

University of North Carolina at Chapel Hill

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

Year 2008

Paper 8

Performance of One-Step Approximation Relative to Exact Cluster Cook's Distance for GEE

John Preisser*

Kunthel By†

Bahjat Qaqish‡

*University of North Carolina at Chapel Hill, jpreisse@bios.unc.edu

†University of North Carolina at Chapel Hill, kby@bios.unc.edu

‡University of North Carolina at Chapel Hill, qaqish@bios.unc.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/uncbiostat/art8>

Copyright ©2008 by the authors.

Performance of One-Step Approximation Relative to Exact Cluster Cook's Distance for GEE

John Preisser, Kunthel By, and Bahjat Qaqish

Abstract

A simulation experiment was conducted to assess the performance of the generalized estimating equations (GEE) approximated cluster Cook's Distance relative to its fully-iterated exact version. Specific interest was in the effect of cluster-deletion on the overall fit of a model for the marginal mean and on the overall fit of a model for the marginal association, in the context of alternating logistic regressions (ALR). The experiment demonstrated that one-step approximated cluster Cook's Distance statistics successfully identify clusters having the greatest influence, as measured by their fully iterated exact counterparts. The regression diagnostic identified the cluster with the greatest influence on the fit of the marginal association model 88% of the time. The cluster with the greatest influence on the fit of the marginal mean model was identified 60% of the time, and the probability of detection increased with increasing size of the diagnostic. Cook's Distance statistics for GEE and recently proposed analogous measures for ALR are useful for the analysis of clustered binary data.

Performance of One-Step Approximation Relative to Exact Cluster Cook's Distance for GEE

John S. Preisser, Kunthel By, and Bahjat F. Qaqish
Department of Biostatistics
University of North Carolina at Chapel Hill

Summary

Abstract

A simulation experiment was conducted to assess the performance of the generalized estimating equations (GEE) approximated cluster Cook's Distance relative to its fully-iterated exact version. Specific interest was in the effect of cluster-deletion on the overall fit of a model for the marginal mean and on the overall fit of a model for the marginal association, in the context of alternating logistic regressions (ALR). The experiment demonstrated that one-step approximated cluster Cook's Distance statistics successfully identify clusters having the greatest influence, as measured by their fully iterated exact counterparts. The regression diagnostic identified the cluster with the greatest influence on the fit of the marginal association model 88% of the time. The cluster with the greatest influence on the fit of the marginal mean model was identified 60% of the time, and the probability of detection increased with increasing size of the diagnostic. Cook's Distance statistics for GEE and recently proposed analogous measures for ALR are useful for the analysis of clustered binary data.

1 Goal

The goal of this simulation study is to assess the performance of the generalized estimating equations (GEE) one-step approximated cluster Cook's distance relative to its fully-iterated exact version. The subject of this report includes the effect of cluster-deletion on the overall fit of a model for the marginal mean as captured by the Cook's distance deletion diagnostic for first-order GEE (Preisser and Qaqish, 1996). Additional interest is in the effect of cluster-deletion on the overall fit of a model for the marginal within-cluster association as captured by a Cook's distance deletion diagnostic for the within-cluster odds ratio introduced recently by Preisser et al. (2008) for the estimating equations procedure known as alternating logistic regressions (ALR) (Carey et al., 1993). The diagnostic for ALR is similar in structure to a computationally fast Cook's distance diagnostic for the marginal within-cluster correlation (Preisser and Perin, 2007) recently introduced for the estimating equations procedure of Prentice (1988). Performance is judged by 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 an earlier 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. In succeeding discussions, we will use the phrases 'probability of identifying the same clusters' and 'probability of detection' interchangeably.

