Biometrics 59, 36–42
March 2003

Marginal Analyses of Clustered Data When Cluster Size Is Informative

John M. Williamson,[1,*] Somnath Datta,[2] and Glen A. Satten[3]

[1]Division of HIV/AIDS Prevention, National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, MS E-37, 1600 Clifton Road, NE, Atlanta, Georgia 30333, U.S.A.
[2]Department of Statistics, University of Georgia, Athens, Georgia 30602, U.S.A.
[3]Division of Laboratory Science, National Center for Environmental Health, Centers for Disease Control and Prevention, MS F-24, 1600 Clifton Road, NE, Atlanta, Georgia 30333, U.S.A.
[*]*email:* jow5@cdc.gov

Summary. We propose a new approach to fitting marginal models to clustered data when cluster size is informative. This approach uses a generalized estimating equation (GEE) that is weighted inversely with the cluster size. We show that our approach is asymptotically equivalent to within-cluster resampling (Hoffman, Sen, and Weinberg, 2001, *Biometrika* **73,** 13–22), a computationally intensive approach in which replicate data sets containing a randomly selected observation from each cluster are analyzed, and the resulting estimates averaged. Using simulated data and an example involving dental health, we show the superior performance of our approach compared to unweighted GEE, the equivalence of our approach with WCR for large sample sizes, and the superior performance of our approach compared with WCR when sample sizes are small.

Key words: Cluster-weighted GEE; Generalized estimating equation (GEE); Within-cluster resampling (WCR).

1. Introduction

Since the widespread use of generalized estimating equations (GEEs), it has become standard to fit marginal models to data with cluster dependence (Liang and Zeger, 1986; Zeger and Liang, 1986). GEEs allow using all members of a cluster when fitting a marginal model, by accounting for the correlation between members of the same cluster through the use of working correlation matrices and sandwich variance estimates. An example where this approach may be used is when studying factors associated with periodontal disease. In this case, data from teeth in the same person are correlated, and GEE could be used to fit a marginal model for factors that are associated with a tooth being diseased.

In the GEE approach, the correlation between cluster members is modeled in order to determine the weight that should be assigned to the data from each cluster. Optimally, data are weighted in a way that minimizes the variance of the marginal parameter estimate. In this viewpoint, the fact that data are clustered is incidental to the definition of the marginal model; clustering only enters the analysis to obtain a valid variance estimate. If the outcome measured among cluster members is independent of cluster size (i.e., if cluster size is uninformative), then this viewpoint is valid. However, if cluster size is informative, then the different ways of weighting the data result in different marginal models. In this case, the choice of a working variance-covariance matrix in the GEE approach can affect which marginal model is being fit, resulting in misleading parameter estimates.

When cluster size is informative, two marginal analyses are of particular interest. In the first marginal analysis, we consider associations between explanatory variables and outcome in the population of all cluster members. This marginal analysis corresponds to weighting each cluster member equally, and can be achieved using a GEE with the independence working correlation. This marginal analysis is not considered further in this paper. The second marginal analysis considers associations between explanatory variables and outcome for a typical member of a typical cluster. Given large enough sample sizes, these two marginal analyses will reach the same conclusion if cluster size is unrelated to the outcome of interest. However, if cluster size is related to outcome (i.e., if cluster size is informative), the two marginal analyses are different.

The distinction between the two marginal analyses is most easily seen in an example. Consider a study of factors associated with periodontal disease, in which we have data on the disease status of each tooth from a sample of individuals, and we wish to estimate the association between explanatory variables and the disease status of a tooth. We expect that the disease status of teeth from the same person will be correlated; however, persons with poor dental health are likely to have fewer teeth than do persons with good dental health, because factors that lead to poor periodontal health also lead to tooth loss. As a result, cluster size (number of teeth) is informative. If our goal is to understand the association between explanatory variables and a randomly selected tooth from the population of teeth, a GEE that uses the independence working correlation will produce the desired result. However, if our goal is to understand associations between explanatory variables and a typical tooth from a randomly selected person, using a GEE will oversample healthy teeth. The association between explanatory variables and a typical tooth from a randomly selected person may be more likely to be biologically associated with periodontal disease than the result obtained using GEEs, because some of the association between explanatory variables and outcome obtained from the GEE will be due to the indirect association of that variable and the number of teeth a person has.

If the response data have a maximum number of subunits in a cluster (e.g., in the dental example, we may assume a maximum of 32 teeth per person, assuming no supernumerary teeth), and if not all clusters have the maximum number of subunits, then the data can be analyzed from two perspectives. We may either consider cluster size to be a random variable, as described previously, or we may consider the "true" cluster size to be the maximum, and regard clusters having fewer members as having missing data. In this situation, weighted estimating equations, such as those proposed by Robins, Rotnitzky, and Zhao (1995), may be used. However, if there is no maximum number of subunits making up a complete cluster, this approach is problematic. For example, in a reproductive toxicology experiment with observations made on each pup in a litter of rodents, there is no maximum number of pups in a litter and the random cluster size perspective seems more natural than does the missing data perspective.

