Nonlinear Regression Analysis

Chapter · October 2022

DOI: 10.1016/b978-0-12-818630-5.10068-5

2 authors:

Hsin-Hsiung Huang
University of Central Florida
55 PUBLICATIONS **469** CITATIONS
SEE PROFILE

Qing He
University of Central Florida
10 PUBLICATIONS **54** CITATIONS
SEE PROFILE

Nonlinear Regression Analysis

Hsin-Hsiung Huang and Qing He

Department of Statistics and Data Science, University of Central Florida, United States

Abstract

Nonlinear regression analysis is a popular and important tool for scientists and engineers. In this article, we introduce theories and methods of nonlinear regression and its statistical inferences using the frequentist and Bayesian statistical modeling and computation. Least squares with the Gauss-Newton method is the most widely used approach to parameters estimation. Under the assumption of normally distributed errors, maximum likelihood estimation is equivalent to least squares estimation. The Wald confidence regions for parameters in a nonlinear regression model are affected by the curvatures in the mean function. Furthermore, we introduce the Newton-Raphson method and the generalized least squares method to deal with variance heterogeneity. Examples of simulation data analysis are provided to illustrate important properties of confidence regions and the statistical inferences using the nonlinear least squares estimation and Bayesian inference.

Keywords: Nonlinear regression, variance heterogeneity, iteratively reweighted least squares, Gauss-Newton, Newton-Raphson, generalized least squares

1. Introduction

1.1. An overview of the nonlinear regression models

Nonlinear regression models have been widely used in various fields including statistics, chemistry, physics, psychology, health science, and biology. Some of them have a linear relationship in the parameters (i.e., linear in the $\boldsymbol{\theta}$). For example, a polynomial regression fits a curved relationship between the response variable and predictors using higher-ordered values of the predictors, but it is linear in terms of the parameters. Statistical inference can be derived from nonlinear regression models (1). The main advantages of nonlinear regression models include interpretability, parsimony, and prediction (Bates and Watts, 1988). In general, nonlinear models are capable of accommodating various mean functions, so nonlinear models are appropriate for applications with parsimonious parameters and easily interpretable due to the fact that the parameters can be associated with meaningful factors.

Assume that there are n observations of responses $y_1, \ldots, y_n$ and predictors $\mathbf{x}_{ik}$ associated with y_i on $k = 1, \ldots, p$ independent variables. In nonlinear regression, the y's and $\mathbf{x}$'s satisfy the nonlinear regression model

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

where f is a function that is nonlinear in the p elements of parameter $\boldsymbol{\theta}$ with the unknown true value $\boldsymbol{\theta}^*$, and the ϵ_i's are independent and identically distributed (i.i.d) random errors with mean zero and constant variance σ^2. The conditional mean response $\mathrm{E}(y_i|\mathbf{x}_i) = f(\mathbf{x}_i, \boldsymbol{\theta})$, so we will refer to f as the mean function. For example, the Beverton-Holt model (Beverton and Holt, 2012) which is similar to the Michaelis-Menten model (Sheiner and Beal, 1980) and used for modeling discrete-time population gives the expected population $y = n_{t+1}$ as a function of the previous population $\mathbf{x} = n_t$ as

$$n_{t+1} = f(\mathbf{x}, (\alpha, \beta)) = \frac{\alpha \mathbf{x}}{1 + \mathbf{x}/\beta},$$

where $\boldsymbol{\theta} = (\alpha, \beta)$ and the parameter α is the slope at 0 and β is the concentration between 0 and the upper limit $\alpha\beta$. It is a nonlinear regression since the two variables are related in a nonlinear (curved) relationship.

1.2. Estimation methods

In order to estimate the parameters in a nonlinear regression model, we minimize the sum of squared residuals, which measures how far the y observations vary from the nonlinear function $f(\mathbf{x}_i, \boldsymbol{\theta})$ used to predict y. Such estimators are called the least squares estimators $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, that is,

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

is used to find

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta}).$$

For example, to estimate α and β in the Beverton-Holt model, we minimize $S(\alpha, \beta) = \sum (y_i - \frac{\alpha \mathbf{x}_i}{1+\mathbf{x}_i/\beta})^2$. We first derive the gradients g_α, g_β and Hessian matrix with elements $H_{i,j},\ i, j \in \{\alpha, \beta\}$ as follows:

$$\begin{aligned}
g_\alpha &= \frac{\partial S(\boldsymbol{\theta})}{\partial \alpha} = -2 \sum \mathbf{x}_i (y_i - \frac{\alpha \mathbf{x}_i}{1 + \mathbf{x}_i/\beta}), \\
g_\beta &= \frac{\partial S(\boldsymbol{\theta})}{\partial \beta} = -2 \sum (y_i - \frac{\alpha \mathbf{x}_i}{1 + \mathbf{x}_i/\beta}) \frac{a \mathbf{x}_i^2}{\beta^2 (1 + \mathbf{x}_i/\beta)^2}, \\
H_{\alpha,\alpha} &= \frac{\partial g_\alpha}{\partial \alpha} = 2 \sum \frac{\mathbf{x}_i^2}{1 + \mathbf{x}_i/\beta}, \\
H_{\alpha,\beta} &= \frac{\partial g_\alpha}{\partial \beta} = 2 \sum \frac{\alpha \mathbf{x}_i^3}{\beta^2 (1 + \mathbf{x}_i/\beta)^2}, \\
H_{\beta,\alpha} &= -2 \sum \frac{\mathbf{x}_i^2 y_i}{\beta^2 (1 + \mathbf{x}_i/\beta)^2} - \frac{\mathbf{x}_i^3}{\beta^2 (1 + \mathbf{x}_i/\beta)^3}, \text{ and} \\
H_{\beta,\beta} &= -2 \sum \frac{-2a \mathbf{x}_i^2 y_i}{(\beta + \mathbf{x}_i)^3} + \frac{2a^2 \mathbf{x}_i^3}{\beta^3 (1 + \mathbf{x}_i/\beta)^3} - \frac{3a^2 \mathbf{x}_i^4}{\beta^4 (1 + \mathbf{x}_i/\beta)^4}.
\end{aligned}$$

Then we can use the Gauss-Newton or Newton-Raphson methods to iteratively estimate α and β introduced in Section 2 and 4, respectively. The Gauss-Newton method is essentially the Newton-Raphson method with the modification that the Gauss-Newton method uses the approximation $2J^T J$, where J is the Jacobian matrix, for the Hessian matrix. Note that when for a scalar nonlinear function its Jacobian matrix is same as the gradient.

