A Diagnostic of Influential Cases in Generalized Linear Mixed Models

Junfeng Shang *
Department of Mathematics and Statistics
Bowling Green State University, USA

Abstract: In the generalized linear mixed modeling (GLMM) framework, we develop a diagnostic for detecting influential cases based on the Information Complexity Criteria (ICOMP). The diagnostic compares the information complexity criteria between the full data set and a case-deleted data set. A real data set of cancer cells is analyzed using the logistic linear mixed model for illustrating the effectiveness of the proposed diagnostic.

Key words: GLMM, ICOMP, case-deletion, Fisher information matrix, logistic regression model.

1 Introduction

In the statistical literature, modeling diagnostics have attracted sufficient attention. In linear regression, measures such as COOK's distance, DFBETAS, DFFITS, studentized residuals, and COVRATIO are popularly applied to reveal influential cases which substantially impact the fitted model and associated results (see Belsley, Kuh, and Welsh, 1980; Cook and Weisberg, 1982). The modeling diagnostics have been furthermore developed in the work of Johnson and Geisser (1983), Johnson (1985), Cavanaugh and Johnson (1999), Cavanaugh and Oleson (2001). Christensen, Pearson, and Johnson (1992) proposed case-deletion diagnostics for detecting influential observations in mixed linear models. Cavanaugh and Shang (2005) developed a predictive influence function (PIF) for discovering the influential cases for the prediction of the random effects in a mixed model.

The identification of influential cases involves the problem of model selection since the detected influential cases may be caused by the simplicity of the model. Bozdogan and Bearse (2003) developed a modeling diagnostic using the information complexity (ICOMP) criteria in dynamic multivariate linear models. In their work, influential case detection and model selection have been addressed jointly. Shang (2008) developed a modeling diagnostic using ICOMP in linear mixed modeling setting.

Little attention has been put on generalized linear mixed modeling diagnostics. The generalized linear models (GLMs) are an extension of the linear modeling process that allows models to be fit to the data that follow probability distributions other than the normal distribution, such as the Poisson, Binomial, Multinomial distributions. GLMs also relax the requirement of equality or constancy of variances that is required for hypothesis tests in traditional linear models. The generalized linear mixed model (GLMM) is named when random effects are involved in GLMs, and the application of GLMMs can be even more extensively carried out than GLMs.

Detecting influential cases in generalized linear mixed models becomes quite crucial because GLMs with random effects provide a wider possible fit to various types of data in practice. For

*Phone: (419) 372-7457. Fax: (419) 372-6092. e-mail: jshang@bgnet.bgsu.edu. Department of Mathematics and Statistics, 450 Math Science Building, Bowling Green State University, Bowling Green, OH 43403.

instance, in biostatistical or medical studies, logit and log-link models are appropriate for a variety types of repeated measurements where a response variable and a collection of covariates are measured on each subject. Influential cases (subjects) can misrepresent the population and further affect the accuracy of statistical inferences. Moreover, the existence of influential cases is practically quite possible. To minimize the negative effect of the influential case, the key is to detect influential cases with the facilitation of diagnostic measures.

We develop a diagnostic for detecting influential cases based on the information complexity criteria in the generalized linear mixed modeling framework. The diagnostic compares the information complexity criteria between the full data set and a case-deleted data set. The information complexity criterion is computed from the Fisher information matrix.A real data set of cancer cells is analyzed using the logistic linear mixed model for illustrating the effectiveness of the proposed diagnostic.

2 The Information Complexity (ICOMP) Criterion in Generalized Linear Mixed Models

2.1 Generalized Linear Models with Random Effects

A generalized linear model with random effects is defined in what follows. Let $Y = (y_1, \cdots, y_N)'$ be a vector of N observations, which can be written as

$$Y = \mu + \epsilon, \tag{2.1}$$

where ϵ is a vector of random errors with zero expectation and covariance matrix V, and V involves parameters to be estimated. Let further $g(.)$ be the link function, which is monotone, such that $g(\mu)$ can be written as the linear model

$$g(\mu) = \eta = X\beta + U\xi. \tag{2.2}$$

Here $X_{N\times p}$ is a known design matrix, the β is a vector of fixed effects, the U is an $N \times q$ known matrix, and the ξ is a $q \times 1$ vector of random effects. If conditionally on μ, the components of Y are independently distributed, and if their distribution is a member of the exponential family, the models (2.1) and (2.2) define a generalized linear model with random effects, also called a generalized linear mixed model (GLMM).

Suppose that the random effects ξ is partitioned as $[\xi_1, \cdots, \xi_r]$ and let $q_1 + \cdots + q_r = q$ and $U = [U_1, \cdots, U_r]$. The random vectors $\xi_1, \cdots, \xi_r$ are assumed to be uncorrelated with zero expectation. The random effects are also assumed to be uncorrelated with ϵ. Further, $cov(\xi_i) = \sigma_i^2 I_{q_i} (i = 1, \cdots, r)$ and $cov(\xi) = D = \text{diag}(\sigma_1^2 I_1, \cdots, \sigma_r^2 I_r)$, where $I_1, \cdots, I_r$ are identity matrices of orders $q_1 \times q_1, \cdots, q_r \times q_r$.

