

Resistant fits for regression with correlated outcomes an estimating equations approach

Bahjat F. Qaqish^{a,*}, John S. Preisser^b

^a *Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, NC 27599-7400, USA*

^b *Section on Biostatistics, Wake Forest University School of Medicine, Medical Center Blvd., Winston-Salem, NC 27157-1063, USA*

Abstract

The generalized estimating equations procedure of Liang and Zeger (1986) can be highly influenced by the presence of unusual data points. A generalization is introduced which yields parameter estimates and fitted values resistant to influential data. A diagonal weight matrix for each cluster is incorporated into the estimating equations which downweights the multivariate response vector element-wise. Efficiency of the procedure is investigated, including the case of correlated binary outcomes. © 1999 Published by Elsevier Science B.V. All rights reserved.

AMS classification: primary 62J12; secondary 62F35

Keywords: Cluster-downweighting; Mallows class; Observation-downweighting; Schweppe class; Resistant

1. Introduction

The generalized estimating equations procedure of Liang and Zeger (1986) is often applied to longitudinal data or to data that are naturally grouped into clusters such that observations within the cluster are correlated, but those from different clusters are assumed to be uncorrelated. Any analysis of clustered data should consider the influence that observations or entire clusters may have on the overall results. Preisser and Qaqish (1999), showed that the parameter estimates from an analysis based on the generalized estimating equations procedure may be highly influenced by a small subset of the data. A resistant fit, on the other hand, is one which is not sensitive to large changes in a few observations (Pregibon, 1982). We modify the estimating equation of Liang and Zeger (1986) by generalizing ideas from Carroll and Pederson (1993) who provide robust estimates in the logistic regression model that are of the

* Corresponding author. E-mail: qaqish@bios.unc.edu.

Mallows class. Estimates of the Mallows class are obtained by downweighting observations with large leverage values. Alternatively, Schweppe class estimates are obtained by downweighting observations with large residuals as in Pregibon (1982) and Künsch et al. (1989). Within both classes, we consider two approaches which we call observation-downweighting and cluster-downweighting. The former downweights individual observations separately, whereas the latter method assigns equal weight to all observations in a cluster based on some aggregate measure of the influence of the entire cluster. Singer and Sen (1985) and Huggins (1993) have proposed robust multivariate methods for continuous responses. Our methods have wider applicability in that they apply to the same class of models for which generalized estimating equations are applied.

Section 2 reviews generalized estimating equations. Section 3 presents a modification of generalized estimating equations which we refer to as resistant generalized estimating equations. Theoretical foundations for resistant generalized estimating equations are presented in Section 4 in which we show the necessary sacrifice of efficiency to obtain robustness. In Section 5, efficiency of resistant generalized estimating equations relative to generalized estimating equations is explored. In Section 6, the special case of correlated binary responses is discussed, including discussion of asymptotic relative efficiency. Concluding remarks are made in the final section, including a discussion of the potential and limitations of resistant generalized estimating equations for other types of outcomes.

2. Generalized estimating equations

To describe the general set-up, let $Y_i \equiv (Y_{i1}, \dots, Y_{i n_i})'$ be a n_i -vector of outcome values for $i = 1, \dots, K$, and $X_i \equiv (x'_{i1}, \dots, x'_{i n_i})'$ is a $n_i \times p$ matrix of covariate values. The class of applicable models are those in which the forms of the first two moments for the marginal distribution of Y_{it} are

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

where i indexes clusters and t indexes observations. In the terminology of generalized linear models (McCullagh and Nelder, 1989, ch. 2), $g(\cdot)$ is the link function which determines the relationship of the mean with the linear predictor η_{it} , $v(\cdot)$ is 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 generalized estimating equations

$$\sum_{i=1}^K D'_i(X_i, \beta) V_i^{-1}(x, \beta) (Y_i - \mu_i(\beta)) = 0, \quad (2)$$

where $D_i \equiv \partial \mu_i / \partial \beta$ is an $n_i \times p$ matrix, $V_i \equiv A_i R_i(\alpha) A_i$, and $A_i = \text{diag}\{v^{1/2}(\mu_{it})\}$ is a $n_i \times n_i$ diagonal matrix. Furthermore, $R_i(\alpha)$ is a $n_i \times n_i$ working correlation matrix that depends on unknown parameter vector α . Solutions to Eq. (2) are obtained by

alternating between estimation of ϕ , α , and β . Liang and Zeger (1986) give consistent estimates of ϕ and α based on squares and cross products of Pearson residuals $r_{it} \equiv (y_{it} - \mu_{it}) / \{v(\mu_{it})\}^{1/2}$. Let $N \equiv \sum n_i$ and define the $N \times 1$ vector $Y \equiv (Y'_1, \dots, Y'_k)'$, the $N \times p$ matrix $X \equiv (X'_1, \dots, X'_k)'$ and the $N \times N$ block diagonal matrix D^* with blocks $D_i^* \equiv \text{diag}\{\partial \eta_{it} / \partial \mu_{it}\}$. Solving for β is done with iteratively reweighted least squares. A current estimate $\hat{\beta}_G$ is updated by regressing the working response vector $Z^* \equiv X\hat{\beta} + D^*(Y - \hat{\mu})$ on X with block diagonal weight matrix W^* whose i th block, corresponding to the i th cluster, is the $n_i \times n_i$ matrix $W_i^* \equiv D_i^{*-1} A_i^{-1} R_i^{-1}(\hat{\alpha}) A_i^{-1} D_i^{*-1}$. A new estimate is obtained by $\hat{\beta}_{\text{new}} = (X'W^*X)^{-1} X'W^*Z^*$, evaluating the right-hand side at the current estimate. Then $\hat{\beta}_{\text{new}}$ is used to update $\hat{\eta} \equiv X\hat{\beta}_{\text{new}} = HZ^*$, where $H \equiv QW^*$ and $Q \equiv X(X'W^*X)^{-1} X'$. The projection matrix, H , maps the current value of Z^* into estimated values of the linear predictor. As in multiple linear regression, the diagonal elements of H , denoted h_{it} , correspond to the amount of leverage of the response on the corresponding fitted value. Under independence, these observation leverages for generalized estimating equations reduce to the leverages of observations in a generalized linear model as given by Williams (1987). 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 i th block along the diagonal in H , and is given by $H_i = Q_i W_i^*$ where $Q_i = X_i(X'W^*X)^{-1} X'_i$ and may be summarized by $\text{tr}(H_i)$ which equals the sum of the observation leverages h_{it} in the cluster.

