Biometrika (1996), **83,** 3, *pp.* 551–562
Printed in Great Britain

Deletion diagnostics for generalised estimating equations

BY JOHN S. PREISSER

Section on Biostatistics, Bowman Gray School of Medicine, Medical Center Boulevard, Winston-Salem, North Carolina 27157-1063, U.S.A.

AND BAHJAT F. QAQISH

Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27599-7400, U.S.A.

SUMMARY

Deletion diagnostics are proposed for generalised estimating equations. The diagnostics consider leverage and residuals to measure the influence of a subset of observations on the estimated regression parameters and on the estimated values of the linear predictor. Computational formulae are provided which correspond to the influence of a single observation and of an entire cluster of correlated observations. Additionally, diagnostics are given which approximate the effect of deletion of an arbitrary subset of observations under a model with general covariance structure and arbitrary link function, extending Proposition 3 of Christensen, Pearson & Johnson (1992). The proposed measures are applied to medical data.

Some key words: Cook's distance; Generalised estimating equations; Generalised linear model; Influence; Leverage; One-step approximation; Regression diagnostics; Residual.

1. INTRODUCTION

The generalised estimating equations procedure of Liang & Zeger (1986) has been applied to a wide range of medical and biological applications in which correlation arises among outcomes which are measured repeatedly on subjects or in which dependence occurs because of clustering. Much of the appeal of generalised estimating equations is due to their broad capabilities which include modelling correlated binary responses, and allowing time-varying covariates. Despite their frequent use, however, there do not exist diagnostics to identify observations or clusters which have a disproportionately large influence on the estimated regression parameters. We introduce deletion diagnostics which account for the leverage and residuals in a set of observations to determine their influence on regression parameter estimates and fitted values. When that set consists of only one observation we call them 'observation-deletion' diagnostics. When it consists of a whole cluster, we call them 'cluster-deletion' diagnostics. Although our primary focus is clustered data, a diagnostic is given which approximates the effect of deletion of an arbitrary subset of observations under a model with a general covariance structure. This diagnostic generalises the case-deletion diagnostics for dependent responses of Christensen et al. (1992) to models with arbitrary link functions. The diagnostics for generalised estimating equations are generalisations of those for generalised linear models (Pregibon, 1981; Williams, 1987; McCullagh & Nelder, 1989, Ch. 12), and they reduce to well-known measures of influence in linear regression found in Cook & Weisberg (1982), Belsley, Kuh & Welsch (1980),

and in a review paper by Chatterjee & Hadi (1986). In § 2, we review generalised estimating equations and introduce some notation. Computational formulae provided in § 3 are based on one-step approximations (Pregibon, 1981), while accounting for within-cluster correlation. Interpretation of the proposed measures is discussed through an illustrative example in § 4.

2. GENERALISED ESTIMATING EQUATIONS

For $i = 1, \ldots, K$, let $Y_i := (Y_{i1}, \ldots, Y_{in_i})'$ be a n_i-vector of response values, and $X_i := (x'_{i1}, \ldots, x'_{in_i})'$ a $n_i \times p$ matrix of covariate values. Throughout the paper, clusters are indexed by i and observations by t. We consider models where the forms of the first two moments for the marginal distribution of Y_{it} are

$$E(Y_{it}) = \mu_{it}, \quad g(\mu_{it}) = \eta_{it} = x_{it}\beta, \quad \text{var}(Y_{it}) = f_{it}\phi. \tag{1}$$

In the terminology of generalised linear models, $g(\mu_{it})$ is the link function, $f_{it} := f(\mu_{it})$ with f_{it} the variance function, β is a $p \times 1$ vector of regression coefficients, and ϕ is the scale parameter, either known or to be estimated. Estimates of β are obtained by solving the generalised estimating equations

$$\sum_{i=1}^{K} (\partial\mu_i/\partial\beta)' \{A_i R_i(\alpha) A_i\}^{-1} (Y_i - \mu_i) = 0, \tag{2}$$

where $\partial\mu_i/\partial\beta$ is a $n_i \times p$ matrix, $A_i := \text{diag}(f_{it}^{1/2})$ is a $n_i \times n_i$ diagonal matrix, and $R_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix that depends on an unknown parameter vector α. A solution is obtained by alternating between estimation of ϕ, α and β, using method of moment estimators for ϕ and α. We define $N := \sum n_i$, the $N \times 1$ vector $Y := (Y'_1, \ldots, Y'_K)'$, the $N \times p$ matrix $X := (X'_1, \ldots, X'_K)'$ assumed to be of full column rank, and $D := \partial\eta/\partial\mu$, a $N \times N$ diagonal matrix with nonzero elements $d_{it} := \partial\eta_{it}/\partial\mu_{it}$. Estimation of β is done with iteratively reweighted least squares by regressing the working response vector $Z := X\hat{\beta} + D(Y - \hat{\mu})$ on X with block diagonal weight matrix W whose ith block, corresponding to the ith cluster, is the $n_i \times n_i$ matrix

$$W_i := D_i^{-1} A_i^{-1} R_i^{-1}(\hat{\alpha}) A_i^{-1} D_i^{-1}, \quad D_i := \text{diag}(d_{i1}, \ldots, d_{in_i}).$$

