Clustered Binomial Data Analysis by Modified Generalized Estimating Equations

Xuemao Zhang[*]

Abstract
Extra-binomial variation in clustered binomial data is frequently observed in biomedical and observational studies. The usual generalized estimating equations (GEE) method treats the extra-binomial parameter as a constant across all clusters. In this paper, a two-parameter variance function modeling the extraneous variance is proposed to account for heterogeneity among clusters. The new approach allows modeling the extra-binomial variation as a function of the mean and binomial size.

Key Words: Clustered binomial data, Extra-binomial variation, Generalized estimating equations, Variance function

1. Introduction

Clustered/longitudinal data arise in many epidemiological and bio-statistical practices in which a number of repeated count/binomial responses are observed on a number of individuals. Consider clustered binomial data arranged in a series of observations $(y_{ij}, \mathbf{x}_{ij})$ from n clusters, where $\mathbf{x}_{ij}$ is a vector of covariates associated with the univariate binomial outcome y_{ij} with size m_{ij}, $j = 1, \dots, T_i$, $i = 1, \dots, n$. Assume all clusters have the same number of observations/individulas $T_i = T$. Now, each y_{ij} is assumed to follow the binomial distribution $B(m_{ij}, \mu_{ij})$; that is

$$P(Y_{ij} = y_{ij}) = \{m_{ij}!/[y_{ij}!(m_{ij} - y_{ij})!]\}\mu_{ij}^{y_{ij}}(1 - \mu_{ij})^{m_{ij}-y_{ij}},$$

$j = 1, \dots, T$, $i = 1, \dots, n$. We suppose that the response probability μ_{ij} depends on the explanatory variable $\mathbf{x}_{ij}$ via a logistic regression model

$$\mu_{ij} = \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})/[1 + \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta})],$$

where $\boldsymbol{\beta}$ is a p-vector of unknown regression coefficients. Now, under the binomial assumption $Var(y_{ij}) = m_{ij}\mu_{ij}(1 - \mu_{ij})$. In presence of extra-binomial variation (over or under dispersion) a more general distribution, such as the beta-binomial distribution, which accounts for such extra-binomial variation, is considered. However, in practice a distributional assumption for the data may not hold. In such a situation, the generalized estimating equation (GEE) approach (Liang and Zeger, 1986, and Zeger and Liang, 1986) can be used for the estimation of the regression parameters.

The main advantage of the GEE method of estimation in clustered data analysis is that the estimators are consistent even if the working correlation structure is misspecified. Although miss-specification of the correlation structure does not affect consistency of the estimates of the regression parameters, it does reduce the efficiency of the regression parameter estimates (Wang and Carey, 2004). As discussed by Wang and Zhao (2007) the GEE approach focuses on correctly modelling the working correlation matrix. However, it treats the variance function to be of a known form. The usual practice is to take the variance function as that obtained from the overdispersed generalized linear models (GLM) (Nelder and Wedderburn, 1972, and McCullagh and Nelder, 1983). Now, the variance function for

[*] Department of Mathematics, East Stroudsburg University, East Stroudsburg, PA 18301

binomial data is $v(\mu) = \mu(1-\mu)$. Therefore, in the framework of a GLM for data with extra-binomial variation $\mathrm{Var}(y_{ij}) = \varphi m_{ij} v(\mu_{ij})$, where φ is the extra-binomial parameter. This form of the variance of clustered binomial response data could be too limited in practice. Heterogeneity between clusters in a clustered data analysis might be another source of the extra-binomial variation.

To apply the GEE methodology to binomial data it would be convenient to deal with data y_{it}/m_{it} in the form of proportion. If y_{it} has a binomial distribution, then $\mathrm{Var}(y_{it}/m_{it}|\mathbf{x}_{it}) = \mu_{it}(1-\mu_{it})/m_{it}$. Suppose y_{it} is a sum of m_{it} correlated binary responses with common success probabilities μ_{it} such that $y_{it} = \sum_{j=1}^{m_{it}} z_{itj}$ and $\mathrm{corr}(z_{itj}, z_{itj'}) = \tau_{it}$ for $j \neq j'$. Then,