2 The Experiment

Adopting notation used in Preisser et al. (2008), let μ_{ij} denote the marginal mean of the binary response for the j th observation from the i th cluster, and let ψ_{ijk} denote the pairwise odds ratio of an event between the j th and k th observations in the i th cluster. The simulation experiment consists of the following sequence of steps. 500 data sets are generated based on Qaqish (2003) using the following model:

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

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

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. Specifically, $x_{1ij} = [2(i-1)/(K-1)] - 1$ is a cluster level covariate taking equally spaced values in the interval $[-1, 1]$ where K denotes the number of clusters and $i = 1, \dots, K$. Similarly, $x_{2ij} = [2(j-1)/(m-1)] - 1$ is an observation level covariate taking equally spaced values in the interval $[-1, 1]$ where m denotes the number of observations in each cluster and $j = 1, \dots, m$. The covariates for the association model are defined as follows: $z_{1ijk} = x_{1ij}$ and $z_{2ijk} = |x_{2ij} - x_{2ik}|$. For each replication, we simultaneously fit models (1) and (2) with the ALR estimating procedure using software developed at the University of North Carolina at Chapel Hill (Zink, 2003; By et al., 2008b); 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 $\boldsymbol{\alpha}$ as defined in Preisser et al. (2008). 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 $(K = 50, m = 5)$ was considered, requiring a total of $500K = 25,000$ applications of ALR. For ease of exposition we will not distinguish between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ unless the potential for confusion warrants doing so. Let

$$D_{(K),Xct}, D_{(K-1),Xct}$$

denote the largest and second largest *exact* cluster Cook's distance respectively and

$$D_{(K),1\text{-step}}, D_{(K-1),1\text{-step}}$$

denote the largest and second largest one-step approximated cluster Cook's distance respectively. Based on these Cook's distance measures, we define the following response variables.

1. $Y_1 = 1$ if $D_{(K),Xct}$ and $D_{(K),1\text{-step}}$ identify the same cluster and $Y_1 = 0$ otherwise.
2. $Y_2 = 1$ if $D_{(K),Xct}$ and $D_{(K),1\text{-step}}$ identify the same cluster **and** $D_{(K-1),Xct}$ and $D_{(K-1),1\text{-step}}$ identify the same cluster. Otherwise, $Y_2 = 0$.
3. $Y_3 = 1$ if at least one of the top two clusters identified by both methods matched; no particular order is required.
4. $Y_4 = 1$ if the clusters identified by $D_{(K),1\text{-step}}$ and $D_{(K-1),1\text{-step}}$ matched those of $D_{(K-1),Xct}$ and $D_{(K),Xct}$; no particular order is required.

In some sense, the Y 's measure the ability of the one-step approximated cluster Cook's distance to identify the same influential clusters as those identified by the exact cluster Cook's distance. In this regard, we may think of the Y 's as a Bernoulli random variable taking on value 1 if the extreme one-step approximated cluster Cook's distance detects the same clusters as the extreme exact cluster Cook's distance.

Examples: For example, if the top two clusters identified by $D_{(K),Xct}$ and $D_{(K-1),Xct}$ are 22 and 20 and the top two clusters identified by $D_{(K),1-step}$ and $D_{(K-1),1-step}$ are 22 and 10, then

$$Y_1 = 1, \quad Y_2 = 0, \quad Y_3 = 1, \quad Y_4 = 0.$$

Similarly, if the top two clusters identified by $D_{(K),Xct}$ and $D_{(K-1),Xct}$ are 94 and 22 and the top two clusters identified by $D_{(K),1-step}$ and $D_{(K-1),1-step}$ are 22 and 94, then