Liang & Zeger (1986) show that, under regularity conditions, as $K \to \infty$, $K^{1/2}(\hat{\beta} - \beta)$ is asymptotically multivariate Gaussian with mean vector 0 and covariance matrix given by

$$J_{\hat{\beta}} := \lim_{K \to \infty} K J_1^{-1} J_2 J_1^{-1}, \tag{3}$$

where

$$J_1 := \sum_{i=1}^{K} (\partial\mu_i/\partial\beta)' \{A_i R_i(\alpha) A_i\}^{-1} (\partial\mu_i/\partial\beta),$$

$$J_2 := \sum_{i=1}^{K} (\partial\mu_i/\partial\beta)' \{A_i R_i(\alpha) A_i\}^{-1} \text{cov}(Y_i) \{A_i R_i(\alpha) A_i\}^{-1} (\partial\mu_i/\partial\beta).$$

The 'robust' or sandwich variance estimate of $\hat{\beta}$ is obtained by replacing $\text{cov}(Y_i)$ by $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$ and β, ϕ and α by their estimates in $J_1^{-1} J_2 J_1^{-1}$. It is robust in the sense that it consistently estimates $J_{\hat{\beta}}$ even if $R(\alpha)$ is misspecified. If $R(\alpha)$ is correctly specified, $\text{cov}(Y_i) = A_i R_i(\alpha) A_i \phi$ and (3) reduces to $\phi \lim K J_1^{-1}$ and the respective 'naive' vari-

ance estimator of $\hat{\beta}$ is

$$\phi\left(\sum_{i=1}^{K} X_i' W_i X_i\right)^{-1},$$

which can be seen by noting that $\partial\mu_i/\partial\beta = (\partial\mu_i/\partial\eta_i)(\partial\eta_i/\partial\beta) = D_i^{-1}X_i$.

A current estimate of β is updated by

$$\hat{\beta}_{\text{new}} = (X'WX)^{-1}X'WZ,$$

evaluating the right-hand side at the current estimate. Then $\hat{\beta}_{\text{new}}$ is used to update $\hat{\eta} = X\hat{\beta}_{\text{new}} = HZ$, where $H := QW$ and $Q := X(X'WX)^{-1}X'$. The asymmetric and idempotent projection matrix H maps the current value of Z into estimated values of the linear predictor. The diagonal elements of H, denoted by h_{it}, correspond to the amount of leverage of the response on the corresponding fitted value. The average of the h_{it} is p/N, which follows from $\text{tr}(H) = p$. The leverage of a cluster is contained in the matrix $H_i := Q_i W_i$ where $Q_i := X_i(X'WX)^{-1}X_i'$, and can be summarised by $\text{tr}(H_i)$ which has the additive property of being the sum of observation leverages.

The estimated adjusted residual vector is

$$E := D(Y - \hat{\mu}) = Z - \hat{\eta} = (I - H)Z. \tag{4}$$

Considering H and the current estimate of β as nonrandom, the variances of E and $\hat{\eta}$ are easily obtained by observing that $\text{var}(Z) = D\,\text{var}(Y)D$. If the correct weights $W = \phi\{\text{var}(Z)\}^{-1}$ are applied, $\text{var}(E) = \phi(W^{-1} - Q)$ and $\text{var}(\hat{\eta}) = \phi Q$; by 'correct' we mean that the working correlation structure is correctly specified, and $\hat{\alpha} = \alpha$. Additionally, $\text{var}(E_i) = \phi(W_i^{-1} - Q_i)$ and $\text{var}(\hat{\eta}_i) = \phi Q_i$, where $E_i = D_i(Y_i - \hat{\mu}_i)$ is the estimated residual of the ith cluster.

3. Measures of influence based on case-deletion

3·1. *General*

We consider the effect of deleting one or more observations on $\hat{\beta}$. Exact formulae would require complete iteration for every subset of observations deleted, which is computationally prohibitive, even in relatively small data sets. We introduce computationally feasible one-step approximations, like those of Pregibon (1981) for generalised linear models. In addition to considering the change in individual $\hat{\beta}_j$'s, we give formulae for the effect of deletion of one or more observations on the estimated values of the linear predictor. Formulae are given for the change caused by deleting an arbitrary subset of observations, and for two useful special cases for clustered data: deleting one observation and deleting one cluster.

3·2. *Assessing the influence of case deletion on the $\hat{\beta}_j$'s*

Let m index the subset of m observations that are to be deleted, let $[m]$ denote the remaining observations, and let $\hat{\beta}_{[m]}$ denote the regression parameter estimate when the set m is removed from the data. Without loss of generality, assume that the observations to be deleted are the first m components of Z, and let W be partitioned as

$$W = \begin{pmatrix} W_m & W_{m[m]} \\ W_{[m]m} & W_{[m]} \end{pmatrix}.$$

All vectors and matrices will be partitioned in a parallel manner. Also, assume a general covariance structure and define $V = W^{-1}$.