$$\mathrm{Var}(y_{it}/m_{it}) = \mu_{it}(1-\mu_{it})[1+\tau_{it}(m_{it}-1)]/m_{it}, \tag{1}$$

with $\tau_{it} \geq -1/(m_{it}-1)$. This variance can also be derived from the beta-binomial distribution (Williams, 1975, and Crowder, 1978). Note that $\mathrm{Var}(y_{it}/m_{it}) = \mu_{it}(1-\mu_{it})/m_{it} + \frac{(m_{it}-1)}{m_{it}}\tau_{it}\mu_{it}(1-\mu_{it})$ consists of two parts. The first part is the variance of a binomial proportion and the second part is the extra-binomial variation. Moore(1987) generalized the model by expressing τ_{it} as $\tau_{it}(\mu_{it}) = \phi_i \mu_{it}^{\delta-1}(1-\mu_{it})^{\delta-1}$ and obtained

$$\begin{aligned}\mathrm{Var}(y_{it}/m_{it}) &= \mu_{it}(1-\mu_{it})\{1+\phi_i(m_{it}-1)[\mu_{it}(1-\mu_{it})]^{\delta-1}]\}/m_{it} \\ &= \mu_{it}(1-\mu_{it})/m_{it} + \phi_i(m_{it}-1)[\mu_{it}(1-\mu_{it})]^{\delta}/m_{it},\ \delta \geq 1.\end{aligned} \tag{2}$$

Note, this also imposes a constraint on ϕ_i as $\phi_i \geq -[\mu_{it}(1-\mu_{it})]^{1-\delta}/(m_{it}-1)$. The parameters ϕ_i and δ, in some way, are over-dispersion parameters; ϕ_i is the usual beta-binomial type over-dispersion parameter and δ is an over-dispersion parameter due to the variation in the proportion parameter μ_i.

In this paper, we propose to use this two-parameter variance function when the GEE method is used to estimate the regression parameters. The form of the variance function is generalized by allowing for negative $\delta - 1$ values. It allows us to model the extra-binomial variation of clustered binomial data as a function of the mean.

The modified GEE approach for clustered binomial data is presented in Section 2. Three methods of estimating the variance parameters are proposed in Section 3. A limited simulation study is conducted in Section 4 to compare the three methods and investigate the effect of the variance function and misspecification of the working correlation structure on the efficiency of the estimates of the regression parameters. A discussion follows in Section 5.

2. Generalized Estimating Equations for Clustered Extra-binomial Data

Consider a clustered data analysis with n clusters and let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ be the binomial responses of size $m_{it}, t = 1, \ldots, T$ for cluster i over T individual in the cluster. Denote the $T \times p$ design matrix for $\mathbf{y}_i$ as $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$, where $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})'$ are the p covariates/predictor variables of interest for individual t, $t = 1, \ldots, T$. We assume that the n clusters are independent while the measurements y_{it}, $t = 1, \ldots, T$ in cluster i are correlated. For each individual t, consider the binomial proportion y_{it}/m_{it} and let

$$\mu_{it} = E(y_{it}/m_{it}|\mathbf{x}_{it}, m_{it}), \quad \mu_{it} = g^{-1}(\mathbf{x}_{it}'\boldsymbol{\beta}), \quad i = 1, \ldots, n, \quad t = 1, \ldots, T,$$

where g is a link function and $\boldsymbol{\beta}$ is the vector of regression parameters of interest. In this paper, we consider the logit link only. That is, $\mathrm{logit}(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta}$. Now, let

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})',\ \mathbf{y}_i = (y_{i1}, \ldots, y_{iT})',\ \text{and}\ \mathbf{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta},\ \ i = 1, \ldots, n.$$