2. Gauss-Newton method for least squares estimation

When f is twice differentiable with respect to $\boldsymbol{\theta}$, we have the gradient and Hessian which can be written as:

$$\begin{aligned}
g_j &= \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} = -2 \sum_{i=1}^{n} (y_i - f_i) \frac{\partial f_i}{\partial \theta_j}, \\
H_{jk} &= \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} = -2 \sum_{i=1}^{n} \left((y_i - f_i) \frac{\partial^2 f_i}{\partial \theta_j \partial \theta_k} - \frac{\partial f_i}{\partial \theta_j} \frac{\partial f_i}{\partial \theta_k} \right)
\end{aligned}$$

where $f_i = f(\mathbf{x}_i, \boldsymbol{\theta})$.

Further, we solve for the least squares solution $\widehat{\boldsymbol{\theta}}$ in the following system of equations

$$\left. \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} = 0;\ j = 1, \ldots, p, \tag{2}$$

which are called the normal equations with

$$\sum_{i=1}^{n} (y_i - f_i) \left. \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}} = 0.$$

Its matrix form is

$$\mathbf{J}(\widehat{\boldsymbol{\theta}})^T\mathbf{r} = 0$$

with (i, j)-th element $J_{i,j}(\widehat{\boldsymbol{\theta}}) = \frac{\partial f_i}{\partial \theta_j}$ and $\mathbf{r} = y - f(\widehat{\boldsymbol{\theta}})$ with the i-th element $r_i = y_i - f_i$. The matrix $\mathbf{J}(\widehat{\boldsymbol{\theta}})$ of size $n \times p$ is called the Jacobian matrix. The $p \times p$ matrices $G_1, \ldots, G_n$ so that the (j, k)-th element of G_i is $\frac{\partial^2 f_i}{\partial \theta_j \partial \theta_k}$, and then the gradient and Hessian are given by

$$g = -2\mathbf{J}^T\mathbf{r},\ \mathbf{H} = 2\mathbf{J}^T\mathbf{J} - 2\sum_{i=1}^{n} r_i G_i. \tag{3}$$

In a linear regression model, the Jacobian is simply the data design matrix $\mathbf{X}$ and does not depend on the parameter values $\boldsymbol{\theta}$.

The normal equations do not have an analytic solution for $\widehat{\boldsymbol{\theta}}$ in most cases, so that numerical iterative procedures are needed. The Gauss-Newton method estimates the parameters in nonlinear regression using the first-order Taylor's expansion

$$f(\mathbf{x}, \boldsymbol{\theta}) \approx f(\mathbf{x}, \boldsymbol{\theta}^*) + \nabla f(\mathbf{x}, \boldsymbol{\theta})^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$$

to approximate a nonlinear regression function (Nadaraya, 1964; Ruppert and Wand, 1994) considering a small neighborhood of $\boldsymbol{\theta}^*$. The linear approximation of $f(\mathbf{x}, \boldsymbol{\theta})$ in the neighborhood of $\boldsymbol{\theta}^*$ results in an approximate residual sum of squares

$$S(\boldsymbol{\theta}) \approx \sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}^*) - \sum_{j}^{p} J_{i,j}^T(\theta_j - \theta_j^*)\right)^2,$$

and

$$S(\boldsymbol{\theta}) - S(\boldsymbol{\theta}^*) \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{J}^T\mathbf{J}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \tag{4}$$

The corresponding normal equations at the t-th iteration are given by

$$\mathbf{J}(\boldsymbol{\theta}^{(t)})^T\mathbf{J}(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) = \mathbf{J}(\boldsymbol{\theta}^{(t)})^T(y - f(\boldsymbol{\theta}^{(t)})).$$

The updating increment

$$\boldsymbol{\delta}^{(t)} = \left(\mathbf{J}(\boldsymbol{\theta}^{(t)})^T\mathbf{J}(\boldsymbol{\theta}^{(t)})\right)^{-1}\mathbf{J}(\boldsymbol{\theta}^{(t)})^T(y - f(\boldsymbol{\theta}^{(t)}))$$

in the t-th iteration, and then $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \lambda_t\boldsymbol{\delta}^{(t)}$ where λ_t is the step size. The update direction $\boldsymbol{\delta}^{(t)}$ is derived from the tangent plane approximation to the solution locus.

The estimate of the asymptotic covariance of the least squares estimate is given by

$$\mathrm{cov}(\widehat{\boldsymbol{\theta}}) = \sigma^2(\mathbf{J}^T\mathbf{J})^{-1}.$$

The statistical inference about $\boldsymbol{\theta}$ is based on the property that $\widehat{\boldsymbol{\theta}}$ follows $N(\boldsymbol{\theta}, \sigma^2(\mathbf{J}^T\mathbf{J})^{-1})$ asymptotically. σ^2 is estimated by $s^2 = S(\widehat{\boldsymbol{\theta}})/(n - p)$, where p is the number of parameters and $\mathbf{J}$ is estimated by $\mathbf{J}(\widehat{\boldsymbol{\theta}})$.

