University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2011 *Paper* 21

Deletion Diagnostics for Alternating Logistic Regressions

John S. Preisser* Kunthel By†
Jamie Perin‡ Bahjat F. Qaqish**

*University of North Carolina at Chapel Hill, jpreisse@bios.unc.edu
†University of North Carolina at Chapel Hill
‡Johns Hopkins Bloomberg School of Public Health, jperin@jhsph.edu
**University of North Carolina, Chapel Hill, qaqish@bios.unc.edu

Deletion Diagnostics for Alternating Logistic Regressions

John S. Preisser, Kunthel By, Jamie Perin, and Bahjat F. Qaqish

Abstract

Deletion diagnostics are introduced for the regression analysis of clustered binary outcomes estimated with alternating logistic regressions, an implementation of generalized estimating equations (GEE) that estimates regression coefficients in a marginal mean model and in a model for the intracluster association given by the log odds ratio. The diagnostics are developed within an estimating equations framework that recasts the esti- mating functions for association parameters based upon conditional resid- uals into equivalent functions based upon marginal residuals. Extensions of earlier work on GEE diagnostics follow directly, including computational formulae for one-step deletion diagnostics that measure the influence of a cluster of observations on the estimated regression parameters and on the overall marginal mean or association model fit. The diagnostic formulae are evaluated with simulations studies and with an application concerning an as- sessment of factors associated with health maintenance visits in primary care medical practices. The application and the simulations demonstrate that the proposed cluster-deletion diagnostics for alternating logistic regressions are good approximations of their exact fully iterated counterparts.

Deletion diagnostics for alternating logistic regressions

John S. Preisser[1], Kunthel By[1], Jamie Perin[2], Bahjat F. Qaqish[1]

[1]University of North Carolina at Chapel Hill

[2]Johns Hopkins Bloomberg School of Public Health

SUMMARY. Deletion diagnostics are introduced for the regression analysis of clustered binary outcomes estimated with alternating logistic regressions, an implementation of generalized estimating equations (GEE) that estimates regression coefficients in a marginal mean model and in a model for the intracluster association given by the log odds ratio. The diagnostics are developed within an estimating equations framework that recasts the estimating functions for association parameters based upon conditional residuals into equivalent functions based upon marginal residuals. Extensions of earlier work on GEE diagnostics follow directly, including computational formulae for one-step deletion diagnostics that measure the influence of a cluster of observations on the estimated regression parameters and on the overall marginal mean or association model fit. The diagnostic formulae are evaluated with simulations studies and with an application concerning an assessment of factors associated with health maintenance visits in primary care medical practices. The application and the simulations demonstrate that the proposed cluster-deletion diagnostics for alternating logistic regressions are good approximations of their exact fully iterated counterparts.

1. Introduction

Many questions in medical research involving association structure among correlated binary responses are suitably addressed with marginal regression models. Consider, for example, cross-sectional observational medical practice data where patient outcomes are nested within physicians that are nested within practices. An earlier analysis of such data described in Preisser and Qaqish (1996) used generalized estimating equations (GEE) to estimate the effects of explanatory variables on the population-averaged probability of whether or not a patient made a health maintenance visit in the prior year. An additional question that may be posed involves modeling the marginal within-practice association structure of the response, and, in particular, characterizing the degree of association within physicians and within practices. When cluster sizes (eg., the number of patients sampled per practice) are small, second-order generalized estimating equations (GEE2) provide estimates of association parameters with good efficiency (Zhao and Prentice, 1990; Liang, Zeger, and Qaqish, 1992). However, GEE2 is not computationally feasible when cluster sizes are large as in the medical practice data where cluster sizes range from 19 to 197. Alternating logistic regressions (ALR), an implementation of GEE for the regression analysis of clustered binary data (Carey *et al.*, 1993) provides more efficient estimation of association parameters than first-order GEE (Liang and Zeger, 1986; Prentice, 1988) and, at least in situations with small cluster sizes, estimation nearly as efficient as GEE2 (Carey *et al.*, 1993; Lipsitz SR and Fitzmaurice, 1996). A practical limitation, however, is that there currently do not exist influence diagnostics for ALR. In the medical practice data, for example, it is natural to assess

whether data from individual practices have a large influence on estimates of within-physician and within-practice clustering.

This paper proposes computationally fast formulae and algorithms for estimating the effect of the deletion of a cluster of observations on the regression parameters in marginal models for intracluster associations. Section 2 describes the estimating equations procedure and introduces computationally efficient formulae to estimate the change in parameter estimates upon deletion of a cluster. The diagnostics are developed within an estimating function framework that recasts the ALR estimating functions for association parameters based upon conditional residuals (Carey, Diggle, Zeger, 1993) into equivalent functions based upon marginal residuals (Zink and Qaqish, 2009). Extensions of first order GEE diagnostics (Preisser and Qaqish, 1996) follow directly, including computational formulae for one-step deletion diagnostics that measure the influence of a cluster of observations on the estimated regression parameters and on the overall marginal mean or association model fit. Section 3 presents two simulation studies to evaluate the performance of the diagnostics. In section 4, the formulae are applied to the medical practice data.