Recently, Hoffman, Sen, and Weinberg (2001) proposed a clever Monte Carlo approach to fitting models to clustered data when the goal is estimation of marginal effects weighted at the cluster level. Their approach, termed "within-cluster resampling" (WCR), is to construct a series of data sets by randomly sampling one person from each cluster; the resulting replicate data sets can be analyzed by using any marginal analysis (e.g., logistic regression) because the observations are independent. Parameters can be estimated by taking the average of the parameters estimated from each replicate data set, and the variance can be estimated as the average of a consistent estimator of variance from each replicate data set, minus the variance-covariance matrix of parameter estimates from the replicate data sets.

WCR has a number of appealing features. First, it lends intuition to the meaning of parameters in the marginal model when the goal is analysis of associations between a typical member of a randomly selected cluster, by constructing a sampling scheme that weights clusters equally. Second, it is a valid estimator for the marginal analysis of a typical cluster member when the size of the cluster is informative. Finally, it does not require specification of a within-cluster correlation structure (or even a "working" correlation structure), because all analyses are conducted on replicate data sets that contain independent observations. However, WCR is a Monte Carlo procedure and is computationally intensive.

We propose an alternative to WCR that preserves the advantages of WCR yet uses standard estimating equation methods. We show that our procedure is asymptotically equivalent to WCR and thereby connects WCR and the GEE approach. We also provide a sandwich-type variance-covariance estimator that can be easily computed. The article is organized as follows. In Section 2, we review WCR, introduce our new approach, and consider estimation of the variance-covariance matrix of parameter estimates. We consider marginal analyses of parameters involving pairs of cluster members in Section 3. In Section 4, we present simulation results, to demonstrate the advantages of the proposed method over GEE and WCR, when cluster size is informative. We apply our approach in Section 5, to data on periodontal disease described above, that were also considered by Hoffman et al. (2001). Finally, in Section 6, we discuss our results. Some technical details are found in the Appendix.

2. WCR and Cluster Weighted Generalized Estimating Equations (CWGEE)

Let i index the clusters, with N total clusters, and let j denote individuals within clusters, with n_i individuals in the ith cluster. Let Y_{ij} and $\boldsymbol{X}_{ij}$ denote the response (dependent variable) and explanatory variables for the jth member of the ith cluster with realization y_{ij} and $\boldsymbol{x}_{ij}$. We assume that $\{n_i, Y_{i1}, \boldsymbol{X}_{i1}, \ldots, Y_{in_i}, \boldsymbol{X}_{in_i}\}$ are independent across the clusters. Let I denote a randomly selected cluster, selected using the discrete uniform distribution on $1, \ldots, N$, and given $I = i$, let K_i be a random variable with realization k_i having a discrete uniform distribution on the integers $1, \ldots, n_i$. Let $\boldsymbol{\beta}$ be a vector of parameters of length L that relates the values of Y_{ij} to values of $\boldsymbol{X}_{ij}$, and let $\boldsymbol{U}_i(Y_{ij},\ \boldsymbol{X}_{ij};\ \boldsymbol{\beta})$ be an estimating function to be used for parameter estimation in the ith cluster. Then, for the marginal model we are considering,

$$E\big\{\boldsymbol{U}_I(Y_{IK_I}, \boldsymbol{X}_{IK_I};\ \boldsymbol{\beta})\big\} = \boldsymbol{0} \tag{1}$$

under the true marginal parameter $\boldsymbol{\beta}$. For example, if Y_{ij} are binary, we may choose a logistic model for regression $\boldsymbol{U}_i(y_{ij},\ \boldsymbol{x}_{ij};\ \boldsymbol{\beta}) = \boldsymbol{x}^{'}_{ij}\{y_{ij} - \mu(\boldsymbol{x}_{ij};\ \boldsymbol{\beta})\}$, where $\mu(\boldsymbol{x}_{ij};\boldsymbol{\beta}) = \exp(\boldsymbol{x}^{'}_{ij} \times \boldsymbol{\beta})/\{1 + \exp(\boldsymbol{x}^{'}_{ij}\boldsymbol{\beta})\}$. Then, our marginal model would specify that $E[\boldsymbol{X}'_{IK_I}\{Y_{IK_I} - \mu(\boldsymbol{X}_{IK_I};\ \boldsymbol{\beta})\}] = \boldsymbol{0}$.