Then, the set of GEEs for estimating $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{n} \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i) = 0, \tag{3}$$

where $\mathbf{m}_i = (m_{i1}, \ldots, m_{iT})'$, $\mathbf{V}_i = \varphi \mathbf{A}_i^{1/2} \mathbf{W}_i^{-1/2} R(\boldsymbol{\alpha}) \mathbf{W}_i^{-1/2} \mathbf{A}_i^{1/2}$, φ is an extra-binomial parameter, $\mathbf{A}_i = \mathrm{diag}(v(\mu_{it}))$ a diagonal matrix with $v(\mu_{it}) = \mu_{it}(1-\mu_{it})$ on its tth diagonal, $\mathbf{W}_i$ is a $T \times T$ diagonal matrix with m_{it} as the tth diagonal and $R(\boldsymbol{\alpha})$ is the working correlation matrix completely defined by the parameter $\boldsymbol{\alpha}$. The main advantage of the GEE approach is that the estimator of $\boldsymbol{\beta}$ is consistent even if the working correlation matrix $R(\boldsymbol{\alpha})$ structure is misspecified, though correct specification of the correlation structure can improve the efficiency of the regression parameter estimate (Wang and Carey 2004).

To account for the size $\mathbf{m}_i$ and binomial probabilities $\boldsymbol{\mu}_i$, we propose the set of GEEs of the form (3) with $\mathbf{V}_i = \mathbf{A}_i^{1/2} R(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$, $\mathbf{A}_i = \mathrm{diag}(v(\mu_{it}))$ a diagonal matrix with its tth diagonal

$$v(\mu_{it}) = \mu_{it}(1-\mu_{it})\{1 + \phi(m_{it}-1)[\mu_{it}(1-\mu_{it})]^{\delta-1}\}/m_{it}, \tag{4}$$

where $-\infty < \delta < \infty$ and $\phi \geq -\min\{[\mu_{it}(1-\mu_{it})]^{1-\delta}/(m_{it}-1), i = 1, \ldots, n, t = 1, \ldots, T\}$.

Note that for $\delta - 1 = 0$, $v(\mu_{it})$ is the beta-binomial variance given in (1); for $\delta - 1 > 0$, $0 < [\mu_{it}(1-\mu_{it})]^{\delta-1} < 1$ and $v(\mu_{it})$ is the generalized variance given in (2) and for $\delta - 1 < 0$, $[\mu_{it}(1-\mu_{it})]^{\delta-1} > 1$ represents a further generalization of (2) that takes into account the over-dispersion of the binomial proportion for each cluster. That is, we have extended the form of the variance function in Moore (1987) by allowing for negative values of $\delta - 1$. In this paper, we focus on the estimation of ϕ and δ to improve the estimation efficiency of regression parameter estimates.

The estimation procedure for $(\boldsymbol{\beta}, \phi, \delta)$ can be described by the following iterative algorithm.

(i) Obtain an initial estimate $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$ from the GEE independence model.

(ii) For given $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(j)}$, estimate the variance parameters ϕ and δ by the method of least squares, weighted least squares or Gaussian estimation method as described in Section 3.

(iii) For given $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(j)}$ obtained in step (i) and $\phi = \hat{\phi}$ and $\delta = \hat{\delta}$ obtained in step (ii), the estimate $\hat{\boldsymbol{\alpha}}$ of the correlation parameter $\boldsymbol{\alpha}$ by the method of moments (Zeger and Liang 1986). For example, the scalar parameter α in the working exchangeable correlation matrix is estimated by

$$\hat{\alpha} = \sum_{i=1}^{N} \sum_{k \neq l} y_{ik}^* y_{il}^* \Big/ \left[(T-1) \sum_{i=1}^{N} \sum_{k=1}^{T} {y_{ik}^*}^2 \right],$$

where $y_{ik}^* = (y_{ik} - \hat{\mu}_{ik})/\sqrt{\hat{\mu}_{ik}(1-\hat{\mu}_{ik})}, \hat{\mu}_{ik} = \mu_{ik}(\hat{\boldsymbol{\beta}}^{(j)})$.

(iv) Update $\hat{\boldsymbol{\beta}}$ according to the modified Fisher scoring formula in the GEE method

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} + \left\{ \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right\}^{-1} \left\{ \sum_{i=1}^{N} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) \right\}, j = 1, 2, \cdots,$$

where $\hat{\mathbf{D}}_i = \partial \mu_i / \partial \boldsymbol{\beta} \big|_{\hat{\boldsymbol{\beta}}^{(j)}}$, $\hat{\mu}_i = \mu_i(\hat{\boldsymbol{\beta}}^{(j)})$ and $\hat{\mathbf{V}}_i = \mathbf{A}_i(\hat{\phi}, \hat{\delta}, \hat{\boldsymbol{\beta}}^{(j)})\, R(\hat{\boldsymbol{\alpha}})\, \mathbf{A}_i(\hat{\phi}, \hat{\delta}, \hat{\boldsymbol{\beta}}^{(j)})$.

