Journal of the Korean Statistical Society (2006), 35: 2, pp 213–224

INFLUENCE ANALYSIS FOR GENERALIZED ESTIMATING EQUATIONS†

KANG-MO JUNG[1]

ABSTRACT

We investigate the influence of subjects or observations on regression coefficients of generalized estimating equations using the influence function and the derivative influence measures. The influence function for regression coefficients is derived and its sample versions are used for influence analysis. The derivative influence measures under certain perturbation schemes are derived. It can be seen that the influence function method and the derivative influence measures yield the same influence information. An illustrative example in longitudinal data analysis is given and we compare the results provided by the influence function method and the derivative influence measures.

AMS 2000 subject classifications. Primary 62J20, 62J12; Secondary 62P10.
Keywords. Derivative influence, diagnostics, generalized estimating equation, influence function, longitudinal data analysis.

1. INTRODUCTION

In longitudinal studies measurements of the same subject are taken repeatedly through time. Longitudinal data has been applied to a wide range of fields, medicine, public health, biology and more. The repeated measurement for each subject requires that the within-subject correlation should be taken into account. Under the condition of non-repeated measurements for each subject it reduces to generalized linear models (McCullagh and Nelder, 1989). There are several methods to extend generalized linear models with the consideration of correlations within subjects: marginal mean models, random-effect models and transition

Received February, 2006; accepted June, 2006.
†This work was supported by Korea Research Foundation Grant (KRF-2005-202-C00076).
[1]Department of Informatics and Statistics, Kunsan National University, Kunsan 573-701, Korea (e-mail: kmjung@kunsan.ac.kr)

models (Diggle *et al.*, 2002). In this article we focus on the first approach which models the average population response with respect to changes in covariates.

Liang and Zeger (1986) suggested the generalized estimating equations (GEE) approach for non-normal response in longitudinal data analysis, which is an extension of quasi-likelihood method to longitudinal data analysis. Quasi-likelihood method requires only second moments assumptions about the distribution of the response variable. The GEE approach is feasible in many situations where maximum likelihood approaches are not, since the full multivariate distribution of the response vector is not necessary (Davis, 2002). The correlation among observations of the same subject is represented by choosing a model, such as the correlation structures corresponding to the AR(1), one-dependent, exchangeable or other situations. It is referred to as the working correlation matrix.

The GEE estimator of the regression coefficients can be obtained by iterative weighted least squares by regressing the working response vector on the covariates. It is well known that least squares estimators are very sensitive to the presence of unusual data. So is the GEE estimator. We have to pay attention to avoid the misleading statistical conclusion due to few outliers. Despite of the popular use of the GEE approach, however, there exist few diagnostics. Preisser and Qaqish (1996) introduced deletion diagnostics which account for the leverage and residuals in a set of subjects or observations to determine their influence on regression parameter estimates and fitted values. They extended the deletion diagnostics for generalized linear models by Christensen *et al.* (1992).

The influence function (Hampel, 1974) for a parameter at a point measures the effect of an infinitesimal contamination at that point on the estimator of the parameter. Hence the influence function can serve as a diagnostic method of detecting influential observations of estimators. One of robustness properties is boundedness of the influence function of an estimator. The results of this article will be used for suggesting a new robust estimator with bounded influence function for GEE. The derivative influence (De Gruttola *et al.*, 1987) measures the differential change in an estimated parameter resulting from a slight perturbation in the weight assigned to a given subject or observation.

In this work we investigate the influence of subjects or observations on the GEE estimators using the influence function and the derivative influence. In Section 2 we review GEE and introduce some notations. In Section 3 we derive the influence function of the GEE estimator and consider three sample versions of the influence function that be used for investigating the influence of subjects. The derivative influence measures for the GEE estimator are derived in Section 4

under perturbation schemes in which a weight is put on the weight of the iterative weighted least squares estimator. Some relationships between the influence function and the derivative influence measures are found. In Section 5 an illustrative example is given and we compare the results provided by the influence function and the derivative influence.