THEOREM 1. *The one-step approximation for* $\hat{\beta}_{[m]}$ *is*

$$\hat{\beta}_{[m]} \doteq \hat{\beta} - (X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{E}_m,$$

where

$$\tilde{X}_m := X_m - V_{m[m]}V_{[m]}^{-1}X_{[m]}, \quad \tilde{Q}_m := \tilde{X}_m(X'WX)^{-1}\tilde{X}'_m,$$
$$\tilde{E}_m := \tilde{Z}_m - \tilde{X}_m\hat{\beta} = E_m - V_{m[m]}V_{[m]}^{-1}E_{[m]}, \quad \tilde{Z}_m := Z_m - V_{m[m]}V_{[m]}^{-1}Z_{[m]}.$$

The proof is in the Appendix. The matrices, $\tilde{X}_m$, W_m^{-1}, $\tilde{Q}_m$, are the building blocks of multiple case-deletion diagnostics for the linear model with general covariance matrix, V, considered by Christensen et al. (1992). The vector $\tilde{E}_m$ is introduced in order to generalise their result to other link functions, giving a one-step approximation for the change in the regression parameter estimates obtained from (2). Recall that $E = (E'_m, E'_{[m]})'$ is the residual of the working response vector given by (4). The dimension of $V_{[m]}^{-1}$ is $(N-m)\times(N-m)$ so that Theorem 1 is not computationally feasible, except for special cases, two of which we now examine.

Now, consider the model described in § 2. First, we introduce a cluster-deletion diagnostic which measures the effect of a single cluster on the estimated regression parameter vector. Let the subscript $[i]$ denote estimates evaluated without the ith cluster.

COROLLARY 1·1. *The one-step approximation for* $\hat{\beta} - \hat{\beta}_{[i]}$ *is*

$$DBETAC_i := (X'WX)^{-1}X'_i(W_i^{-1} - Q_i)^{-1}E_i. \tag{5}$$

Proof. Applying Theorem 1, $V_{i[i]} = 0$ implies $\tilde{E}_i = E_i$, $\tilde{X}_i = X_i$ and $\tilde{Q}_i = Q_i$. □

The simple structure of (5), resulting from the fact that W and V are block diagonal matrices, provides a formula which is computationally feasible. An alternative expression to (5) is given by observing that $(W_i^{-1} - Q_i)^{-1} = W_i(I - H_i)^{-1}$.

Next, we introduce an observation-deletion diagnostic. Let $\hat{\beta}_{[it]}$ denote estimates evaluated using all the data except the tth observation of the ith cluster, and let the subscript $i[t]$ denote matrices corresponding to the ith cluster without the tth observation. Let the matrices W_i and V_i be partitioned as

$$W_i = \begin{pmatrix} W_{it} & W_{it[t]} \\ W_{i[t]t} & W_{i[t]} \end{pmatrix}, \quad V_i = W_i^{-1} = \begin{pmatrix} V_{it} & V_{it[t]} \\ V_{i[t]t} & V_{i[t]} \end{pmatrix}.$$

COROLLARY 1·2. *Let*

$$E_{it} := D_{it}(Y_{it} - \hat{\mu}_{it}), \quad E_{i[t]} := D_{i[t]}(Y_{i[t]} - \hat{\mu}_{i[t]}).$$

Then the one-step approximation for $\hat{\beta} - \hat{\beta}_{[it]}$ *is*

$$DBETAO_{it} := (X'WX)^{-1}\tilde{X}'_{it}\frac{\tilde{E}_{it}}{W_{it}^{-1} - \tilde{Q}_{it}}, \tag{6}$$

where

$$\tilde{X}_{it} := X_{it} - V_{it[t]}V_{i[t]}^{-1}X_{i[t]}, \quad \tilde{Q}_{it} := \tilde{X}_{it}(X'WX)^{-1}\tilde{X}'_{it}, \quad \tilde{E}_{it} := E_{it} - V_{it[t]}V_{i[t]}^{-1}E_{i[t]}.$$

Note that W_{it}, $\tilde{Q}_{it}$ and E_{it} are scalars.

Proof. Applying Theorem 1,

$$V_{[m]m} = \begin{pmatrix} V_{i[t]t} \\ 0 \end{pmatrix}$$

and because V is block diagonal we get

$$V_{m[m]} V_{[m]}^{-1} X_{[m]} = V_{it[t]} V_{i[t]}^{-1} X_{i[t]}, \quad V_{m[m]} V_{[m]}^{-1} E_{[m]} = V_{it[t]} V_{i[t]}^{-1} E_{i[t]}.$$

The result follows. □

The influence diagnostics in Theorem 1, Corollary 1·1 and Corollary 1·2 are generalisations of a number of known results in which the estimating equations are special cases of (2). If independence is assumed, the distinction between one cluster and one observation vanishes as W becomes a diagonal matrix, and the result in either of the corollaries above reduces to the one-step approximation for generalised linear models given in § 3 of Williams (1987). Applying Corollary 1·1 with all cluster sizes equal to 1, we obtain

$$\hat{\beta} - \hat{\beta}_{[i]} \doteq (X'WX)^{-1} X_i' W_i^{1/2} (1 - h_i)^{-1/2} r_{pi}, \tag{7}$$

where W_i is a scalar, h_i is the ith diagonal element of $H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}$, and $r_{pi} := (y_i - \mu_i)\{f_i(1 - h_i)\}^{-1/2}$ is a version of the Pearson residual. Pregibon (1981) introduced (7) for the logistic regression model, in which case a simpler form is obtained because $W_i = f_i$ due to the canonical link.