(v) Iterate between (ii) and (iv) until a desired convergence criterion (for example max$|\hat{\boldsymbol{\beta}}^{(j+1)} - \hat{\boldsymbol{\beta}}^{(j)}| < 0.001$) for $\boldsymbol{\beta}$ is satisfied. At convergence, the final estimates of $\boldsymbol{\alpha}$, ϕ and δ are given by $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$, $\phi = \hat{\phi}$ and $\delta = \hat{\delta}$ used in the last step of iteration.

3. Estimation of Variance Parameters

The GEE estimate of the regression parameter $\boldsymbol{\beta}$ is obtained by an iterative procedure given above in which the variance parameters ϕ and δ in (4) need to be estimated. In this section we discuss three methods of estimation of ϕ and δ for given estimate $\hat{\boldsymbol{\beta}}$.

3.1 Least squares

Let $\mathbf{e}_i = \mathbf{y}_i/\mathbf{m_i} - \boldsymbol{\mu}_i$ for given estimate of $\boldsymbol{\beta}$. Then e_{it} has mean zero and $E(e_{it}^2)$ is approximately $v(\mu_{it})$, $t = 1, \ldots, T$. By the method of least squares, ϕ and δ are obtained by minimizing

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it}^2 - v(\mu_{it}))^2.$$

That is, estimates of ϕ and δ are obtained by solving the following system of equations simultaneously.

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it}^2 - v(\mu_{it}))[\partial v(\mu_{it})/\partial\phi] &=0,\\ \sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it}^2 - v(\mu_{it}))[\partial v(\mu_{it})/\partial\delta] &=0.\end{aligned}$$

Note that the least square method involves a nonlinear variance function and there is no closed-form solution. Thus the convergence depends on the choice of initial values. Furthermore, when the binomial sizes $\mathbf{m}_i$'s are small, $v(\boldsymbol{\mu}_i)$ could be very small while the observed values of the square of the residuals are large. In this case, the procedure could diverge or the estimates of ϕ and δ may not be reliable even if convergence can be obtained. Therefore, the estimators by the least square method are consistent only if there is a convergent solution to the above minimization problem.

3.2 Weighted least squares

As is well-known, for normal data, e_{it}^2 has approximate variance $v^2(\mu_{it})$. This suggests that the weighted least square estimator of ϕ and δ can be obtained by minimizing

$$\sum_{i=1}^{n}\sum_{t=1}^{T}(e_{it}^2 - v(\mu_{it}))^2/v^2(\mu_{it}),$$

or by solving the following system of equations simultaneously

$$\begin{aligned}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial v(\mu_{it})}{\partial\phi}\Big[1 + (e_{it}^2 - v(\mu_{it}))/v(\mu_{it})\Big](e_{it}^2 - v(\mu_{it}))/v^2(\mu_{it}) &=0,\\ \sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial v(\mu_{it})}{\partial\delta}\Big[1 + (e_{it}^2 - v(\mu_{it}))/v(\mu_{it})\Big](e_{it}^2 - v(\mu_{it}))/v^2(\mu_{it}) &=0.\end{aligned}$$

3.3 Gaussian estimation

Whittle (1961) introduces the Gaussian estimation procedure which uses the normal log-likelihood, without assuming that the data are normally distributed. Then, the Gaussian log-likelihood apart from a constant is given by

$$l = -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\det[2\pi\mathbf{V}_i] + (\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)\right\}. \tag{5}$$

Wang and Zhao (2007) showed that when $R(\boldsymbol{\alpha})$ is an identity matrix, the estimates of the parameters ϕ and δ obtained by maximizing (5) are consistent. Therefore, we choose $\mathbf{V}_i = \mathbf{A}_i$ in (5) and estimates ϕ and δ by solving the following Gaussian score equations simultaneously.