Liang and Zeger (1986) show that, for large K , $K^{1/2}(\hat{\beta}_G - \beta)$ is asymptotically multivariate Gaussian with zero mean. The variance of $\hat{\beta}_G$ can be estimated consistently by the ‘robust’ or sandwich variance estimate,

$$\left(\sum_{i=1}^K D'_i V_i^{-1} D_i \right)^{-1} \left\{ \sum_{i=1}^K D'_i V_i^{-1} (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' V_i^{-1} D_i \right\} \times \left(\sum_{i=1}^K D'_i V_i^{-1} D_i \right)^{-1}, \tag{3}$$

where β , ϕ , and α are replaced by their estimates. It is robust in the sense that it consistently estimates the variance of $\hat{\beta}_G$ even if $R(\alpha)$ is misspecified.

3. Resistant generalized estimating equations

3.1. General

In order to provide robust estimation of the broad class of models considered by Liang and Zeger (1986), we introduce a diagonal weight matrix $W_i \equiv W_i(X_i, Y_i, \alpha, \beta, \phi)$ and define resistant generalized estimating equations as

$$\sum_{i=1}^K D'_i(X_i, \beta) V_i^{-1}(\alpha, \beta) [\psi_i - c_i] = 0, \tag{4}$$

where D_i and V_i are defined as in Eq. (2) and μ_i is parametrized as in Eq. (1), $\psi_i \equiv W_i(Y_i - \mu_i)$ and $c_i = E[\psi_i]$. The generalized estimating equations given by Eq. (2) are a special case of Eq. (4) in which $W_i = I$ and $c_i = 0$. In general, however, W_i is a diagonal matrix for the i th cluster that contains weights, w_{it} , $t = 1, \dots, n_i$, that correspond to the elements of the response vector, Y_i . Downweighting may be done based on the covariates only, the so-called Mallows class, or on the responses as well, the Schweppe class. For the Mallows class the weights W_i are non-random and the l.h.s. of Eq. (4) remains unbiased with $c_i \equiv 0$. For the Schweppe class c_i is determined so that the l.h.s. of Eq. (4) is an unbiased estimating function. The w_{it} are between 0 and 1 and are analogous to the ψ -functions in M -estimation (Hampel et al., 1986) in that they determine the robustness and efficiency of $\hat{\beta}_R$. Most observations will have a weight near 1, but those observations which are determined to have large influence on the estimation of β will receive a smaller weight. The following theorem gives asymptotic results for Eq. (4) under the model in Eq. (1).

Theorem. Assume that:

- (i) $\hat{\alpha}$ is $K^{1/2}$ – consistent given β and ϕ ;
- (ii) $\hat{\phi}$ is $K^{1/2}$ – consistent given β ;
- (iii) $\text{Var}(\psi_i) < \infty$;
- (iv) ψ_i is absolutely continuous in β such that the derivative with respect to μ_i , denoted $\dot{\psi}_i$, exists, and $E\|\dot{\psi}_i\| < \infty$;

Under additional regularity conditions, $K^{1/2}(\hat{\beta}_R - \beta)$ is asymptotically Gaussian with zero mean and covariance matrix V_R given by

$$\lim_{K \rightarrow \infty} K \left(\sum_{i=1}^K D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \left\{ \sum_{i=1}^K D_i' V_i^{-1} \text{Var}(\psi_i) V_i^{-1} D_i \right\} \times \left\{ \left(\sum_{i=1}^K D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \right\}' \tag{5}$$

where

$$\Gamma_i = E\dot{\psi}_{ki} - \dot{c}_i, \quad \dot{\psi}_{ki} = \frac{\partial}{\partial \mu_i} \psi_i(\mu_i) \quad \text{and} \quad \dot{c}_i = \frac{\partial}{\partial \mu_i} c_i.$$

The additional regularity conditions are that derivatives, $\partial \hat{\alpha}(\beta, \phi) / \partial \phi$, $\partial \hat{\alpha}\{\beta, \hat{\phi}(\beta)\} / \partial \beta$, and $\partial \hat{\phi}(\beta) / \partial \beta$, are bounded in variation. A sketch of the proof is given in the appendix. The Theorem implies that use of Eq. (4) will result in some loss of efficiency under model (1).

Estimation of β is done with iteratively reweighted least squares by regressing the working response vector $Z = X\hat{\beta} + D^*(\psi - c)$ on X with the original weight matrix

W^* from generalized estimating equations. The variance of $\hat{\beta}_R$ is estimated by

$$\left(\sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \left\{ \sum_{i=1}^k D_i' V_i^{-1} (\psi_i - c_i) (\psi_i - c_i)' V_i^{-1} D_i \right\} \times \left\{ \left(\sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i \right)^{-1} \right\}' \tag{6}$$

evaluated at $\hat{\beta}_R$, $\hat{\psi}$ and $\hat{\alpha}$. The ‘robust’ variance estimator, Eq. (3), is obtained by setting $W_i = I$ in Eq. (6).

3.2. Classes of estimates

Estimates of the Mallows observation-downweighting class are obtained by specifying weights $w_{it} = w_{it}(h_{it})$ which are updated at each iteration. An alternative way of measuring leverage in which weights are calculated only once is given by Carroll and Pederson (1993) in the logistic regression case. In the Schweppe observation-downweighting class, observations are downweighted according to their residual, and possibly, their leverage as well by specifying $w_{it} \equiv w_{it}(x_{it}, X, y_{it}, \alpha, \beta)$. In this case, c_i is determined by the individual marginal distributions of Y_{it} , so that Eq. (4) yields consistent estimates of β under the Theorem above. In particular, we consider weights $w(r_{it}/\sqrt{\hat{\phi}})$ that are a function of the scaled Pearson residual.