Next consider an identity link, $g(\mu_{it}) \equiv \mu_{it}$, for the model in (1) with a general covariance structure, V. The identity link implies $D = I$ which implies $\tilde{Z}_m = \tilde{Y}_m$, where

$$\tilde{Y}_m := Y_m - V_{m[m]} V_{[m]}^{-1} Y_{[m]}.$$

A formula identical to that of Theorem 1 is obtained, but with $W = V^{-1}$ and $\tilde{E}_m = \tilde{Y}_m - \tilde{X}_m \hat{\beta}$. For observation deletion, proposition 3 of Christensen et al. (1992) is obtained as a special case of Corollary 1·2.

Consideration of independence, identity link, and constant variance simultaneously gives well-known results for the multiple linear regression model. In this case, $W = I$ and $\tilde{Q}_m = X_m (X'X)^{-1} X_m'$. For multiple case deletion, Theorem 1 reduces to $DBETA_m$ (Atkinson, 1985, p. 20). For observation deletion it reduces to $DBETA_i$ as given in Cook (1977) and Belsley et al. (1980, p. 13).

Other classes of influence diagnostics have been proposed for multivariate linear models. Barrett & Ling (1992) consider various diagnostics for the multivariate linear regression model. They consider diagnostics for clusters only, which they call cases, whereas we consider diagnostics for observations as well.

The proposed diagnostics can be standardised using the variances of $\hat{\beta}$ based on the complete data or with the subset m omitted from the calculation. For example, the standardised one-step approximation for the changes in $\hat{\beta}_j$ due to the deletion of the ith cluster is

$$DBETACS_{ij} := DBETAC_{ij} / \{\hat{\phi}(X'WX)^{-1}\}_{jj}^{1/2}. \tag{8}$$

For single-case deletion in multiple linear regression this measure is equal to $DBETAS_{ij}$ of Belsley et al. (1980, p. 13) if ϕ is estimated by $s_{[i]}^2$ which is the usual estimate of variance calculated without the ith case. Notice that standardisation in (8) is achieved by dividing by the naive standard error of the jth coefficient based on all the data. Alternatively, the studentised approximation for the change in β_j is obtained by dividing by the square root of $\{\hat{\phi}(X_{[i]} W_{[i]} X_{[i]})^{-1}\}_{jj}$ which omits the effect of the ith cluster on the standard error of

$\hat{\beta}_j$. Both of these scaled measures for the change in $\hat{\beta}$ use the naive variance estimate which is a consistent estimate of the true variance of $\hat{\beta}$ only if the working correlation structure is correctly specified. Alternatively, scaling may be done by the robust variance estimate which is consistent even under misspecification of $R(\alpha)$ (Liang & Zeger, 1986). We prefer the naive variance estimate. The robust variance estimate, with squared residuals in the middle of its sandwich formula, is inflated by large residuals, so its use for standardisation would mask observations with large residuals.

3·3. *Assessing the influence of case deletion on the fitted values*

Diagnostics analogous to those of Cook (1977, 1979) can be obtained to measure the influence of observations on estimated values of the linear predictor, and hence on the fitted values. These diagnostics measure the influence of the deleted subset m on the overall fit. The change in the overall fit is given by either $X\Delta\hat{\beta}_m$ or $X_{[m]}\Delta\hat{\beta}_m$, where $\Delta\hat{\beta}_m := \hat{\beta} - \hat{\beta}_{[m]}$, depending upon whether one's view is that of deleting or adding a subset of observations. A class of norms which are location and scale invariant is given by $D_m(M; c) := (\Delta\hat{\beta}_m)' M(\Delta\hat{\beta}_m)/c$. First, norms with $c = p\phi$ and $M = X'WX$ are considered.

THEOREM 2. *A measure of the standardised influence of the subset m of observations on the linear predictor is given by*

$$D_m(X'WX;\, p\hat{\phi}) = (\hat{\beta} - \hat{\beta}_{[m]})'(X'WX)(\hat{\beta} - \hat{\beta}_{[m]})/(p\hat{\phi}) \tag{9}$$

$$= \tilde{E}_m'(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{Q}_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{E}_m/(p\hat{\phi}). \tag{10}$$

As in § 3·1, we consider cluster deletion and observation deletion under the model described in § 2.