For notational simplicity, we will sometimes write $\boldsymbol{U}_{ij}(\boldsymbol{\beta})$ for $\boldsymbol{U}_i(y_{ij},\ \boldsymbol{x}_{ij};\ \boldsymbol{\beta})$. Because $\boldsymbol{U}_I(Y_{IK_I}, \boldsymbol{X}_{IK_I};\ \boldsymbol{\beta})$ has mean zero, $\boldsymbol{\beta}$ may be estimated by solving the estimating equation $\sum_i \boldsymbol{U}_i(y_{ik_i}, \boldsymbol{x}_{ik_i}; \boldsymbol{\beta}) = \boldsymbol{0}$. However, because each cluster contains n_i observations and choice of the cluster member k_i is arbitrary, this estimator is unsatisfactory. In particular, randomly relabeling cluster members will result in a different estimator (although all of these estimators will converge to β as the number of clusters increases); also, each of these estimators only uses data from a single cluster member. In response to these issues, Hoffman et al. (2001) proposed WCR, a Monte Carlo method based on resampling replicate data sets, each containing one person from each cluster. If Q such data sets are generated, and if $\widehat{\boldsymbol{\beta}}_q$ and $\widehat{\boldsymbol{\Sigma}}_q$ are estimates of the parameter and asymptotic variance-covariance matrices of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_q - \boldsymbol{\beta})$, then the WCR estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{wcr} := Q^{-1}\sum_{q=1}^{Q}\widehat{\boldsymbol{\beta}}_q$, and the WCR asymptotic variance-covariance estimator of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{wcr} - \boldsymbol{\beta})$ is

$$\widehat{\boldsymbol{V}}_{wcr} = \frac{1}{Q}\sum_{q=1}^{Q}\widehat{\boldsymbol{\Sigma}}_q - \frac{1}{Q}\sum_{q=1}^{Q}(\widehat{\boldsymbol{\beta}}_q - \widehat{\boldsymbol{\beta}}_{wcr})(\widehat{\boldsymbol{\beta}}_q - \widehat{\boldsymbol{\beta}}_{wcr})^T.$$

Note that each $\widehat{\boldsymbol{\beta}}_q$ is the solution to a score equation $\boldsymbol{S}_q(\boldsymbol{\beta}) = \boldsymbol{0}$, where $\boldsymbol{S}_q(\boldsymbol{\beta}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\boldsymbol{U}_{ij}(\boldsymbol{\beta})I[(i,j) \in r_q]$, where r_q is the set of indices $(i,\ j)$ that are sampled in the qth data set. Let $\boldsymbol{H}_q(\boldsymbol{\beta}) = \partial \boldsymbol{S}_q(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and let

$$\boldsymbol{H}(\boldsymbol{\beta}) = N^{-1}E\{\boldsymbol{H}_q(\boldsymbol{\beta})\}. \qquad (2)$$

Because $\widehat{\boldsymbol{\beta}}_q \approx \boldsymbol{\beta} - N^{-1}\boldsymbol{S}_q(\boldsymbol{\beta})\boldsymbol{H}^{-1}(\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the true parameter value, instead of averaging the $\widehat{\boldsymbol{\beta}}_q$'s as in WCR, we could average the score functions and estimate $\boldsymbol{\beta}$ by solving $Q^{-1}\sum_{q=1}^{Q} \boldsymbol{S}_q(\boldsymbol{\beta}) = \boldsymbol{0}$. Clearly, as $Q \rightarrow \infty$ (as in the WCR proposal), this average converges to its expected value with respect to the resampling distribution, given the original sample. However, because sampling is uniform in each cluster, we may determine analytically the average of $\boldsymbol{S}_q(\boldsymbol{\beta})$ over replicate data sets to be

$$\boldsymbol{\mathcal{U}}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\boldsymbol{\beta}). \qquad (3)$$

Hence, our proposal is to solve

$$\boldsymbol{\mathcal{U}}(\boldsymbol{\beta}) = \boldsymbol{0} \qquad (4)$$

to estimate $\boldsymbol{\beta}$. Note that equations (3) and (4) represent natural implementations of equation (1), which defines the marginal model. We will denote the solution to (4) as $\widehat{\boldsymbol{\beta}}$. If we define $\boldsymbol{\beta}^{*}_{wcr} = \lim_{Q\rightarrow\infty} \widehat{\boldsymbol{\beta}}_{wcr}$, the heuristic argument above can be formalized (see Appendix) to show that $\boldsymbol{\beta}^{*}_{wcr} - \widehat{\boldsymbol{\beta}} \rightarrow \boldsymbol{0}$, as $N \rightarrow \infty$. However, $\widehat{\boldsymbol{\beta}}$ can be obtained by solving a single weighted score function, whereas obtaining $\widehat{\boldsymbol{\beta}}_{wcr}$ requires the computationally intensive WCR approach. We will refer to (4) as the cluster weighted generalized estimating equation (CWGEE).

In the Appendix, we provide a proof of asymptotic normality of $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ obtained by solving (4). The variance-covariance matrix of $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, denoted $\boldsymbol{\mathcal{V}}$, has a sandwich form, and can be consistently estimated by

$$\widehat{\boldsymbol{\mathcal{V}}} = \widehat{\boldsymbol{H}}^{-1}\widehat{\boldsymbol{V}}\widehat{\boldsymbol{H}}^{-1},$$

where

$$\widehat{\boldsymbol{H}} = N^{-1} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \boldsymbol{U}_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$$

and

$$\widehat{\boldsymbol{V}} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\widehat{\boldsymbol{\beta}}) \right\} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\widehat{\boldsymbol{\beta}}) \right\}^{T}.$$

3. Marginal Analysis of Correlation Structure

In Section 2, we described a marginal analysis of a quantity that is defined for single cluster members. We may also wish to consider a marginal analysis of quantities that are defined for pairs of cluster members or even larger groupings. Here, we consider pairwise measures only; generalization to larger groupings should be clear.