In both Mallows and Schweppe classes, it is possible to downweight clusters instead of observations, by assigning all the observations in a cluster equal weight. Mallows cluster-downweighting may be achieved by assigning $w_{it} = w(\text{tr}(H_i))$. For Schweppe cluster-downweighting, one possibility is to summarize the lack of fit of the observations in the cluster by downweighting clusters as a function of $r_i' R^{-1} r_i / (\hat{\phi} n_i)$ where $r_i = (r_{i1}, \dots, r_{in_i})'$.

The following remarks can be made about the different classes of resistant generalized estimating equations:

(1) The Mallows class, where the weights are nonrandom, does not require additional assumptions about the univariate marginal distributions beyond Eq. (1) and applies to any situation that generalized estimating equations might be used, including modelling binary, count, and continuous responses. The Mallows class, either observation-downweighting or cluster-downweighting, is a special case of the Theorem in which $\Gamma_i = W_i$ and $\text{Var}(\psi_i) = W_i \text{Var}(Y_i) W_i$.

(2) The Schweppe observation-downweighting class generally requires full specification of the marginal univariate distributions for estimation of β and ϕ . This applies to independent responses too (Morgenthaler, 1992); an exception is the location/scale family of distributions.

(3) In the Schweppe cluster downweighting class, the weights depend on the full response vector for the cluster and, thus, the full multivariate distribution is required in order to calculate the debiasing factor, c_i , in Eq. (4). Unfortunately, multivariate generalizations of the Poisson and Gamma distributions and many other distributions

in the exponential family present two limitations. First, the debiasing factor, c_i , in Eq. (4) is difficult to compute and generally does not have a closed form expression. Second, additional parameters beyond β , α , and ϕ are required for calculation of c_i . Notable exceptions are the bivariate binary distribution and the multivariate normal distribution which are completely specified by β , α , and ϕ . However, even in the latter case, the calculation of c_i may be formidable.

4. Optimality considerations

The theory of estimating functions provides further perspective on the asymptotic properties of resistant generalized estimating equations. Consider the non-robust elementary estimating functions, $g_{it} = Y_{it} - \mu_{it}(\beta)$, $t = 1, \dots, n_i$, and $i = 1, \dots, K$. The theory of Godambe and Heyde (1987) and Morton (1981) states that the optimal estimating function is given by

$$EF = \sum_{i=1}^K C_i' V_i^{-1} g_i,$$

where $g_i = (g_{i1}, \dots, g_{in_i})'$, $C_i = E_{\beta}[\partial g_i / \partial \beta]$ and $V_i = \text{Var}(g_i)$. The solution, $\hat{\beta}$ of the estimating equation $EF = 0$ is optimal in the sense that it has the smallest asymptotic variance among all estimating functions that are linear combinations of the elementary estimating functions. Its variance is said to be smaller if $\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ is positive definite for all possible solutions $\tilde{\beta}$ in the class of estimating functions considered. It follows that Mallows resistant generalized estimating equations is suboptimal because it is in the same class of estimating functions as generalized estimating equations, i.e., those that can be expressed as linear functions of Y_i , $i = 1, \dots, K$. The question of Schweppe resistant generalized estimating equations is not as straightforward because these functions are not linear in Y_i . Intuitively, one might expect a loss in efficiency for a gain in robustness. Indeed, this is the case for Schweppe observation-downweighting, applied to correlated binary responses. The robust elementary estimating function is $g_{it}^* = w_{it}(r_{it}, a)(Y_{it} - \mu_{it}) - c_{it}$, where w is a smooth weight function with tuning constant a , r_{it} is the Pearson residual and $c_{it} = E[w_{it}(r_{it}, a)(Y_{it} - \mu_{it})]$. The optimal estimating function based on g_i^* is

$$EF^* = \sum_{i=1}^K C_i^* (V_i^*)^{-1} g_i^*,$$

where $g_i^* = (g_{i1}^*, \dots, g_{in_i}^*)'$, $C_i^* = E_{\beta}[\partial g_i^* / \partial \beta]$ and $V_i^* = \text{Var}(g_i^*)$. It is easy to show that for binary responses $EF^* = EF$. Thus, in optimizing the efficiencies of the robust elementary estimating function g_i^* , the robustness is lost.

Eq. (4) is essentially of the form

$$\sum_{i=1}^K C_i' V_i^{-1} g_i^*,$$

which is necessarily less efficient than Eq. (2). Robustness can generally be gained at the cost of efficiency, both of which are determined by w and a in resistant generalized estimating equations.

5. Efficiency considerations

5.1. General

The resistant generalized estimating equations procedure provides protection in the form of robustness against observations or clusters which deviate from the model. When all the data follow (1), i.e., when the model is not contaminated, the discussion in the previous section suggests that resistant generalized estimating equations will be less efficient than generalized estimating equations. This section defines and examines the asymptotic relative efficiency of resistant generalized estimating equations to generalized estimating equations. The efficiency loss of resistant generalized estimating equations is given by the generalized asymptotic relative efficiency of $\hat{\beta}_R$ to $\hat{\beta}_G$ which is defined to be

$$\text{ARE}_{R:G} := |\text{var}(\hat{\beta}_G)\{\text{var}(\hat{\beta}_R)\}^{-1}|^{1/p} = \left(\prod_{j=1}^p \lambda_j \right)^{1/p}, \tag{7}$$

where λ_j is the j th eigenvalue of $\text{var}(\hat{\beta}_G)\{\text{var}(\hat{\beta}_R)\}^{-1}$. Because Eq. (7) is generally complicated, some insight is gained by restricting to equal cluster sizes $n_i = n$, and exchangeable correlation matrix

$$R_i = [1 + (n - 1)\rho] \frac{1}{n} J + (1 - \rho) \left(I - \frac{1}{n} J \right),$$

where $J = 11'$ is a $n \times n$ matrix of ones, with inverse

$$R_i^{-1} = [1 + (n - 1)\rho]^{-1} \frac{1}{n} J + (1 - \rho)^{-1} \left(I - \frac{1}{n} J \right). \tag{8}$$