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left\{\left[\mathbf{A}_i^{-1}(\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)' - I_T\right]\mathbf{A}_i^{-1}\frac{\partial\mathbf{A}_i}{\partial\phi}\right\} &= 0,\\
\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left\{\left[\mathbf{A}_i^{-1}(\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i/\mathbf{m}_i - \boldsymbol{\mu}_i)' - I_T\right]\mathbf{A}_i^{-1}\frac{\partial\mathbf{A}_i}{\partial\delta}\right\} &= 0,
\end{aligned}$$

where I_T is a T dimensional identity matrix.

4. Simulation study

In this section we compare, by simulations, the estimators of the regression parameters by the four methods: GEE, GEELS, GEEWLS and GEEG. That is, we compare the usual GEE and GEE with ϕ and δ estimated by the method of least squares (LS), weighted least squares (WLS) and Gaussian (G) method, respectively. We investigate the effect of the variance function and misspecification of the working correlation structure on the efficiency of the estimates of the regression parameters.

In our limited simulation study, the extra-binomial responses are generated by adding several correlated binary random variables. The detailed data generation procedure is described as in the following. For the ith cluster, given a $T \times p$ design matrix $\mathbf{X}_i$ and the regression parameter $\boldsymbol{\beta}$, denote $\boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\beta} = (\eta_{i1}, \ldots, \eta_{iT})'$. First, T correlated binary variables $(z_{i11}, \ldots, z_{iT1})'$ with marginal mean $\boldsymbol{\mu}_i = \exp(\boldsymbol{\eta}_i)/[1 + \exp(\boldsymbol{\eta}_i)]$ and correlation matrix $\Omega(\gamma)$ are generated using the method of Qaqish(2003), which accounts for the lower and upper bounds of the correlation parameters. Secondly, we generate $z_{ij2}, \ldots, z_{ijm_{ij}}$, again by the method of Qaqish(2003), such that the m_{ij} binary responses $\mathbf{z}_{ij} = (z_{ij1}, \ldots, z_{ijm_{ij}})$ are equally correlated with correlation strength ρ_i, a uniform random variable between 0 and 0.7, $j = 1, \ldots, T$, $i = 1, \ldots, n$. Here, the strength of ρ_i is chosen to be not too strong to avoid data generation difficulties because of natural restrictions imposed by the marginal means (Qaqish, 2003).

Finally, the correlated extra-binomial response variables are given by $y_{ij} = \sum_{k=1}^{m_{ij}} z_{ijk}$, $j = 1, \ldots, T$, $i = 1, \ldots, n$. Thus, the response variables $(y_{i1}, \ldots, y_{iT})'$ generated for cluster i in this way are correlated and each y_{ij} is an extra-binomial variable, $j = 1, \ldots, T$. Note that the correlation structure among the binomial responses cannot in general be the same as that of $\Omega(\gamma)$. Therefore, even if the working correlation structure in the GEE estimation is the same as that of $\Omega(\gamma)$, the correlation structure might have been misspecified.

To investigate the effect of the variance function on the efficiency of the estimates of the regression parameters, we choose large samples in the simulations. We consider sample size $n = 100$ with each cluster having $T = 4$ observations with binomial size

$m_{ij} = m_i, i = 1, \ldots, 4$, a discrete uniform random variable between 30 and 100, $p = 2$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ with $\beta_0 = 0.5$, $\beta_1 = 1.0$. The covariates x_{ij} are generated as a uniform random variable from -3 to 3, $j = 1, \ldots, 4$. That is, the design matrix for each cluster i is of the form

$$\mathbf{X}_i = \begin{pmatrix} 1 & \cdots & 1 \\ x_{i1} & \cdots & x_{i4} \end{pmatrix}'.$$

Simulations were conducted with true exchangeable correlation structure $\Omega(\gamma)$. We take $\gamma = 0.0$, 0.1, 0.4 and 0.7 in $\Omega(\gamma)$ to account for the weak, moderate and strong correlations among the binomial responses. A working exchangeable correlation structure is used in the first set of simulations. Simulation results are summarized in Table 1.