2. Statistical methods

2.1 *Alternating logistics regressions based upon marginal residuals*

The development of deletion diagnostics in this paper is based upon a new representation of the ALR method through marginal residuals proposed by Zink and Qaqish (2009). Let $\mathbf{Y}_i$ be the response vector for the ith cluster where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is the vector of responses from the n_i observations in the ith cluster, $i = 1, \ldots, K$. Let $\boldsymbol{\mu}_i$ be the vector of population marginal

means, $\mathrm{E}[\mathbf{Y}_i] = \boldsymbol{\mu}_i$. A generalized linear model is $g_1(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta}$ where $g(\cdot)$ is the link function and $\mathbf{X}_{ij} = (x_{0ij}, x_{1ij}, \ldots, x_{p-1,ij})'$ is the $p \times 1$ vector of covariates for the j–th observation in the i–th cluster. Estimation by the ALR procedure is performed by iteratively solving two estimation equations, one for the marginal mean model parameters $\boldsymbol{\beta}$, and the other for the marginal bivariate association parameters $\boldsymbol{\alpha}$. The estimating equations for $\boldsymbol{\beta}$ are:

$$\sum_{i=1}^{K} \mathbf{D}'_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0 \tag{1}$$

where $\mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{diag}(\sigma_{ijj}^{1/2})\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{diag}(\sigma_{ijj}^{1/2})$, $\sigma_{ijj} = \mu_{ij}(1 - \mu_{ij})$, and $\mathbf{R}_i(\boldsymbol{\alpha})$ is a working correlation matrix.

The second set of estimating equations correspond to $\boldsymbol{\alpha}$. Let $m_i = n_i(n_i - 1)/2$ and j and k index observations within a cluster. Let $W_{ijk} = Y_{ij}Y_{ik}$ and define $\mu_{ijk} = E[W_{ijk}] = \mathrm{pr}(Y_{ij} = Y_{ik} = 1)$. The dependence or association between Y_{ij} and Y_{ik} can be represented by the pairwise odds ratio (Carey, Zeger, and Diggle, 1993; Lipsitz, Laird, and Harrington, 1991)

$$\psi_{ijk} = \frac{\mu_{ijk}(1 - \mu_{ij} - \mu_{ik} + \mu_{ijk})}{(\mu_{ij} - \mu_{ijk})(\mu_{ik} - \mu_{ijk})}.$$

A log pairwise odds ratio model is specified for the association,

$$\log[\psi_{ijk}(\mu_{ij}, \mu_{ij}, \mu_{ijk})] = \mathbf{Z}'_{ijk}\boldsymbol{\alpha}$$

where $\mathbf{Z}_{ijk} = (z_{0ijk}, z_{1ijk}, \ldots, z_{q-1,ijk})'$ is a covariate q-vector associated with the pair (Y_{ij}, Y_{ik}), and $\boldsymbol{\alpha}$ is a vector of association parameters. The method of Zink and Qaqish uses an m_i-vector $\mathbf{T}_i$ with elements T_{ijk} obtained as the residuals from the linear regression of W_{ijk} on Y_{ij} and Y_{ik}. Specifically,

$$T_{ijk} = W_{ijk} - \{\mu_{ijk} + b_{ijk:j}(Y_{ij} - \mu_{ij}) + b_{ijk:k}(Y_{ik} - \mu_{ik})\},$$

where

$$b_{ijk:j} = \mu_{ijk}(1-\mu_{ik})(\mu_{ik}-\mu_{ijk})/d_{ijk},$$

$$b_{ijk:k} = \mu_{ijk}(1-\mu_{ij})(\mu_{ij}-\mu_{ijk})/d_{ijk},$$

$d_{ijk} = \sigma_{ijj}\sigma_{ikk} - \sigma_{ijk}^2$, and $\sigma_{ijk} = \text{cov}(Y_{ij}, Y_{ik}) = \mu_{ijk} - \mu_{ij}\mu_{ik}$. The reformulated ALR estimating equations based upon marginal residuals are defined by

$$U_{\boldsymbol{\alpha}} = \sum_{i=1}^{K} E\left[\frac{-\partial \mathbf{T}_i{}'}{\partial \boldsymbol{\alpha}}\right] \mathbf{P}_i^{-1}\mathbf{T}_i = \sum_{i=1}^{K} \mathbf{C}_i'\mathbf{P}_i^{-1}\mathbf{T}_i \tag{2}$$

where $\mathbf{P}_i = \mathbf{diag}\{v_{ijk}\}$ and

$$v_{ijk} := \text{var}(T_{ijk}) = \frac{\mu_{ijk}(\mu_{ij}-\mu_{ijk})(\mu_{ik}-\mu_{ijk})(1-\mu_{ij}-\mu_{ik}+\mu_{ijk})}{\mu_{ij}\mu_{ik}(1-\mu_{ij}-\mu_{ik}+2\mu_{ijk}) - \mu_{ijk}^2}.$$

Expression (2) is equivalent to the ALR estimating equations based upon conditional residuals given by (7) of Carey et al. (1993). Zink and Qaqish (2009) allow $\mathbf{P}_i$ to be non-diagonal in a procedure they call orthogonalized residuals, thereby generalizing the ALR procedure and increasing efficiency. They show, following standard arguments (Liang and Zeger, 1986; Prentice, 1988), the asymptotic distribution of $K^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})'$, is multivariate Gaussian with mean zero. A practical advantage of (2) is that, unlike the case based upon conditional residuals, the robust variance estimator of the asymptotic covariance matrix corresponding to the association model is invariant to permutations of the i-th subject's response vector $\mathbf{Y}_i$. Kuk (2004) proposed a modified symmetrized version of the ALR equations based upon conditional residuals that have permutation invariance for the standard errors. By, Qaqish, and Preisser (2008) provide an R package for ALR and orthogonalized residuals, which includes the regression diagnostics proposed in the next section. A SAS macro is available at http://www.bios.unc.edu/∼qaqish/software.htm.

2.2 *Cluster–deletion diagnostics*

The regression diagnostics proposed in this section are computationally fast formulae because all matrix components of the diagnostics are available at convergence of the iteratively reweighted least squares algorithm. They are one-step deletion diagnostics because the computational formulae are equivalent to deleting the cluster and computing one more iteration of (1) and (2).