Let L_i denote the diagonal matrix of iterative weights, $l_{ii} = (\partial \mu_{ii} / \partial \eta_{ii}) \{v_{ii}(\mu_{ii})\}^{-1/2}$. The asymptotic variance matrix of $\hat{\beta}_G$ is

$$\text{var}(\hat{\beta}_G) = \left(\sum_{i=1}^K D_i' V_i^{-1} D_i \right)^{-1} \phi = \left(\sum_{i=1}^K X_i' L_i R_i^{-1} L_i X_i \right)^{-1} \phi. \tag{9}$$

The iterative weights will be constant, $l_{ii} = l$ for the identity link function with constant variance and more generally for generalized linear models with link function equal to the variance stabilizing transformation (McCullagh and Nelder, 1989). In that case Eq. (9) can be written as

$$\text{var}(\hat{\beta}_G) = l^{-2} (e_n^{-1} \text{SS}_{uc} + f^{-1} \text{SS}_{wc})^{-1} \phi, \tag{10}$$

where $e_n := 1 + (n - 1)\rho$, $f := 1 - \rho$, and SS_{uc} and SS_{wc} are the between-cluster (uncorrected) and within cluster sum of squares and cross products matrices of x . Specifically,

$$SS_{uc} := \frac{1}{n} \sum_{i=1}^K X_i' J X_i, \quad SS_{wc} := \sum_{i=1}^K \left(X_i' X_i - \frac{1}{n} X_i' J X_i \right).$$

Mancl and Leroux (1996) used a similar setup to evaluate the loss of efficiency of generalized estimating equations due to misspecification of the correlation matrix. If $\text{var}(\psi_i) = \Gamma_i \text{var}(Y_i) \Gamma_i$, the asymptotic variance matrix of \hat{B}_R is

$$\text{var}(\hat{\beta}_R) = H_1^{-1} H_2 \{H_1^{-1}\}^T \phi, \tag{11}$$

$$H_1 = \sum_{i=1}^K D_i' V_i^{-1} \Gamma_i D_i \quad \text{and} \quad H_2 = \left(\sum_{i=1}^K D_i' V_i^{-1} \Gamma_i V_i \Gamma_i V_i^{-1} D_i \right).$$

In the next section, the roles of n , ρ and X are considered for special cases.

5.2. Cluster-level covariates

If all covariates are cluster level, $\Gamma_i = b_i I$, where b_i is a scalar quantity. This follows for Mallows cluster-downweighting because all observations receive equal weight by definition. For observation downweighting, Mallows or Scheppe, $b_{it} = b_i$ follows from $\mu_{it} = \mu_i$. Cluster level covariates imply $l_{i1} = l_{i2} = l_{im_i} := l_i$, and $b_{i1} = b_{i2} = b_{im_i} := b_i$ because $\mu_{i1} = \mu_{i2} = \mu_{im_i}$, and in Eq. (11),

$$H_1 = \sum_{i=1}^K b_i l_i^2 X_i' R_i^{-1} X_i = e_n^{-1} SS_{uc}^{(12)},$$

$$H_2 = \sum_{i=1}^k b_i^2 l_i^2 X_i' R_i^{-1} X_i = e_n^{-1} SS_{uc}^{(22)},$$

where

$$SS_{uc}^{(12)} = \frac{1}{n} \sum_{i=1}^K b_i l_i^2 X_i' J X_i, \quad SS_{uc}^{(22)} = \frac{1}{n} \sum_{i=1}^K b_i^2 l_i^2 X_i' J X_i.$$

Hence

$$\text{ARE}_{R:G} = |SS_{uc}^{(02)-1} SS_{uc}^{(12)} SS_{uc}^{(22)-1} SS_{uc}^{(12)}|^{1/p}, \tag{12}$$

where

$$SS_{uc}^{(02)} = \sum_{i=1}^K \frac{1}{n} l_i^2 X_i' J X_i.$$

In summary, if all the covariates are cluster level, the efficiency is a function of the resistant generalized estimating equations weights, but not of the common cluster size and dependence upon the correlation is only through these weights.

5.3. Within-cluster covariates

In general, when covariates vary within cluster, Eq. (11) is complicated and depends upon the correlation and the common cluster size. A special case is for constant iterative weights, i.e., $l_{it} = l$ in the context of Mallows cluster-downweighting which implies $\Gamma_i = b_i I$. Then,

$$H_1 = l^2 \sum_{i=1}^K b_i X_i' R_i^{-1} X_i = l^2 (e_n^{-1} SS_{uc}^{(10)} + f^{-1} SS_{wc}^{(10)})$$

$$H_2 = l^2 \sum_{i=1}^K b_i^2 X_i' R_i^{-1} X_i = l^2 (e_n^{-1} SS_{uc}^{(20)} + f^{-1} SS_{wc}^{(20)}),$$

where

$$SS_{uc}^{(10)} = \frac{1}{n} \sum_{i=1}^K b_i X_i' J X_i, \quad SS_{uc}^{(20)} = \frac{1}{n} \sum_{i=1}^K b_i^2 X_i' J X_i,$$

$$SS_{uc}^{(10)} = \sum_{i=1}^K \left(b_i X_i' X_i - \frac{1}{n} b_i X_i' J X_i \right) \quad \text{and} \quad SS_{wc}^{(20)} = \sum_{i=1}^K \left(b_i^2 X_i' X_i - \frac{1}{n} b_i^2 X_i' J X_i \right).$$

For nonconstant iterative weights, the matrices resulting from reexpressing Eqs. (10) and (11) are difficult to interpret because they depend not only on the design, but also on the values of ρ and β . When cluster sizes vary, the asymptotic variance of $\hat{\beta}_G$ will, generally, depend on the correlation, average cluster size, and coefficient of variation of the cluster sizes. Mancl and Leroux (1996) have similar findings in a different context.