$$Y_1 = 0, \quad Y_2 = 0, \quad Y_3 = 1, \quad Y_4 = 1.$$

3 Methods for Assessment of Diagnostics

To assess the performance of one-step approximated cluster Cook's distance, the means of the Y 's are computed. A high average indicates that the one-step approximated cluster Cook's distance performs well relative to the exact cluster Cook's distance. A low average indicates that the one-step approximated cluster Cook's distance performs poorly. In addition, if the performance of the one-step approximated cluster Cook's distance is not strong, we want to consider whether the odds of detection improves if the *exact* extreme cluster Cook's distance gets larger, where by odds of detection we mean the odds that the one-step approximated cluster Cook's distance identifies the same cluster as the exact cluster Cook's distance. The rationale for this latter exploration is as follows. If the extreme cluster Cook's distance is small, it doesn't matter whether the one-step does poorly. Even if one is able to identify the most influential cluster whether through the one-step method or the exact method, if the extreme Cook's distance is very small, the information is not very useful. What matters most is being able to identify the correct influential cluster when the particular cluster has a large effect on Cook's distance.

Models under consideration Let $\mu_r = \Pr(Y_r = 1)$, $r = 1, \dots, 4$. Let \bar{D}_{Xct} denote the average of $D_{(K),Xct}$ and $D_{(K-1),Xct}$ and let Δ_{Xct} denote the difference of $D_{(K),Xct}$ and $D_{(K-1),Xct}$. For the response Y_1 , we hypothesized a strong association between detection and the size of the exact cluster Cook's distance. As such, the following models were considered:

$$\text{logit } \mu_1 = \beta_0 \tag{3}$$

$$\text{logit } \mu_1 = \beta_0 + \beta_1 D_{(K),Xct} \tag{4}$$

$$\text{logit } \mu_1 = \beta_0 + \beta_1 G_{K2} + \beta_2 G_{K3} + \beta_3 G_{K4} \tag{5}$$

$$\text{logit } \mu_1 = \beta_0 + \beta_1 D_{(K),Xct} + \beta_2 D_{(K),Xct}^2 \tag{6}$$

where

$$G_{Kq} = \begin{cases} 1, & \text{if } Q_{K,q-1} < D_{(K),Xct} \leq Q_{K,q} \\ 0, & \text{otherwise} \end{cases}$$

$q = 2, 3, 4$ and $Q_{K,q}$ denotes the q -th quantile of the largest exact Cook's distance for β . Since Y_2 , Y_3 and Y_4 measure detection of the top 2 influential clusters, the following models were considered:

$$\text{logit } \mu_r = \beta_0 \tag{7}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 \bar{D}_{Xct} \tag{8}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 \Delta_{Xct} \tag{9}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 \bar{D}_{Xct} + \beta_2 \bar{D}_{Xct}^2 \tag{10}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 \Delta_{Xct} + \beta_2 \Delta_{Xct}^2 \tag{11}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 D_{(K-1),Xct} \tag{12}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 D_{(K-1),Xct} + \beta_2 D_{(K-1),Xct}^2 \tag{13}$$

$$\text{logit } \mu_r = \beta_0 + \beta_1 G_{K12} + \beta_2 G_{K13} + \beta_3 G_{K14} \tag{14}$$

for $r = 2, 3, 4$, where

$$G_{K1q} = \begin{cases} 1, & D_{(K-1),Xct} \text{ is in the } q\text{-th quantile} \\ 0, & \text{otherwise} \end{cases}$$

$q = 2, 3, 4$ and G_{K1q} denotes the q -th quantile of the second largest exact Cook's distance.

4 Results

Table 1 summarizes the agreement between the one-step and the exact cluster Cook's distance in identifying the most influential clusters based upon models (3) and (7). The first column provides the average of the four responses corresponding to correct identification of cluster Cook's distance for α and the second column corresponds to cluster Cook's distance for β . One striking feature

Table 1: Proportion of influential sets of clusters correctly identified by the diagnostics