Let $\hat{\boldsymbol{\beta}}_{[i]}$ denote the estimate with the i-cluster deleted. The one-step GEE deletion diagnostic to approximate $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]}$ is

$$\text{DBETAC}_i = (\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1}(\mathbf{I}_{n_i} - \mathbf{H}_{1i})^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) \tag{3}$$

where $\mathbf{I}_d$ represents the identity matrix of dimension d and

$$\mathbf{H}_{1i} = \mathbf{D}_i(\mathbf{D}'\mathbf{V}^{-1}\mathbf{D})^{-1}\mathbf{D}_i'\mathbf{V}_i^{-1}. \tag{4}$$

Preisser and Qaqish (1996) gave a proof for a formula that is equivalent to (3) (see also Ziegler and Arminger, 1996). Hammill and Preisser (2006) proposed (3) in light of its connections, on a matrix component basis, to (1) and they showed its algebraic equivalency to expression (5) of Preisser and Qaqish (1996). Expression (4) is the leverage matrix (corresponding to $\boldsymbol{\beta}$) for cluster i (Mancl and DeRouen, 2001). The leverage of a cluster may be defined as the trace of H_{1i}, that is the sum of the diagonal elements which may individually be viewed as leverages of observations. Preisser and Qaqish (1996) gave a slightly different formula for the cluster leverage matrix.

Using arguments similar to their derivation of DBETAC_i, the one-step formula to approximate $\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{[i]}$ is

$$\text{DALPHAC}_i = (\mathbf{C}'\mathbf{P}^{-1}\mathbf{C})^{-1}\mathbf{C}_i'\mathbf{P}_i^{-1}(\mathbf{I}_{m_i} - \mathbf{H}_{2i})^{-1}\mathbf{T}_i \tag{5}$$

where

$$\mathbf{H}_{2i} = \mathbf{C}_i(\mathbf{C}'\mathbf{P}^{-1}\mathbf{C})^{-1}\mathbf{C}_i'\mathbf{P}_i^{-1} \tag{6}$$

is the cluster leverage matrix corresponding to $\boldsymbol{\alpha}$. Standardized versions of DBETAC_i and DALPHAC_i are obtained by dividing their components by their respective standard errors.

It is worth mention that computation of (5) is non-trivial for large cluster sizes. For the medical practice data analyzed in section 4, $max(n_i) = 197$, and, thus, $max(m_i) = 19,306$. Fortunately, due to its special structure, the matrix in (5) of this dimension involving H_{2i} can be easily inverted using an algorithm based upon the Sherman-Morrison-Woodbury formula (Sherman and Morrison, 1950). Details of the computational approach are provided by Preisser, Qaqish and Perin (2008).

The assessment of the influence of a cluster of observations on the overall model fit may be carried out with diagnostic measures that are extensions of Cook's distance for linear regression (Cook and Weisberg, 1982). Cluster level Cook's Distance for $\boldsymbol{\beta}$ is defined as for GEE1 (Corollary 2.1 of Preisser and Qaqish, 1996):

$$\text{DCLS}_{\boldsymbol{\beta},i}(p) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})'\widehat{\text{var}}^{-1}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[i]})/p \tag{7}$$

Analogously, Cook's Distance describing the influence of the i-th cluster on the overall fit of the model for $\boldsymbol{\alpha}$ is defined as

$$\text{DCLS}_{\boldsymbol{\alpha},i}(q) = (\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{[i]})'\widehat{\text{var}}^{-1}(\hat{\boldsymbol{\alpha}})(\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{[i]})/q \tag{8}$$

This is the measure introduced by Ziegler and Arminger (1996) in the context of modeling within-cluster correlations using the GEE approach of Prentice

(1988). In that context, Preisser and Perin (2007) provided computationally fast formula for the influence of the i-th cluster on the overall fit of the correlation model. A similar computationally fast formula for the influence of the i-th cluster on the overall fit of the within-cluster log odds ratio model, estimated with alternating logistic regressions, may be obtained by substituting DALPHAC_i in for $\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{[i]}$.

Interpretations for cluster diagnostics are not straightforward when cluster sizes vary. Generally, one might expect that larger clusters tend to have larger influence, so plots of cluster diagnostics against cluster size are recommended to assess their influence.

3. Simulation studies

3.1 *The performance of the one-step approximation*

The first of two simulation studies was conducted to determine the extent to which the clusters with the most extreme exact cluster Cook's distance are identified by the one-step cluster Cook's distance. As in a study on the performance of Cook's Distance in the generalized linear mixed model (Xiang *et al.* 2002), the simulation study assessed the diagnostics ability to identify the clusters with the largest and second largest exact Cook's distances. We have two *a priori* expectations. First, we expect the one-step approximation, to a large extent, will identify the same clusters as those identified by the exact cluster Cook's distance. Second, we expect the probability of identifying the same clusters to increase as the value of the exact cluster Cook's distance increases. The simulation experiment was based upon 500 data sets generated from the following model using the algorithm of Qaqish (2003):

$$\text{logit}(\mu_{ij}) \quad = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \tag{9}$$

$$\log \psi_{ijk} \quad = \alpha_0 + \alpha_1 z_{1ijk} + \alpha_2 z_{2ijk} \tag{10}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -0.8 \\ 0.27 \\ 0.20 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 1.05 \\ 0.35 \\ -0.35 \end{bmatrix}.$$

and x_{1ij}, x_{2ij}, z_{1ijk}, and z_{2ijk} are continuous covariates: $x_{1ij} = [2(i-1)/(K-1)] - 1$ is a cluster level covariate taking equally spaced values in the interval [-1,1], $i = 1, \ldots, K$; $x_{2ij} = [2(j-1)/(n-1)] - 1$ is an observation level covariate taking equally spaced values in the interval [-1,1] where n denotes the number of observations in each cluster, $n = n_i$ for all i; $j = 1, \ldots, n$; $z_{1ijk} = x_{1ij}$; and $z_{2ijk} = |x_{2ij} - x_{2ik}|$. These parameter values were chosen so as to induce response vectors with positive within-cluster association that decreased over time, akin to autoregressive correlation in longitudinal data settings. For each replication, we simultaneously fit models (9) and (10) with the ALR estimating procedure given by equations (1) and (2). Within a replication, for each cluster, we computed both exact cluster Cook's distance and one-step approximated cluster Cook's distance for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Note that the computation of exact cluster Cook's distance for all clusters required an additional K applications of ALR per replication to obtain fully iterated parameter estimates after deletion of a single cluster. Due to the computational intensity of the experiment, only the combination of $(K = 50, n = 5)$ was considered, requiring a total of 500K =25,000 applications of ALR.