The efficiency of parameter estimates excluding the intercept is often of interest, instead of Eq. (7). An expression like Eq. (10) is given by Mancl and Leroux (1996) with SS_{uc} replaced by

$$\sum_{i=1}^K \frac{1}{n} X_i' J X_i - \frac{1}{nK} \left(\sum_{j=1}^K X_j' 1 \right) \left(\sum_{l=1}^K 1' X_l \right).$$

This result is obtained by applying a standard matrix algebra result for the inverse of a partitioned matrix. For resistant generalized estimating equations, however, the analogous formula for Eq. (12) is algebraically complex. As illustrated in the next section, however, the conclusions drawn for the entire regression parameter appear to apply to the parameter vector which omits the intercept.

6. Resistant generalized estimating equations for correlated binary outcomes

6.1. General

The Scheppe observation-downweighting theory is easily applied to correlated binary outcomes because the marginal distributions and the bivariate distributions depend only on α and β . Consider a cluster of arbitrary size, and, without loss of generality,

consider the first two elements in the response vector. Define $\pi_{jk} = \Pr(Y_1 = j, Y_2 = k)$. The bivariate distribution is a multinomial distribution with cell probabilities $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ which are completely determined by the marginal means $\mu_1 = \Pr(Y_1 = 1)$, $\mu_2 = \Pr(Y_2 = 1)$ and the correlation between Y_1 and Y_2 denoted by ρ_{12} . Specifically, $\pi_{11} = \mu_1\mu_2 + \rho_{12}v_1^{1/2}v_2^{1/2}$, $\pi_{10} = \mu_1 - \pi_{11}$, $\pi_{01} = \mu_2 - \pi_{11}$, and $\pi_{00} = 1 - \mu_1 - \mu_2 + \pi_{11}$. The variance function is $v_t = v(\mu_t) = \mu_t(1 - \mu_t)$, $t = 1, 2$, and $\phi = 1$.

Let the observation weight w_t be a function of the corresponding residual r_t . To solve the Schweppe observation-downweighting resistant generalized estimating equations, we need to find c_t from the marginal Bernoulli distribution. It can be shown that $c_t = v_t(w_t^{(1)} - w_t^{(0)})$, where $w_t^{(j)}$ is the weight for the t th observation in the cluster evaluated at $y_t = j$. Now, because each w_t is a function of its corresponding residual and not the entire residual vector for the cluster, it follows that $\dot{\psi}$, and thus Γ are diagonal matrices. It can be shown that $\Gamma = \text{Diag}\{-b_t\}$, $\text{Var}(\psi_t) = v_t b_t^2$, and $\text{Cov}(\psi_t, \psi_{t'}) = \rho_{tt'}v_t^{1/2}v_{t'}^{1/2}b_t b_{t'}$, where $b_t = (1 - \mu_t)w_t^{(1)} + \mu_t w_t^{(0)}$, and $\rho_{tt'}$ is the correlation between Y_t and $Y_{t'}$.

Schweppe cluster-downweighting for correlated binary data is difficult to implement beyond clusters of size two because of the complexity of the multivariate distribution. The quadratic exponential distribution has been suggested as an alternative (Prentice and Zhao, 1991), but the computing seems to be prohibitive except for small cluster sizes.

Finally, for binary data, our result generalizes existing theory for robust logistic regression. For a logit link applied to independent binary data, the Mallows class theory from the Theorem in Section 3.1 is equivalent to Eq. (2.3) in Carroll and Pederson (1993). For the Schweppe class, our approach is related to that of Künsch et al. (1989).

6.2. Efficiency

In order to further study the roles of ρ and the design matrix in the asymptotic relative efficiency of resistant generalized estimating equations relative to generalized estimating equations, Eq. (7) was evaluated in a model for correlated binary responses with logit link. The model has a single covariate which is either constant within cluster or varying within cluster but mean-balanced with $\bar{x}_i = 0$. Furthermore, we assume the intercept parameter is $\beta_0 = -2$ and the slope is $\beta_1 = 0.8$. Primary interest is in the asymptotic relative efficiency of $\hat{\beta}_{IR}$ defined by $\text{ARE}_{\beta_1} = \text{var}(\hat{\beta}_{IG})/\text{var}(\hat{\beta}_{IR})$. For comparative purposes, $K = 50$ and $n = 4$ throughout, while ρ and the design are varied.

Two designs are considered and they are summarized below. They will be referred to as designs A and B. In design A, the covariate was constant within cluster, while in design B, it varied within cluster as shown in Table 1.

For each design, ARE_{β_1} was evaluated for $\rho = 0.3$ and 0.7 giving a total of four scenarios for Schweppe observation-downweighting (Table 2), Mallows observation-downweighting (Table 3) and Mallows cluster-downweighting (Table 4). For each scenario in each table, the asymptotic relative efficiency is given for a range of tuning constants, a , applied to the weight function, $w_{it}(v_{it}, a) = \exp\{-(v_{it}/a)^2\}$,

Table 1

		Cluster level $\delta_i = -1 + (2(i - 1))/(k - 1)$	Vary within cluster $\Delta_i = i/K$
$n = 4$	Design A	$\begin{pmatrix} 1 & \delta_i \\ 1 & \delta_i \\ 1 & \delta_i \\ 1 & \delta_i \end{pmatrix}$	Design B
			$\begin{pmatrix} 1 & \Delta_i \\ 1 & \Delta_i/3 \\ 1 & -\Delta_i/3 \\ 1 & -\Delta_i \end{pmatrix}$