Table 1: Biases, MSEs and variances of regression parameter GEE estimates; exchangeable $\Omega(\gamma)$ for $\gamma = 0.0$, 0.1 0.4 and 0.7; $n = 100$, $T = 4$ and $\beta_0 = 0.5, \beta_1 = 1.0$; based on $1,000$ replications.

		$n\times$Bias		$n\times$MSE		$n\times$Variance	
		β_0	β_1	β_0	β_1	β_0	β_1
$\gamma = 0.0$	GEE	0.102	0.298	0.677	0.368	0.593	0.322
	GEELS	0.157	0.536	0.615	0.343	0.524	0.343
	GEEWLS	0.341	1.310	0.848	0.574	0.795	0.726
	GEEG	0.802	1.086	0.637	0.367	0.538	0.303
$\gamma = 0.1$	GEE	0.310	0.661	0.722	0.452	0.757	0.522
	GEELS	0.372	0.833	0.660	0.410	0.696	0.438
	GEEWLS	0.760	1.392	0.842	0.661	0.796	0.592
	GEEG	0.924	1.476	0.691	0.449	0.710	0.438
$\gamma = 0.4$	GEE	0.628	0.877	0.922	0.602	0.777	0.472
	GEELS	0.833	1.154	0.853	0.547	0.700	0.500
	GEEWLS	0.974	1.418	1.042	0.738	0.932	0.751
	GEEG	1.440	1.873	0.903	0.606	0.717	0.520
$\gamma = 0.7$	GEE	0.012	0.394	1.103	0.698	1.323	0.494
	GEELS	0.069	0.550	0.967	0.640	1.135	0.472
	GEEWLS	0.762	1.621	1.343	1.010	1.445	0.682
	GEEG	1.117	1.452	1.028	0.713	1.139	0.476

We see from the results in Table 1 that the biases of the estimates of β_0 and β_1 by all methods are small showing that the estimators by the four methods are consistent. GEELS and GEEG estimates have a little larger biases compared to GEE and GEELS method while GEE method results in smallest biases. Furthermore, we compare the MSE's (mean squared error) of the estimates, where MSE is defined as the average of the squares of the differences between the true values of the parameters and their estimates. GEELS estimates of β_0 and β_1 have smallest MSE and variance estimates and GEEWLS estimates of β_0 and β_1 have the largest MSE and variance estimates. The usual GEE method leads to estimates of β_0 and β_1 with a somewhat larger MSE and variance estimates compared with GEEG.

To study the effect of misspecification of the working correlation structure, a further set of simulations were conducted with true AR(1) correlation structure $\Omega(\gamma)$ with the values of γ as 0.0, 0.1, 0.4 and 0.7 and working exchangeable correlation structure. Simulation results are summarized in Table 2.

We see from Table 2 that, again, all methods produce consistent estimators and GEE has least biased estimators; GEEWLS performs worst in terms of bias, MSE and variance estimates. In general, GEELS estimates of β_0 and β_1 have the smallest MSE and GEEG performs the second best in terms of MSE. For variance estimates of β_0 and β_1, GEEG performs best and the second best is GEELS.

Table 2: Biases, MSEs and variances of regression parameter estimates; AR(1) $\Omega(\gamma)$ for $\gamma = 0.0, 0.1, 0.4$ and 0.7; $n = 100$, $T = 4$ and $\beta_0 = 0.5, \beta_1 = 1.0$; based on $1,000$ replications.