2.1. Maximum likelihood estimation

Considering a normal distributed error term

$$y_i - f(\mathbf{x}_i, \boldsymbol{\theta}) \overset{i.i.d.}{\sim} N(0, \sigma^2),$$

the log-likelihood function of $\boldsymbol{\theta}$ and σ^2 is

$$l(\boldsymbol{\theta}, \sigma^2) = -\frac{S(\boldsymbol{\theta})}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2).$$

Maximum likelihood estimation (MLE) is a method of estimating the parameters by maximizing $l(\boldsymbol{\theta}, \sigma^2)$ with respect to $\boldsymbol{\theta}$ and σ^2 for which we need the following derivatives

$$\frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^4}S(\boldsymbol{\theta}) - \frac{n}{2\sigma^2}, \tag{5}$$

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = -\frac{1}{2\sigma^2}\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{6}$$

Setting equations (5), (6) equal to 0, we have $\widehat{\sigma}^2_{\text{MLE}} = \frac{S(\widehat{\boldsymbol{\theta}})}{n}$ and the normal equation (2). As a result, $\widehat{\boldsymbol{\theta}}_{\text{MLE}} = \widehat{\boldsymbol{\theta}}$. In summary, when the noise follows a normal distribution, the Gauss-Newton iterations to minimize the least squares $S(\boldsymbol{\theta})$ are exactly the same as the Newton-Raphson iterations to maximize the log-likelihood function with respect to $\boldsymbol{\theta}$, and hence the maximum likelihood estimator is equivalent to the least squares estimator (Bates and Watts, 1988).

2.2. Asymptotic statistical inference

Let us consider the nonlinear regression model (1) where the $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. We note that as $n \to \infty$, the least squares estimate $\widehat{\boldsymbol{\theta}}$ is asymptotically $N\left(\boldsymbol{\theta}^*, \sigma^2(\mathbf{J}^T\mathbf{J})^{-1}\right)$, where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$, and the matrix $\mathbf{J} = [(\partial f(x_i;\boldsymbol{\theta})/\partial\theta_j]_{\boldsymbol{\theta}^*}$ plays the same role as X in the linear regression (Saber and Wild, 1989). In particular, by analogy with the linear confidence region

$$\{\boldsymbol{\theta} \mid (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\widehat{\mathbf{J}}^T\widehat{\mathbf{J}}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \le ps^2F^{\alpha}_{p,n-p}\}, \tag{7}$$

is an approximate $100(1-\alpha)\%$ confidence region for $\boldsymbol{\theta}$. Here $s^2 = S(\widehat{\boldsymbol{\theta}})/(n-p)$, $\widehat{\mathbf{J}} = \mathbf{J}(\widehat{\boldsymbol{\theta}})$, and $F^{\alpha}_{p,n-p}$ is the upper a critical value of the $F_{p,n-p}$ distribution. As the linear approximation is valid asymptotically, (7) will have the correct confidence level of $1-\alpha$ asymptotically. As α varies, the regions (7) are ellipsoid contours of the approximate multivariate normal density function of $\widehat{\boldsymbol{\theta}}$ (with $\mathbf{J}$ replaced by $\widehat{\mathbf{J}}$). Since $S(\boldsymbol{\theta})$ measures how the observations are close to the fitted equation for any $\boldsymbol{\theta}$, it would seem appropriate to also base confidence regions for $\boldsymbol{\theta}$ on the contours of $S(\boldsymbol{\theta})$. Such a region could take the form

$$\left\{\boldsymbol{\theta} \mid S(\boldsymbol{\theta}) \le cS(\widehat{\boldsymbol{\theta}})\right\}$$

for some c, $c > 1$. Regions of this type are often called the exact confidence regions, as they are not based on any approximations. However, the confidence levels, or coverage probabilities, of such regions are generally unknown, though approximate levels can be obtained from asymptotic theory.

For large enough n, the consistent estimator $\widehat{\boldsymbol{\theta}}$ will be sufficiently close to $\boldsymbol{\theta}^*$ for the approximation (4) to hold with $F^{\alpha}_{p,n-p}$ in (7). Therefore, substituting

$$S(\boldsymbol{\theta}^*) - S(\boldsymbol{\theta}) \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\widehat{\mathbf{J}}^T\widehat{\mathbf{J}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \tag{8}$$

from (4), we obtain the confidence region

$$\left\{\boldsymbol{\theta} \mid S(\boldsymbol{\theta}) \le S(\widehat{\boldsymbol{\theta}})\left(1 + \frac{p}{n-p}F^{\alpha}_{p,n-p}\right)\right\} \tag{9}$$

that has the required asymptotic confidence level of $100(1-\alpha)\%$. The regions (7) and (9) are asymptotically the same, and they are identical for linear models. However, for finite n, these regions may be very different, and it indicates inadequacy of the linear approximation (4). Under the assumption of normal errors we see that the above clash between the two confidence regions embodies a wider principle. We have a choice between confidence regions based on the asymptotic normality of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$, which is also the least squares estimator, and those based on the contours of the likelihood function via the likelihood ratio test.

3. Parameterization and curvature

There are some nonlinear models which are intrinsically linear because they can be made linear in the parameters by a simple transformation. For example, the reciprocal of the Michaelis-Menten regression mean function

$$f(x, \mathbf{b}) = \frac{b_1 x}{b_2 + x} \tag{10}$$

results in the model

$$\begin{aligned}\frac{1}{y} &= \frac{1}{b_1} + \frac{b_2}{b_1}\frac{1}{x} + \epsilon \\ &= \beta_1 + \beta_2\frac{1}{x} + \epsilon,\end{aligned}$$