Suppose that we wish to estimate a quantity such as a correlation coefficient, but again wish the average to be of the population of clusters, rather than individuals. We will take a similar approach to that of Prentice (1988) and use two sets of estimating equations: the first set to estimate $\boldsymbol{\beta}$, and the second set to estimate the correlation parameters, denoted by $\boldsymbol{\alpha}$. Suppose that $\boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}})$ is the contribution of the pair of observations $(i,\ j)$ and $(i,\ j')$ to an unbiased estimating function for parameters $\boldsymbol{\alpha}$, where $\widehat{\boldsymbol{\beta}}$ solves equation (4). Then, analogous to our results in Section 2, we propose to estimate $\boldsymbol{\alpha}$ by using

$$\boldsymbol{\mathcal{U}}_2(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}) = \boldsymbol{0},$$

where $\binom{n_i}{2}$ is the number of ways of picking the two distinct cluster members, $(i,\ j)$ and $(i,\ j')$, from the n_i total members in the ith cluster. If $\boldsymbol{\alpha}$ is d-dimensional, then in analogy with Section 2, $\sqrt{N}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is asymptotically normally distributed, with a variance-covariance matrix that can be estimated by the lower $d \times d$ block of $\widehat{\boldsymbol{\mathcal{V}}}_{\mathbf{2}}$, where

$$\widehat{\boldsymbol{\mathcal{V}}}_{\mathbf{2}} = \begin{pmatrix} \widehat{\boldsymbol{A}} & \boldsymbol{0} \\ \widehat{\boldsymbol{B}} & \widehat{\boldsymbol{C}} \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\boldsymbol{\Delta}}_{11} & \widehat{\boldsymbol{\Delta}}_{12} \\ \widehat{\boldsymbol{\Delta}}_{21} & \widehat{\boldsymbol{\Delta}}_{22} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{A}}^{T} & \widehat{\boldsymbol{B}}^{T} \\ \boldsymbol{0} & \widehat{\boldsymbol{C}}^{T} \end{pmatrix}^{-1}$$

and where

$$\widehat{\boldsymbol{A}} = N^{-1} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\partial \boldsymbol{U}_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

$$\widehat{\boldsymbol{B}} = N^{-1} \sum_{i=1}^{N} \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \frac{\partial \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \bigg|_{(\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})},$$

$$\widehat{\boldsymbol{C}} = N^{-1} \sum_{i=1}^{N} \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \frac{\partial \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} \bigg|_{(\boldsymbol{\alpha},\boldsymbol{\beta})=(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})},$$

$$\widehat{\boldsymbol{\Delta}}_{11} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\widehat{\boldsymbol{\beta}}) \right\} \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\widehat{\boldsymbol{\beta}}) \right\}^{T},$$

$$\begin{aligned} \widehat{\boldsymbol{\Delta}}_{12} = N^{-1} \sum_{i=1}^{N} & \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{U}_{ij}(\widehat{\boldsymbol{\beta}}) \right\} \\ & \times \left\{ \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \right\}^{T}, \end{aligned}$$

$$\begin{aligned} \widehat{\boldsymbol{\Delta}}_{22} = N^{-1} \sum_{i=1}^{N} & \left\{ \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \right\} \\ & \times \left\{ \frac{1}{\binom{n_i}{2}} \sum_{\substack{j,j'=1\\ j\neq j'}}^{n_i} \boldsymbol{U}_{i;jj'}(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}; \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) \right\}^{T}, \end{aligned}$$

and $\widehat{\boldsymbol{\Delta}}_{21} = \widehat{\boldsymbol{\Delta}}_{12}^{T}$.

It should be noted that the definition (1) of the marginal model may have to be modified if clusters of size less than two are observed, since for these clusters, a correlation parameter is not defined. If this is the case, then the averaging over the distribution of cluster sizes implied in (1) is understood to be restricted to clusters of size greater than or equal to two.

4. Simulations

To assess the performance of our cluster-weighted estimating equation and to compare its performance with the WCR estimator, we conducted several studies with simulated data. The simulation we considered is similar to the example used by Hoffman et al. (2001) to show inconsistency of the (unweighted) GEE approach. Clustered binary data with informative cluster size were simulated by first assigning each cluster a baseline level of risk from a beta(a, b) distribution. The cluster size was then chosen on the basis of the cluster baseline risk. Cluster sizes for clusters with baseline risk of less than $a/(a + b)$ were simulated from a truncated binomial(0.75, 9) distribution, where clusters of size zero, one, eight, or nine were discarded—whereas cluster sizes for the remaining clusters were simulated from a truncated binomial(0.25, 9) distribution, where clusters of size zero, one, eight, or nine were also discarded.