COROLLARY 2·1. *The effect of the ith cluster on the overall fit is*

$$DCLS_i := E_i'(W_i^{-1} - Q_i)^{-1}Q_i(W_i^{-1} - Q_i)^{-1}E_i/(p\hat{\phi}). \tag{11}$$

COROLLARY 2·2. *The effect of the tth observation in the ith cluster on the overall fit is*

$$DOBS_{it} := \frac{\tilde{E}_{it}^2\tilde{Q}_{it}}{p\hat{\phi}(W_{it}^{-1} - \tilde{Q}_{it})^2}.$$

The results in Theorem 2, Corollary 2·1 and Corollary 2·2 follow directly from Theorem 1, Corollary 1·1 and Corollary 1·2.

The Cook statistic for generalised linear models is given by McCullagh & Nelder (1989, p. 407) as in (9) but with a diagonal W, and is equivalent to $D_i = p^{-1}h_i(1-h_i)^{-1}r_{pi}^2$ (Williams, 1987). Kay & Little (1986) and Johnson (1985) have applied (9) to data fitted by logistic regression models in order to identify points having a large effect on the fitted probabilities. Under the identity link, but allowing dependence, the result in Corollary 2·2 is analogous to the Cook-type measure of Christensen et al. (1992); they refer to $\tilde{Q}_{it}W_{it}$ as the generalised leverage. For the multiple linear regression model, (10) reduces to the multiple-case-deletion Cook's distance (Cook & Weisberg, 1982, p. 136). For a single observation, this is equivalent to the original Cook's distance (Cook, 1977, 1979). A value of about 1·0 is generally considered large (Kleinbaum, Kupper & Muller, 1988, p. 201). In § 4, we consider the presence of correlation in interpreting (11) through an illustrative example.

An alternative measure to that of Theorem 2 can be achieved by scaling with the subset m deleted from the covariance matrix by letting $M = X_{[m]}'W_{[m]}X_{[m]}$ or $M = X_{[m]}'V_{[m]}^{-1}X_{[m]}$.

In general, these two choices are different because $V_{[m]}^{-1} = W_{[m]} - W_{[m]m} W_m^{-1} W_{m[m]}$. The exception is cluster deletion where $W_{[m]} = V_{[m]}^{-1}$, in which case, applying Corollary 1·1 and the relation

$$X'_{[i]} W_{[i]} X_{[i]} = X' W X - X'_i W_i X_i,$$

the studentised distance for the influence of the ith cluster on the overall fit is

$$MCLS_i := E'_i (W_i^{-1} - Q_i)^{-1} H_i E_i / (p\hat{\phi}). \tag{12}$$

Notice that $MCLS_i$ is expressed as the product of the cluster leverage and the squared residual scaled by $\mathrm{var}(E_i)$ as defined in § 2. The analogous measure for observation deletion depends on the choice of M and is a bit more complicated algebraically. For multiple linear regression, $D_m(X'_{[m]} X_{[m]}; 1)$ gives the multiple case-deletion diagnostic, $MDFIT_m$, and $D_i(X'_{[i]} X_{[i]}; 1)$ gives the single-case-deletion $DFIT_i$ of Belsley et al. (1980, p. 32 and p. 15, respectively).

In summary, diagnostics for assessing the influence of observations on the fitted values of the linear predictor may be scaled by the variance estimate of $\hat{\beta}$ based on all the observations as in (11) or on all but the subset of observations to be deleted as in (12). The former has the attraction that the comparison of distances between observations is meaningful because they refer to the same metric, as pointed out by Cook & Weisberg (1982) and Chatterjee & Hadi (1986) for linear regression. On the other hand, since the deleted case influences the estimate of the variance, its inclusion may decrease the magnitude of the diagnostic and, to some degree, may hide influence. Welsch (1986) prefers the studentised version, because of its 'robustness'. Cook (1986) points out, however, that the studentised diagnostic has a different interpretation than the standardised version. Thus the diagnostic given in (11), which is a generalisation of Cook's distance (1977), has the interpretation of the influence of a subset of observations on $\hat{\beta}$, whereas (12), which is a generalisation of *DFITS* of Belsley et al. (1980, p. 15), has the interpretation of the influence of a subset of observations on $\hat{\beta}$ and the variance estimate of $\hat{\beta}$ simultaneously. For this reason, the question 'influence on what?' should be the determining factor in choice of (11) or (12).

4. Illustration

Data from the North Carolina Early Cancer Detection Program at the Lineberger Comprehensive Cancer Center are used to illustrate the use of the diagnostics presented in § 3. Chart review data were collected from a random sample of $N = 3889$ medical charts in $K = 57$ medical practices. A practice constitutes a cluster. The number of charts per practice ranges from 19 to 197. The response $Y_{it} = 1$ if the tth chart in the ith practice indicates that the patient made at least one 'health maintenance visit' during the years 1990 and 1991, and $Y_{it} = 0$ otherwise. An exchangeable correlation structure is assumed: $\mathrm{corr}(Y_{it}, Y_{it'}) = \rho$ for $i = 1, \ldots, K$, $t \neq t' = 1, \ldots, n_i$. We specify a logit link in (1) and consider the covariates in Table 1.