Mean	α	β
$\Pr(Y_1 = 1)$	0.88	0.60
$\Pr(Y_2 = 1)$	0.74	0.23
$\Pr(Y_3 = 1)$	0.99	0.88
$\Pr(Y_4 = 1)$	0.82	0.31

displayed in Table 1 is that when the cluster Cook's distance for α is used as the metric, the ability of the one-step to detect the most influential cluster is strong. For example, $\Pr(Y_1 = 1) = 0.88$ means that eighty eight percent of the time, the one-step correctly identifies the most influential cluster. $\Pr(Y_2 = 1) = 0.74$ means that seventy four percent of the time, the one-step correctly identifies the top two most influential clusters in the same order as that given by the exact cluster Cook's distance for α . $\Pr(Y_3 = 1) = 0.99$ means that ninety-nine percent of the time, the one-step approximation correctly identifies at least one of the top two most influential clusters but not necessarily in the same order as that given by the exact cluster Cook's distance for α . Lastly, $\Pr(Y_4 = 1) = 0.82$ means that eighty-two percent of the time, the one-step approximation correctly

identifies the top two most influential clusters but possibly in the reversed order as that suggested by the exact cluster Cook's distance. Overall, the performance of the cluster Cook's distance for α in detecting the appropriate influential clusters is strong.

This picture is not shared when we look at the agreement in terms of the cluster Cook's distance for β . For all four response variables, the ability of the one-step to detect is much smaller than that seen for α . The saving grace is the third response. There, we see that eighty-eight percent of the time, the one-step approximation correctly identifies at least one of the top two most influential clusters. Since this probability is high, we chose not to consider further models for Y_3 . Instead, we focused our modeling efforts on (4) through (6) for Y_1 and (8) through (14) with $r = 2$ and with $r = 4$ (and not with $r = 3$). We rescaled the extreme Cook's distances by multiplying by 100. The original values were too small to make parameter estimates meaningful in the odds scale.

In fitting the above models, likelihood ratio tests suggest that including the quadratic terms does not contribute to substantial information gains (all p values are bigger than 0.10). Therefore, succeeding discussions will exclude all models with quadratic terms (i.e., we will not further consider models (6), (10), (11), and (13)). Table 2 provides some summary statistics for the covariates.

Table 2: Summary statistics for the data generated by the 500 simulations for Cook's distance for β after rescaling by multiplying by 100. The first and third quartiles are Q_1 and Q_3 whereas P_{90} and P_{95} denote the 90th and 95th percentiles, respectively

Variable	Minimum	Q_1	Median	Q_3	P_{90}	P_{95}	Maximum
$D_{(K),Xct}$	2.73	4.7950	5.995	7.655	10.0050	11.655	32.100
$D_{(K-1),Xct}$	2.19	3.7100	4.350	5.160	6.1550	6.830	12.870
Δ_{Xct}	0.00	0.4150	1.215	2.695	5.0050	6.360	27.470
\bar{D}_{Xct}	2.46	4.3675	5.270	6.410	7.7625	8.885	18.365

Response Y_1 : Table 3 gives estimates under model (4). The odds ratio estimate is 1.47 with

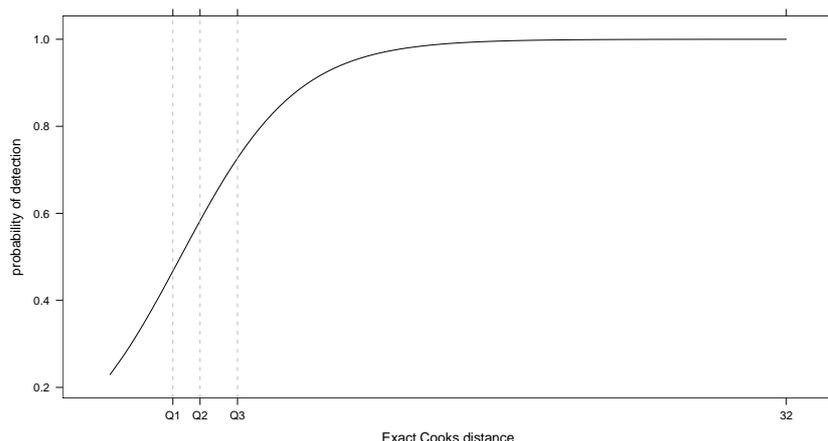
Table 3: Parameter estimates for model (4), response Y_1