For the purpose of further analysis, the Y data can be linearized (McCullagh and Nelder, 1989, p. 31), and then the link function $g(.)$ is re-written by providing the first order as

$$g(Y) = g(\mu) + (Y - \mu)g'(\mu) = Z. \tag{2.3}$$

and therefore Z is called the adjusted dependent variable. Correspondingly, let $Z = (z_1, \cdots, z_N)'$ be a vector of N observations. From now on, instead of using Y, we will utilize the Z to propose the diagnostic of influential cases in modeling. From (2.1) and (2.2) to (2.3), we can have a linear random effects model for Z is

$$Z = X\beta + U\xi + \epsilon g'(\mu). \tag{2.4}$$

We know that $E(Z) = X\beta$ and $cov(\xi) = D$. Therefore, $cov(\epsilon g'(\mu)) = V(\partial\eta/\partial\mu)^2 = W$, $cov(Z) = W + UDU' = \Sigma$, and $W = V(\partial\eta/\partial\mu)^2 = V\,\mathrm{diag}\{(\partial\eta_i/\partial\mu_i)^2\}$, $i = 1, \cdots, N$.

Model (2.4) is a linear random effects model with the adjusted dependent variable Z instead of Y and therefore is considered as a linear mixed model. However, the covariance matrix Σ here is not as simple as that in a linear mixed model because it is a function of the fixed effects β.

2.2 The Information Complexity (ICOMP) in Generalized Linear Mixed Models

Prior to the introduction of the information complexity (ICOMP) criterion, the most recognized model selection criterion, the Akaike Information Criterion (AIC, Akaike, 1973, 1974), is presented and discussed for the purpose of comparison with the ICOMP criterion.

Akaike's (1973) original AIC is given by

$$\mathrm{AIC} = -2\,\log L(\hat{\theta} \mid Z) + 2k,$$

where $L(\hat{\theta} \mid Z)$ is the maximized likelihood function, and k represents the dimension of estimated parameter $\hat{\theta}$ under the given model.

For model (2.4), let θ denote the vector containing β and the parameters in the D. Note that the W is a function of μ, and μ is a function of β and random effects ξ with covariance D as shown in model (2.2). Here, the "goodness of fit" term, $-2\,\log L(\hat{\theta} \mid Z)$, gauges how well the model fits the data, and the penalty term, $2k$, measures the complexity that compensates for the bias in the lack of fit when the maximum likelihood estimators are used. The success of AIC depends on its approximation to the bias adjustment by $2k$ for large samples.

Suppose the generating model or the true model, which presumably gave rise to the data. Also, suppose that a candidate or approximating model is a model that could potentially be used to describe the data and a fitted model is a candidate model that has been fit to the data. AIC is justified as an asymptotic unbiased estimator of Kullback-Leibler discrepancy between the generating model and a fitted model. The formation of AIC reflects an underlying principle for model selection criteria, that is, a model selection criterion involves both a goodness of fit term gauging how well the model fits the data and a penalty term measuring the model complexity. AIC penalizes the complexity of model by two times of the number of estimated parameters.

Similar to AIC, the Information Complexity Criterion (ICOMP) criterion combines a goodness-of-fit term with a term for measuring the complexity of model. In what follows, we will see that instead of penalizing the number of estimated parameters, the ICOMP criterion penalizes the covariance complexity of model.

The ICOMP criterion is based on the covariance complexity index of van Emdan (1971) in parametric estimation. The ICOMP criterion is defined as

$$ICOMP = -2\log L(\hat{\theta} \mid Z) + 2\,C(\hat{Q}), \tag{2.5}$$

where $L(\hat{\theta} \mid Z)$ represents the maximized likelihood function, $\hat{\theta}$ represents the maximum likelihood estimator of the unknown parameter θ, C represents a complexity measure, Q represents the covariance matrix of the estimated parameters for the model, and correspondingly $\hat{Q}$ represents the estimated covariance matrix of Q. Note that in original definition of the ICOMP criterion, the $\hat{\theta}$ could be any estimator of θ. In this paper, we utilize the maximum likelihood estimator (MLE) of θ.

Note that the ICOMP criterion and the AIC share the similarity in containing two terms, one is the goodness of fit term, $-2\log L(\hat{\theta} \mid Z)$; the other one is the penalty term. However, the penalty term of AIC is $2k$, two times of the number of estimated parameters, whereas the penalty term of the ICOMP criterion is the measure of the covariance complexity for model.

To evaluate the complexity measure of the ICOMP criterion, Bozdogan (1988, 1990, 1993, 1994) proposed a maximal information complexity measure which is expressed as

$$C_m(Q) = \frac{m_k}{2}\log\frac{\mathrm{tr}(Q)}{m_k} - \frac{1}{2}\log|Q|, \tag{2.6}$$

where m_k is the dimension of Σ. This measure is optimal in that it is invariant with respect to scaler multiplication and orthonormal transformation and in that it is a monotonically increasing function of the dimension m_k of Q. (See Bozdogan, 1988, 1990 for details.)