2. GENERALIZED ESTIMATING EQUATIONS

For $i = 1, \ldots, K$, let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ be a n_i-vector of response variables, and $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in_i})'$ a $n_i \times p$ matrix of covariate values. In longitudinal data analysis K denotes the number of subjects and n_i is the number of repeated measurements for the i^{th} subject, that is, the number of observations. In marginal models, the marginal expectation $\mu_{ij} = E(Y_{ij})$ is modeled as a function of covariates $\mathbf{X}_{ij}$. The GEE approach is a marginal model that was proposed by Liang and Zeger (1986). The GEE approach is an extension of quasi-likelihood method to longitudinal data analysis. The forms of the first two moments for the marginal distribution of Y_{ij} are given by

$$E(Y_{ij}) = \mu_{ij}, \quad g(\mu_{ij}) = \eta_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta}, \quad \text{Var}(Y_{ij}) = V(\mu_{ij})\phi,$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, $g(\cdot)$ is the link function, $V(\cdot)$ is the variance function and ϕ is the scale parameter. Let denote $\mathbf{R}_i(\boldsymbol{\alpha})$ the $n_i \times n_i$ working correlation matrix for each $\mathbf{Y}_i$. The elements of $\mathbf{R}_i(\boldsymbol{\alpha})$ are the known, hypothesize, or estimated correlation between Y_{ij} and $Y_{ij'}$ for unknown parameter $\boldsymbol{\alpha}$. Estimates of $\boldsymbol{\beta}$ are obtained by solving the GEE

$$\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' [\mathbf{V}_i(\hat{\boldsymbol{\alpha}})]^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}_p, \tag{2.1}$$

where $\mathbf{V}_i(\boldsymbol{\alpha}) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, $\mathbf{A}_i$ is the $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ as the j^{th} diagonal element, and $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. See Davis (2002) for details.

A solution of (2.1) can be obtained by alternating between estimation of $\phi, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using method of moment estimators for ϕ and $\boldsymbol{\alpha}$. Let define $N = \sum_{i=1}^{K} n_i$, $\mathbf{Y} = (\mathbf{Y}'_1, \ldots, \mathbf{Y}'_K)'$, the $N \times p$ matrix $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_K)'$ assumed to be of full column rank, and $\mathbf{D} = \partial \boldsymbol{\eta} / \partial \boldsymbol{\mu}$, an $N \times N$ diagonal matrix with nonzero elements $d_{ij} = \partial \eta_{ij} / \partial \mu_{ij}$. Estimation of $\boldsymbol{\beta}$ is done with iteratively reweighted least squares by regressing the working response vector $\mathbf{Z} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{D}(\mathbf{Y} - \hat{\boldsymbol{\mu}})$ and $\mathbf{X}$ with block

diagonal weight matrix $\mathbf{W}$ whose i^{th} block, corresponding to the i^{th} subject, is the $n_i \times n_i$ matrix

$$\mathbf{W}_i = \hat{\phi}\mathbf{D}_i^{-1}\mathbf{V}_i^{-1}(\hat{\alpha})\mathbf{D}_i^{-1}, \quad \mathbf{D}_i = \text{diag}(d_{i1}, \ldots, d_{in_i}).$$

And the $n_i \times 1$ vector $\mathbf{Z}_i$ corresponds to the i^{th} subject of $\mathbf{Z}$. A current estimate of $\boldsymbol{\beta}$ is updated by

$$\hat{\boldsymbol{\beta}}_{new} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Z}, \tag{2.2}$$

evaluating the right-hand side at the current estimate. See Preisser and Qaqish (1996) for details.

3. INFLUENCE FUNCTION

The GEE estimator $\hat{\boldsymbol{\beta}}$ in (2.2) is a least squares estimator. It is known that the least squares estimator is very sensitive to outliers and it is caused by a unbounded influence function of the estimate. In this section we derive the influence function for the GEE estimator $\hat{\boldsymbol{\beta}}$ of (2.2) and three sample versions of the influence function that are used for investigating the influence of subjects on $\hat{\boldsymbol{\beta}}$. In this article we focus the influence analysis for regression estimator. Assume that the working correlation matrix and the nuisance parameters are fixed.