Each simulated data set consisted of two groups (exposed or unexposed) of observations, each corresponding to a different beta distribution. For the unexposed group ($x = 0$), the beta parameters were such that the mean response was 0.25, and the within-cluster correlation was 0.15. For the exposed group ($x = 1$), the beta parameters were such that the mean response was 0.35, and the within-cluster correlation was 0.25. Two sets of simulations were conducted: one with data sets of size $N = 50$, and one with data sets of size $N = 500$. There were an equal number of exposed and unexposed clusters in each data set. For each set of simulations, 10,000 data sets were generated and analyzed with the proposed method, with the WCR method, and with two sets of generalized estimating equations. For all simulations, the marginal probability of response was modeled with a logistic function:

$$\text{logit}\{Pr(Y_{ij} = 1)\} = \beta_0 + \beta_1 x_{ij},$$

where i denoted the cluster $i = 1, \ldots, N$ (N =50, 500*)*, and j denoted the subunit within the cluster ($j = 1, \ldots, n_i$). The covariate x_{ij} was an indicator of whether the ith cluster was exposed, and as such was the same for all cluster members. The within-cluster correlation (denoted by ρ) was also modeled with our approach in a second set of weighted generalized estimating equations and with the usual second set of generalized estimating equations. We used the model $\rho = \alpha_0 + \alpha_1 x_{ij}$ with both approaches. The identity working correlation matrix was used when we analyzed the data with the GEE approach. We used Q = 10,000 resamples when analyzing the data sets with the WCR method.

The simulation results are presented in Table 1. For both sets of simulations (N = 50, 500), the usual GEE approach resulted in severely biased estimates of both the marginal regression and association parameters, as expected. The WCR approach was a notable improvement over the GEE approach for estimating the marginal parameters, although there was still noticeable bias when N = 50. The CWGEE approach resulted in a strong improvement over the WCR method for estimating $\boldsymbol{\beta}$ when N = 50, but only a slight improvement when N = 500. The different performance of CWGEE and WCR at N = 50, as compared to their similar performance at N = 500, is not due to the number of resamples (Q = 10,000), but to the requirement of a large sample size for WCR to approximate CWGEE well.

Table 1
Simulation results of analysis of nonignorable cluster sizes

	True	GEE(i)[a]	WCR[a]	CWGEE(i)[a]
		N = 50		
β_0	−1.099	−1.382	−1.165	−1.126
$se(\widehat{\beta}_0)$		0.273	0.288	0.290
$ese(\widehat{\beta}_0)$[b]		0.282	0.318	0.302
β_1	0.480	0.429	0.518	0.495
$se(\widehat{\beta}_1)$		0.401	0.413	0.414
$ese(\widehat{\beta}_1)$[b]		0.412	0.445	0.427
α_0	0.15	−0.002		0.129
$se(\widehat{\alpha}_0)$		0.064		0.133
$ese(\widehat{\alpha}_0)$[b]		0.072		0.149
α_1	0.10	0.006		0.102
$se(\widehat{\alpha}_1)$		0.106		0.198
$ese(\widehat{\alpha}_1)$[b]		0.116		0.213
		N = 500		
β_0	−1.099	−1.366	−1.104	−1.101
$se(\widehat{\beta}_0)$		0.087	0.093	0.093
$ese(\widehat{\beta}_0)$[b]		0.087	0.093	0.093
β_1	0.480	0.418	0.485	0.483
$se(\widehat{\beta}_1)$		0.128	0.132	0.132
$ese(\widehat{\beta}_1)$[b]		0.129	0.133	0.133
α_0	0.15	0.006		0.149
$se(\widehat{\alpha}_0)$		0.023		0.047
$ese(\widehat{\alpha}_0)$[b]		0.023		0.048
α_1	0.10	0.006		0.100
$se(\widehat{\alpha}_1)$		0.037		0.067
$ese(\widehat{\alpha}_1)$[b]		0.037		0.068

[a] Average of 10,000 simulated data sets.
[b] Empirical standard error of parameter estimate (e.g., $\{\sum_{l=1}^{10,000} (\widehat{\beta}_l - \widehat{\beta}_.)^2/9999\}^{1/2}$, where $\widehat{\beta}_. = \sum_{l=1}^{10,000} \widehat{\beta}_l/10,000$).

The CWGEE approach was a notable improvement over the usual GEE approach when estimating the common correlation parameter, and the bias of the CWGEE estimates were relatively small, even when N = 50. The average standard errors of the parameter estimates were virtually the same for the WCR and CWGEE approaches when N = 500, and very similar when N = 50. The average standard errors of the parameter estimates were also very similar to the empirical standard errors for both approaches when N = 500, and somewhat similar when N = 50.