In the linear mixed model (2.4), recall that $cov(\epsilon g'(\mu)) = V(\partial\eta/\partial\mu)^2 = W$, $cov(Z) = W + UDU' = \Sigma$, and $W = V(\partial\eta/\partial\mu)^2 = V\mathrm{diag}\{(\partial\eta_i/\partial\mu_i)^2\}$, $i = 1, \cdots, r$. V is the covariance matrix of Y conditional on μ. Write $U = [U_1, \cdots, U_r]$ and $\xi' = [\xi_1', \cdots, \xi_r']$ with $Cov(\xi_i) = \sigma_i^2 I_{q_i}$ and $Cov(\xi_i, \xi_j) = 0$. Let q_i denote the number of columns in ξ_i and then I_{q_i} is a $q_i \times q_i$ identity matrix. Therefore, the covariance matrix of ξ is a block diagonal matrix with blocks $\sigma_i^2 I_{q_i}$. In model (2.4), the covariance of Z can be re-written as

$$\Sigma = \sum_{i=1}^{r} \sigma_i^2 U_i U_i' + W.$$

Now, the unknown parameter vector θ consists of the elements of the vector β and the scalars $\sigma_1^2, \cdots, \sigma_r^2$. Instead of estimating the matrix D, we need to estimate scalars $\sigma_1^2, \cdots, \sigma_r^2$. We have $p + r$ parameters to evaluate. Let $\hat{\theta}$ denote the MLE of θ, and $\hat{\theta} = (\hat{\beta}', \hat{\sigma}_1^2, \cdots, \hat{\sigma}_r^2)$. We utilize Schall's method (1991) to estimate the MLE's.

In model (2.4), the covariance matrix of the estimated parameters Q is unknown in closed form, we therefore employ the estimated inverse-Fisher information matrix to assess the complexity. Let F represents the Fisher information matrix for the model, then let F^{-1} denote the inverse of F. The estimated inverse-Fisher information matrix $\hat{F}^{-1}$ is obtained with $\hat{\theta}$ in place of θ in the matrix F^{-1}.

By the expressions (2.5) and (2.6), we therefore re-write the ICOMP criterion for model (2.4) as

$$\begin{aligned} ICOMP &= -2\log L(\hat{\theta} \mid Z) \\ &\quad + m_k \log\frac{\mathrm{tr}(\hat{F}^{-1})}{m_k} - \log|\hat{F}^{-1}| \\ &= N\log(2\pi) + \log|\hat{\Sigma}| \\ &\quad + (Z - X\hat{\beta})'\hat{\Sigma}^{-1}(Z - X\hat{\beta}) \\ &\quad + m_k \log\frac{\mathrm{tr}(\hat{F}^{-1})}{m_k} - \log|\hat{F}^{-1}|. \end{aligned} \tag{2.7}$$

To derive the matrix F, the second derivative of the log-likelihood is needed, and its derivation is not trivial. Since the link functions in the GLMMs are different, the F formats are different. In the paper, the example of a logistic regression model is utilized, so the F matrix for the logistic link function is derived and the derivation will not be presented in the paper.

2.3 A Diagnostic of Influential Cases Based on the ICOMP Criterion

For the identification of influential cases, the idea of leave-one-out method is typically utilized to develop measures for identifying influential cases. This idea compares inferential quantities such as regression parameter estimates, fitted values, and estimated variances based on a fitted model to

the full data set with those based on fitting a model to the data set with a case deleted. For instance, Cook (1977, 1979) effectively applied leave-one-out method and developed numerous measures for detecting influential observations in linear regression modeling framework.

We propose a diagnostic which makes use of the deletion of cases at a time based on the ICOMP criteria. The diagnostic is defined by the discrepancy of the two ICOMP criteria, one is computed based on the full data; the other one is computed based on a case-deleted data set. We define the diagnostic as

$$\delta_{ICOMP}(i) = ICOMP_{\text{Full-Data}} - ICOMP^{(i)}, \tag{2.8}$$

where $ICOMP_{\text{Full-Data}}$ is the ICOMP criterion value for a fitted mixed model when the full data set is utilized; $ICOMP^{(i)}$ is the ICOMP criterion value for the same fitted mixed model when the i^{th} case is deleted.

Straightforwardly, the magnitude of $\delta_{ICOMP}(i)$ reflected on definition (2.8) evaluates the influence of y_i on the ICOMP criterion. We recall that the ICOMP criterion consists of two terms and essentially takes into account of both goodness of fit and model complexity. The magnitude of $\delta_{ICOMP}(i)$ therefore combines the influences of y_i on both goodness of fit and on model complexity.

For the evaluated $\delta_{ICOMP}(i)$ values, we need a standard to determine which cases are influential. Suppose each case is potentially abnormal or influential. Once this case is removed, the leave-one-out data will make the model better fit. In this sense, the leave-one-out ICOMP criterion, i.e., $ICOMP^{(i)}$, will shrink compared to the ICOMP criterion under the full data set. Thus, the $\delta_{ICOMP}(i)$ value is positive. However, positive diagnostics only indicate that the corresponding cases are potentially influential. When the diagnostic values are positive for some cases, we hope to reveal the most influential cases. As a result, among all the evaluated $\delta_{ICOMP}(i)$ for the cases in a data set, the outstanding positive ones specify the influential cases.