Let $\theta = \theta(F)$ be a parameter of interest which is a functional of the distribution F. Assume that the distribution is perturbed as $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{\mathbf{u}}$ for $0 \le \varepsilon \le 1$, where $\delta_{\mathbf{u}}$ denotes the distribution having unit mass at the point $\mathbf{u}$. The perturbation of θ at $\mathbf{u}$ is $\theta(F_\varepsilon)$. The influence function for θ at $\mathbf{u}$ (Hampel, 1974) is defined by

$$IF(\theta) = \lim_{\varepsilon \to 0} \frac{\theta(F_\varepsilon) - \theta(F)}{\varepsilon}. \tag{3.1}$$

The influence function for a parameter at $\mathbf{u}$ measures the effect of an infinitesimal contamination at $\mathbf{u}$ on the estimator of the parameter.

Let the $(p+1) \times n_i$ generic matrix $(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}})'$ have a joint distribution function with

$$E[\begin{pmatrix} \tilde{\mathbf{X}}' \\ \tilde{\mathbf{Z}}' \end{pmatrix} \tilde{\mathbf{W}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}})] = \begin{pmatrix} \boldsymbol{\Sigma}_{11}(F) & \boldsymbol{\Sigma}_{12}(F) \\ \boldsymbol{\Sigma}_{21}(F) & \boldsymbol{\Sigma}_{22}(F) \end{pmatrix},$$

where $\tilde{\mathbf{W}}$ is the $n_i \times n_i$ generic matrix corresponding to $\mathbf{W}_i$. The functional corresponding to the weighted least squares estimator (2.2) of $\boldsymbol{\beta}$ can be rewritten as

$$\boldsymbol{\beta}(F) = \boldsymbol{\Sigma}_{11}^{-1}(F)\boldsymbol{\Sigma}_{12}(F),$$

assuming, of course, $\boldsymbol{\Sigma}_{11}(F)$ is nonsingular. The above equation reduces to the functional for linear regression if $n_i = 1$ for all i. See Eq. (3.3.2) of Cook and Weisberg (1982).

The p-dimensional influence function of $\boldsymbol{\beta}$ as a function of $(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}})$ is defined pointwise. The explicit formula is obtained by perturbing the distribution function F. Under the perturbed distribution function F_ε the functional of $\boldsymbol{\Sigma}_{11}$ becomes

$$\boldsymbol{\Sigma}_{11}(F_\varepsilon) = \boldsymbol{\Sigma}_{11}(F) + (\tilde{\mathbf{X}}'\tilde{\mathbf{W}}\tilde{\mathbf{X}} - \boldsymbol{\Sigma}_{11}(F))\varepsilon + O(\varepsilon^2).$$

The fact that $\boldsymbol{\Sigma}_{11}(F_\varepsilon)\boldsymbol{\Sigma}_{11}^{-1}(F_\varepsilon) = \mathbf{I}$ yields the expansion

$$\boldsymbol{\Sigma}_{11}^{-1}(F_\varepsilon) = \boldsymbol{\Sigma}_{11}^{-1}(F) - \left\{\boldsymbol{\Sigma}_{11}^{-1}(F)\tilde{\mathbf{X}}'\tilde{\mathbf{W}}\tilde{\mathbf{X}}\boldsymbol{\Sigma}_{11}^{-1}(F) - \boldsymbol{\Sigma}_{11}^{-1}(F)\right\}\varepsilon + O(\varepsilon^2).$$

Under the perturbed distribution we get the functional

$$\begin{aligned}\boldsymbol{\beta}(F_\varepsilon) &= \boldsymbol{\Sigma}_{11}^{-1}(F_\varepsilon)\boldsymbol{\Sigma}_{12}(F_\varepsilon) \\ &= \boldsymbol{\beta}(F) + [\boldsymbol{\Sigma}_{11}^{-1}(F)\tilde{\mathbf{X}}'\tilde{\mathbf{W}}\tilde{\mathbf{D}}\{\tilde{\mathbf{Y}} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\beta}(F))\}]\varepsilon + O(\varepsilon^2),\end{aligned}$$

from the equality $\tilde{\mathbf{Z}} = \tilde{\mathbf{X}}\boldsymbol{\beta}(F) + \tilde{\mathbf{D}}(\tilde{\mathbf{Y}} - \tilde{\boldsymbol{\mu}}(F))$. Thus the influence function of the GEE estimator (2.2) at $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})'$ is

$$IF(\boldsymbol{\beta}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \boldsymbol{\Sigma}_{11}^{-1}(F)\tilde{\mathbf{X}}'\tilde{\mathbf{W}}\tilde{\mathbf{D}}\{\tilde{\mathbf{Y}} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\beta}(F))\}. \tag{3.2}$$

Next we consider three sample versions as in Critchley (1985): the empirical influence function (EIF), the sample influence function (SIF) and the deleted empirical influence function (DIF). A large absolute value of each sample version indicates that the corresponding subject is influential.