Effects	Estimate	Std Error	Chi Square	p Value
Intercept	-1.9872	0.3395	34.2598	< 0.0001
$D_{(K),Xct}$	0.3874	0.0550	49.5672	< 0.0001

95% confidence interval (1.323, 1.641). Figure 1 plots the probability of detection as a function of cluster Cook's distance for β . Q_1, Q_2, Q_3 in Figure 1 are the first, second, and third quartiles of the largest cluster Cook's distance for β respectively. These values can be gleaned from Table 2. We see that for values of Cook's distance around the third quartile, the probability of detecting the most influential cluster is approximately 75%. Contrast this with the unadjusted probability of detection of 60% in Table 1.

The results for model (5) are provided in Table 4. We see that the log odds of detection is strong if the cluster Cook's distance is in the group G_{K4} . The probability of detection is 85.48% for cluster Cook's distance in this group. For values in G_{K3} , the probability of detection is approximately

Figure 1: Probability of detection as a function of cluster Cook's distance for β for response Y_1 . Q1, Q2, and Q3 represents the appropriate quantiles of the largest cluster Cook's distance



63.5%.

Table 4: Parameter estimates of model (5) for response Y_1 where the predictor is a classification of the values of the largest cluster Cook's distances into groups based on quantiles

Effects	Estimate	Std Error	Chi Square	p Value
Intercept	-0.4389	0.1832	5.7393	0.0166
G_{K2}	0.5832	0.2564	5.1737	0.0229
G_{K3}	0.9923	0.2604	14.5218	0.0001
G_{K4}	2.2117	0.3139	49.6397	< 0.0001

Response Y_2 : Recall that Y_2 measures the agreement between the one-step cluster Cook's distance and the exact cluster Cook's distance in identifying the top two most influential clusters in the same order as that identified by the exact Cook's distance. For all models associated with the second response (models (8) through (14) with $r = 2$), the odds of detection do not appreciably improve with increasing size of cluster Cook's distance. Figure 2 plots the probability of detection against the various Cook's distance summaries. The probability of detection is not impressive. Table 5 gives estimates under model (14). Observations based on the group G_{K14} does not show any noticeable improvement in the probability of detection. In fact, the probability of detection is only 33.87%.

Figure 2: Probability of detecting as a function of cluster Cook's distance for β for response Y_2 described in models (8), (9), and (12), $r = 2$

