ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 12, No. 2, 2017, *pp*.107-112

Testing for outliers in nonlinear longitudinal data models based on M-estimation

Huihui Sun[1]
[1] *School of Mathematics and Statistics, Yancheng Teachers University, Yancheng, 224002, China, E-mail: sunhuihui12@163.com.*
(Received September 21, 2016, accepted January 04, 2017)

Abstract. In this paper we propose and analyze nonlinear mixed-effects models for longitudinal data, obtaining robust maximum likelihood estimates for the parameters by introducing Huber's function in the log-likelihood function. Furthermore, the test for outliers in the model based on robust estimation is investigated through generalized Cook's distance. The obtained results are illustrated by plasma concentrations data presented in Davidian and Giltiman, which was analyzed under the non-robust situation.

Keywords: M-estimation; nonlinear mixed models; longitudinal data; testing for outliers; generalized Cook's distance.

1. Introduction

Nonlinear mixed models can model mechanistic relationships between independent and dependent variables and can estimate more physically interpretable parameters (Pinheiro and Bates 2000), which are important to the analysis of longitudinal data, multi-level data and repeated survey data and widely used in the field of economics, bio-pharmaceuticals, agriculture. Recently, some different nonlinear mixed effects models and inference procedures have been proposed. Russo et al. (2009) proposed nonlinear elliptical models for longitudinal data and presented diagnostic results based on residual distances and local influence (Cook, 1986 and 1987). Wei and Zhong (2001) focused on influence analysis in nonlinear models with random effects.

In standard analysis of data well-modeled by a nonlinear mixed model, an outlying observation can greatly distort parameter estimates and subsequent standard errors. Consequently, inferences about parameters are misleading. Then, a robust procedure is needed for accurate results in the presence of outliers. M-estimation is the most widely used robust estimation method, which was firstly introduced in Huber's article (1981) on regression. Mancini et al. (2005) and Muler and Yohia (2008) proposed a robust M-estimator that assigns a much lower weight to outliers than traditional maximum likelihood estimators does. Pinheiro et al. (2001) and Staudenmayer et al. (2009) studied robust estimation techniques in which both random effects and errors have multivariate Student-t distributions. While, less alternatives have been studied for outlier accommodation in the context of nonlinear mixed-effects models. Yeap and Davidian (2001), who proposed a two-stage robust estimation in nonlinear mixed-effects models when outliers are presented, is one of the few references that address this case. Meza et al. (2012) presented an extension of a Gaussian nonlinear mixed-effects model using heavy-tailed multivariate distributions for both random effects and residual errors. James et al. (2015) proposed an outlier robust method based on linearization to estimate fixed effects parameters and variance components in nonlinear mixed model. However, little attention has been paid to the influence diagnostic for nonlinear mixed models in the current literature. In this article, we introduce a robust method by utilizing a robust version of the log-likelihood for the nonlinear mixed model and investigate the test for outliers in the model based on robust estimation by generalized Cook's distance, extending and expanding the studies of Gill (2000) and James et al. (2015). Our results show that the generalized Cook's distance based on robust estimation can successfully detect the masking effects that appear in the data set.

The rest of the article is organized as follows. Section 2 introduces the nonlinear mixed effects model discussed in this paper and uses Fisher scoring method to get M-estimation of parameters. And the asymptotic properties is also established. In Section 3, we investigate the test for outliers in nonlinear model based on robust estimates. In Section 4, as an illustration, we apply the proposed method to analyze an observational data set. Finally, some conclusions are given and possible future work is discussed in Section 5.

2. Model and robust estimation

Assume that response measurements are collected on N subjects and the k-th subject being observed on n_k time points, thus $M = \sum_{k=1}^{N} n_k$ is the total number of measurements. In the matrix notation, the model for measurements from subject k is

$$y_k = f(X_k, \beta) + C_k \tau_k + e_k, \quad k = 1, 2, \cdots, N, \tag{2.1}$$

where $y_k = (y_{k1}, \cdots, y_{kn_k})^T$ is a vector of length n_k containing observable response variable from subject k; $f(\cdot,\cdot)$ is a known second order differentiable nonlinear function of the regression vector β, which is a vector of q unknown but fixed parameters with known design matrix X_k, and $X_k = (x_{k1}, \cdots, x_{kn_k})^T$; C_k is the $n_k \times r$ design matrix for the random effects of subject k, τ_k is a $r \times 1$ vector of random effects assumed to be sampled from a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2\Gamma$. $e_k = (e_{k1}, \cdots, e_{kn_k})^T$ is an $n_k \times 1$ unobservable random error and $e_k \sim N(0, \sigma^2\Omega_k)$. It is also assumed that τ_k and e_k are independent from each other. Then, $cov(y_k) = \sigma^2\Sigma_k = \sigma^2 C_k \Gamma C_k^T + \sigma^2\Omega_k$.