The results were as follows: 88% of the time, the one-step formula (8) correctly identified the most influential cluster on $\boldsymbol{\alpha}$; 82% of the time, diagnostic (8) correctly identified the top two most influential clusters on $\boldsymbol{\alpha}$ (possibly in the reversed order); 60% of the time, the one-step formula (7)

correctly identified the most influential cluster on $\boldsymbol{\beta}$; 31% of the time, the diagnostic (7) correctly identified the top two most influential clusters on $\boldsymbol{\beta}$. While the latter result was not very good, we note that $DCLS_{\boldsymbol{\beta},i}$ was able to identify at least one of the two most influential clusters 88% of the time.

To determine if the probability of correctly identifying the most influential clusters with $\text{DCLS}_{\beta,i}$ increased as the influence of those clusters increased, as measured by the magnitude of the exact fully-iterated Cook's distance, two logistic regressions were carried out. First, the binary indicator for whether the cluster with largest exact cluster Cook's distance matched with the cluster with the largest $\text{DCLS}_{\beta,i}$ was regressed on the exact cluster Cook's distance. A significant monotonically increasing relationship (figure 1a) shows that for a value of Cook's distance at the third quantile (with respect to the 500 simulated largest exact cluster Cook's distance values), the probability of detecting the most influential cluster is approximately 75%. Next, the binary indicator for whether the two clusters with largest exact cluster Cook's distance were the same as the two clusters with largest $\text{DCLS}_{\beta,i}$ was regressed on the 2nd largest exact cluster Cook's distance. A significant monotonically increasing relationship (figure 1b) shows that for values of the second largest Cook's distance at the 75th, 90th, and 95th percentiles, respectively, the estimated probabilities of detecting the two most influential clusters are approximately 35%, 42%, and 47%, respectively. A more extensive report of the simulations is available as a technical report (Preisser, By, and Qaqish, 2008).

3.2 *Responsiveness of diagnostics to contamination models*

The aim of the second simulation study is to consider the distribution of extreme cluster Cook's distance for β and α under binary response contaminated data models. The simulation study investigates whether the Cook's distance measures are responsive to contamination of the response data? Second, if it is responsive to contamination, does it behave in some predictable way relative to the Cook's distance for the uncontaminated data? We expect to see a shift in the distribution Cook's distance under contamination relative to the uncontaminated data. Furthermore, we expect this shift to grow (to a point) as the level of contamination increases.

Data are generated from the models (9) and (10) with constant cluster size n using the same values of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ from the previous section. Contamination is considered under two scenarios: (1) random contamination (RC) and (2) cluster concentrated (CC) contamination. Under random contamination, each of the Kn observations is contaminated with probability p_c. Letting $Y_{ij,c}$ be the contaminated observation and Y_{ij} be the original observation, the contamination is done as follows:

$$Y_{ij,c} = \begin{cases} 1 - Y_{ij} & \text{with probability } p_c \\ Y_{ij} & \text{with probability } 1 - p_c \end{cases}$$

Under cluster contamination, each cluster is first chosen with probability $2p_c$. Once the cluster is chosen, each of the n observations within the chosen cluster is contaminated with probability 0.50. The simulation experiment was conducted to investigate 63 scenarios: $K = \{50, 100, 200\}$, $n = \{5, 20, 50\}$, $p_c = \{0, 0.02, 0.05, 0.10\}$ and $cc = \{\text{RC}, \text{CC}\}$ (when $p_c > 0$).

To address whether the one-step cluster Cook's distance diagnostics for

$\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are sensitive to contamination, the empirical distributions of their largest order statistics under models of contamination were examined relative to empirical distributions of the largest Cook's statistic when their was no contamination. This was accomplished visually with QQ plots for each combination of K and n. Figure 2 compares the QQ plots of the largest order statistic for cluster Cook's distance for α under contaminated data relative to uncontaminated. For cluster size 5 (lower panels of Figure 2), there is not any observable difference between the empirical distribution of the cluster Cook's distance under contamination relative to no contamination. However, for cluster size $n = 50$ (as well as for $n = 20$, not shown), the spacing (or clear separation) in the loess fits, and the fact that the curves generally lie above the 45 degree line, indicates that the distibution of Cook's distance under contamination is shifted to the right, and that their values increase monotonically with the level of contamination. Plots for $K = 200$ (not shown) are similar to plots for $K = 100$ for a given value of n. Thus, at least for α, cluster Cook's distance under random contamination behaves in the manner that is expected by shifting to the right as the level of contamination increases. The second largest, third largest, and fourth largest cluster Cook's distance for $\boldsymbol{\alpha}$ also exhibit this behavior (plots not shown).

This behavior is not consistent in general. In fact, for every other situation, the QQ plots either show that the distribution of the contaminated Cook's distance is no different than the uncontaminated or that if a shift is present, it is shifting to the left as the contamination increases. For example, for β under random contamination and small cluster sizes ($n = 5$), the distribution has a tendency to shift further and further to the left as

the level of contamination increases, opposite as was expected (plots not shown). For the other cluster sizes ($n = 20$ and $n = 50$) as they pertain to $\boldsymbol{\beta}$ under random contamination, there is no apparent difference between contaminated and uncontaminated distributions. A report of the entire simulation experiment, including exhaustive displays of QQ plots for both random and cluster-concentrated contamination cases, is available (By, Preisser and Qaqish, 2008).