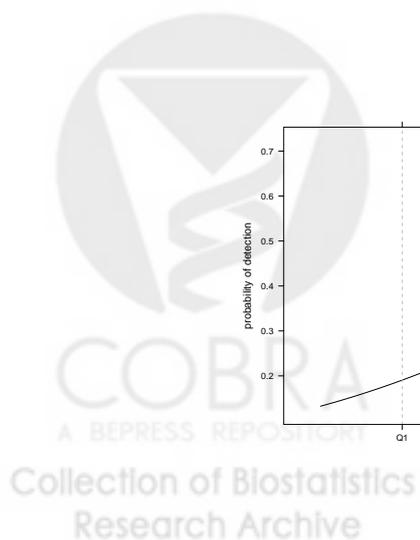
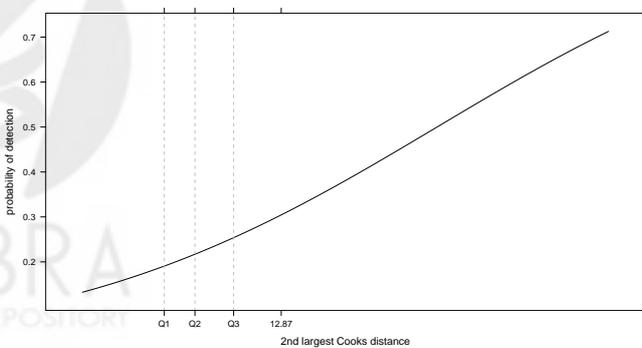
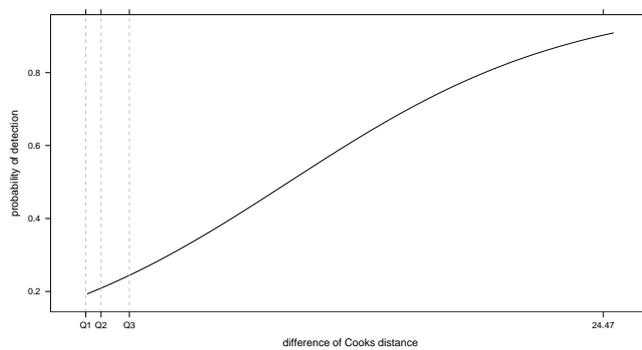
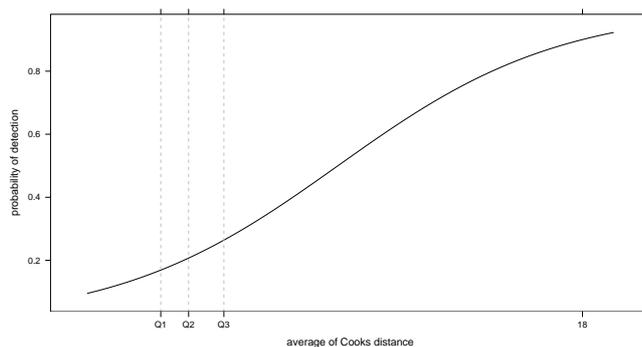


Table 5: Parameter estimates of model (14) for response Y_2 where the predictor is a classification of the values of the second largest cluster Cooks distance into groups based on quantiles

Effects	Estimate	Std Error	Chi Square	p Value
Intercept	-1.4567	0.2267	41.3019	< 0.0001
G_{K12}	0.0396	0.3212	0.0152	0.9018
G_{K13}	0.1096	0.3160	0.1203	0.7287
G_{K14}	0.7876	0.2956	7.0995	0.0077

Response Y_3 : For the third response, further analyses were not conducted because the probability of detection (reported in Table 1) was high.

Response Y_4 : With the fourth response, the picture is similar to the second response. Probability plots (Figure 3) associated with models (8), (9), and (12) for $r = 4$ show weak probabilities of detection similar to those seen in Figure 2. Model (12) provides the strongest association between the log-odds of detection and the predictor relative to models (8) and (9). From Table 6, the probability of detection when $D_{(K-1),X_{ct}} = 5.160$ (the 3rd quantile) is 34.54%. Similarly, the probabilities of detection at the 90th and 95th percentiles are 41.59% and 46.59%, respectively. Estimates for model (14) where we categorized observations into groups based on the quantiles of the second largest cluster Cook's distance for β are provided in Table 7. The probability of detection for an observation in group G_{K14} is 42.74%, a slight improvement over the unadjusted probability of detection of 0.31 as seen in Table 1.

