PAPER • OPEN ACCESS

Computation of nonlinear least squares estimator and maximum likelihood using principles in matrix calculus

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042125 doi:10.1088/1757-899X/263/4/042125

Computation of nonlinear least squares estimator and maximum likelihood using principles in matrix calculus

B Mahaboob[1], B Venkateswarlu[2], J Ravi Sankar[1] and P Balasiddamuni[3]
[1]Department of Mathematics, Swetha Institute of Technology, and Science, Tirupati, Andhra Pradesh, India
[2]Department of Mathematics, School of Advanced Sciences, VIT University, Vellore-632014, India
[3]Department of Statistics, S.V.University,Tirupati, Andhra Pradesh, India

E-mail: venkatesh.reddy@vit.ac.in

Abstract. This paper uses matrix calculus techniques to obtain Nonlinear Least Squares Estimator (NLSE), Maximum Likelihood Estimator (MLE) and Linear Pseudo model for nonlinear regression model. David Pollard and Peter Radchenko [1] explained analytic techniques to compute the NLSE. However the present research paper introduces an innovative method to compute the NLSE using principles in multivariate calculus. This study is concerned with very new optimization techniques used to compute MLE and NLSE. Anh [2] derived NLSE and MLE of a heteroscedatistic regression model. Lemcoff [3] discussed a procedure to get linear pseudo model for nonlinear regression model. In this research article a new technique is developed to get the linear pseudo model for nonlinear regression model using multivariate calculus. The linear pseudo model of Edmond Malinvaud [4] has been explained in a very different way in this paper. David Pollard et.al used empirical process techniques to study the asymptotic of the LSE (Least-squares estimation) for the fitting of nonlinear regression function in 2006. In Jae Myung [13] provided a go conceptual for Maximum likelihood estimation in his work “Tutorial on maximum likelihood estimation

1. Introduction

A large number of problems in nonlinear model building are concerned with the inferential aspects including estimating the parameters and testing the hypothesis about the parameters of the nonlinear regression models. Now–a-days efficient estimation of the nonlinear models has received little attention. Several researches in applied mathematics are very often interested in inferential aspects of the nonlinear regression models. These inferential aspects have been studied intensively for the last three decades. The nonlinear inferential methods and the error assumptions are generally analogous to those made for the linear regression models. The literature on nonlinear methods of estimation has been grown enormously for the past four decades. This research paper gives new procedures to compute nonlinear least squares estimator, maximum likely hood estimator and a linear pseudo model for general nonlinear regression model using principles in matrix calculus.

2. Productive efficiency of DMU

For a sample of n observations on the dependent variable, consider the general nonlinear regression model in vector notation as

$$Y_{n\times 1} = f_{n\times 1}\left(X_{n\times k}, \beta_{p\times 1}\right) + \varepsilon_{n\times 1} \tag{2.1}$$

Where, $Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}$, $f\left(X,\beta\right) = \begin{bmatrix} f\left(X_1,\beta\right) \\ f\left(X_2,\beta\right) \\ . \\ . \\ f\left(X_n,\beta\right) \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$

Here, Y, $f\left(X,\beta\right)$ and ε are $\left(n\times 1\right)$ vectors;
β is $\left(p\times 1\right)$ vector of unknown parameters.
X_i is $\left(k\times 1\right)$ vector
Assume that ε_i 's are independently identically distributed with

$$E\left(\varepsilon_i\right) = 0,\ \forall i = 1,2,...,n$$

and $E\left(\varepsilon_i\varepsilon_j\right) = \sigma^2,\ \forall i = j = 1,2,...,n$

$$= 0,\ \forall i \neq j$$

(or) $\varepsilon \overset{\text{i.i.d}}{\sim} \left(O, \sigma^2 I_n\right)$
The exact form of distribution of ε is unknown.
Under nonlinear least squares method of estimation, suppose that one may choose the vector β^* for β that minimizes the residual sum of squares.
Write the residual sum of squares function for the minimization as

$$\phi\left(\beta^*\right) = \left[Y - f\left(X,\beta^*\right)\right]'\left[Y - f\left(X,\beta^*\right)\right] \tag{2.2}$$

or $\phi\left(\beta^*\right) = \sum_{i=1}^{n}\left[Y_i - f\left(X_i,\beta^*\right)\right]^2$ (2.3)

By minimizing $\phi\left(\beta^*\right)$ with respect to β^*, the first order condition gives the following p-nonlinear normal equations.