4. Application of Diagnostics to Medical Practice Data

The proposed cluster deletion diagnostics are illustrated with medical practice data. In 1990-1991, chart review data were collected from a random sample of 3889 medical charts in 57 medical practices (clusters). The cluster sizes (number of patients per practice) ranged from 19 to 197 with a mean of 68. A logistic regression model was specified for the probability that the j-th patient in the i-th practice made at least one maintenance visit during the years 1990 and 1991. Preisser and Qaqish (1996) introduced and applied to the medical practice data computationally efficient formulae for cluster deletion diagnostics to identify practices with the largest influences on regression coefficients. ALR may be used to fit the same logistic regression model for the marginal mean while specifying an additional model for the within-practice association. The pairwise odds ratio model has the form of (10) where $z_{1ijk} = 1$ if patients j and k in practice i have the same physician (and $z_{1ijk} = 0$, otherwise); and $z_{2ijk} = (n_i - 68)/50$. Note α_0 is the log pairwise odds ratio of health maintenence visit for two patients from the same medical practice who saw different doctors for a practice with cluster size $n_i = 68$; α_1 is the change in the log pairwise odds ratio between two patients

who saw the same doctor, relative to the association between two patients who saw different doctors; and α_2 is the change in the within-practice log pairwise odds ratio for two patients comparing two practices that differ in cluster size by 50 patients.

The first column of results in Table 1 shows the ALR parameter estimates applied to the full data set where all three association model covariates are statistically significant at the 0.05 level. Recall that the standard errors in Table 1, based upon the reformulated ALR of Zink and Qaqish given in equation (2), are invariant to the ordering of a subjects' responses. For a practice of mean cluster size, the estimated between-physician within-practice odds ratio is 1.71, and the within-physician odds ratio is 2.29. These associations decrease with increasing cluster size. It is natural to inquire whether certain practices have an undue influence on these estimates. The remaining columns of Table 1 show the fully iterated results obtained upon deleting selected clusters, suggesting that some clusters have a moderate influence on estimates of within-practice and within-physician clustering.

Figure 3 presents cluster deletion diagnostic statistics for medical practices for the ALR procedure given by (1) and (2). Plots (a), (b) and (c) depict the difference, given by the vertical distance between a point and the 45 degree line, between the fully-iterated parameter estimate after removal of a medical practice ('exact') and the approximate change given by $DALPHAC_i$, for α_0, α_1, and α_2, respectively. These plots show that the proposed one-step diagnostics provide good approximations of the exact change, as nearly all the points fall close to the 45 degree line. The most infuential clusters are identified in the figure, practice # 34 for $\hat{\alpha}_0$ and $\hat{\alpha}_2$ (Figures

3(a) and 3(c), respectively), and practice # 15 for $\hat{\alpha}_1$ (Figure 3(b)). Figure 3(d) shows the three clusters with the most influence on the overall fit of the within-practice association model; practice # 34 has the second greatest influence by this measure. Table 2 gives the actual values of the deletion diagnostic statistics for selected practices. It is interesting that Preisser and Qaqish (1996) using $\text{DCLS}_{\beta,i}$ (see their expression (9)) identified cluster # 5 as having the greatest influence among clusters on the overall fit of the marginal mean model. This was also true for the ALR analysis of the influence of clusters on β (not shown). However, as shown in Table 2, cluster # 5 was not particularly notable for its overall influence on the fit of model (10) despite its having the second largest influence on α_1.

5. Discussion

Application to the medical practice data, as well as results from the first simulation study (section 3.1), demonstrate that the proposed cluster-deletion diagnostics for alternating logistic regressions are good approximations of their exact counterparts. On the other hand, results from the second simulation failed to provide convincing evidence that the extreme Cook's Distance diagnostics respond in a consistent manner to contaminated binary response models. Besides their relevance to alternating logistic regressions, these are the first published results of their kind concerning the behavior in practical situations of $DCLS_{\boldsymbol{\beta},i}$ (Preisser and Qaqish, 1996) for generalized estimating equations (Liang and Zeger, 1986).

The proposed diagnostic formulae are computationally fast. Computation of the cluster diagnostics for the medical practice data of section 4 took 13 minutes on dual 700 MHz SPARC processors, compared to over 12 hours for

computation of their fully iterated 'exact' counterparts. Qualitatively similar computational savings using formulae similar to (5)- (8), in the context of modelling intracluster correlations using the estimating equations approach of Prentice (1988), have been reported by Preisser and Perin (2007), for four data sets from medicine and public health. One-step formulae that have structure similar to those presented here could be developed for other estimating equation procedures for correlated binary data (Kuk and Nott, 2000; le Cessie and van Houwelingen, 1994; Lipsitz and Fitzmaurice, 1996).

Although cluster-deletion diagnostics seem to be the most useful, diagnostic formulae for other kinds of subset deletion could be developed. Preisser and Qaqish (1996) proposed a one-step approximation to $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[m]}$ for GEE where m denotes an arbitrary subset of observations to be deleted. A similar formula for $\hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\alpha}}_{[m]}$ could be easily developed with derivations similar to those found in their appendix. Besides cluster-deletion diagnostics, we have developed and implemented in SAS/IML software observation-deletion diagnostics for ALR that approximate the change in regression coefficients when a single observation (eg., patient) is deleted. The formulae, not presented here, are similar in form to observation-deletion diagnostics of Preisser and Perin (2007). When applied to the medical practice data, however, they reveal that no single patient was found to have large influence (results not shown).