Let α denote the vector of unknown parameters in Σ_k, and the log-likelihood for the nonlinear mixed model is

$$l(\beta, \alpha \mid y) = cons\tan t - \frac{1}{2} M \log\sigma^2 - \frac{1}{2}\sum_{k=1}^{N} \log|\Sigma_k| - \sum_{k=1}^{N} \frac{1}{2} \varepsilon_k^T \varepsilon_k, \tag{2.2}$$

which is obtained from the marginal model from (2.1) (Russo et al., 2009), where $\varepsilon_k = \sigma^{-1}\Sigma_k^{-1/2}(y_k - f(X_k, \beta))$. Note that the last term of (2.2) is a half sum of squares and grows quickly. Following the M-estimation theory expressed in Huber (1981), the log-likelihood function l is robustified by replacing it with a function that increases much slower.

In this paper, Huber ρ function is chosen to bound the influence of outlying observations on the estimation, which is defined by

$$\rho(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2 & if \mid \varepsilon \mid \le c \\ c \mid \varepsilon \mid - \frac{1}{2}c^2 & if \mid \varepsilon \mid > c \end{cases},$$

where c is the Huber tuning constant and usually $c \in [0.7, 2]$, here c=1.345 (Ripley, 2004). The first derivative of $\rho(\varepsilon)$ with respect to ε, is given by

$$\psi(\varepsilon) = \partial\rho(\varepsilon)/\partial\varepsilon = \begin{cases} \varepsilon & if \mid \varepsilon \mid \le c \\ csign(\varepsilon) & if \mid \varepsilon \mid > c \end{cases}.$$

Therefore, the robustified version of (2.2) is given by

$$\eta(\beta, \alpha \mid y) = cons\tan t - \frac{1}{2}\kappa_1 M \log\sigma^2 - \frac{1}{2}\kappa_1 \sum_{k=1}^{N} \log|\Sigma_k| - \sum_{k=1}^{N}\sum_{j=1}^{n_k} \rho(\varepsilon_{jk}), \tag{2.3}$$

where $\kappa_1 = E(\varepsilon\psi(\varepsilon)) = Pr(\mid \varepsilon \mid \le c)$ is the consistency correction factor and is needed to make the estimating equations have zero expectation.

Then, we can obtain robust maximum likelihood estimation (RMLE) through Fisher scoring method (Gill, 2000 and James et al., 2015) based on (2.3). Note that the robust maximum likelihood estimation becomes the classical maximum likelihood estimation as c approaches infinity in Huber function. Now we study the asymptotic properties of RMLE. Let $\vartheta = (\beta^T, \sigma^2, \alpha^T)^T$, by Domowitz and White (1982), we consider that the RMLE ϑ_N of ϑ is obtained by maximizing an objective function in the form

$$G_N(y, \vartheta) = \frac{1}{N}\sum_{i=1}^{N} g(y_i, \vartheta),$$

and the estimating equation for ϑ is as follows

$$G_N^{'}(y,\vartheta)=\frac{1}{N}\sum_{i=1}^{N}g'(y_i,\vartheta)=0,$$

where $g'(y_i,\vartheta)=\partial g(y_i,\vartheta)/\partial\vartheta$. Consider that ϑ_0 is obtained by maximizing the objective function

$$\bar{G}_N(\vartheta)=E[G_N(y,\vartheta)].$$

So we have $E[G_N^{'}(y,\vartheta_0)]=0$. Now using the mean value theorem, we get

$$G_N^{'}(y,\vartheta_N)\approx G_N^{'}(y,\vartheta_0)+G_N^{''}(y,\vartheta)(\vartheta_N-\vartheta_0) \tag{2.4}$$

where $\vartheta=\alpha'\vartheta_N+(1-\alpha')\vartheta_0, 0<\alpha'<1$. Because $G_N^{'}(y,\vartheta_N)=0$, (2.4) gives

$$\sqrt{N}(\vartheta_N-\vartheta_0)=[-G_N^{''}(y,\vartheta)]^{-1}(\sqrt{N}G_N^{'}(y,\vartheta_0)).$$

Property 2.1 shows that the M-estimation for ϑ has the same properties as MLE.

