Published in final edited form as:

Biometrics. 2011 September ; 67(3): 843–851. doi:10.1111/j.1541-0420.2010.01542.x.

Informative Cluster Sizes for Sub-Cluster Level Covariates and Weighted Generalized Estimating Equations

Ying Huang^{1,*} and Brian Leroux²

¹ Department of Biostatistics, Columbia University, New York, NY, USA

² Department of Biostatistics, University of Washington, Seattle, WA, USA

Summary

Williamson et al. (2003)'s cluster-weighted generalized estimating equations (CWGEE) are effective at adjusting for bias due to informative cluster sizes for cluster-level covariates. We show that CWGEE may not perform well, however, for covariates that can take different values within a cluster if the numbers of observations at each covariate level are informative. On the other hand, inverse probability of treatment weighting accounts for informative treatment propensity but not for informative cluster size. Motivated by evaluating the effect of a binary exposure in presence of such types of informativeness, we propose several weighted generalized estimating equation estimators, with weights related to the size of a cluster as well as the distribution of the binary exposure within the cluster. Choice of the weights depends on the population of interest and the nature of the exposure. Through simulation studies, we demonstrate the superior performance of the new estimators compared to existing estimators such as from GEE, CWGEE, and inverse probability of treatment weighted GEE. We demonstrate the use of our method using an example examining covariate effects on the risk of dental caries among small children.

Keywords

Cluster; GEE; Informative cluster size; IPTW; Propensity Score; Weighting

1. Introduction

In clinical dental research, data are often available on the outcomes of teeth from a sample of patients. Here, a patient can be thought of as a cluster, and tooth outcomes within each cluster are correlated. The Generalized Estimating Equation (GEE) approach has been widely used to estimate marginal covariate effects in such data with cluster dependence. GEE gives consistent estimates of marginal parameter values and their variances when the estimating equation is unbiased.

However, when cluster size is “informative”, i.e., when disease outcome is related to the cluster size, GEE can give biased estimates of marginal mean parameters (Hoffman et al., 2001). This issue is important to dentistry, because disease risk in adults is often inversely related to the number of surviving teeth. As pointed out by Williamson et al. (2003), when cluster size is informative, there are two kinds of marginal analysis that inference could be based on. The first is to explore the association between the covariate and the outcome in the

*yhuang124@gmail.com.

Supplementary Materials

Web Appendix A, referenced in Sections 1 and 5, is available under the Paper Information link at the Biometrics website <http://www.biometrics.tibs.org>.

population of all cluster members (e.g., teeth), where ordinary GEE could be applied with a chosen working correlation matrix. The second is to study the covariate effect for a *typical member of a typical cluster*, where ordinary GEE would over-weight members in larger clusters. When interest is based on the second type of marginal analysis, Hoffman et al. (2001) proposed a within-cluster resampling method for unbiased estimation of the covariate effect. Specifically, they applied independent-observation methods to datasets formed by sampling single observations from each cluster, and then combined the resulting estimates to form an overall estimate. The resampling method is intuitive but computationally intensive and is subject to sampling variability. Later, Williamson et al. (2003) developed the “Cluster-Weighted GEE” method (CWGEE) using inverse weighting based on the overall cluster size. They justified the use of CWGEE by demonstrating its asymptotic equivalence to the within-cluster resampling method.

When the covariate of interest varies within a cluster, a different type of “informativeness” may arise: the expected outcome can be related to the probability of a unit within a cluster having a certain covariate level. When the covariate is a treatment, it is well known that ignoring this type of informativeness can lead to biased estimation of treatment effect. To deal with this kind of informativeness, an inverse probability of treatment weighted (IPTW) method has been proposed where the contribution from a unit is inversely weighted by the probability that one receives its treatment (Robins et al., 2000). Alternatively, between and within cluster components of the covariate can be separated during estimation (Neuhaus and Kalbfleisch, 1998; Mancl et al., 2000; Neuhaus and McCulloch, 2006).

In practice, one can have both informativeness of the total cluster size and informativeness of the covariate distribution. The covariate of interest might be a characteristic inherent to the study unit or a treatment applied to it. The objectives of this paper are to (1) clarify the population parameters of interest in regression models for clustered data in the presence of informativeness for different scenarios of interest; and (2) to propose new weighted GEE methods for unbiased estimation of the population parameters in these scenarios. Throughout this manuscript, we focus on a binary covariate of interest whose distribution within a cluster might be informative, which is not uncommon in clinical practice. The extension to exposure with multiple levels is straightforward and is provided in the Web Supplementary Appendix A. From now on, we call the covariate of interest an “exposure” to differentiate it from other covariates in the regression model and refer to a unit being “exposed” or “unexposed” to indicate the value of the exposure variable.