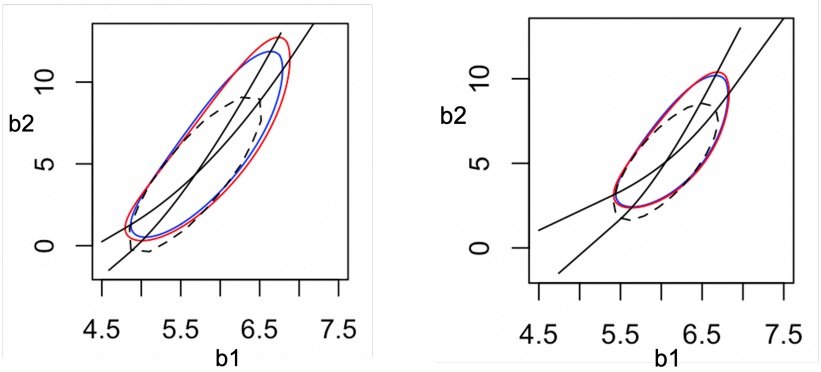
which is linear in the transformed parameters b_1 and b_2. In such cases, transforming a model to its linear form often provides better inference procedures and confidence intervals. However, parameterization may affect the asymptotic variance and convergence of the estimation algorithm (Bates and Watts, 1988; Huang et al., 2010). The re-parameterization

is an important issue. Although the shape of the solution locus is fixed, the performance of the Gauss-Newton estimator varies with respect to different parameterizations. This depends on the nonlinearity in the model. There are two kinds of nonlinearity–intrinsic nonlinearity and parameter-effects nonlinearity (Hamilton and Watts, 1985; Bates and Watts, 1988). The intrinsic nonlinearity is associated with the modeling and is invariant under parameterization. The parameter-effects nonlinearity, however, can be lessened through a proper parameterization. If either component of the nonlinearity is large, the least squares estimate may not converge. Furthermore, the estimated covariance for $\widehat{\mathbf{b}}$ given by $\widehat{V}(\widehat{\mathbf{b}}) = \widehat{\sigma}^2 \left(\mathbf{J}(\widehat{\mathbf{b}})^T \mathbf{J}(\widehat{\mathbf{b}})\right)^{-1}$, would change greatly in each step of the iteration, and the statistical inference based on the asymptotic normality becomes unreliable. In other words, the least squares estimator in nonlinear regression depends on the curvedness of the underlying model as well as the parameterization adopted.

The quality of the linear approximation can be summarized by intrinsic curvature $\kappa_d^N = \|d^T C_N d\|$ and parameter-effects curvature $\kappa_d^T = \|d^T C_T d\|$ (Bates and Watts, 1988), where d has unit length, C_N is the normal direction curvature matrix, and C_T is the tangent direction curvature matrix. For details on curvatures, please read the Appendix in Huang et al. (2010).

The two curvatures correspond to two assumptions of linear approximation of a nonlinear mean function. First, the planar assumption ensures that the nonlinear mean function is approximated by its tangent plane at a given point. Second, the uniform coordinate assumption means that an equi-spaced, straight and parallel linear coordinate system is placed on the approximation tangent plane. Intrinsic curvature is related to the planar assumption, and it depends on the dataset considered and the mean function but not on the parameterization used in the mean function. Parameter-effects curvature is related to the uniform coordinate assumption, and it depends on all aspects of the model, including the parameterization. Large values of these two curvature measures indicate a poor linear approximation. The function rms.curv() in the R package MASS can be used to calculate the two measures for a given nls() model fit (Venables and Ripley, 2002). The intrinsic curvature is generally relatively negligible compared with the parameter-effects curvature (Bates and Watts, 1988), and in such cases the profile likelihood approach is useful.

The Wald, likelihood ratio (LR) and score tests are asymptotically equivalent. They all have the same sampling distribution which is a chi-square distribution with the same degrees of freedom. However, the Wald test statistics $(\widehat{\mathbf{b}} - \mathbf{b}_0)^T \left[\widehat{V}(\widehat{\mathbf{b}})\right]^{-1} (\widehat{\mathbf{b}} - \mathbf{b}_0)$, which results in an ellipsoidal confidence region, uses $\widehat{\mathbf{b}}$ and depends on curvature of the likelihood at $\widehat{\mathbf{b}}$. The score test depends on the slope and curvature at $\widehat{\mathbf{b}}_0$, while the LR test uses information from both $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{b}}_0$. The differences among them vanish in large samples if the null hypothesis is true. If the null hypothesis is false, they may take very different values. The Wald procedures are influenced by both intrinsic curvature and parameter-effects curvature. A model with a lower parameter-effects curvature is preferable.



(a) n=50 **(b)** n=100

Figure 1: The profile plots of the simulation data analysis. The 95% confidence region of b_1 and b_2 using the Wald statistic (dashed), the likelihood ratio (blue) and Skovgaard's approximation (red) with simulated data from the Michaelis-Menten model. The left panel is with sample size $n = 50$, and the right panel is with sample size $n = 100$.

The curvature can affect the quality of statistical inferences based on the asymptotic normality. When the mapped parameter curves onto the tangent plane are not uniform grid lines, the resulting confidence region may not be reliable. We use the Michaelis-Menten model (10) as an example by simulating data with $\epsilon_i \overset{i.i.d.}{\sim} N(0, 1^2)$, $b_1 = 6, b_2 = 5$, and $\mathbf{x}$ generated from the uniform distribution $(1, 100)$. We demonstrate how confidence regions of b_1 and b_2 are affected

by curvatures with two different sample sizes $n_1 = 50$ and $n_2 = 100$ displayed in Figure 1 using the contour method of the nlreg package (Brazzale, 2005). The dashed, blue and red lines represent the confidence regions of (b_1, b_2) obtained with the Wald test, the likelihood ratio test, and Skovgaard's approximation, respectively (Skovgaard, 1996). The confidence regions of the latter two methods are closer to the exact one than the Wald (Bates and Watts, 1988; Skovgaard, 1996). When the sample size is small, the shapes and coverage between the Wald confidence region and the two others are very different. As the sample size grows, they become less different.