Property 2.1 Under some assumptions (Domowitz and White, 1982), according to Sanjoy (2004), the following consistency and asymptotic normality property can be got

$(a)\vartheta_N\rightarrow\ \vartheta_0\quad a.s.$

$(b)\sqrt{N}(\vartheta_N-\vartheta_0)\rightarrow N(\mathbf{0},\mathbf{J}_N^{-1}(\vartheta_0))$

where $\mathbf{J}_N^{-1}(\vartheta_0)=M_N^{-1}Q_NM_N^{-1}, M_N=-\bar{G}_N''(\vartheta_0), \bar{G}_N''(\vartheta)=E[G_N''(y,\vartheta)], Q_N=var[\sqrt{N}G_N'(y,\vartheta_0)]$, Domowitz and White (1982) gives the detailed proof.

3. Case deletion model and generalized Cook's distance

The data deletion model is one of the most basic statistical diagnosis model. The basic method commonly used to study outliers or strong influence points is to compare the difference of the estimator between the data point deleted before and after. For nonlinear mixed effects model of longitudinal data, there are two types of data deletion: one is to delete an observation value of an individual, that is (y_{ij},x_{ij}), and the other is to delete the whole set of observations of an individual, i.e. (y_i^T,x_i^T). For convenience, this paper only discusses the second case.

When the whole data set of the *i*-th individual is deleted, model (2.1) becomes

$$y_j=f(X_j,\ \beta)+C_j\tau_j+e_j,\quad j\neq i. \tag{3.1}$$

The robustified log-likelihood function is given by

$$\eta_{[i]}(\beta,\alpha,\sigma^2)=constant-\frac{1}{2}\kappa_1(M-n_i)log\sigma^2-\frac{1}{2}\kappa_1\sum_{k=1,k\neq i}^{N}log|\Sigma_k|-\sum_{k=1,k\neq i}^{N}\sum_{j=1}^{n_k}\rho(\varepsilon_{jk}), \tag{3.2}$$

where the subscript [*i*] denotes the deletion of observations of the *i*-th individual. Let $\vartheta_{[i]}$ be the robust estimate of ϑ after deleting the *i*-th individual. The difference between $\vartheta_{[i]}$ and ϑ can be compared to measure the influence of the *i*-th individual on ϑ. If the difference is large, the *i*-th individual may be a strong influence individual. According to Cook and Weisberg (1982), define the generalized Cook's distance to measure $\vartheta-\vartheta_{[i]}$.

Definition 3.1 The generalized Cook's distance based on the Fisher information matrix of ϑ for model (2.1) and its case deletion model (3.1) is given by

$$GC_i=(\vartheta-\vartheta_{[i]})^T E\left(-\frac{\partial^2\eta}{\partial\vartheta\partial\vartheta^T}\right)\Bigg|_{\vartheta}(\vartheta-\vartheta_{[i]}). \tag{3.3}$$

If the value of generalized Cook's distance corresponding to the *i*-th individual GC_i is larger than other values, the *i*-th individual is considered to be a strong influential individual.

4. An illustrative example

Plasma drug penetration data was collected by the following experiment: Firstly, six volunteers were injected the same dose of a drug through veins. Then, the drug concentration in their plasma was measured at eleven times within eight hours. Davidian and Giltiman (1995) applied the following double exponential model to describe the data:

$$y_{ij} = e^{\beta_1}\exp(-e^{\beta_2}x_{ij}) + e^{\beta_3}\exp(-e^{\beta_4}x_{ij}) + \tau_i + e_{ij}, i = 1,2,\cdots,6; j = 1,2,\cdots,11,$$

where y_{ij} denotes the drug concentration of the *i*-th volunteer at the *j*-th time; x_{ij} is the time interval of the *i*-th volunteer measure the *j*-th time drug concentration; τ_i is the drug effect of the *i*-th volunteer and $\tau_i \sim N(0,\sigma^2\Gamma)$; e_{ij} is the random error that independent with τ_i and we suppose $e_{ij} \sim N(0,\sigma^2)$. Design matrix

$$X_i = (0.25, 0.50, 0.75, 1.00, 1.25, 2.00, 3.00, 4.00, 5.00, 6.00, 8.00)^T.$$

First, we get the RMLE of parameters by the method we proposed in the paper. The results are shown in Table 1.