Table 2(a) shows parameter and standard error estimates. Because of large cluster sizes no single observation had a large influence, so we present cluster-deletion diagnostics only. Exact cluster-deletion diagnostics were obtained by iterating the fitting algorithm, including estimation of ρ, to convergence.

Figure 1 shows plots of one-step approximations versus their exact counterparts. Figures 1(a) and 1(b) show *DBETAC* for M3 and SPECLTY, respectively, while Fig. 1(c) shows the

Table 1. *Covariates for North Carolina Early Cancer Detection Program illustration*

INTERCEPT	
SPECLTY	Doctor's specialty: 0 if family or general practice, 1 if internal medicine
MDAGE	(Doctor's age in years − 45)/10
MDSEX	Doctor's sex: 0 if male, 1 if female
PATAGE	(Patient's age in years − 65)/10
NOINSUR	Health insurance: 0 if insured, 1 if not insured
NBRMDS	Number of doctors in practice − 1
M3	(Number of patients over 50 years old seen per day − 15)/10
MDFLU	Doctor's flu vaccination: 0 if in the last two years, 1 if 3 to 5 years ago, 2 if never
MALEPAT	Patient's sex: 0 if female, 1 if male
BLACKPAT	Patient's race: 0 if white, 1 if black

Table 2. *Parameter estimates and naive standard errors with and without cluster 5 for North Carolina Early Cancer Detection Program illustration*

	(a) *All data*			(b) *Cluster 5 deleted*		
	$\hat{\beta}_j$	SE	Z	$\hat{\beta}_j$	SE	Z
INTERCEPT	−0·044	0·164	−0·269	0·045	0.174	0·258
SPECLTY	−0·475	0·152	−3·133	−0·597	0·163	−3·670
MDAGE	−0·311	0·055	−5·631	−0·281	0·061	−4·606
MDSEX	0·533	0·154	3·451	0·350	0·158	2·213
PATAGE	−0·104	0·031	−3·383	−0·103	0·031	−3·307
NOINSUR	−0·447	0·102	−4·386	−0·461	0·103	−4·475
NBRMDS	0·025	0·059	0·431	0·037	0·066	0·562
M3	0·480	0·090	5·316	0·214	0·111	1·932
MDFLU	0·001	0·061	0·022	−0·034	0·065	−0·530
MALEPAT	−0·400	0·067	−5·959	−0·424	0·068	−6·245
BLACKPAT	−0·441	0·126	−3·490	−0·369	0·133	−2·778

(a) Based on all data: $N = 3889$ patients, $K = 57$ medical practices, $\hat{\rho} = 0{\cdot}154$.
(b) Cluster 5 deleted: $N = 3698$ patients, $K = 56$ medical practices, $\hat{\rho} = 0{\cdot}171$.

Cook's distance, or *DCLS*. First, the plots show that the one-step approximations given by (5) are in good agreement with the exact diagnostics. This was the case for the other coefficients, and with respect to (6), as well. Secondly, cluster 5 has a large influence on the slope for M3; and clusters 19 and 29 have a large influence on the slope for SPECLTY. Figure 1(c) shows that clusters 5, 15, 19 and 29 had the largest influence on the fitted values; the log scale is used for clarity.

Figure 2(a) shows a plot of *DCLS* versus cluster size. Clearly, clusters 5, 15 and 29 have a large influence relative to their size. Figure 2(b) shows that cluster 29 had less leverage than clusters 15 and 19 but about the same influence, indicating that it had large residuals; that is it was poorly fitted. Figure 3 shows the relationship of leverage to cluster size; cluster 5 had the largest leverage and cluster 48, although not influential with respect to (5), had a large leverage relative to its size, which was the smallest of all clusters.

Estimates obtained after deleting cluster 5 are shown in Table 2(b). The estimates for M3 changed by 2·94 standard errors, while for MDSEX it changed by 1·18 standard errors. The one-step approximations, (8), shown in Table 3, are 3·21 and 1·15, respectively. Table 3 also shows important summary information for the most influential clusters; only clusters 5, 19, 15 and 29 have at least one value of (8) greater than 1·0 in absolute terms.