Table 2

The asymptotic relative efficiency of Schweppe observation downweighting REGEE to GEE for correlated binary responses

a	5	4	3.5	3	2.5	2.25	2	1.75	1.5	1	0.5
ratio	1.65	2.16	2.68	3.62	5.39	6.62	7.65	7.75	6.70	3.69	1.67
Design A											
ARE_{β_1}	0.982	0.959	0.935	0.894	0.829	0.791	0.758	0.747	0.774	0.890	0.982
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	0.981	0.956	0.929	0.883	0.809	0.766	0.730	0.722	0.759	0.887	0.977
Design B, $P = 0.3$											
ARE_{β_1}	0.967	0.930	0.893	0.837	0.758	0.715	0.680	0.664	0.685	0.821	0.970
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	0.967	0.929	0.892	0.835	0.756	0.713	0.677	0.662	0.682	0.817	0.969
Design B, $P = 0.7$											
ARE_{β_1}	0.922	0.840	0.769	0.673	0.558	0.504	0.463	0.446	0.467	0.642	0.933
λ_1	1	1	1	1	1	1	1	1	1	1	1
λ_2	0.910	0.819	0.741	0.639	0.521	0.467	0.425	0.408	0.428	0.603	0.924

a = Tuning constant which is applied to weight function, $\exp\{- (r_{it}/a)^2\}$.

ratio = $\max(b_{it})/\min(b_{it})$.

$ARE_{\beta_1} = \text{var}(\hat{\beta}_{1G})/\text{var}(\hat{\beta}_{1R})$.

λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\hat{\beta}_G)\text{var}^{-1}(\hat{\beta}_R)$.

(Holland and Welsch, 1977) where v_{it} is specified later for each class. Also provided for each scenario is a measure, $\text{ratio}_b = \max(b_{it})/\min(b_{it})$, evaluated over $t = 1, \dots, n$, and $i = 1, \dots, K$. It is a crude measure of loss of efficiency, since larger values indicate a wider range of nonoptimal weights applied to the observations. Finally, the eigenvalues of $\text{var}(\hat{\beta}_G)\text{var}^{-1}(\hat{\beta}_R)$, denoted λ_1 for the largest and λ_2 for the smallest are given for each scenario. The eigenvalue, λ_1 , corresponds to the maximum efficiency obtained among all linear combinations of β_0 and β_1 , and λ_2 corresponds to the minimum. It was found that the efficiency was generally closer to λ_2 than λ_1 . Note throughout that as a approached infinity which represents generalized estimating equations, the weights and thus ratio_b , ARE_{β_1} , and $ARE_{R:G}$ approached 1. The following values of the tuning constant a were considered: 0.5, 1, 1.5, 2, 2.5, 3, 4 and 5.

6.2.1. Results for Schweppe observation downweighting

The downweighting function, w_{it} , was applied to the Pearson residuals, i.e., $v_{it} = r_{it}$, to obtain the efficiencies shown in Table 2. For design A in which the covariate was cluster-level, ARE_{β_1} does not depend on correlation. In addition, for any common

Table 3

The asymptotic relative efficiency of Mallows observation downweighting REGEE to GEE for correlated binary responses

<i>a</i>	5	4	3	2.5	2	1.5	1	0.5
Design A								
ratio	1.26	1.44	1.91	2.53	4.26	13.2	331	++
ARE $_{\beta_1}$	0.997	0.993	0.978	0.958	0.912	0.802	0.565	0.154
λ_1	1	1	1	1	1	1	0.904	0.515
λ_2	0.994	0.987	0.961	0.927	0.852	0.698	0.448	0.131
Design B, $\rho = 0.3$								
ratio	1.97	2.87	6.53	14.9	68.2	++	++	++
ARE $_{\beta_1}$	0.966	0.929	0.838	0.757	0.647	0.507	0.322	0.138
λ_1	1	1	1	1	1	0.983	0.919	0.725
λ_2	0.956	0.908	0.797	0.703	0.583	0.446	0.283	0.124
Design B, $\rho = 0.7$								
ratio	1.84	2.59	5.42	11.4	44.8	863	++	++
ARE $_{\beta_1}$	0.948	0.892	0.770	0.673	0.561	0.450	0.325	0.136
λ_1	1	1	1	1	1	0.991	0.942	0.694
λ_2	0.930	0.859	0.710	0.600	0.480	0.371	0.261	0.113

++ = Greater than 1000.

tuning constant is *a*.

ratio = $\max(b_{it})/\min(b_{it})$.

ARE $_{\beta_1}$ = $\text{var}(\hat{\beta}_{1G})/\text{var}(\hat{\beta}_{1R})$.

λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\hat{\beta}_G)\text{var}^{-1}(\hat{\beta}_R)$.

Table 4

The asymptotic relative efficiency of Mallows cluster downweighting REGEE to GEE for correlated binary responses

Design B, $\rho = 0.3$								
ratio	1.15	1.25	1.49	1.77	2.44	4.89	35.5	++
ARE $_{\beta_1}$	0.998	0.996	0.987	0.974	0.940	0.838	0.549	0.145
λ_1	0.998	0.996	0.989	0.980	0.956	0.896	0.749	0.473
λ_2	0.998	0.996	0.987	0.973	0.938	0.832	0.532	0.132
Design B, $\rho = 0.7$								
ratio	1.15	1.25	1.49	1.77	2.44	4.90	35.7	++
ARE $_{\beta_1}$	0.998	0.996	0.987	0.974	0.941	0.842	0.562	0.155
λ_1	0.998	0.996	0.989	0.979	0.955	0.894	0.743	0.464
λ_2	0.998	0.996	0.987	0.973	0.938	0.832	0.532	0.133

++ = Greater than 1000.

a = Tuning constant which is applied to weight function, $\exp\{- (r_{it}/a)^2\}$.

ratio = $\max(b_{it})/\min(b_{it})$.

ARE $_{\beta_1}$ = $\text{var}(\hat{\beta}_{1G})/\text{var}(\hat{\beta}_{1R})$.