Acknowledgements

This research was supported by grant CA101901 from the U.S. National Institutes of Health.

REFERENCES

By, K., Preisser, J., and Qaqish, B. (2008). A simulation experiment to investigate the distributional behavior of extreme Cook's Distance for GEE to models with contaminated binary responses. *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series.* Working Paper 9.
http://biostats.bepress.com/uncbiostat/papers/art9.

By, K., Qaqish, B, and Preisser J. (2008). *orth: Multivariate logistic regression using orthogonalized residuals.* R package version 1.5, URL http://cran.r-project.org.

Carey, V., Zeger, S.L. and Diggle, P. (1993). Modeling multivariate binary data with alternating logistic regressions. *Biometrika* 80, 517-526.

Hammill, B.G. and Preisser, J.S. (2006). A SAS/IML Program for GEE and Regression Diagnostics. *Computational Statistics and Data Analysis* 51, 1197-1212.

Kuk, A.Y.C. (2004). Permutation invariance of alternating logistic regressions for multivariate binary data. *Biometrika* 91, 758-761.

Kuk, A.Y.C. and Nott, D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters* 47, 329-35.

le Cessie, S. and van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics* 43, 95–108.

Liang, K.-Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13-22.

Liang, K.-Y., Zeger, S.L. and Qaqish, B.F. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society B* 54, 3-40.

Lipsitz, S.R. and Fitzmaurice, G.M. (1996). Estimating equations for measures of association between repeated binary responses. *Biometrics* 52, 903-12.

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika* 78, 153-60.

Mancl, L.A., and DeRouen, T.A. (2001). A covariance estimator for GEE with improved small-sample properties. *B*iometrics 57, 126-134.

Preisser, J.S. and Qaqish, B.F. (1996). Deletion diagnostics for generalized estimating equations. *Biometrika* 83, 551-562.

Preisser, J., By, K., and Qaqish, B. (2008). Performance of one-step approximation relative to exact cluster Cook's Distance for GEE. *The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series.* Working Paper 8. http://biostats.bepress.com/uncbiostat/papers/art8.

Preisser, J.S. and Perin, J. (2007). Deletion diagnostics for marginal mean and correlation model parameters in estimating equations. *Statistics and Computing* 17, 381-393.

Preisser, J.S., Qaqish, B.F., and Perin J. (2008). A note on deletion diagnostics for estimating equations. *Biometrika* 95, 509-513.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* 44, 1033-1048.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrka*, 90, 455–463.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics,* 21, 124-7.

Xiang, L.M. Tse, S.K., and Lee, A.H. (2002). Influence diagnostics for generalized linear mixed models: application to clustered data. *Computational Statistics and Data Analysis* 40, 759-774.

Zhao, L. P. and Prentice R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* 77, 642-648.

Ziegler, A. and Arminger, G. (1996). Parameter estimation and regression diagnostics using generalized estimating equations. In Faulbaum F. & Bandilla, W. (editors): *StatSoft 95 - Advances in statistical software* **5**, 229-37. Lucius & Lucius, Stuttgart.

Ziegler, A., Blettner, M., Kastner, C. and Chang-Claude. J, (1998). Identifying influential families using regression diagnostics for generalized estimating equations. *Genetic Epidemiology* 15, 341-353.

Zink R.C. and Qaqish B.F. (2009). Correlated binary regression using orthogonalized residuals. *COBRA Preprint Series.* Working Paper 51. http://biostats.bepress.com/cobra/ps/art51/.

Table 1

Model parameter estimates (est.) with robust standard errors (se) for the logistic model of the marginal probability a patient made a health maintenance visit during the years 1990 and 1991 (β), and for the log odds ratio model of within-practice association (α). Data are from the North Carolina Early Cancer Detection Program at the Lineberger Comprehensive Cancer Center. Results are presented for the full data, and based upon selected cluster deletions.*

	Full Data		Without #15		Without #19		Without #34	
Parameter	est.	(se)	est.	(se)	est.	(se)	est.	(se)
β‡								
INTERCEPT	-0.106	(0.173)	-0.070	(0.182)	-0.178	(0.170)	-0.155	(0.167)
NBRMDS	-0.034	(0.039)	-0.013	(0.043)	-0.018	(0.053)	-0.030	(0.039)
M3	0.249	(0.170)	0.340	(0.161)	0.244	(0.171)	0.228	(0.173)
SPECLTY	-0.078	(0.249)	-0.178	(0.263)	0.056	(0.250)	-0.075	(0.239)
MDAGE	-0.264	(0.064)	-0.297	(0.059)	-0.268	(0.066)	-0.227	(0.065)
MDSEX	0.424	(0.262)	0.511	(0.266)	0.476	(0.283)	0.472	(0.255)
MDFLU	-0.072	(0.099)	-0.123	(0.087)	-0.092	(0.104)	-0.073	(0.102)
PATAGE	-0.097	(0.034)	-0.103	(0.035)	-0.101	(0.035)	-0.103	(0.035)
BLACKPAT	-0.395	(0.123)	-0.413	(0.122)	-0.401	(0.124)	-0.426	(0.124)
MALEPAT	-0.411	(0.065)	-0.422	(0.067)	-0.431	(0.066)	-0.421	(0.068)
NOINSUR	-0.416	(0.119)	-0.422	(0.122)	-0.439	(0.123)	-0.393	(0.120)
α‡‡								
INTERCEPT	0.538	(0.173)	0.613	(0.172)	0.524	(0.156)	0.410	(0.147)
SAMEMD	0.290	(0.112)	0.207	(0.098)	0.256	(0.107)	0.347	(0.113)
CLSIZE	-0.179	(0.062)	-0.178	(0.069)	-0.147	(0.072)	-0.139	(0.052)