$$\frac{\partial\phi\left(\beta^*\right)}{\partial\beta^*} = 0 \tag{2.4}$$

$$\left.\begin{array}{l} \Rightarrow \dfrac{\partial\phi\left(\beta^*\right)}{\partial\beta_1^{\;*}} = 0 \Rightarrow -2\sum_{i=1}^{n}\left[Y_i - f\left(X_i,\beta^*\right)\right]\left[\dfrac{\partial f\left(X_i,\beta^*\right)}{\partial\beta_1^{\;*}}\right] = 0 \\ \quad\vdots \qquad\qquad\qquad \vdots \\ \Rightarrow \dfrac{\partial\phi\left(\beta^*\right)}{\partial\beta_p^{\;*}} = 0 \Rightarrow -2\sum_{i=1}^{n}\left[Y_i - f\left(X_i,\beta^*\right)\right]\left[\dfrac{\partial f\left(X_i,\beta^*\right)}{\partial\beta_p^{\;*}}\right] = 0 \end{array}\right\} \quad (2.5)$$

where $\dfrac{\partial\phi\left(\beta^*\right)}{\partial\beta^{\,*}}$ is the partial derivative of $\phi\left(\beta^*\right)$ with respect to β^* evaluated at β.

Given $f\left(X_i,\beta\right)$ is nonlinear function in β_j 's, j=1,2,..,p, the nonlinear normal equations are nonlinear both in variables X_i 's and parameters β_j^* 's.

Here, $\hat{\beta}$ will not be a linear function of Y and in general, it is difficult or impossible to derive its small sample properties. One may show that the nonlinear least squares estimator β^* will not be the best linear unbiased estimator (BLUE) for β. Under certain regularity conditions to be stated, it can be shown that the nonlinear least squares estimator β^* will be consistent and asymptotically normally distributed.

A reasonable estimator for unknown error variance σ^2 is given by

$$\overset{*}{\sigma^2} = \frac{\phi\left(\beta^*\right)}{n-p} \qquad (2.6)$$

3. Maximum likelihood estimator

By introducing the assumption that ε_i 's are independently, identically normally distributed such that ε follows multivariate normal distribution with zero mean vector and variance-covariance matrix Ω, one may apply the maximum likelihood method of estimation to estimate β and error covariance matrix Ω in the nonlinear regression model :

$$Y_{n\times1} = f_{n\times1}\left(X_{n\times k},\beta_{k\times1}\right) + \varepsilon_{n\times1} \qquad (3.1)$$

and $\varepsilon \sim N\left(O,\Sigma\right)$

For a random sample of n observations, the likelihood function of the observation vector Y may be written as

$$L\left(\beta,\Sigma\right) = \frac{1}{\left(2\pi\right)^{\frac{n}{2}}\left|\Sigma\right|^{\frac{n}{2}}}\exp\left\{-\frac{1}{2}\left[Y - f\left(X,\beta\right)\right]'\Sigma^{-1}\left[Y - f\left(X,\beta\right)\right]\right\}$$

$$\text{or}\ \mathrm{Ln}\,L\left(\beta,\Sigma\right) = \frac{-n}{2}\mathrm{Ln}2\pi - \frac{n}{2}\mathrm{Ln}\left|\Sigma\right| - \frac{1}{2}\left[Y - f\left(X,\beta\right)\right]'\Sigma^{-1}\left[Y - f\left(X,\beta\right)\right] \qquad (3.2)$$

Under the maximum likelihood method of estimation, the maximum likelihood estimators $\hat{\beta}$

of β and $\hat{\Sigma}$ of Σ can be obtained by maximizing the $\operatorname{Ln} L(\beta,\Sigma)$ or minimizing $\operatorname{Ln} L(\beta,\Sigma)$ with respect to $\hat{\beta}$ and $\hat{\Sigma}$.