Although AIC is similar to the ICOMP criterion for model selection, the analogous criterion based on AIC cannot provide a diagnostic of influential cases as effective as $\delta_{ICOMP}(i)$ because the dimension of estimated parameters is identical for both the full data set and a case-deleted data set, and the difference of the model complexity cannot be measured.

3 The Information Complexity (ICOMP) in the Logistic Linear Regression with Random Effects

3.1 An Application of the Proposed Diagnostic in the Logistic Linear Regression Model with Random Effects

In what follows, we consider the logistic linear regression with random effects in the setting of generalized linear models as an example to illustrate the perfomance and effectiveness of the proposed diagnostic for distinguishing the influential cases in the data set for the model. Let $Y = (y_1, \cdots, y_N)'$ be a vector of N observations, which can be written as

$$Y = \mu + \epsilon, \tag{3.1}$$

where ϵ is a vector of random errors with zero expectation and covariance matrix V given the μ, and V is a diagonal matrix with the element $V_i = \frac{\mu_i}{n_i}(n_i - \mu_i)$, $i = 1, \cdots, N$, the n_i is the total number of trials for the binomial distribution. Let the link function be the logit, which is monotone, such that $g(\mu)$ can be written as the linear model

$$g(\mu) = \eta = X\beta + U\xi. \tag{3.2}$$

Here $g(\mu)$ is an $N \times 1$ vector containing the element $\eta_i = \log \frac{\mu_i}{n_i - \mu_i}$, $i = 1, \cdots, N$. Here $X_{N\times p}$ is a known design matrix, the β is a vector of fixed effects, the U is an $N \times q$ known matrix, and the ξ is $q \times 1$ vector of random effects. If conditionally on μ the components of Y are independently distributed.

Suppose that random effects ξ is partitioned as $[\xi_1, \cdots, \xi_r]$ and let $q_1 + \cdots + q_r = q$ and $U = [U_1, \cdots, U_r]$. The random vectors $\xi_1, \cdots, \xi_r$ are assumed to be uncorrelated with zero expectation. The random effects are also assumed to be uncorrelated with ϵ. Further, $cov(\xi_i) = \sigma_i^2 I_{q_i} (i = 1, \cdots, r)$ and $cov(\xi) = D = \text{diag}(\sigma_1^2 I_1, \cdots, \sigma_r^2 I_r)$, where $I_1, \cdots, I_r$ are identity matrices of orders $q_1 \times q_1, \cdots, q_r \times q_r$.

Again, the link function $g(.)$ is applied to the data Y (McCullagh and Nelder, 1989, p.31) is linearized, providing the first order by

$$g(Y) = g(\mu) + (Y - \mu)g'(\mu) = Z. \tag{3.3}$$

Z is called the adjusted dependent variable. Correspondingly, let $Z = (Z_1, \cdots, Z_N)'$ be a vector of N observations. From (3.1) and (3.2) to (3.3), we can have a linear random effects model for Z is

$$Z = X\beta + U\xi + \epsilon g'(\mu). \tag{3.4}$$

We know that $E(Z) = X\beta$ and $cov(\xi) = D$. Therefore, $cov(\epsilon g'(\mu)) = V(\partial\eta/\partial\mu)^2 = W$, $cov(Z) = W + UDU' = \Sigma$, and $W = V(\partial\eta/\partial\mu)^2 = V\text{diag}\{(\partial\eta_i/\partial\mu_i)^2\} = \text{diag}\{\frac{n_i}{\mu_i(n_i-\mu_i)}\}$, $i = 1, \cdots, N$.

For the illustration of the effectiveness of the diagnostic, we apply the diagnostic to a data set which comes from an experiment to measure the mortality of cancer cells under radiation from Schall's paper (1991). For this data set, four hundred cells ($n_i = 400$) were placed on a dish, and three dishes were irradiated at a time, or occasion. After the cells were irradiated, the surviving cells were counted. Since cells would also die naturally, dishes with cells were put in the radiation chamber without being irradiated, to establish the natural mortality. Taking the difference for the two mortalities will be the one of the cancer. This data set can be described by model (3.1) and (3.2), and the models can be linearized by (3.3) and (3.4). The results will demonstrate that the proposed diagnostic can effectively flag the influential cases in the mixed model which is rewritten from the logistic linear model.

To describe the cancer cell data and to avoid the presence of extra-binomial variation, the model is written as

$$\log \frac{\mu_{ij}}{n - \mu_{ij}} = \beta + \xi_{1i} + \xi_{2ij}, \tag{3.5}$$

where $i = 1, \cdots, 9$, $j = 1, \cdots, 3$. The number of locations is 9, and ξ_{1i} are the random effects originating from the location. The number of dishes for each location is 3, and ξ_{2ij} are the random effects coming from the dish of the location. That is, the random effects are initiated from the location and the error term for each dish. Therefore the total observed y_i is 27, i.e., $N = 27$ and $n_i = 400$ for the binomial distribution.

For model (3.5), β is the fixed effect, so $p = 1$, and the corresponding design matrix X is a 27×1 vector consisting of all 1's. The dimension of random effects is 2, so $r = 2$.