3.1. Empirical Influence Function

The EIF for $\boldsymbol{\beta}$ is obtained by substituting the empirical distribution function $\hat{F}$ for F in (3.2). Under the distribution $\hat{F}$ the regression coefficients $\boldsymbol{\beta}$ becomes $\hat{\boldsymbol{\beta}}$ in (2.2). And also the functional $\boldsymbol{\Sigma}_{11}(\hat{F})$ becomes the average of $\mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i$ for $i = 1, \ldots, K$. It is easily seen that the EIF for $\boldsymbol{\beta}$ at $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})'$ is

$$EIF(\boldsymbol{\beta}, \tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = K(\sum_{i=1}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{W}}\tilde{\mathbf{D}}\{\tilde{\mathbf{Y}} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\beta}(\hat{F}))\}.$$

Thus it leads to

$$EIF_i = EIF(\boldsymbol{\beta}, \mathbf{X}_i, \mathbf{Y}_i) = K(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'\mathbf{W}_i\mathbf{D}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i), \tag{3.3}$$

where $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$.

The estimating equation (2.2) has similar to that of $M-$estimate. Under the regularity condition the covariance matrix of $\hat{\boldsymbol{\beta}}$ converges asymptotically to

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{M}_0^{-1}\mathbf{M}_1\mathbf{M}_0^{-1},$$

where

$$\mathbf{M}_0 = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' \mathbf{V}_i^{-1}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)$$

and

$$\mathbf{M}_1 = \lim_{K\to\infty} \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'\mathbf{V}_i^{-1}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right).$$

Similar to Eq.(6.3.6) of Hampel *et al.* (1986) we may conjecture that the influence function of $\boldsymbol{\beta}$ has the form as

$$\mathbf{M}_0^{-1}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)' \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i).$$

Some algebra clearly forces the same result as (3.3).

3.2. Sample Influence Function

The SIF can be obtained by setting $F = \hat{F}$ and taking $\varepsilon = -1/(K-1)$ in the definition of the influence function (3.1) instead of taking a limit. Then the SIF for a parameter $\boldsymbol{\beta}$ at $(\mathbf{X}_i, \mathbf{Y}_i)'$ can be rewritten as $(K-1)[\boldsymbol{\beta}(\hat{F}) - \boldsymbol{\beta}(\hat{F}_{(i)})]$, where $\hat{F}_{(i)} = (1+(K-1)^{-1})\hat{F} - (K-1)^{-1}\delta_{(\mathbf{X}_i,\mathbf{Y}_i)}$ is the deleted version of $\hat{F}$ with the ith subject deleted. Hereafter we denote by the subscript (i) the estimator based on the reduced data set with the deletion of the i^{th} subject.

The weighted covariance matrix $\boldsymbol{\Sigma}_{11}$ evaluated at $\hat{F}_{(i)}$ is given by

$$\boldsymbol{\Sigma}_{11}(\hat{F}_{(i)}) = \frac{1}{K-1}(\mathbf{X}'\mathbf{W}\mathbf{X} - \mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i).$$

Thus we get

$$\boldsymbol{\Sigma}_{11}^{-1}(\hat{F}_{(i)}) = (K-1)\left\{(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{X}_i(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right\}$$

and

$$\boldsymbol{\Sigma}_{12}(\hat{F}_{(i)}) = \frac{1}{K-1}(\mathbf{X}'\mathbf{W}\mathbf{Z} - \mathbf{X}_i'\mathbf{W}_i\mathbf{Z}_i),$$

where $\mathbf{Q}_i = \mathbf{X}_i(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'$. Thus the functional of $\boldsymbol{\beta}$ under the distribution function $\hat{F}_{(i)}$ can be written by

$$\boldsymbol{\beta}(\hat{F}_{(i)}) = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{D}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i).$$

Therefore the SIF of $\boldsymbol{\beta}$ at $(\mathbf{X}_i, \mathbf{Y}_i)'$ becomes

$$SIF_i = SIF(\boldsymbol{\beta}, \mathbf{X}_i, \mathbf{Y}_i) = (K-1)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{D}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i), \quad (3.4)$$

which is proportional to the change in the estimate of $\boldsymbol{\beta}$ when the ith subject is deleted. This result is equivalent to Equation (5) of Preisser and Qaqish (1996).