		$n\times$Bias		$n\times$MSE		$n\times$Variance	
		β_0	β_1	β_0	β_1	β_0	β_1
$\gamma = 0.0$	GEE	0.295	0.336	0.723	0.352	0.751	0.558
	GEELS	0.308	0.525	0.680	0.318	0.728	0.516
	GEEWLS	0.752	1.200	0.880	0.542	1.236	0.792
	GEEG	0.976	1.114	0.722	0.343	0.688	0.507
$\gamma = 0.1$	GEE	0.167	0.258	0.701	0.378	0.690	0.351
	GEELS	0.358	0.440	0.639	0.346	0.669	0.312
	GEEWLS	0.875	0.884	0.798	0.517	0.758	0.397
	GEEG	0.974	1.076	0.673	0.373	0.667	0.312
$\gamma = 0.4$	GEE	0.301	0.504	0.763	0.541	0.624	0.519
	GEELS	0.439	0.806	0.674	0.491	0.595	0.497
	GEEWLS	0.634	1.306	0.803	0.678	0.846	0.854
	GEEG	0.975	1.499	0.707	0.540	0.585	0.491
$\gamma = 0.7$	GEE	0.776	1.187	1.085	0.596	1.175	0.780
	GEELS	0.914	1.439	1.029	0.573	1.104	0.747
	GEEWLS	1.064	1.793	1.281	0.843	1.094	1.023
	GEEG	1.723	2.272	1.109	0.627	1.132	0.725

The limited simulations show that GEELS and GEEG perform better than the usual GEE method in terms of MSE and variance estimates. The reason may be that the extra-binomial variances of correlated binomial responses for different clusters cannot be accounted for by one parameter only adopted by the usual GEE method. The GEEWLS performs worst in terms of MSE and variance estimates. This may be partly because that the weight chosen is based on normal data. The incorrect weight may exaggerate the heterogeneity among clusters. The problem of choosing an appropriate weight is left for future research.

In the above limited simulation study, the over-dispersion of the binomial responses is a result of adding several correlated binary responses. Therefore, there is no way to get any bias information of the estimates of ϕ and δ in the proposed model. The simulation of correlated over-dispersed binomial variables with specified variance function is left as a future research.

5. An example

We now analyze a biological data from Alderdice and Forrester (1968) analyzed by Chaganty, Sabo and Deng(2012). The data set is reproduced in Table 3. The purpose of the study is to model the effects of salinity and temperature on the proportion of hatched English sole eggs. The number of hatched eggs was recorded at seven salinity and five temperature levels. Measurements were taken in four separate tanks for each combination of salinity and temperature, and for each tank the number of fish eggs and the number hatched were recorded. Therefore, the tanks represent the repeated measure component for this binomial data set.

The goal of the analysis is to study the dependence of the proportion of eggs hatched on the salinity and temperature. Thus, we consider the following marginal model with a logit link

$$\text{logit}(\mu) = \beta_0 + \beta_1\text{Temp} + \beta_2\text{Salinity}. \qquad (6)$$

Table 3: Number of hatched and total eggs of English sole at different salinity and temperature levels in sea water.

		Tank 1		Tank 2		Tank 3		Tank 4	
Salinity (‰)	Temp.(C)	Hatch	Total	Hatch	Total	Hatch	Total	Hatch	Total
15	4	236	666	203	724	183	764	212	723
	8	600	656	697	747	615	746	641	703
	12	407	566	343	603	365	560	302	394
25	4	203	717	177	782	155	852	138	590
	8	591	621	564	640	714	754	532	570
	12	475	622	465	645	506	608	415	532
35	4	1	738	3	655	10	742	3	763
	8	526	616	419	467	410	484	374	606
	12	272	362	352	478	392	590	382	459
10	10	303	681	329	710	262	611	301	700
	6	277	757	234	681	263	647	287	801
40	10	387	450	389	553	388	564	318	604
	6	276	662	247	542	248	527	149	591
20	10	351	391	559	650	527	603	476	548
	6	585	643	620	671	437	497	667	771
30	10	484	532	538	605	507	563	508	559
	6	563	666	600	704	562	656	615	723

Here, $\mu_{ij} = E(y_{ij}/m_{ij})$, where y_{ij} is the number of eggs hatched out of the total m_{ij} in the jth tank at the ith combination of salinity and temperature, $j = 1, \ldots, 5, i = 1, \ldots, 17$. We consider the exchangeable correlation structure only among the binomial responses due to convergence issues (Chaganty, Sabo and Deng, 2012). The convergence rates of the three proposed methods using the iterative procedure in section 2 involve different minimization problems. Therefore, they need a little more time for convergence compared to the GEE method.