Further, $cov(\xi_i) = \sigma_i^2 I_{q_i} (i = 1, 2)$ and $cov(\xi) = D = \text{diag}(\sigma_1^2 I_1, \sigma_2^2 I_2)$, where I_1, I_2 are identity matrices of orders $q_1 \times q_1, q_2 \times q_2$, and in this data set, q_1=9 and q_2=27. We can write $U = [U_1, U_2]$, and the U_1 is a 27×9 block matrix and the U_2 is a 27×1 matrix. Also, we can write $\xi = [\xi_1', \xi_2']'$ with $Cov(\xi_i) = \sigma_i^2 I_{q_i}$ and $Cov(\xi_1, \xi_2) = 0$. Note that ξ_1 is a vector of 9×1 and ξ_2 is a vector of 27×1.

For the calculation of $\delta_{ICOMP}(i)$ in (2.8), we need to find out the Fish information matrix of model (3.4), and its derivation is illustrated in the Appendix.

3.2 The Estimation of the Parameters in the Logistic Linear Regression Model with Random Effects

For model (2.4), to estimate the maximum likelihood estimation in the normal variance components, we utilize the Schall's estimation method, and its iteration algorithm is described as follows:

1. Given estimates $\hat{\sigma}^2$ and $\hat{\sigma}_1^2, \cdots, \hat{\sigma}_r^2$ for β and $\xi_1, \cdots, \xi_r$ as least-squares solutions to the set of overdetermined linear equations

$$C \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} W^{-\frac{1}{2}}X & W^{-\frac{1}{2}}U \\ 0 & D^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} W^{-\frac{1}{2}}Z \\ 0 \end{bmatrix}, \tag{3.6}$$

where W and D are evaluated at the current estimates of variance components.

2. Let T be the inverse of the matrix formed by the last q rows and columns of $C'C$, partitioned conformbly with D as $\begin{bmatrix} T_{11} & \cdots & T_{1r} \\ \vdots & \ldots & \vdots \\ T_{r1} & \cdots & T_{rr} \end{bmatrix}$.

Given estimates $\hat{\beta}$ and $\hat{\xi}_1, \cdots, \hat{\xi}_r$, compute estimates $\hat{\sigma}^2$ and $\hat{\sigma}_1^2, \cdots, \hat{\sigma}_r^2$ for σ^2 and σ^2 and $\sigma_1^2, \cdots, \sigma_r^2$ as

$$\hat{\sigma}^2 = (Z - X\hat{\beta} - U\hat{\xi})'(Z - X\hat{\beta} - U\hat{\xi})/\{N - \sum_{i=1}^{r}(q_i - v_i)\}, \hat{\sigma}_i^2 = \frac{\hat{b}_i'\hat{b}_i}{q_i - v_i}, \tag{3.7}$$

where $v_i = \mathrm{tr}(T_{ii})/\sigma_i^2$ is evaluated at the current estimates of σ_i^2. Note that σ^2 is the residual variance of the model, and for the logistic linear model and without the extra-binomial variation, its estimate $\hat{\sigma}^2$ should be close to 1. Otherwise, this value will be larger than one. In addition, it is mentioned earlier that this algorithm yields maximum likelihood estimates of the parameters.

3.3 The Presentation of the Application Results

As described previously, the cancer cell data have been used in the paper of Schall (1991). The data were collected from the 9 locations, and each location contains 3 dishes. It could be observed that the numbers of cells surviving out of 400 placed are mostly around 110-145, some are about 170-180. Only for cases (locations) 3 and 8, the surviving cell numbers are very far from the other data. Intuitively, these two are may be influential.

In fact, the $\delta_{ICOMP(i)}$ values are calculated and plotted versus the index of case numbers in Figure 1. The values of $\delta_{ICOMP(i)}$ for cases 3 and 8 are quite outstanding, they are therefore justified as influential, and labeled by the symbol of "#".

For this cancer cell data set, we may want to see the estimates when the two influential cases are individually eliminated from the data set. Figure 2 features the plot of case-deleted parameter estimates β versus σ_1^2 in (a) and the plot of β versus σ_2^2 in (b). The dots for cases 3 and 8 are labeled. It is easy to see that these estimates are quite away from the others, indicating that when case 3 or 8 is eliminated from the data set, the parameter estimates are significantly changed. As a result, the corresponding $\delta_{ICOMP(i)}$ values are very large. The results therefore demonstrate that the diagnostic $\delta_{ICOMP(i)}$ can effectively detect the comprehensive change of the parameter estimates

when case is removed from the data set and further can successfully evaluate the magnitude of the influence of each case.

Table 1 features the actual parameter estimates for the full and case-deleted data sets. It is observed that the binomial-variation scale parameter estimates $\hat{\sigma}^2$ are all quite close to 1. From Table 1, it is observed that the parameter estimates significantly or moderately change when case 3 or 8 is deleted from the data set. Note that the change of the parameter estimates can only partially reflect the possibility of being an influential case because this change can not determine which case is influential. Different from the parameter estimates, the proposed diagnostic $\delta_{ICOMP(i)}$ is the one who aims to detect the change of the ICOMP caused by the removal of a case and who can summarize the complete change of the parameter estimates produced by the deletion of a case. Therefore, the magnitude of the proposed diagnostic $\delta_{ICOMP(i)}$ indicates whether one case is influential or not.