4. Newton-Raphson Method

The Newton-Raphson method (also called Newton's method) which has been widely used for estimating the parameters of nonlinear regression models due to its advantage of fast convergence using the second order Taylor's expansion to approximate the nonlinear function $l(\boldsymbol{\theta})$ as a quadratic polynomial as follows:

$$l(\boldsymbol{\theta}) \approx l(\boldsymbol{\theta}^{(t)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^T \nabla l(\boldsymbol{\theta}^{(t)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)})^T \nabla^2 l(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}), \tag{11}$$

where $\boldsymbol{\theta}^{(t)}$ is in a neighborhood of the true $\boldsymbol{\theta}$, ∇ is the first derivative and ∇^2 is the second derivative with respect to $\boldsymbol{\theta}$. The Newton-Raphson method can be computationally challenging and heavily dependent on good starting values (Gill et al., 1986). To maximize the quadratic approximation, we set the first derivative of the approximate quadratic function (11) equal to 0 at a new point $\boldsymbol{\theta}^{(t+1)}$

$$\nabla l(\boldsymbol{\theta}^{(t)}) + [\nabla^2 l(\boldsymbol{\theta}^{(t)})](\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}) = \mathbf{0}_p,$$

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - [\nabla^2 l(\boldsymbol{\theta}^{(t)})]^{-1} \nabla l(\boldsymbol{\theta}^{(t)}) \\ &= \boldsymbol{\theta}^{(t)} + [-\nabla^2 f(\boldsymbol{\theta}^{(t)})]^{-1} \nabla f(\boldsymbol{\theta}^{(t)}) \end{aligned}$$

which approximates $\boldsymbol{\theta}$ in each iteration t. The value $\boldsymbol{\theta}^{(t+1)}$ should be a better guess than the initial point $\boldsymbol{\theta}^{(t)}$ is.

Here, we focus on deriving the Newton-Raphson method for the exponential family (Nelder and Wedderburn, 1972) including the Gaussian (normal) distributions for continuous variables, binomial distributions for binary response variables, Poisson distributions for count variables, and the gamma and inverse-Gaussian families of distributions for modeling positive continuous data, where the conditional variance of Y_i increases with its expectation. The log-likelihood for n independent observations from an exponential family is given by

$$L_n(\boldsymbol{\theta}, \phi; Y) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}. \tag{12}$$

Assume that the mean μ_i depends linearly on the covariates through a link function g as follows:

$$g(\mu_i) = \eta_i = X_i^T \boldsymbol{\beta}.$$

Because the link function is invertible, we can also write

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(X_i^T \boldsymbol{\beta}).$$

Note that $\mu_i = b'(\theta_i), \theta_i = (b')^{-1} \circ g^{-1}(X_i^T \beta) := h(X_i^T \boldsymbol{\beta})$ where

$$b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}, \; g'(\mu_i) = \frac{\partial g(\mu_i)}{\partial \mu_i}, \text{ and } h'(\boldsymbol{\beta}) = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

According to the chain rule, we have

$$\begin{aligned} \frac{\partial L_n}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} \\ &= \sum_{i}^{n} \frac{Y_i - \mu_i}{\phi} h'(X_i^T \boldsymbol{\beta}) X_{i,j} \\ &:= \sum_{i}^{n} (Y_i - \mu_i) W_i X_{i,j}, \end{aligned}$$

where $W_i = h^{'}(X_i^T\boldsymbol{\beta})/g^{'}(\mu_i)\phi$.
Define $\mathbf{W} = \text{diag}(W_1, \ldots, W_n)$, then the gradient is

$$\nabla L_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)\mu_i'(\eta_i)}{\sigma_i^2} X_i = \mathbf{X}^T\mathbf{W}(\mathbf{Y} - \boldsymbol{\mu}).$$

For the Hessian, we have

$$\frac{\partial^2 L_n}{\partial\beta_j\beta_k} = \sum_i \frac{Y_i - \mu_i}{\phi} h^{''}(X_i^T\boldsymbol{\beta})X_{i,j}X_{i,k} - \frac{1}{\phi}\sum_i \left(\frac{\partial\mu_i}{\partial\beta_k}\right) h^{'}(X_i^T\boldsymbol{\beta})X_{i,j}.$$

Note that

$$\frac{\partial\mu_i}{\partial\beta_k} = \frac{\partial b^{'}(\theta_i)}{\partial\beta_k} = \frac{\partial b^{'}(h(X_i^T\boldsymbol{\beta}))}{\partial\beta_k} = b^{''}(\theta_i)h^{'}(X_i^T\boldsymbol{\beta})X_{i,k}.$$

So,

$$-E[\mathbf{H}_{L_n(\boldsymbol{\beta})}] = \mathbf{X}^T\mathbf{W}\mathbf{X},$$

where $\mathbf{W} = \text{diag}\left(\frac{h^{'}(X^T\boldsymbol{\beta})}{g^{'}(\mu_i)}\right)$. It leads to the Hessian matrix $\mathbf{H}(\boldsymbol{\theta}) = -\mathbf{X}^T\mathbf{W}\mathbf{X}$. Fisher scoring is equivalent to the Newton-Raphson method when we use the canonical link in logistic regression. Since $y_i \in (0, 1)$, $-\mathbf{X}^T\mathbf{W}\mathbf{X}$ is strictly negative definite. When y_i is too close to 0 or 1, weights are close to 0 and $\mathbf{H}$ may have singular values close to 0 and therefore may be computationally singular. The Newton-Raphson algorithm (Green, 1984; Cleveland and Devlin, 1988; Simonoff and Tsai, 1991) iterates according to

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(y - \mu) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(y - \mu + \mathbf{X}\beta^{(t)}),\end{aligned} \tag{13}$$

which is the iteratively reweighted least squares (IRLS) algorithm (Ruppert and Wand, 1994).