∗ Standard errors are from the reformulated ALR based upon marginal residuals.
‡ NBRMDS = number of doctors in practice minus one; M3 = (The number of patients over 50 years old seen per day minus 15)/10. SPECLTY = doctor's specialty: 0 if family or general practice, 1 if internal medicine; MDAGE = (doctor's age in years minus 45)/10; MDSEX=1 if female, 0 if male; MDFLU = Doctor's flu vaccination: 0 in the last two years, 1 if 3 to 5 years ago, 2 if never; PATAGE = (patient's age in years - 65)/10; BLACKPAT = 1 if black, and 0 if; white; MALEPAT = 1 if patient is male, and 0 if female; and NOINSUR = 1 if patient is not insured, and 0 if insured.
‡‡ SAMEMD = 1 if two patients have the same doctor, and 0 otherwise; CLSIZE is the size of the cluster centered at 68 scaled by 50.

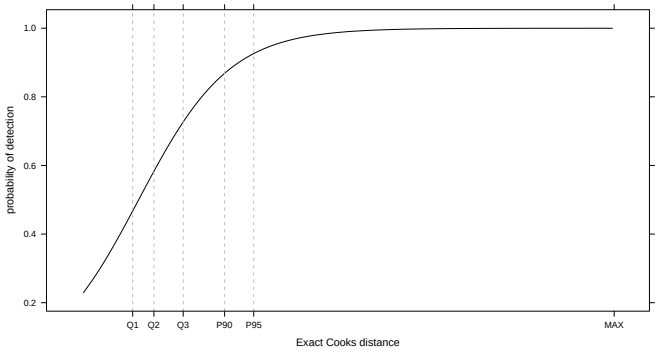
Table 2

One-step approximated cluster-deletion diagnostic values for selected clusters. The values are based upon the "full data" analysis in Table 1 using data from the North Carolina Early Cancer Detection Program. Ranks are shown in parentheses.

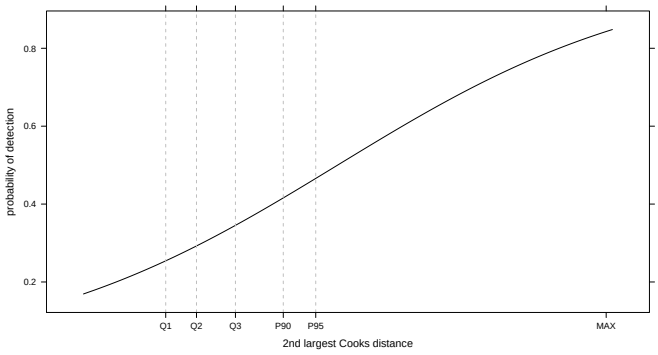
					DALPHA*					
Cluster #	n_i		$DCLS_\alpha^\ddagger$		INTERCEPT		SAMEMD		CLSIZE	
5	191	(2)	0.059	(7)	-0.121	(13)	0.367	(2)	-0.123	(17)
15	158	(3)	0.064	(5)	-0.197	(8)	0.417	(1)	-0.050	(43)
19	197	(1)	0.126	(3)	-0.039	(32)	0.284	(6)	-0.452	(2)
34	100	(12)	0.143	(2)	0.629	(1)	-0.359	(3)	-0.521	(1)
50	120	(9)	0.061	(6)	0.356	(3)	-0.150	(12)	-0.080	(27)
52	140	(6)	0.146	(1)	0.358	(2)	-0.333	(4)	0.235	(6)

∗ Rankings are based on the absolute value of DALPHA.

‡ Cluster level Cook's distance for α.



(a) Response: detect largest $\text{DCLS}_{\beta,i}$



(b) Response: detect two largest $\text{DCLS}_{\beta,i}$

Figure 1. Probability of detecting the most influential clusters with respect to β. Plot 1(a) shows the probability of detecting the cluster with the largest exact Cook's distance. Plot 1(b) shows the probability of detecting the two clusters with largest exact Cook's distance. Q_1, Q_2, and Q_3 represent the appropriate quantiles of exact cluster Cook's distance from the 500 simulationss while P_{90} and P_{95} denote the 90-th and 95-th percentiles respectively.