To further examine the effectiveness of the proposed diagnostic, we then artificially changed the values for cases 3 and 8. Values for cases 3 and 8 are originally 66, 75, 80 and 88, 76, 90 respectively. For Version 1, we changed them to 104, 105, 116 and 120, 110, 117. Then the $\delta_{ICOMP(i)}$ values are featured in Figure 3. Since the data for the two influential cases are changed to close to the most of the other data, the $\delta_{ICOMP(i)}$ values for cases 3 and 8 are not significantly larger any more, and they become the normal sizes compared to the other $\delta_{ICOMP(i)}$ values. The two $\delta_{ICOMP(i)}$ values are marked by "#" in Figure 3, however, they are not influential this time.

For Version 2, we changed them to 120, 111, 130 and 123, 134, 125 respectively, which are closer to the other data compared to those in Version 1. Then the $\delta_{ICOMP(i)}$ values are featured in Figure 4. The $\delta_{ICOMP(i)}$ values generally become smaller than those for Version 1. The two $\delta_{ICOMP(i)}$ values for cases 3 and 8 are specially marked by "#" in Figure 4 and it is easy to report that they are not influential either this time.

From the previous application results, it is exhibited that the proposed diagnostic $\delta_{ICOMP(i)}$ performs well in detecting an influential case in the GLMMs.

4 Concluding Remarks

We develop a diagnostic for detecting influential cases based on the ICOMP criteria in generalized linear mixed models. The ICOMP is a model selection criterion taking into account of both goodness-of-fit and model complexity. The diagnostic is defined for revealing influential cases as the discrepancy of the ICOMP criteria based on the full data set and a case-deleted data set.

Given the generalized linear mixed model (GLMM), it can be linearized using the Taylor expansion, and then the GLMMs are simplified to the linear mixed models, and correspondingly the focus on the response variable is shifted to the adjusted dependent variable. Based on the linearized mixed model, the diagnostic can be computed for detecting the influential case among the data where the GLMM can be utilized.

Since the covariance matrix of estimated parameters in the linear mixed modeling framework is unknown, the Fisher information matrix is employed to compute the ICOMP criterion. The Fisher information matrix is derived for the logistic linear mixed model in the Appendix. Since the covariance of the adjusted dependent variable is a function of the fixed effects in the mixed model, the derivation of the Fisher information matrix is not trivial, yet it is feasible.

To demonstrate the effectiveness of the proposed diagnostic, an application on a cancer cell data is carried out. The data are described by the logistic linear mixed model. The application results verifies that the proposed procedure effectively performs in detecting the influential case.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F.Csaki ed. *2nd International Symposium on Information Theory* 267–281. Akademia Kiado, Budapest.

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley and Sons.

Bozdogan, H., 1988. ICOMP: A new model selection criterion. *In: H.H. Bock(Ed.), Classification and Related Methods of Data Analysis, North-Holland, Amsterdam,* 599–608.

Bozdogan, H., 1990. On the information-based measure of covariance complexity and its application on the evaluation of multivariate linear models. *Communications in Statistics, Theory Methods* **19(1)**, 221–278.

Bozdogan, H., 1993. Choosing the number of component clusters in the mixture model using a new informational complexity criterion of the inverse Fisher information matrix. *In: O.Opitz, B. Lausen, and R.Klar (Eds.), Studies in Classification, Data Analysis, and Knowledge Organization,* Springer, Heidelberg, pp. 40–54.

Bozdogan, H., 1994. Mixture-model cluster analysis using a new informational complexity and model selection criteria. *In: H. Bozdogan (Ed.), Multivariate Statistical Modeling,* Vol 2, Proc. 1st US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, Kluwer Academic Publishers, the Netherlands, Dordrecht, pp. 69–113.

Bozdogan, H., Bearse, P., 2003. Information complexity criteria for detecting influential observations in dynamic multivariate linear models using the genetic algorithm. *Journal of Statistical Planning and Inference* **114**, 31–44.

Cavanaugh, J.E., Johnson, W.O. 1999. Assessing the predictive influence of cases in a state-space process. *Biometrika* **86**, 183–190.

Cavanaugh, J.E., Oleson, J.J., 2001. A diagnostic for assessing the influence of cases on the prediction of missing data. *Journal of the Royal Statistical Society, Series D* **50**, 427–440.

Cavanaugh, J.E., Shang, J., 2005. A diagnostic for assessing the influence of cases on the prediction of random effects in a mixed model. *Journal of Data Science* **3**, 137–151.

Christensen, R., Pearson, L.M., Johnson, W., 1992. Case-deletion diagnostics for mixed models. *Technometrics* **34**, 38–45.

Cook, R.D., 1977. Detection of influential observations in linear regression. *Technometrics* **19**, 15–18.

Cook, R.D., 1979. Influential observations in linear regression. *Journal of the American Statistical Association* **74**, 169–174.

Cook, R.D., Weisberg, S., 1982. *Residuals and Influence in Regression.* London: Chapman and Hall.

Johnson, W., 1985. Influence measures for logistic regression: Another point of view. *Biometrika* **72**, 59–65.

Johnson, W., Geisser, S., 1983. A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association* **78**, 137–144.