3.3. Deleted Empirical Influence Function

The DIF is obtained by replacing F with $\hat{F}_{(i)}$ in (3.2) and it measures the effect on the estimator with the deletion of the i^{th} subject. That is, we have

$$DIF_i = DIF(\boldsymbol{\beta}, \mathbf{X}_i, \mathbf{Y}_i) = \boldsymbol{\Sigma}_{11}^{-1}(\hat{F}_{(i)})\mathbf{X}_i'\mathbf{W}_i\mathbf{D}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i(\hat{F}_{(i)}).$$

Using $\boldsymbol{\Sigma}_{11}^{-1}(\hat{F}_{(i)})$ and $\boldsymbol{\beta}(\hat{F}_{(i)})$ derived in the previous subsection, the result is

$$DIF_i = (K-1)(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_i'(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{W}_i^{-1}(\mathbf{W}_i^{-1} - \mathbf{Q}_i)^{-1}\mathbf{D}_i(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i). \quad (3.5)$$

Since the influence measures Equations (3.3) to (3.5) do not have simple forms, their distribution properties can not be obtained easily. Thus we consider a graphical approach to detect influential observations.

The EIF measures the instantaneous rate of change in the estimator as a single subject is added to the data (Cook and Weisberg, 1982). The interpretation of DIF is analogous to that for EIF, but the DIF is computed based on the data without a subject. The SIF is a compromise between EIF and DIF. Three sample versions of the influence function for $\boldsymbol{\beta}$ have similar values and as the size of subjects tends to be large, their values give the same degree of influence. Hence the use of only one version is enough to investigate the influence of subjects on the regression parameter $\boldsymbol{\beta}$.

4. DERIVATIVE INFLUENCE MEASURE

Let $\hat{\theta}$ be an estimator of the parameter θ. The derivative influence is derived by considering a perturbation scheme represented by t in which the distribution for only one subject is perturbed. When we denote by $\hat{\theta}(t)$ the estimator $\hat{\theta}$ under

this perturbation scheme, we assume that $\hat{\theta} = \hat{\theta}(t_0)$ for some t_0 called the null point. The derivative influence (De Gruttola *et al.*, 1987) is defined by

$$\nabla\hat{\theta}(t_0) = \left.\frac{d\hat{\theta}(t)}{dt}\right|_{t=t_0}.$$

First we consider the perturbation t such that the only l^{th} response vector $\mathbf{Y}_l$ is drawn from a perturbed distribution with the mean vector $\boldsymbol{\mu}_l$ and the covariance matrix $\boldsymbol{\Sigma}_l/t$. Then under this perturbation scheme, the GEE estimator of $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}}(t) = (\sum_{i\neq l}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i + t\mathbf{X}_l'\mathbf{W}_l\mathbf{X}_l)^{-1}(\sum_{i\neq l}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{Z}_i + t\mathbf{X}_l'\mathbf{W}_l\mathbf{Z}_l). \tag{4.1}$$

When $t = 1$ the perturbed estimator $\hat{\boldsymbol{\beta}}(1)$ reduces to the unperturbed estimator $\hat{\boldsymbol{\beta}}$. Thus $t = 1$ is the null point. For $t = 0$ the estimator $\hat{\boldsymbol{\beta}}(0)$ becomes the estimator based on the data without the l^{th} subject. That is, from (3.4) $\hat{\boldsymbol{\beta}}(0) = \hat{\boldsymbol{\beta}} - SIF_l/(K-1)$.

