{j0})$ and in the sum of squares as θ_{j0} varies.

Robust and Generalized Nonlinear Regression

This entry has concentrated on least squares estimation, but it may be of interest to consider other estimation criteria in order to accommodate **outliers** or non-normal responses. Stromberg and Ruppert [11, 12] have considered high-breakdown nonlinear regression. Wei [13] gives an extensive treatment of generalized nonlinear regression models with exponential family responses. In particular, Wei [13] extends curvature measures of nonlinearity to this more general context and uses them for second-order asymptotics.

References

[1] Bache, C.A., Serum, J.W., Youngs, W.D. & Lisk, D.J. (1972). Polychlorinated biphenyl residues: accumulation in Cayuga Lake trout with age, *Science* **117**, 1192–1193.

[2] Bates, D.M. & Watts, D.G. (1980). Relative curvature measures of nonlinearity (with discussion), *Journal of the Royal Statistical Society, Series B* **42**, 1–25.

[3] Bates, D.M. & Watts, D.G. (1988). *Nonlinear Regression Analysis and Its Applications*, Wiley, New York.

[4] Hamilton, D.C., Watts, D.G. & Bates, D.M. (1982). Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions, *The Annals of Statistics* **10**, 386–393.

[5] Grimvall, A. & Stålnacke, P. (1996). Statistical methods for source apportionment of riverine loads of pollutants, *Environmetrics* **7**, 201–213.

[6] Jennrich, R.I. (1969). Asymptotic properties of nonlinear least squares estimators, *The Annals of Mathematical Statistics* **40**, 633–643.

[7] Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in Fortran*, Cambridge University Press, Cambridge (also available for C, Basic and Pascal).

[8] Ratkowsky, D.A. (1983). *Nonlinear Regression Modelling: A Unified Practical Approach*, Marcel Dekker, New York.

[9] Seber, G.A.F. & Wild, C.J. (1989). *Nonlinear Regression*, Wiley, New York.

[10] Smyth, G.K. (1996). Partitioned algorithms for maximum likelihood and other nonlinear estimation, *Statistics and Computing* **6**, 201–216.

[11] Stromberg, A.J. (1993). Computation of high breakdown nonlinear regression parameters, *Journal of the American Statistical Association* **88**, 237–244.

[12] Stromberg, A.J. & Ruppert, D. (1992). Breakdown in nonlinear regression, *Journal of the American Statistical Association* **87**, 991–997.

[13] Wei, B.-C. (1998). *Exponential Family Nonlinear Models*, Springer-Verlag, Singapore.

(*See also* **Generalized linear mixed models**; **Nonparametric regression model**)

Gordon K. Smyth