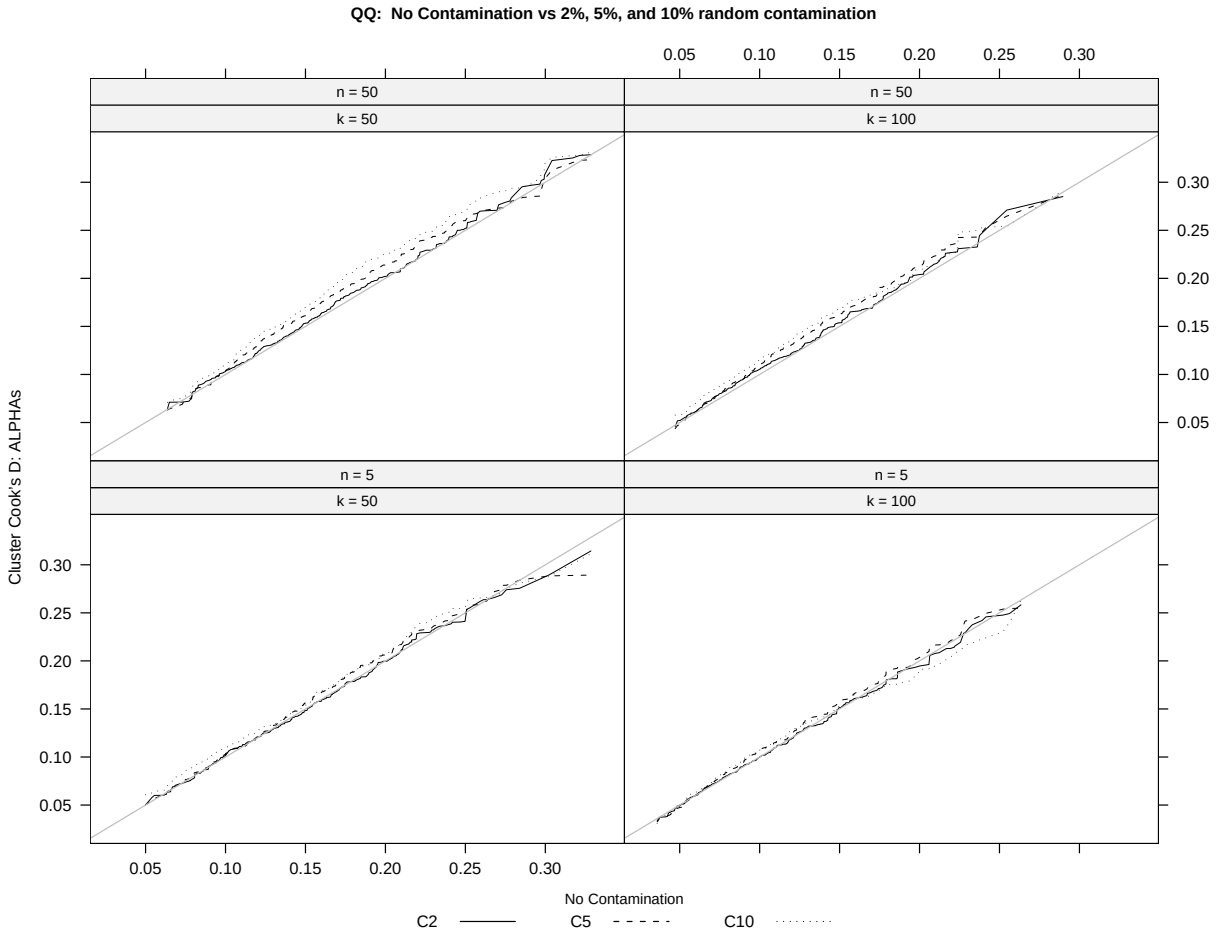
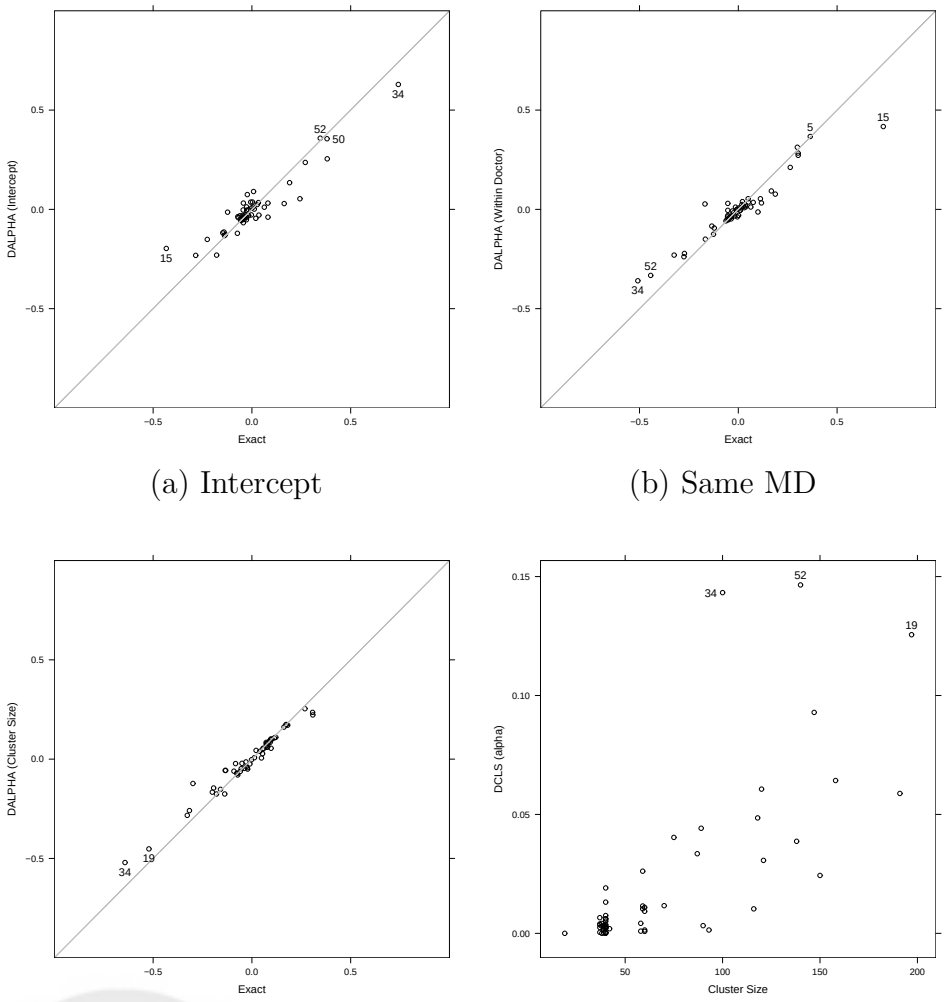


Figure 2. QQ plot of cluster Cook's distance for α. Vertical axis denotes cluster Coook's distance under random contamination. C2, C5 and C10 denotes 2, 5 , and 10 percent contamination respectively. The horizontal axis denotes cluster Cook's distance under no contamination.



(a) Intercept (b) Same MD

(c) Scaled cluster size (d) Cluster cook's distance (α)

Figure 3. Cluster deletion diagnostics for within-cluster association model. DALPHA's and the exact deletion diagnostic for α in plots 3(a), 3(b), and 3(c) are standardized by the appropriate robust standard errors.