Since $\mathbf{A}^{-1}(t) = -\mathbf{A}^{-1}(t)[d\mathbf{A}(t)/dt]\mathbf{A}^{-1}(t)$, we have

$$\frac{d}{dt}\hat{\boldsymbol{\beta}}(t) = (\sum_{i\neq l}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i + t\mathbf{X}_l'\mathbf{W}_l\mathbf{X}_l)^{-1}\mathbf{X}_l'\mathbf{W}_l(\mathbf{Z}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}(t)). \tag{4.2}$$

The cases $t = 0$ and $t = 1$ of the above derivative will be the ones of interest to us. At $t = 1$ the perturbed estimator $\hat{\boldsymbol{\beta}}(t)$ reduces to the unperturbed case. Thus the derivative at $t = 1$ measures the local change in the estimator caused by the i^{th} subject. The larger value of $\nabla\hat{\boldsymbol{\beta}}(1)$ indicates that the slope of $\hat{\boldsymbol{\beta}}(t)$ at the null point changes rapidly. In the other case $t = 0$, we obtain $\hat{\boldsymbol{\beta}}(0)$ which is the GEE estimator based on the remaining data with deletion of the l^{th} subject. Thus the derivative $\nabla\hat{\boldsymbol{\beta}}(0)$ at $t = 0$ measures the rate of change in the estimator when the l^{th} subject is omitted.

For $t = 0$ and $t = 1$ we have

$$\nabla\hat{\boldsymbol{\beta}}(1) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_l'\mathbf{W}_l(\mathbf{Z}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}) \tag{4.3}$$

and

$$\nabla\hat{\boldsymbol{\beta}}(0) = (\sum_{i\neq l} \mathbf{X}_i'\mathbf{W}_i\mathbf{X}_i)^{-1}\mathbf{X}_l'\mathbf{W}_l(\mathbf{Z}_l - \mathbf{X}_l\hat{\boldsymbol{\beta}}(0)). \tag{4.4}$$

Comparing (4.3) and (3.3) yields the identity $\nabla\hat{\boldsymbol{\beta}}(1) = EIF_l/K$. It implies that $\nabla\hat{\boldsymbol{\beta}}(1)$ and EIF_l have the same influence information. Also from (4.4) and (3.5)

it is easy to verify $\nabla\hat{\boldsymbol{\beta}}(0) = DIF_l/(K-1)$. Thus the influence measures using $\nabla\hat{\boldsymbol{\beta}}(0)$ and DIF_l are equivalent. Since $\nabla\hat{\boldsymbol{\beta}}(t)$ is differentiable over the range of t, the average of $\nabla\hat{\boldsymbol{\beta}}(t)$ can be written as

$$\int_0^1 \nabla\hat{\boldsymbol{\beta}}(t)dt = \hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(0) = SIF_l/(K-1). \tag{4.5}$$

The average of $\nabla\hat{\boldsymbol{\beta}}(t)$ gives the same influence information as the SIF. Thus to investigate the influence of observations on the GEE estimator we may use the derivative influence measures or the sample versions of the influence function.

Second we consider the perturbation t of the only j^{th} observation in the l^{th} subject such that the response variable Y_{lj} is drawn from a perturbed distribution with the variance $\Sigma_{l,jj}/t$, where $\Sigma_{l,jj}$ is the $(j,j)^{th}$ element of the covariance matrix $\boldsymbol{\Sigma}_l$. That is, $\mathbf{Y}_l$ is followed as the mean vector $\boldsymbol{\mu}_l$ and the covariance vector $\boldsymbol{\Sigma}_l\mathbf{T}_l^{-1}$, where $\mathbf{T}_l$ is the $n_l \times n_l$ diagonal matrix with all elements equal to one except the j^{th} element t. Then under this perturbation scheme the GEE estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}}(t) = (\sum_{i=1}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{T}_i\mathbf{X}_i)^{-1} \sum_{i=1}^{K} \mathbf{X}_i'\mathbf{W}_i\mathbf{T}_i\mathbf{Z}_i,$$

where $\mathbf{T}_i = \mathbf{I}$ if $i \neq l$. It is clear that $t = 1$ is the null point. When $t = 0$ the estimator $\hat{\boldsymbol{\beta}}(0)$ becomes the estimator when the j^{th} observation of the l^{th} subject is deleted. We get

$$\hat{\beta}(0) = (\mathbf{X}'\mathbf{W}\mathbf{X} - \mathbf{X}'\mathbf{W}_{lj}\mathbf{X}_{lj}^r)^{-1}(\mathbf{X}'\mathbf{W}\mathbf{Z} - \mathbf{X}_l'\mathbf{W}_{lj}Z_{lj}),$$

where $\mathbf{W}_{lj}$ are the j^{th} column vector of $\mathbf{W}_l$ and Z_{lj} is the j^{th} component of the vector $\mathbf{Z}_l$. Here $\mathbf{X}_{lj}^r$ is the j^{th} row vector of $\mathbf{X}_l$.