Table 1. RMLE of parameters

Parameters	β_1	β_2	β_3	β_4	σ^2	Γ
RMLE	1.3419	0.8294	-0.4201	-0.9637	0.0369	0.6491
MSE	0.1902	0.1837	0.2436	0.2010	0.2568	0.3741

Now calculate the generalized Cook's distance for each individual using formula 3.3. Consider the scatter plot of the generalized Cook's distance as shown in Figure 1.

The generalized Cook's distance is calculated based on RMLE in Figure 1(a) and based on MLE in Figure 1(b). As can be seen from the two figures, the generalized Cook's distance for the third individual is the largest, followed by the forth individual. So it can be considered that the third individual is a strong influential individual. In addition, we can see from Figure 1(a) and 1(b) that the generalized Cook's distance for the forth individual is larger in Figure 1(a). Therefore, the generalized Cook's distance presented in the paper based on M-estimation may identify potential outliers, which could overcome the masking phenomenon.

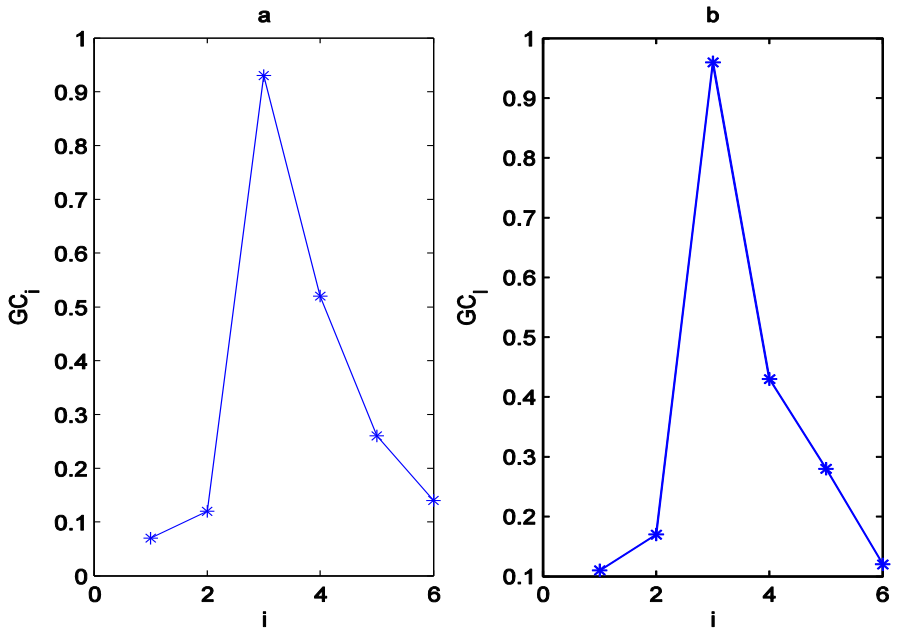

Figure 1. generalized Cook's distance

5. Conclusion

Nonlinear mixed effects models are very useful statistical tools in analyzing nonlinear data with random effects. However, when there are outliers in the data, current nonrobust estimation methods may produce misleading results, so a robust method is needed. In this article, the log-likelihood function of the nonlinear mixed effects model is altered by incorporating an appropriate loss function, ρ, of the residuals. Fisher scoring method is applied to get M-estimation of the parameters for the means model and for the variance components. Then, generalized Cook's distance based on M-estimation is presented in the paper. Here Huber ρ function is used to bound the influence of outlying observations. An alternative function is the one suggested by Andrews et al.(1972), called the Bisquare function, is

$$\rho(\varepsilon)=\begin{cases} -\frac{c^2}{6}\left(\left(1-(\frac{\varepsilon}{c})^2\right)^3-1\right) & if\ |\varepsilon|\le c \\ \frac{c^2}{6} & if\ |\varepsilon|>c \end{cases},$$

where *c* is the tuning constant. The first derivative of $\rho(\varepsilon)$ with respect to ε, denoted $\psi(\varepsilon)=\partial\rho(\varepsilon)/\partial\varepsilon$, is given by

$$\psi(\varepsilon)=\begin{cases} \varepsilon\left(1-\left(\frac{\varepsilon}{c}\right)^2\right)^2 & if\ |\varepsilon|\le c \\ 0 & if\ |\varepsilon|>c \end{cases}.$$