Remarks: IRLS solves a generalized linear model in each iteration

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}z^{(t)},$$

where $z^{(t)} = y - \mu + \mathbf{X}\beta^{(t)}$ is the residual. When $\mathbf{W}$ is symmetric positive definite this is equivalent to the solution of the weighted least squares problem:

$$\min_{\beta} \sum_{i=1}^{n} \left(\mathbf{W}^{1/2}(X_i^T\boldsymbol{\beta} - y_i)\right)^2.$$

Note that in the logistic regression case $g'(\mu) = g(\mu)(1 - g(\mu))$ is never zero, so that the diagonal entries of $\mathbf{W}$ are never zero. However, other link functions might lead to degenerate cases with indefinite weighting matrices. When the link function $g(\mu)$ is the identity function and the error follows Gaussian with mean zero and variance one ($N(0, 1)$), then the IRLS algorithm becomes

$$\begin{aligned}\beta^{(1)} &= 0, \\ \mu &= \mathbf{X}\beta^{(1)} = 0, \\ z &= \eta + \frac{y - g(\mu)}{g'(\mu)} = 0 + \frac{y - 0}{1} = y, \\ \mathbf{W} &= \text{diag}\left(\frac{g'(\mu)^2}{(g'(\mu)}\right) = \text{diag}(1^2/1) = I_n, \\ \beta^{(2)} &= 0 + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y,\end{aligned}$$

which is the ordinary least squares estimator.

4.1. Variance heterogeneity

When the residual errors show a trend (e.g., increasing variability as the explanatory variable increases), this can be addressed by modeling the variance as a function of the independent variable or the fitted values. Different methods such as the likelihood ratio test, score test, nonparametric test, and certain modelling approaches can be used for the diagnosis of heteroscedasticity in normal nonlinear regression models (Tsai, 1986; Simonoff and Tsai, 1991; Diblasi and Bowman, 1997; Lin and Wei, 2003). For non-normal models, the test of departure from the nominal dispersion, including over-dispersion and under-dispersion were also discussed (Dean, 1992; Wei et al., 1998).

If variance heterogeneity is ignored, the parameter estimates may not be influenced much, but it may result in severely misleading confidence and prediction intervals (Carroll and Ruppert, 2017). One way of taking into account variance heterogeneity is by explicitly modelling it by formulating a regression model for the variance. For example, Chung and Park (2007) investigated regression models in which the conditional variance of the error is an unscaled function of an integrated time series.

The generalized least squares (GLS) method (Nelder and Wedderburn, 1972) is typically used for correlated residuals with heterogeneous variances using the IRLS (13). Formally, we replace the assumption of the noise distribution $\epsilon \sim N(0, \sigma^2 I_n)$ with the assumption of the noise distribution as $\epsilon \sim N(0, \Sigma)$ where $\Sigma = V^{1/2} R V^{1/2}$. Here V is a diagonal matrix of potentially different (heterogeneous) variance terms and R is a correlation matrix. V has a constant diagonal and R can be modelled by a variance function (e.g., $a_i(\phi)$ in an exponential family in (12)) or an autoregressive–moving-average temporal structure. Additionally, Gaussian processes (Rasmussen, 2003) with the Bayesian nonparametric inference have been widely used for heteroscedastic nonlinear regression in which we have $f \sim N(\mu, K)$. The covariance matrix K has elements $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, and function κ measures similarity between two observations $\mathbf{x}_i$ and $\mathbf{x}_j$. For example, the radial basis function kernel has $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{-\|x_i - x_j\|^2/\gamma\right\}$ with a hyperparameter γ.

5. Model building

Model building aims at finding more realistic ways to describe the stochastic behavior observed in data. We listed a few widely used models, which are classified by their shape and applied to real data analysis in various fields (Farebrother, 1986).

- Polynomial models. Linear and quadratic polynomial regression have been widely applied to solve real-world data problems for scientists and engineers. Quadratic or higher-order polynomials are curved with respect to the predictors, but they are linear in the parameters so that they can be fitted by using linear regression. However, they tend to overfit the responses with a large number of predictors. As the availability of nonlinear regression algorithms increased, the use of polynomials has sensibly decreased. Linear or quadratic polynomials are mainly used to approximate the observed response within a small range of each predictor for understanding the relationship with the response.

- Concave/Convex curves models. Concave/convex curves describe nonlinear relationships with asymptotes and without inflection points.

    1. Exponential model. The exponential equation describes an increasing or decreasing trend with constant relative rate. The most common parameterization is

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

    which has the slope of the tangent line through x is θ_2. Therefore, the response increases by an amount that is proportional to x positively when $\theta_1 > 0$ (exponential growth), and negatively when $\theta_2 < 0$ (exponential decay). The exponential curve is used to describe the growth of a population in unconstrained environmental conditions. The exponential function is nonlinear and can be fitted by using nls() or drm() functions in R (Team et al., 2018).

    2. Asymptotic model. The asymptotic regression model describes a limited growth, where y approaches a horizontal asymptote as $x \to \infty$. There are several different parameterizations known as the monomolecular growth model (Fekedulegn et al., 1999). The most widely used parameterization is

    $$y = \theta_1 - (\theta_1 - \theta_2) \exp(-\theta_3 x) + \epsilon,$$

    where θ_1 is the maximum attainable y, θ_2 is y at $x = 0$, and θ_3 is proportional to the relative rate of y increase while x increases.

3. Negative exponential model. This model can be given by

$$y = \theta_1 \left(1 - \exp(-\theta_2 x)\right),$$

which has a similar shape to the asymptotic regression, but $y = 0$ when $x = 0$, and is often used to model absorbed photosynthetically active radiation (Kiniry et al., 1989).

4. Power curve model. The power curve is also known as the Freundlich equation or allometric equation (Thommes et al., 2015) and the most common parameterization is

$$y = \theta_1 x^{\theta_2} + \epsilon.$$

This curve is equivalent to an exponential curve on the logarithm of x since $\theta_1 x^{\theta_2} = \theta_1 \exp\left(\log(x^{\theta_2})\right) = \theta_1 \exp\left(\theta_2 \log(x)\right)$, and it does not have an asymptote for $x \to \infty$.

5. Logarithmic model. After x is log-transformed,

$$y = \theta_1 + \theta_2 \log(x)$$

is a linear model with $x > 0$. The parameter θ_2 dictates the shape, as in the exponential equation, Indeed, if $\theta_2 > 0$, the curve is convex up and y increases as x increases. If $\theta_2 < 0$, the curve is concave up and y decreases as x increases.