Similar to (4.2) we obtain

$$\frac{d}{dt}\hat{\boldsymbol{\beta}}(t) = (\sum_i \mathbf{X}_i'\mathbf{W}_i\mathbf{T}_i\mathbf{X}_i)^{-1}\mathbf{X}_l'\mathbf{W}_l\mathbf{1}_{lj}\mathbf{1}_{lj}'(\mathbf{Z}_l - \mathbf{X}_l'\hat{\boldsymbol{\beta}}(t)),$$

where $\mathbf{1}_{ij}$ is the $n_i \times 1$ vector with its j^{th} element equal to one and the others being zero. From the definition of the derivative influence it follows that

$$\nabla\hat{\boldsymbol{\beta}}(1) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}_l'\mathbf{W}_{lj}(Z_{lj} - \mathbf{X}_{lj}\hat{\boldsymbol{\beta}}). \tag{4.6}$$

Moreover, we have

$$\nabla\hat{\boldsymbol{\beta}}(0) = (\mathbf{X}'\mathbf{W}\mathbf{X} - \mathbf{X}_l'\mathbf{W}_{lj}\mathbf{X}_{lj}^r)^{-1}\mathbf{X}_l'\mathbf{W}_{lj}(Z_{lj} - \mathbf{X}_{lj}\hat{\boldsymbol{\beta}}(0)). \tag{4.7}$$

As the influence information on the j^{th} observation of the l^{th} subject (4.6) and (4.7) provide EIF_{lj} and DIF_{lj}, respectively, similar to (4.3) and (4.4). Furthermore, the average of the derivative influence is $\hat{\boldsymbol{\beta}}(1) - \hat{\boldsymbol{\beta}}(0)$.

5. NUMERICAL EXAMPLE

The influence function method and the derivative influence measure are applied to the Epileptic data (Thall and Vail, 1990). It is from a panel study in which four successive two-week counts of seizures were recorded for each epileptic patient. The covariates are the Progabide treatment indicator (X_1), the followup indicator (X_2) and an interaction of these covariates (X_3). And the response is the seizures Y. The objective of the study was to determine whether progabide reduces the rate of seizures in subjects. Fifty nine subjects and five observations in a cluster are obtained. Diggle *et al.* (2002) concluded that the patient number 207 had unusual seizure counts.

To estimate the overall treatment effect we used the following model:

$$logE(Y_{ij}) = logt_{ij} + \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3}, \quad j = 0, 1, \ldots, 4, \quad i = 1, \ldots, 59.$$

Here $t_{ij} = 8$ if $j = 0$ and $t_{ij} = 2$ if $j = 1, \ldots, 4$. The model was fitted using GEE assuming exchangeable working correlation and we obtained the regression coefficients $\hat{\beta}_0 = 1.35$, $\hat{\beta}_1 = 0.028$, $\hat{\beta}_2 = 0.11$, $\hat{\beta}_3 = -0.10$. On the remaining data with deleting patient number 207 the regression coefficients become $\hat{\beta}_0 = 1.35$, $\hat{\beta}_1 = -0.11$, $\hat{\beta}_2 = 0.11$, $\hat{\beta}_3 = -0.30$. The coefficient of interest β_3 represents the difference in the logarithm of the post- to pre-treatment ratio between the progabide and placebo groups (Diggle *et al.*, 2002). We observed that the maximum of absolute values of (4.6) is 0.126. It implies that no single observation had a large influence. So we present influence analysis based on the subject only.

Three sample versions of the influence function for the regression coefficient are given in Figure 5.1. In this plot, y-axis indicates EIF, SIF, DIF for β_3. From Figure 5.1, we may conclude that subject 49 (patient number 207) is the most influential on β_3 and that three sample versions yield the same influence information. Furthermore we observed that $|EIF| < |SIF| < |DIF|$ as described in Section 4.