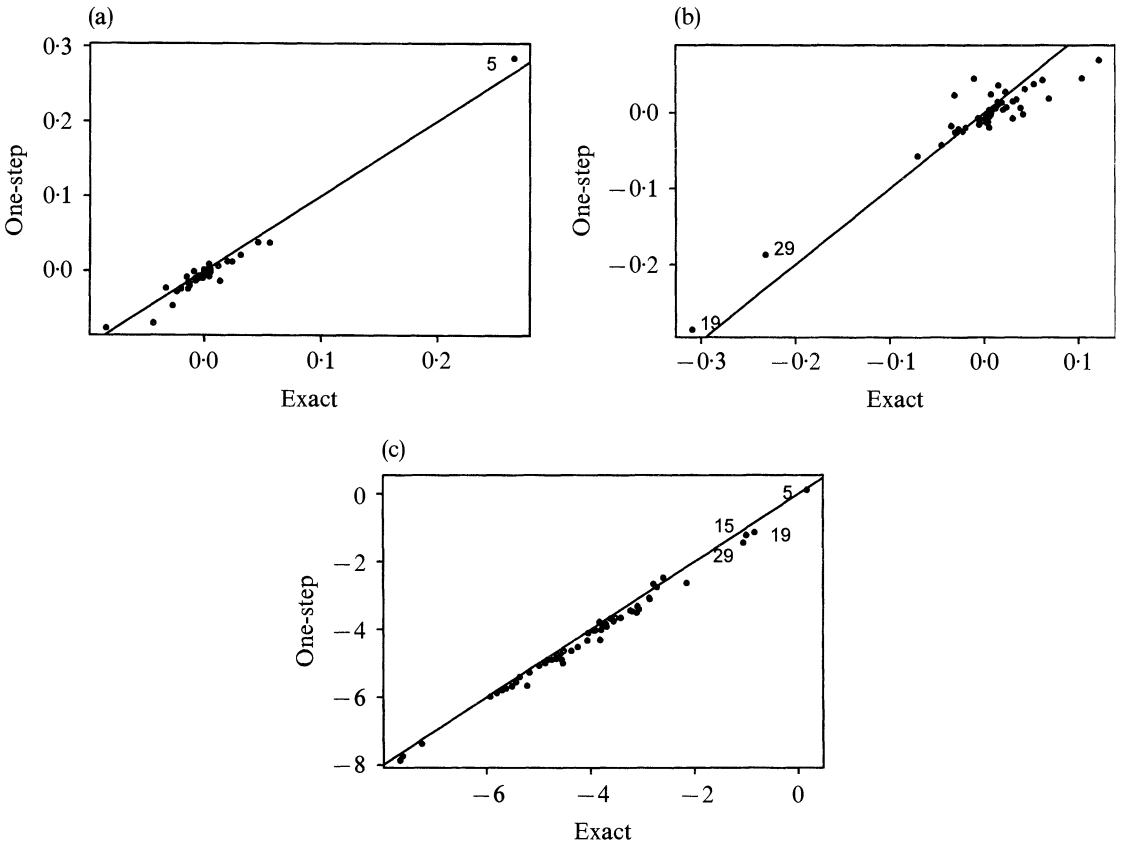

Fig. 1. One-step approximation versus (a) exact change in estimated coefficient for M3, (b) exact change in estimated coefficient for SPECLTY, (c) exact value of log of Cook's distance.

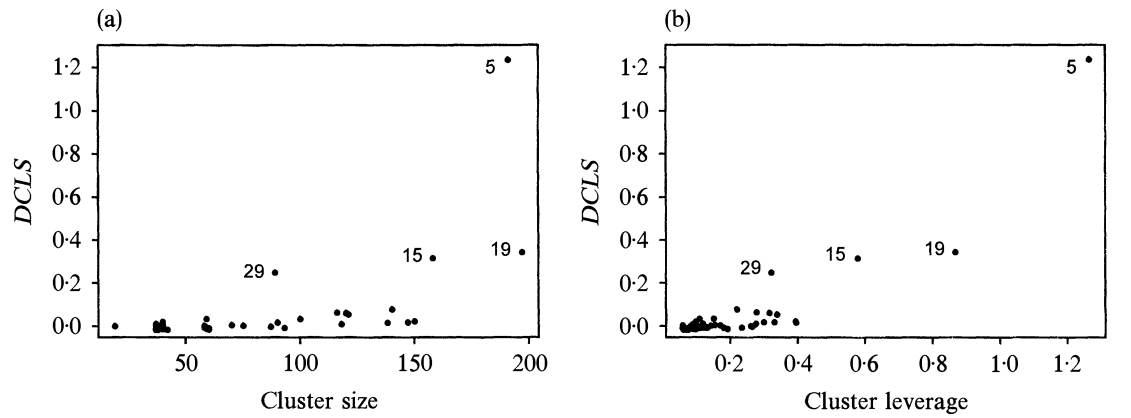

Fig. 2. Cook's distance versus (a) cluster size, (b) cluster leverage.

5. Discussion

In this paper, one-step deletion diagnostics were introduced which identify influential data in the generalised estimating equations procedure. The diagnostics are good approximations of their exact counterparts, and their computation is fast, so that they may be

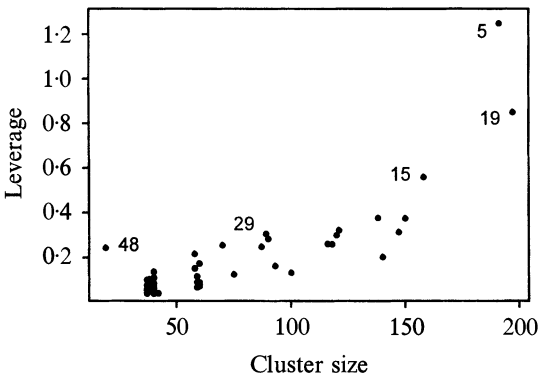

Fig. 3. Cluster leverage versus cluster size.

