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