We can also use perturbation diagnostics based on M-estimation to overcome the masking phenomenon when identifying potential outliers, such as mean shift perturbation, case weights perturbation, perturbation of covariate in random effects and so on. The key point of all the above perturbations is to find $-\ddot{F}$.

6. Acknowledgement

The project supported by National Statistical Science Research Project of China (2015LZ27).

7. References

[1] Andrews D. F., Bickel P. J., Hampel F. R., Huber P. J., Rogers W. H. and Tukey J. W. Robust Estimates of Location: Survey and Advances. Princeton University Press: New Jersey. (1972).

[2] Banerjee M., Frees E. W. Influence diagnostics for longitudinal models. Journal of the American Statistical Association. 439: 999-1005(1997).

[3] Cook, D. Assessment of local influence. Journal of the Royal Statistical Society - Series B. 48(2): 133-169(1986).

[4] Cook, D. Influence assessment. Journal of Applied Statistics. 14(2):117-131(1987).

[5] Diggle, P. J., Liang, K. Y. and Zeger, S. T. Analysis of longitudinal data. New York: Oxford University Press. (2002).

[6] Davidian M., Giltiman D. M. Nonlinear Models for Repeated Measurement Data. London: Chaman and Hall. (1995).

[7] Domowitz, I., White, H. Misspecified models with dependent observations. Journal of econometrics. 20: 35-58(1982).

[8] Fung W.K.,Kwan C.W. A note on local influence based on normal curvature. J R Statist Soc B. 59: 839-843(1997).

[9] Gill, P. S. A robust mixed linear model analysis for longitudinal data. Statistics in Medicine. 19: 975-987(2000).

[10] Huber, P.J. Robust Statistics. New York: Wiley. (1981).

[11] James D.W.,Jeffrey B. B. and Abdel-Salam G. A. Outlier robust nonlinear mixed model estimation. Statistics in Medicine. 34: 1304-1316(2015).

[12] Lesaffre E., Verbeke G. Local influence in linear mixed models. Biometrics. 54: 570-582(1998).

[13] Mancini L., Ronchetti E., Trojani F. Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. Journal of the American Statistical Association. 105: 628–641(2005).

[14] Muler N., Yohai V.J. Robust estimates for GARCH models. Journal of Statistical Planning and Inference. 138: 2918–2940(2008).

[15] Meza A., Osorio F., De la Cruz R. Estimation in nonlinear mixed-effects models using heavy-tailed distributions. Statistics and Computing. 22: 121–139(2012).

[16] Pinheiro J. C., Liu C., Wu Y. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t-distribution. Journal of Computational and Graphical Statistics. 10: 249–276(2001).

[17] Pinheiro, J. C., Bates, D. M. Mixed-effects models in S and S-plus. New York: Springer. (2000).

[18] Russo,C.M., Paula,G.A. and Aoki,R. Influence diagnostics in nonlinear mixed effects elliptical models. Computational Statistics & Data Analysis. 53: 4143-4156(2009).

[19] Ripley B. D. Robust statistics. M. Sc. in Applied Statistics. 2: 1-3(2004).

[20] Sinha S.K. Robust analysis of generalized linear mixed models. Journal of the American Statistical Association. 466: 451-460(2004).

[21] Staudenmayer J., Lake E.E., Wand M.P. Robustness for general design mixed models using the t-distribution. Statistical Modelling. 9: 235–255(2009).

[22] Vanegas,L.H., Cysneiros, F.J.A. Assessment of diagnostic procedures in symmetrical nonlinear regression models. Computational Statistics & Data Analysis. 54: 1002-1016(2010).

[23] Wei B. C., Zhong X. P. Influence analysis in nonlinear models with random effects. Applied Mathematics: A Journal of Chinese Universities. 16(1): 35-44(2001).

[24] Yeap B.Y., Davidian M. Robust two-stage estimation in hierarchical nonlinear models. Biometrics. 57: 266–272(2001).