λ_1 and λ_2 are, respectively, largest and smallest eigenvalue of $\text{var}(\hat{\beta}_G)\text{var}^{-1}(\hat{\beta}_R)$.

cluster size the efficiencies are the same as those shown for Design A in Table 2. Resistant generalized estimating equations was less efficient when the covariate varied within cluster. For these designs the ARE $_{\beta_1}$ depended on sample size and correlation, and in particular, resistant generalized estimating equations was less efficient when $\rho = 0.7$ than when $\rho = 0.3$. Interestingly, for all scenarios considered, the greatest loss in efficiency occurred at approximately $a = 1.75$. Similarly, ratio $_b$, attained its greatest value at 1.75. Fig. 1 plots ARE $_{\beta_1}$ and ratio $_b$ versus *a*, for design, A. As *a* became smaller the efficiency actually increased due to excessive downweighting resulting in

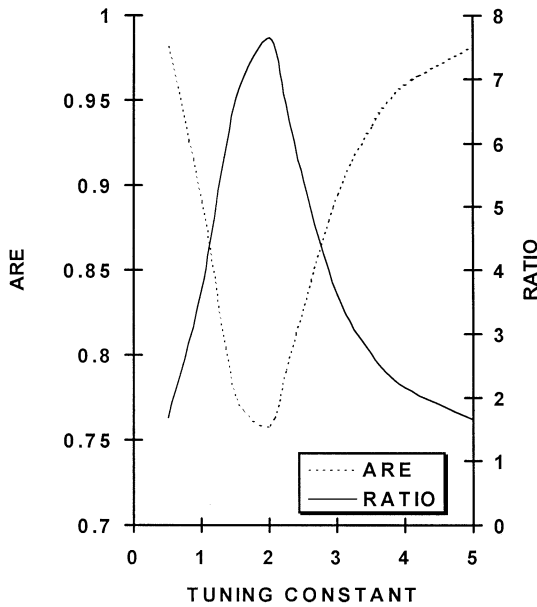


Fig. 1. ARE_{β_1} and $ratio_b$ versus tuning constant for cluster level designs.

a loss of discriminating power to detect the data which was the most influential. Since $b_{it} = b_i(\mu_{it})$ is a function of the mean only, $ratio_b$ is the same for every design considered since $\min(\mu_{it}) = 0.06$ and $\max(\mu_{it}) = 0.23$. In general, $ratio_b$ indicates that the loss of efficiency increases with increasing variation in the covariates, X , and with increasing magnitude of β in absolute terms. The relationship between $ratio_b$ and efficiency for the cluster level designs is also illustrated by Fig. 2. This shows that two very different tuning constants may give very similar efficiency, but one represents excessive downweighting, and the other does not. Both figures show that as robustness increases, efficiency decreases. The relationship of efficiency and tuning constant is illustrated for design B in Fig. 3.

6.2.2. Results for Mallows observation downweighting

The weight function was applied to the observation leverages by setting $v_{it} = h_{it}N/P$. Generally, efficiency increased as a increased. For design B resistant generalized estimating equations was less efficient when $\rho = 0.7$ than when $\rho = 0.3$.

6.2.3. Results for Mallows cluster downweighting

The weight function was applied to the cluster leverages by setting $v_{it} = H_i N/n\rho = 25 H_i$. In Table 4 there is no entry for design A as it is identical to that in Table 3. For designs in which the covariate varied within cluster, the ARE_{β_1} depended on n and ρ , but varied little.

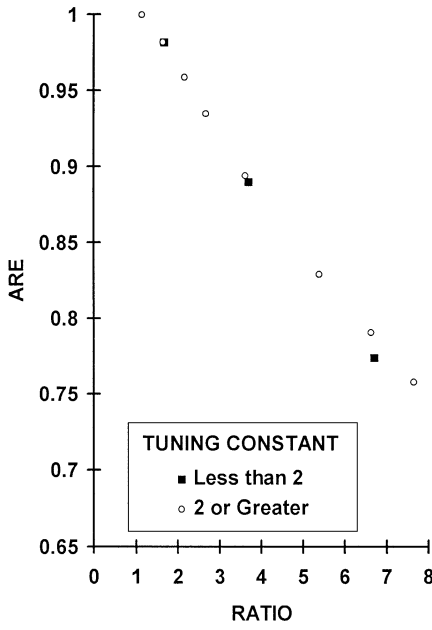


Fig. 2. ARE_{β_1} versus $ratio_b$ for cluster level designs.

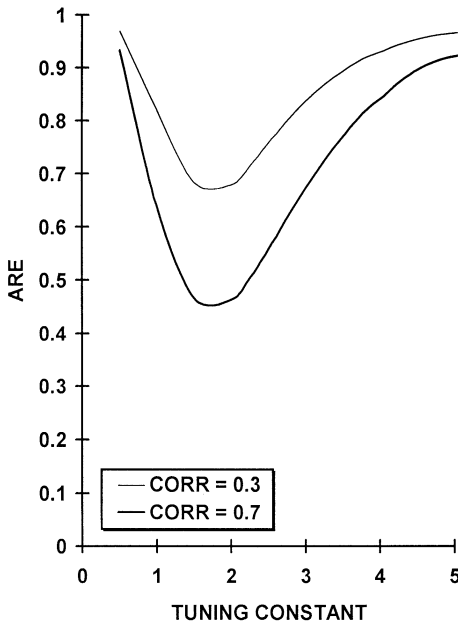


Fig. 3. ARE_{β_1} versus tuning constant for Design B for $\rho = 0.3$ and 0.7 .