5. Analysis of Dental Data

We analyzed data from the Intergenerational Epidemiologic Study of Periodontitis (Gansky et al., 1998, 1999), also analyzed by Hoffman et al. (2001), to illustrate our proposed method. Family members spanning three generations were recruited from the Piedmont region of North Carolina. Several dentists examined four sites on every tooth of each participant for periodontal disease (mean clinical attachment level per tooth > 3 mm). Other variables, such as demographic and dental hygiene information, were also recorded.

For this analysis, only the premolars and molars (excluding third molars and wisdom teeth) of the eldest family member

Table 2

Percentage of teeth with periodontal disease by number of teeth

# of teeth	# of persons	% of teeth with periodontal disease
1–8	49	67
9–12	51	50
13–15	49	31
16	57	20
total	206	35

Table 3

Analysis of dental data, standard errors in parentheses

Covariate	GEE(i)	WCR	CWGEE(i)
Intercept	−1.657	−1.447	−1.409
	(0.243)	(0.280)	(0.275)
Age in years	0.049	0.049	0.047
	(0.014)	(0.014)	(0.016)
Ex-smoker*	0.639	0.419	0.410
	(0.237)	(0.241)	(0.240)
Current smoker*	1.678	1.625	1.590
	(0.271)	(0.277)	(0.277)
Plaque*	0.487	0.645	0.623
	(0.156)	(0.189)	(0.185)
Rural residence*	0.368	0.389	0.381
	(0.216)	(0.205)	(0.212)
White*	−0.784	−0.832	−0.816
	(0.211)	(0.212)	(0.213)
Correlation	0.251		0.326
	(0.035)		(0.041)

* 1 = yes, 0 = no.

from the second generation were included. Here, the teeth of the selected individual make up the cluster, and the observation within the cluster is a single tooth. Therefore, the maximum cluster size was 16. The percentage of teeth with periodontal disease increases with the number of teeth per person, indicating that cluster size may be nonignorable for this data set (Table 2). We modeled tooth-specific periodontal disease as a binary variable and considered the effect of a variety of covariates.

Let i denote the person sampled, and let j denote the tooth, $j = 1, \ldots, n_i$, where n_i is the number of teeth in the ith mouth. Denote the presence of periodontal disease in the jth tooth of the ith mouth by $Y_{ij} = 0$, 1, and let $\boldsymbol{x}_{ij}$ denote the corresponding tooth-specific covariate vector. We first arrived at a final model, using our new approach, and then fit the same model with WCR and the usual GEE approach (Table 3). We modeled the probability of periodontal disease in a randomly selected tooth, in a randomly selected person with a logit model:

$$\text{logit}\{Pr(Y_{ij} = 1)\} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij},$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, where x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij}, x_{5ij}, and x_{6ij} denote the covariates age, ex-smoker status, current smoker status, indicator for plaque on the tooth, rural residence, and white race, respectively. Note that only x_{4ij} (plaque on the tooth) varies among all members of the same cluster. To allow for easier comparison with WCR and our proposal, we used the identity “working” correlation matrices when analyzing the data with the usual GEE approach. We used $Q = 500{,}000$ resamples in the WCR analysis.

For our new approach and the usual GEE approach, we presented the empirically corrected standard errors of the parameter estimates (Table 3). After deleting observations with missing covariable data, we had data on 2396 teeth from 206 persons. In general, results from WCR and cluster-weighted GEE are similar; the results from the standard GEE differ by up to 50% (for the effect of being an ex-smoker). We also modeled the correlation in tooth-specific periodontal disease status between two teeth for a single individual. The estimated correlation using cluster-weighted GEE was 30% higher than the estimate obtained by using the standard GEE approach, presumably because the standard GEE gives more weight to teeth from persons with more teeth (hence better overall dental health).

6. Discussion

Cluster size is informative when the response among observations in a cluster is associated with the cluster size. When cluster size is informative, a standard GEE will provide parameter estimates that are weighted by clusters. If marginal analyses are desired with the cluster as the sampling unit, the GEE must be modified so that individuals in large or informative clusters are not overweighted. We have shown how weighting estimating equations by the inverse of the cluster size results in unbiased estimation in this situation. This work was motivated by within-cluster resampling, a computationally intensive Monte Carlo procedure proposed by Hoffman et al. (2001). We have shown that performance equal to that of WCR can be obtained by simpler means.

In the usual GEE approach, the correlation between cluster members is often modeled (choosing a working correlation matrix close to the true one) to increase the efficiency of parameter estimates. However, this approach may lead to difficulties with CWGEE, because the weighting scheme resulting from the inclusion of a working correlation matrix may change the weights given to each cluster. If this occurs, the marginal model is altered and biased parameter estimates (compared to the marginal model of interest) will result. Note also that the marginal model we are considering also depends on the distribution of cluster sizes in the population, because the association between the explanatory variables and the subunit response in a cluster varies with cluster size. Therefore, if the distribution of cluster sizes is different in two populations (e.g., the distribution of the number of teeth may differ due to socioeconomic factors that differentiate the two populations), then the corresponding marginal models will be different, even if the relationship between explanatory and outcome variables conditional on cluster size is the same in the two populations. Finally, it should be noted that when cluster size is actually uninformative, use of the CWGEE will result in a loss of efficiency, compared with the standard GEE with a properly-identified working correlation matrix.