McCullagh, P., Nelder, J. A., 1989. *Generalized Linear Models.* Second ed. London: Chapman and Hall.

Schall, R., 1991. Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727

Shang, J., 2008. A diagnostic of influential cases based on the Information Complexity Criterion in mixed models. Shang, J. *Proceedings of the International Conference on Information Theory and Statistical Learning (ITSL)* 36–41.

van Emden, M., 1971. An analysis of complexity. Mathematisch Centrum Tracts 35, Mathematisch Centrum, Amsterdam.

Table 1: Parameter Estimates for Cancer Cell Data

Estimates	β	σ^2	σ_1^2	σ_2^2
Full data	-0.7579	1.1294	0.1958	0.0057
Case 1 deleted	-0.8437	1.1165	0.1516	0.0038
Case 2 deleted	-0.7330	1.0416	0.2151	0.0109
Case 3 deleted	-0.6648	1.0543	0.1454	0.0093
Case 4 deleted	-0.7498	1.1792	0.2230	0.0062
Case 5 deleted	-0.7773	1.0097	0.2156	0.0038
Case 6 deleted	-0.7495	1.1211	0.2228	0.0076
Case 7 deleted	-0.8316	1.0950	0.1696	0.0069
Case 8 deleted	-0.6882	1.1500	0.1771	0.0062
Case 9 deleted	-0.7570	1.1347	0.2193	0.0081