7. Discussion

This work addressed the problem that a few influential observations or clusters can have a large effect on regression parameter estimates and fitted values. Preisser and Qaqish (1996) introduced deletion diagnostics which estimate the effect of the deletion of an observation or a cluster. This paper introduced a modification of the generalized estimating equations procedure called resistant generalized estimating equations which gives regression parameter estimates that are resistant to the influence of a small subset of the data. This is achieved by the automatic downweighting of influential observations in the estimating function. Although, in principle, resistant generalized estimating equations applies to the same class of models considered by generalized estimating equations, limitations exist in its actual implementation. For Schweppe observation downweighting, the full marginal univariate distributions are required. Except for bivariate binary or normal data, Schweppe cluster downweighting is generally prohibitive because the full multivariate distribution is needed in order to estimate the regression parameters consistently. In the Mallows class, however, no additional distributional assumptions are required. The robustness of resistant generalized estimating equations was illustrated by Preisser and Qaqish (1999), though an example of medical practice data with widely varying cluster sizes.

Generally, robust regression methods require a robust estimate of the scale parameter. This issue is not addressed in this paper. However, for binary the scale is not estimated because it is equal to one. The same argument applied to count data that exhibits no extra-poisson variation.

Appendix A

A.1. Outline of Proof of Theorem in Section 4.1. The sketch here considers α known. The REGEE in Schweppe class can be written as

$$U(\beta) = \sum_{i=1}^K U_i(\beta) = \sum_{i=1}^K D_i'(\beta) V_i^{-1}(\beta) [\psi_{ki} - c_i],$$

where $\psi_{ki} = W_{ki}(Y_i, \beta)(Y_i - \mu_i(\beta))$ and $c_i = E(\psi_{ki})$.

A Taylor expansion gives

$$U(\hat{\beta}) = 0 = U(\beta) + \partial U(\beta) / \partial \beta |_{\beta = \tilde{\beta}} (\beta - \hat{\beta})$$

for some $\tilde{\beta} = \lambda \beta + (1 - \lambda) \hat{\beta}$, $\lambda \in (0, 1)$. It follows that

$$k^{1/2}(\hat{\beta} - \beta) = \left[\frac{\partial U(\tilde{\beta}) / \partial \beta}{k} \right]^{-1} (U(\beta) / k^{1/2}).$$

By the Central limit theorem for triangular arrays (Theorem 3.3.5 in Sen and Singer, 1993) in conjunction with the Cramer–Wold device, and regularity conditions, and

assumption (iii),

$$U(\beta)/k^{1/2} \sim \text{MVN}\left(0, \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k D_i' V_i^{-1} \text{Var}(\psi_i) V_i^{-1} D_i\right).$$

Note that assumption (iii) follows from: (1) the fact that the weight matrix W_{ki} is bounded between 0 and 1, and (2) the usual generalized estimating equations assumptions of the finiteness of the second moment of Y_i . Next, it can be shown that

$$E[\partial U(\beta)/\partial \beta] = \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i.$$

To see this note that

$$\partial U_i(\beta)/\partial \beta = \frac{\partial}{\partial \beta} \{D_i' V_i^{-1}\} [\psi_{ki} - c_i] + \{D_i' V_i^{-1}\} \frac{\partial}{\partial \beta} [\psi_{ki} - c_i].$$

The first part has expectation zero, and in the second part, under assumption (iv), apply,

$$\frac{\partial}{\partial \beta} [\psi_{ki} - c_i] = \frac{\partial}{\partial \mu_i} [\psi_{ki} - c_i] \frac{\partial \mu_i}{\partial \beta} = [\dot{\psi}_{ki} - \dot{c}_i] D_i.$$

Then by the Markow weak law of large numbers,

$$\frac{1}{k} \partial U(\beta)/\partial \beta \xrightarrow{p} \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i.$$

Since, under certain regularity conditions,

$$\frac{1}{k} \|\partial U(\tilde{\beta})/\partial \beta - \partial U(\beta)/\partial \beta\| \xrightarrow{p} 0,$$

it follows that,

$$\frac{1}{k} \partial U(\tilde{\beta})/\partial \beta \xrightarrow{p} \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^k D_i' V_i^{-1} \Gamma_i D_i.$$

Lastly, by Slutsky's Theorem, $k^{1/2}(\hat{\beta} - \beta) \sim \text{MVN}(0, V_R)$. \square

References

- Carroll, R.J., Pederson, S., 1993. On robustness in the logistic regression model. *J. R. Statist. Soc. B*, 55, 693–706.
- Godambe, Heyde, 1987. Quasi-likelihood and optimal estimation. *Int. Statist. Rev.* 55, 231–244.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- Holland, P.W., Welsch, R.E., 1977. Robust regression using iteratively reweighted least-squares. *Commun. Statist. Theor. Meth.* A6 (9), 813–827.
- Huggins, R.M., 1993. A robust approach to the analysis of repeated measures. *Biometrics* 49, 715–720.
- Kunsch, H.R., Stefanski, L.A., Carroll, R.J., 1989. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *J. Amer. Statist. Assoc.* 84, 460–466.

- Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.
- Mancl, L., Leroux, B., 1996. Efficiency of regression estimates for clustered data. *Biometrics* 52, 500–511.
- McCullagh, P., Nelder, J.A., 1989. *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.
- Morgenthaler, S., 1992. Least-absolute-deviations fits for generalized linear models. *Biometrika* 79, 747–754.
- Morton, 1981. Efficiency of estimating equations and the use of pivots. *Biometrika* 68, 227–233.
- Preisser, J., Qaqish, B., 1996. Deletion diagnostics for generalized estimating equations. *Biometrika* 83, 551–562.
- Preisser, J., Qaqish, B., 1999. Robust regression for clustered data with application to binary responses. *Biometrics* (in press).
- Pregibon, D., 1982. Resistant fits for some commonly used logistic models with medical applications. *Biometrics* 38, 485–498.
- Prentice, R., Zhao, 1991. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* 47, 825–839.
- Sen, P.K., Singer, J.M., 1993. *Large Sample Methods in Statistics*. Chapman & Hall, New York.
- Singer, J.M., Sen, P.K., 1985. M-methods in multivariate linear models. *J. Multivariate Anal.* 17, 168–184.
- Williams, D.A., 1987. Generalized linear model diagnostics using the deviance and single case deletions. *Appl. Statist.* 36, 181–191.