Acknowledgements

We thank Dr. Stuart Gansky and Dr. Jane Weintraub of the University of California at San Francisco for the use of the periodontal data. We also thank the associate editor and two anonymous referees for their critical comments on an earlier version of the manuscript.

Résumé

Nous proposons une nouvelle approche pour ajuster des modèles marginaux sur des données en "clusters" lorsque la taille de ceux-ci est informative. Cette approche utilise une estimation d'équation généralisée (GEE) pondérée par l'inverse de la taille des "clusters". Nous montrons que notre approche est asymptotiquement équivalente au ré-échantillonnage intra "cluster" (WCR) (Hoffman, Sen and Weinberg, 2001, *Biometrika*), une approche coûteuse en temps de calcul dans laquelle sont analysées des répétitions d'ensemble de données contenant une observation tirée au sort à partir de chaque "cluster" et où les estimations résultantes sont moyennées. En utilisant des données simulées et un exemple concernant des soins dentaires, nous montrons la supériorité en terme de performances de notre approche comparée à la GEE non pondérée, son équivalence avec la méthode WCR pour les grands échantillons et de meilleures performances par rapport à cette dernière lorsque les échantillons sont petits.

References

Datta, S. and Hannan, J. F. (1997). A uniform L_1 law of large numbers for functions on a totally bounded metric space. *Sankhya* **59,** 167–174.

Gansky, S. A., Weintraub, J. A., Shain, S., and the Multi-Pied Investigators (1998). Parental periodontal predictors of oral health in adult children of a community cohort. *Journal of Dental Research* **77** (Spec Iss B), 707.

Gansky, S. A., Weintraub, J. A., Shain, S., and the Multi-Pied Investigators (1999). Family aggregation of periodontal status in a two-generation cohort. *Journal of Dental Research* **78** (Special Issue B), 123.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88,** 1121–1134.

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

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

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90,** 106–121.

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

Received July 2001. *Revised September* 2002.
Accepted September 2002.

Appendix

Equivalence between WCR and CWGEE estimators
Under appropriate regularity conditions (as in WCR),

$$\widehat{\boldsymbol{\beta}}_q = \boldsymbol{\beta} - N^{-1}\boldsymbol{S}_q(\boldsymbol{\beta})\boldsymbol{H}^{-1}(\boldsymbol{\beta}) + o_p(1/\sqrt{N}), \qquad \text{(A.1)}$$

where $\boldsymbol{H}(\boldsymbol{\beta})$ is defined in (2) and where the $o_p(1/\sqrt{N})$ term is uniform in $q \geq 1$. Averaging (A.1) w.r.t. k, we get

$$\widehat{\boldsymbol{\beta}}_{wcr} = \boldsymbol{\beta} - N^{-1}\left(Q^{-1}\sum_{q=1}^{Q}\boldsymbol{S}_q(\boldsymbol{\beta})\right)\boldsymbol{H}^{-1}(\boldsymbol{\beta}) + o_p(1/\sqrt{N}) \qquad \text{(A.2)}$$

uniformly in $Q \geq 1$. On the other hand, from the asymptotic normality arguments of $\widehat{\boldsymbol{\beta}}$ given below, we get

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} - N^{-1}\boldsymbol{\mathcal{U}}(\boldsymbol{\beta})\boldsymbol{H}^{-1}(\boldsymbol{\beta}) + o_p(1/\sqrt{N}). \qquad \text{(A.3)}$$

Therefore, from (A.2) and (A.3),

$$\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{wcr}\| \leq N^{-1}\left\|Q^{-1}\sum_{q=1}^{Q}\boldsymbol{S}_q(\boldsymbol{\beta}) - \boldsymbol{\mathcal{U}}(\boldsymbol{\beta})\right\|\|\boldsymbol{H}^{-1}(\boldsymbol{\beta})\| + o_p(1/\sqrt{N}\,) \qquad \text{(A.4)}$$

uniformly in Q. Now, by the law of large numbers applied to the resampling distribution, the first term of the right hand side of (A.4) converges to zero in probability as $Q \to \infty$, and hence after taking limit of (A.4) as $Q \to \infty$, we get $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^*_{wcr} = o_p(1/\sqrt{N})$. This shows that $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^*_{wcr} \to \mathbf{0}$, as $N \to \infty$; furthermore, $\widehat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\beta}}^*_{wcr}$ have the same asymptotic distribution, as $N \to \infty$.

Asymptotic normality of $\widehat{\boldsymbol{\beta}}$
We will show that under reasonable regularity conditions, $\widehat{\boldsymbol{\beta}}$, the consistent solution of the score equation (4), is asymptotically normal.