6. Rectangular hyperbola model. The Michaelis-Menten equation is a rectangular hyperbola, often parameterized as

$$y = \frac{\theta_1 x}{\theta_2 + x} + \epsilon,$$

which is a convex-up curve and y increases as x increases up to a plateau level. The parameter θ_1 represents the asymptote as $x \to \infty$, while θ_2 is the x value giving a response equal to $\theta_1/2$. This is because $\theta_2 = x_{50}$ when $\theta_1/2 = \frac{\theta_1 x_{50}}{\theta_2 + x_{50}}$ where x_{50} is the median.

- Sygmoidal curve model. Sygmoidal curves are S-shaped and may be increasing, decreasing, or symmetric around the inflection point. We show a common parameterization.

1. Logistic model. The logistic curve is derived from the cumulative logistic distribution function; the curve is symmetric around the inflection point and it may be parameterized as

$$y = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp(\theta_3(x - \theta_4))} + \epsilon,$$

where θ_2 is the upper asymptote, θ_1 is the lower asymptote, θ_4 is the x value producing a response halfway between θ_2 and θ_1. θ_3 is the slope around the inflection point, and it can be positive or negative and, consequently. Hence, y may increase or decrease as x increases.

2. Gompertz model. The Gompertz curve has the following parameterization

$$y = \theta_1 + (\theta_2 - \theta_1) \exp\{-\exp[\theta_3(x - \theta_4)]\} + \epsilon,$$

where the parameters have the same meaning as those in the logistic model, but the difference is that this curve is not symmetric around the inflection point.

- Log-logistic models. The sigmoidal response curve is symmetric on the logarithm of x, which requires a log-logistic curve. For example, in biologic assays, the log-logistic curve is defined as follows:

$$y = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp\{\theta_3[\log(x) - \log(\theta_4)]\}} + \epsilon,$$

where the parameters have the same meaning as the logistic model.

1. Weibull-type 1 model. The Weibull distribution is a widely used lifetime distribution in reliability and life data analysis. The type-1 Weibull curve is an alternative for the Gompertz curve like the log-logistic curve is for the logistic curve. The model function is as follows:

$$f(x, \boldsymbol{\theta}) = \theta_1 + (\theta_4 - \theta_1)\{1 - \exp\{-\exp[\theta_2(\log(x) - \log(\theta_3))]\}\}.$$

The parameters have the same meaning as in the sygmoidal curves.

2. Weibull-type 2 model. The type-2 Weibull curve is for the Gompertz curve what the log-logistic curve is for the logistic curve. The equation is as follows:

$$f(x, \boldsymbol{\theta}) = \theta_1 + (\theta_4 - \theta_1) \exp\{-\exp[\theta_2(\log(x) - \log(\theta_3))]\},$$

again with the same parameter interpretation as in the sygmoidal curves.

6. Simulation data analysis

We compare the nonlinear least squares (NLS) estimation with different initial values of the parameters in the nonlinear regression using a simulated dataset where the error term ϵ is not normally distributed. We show that the Bayesian approach is more appropriate to find reliable estimates as well as the confidence regions. The simulation setting is similar to the one shown in Figure 1. We use the Michaelis-Menten model (10) to simulate 100 data points with $\epsilon \overset{i.i.d.}{\sim} Gamma(10, 0.25)$, $\theta_1 = 100$, and $\theta_2 = 0.05$. The covariates $\mathbf{X}$ are generated from the absolute values of $N(0, 20^2)$.

We implement the least squares method through the nls() function in R (Baty et al., 2015). Two sets of starting values for the parameters θ_1 and θ_2 are tried, $(10, 3)$ and $(50, 0.1)$, to validate if least squares estimates can converge to the true values given bad starting values.

Both the estimates and 95% confidence intervals for θ_1 and θ_2 using the least squares and Bayesian approaches are shown in Table 1. With good initial values, both the least squares and Bayesian method give good estimates and confidence regions for the parameters. However, when the starting values are bad, the estimates are far from the true values for both methods. The 95% confidence regions still contain the true values for Bayesian modeling, but the bands are too wide since the posterior samples tend to converge to local optimums.

Model	Initals	$\theta_1 = 100$	$\theta_2 = 0.05$
NLS	(10, 3)	81.78 (73.94, 89.61)	-0.22 (-0.23, -0.21)
	(50,0.1)	102.53 (102.38, 102.69)	0.05 (0.05, 0.05)
Bayesian modeling	(10, 3)	84.74 (25.48, 102.67)	-0.25 (-1.18, 0.05)
	(50,0.1)	102.53 (102.37, 102.69)	0.05 (0.05, 0.05)

Table 1: The estimates of θ_1 and θ_2 with the 95% confidence intervals using least squares and Bayesian approach with two different settings of initial values.

7. Further reading

- Levenberg-Marquardt algorithm. The use of the Gauss-Newton algorithm with Levenberg-Marquardt modifications (Levenberg, 1944; Marquardt, 1963) is to speed up the convergence as well as to stabilize the computation for near-singular $\mathbf{J}^T\mathbf{J}$ in the least squares normal equations. The Levenberg-Marquardt method compromises the Gauss-Newton method and the steepest descent method.

- Multicollinearity. This occurs when some columns in the Jacobian matrix J are highly correlated and leads to an ill-conditioned matrix of normal equations. This implies that the model may be over-parameterized, so that a simpler model or a transformation of the predictors or parameters may be considered.

- Nonparametric and semiparametric regression. If the target of data analysis is to fit the response curve on the explanatory variables, a nonparametric regression could be a better alternative than a nonlinear regression model. However, the fitted curve may be less interpretable. Semiparametric regression which consists of both parametric and nonparametric components is an alternative to nonlinear regression modelling.