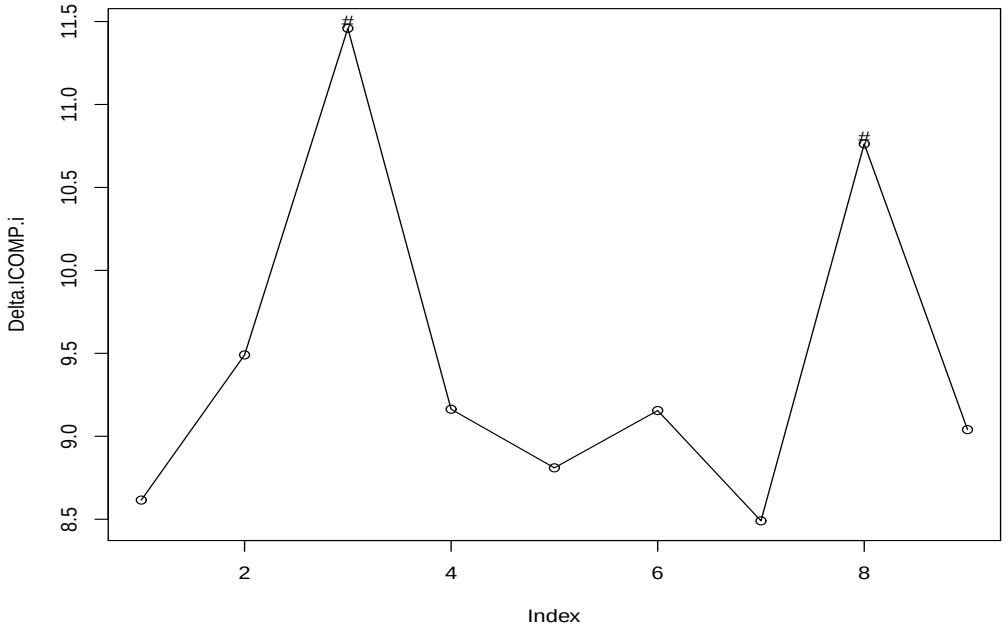

Figure 1: $\delta_{ICOMP(i)}$ vs. Case Index i for Cancer Cell Data

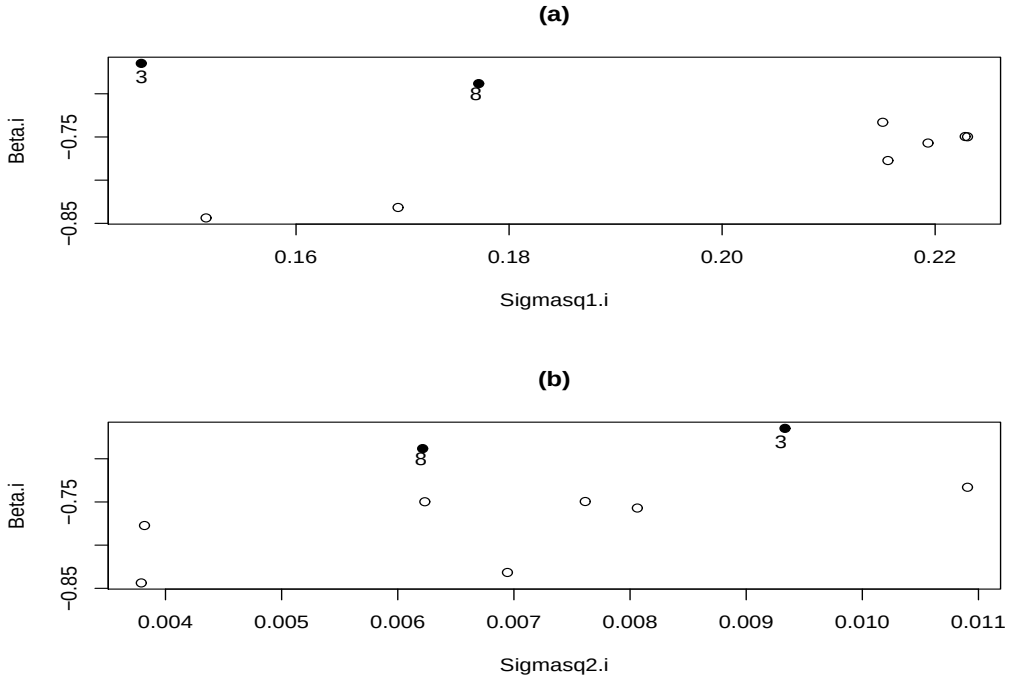

Figure 2: Parameter Estimates for Cancer Cell Data

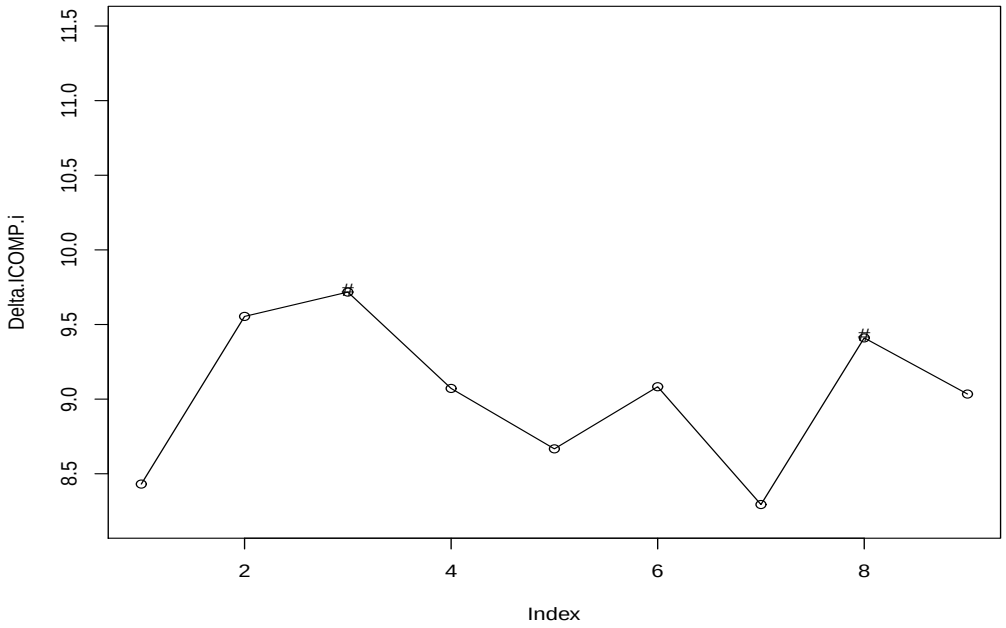

Figure 3: $\delta_{ICOMP(i)}$ vs. Case Index i for Changed Cancer Cell Data Version 1

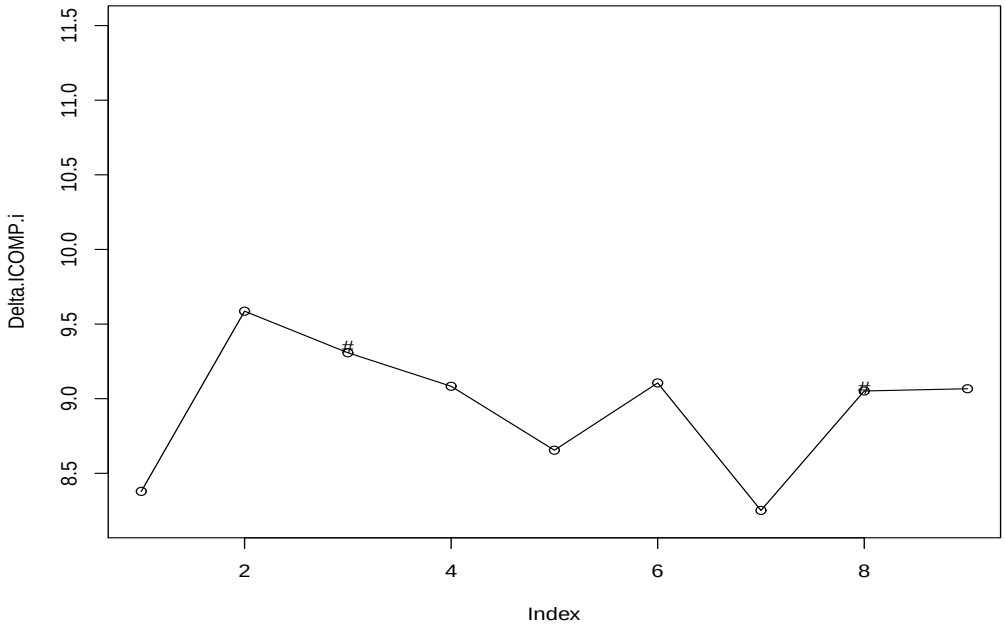

Figure 4: $\delta_{ICOMP(i)}$ vs. Case Index i for Changed Cell Data Version 2