Let $\boldsymbol{U}_i(\boldsymbol{\beta}; Y, \boldsymbol{X})$ be the score of an individual with response Y and covariate $\boldsymbol{X}$ in cluster i. Suppose that for subunit j in cluster i, their contribution to a marginal score function is $\boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \boldsymbol{U}_i(\boldsymbol{\beta};\, Y_{ij},\, \boldsymbol{X}_{ij})$. Let $\overline{\boldsymbol{U}}_i(\boldsymbol{\beta}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{U}^c_{ij}(\boldsymbol{\beta})$, where $\boldsymbol{U}^c_{ij}(\boldsymbol{\beta}) = \boldsymbol{U}_{ij}(\boldsymbol{\beta}) - E\,\boldsymbol{U}^c_{ij}(\boldsymbol{\beta})$. Note that under the true marginal parameter $\boldsymbol{\beta}$, $\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \sum_{i=1}^{N}\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})$.

Note that $\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})$, $i = 1, \ldots, N$, are independent zero mean random vectors. Note that

$$\begin{aligned}
&E\|\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})\|^{2+\alpha}\\
&\quad = E\left\{E\left(\left\|\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})\right\|^{2+\alpha}\middle| n_i\right)\right\}\\
&\quad \leq 2E\left\{\left(\frac{1}{n_i}\sum_{j}\left\{E\left(\|\boldsymbol{U}_{ij}(\boldsymbol{\beta})\|^{2+\alpha}\,\middle|\, n_i\right)\right\}^{\frac{1}{2+\alpha}}\right)\right\}^{2+\alpha},
\end{aligned}$$

which is bounded in N, provided $E(\|\boldsymbol{U}_{ij}(\boldsymbol{\beta})\|^{2+\alpha}\,|\,n_i) \leq A_i$, with $\sup_i\, E(A_i) < \infty$.

Assume further that the average variance-covariance matrix $N^{-1}\sum_{i=1}^{N}Var(\overline{\boldsymbol{U}}_i(\boldsymbol{\beta}))$ converges to a positive definite matrix $\boldsymbol{V}(\boldsymbol{\beta})$. Hence, for any $d_1 \times 1$ vector $\boldsymbol{a}$,

$$\frac{\sum_{i=1}^{N} E\left|\boldsymbol{a}^T\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})\right|^{2+\alpha}}{\left(\sum_{i=1}^{N}\mathrm{Var}\left(\boldsymbol{a}^T\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})\right)\right)^{\frac{2+\alpha}{2}}} = O(N^{-\alpha/2}) \to 0.$$

Therefore by Lyapunov's CLT,

$$\frac{1}{\sqrt{N}}\boldsymbol{\mathcal{U}}(\boldsymbol{\beta}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\boldsymbol{U}}_i(\boldsymbol{\beta}) \xrightarrow{d} MN(\mathbf{0}, \boldsymbol{V}). \qquad \text{(A.5)}$$

Under a similar conditional moment condition as above, where we replace $\boldsymbol{U}_{ij}(\boldsymbol{\beta})$ by its partial derivative $(\partial/\partial\beta)$ $\boldsymbol{U}_{ij}(\boldsymbol{\beta})$, it follows that $\sup_i E\|-\partial\overline{\boldsymbol{U}}_i/\partial\boldsymbol{\beta}\|^{1+\alpha} < \infty$, and hence $-\partial\overline{\boldsymbol{U}}_i/\partial\boldsymbol{\beta}$ are uniformly integrable. Therefore, by the law of large numbers,

$$N^{-1}\sum_{i=1}^{N}\left(-\partial\overline{\boldsymbol{U}}_i/\partial\boldsymbol{\beta}\right) \xrightarrow{P} \boldsymbol{H}, \qquad \text{(A.6)}$$

where we assume that $\boldsymbol{H}(\boldsymbol{\beta})$, the limit of $N^{-1}\sum_{i=1}^{N} E(-\partial\overline{\boldsymbol{U}}_i/\partial\boldsymbol{\beta})$ exists and is nonsingular. Moreover, under further regularity conditions on $\partial\boldsymbol{U}_{ij}/\partial\boldsymbol{\beta}$ (see, e.g., Datta and Hannan, 1997), the convergence in (A.6) is uniform in $\boldsymbol{\beta}$ on K.

Finally, expanding the estimating function by Taylor expansion one gets

$$\sqrt{N}\,(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left\{\frac{1}{\sqrt{N}}\ \sum_{i=1}^{N}\overline{\boldsymbol{U}}_i(\boldsymbol{\beta})\right\}\left\{-\frac{1}{N}\sum_{i=1}^{N}(\partial\overline{\boldsymbol{U}}_i/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}^*})\right\}$$

$$\xrightarrow{d} MN\left(\mathbf{0}, \boldsymbol{\mathcal{V}} = \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}\right)$$

from (A.5) and the uniform version of (A.6).