The first order conditions for a minimum of $R^* = -\operatorname{Ln} L(\beta,\Sigma)$ with respect to $\hat{\beta}$ and $\hat{\Sigma}^{-1}$ yield the following maximum likelihood equations:

$$\left.\begin{aligned} &\frac{\partial R^*}{\partial \hat{\beta}_1} = 0 \\ &\frac{\partial R^*}{\partial \hat{\beta}_2} = 0 \\ &\quad . \\ &\quad . \\ &\quad . \\ &\frac{\partial R^*}{\partial \hat{\beta}_p} = 0 \end{aligned}\right\} \tag{3.3}$$

and $\dfrac{\partial R^*}{\partial \hat{\Sigma}^{-1}} = 0$ (3.4)

The solutions of first p-maximum likelihood equations in (3.3) give the vector of maximum likelihood estimators $\hat{\beta}$ for β. It can be easily shown that $\hat{\beta}$ equals the nonlinear least squares estimator β^* for β. The equation (3.4) yields

$$\frac{n}{2}\left(\hat{\Sigma}\right) - \frac{1}{2}\left[Y - f\left(X,\hat{\beta}\right)\right]\left[Y - f\left(X,\hat{\beta}\right)\right]' = 0$$

$$\text{or } \frac{n}{2}\left(\hat{\Sigma}\right) - \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - f\left(X_i,\hat{\beta}\right)\right]\left[Y_i - f\left(X_i,\hat{\beta}\right)\right]' = 0$$

or the maximum likelihood estimator of error covariance matrix Σ is given by

$$\hat{\Sigma} = \frac{\left[Y - f\left(X,\hat{\beta}\right)\right]\left[Y - f\left(X,\hat{\beta}\right)\right]'}{n}$$

$$\text{or } \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - f\left(X_i,\hat{\beta}\right)\right]\left[Y_i - f\left(X_i,\hat{\beta}\right)\right]'$$

It is known that, under general regularity conditions, the maximum likelihood estimators are asymptotically optimal in the sense that they are consistent and have the asymptotic distribution with smallest variance. Further, these properties do not depend on the linearity of the model. Thus, if the model errors ε_i's are independently, identically normally distributed such that ε follows $N\left(O,\hat{\Sigma}\right)$ then the maximum likelihood estimators $\hat{\beta}$ and $\hat{\Sigma}$ are consistent, sufficient, asymptotically efficient and follow asymptotic normal distribution. From the asymptotic theory, it can be shown that maximum likelihood estimator $\hat{\beta}$ has multivariate normal distribution with mean vector $\mathrm{E}\left(\hat{\beta}\right)=\beta$ and variance covariance matrix $\mathrm{Var}\left(\hat{\beta}\right)=\left[\mathrm{Z}\left(\hat{\beta}\right)'\hat{\Sigma}^{-1}\mathrm{Z}\left(\hat{\beta}\right)\right]^{-1}$

$$\text{where } \mathrm{Z}\left(\hat{\beta}\right)=\begin{bmatrix} \frac{\partial \mathrm{f}\left(\mathrm{X}_1,\hat{\beta}\right)}{\partial\hat{\beta}_1} & . & . & \frac{\partial \mathrm{f}\left(\mathrm{X}_1,\hat{\beta}\right)}{\partial\hat{\beta}_p} \\ . & & & . \\ . & & & . \\ \frac{\partial \mathrm{f}\left(\mathrm{X}_n,\hat{\beta}\right)}{\partial\hat{\beta}_1} & . & . & \frac{\partial \mathrm{f}\left(\mathrm{X}_n,\hat{\beta}\right)}{\partial\hat{\beta}_p} \end{bmatrix}$$

Remark:

1. If the model errors ε_i 's are independently, identically normally distributed such that ε follows $\mathrm{N}\left(\mathrm{O},\sigma^2\mathrm{I}_n\right)$ then, for a sample of n observations, the likelihood function of Y is given by
$$\mathrm{L}\left(\beta,\sigma^2\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left\{-\frac{\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\beta\right)\right]'\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\beta\right)\right]}{2\sigma^2}\right\} \text{or}$$
$$\mathrm{Ln}\,\mathrm{L}\left(\beta,\sigma^2\right)=\frac{-n}{2}\mathrm{Ln}2\pi-\frac{n}{2}\mathrm{Ln}\sigma^2-\frac{1}{2\sigma^2}\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\beta\right)\right]'\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\beta\right)\right]$$
The maximum likelihood estimators for β and σ^2 are respectively given by $\hat{\beta}$ and
$$\hat{\sigma}^2=\frac{\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\hat{\beta}\right)\right]'\left[\mathrm{Y}-\mathrm{f}\left(\mathrm{X},\hat{\beta}\right)\right]}{n}$$
2. From the asymptotic theory, the maximum likelihood estimator $\hat{\beta}$ has multivariate normal distribution with mean vector β and variance covariance matrix
$$\mathrm{Var}\left(\hat{\beta}\right)=\hat{\sigma}^2\left[\mathrm{Z}\left(\hat{\beta}\right)'\mathrm{Z}\left(\hat{\beta}\right)\right]^{-1}$$