- Kernel estimator. A one-dimensional smoothing kernel is any smooth, symmetric function $K : \mathbb{R} \to \mathbb{R}$ such that $K(x) \geq 0$ and $\int K(x)dx = 1$, $\int xK(x)dx = 0$, $0 < \int x^2K(x)dx < \infty$. Given a bandwidth $h > 0$, the Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964) is defined as

$$\widehat{f_h}(x) = \frac{\sum_{i=1}^n K\left(\frac{\|x-x_i\|}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{\|x-x_i\|}{h}\right)} = \sum_{i=1}^n w_i(x) y_i,$$

where $w_i(x) = \frac{K(\|x-X_i\|/h)}{\sum_{j=1}^n K(\|x-X_j\|/h)}$. Therefore, $\widehat{f_h}(x)$ is a local weighted average of the y_i's.

- Initial values for parameter estimation. For a linear model the Gauss-Newton method finds the minimum in one iteration from any initial parameter estimates. If the nonlinear mean function is nearly linear, the convergence of the Gauss-Newton method is fast and does not depend heavily on the initial parameter estimates. However, as the magnitude of model nonlinearity is large, convergence can be slow or even may not occur, and the resulting parameter estimates may not be reliable. In that case, good initial values are important.
- Differential geometric view. The measures of curvature described above are related to the parameterization. When a normal error assumption is assumed, a Euclidean metric is imposed. Changing the error distribution results in a different metric and geometric structures of the distributions. The study of probability and information using differential geometry is called information geometry. We refer the reader to Amari (2016) for theory and applications of information geometry.

Acknowledgement

Hsin-Hsiung Huang is supported by the National Science Foundation grants (DMS-1924792).

References

S.-i. Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

D. Bates and D. Watts. *Nonlinear regression analysis and its applications*. Wiley series in probability and mathematical statistics. Wiley, New York [u.a.], 1988.

F. Baty, C. Ritz, S. Charles, M. Brutsche, J.-P. Flandrois, and M.-L. Delignette-Muller. A toolbox for nonlinear regression in R: The package nlstools. *Journal of Statistical Software*, 66(5):1–21, 2015.

R. J. Beverton and S. J. Holt. *On the dynamics of exploited fish populations*, volume 11. Springer Science & Business Media, 2012.

A. R. Brazzale. hoa: An R package bundle for higher order likelihood inference. *Rnews*, 5/1 May 2005:20–27, 2005. ISSN 609-3631.

R. J. Carroll and D. Ruppert. *Transfor mation and Weighting in Regression*. Chapman and Hall/CRC, 2017.

H. Chung and J. Y. Park. Nonstationary nonlinear heteroskedasticity in regression. *Journal of Econometrics*, 137(1): 230–259, 2007.

W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

C. B. Dean. Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, 87(418):451–457, 1992.

A. Diblasi and A. Bowman. Testing for constant variance in a linear model. *Statistics & Probability Letters*, 33(1): 95–103, 1997.

R. W. Farebrother. Nonlinear regression modelling: A unified practical approach: David a. ratkowsky, (dekker, new york, 1983). *International Journal of Forecasting*, 2(1):125–125, 1986.

D. Fekedulegn, M. M. Siurtain, and J. Colbert. Parameter estimation of nonlinear growth models in forestry. *Silva Fennica*, 33, 1999.

P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright. User's guide for npsol (version 4.0): A fortran package for nonlinear programming. Technical report, Stanford Univ CA Systems Optimization Lab, 1986.

P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, 46:149–192, 1984.

D. C. Hamilton and D. G. Watts. A quadratic design criterion for precise estimation in nonlinear regression models. *Technometrics*, 27(3):241–250, 1985. doi: 10.1080/00401706.1985.10488048.

H. Huang, C. Hsiao, and S. Huang. *Nonlinear Regression Analysis*. International Encyclopedia of Education, Elsevier Science, 2010. ISBN 9780080448930.

J. Kiniry, C. Jones, J. O'toole, R. Blanchet, M. Cabelguenne, and D. Spanel. Radiation-use efficiency in biomass accumulation prior to grain-filling for five grain-crop species. *Field Crops Research*, 20(1):51–64, 1989.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.

J.-G. Lin and B.-C. Wei. Testing for heteroscedasticity in nonlinear regression models. *Communications in Statistics-Theory and Methods*, 32(1):171–192, 2003.

D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441, 1963. doi: 10.1137/0111030.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384, 1972.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

D. Ruppert and M. P. Wand. Multivariate locally weighted least squares regression. *The Annals of Statistics*, 22: 1346–1370, 1994.

G. A. F. Saber and C. J. Wild. Nonlinear regression. Technical report, 1989.

L. B. Sheiner and S. L. Beal. Evaluation of methods for estimating population pharmacokinetic parameters. i. michaelis-menten model: routine clinical pharmacokinetic data. *Journal of pharmacokinetics and biopharmaceutics*, 8(6):553–571, 1980.

J. S. Simonoff and C.-L. Tsai. *Improved tests for nonconstant variance in regression based on the modified profile likelihood*. Leonard N. Stern School of Business, New York University, 1991.

I. M. Skovgaard. An explicit large-deviation approximation to one-parameter tests. *Bernoulli*, 2(2):145–165, 06 1996. doi: 10.3150/bj/1193839221.

R. C. Team et al. R: A language and environment for statistical computing; 2018, 2018.

M. Thommes, K. Kaneko, A. V. Neimark, J. P. Olivier, F. Rodriguez-Reinoso, J. Rouquerol, and K. S. Sing. Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (iupac technical report). *Pure and Applied Chemistry*, 87(9-10):1051–1069, 2015.

C.-L. Tsai. Score test for the first-order autoregressive model with heteroscedasticity. *Biometrika*, 73(2):455–460, 1986.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

G. S. Watson. Smooth regression analysis. *Sankhyā Ser.*, 26:359–372, 1964.

B.-C. Wei, J.-Q. Shi, W.-K. Fung, and Y.-Q. Hu. Testing for varying dispersion in exponential family nonlinear models. *Annals of the Institute of Statistical Mathematics*, 50(2):277–294, 1998.