The estimates of (ϕ, δ) in the proposed variance function (4) are $(0.129, 0.585)$, $(0.028, -0.621)$ and $(0.677, 1.629)$ for the method GEELS, GEEWLS and GEEG respectively. Note that the estimate of (ϕ, δ) by GEEWLS, with negative δ, is very different from those by GEELS and GEEG. This affects the estimates of the regression parameters of model (6) as well. Estimates of the regression parameters and their standard errors by the four methods are given in Table 4.

Table 4: Parameter estimation by GEE, GEELS, GEEWLS and GEEG of β_0, β_1, and β_2.

	Estimates			Standard error		
	β_0	β_1	β_2	β_0	β_1	β_2
GEE	-2.013	0.348	-0.00324	0.743	0.105	0.0284
GEELS	-1.948	0.359	-0.00700	0.757	0.101	0.0273
GEEWLS	-3.696	0.687	-0.00860	1.098	0.175	0.0305
GEEG	-1.796	0.323	-0.00583	0.775	0.0786	0.0275

There does not appear to be a lot difference among the estimates or the regression parameters by GEE, GEELS and GEEG. The estimates by GEEWLS seem to be unusual since β_0 is estimated much smaller and β_1 is estimated much larger. However, the signs of the estimates by the four methods are consistent. Positive β_1 shows that salinity in the range 15-40‰ has positive effect on the hatching efficiency while negative β_2 shows that temperature in the range 4-12C has positive effect on the hatching efficiency. GEELS and GEEG produce estimates of β_1 and β_2 with the smallest standard errors. The usual GEE method leads to the smallest standard error for the estimate of β_0. The results obtained from this example are consistent with the findings in the simulation study.

6. Discussion

In this paper we develop a model for correlated extra-binomial data where two variance parameters are used to account for the heterogeneity amongst clusters in clustered data analysis. We modify the usual GEE approach to estimate the regression parameters in which the variance parameters are estimated by the method of least squares, weighted least squares and Gaussian estimation. Simulations do not show a clear cut conclusion as to which method is the best overall. However, the GEE procedure still shows best overall performance in terms of bias. The GEEWLS performs worst in terms of MSE and variance estimates. Regardless of the working correlation structure, the GEE procedure, where the variance parameters are estimated by the least squares method, performs best in terms of MSE and variance estimates. The GEEG procedure performs best overall in terms of estimated variance and worst in terms of bias. A bias corrected GEEG could perform better in terms of all criteria. This issue and whether any other working correlation structure affect estimation efficiency will be the subject of a future study.

REFERENCES

Alderdice, D.F., and Forrester, C.R. (1968), "Some effects of salinity and temperature on early development and survival of the English sole (Parophrys vetulus)," *Journal of the Fisheries and Research Board of Canada*, 25, 495–521.

Chaganty, N.R., Sabo, R., and Deng, Y. (2012), "Alternatives to mixture modelanalysis of correlated binomial data," *ISRN Probability and Statistics*, doi:10.5402/2012/896082.

Crowder, M.J.(1978), "Beta-binomial anova for proportions," *Applied Statistics*, 27, 34–37.

Liang, K.Y., and Zeger, S.L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 45–51.

McCullagh, P., and Nelder, J.A. (1983), *Generalized Linear Models.* London, UK: Chapman & HALL.

Moore, D.F. (1987), "Modelling the extraneous variance in the presence of extra-binomial variation," *Applied Statistics*, 36, 8–14.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized linear models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Qaqish, B.F. (2003), "A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations," *Biometrika*, 92, 455–463.

Wang, Y.G., and Carey, V.J. (2004), "Unbiased estimating equations from working correlation models for irregularly timed repeated measures," *Journal of the American Statistical Association*, 99, 845–853.

Wang, Y.G., and Zhao, Y. (2007), "A modified pseudolikelihood approach for analysis of longitudinal data," *Biometrics*, 63, 681–689.

Whittle, P. (1961), "Gaussian estimation in stationary time series," *Bulletin of the international statistical institute*, 39, 1–26.

Williams, D.A. (1975), "The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity," *Biometrics*, 31, 949–952.

Zeger, S.L., and Liang, K.Y. (1986), "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, 42, 121–130.