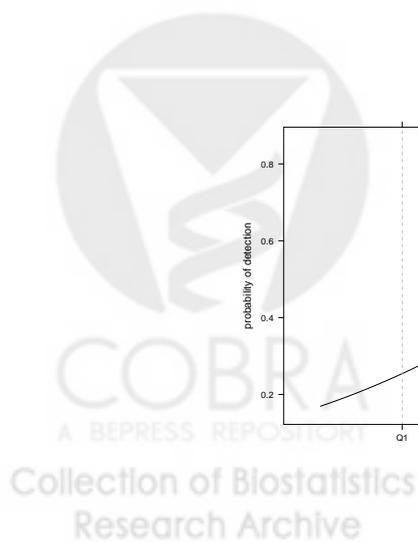
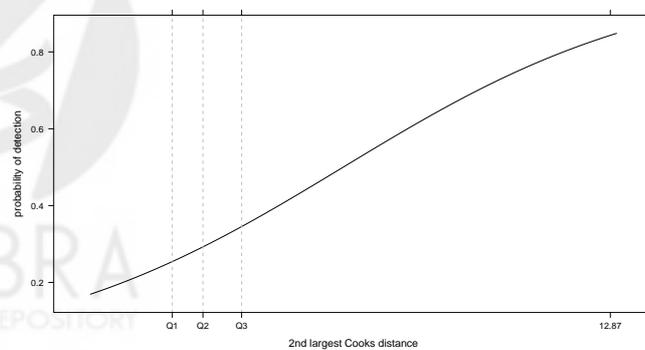
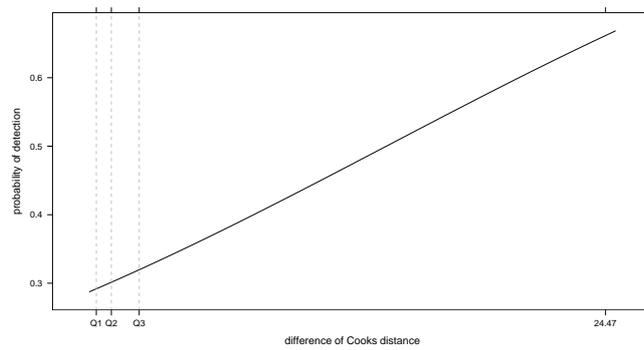
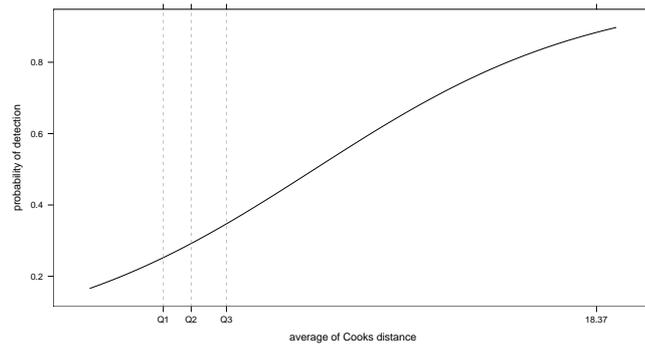
Table 6: Parameter estimates for model (12) for response Y_4 . $D_{(K-1),X_{ct}}$ is scaled from the original values by multiplication by 100

Effects	Estimate	Std Error	Chi Square	p Value
Intercept	-2.1930	0.3633	36.4459	< 0.0001
$D_{(K-1),X_{ct}}$	0.3011	0.0744	16.3985	0.0001

Table 7: Parameter estimates of model (14) for response Y_4 where the predictor is a classification of the values of the second largest cluster Cooks distance into groups based on quantiles

Effects	Estimate	Std Error	Chi Square	p Value
Intercept	-1.1302	0.2066	29.9357	< 0.0001
G_{K12}	0.0851	0.2914	0.0854	0.7701
G_{K13}	0.3648	0.2816	1.6778	0.1952
G_{K14}	0.8379	0.2750	9.2847	0.0023

Figure 3: Probability of detection as a function of the second largest cluster Cook's distance for β for response Y_4 described in models (8), (9), and (12), $r = 4$