$$\text{where } Z\left(\hat{\beta}\right) = \begin{bmatrix} \frac{\partial f\left(X_1,\hat{\beta}\right)}{\partial \hat{\beta}_1} & . & . & \frac{\partial f\left(X_1,\hat{\beta}\right)}{\partial \hat{\beta}_p} \\ . & & & . \\ . & & & . \\ \frac{\partial f\left(X_n,\hat{\beta}\right)}{\partial \hat{\beta}_1} & . & . & \frac{\partial f\left(X_n,\hat{\beta}\right)}{\partial \hat{\beta}_p} \end{bmatrix}$$

3. By defining, $\theta = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}_{(p+1)\times 1}$, under fairly general conditions the maximum likelihood estimator $\hat{\theta}$ is asymptotically efficient and its asymptotic variance covariance matrix is the Cramer-Rao lower bound which is the inverse of the information matrix.

$$I\left(\hat{\theta}\right) = -E\left[\frac{\partial^2 \mathrm{Ln}\, L\left(\hat{\beta},\hat{\sigma}^2\right)}{\partial\hat{\theta}\partial\hat{\theta}'}\right]$$

$$\text{or } I\left(\hat{\theta}\right) = -E\left[\frac{\partial^2 \mathrm{Ln}\, L\left(\hat{\beta},\hat{\sigma}^2\right)}{\partial\hat{\theta}_i\partial\hat{\theta}_j}\right]_{i,j=1,2,..,(p+1)}$$

$$\text{or } I\left(\hat{\theta}\right) = \begin{bmatrix} \frac{Z\left(\hat{\beta}\right)' Z\left(\hat{\beta}\right)}{\hat{\sigma}^2} & O \\ O & \frac{n}{2\hat{\sigma}^4} \end{bmatrix}$$

This yields, $\mathrm{Var}\left(\hat{\beta}\right) = \hat{\sigma}^2\left[Z\left(\hat{\beta}\right)' Z\left(\hat{\beta}\right)\right]^{-1}$ and $\mathrm{Var}\left(\hat{\sigma}^2\right) = \frac{2\hat{\sigma}^4}{n}$

4. Linear pseudo model for nonlinear regression

Consider the general nonlinear regression model in vector notation as

$$Y_{n\times 1} = f_{n\times 1}\left(X_{n\times k}, \beta_{p\times 1}\right) + \varepsilon_{n\times 1} \tag{4.1}$$

such that $\varepsilon_i \overset{\text{i.i.d}}{\sim} \left(0,\sigma^2\right), i = 1,2,..,n$ or ε independently follows $\left(O,\sigma^2 I_n\right)$.

One may approximate $f\left(X,\beta\right)$ in β^* by the linear term of the Taylor series expansion such that neglecting the terms from second order derivatives, gives the best linear approximation to the nonlinear regression model. That is,

$$f\left(X,\beta\right) \simeq f\left(X,\beta^*\right) + Z\left(\beta^*\right)\left(\beta-\beta^*\right) \tag{4.2}$$

where $Z(\beta^*) = \begin{bmatrix} \left.\dfrac{\partial f(X_1,\beta)}{\partial \beta_1}\right|_{\beta^*} & . & . & \left.\dfrac{\partial f(X_1,\beta)}{\partial \beta_p}\right|_{\beta^*} \\ . & & & . \\ . & & & . \\ \left.\dfrac{\partial f(X_n,\beta)}{\partial \beta_1}\right|_{\beta^*} & . & . & \left.\dfrac{\partial f(X_n,\beta)}{\partial \beta_p}\right|_{\beta^*} \end{bmatrix}_{n\times p}$

For instance, the i^{th} nonlinear function $f(X_i,\beta)$ can be approximated by a linear function as

$$f(X_i,\beta) \simeq f(X_i,\beta^*) + \left[\left.\frac{\partial f(X_i,\beta)}{\partial \beta_1}\right|_{\beta^*} \cdots\cdots \left.\frac{\partial f(X_i,\beta)}{\partial \beta_p}\right|_{\beta^*} \right] (\beta - \beta^*) \tag{4.3}$$

Here, $\left[\left.\dfrac{\partial f(X_i,\beta)}{\partial \beta_1}\right|_{\beta^*} \cdots\cdots \left.\dfrac{\partial f(X_i,\beta)}{\partial \beta_p}\right|_{\beta^*} \right]$ is the i^{th} row of $Z(\beta^*)$