Table 3. *Cluster size, leverage and influence diagnostics for selected clusters and covariates*

					$DBETACS_{ij}$		
Cluster	n_i	$\mathrm{tr}(H_i)$	$DCLS_i$	SPECLTY	MDSEX	M3	MDFLU
5	191	1·26	1·25	0·49	1·15	3·21	0·74
19	197	0·87	0·36	−1·85	−0·76	0·29	0·27
15	158	0·58	0·33	0·33	−0·37	−0·78	1·09
29	89	0·32	0·27	−1·20	1·07	0·47	0·23
48	19	0·26	0·02	−0·01	−0·16	−0·20	−0·19

routinely used in data analysis. In particular, for the example in § 4, (5) and (6) were approximately 25 times faster to compute than the exact quantities. Typically, diagnostics may identify faulty data which are subsequently removed from the analysis, otherwise the influence is tolerated. For the medical practice data in § 4, an intermediate solution of downweighting influential data using a robust regression may be better than removing 5% of the data corresponding to cluster 5. Such a procedure, like that of Pregibon (1982) but for correlated outcomes, will be pursued elsewhere.

ACKNOWLEDGEMENT

The work of J. S. Preisser was partially supported by a National Institute of Environmental Health Sciences Training Grant. We are grateful to an associate editor and referee whose comments improved the paper and to the North Carolina Early Cancer Detection Program and the Lineberger Comprehensive Cancer Center for allowing the use of their data. The North Carolina Early Cancer Detection Program is funded by the National Cancer Institute.

APPENDIX

Proof of Theorem 1

The proof is obtained by establishing three results for matrices which generalise those in the Appendix of Christensen et al. (1992) to models with a general link function and to influence of multiple observations.

LEMMA 1. *We have*

$$X'_{[m]}V^{-1}_{[m]}X_{[m]} = X'WX - \tilde{X}'_m W_m \tilde{X}_m.$$

Proof. The proof follows upon writing

$$X'WX = (X'_m, X'_{[m]})V^{-1}\begin{pmatrix} X_m \\ X_{[m]} \end{pmatrix}$$

and applying the usual formula for the inverse of a partitioned matrix (Searle, 1982, p. 260, eqn (14)). □

LEMMA 2. *We have*

$$(X'_{[m]}V^{-1}_{[m]}X_{[m]})^{-1} = (X'WX)^{-1} + (X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{X}_m(X'WX)^{-1}.$$

Proof. The proof follows directly by applying the inverse formula (Searle, 1982, p. 261, eqn (17)) to Lemma 1. □

LEMMA 3. *We have*

$$X'_{[m]}V^{-1}_{[m]}Z_{[m]} = X'WZ - \tilde{X}'_m W_m \tilde{Z}_m.$$

Proof. The proof is similar to that of Lemma 1. □

Theorem 1 follows on applying Lemmas 2 and 3 to show

$$\begin{aligned}\hat{\beta}_{[m]} &= (X'_{[m]}V^{-1}_{[m]}X_{[m]})^{-1}X'_{[m]}V^{-1}_{[m]}Z_{[m]} \\ &= \hat{\beta} + (X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{X}_m\hat{\beta} - (X'WX)^{-1}\tilde{X}'_m W_m \tilde{Z}_m \\ &\quad - (X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}\tilde{Q}_m W_m \tilde{Z}_m.\end{aligned}$$

Rewrite the third term in the last expression as

$$(X'WX)^{-1}\tilde{X}'_m(W_m^{-1} - \tilde{Q}_m)^{-1}(W_m^{-1} - \tilde{Q}_m)W_m\tilde{Z}_m.$$

Straightforward matrix algebra gives the result. □

REFERENCES

ATKINSON, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Oxford: Clarendon Press.

BARRETT, B. E. & LING, R. F. (1992). General classes of influence measures for multivariate regression. *J. Am. Statist. Assoc.* **87**, 184–91.

BELSLEY, D. A., KUH, E. & WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley.

CHATTERJEE, S. & HADI, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression (with Discussion). *Statist. Sci.* **1**, 379–416.

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

COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15–8.

COOK, R. D. (1979). Influential observations in linear regression. *J. Am. Statist. Assoc.* **74**, 169–74.

COOK, R. D. (1986). Discussion of paper by S. Chatterjee and A. S. Hadi. *Statist. Sci.* **1**, 393–7.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

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

KAY, R. & LITTLE, S. (1986). Assessing the fit of the logistic model: A case study of children with the haemolytic uraemic syndrome. *Appl. Statist.* **35**, 16–30.

KLEINBAUM, D. G., KUPPER, L. L. & MULLER, K. E. (1988). *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. Boston, MA: PWS-Kent.

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

McCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.
PREGIBON, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705–24.
PREGIBON, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrika* **38**, 485–98.
SEARLE, S. R. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley.
WILLIAMS, D. A. (1987). Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.* **36**, 181–91.
WELSCH, R. E. (1986). Discussion of paper by S. Chatterjee and A. S. Hadi. *Statist. Sci.* **1**, 403–5.

[*Received August* 1994. *Revised November* 1995]