5 Conclusions

The simulation study demonstrated that one-step approximated cluster Cook's distance statistics successfully identify those clusters that have the greatest influence, as measured by the fully iterated exact cluster Cook's distance, on parameter estimates in an analysis based upon the alternating logistic regressions estimating procedure for clustered binary data. This conclusion especially applies to the one-step cluster Cook's distance for the model for the within-cluster odds ratio introduced recently (Preisser et al., 2008). In particular, the results of the simulation found that the most influential cluster, as defined by the one with the largest exact cluster Cook's distance, was correctly identified by the diagnostics 88% of the time. The two most influential clusters, with respect to α , were identified 82% of the time. Results for β were promising, but not as good as those for α . In particular, the one-step approximated cluster Cook's distance measures for the marginal mean regression model (Preisser and Qaqish, 1996) correctly identified the most influential cluster 60% of the time. However, the two most influential clusters, irrespective of order of influence, were only identified 31% of the time. While the latter result is not very good, the simulation showed that the probability of identifying the most influential clusters increases as the value of the exact cluster Cook's distance increases. In particular, when the largest exact cluster Cook's distance for β was in the largest quartile among the 500 replications of the experiment, the largest one-step diagnostic was associated with the cluster with the largest exact cluster Cook's distance 85% of the time. When the second largest exact cluster Cook's distance was in the upper quartile, the one-step diagnostic correctly identified the two most influential clusters 43% of the time, a notable improvement relative to 31% based on all 500 replications. While the diagnostics had some difficulty in correctly identifying the two most influential clusters for β in model (1), Table 1 reported that at least one of the two most influential clusters were identified 88% of the time. These results are clearly limited since they are based upon one combination of β, α, K and m .

This report adds to the evidence that the GEE deletion diagnostics are useful for the analysis of clustered data. Using binary outcome data from 57 medical practices, Preisser and Qaqish (1996) showed that the one-step cluster Cook's distance for β are good approximations to their exact counterparts. Preisser and Perin (2007) used binary response data from a cluster trial aimed at reducing underage drinking to further illustrate the goodness of the approximation. They also illustrated the goodness of the approximation of the one-step cluster Cook's distance with respect to α from a model for the within-cluster correlation with four data sets drawn from statistical practice. Preisser et al. (2008) demonstrated the goodness of the approximation of the one-step diagnostic for α in an ALR analysis using the same data considered by Preisser and Qaqish (1996). Hammill and Preisser (2006) presented a SAS/IML software program that implements the one-step cluster Cook's distance diagnostics for first-order GEE (Liang and Zeger, 1986). By et al. (2008a,b) described free software for computing ALR diagnostics.

References

- K. By, B.F. Qaqish, and J.S. Preisser. *orth: Multivariate logistic regression using orthogonalized residuals*, 2008a. URL <http://cran.r-project.org>. R package version 1.5.
- K. By, B.F. Qaqish, J.S. Preisser, J. Perin, and R.C. Zink. ORTH: R and SAS software for regression models of correlated binary data based on orthogonalized residuals. (Submitted), 2008b.
- V. Carey, S.L. Zeger, and P. Diggle. Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80:517–526, 1993.

- B.G. Hammill and J.S. Preisser. A SAS/IML software program for GEE and regression diagnostics. *Computational Statistics & Data Analysis*, 51:1197–1212, 2006.
- K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.
- J.S. Preisser and J. Perin. Deletion diagnostics for marginal mean and correlation model parameters in estimating equations. *Statistics and Computing*, 17:381–393, 2007.
- J.S. Preisser and B.F. Qaqish. Deletion diagnostics for generalized estimating equations. *Biometrika*, 83:551–562, 1996.
- J.S. Preisser, K. By, J. Perin, and B.F. Qaqish. Regression diagnostics for alternating logistic regressions. (Submitted), 2008.
- R.L. Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44:1033–1048, 1988.
- B.F. Qaqish. A family of multivariate binary distributions for simulating correlated binary with specified marginal means and correlations. *Biometrika*, 90:455–463, 2003.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL <http://www.R-project.org>.
- L. Xiang, S. Tse, and A.H. Lee. Influence diagnostics for generalized linear mixed models: applications to clustered data. *Computational Statistics & Data Analysis*, 40:759–774, 2002.
- R.C. Zink. *Correlated binary regression using orthogonalized residuals*. PhD thesis, University Of North Carolina, Chapel Hill, 2003.