From the equations (4.1) and (4.2) one may obtain the linear pseudo model (Due to Malinvaud [7]) for nonlinear regression model as

$$Y = f(X,\beta^*) + Z(\beta^*)(\beta - \beta^*) + \varepsilon \tag{4.4}$$

or $$Y^* = Z(\beta^*)\beta + \varepsilon \tag{4.5}$$

where $$Y^* = Y - f(X,\beta^*) + Z(\beta^*)\beta^* \tag{4.6}$$

Since, β^* is unknown one cannot use the linear pseudo model (4.5) directly for parameter estimation. If $Z(\beta^*)$ and Y^* are known, then the least squares estimator for β is given by

$$\hat{\beta}_{LS} = \left[Z(\beta^*)' Z(\beta^*) \right]^{-1} \left[Z(\beta^*)' Y^* \right] \tag{4.7}$$

Further, if the errors $\varepsilon_i \overset{i.i.d}{\sim} \left(0, \overset{*}{\sigma}{}^2\right)$ under certain conditions, the least squares estimator $\hat{\beta}_{LS}$ of β will be consistent and asymptotically normally distributed.

Thus, the variance-covariance matrix of $\hat{\beta}_{LS}$ is given by

$$\mathrm{Var}(\hat{\beta}_{LS}) = \overset{*}{\sigma}{}^2 \left[Z(\beta^*)' Z(\beta^*) \right]^{-1} \tag{4.8}$$

Since, the model is only approximately correct, this holds approximately only. Suppose that the nonlinear least squares estimator $\hat{\beta}_{LS}$ is sufficiently close to β^*, then one may obtain an estimate of $\mathrm{Var}(\hat{\beta}_{LS})$ as

$$\hat{\mathrm{Var}}(\hat{\beta}_{LS}) = \hat{\sigma}^2_{LS} \left[Z(\hat{\beta}_{LS})' Z(\hat{\beta}_{LS}) \right]^{-1} \tag{4.9}$$

Here, $\hat{\sigma}^2_{LS}$ is least squares estimate of $\overset{*}{\sigma}{}^2$. For instance,

$$\hat{\sigma}^2_{LS} = \frac{\left[Y - f\left(X, \hat{\beta}_{LS}\right)\right]' \left[Y - f\left(X, \hat{\beta}_{LS}\right)\right]}{n - p} \tag{4.10}$$

Assuming that (4.5) is the correct model, and $\varepsilon \sim N\left(O, \overset{*}{\sigma}{}^2 I_n\right)$; under certain regularity conditions for the asymptotic theory,

$$\left[\frac{(n-p)\hat{\sigma}^2_{LS}}{\overset{*}{\sigma}{}^2}\right] = \frac{\varepsilon' \left\{I - Z(\beta^*)\left[Z(\beta^*)' Z(\beta^*)\right]^{-1} Z(\beta^*)'\right\}\varepsilon}{\overset{*}{\sigma}{}^2} \tag{4.11}$$

has a χ^2 distribution with $(n-p)$ degrees of freedom.

5. Conclusions

In the above research study very different methods of estimation of NLLS estimator, maximum likelihood estimators are proposed. In addition to this linear pseudo model for general nonlinear regression model has been derived.

References

[1] David Pollard and Peter Radchenko 2006 *Journal of Multivariate Analysis* **97** 548-562
[2] Anh V V 1998 Stochastic processes and their applications **29** 317-333
[3] OLemcoff N 1977 *The chemical engineering journal* **13**1 71-74
[4] Edmond Malinvaud A tribute to his contributions in Econometrics Peter C B PhilipsYale University Frank Nielson 2013 Camel University
[5] A kaike H 1973I nformation theory and an extension of the maximum likelihood principle
[6] Malinvaud E1970 Annuals of Mathematical Statistics **41** 956- 969
[7] Prakasa Rao B L S 1984 Journal of multivariate analysis**14(**13) 315-322
[8] Ivanov AV theory of probability & its applications **21(** 3) 557-570
[9] Chien Fu Wu 1981 the annals of statistics **9**(3) 501 -513
[10] David pollard and peter Radchenko 2006 *Journal of Multivariate Analysis* **97**548-562
[11] Mu B, Bai E W, Zheng W X and Zhu Q 2017 *Automatic a* **77**322-335
[12] Jae Myung 2003 *Journal of Mathematical Psychology* **47** (1)90-100