We conducted the influence analysis using three derivative influence measures $\nabla\hat{\beta}(1)$, $\hat{\beta}(1) - \hat{\beta}(0)$ and $\nabla\hat{\beta}(0)$ using (4.3) to (4.5). We observed that $\hat{\beta}(1) - \hat{\beta}(0)$ is the average of $\nabla\hat{\beta}(1)$ and $\nabla\hat{\beta}(0)$ as explained in Section 4. And we checked the

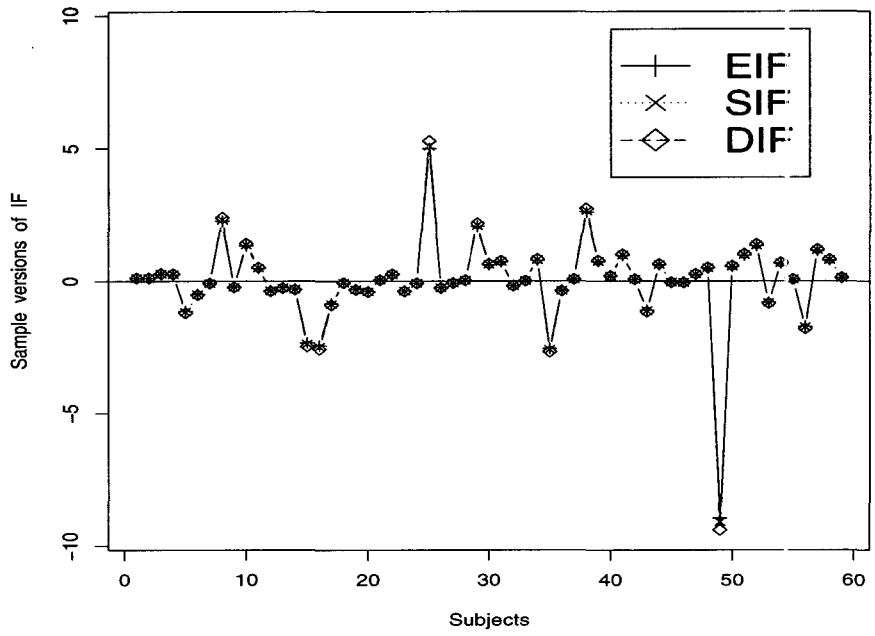

FIGURE 5.1 *Sample versions of the influence function for* β_3.

influence measures EIF, SIF and DIF are equivalent to $\nabla\hat{\beta}(1)$, $\hat{\beta}(1) - \hat{\beta}(0)$ and $\nabla\hat{\beta}(0)$, respectively, as expected. These are also confirmed by observing that the sample correlation coefficients between corresponding measures are all one. It implies that the influence function method and the derivative influence measures give the identical results.

6. CONCLUDING REMARKS

The influence function method and the derivative influence were applied to the investigation of the influence of observations on GEE estimates which is a popular estimating method in longitudinal data analysis. The results based on the former are in parallel with the latter. The derivative influence measure is more flexible than the influence function method because the derivative influence measure allows us to take diverse perturbation schemes according to the purpose of investigation. However the influence function method provides a robust GEE estimator with the bounded influence function which is a good robustness

property.

REFERENCES

CHRISTENSEN, R., PEARSON, L. M. AND JOHNSON, W. (1992). "Case-deletion diagnostics for mixed models", *Technometrics*, **34**, 38–45.

COOK, R. D. AND WEISBERG, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, London.

CRITCHLEY, F. (1985). "Influence in principal components analysis", *Biometrika*, **72**, 627–636.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. AND ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, Oxford.

DAVIS, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*, Springer-Verlag, New York.

DE GRUTTOLA, V., WARE, J. H. AND LOUIS, T. A. (1987). "Influence analysis of generalized least squares estimators", *Journal of the American Statistical Association*, **82**, 911–917.

HAMPEL, F. R. (1974). "The influence curve and its role in robust estimation", *Journal of the American Statistical Association*, **69**, 383–393.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. AND STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

LIANG, K.-Y. AND ZEGER, S. L. (1986). "Longitudinal data analysis using generalized linear models", *Biometrika*, **73**, 13–22.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized linear models*, Chapman & Hall, New York.

PREISSER, J. S. AND QAQISH, B. F. (1996). "Deletion diagnostics for generalized estimating equations", *Biometrika*, **83**, 551–562.

THALL, P. F. AND VAIL, S. C. (1990). "Some covariance models for longitudinal count data with overdispersion", *Biometrics*, **46**, 657–671.