2. Informative cluster size for sub-cluster exposure and DWGEE

Let i index the cluster and let j index the unit within a cluster. For the j^{th} unit in the i^{th} cluster, let Y_{ij} denote its outcome and let $\mathbf{X}_{ij}$ denote a vector of predictors. Let β be a vector of parameters that relates values of the marginal mean of Y to $\mathbf{X}$ through a generalized linear model (GLM) with link function g , i.e., $g\{E(Y_{ij}|\mathbf{X}_{ij})\} = \mathbf{X}_{ij}^T \beta$. In addition, with a slight abuse of notation, we use X to denote the binary exposure of particular interest whose distribution within a cluster is likely to be related to the outcome. Note that X is a subset of $\mathbf{X}$.

Consider a simple example where $\mathbf{X} = X$ is a binary exposure taking value 0 and 1 and its value varies within a cluster. Suppose the relationship between X and Y can be characterized by a simple linear model:

$$\mu_{ij} = E(Y_{ij}|X_{ij}) = \beta_0 + X_{ij}\beta_1.$$

Let m be the total number of clusters in the sample and let n_i be the number of units in cluster i . Using an independent working correlation matrix, the ordinary GEE with an identity link function solves for

$$U(\hat{\beta}) = \sum_{i=1}^m \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}) = 0,$$

with

$$U_{ij}(\beta) = \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} (Y_{ij} - \beta_0 - \beta_1 X_{ij}).$$

The corresponding estimates of group means are

$$\hat{\mu}_x^{GEE} = \sum_{i=1}^m n_{ix} \bar{y}_{ix} / \sum_{i=1}^m n_{ix}, \quad x=0, 1,$$

where n_{ix} and $\bar{y}_{ix}$ are the number of and the average outcome for units with exposure status $X = x$ in cluster i respectively. The CWGEE (Williamson et al., 2003) inversely weights each unit's contribution by the size of the cluster it belongs to

$$U(\hat{\beta}) = \sum_{i=1}^m \frac{1}{n_i} \sum_{j=1}^{n_i} U_{ij}(\hat{\beta}) = 0. \quad (1)$$

The corresponding mean estimates are given by

$$\hat{\mu}_x^{CWGEE} = \sum_{i=1}^m \frac{n_{ix}}{n_i} \bar{y}_{ix} / \sum_{i=1}^m \frac{n_{ix}}{n_i}, \quad x=0, 1.$$

Note that if the expected outcome of a unit either unexposed or exposed within a cluster is associated with the probability of being exposed in this cluster, then both $\hat{\mu}_1^{GEE} - \hat{\mu}_0^{GEE}$ and $\hat{\mu}_1^{CWGEE} - \hat{\mu}_0^{CWGEE}$ are misleading for the exposure effect β_1 . For instance, if probability of having $X = 1$ is larger with poorer outcomes, then the exposure effect will appear to be worse than it actually is.

Next we propose estimators for covariate effects when informativeness might occur at both cluster and sub-cluster levels. That is, both the size of a cluster and the distribution of a binary exposure within the cluster could be related to outcome. The binary exposure of interest could either be an inherent property of the unit that can not be manipulated, or a treatment applied to it. Next we investigate the two scenarios separately. For non-manipulable exposures, two sub-scenarios are presented.

2.1 Non-manipulable exposure

First we consider an exposure that is a characteristic of the unit under study. Being an inherent property of a unit, this type of exposure is not manipulable. Furthermore the size of a cluster and the distribution of the exposure within the cluster can be observed simultaneously. An example is presented in Section 4, where we analyze data from an observational study of caries in children. There the cluster is an individual child whose teeth observed at year 1 followup are units under study; the binary exposure is the presence of an incipient lesion at baseline, and the outcome is the presence of dental caries at year 1 followup. In practice, it is important to distinguish between settings where all clusters are expected to have both types of exposure and settings where only one level of exposure is present in some clusters. Next we study the two settings separately.

2.1.1 Both Exposure Levels Present in Each Cluster—Suppose with probability one each sampled cluster will have both levels of exposure present. This assumption was made in Rieger and Weinberg (2002) when using within-cluster paired resampling. For example, the comparison between anterior and posterior teeth might be of interest among a cohort with relatively good dental health, such that each of them are expected to have both types of teeth.

We are interested in making inference for a *typical* unit within a *typical* cluster. Here the meaning of *typical* is different when a unit is concerned compared to when a cluster is concerned. The latter corresponds to simple random selection, whereas for the former, *typical* is conditional on the value of the binary exposure X whose distribution within a cluster is likely to be related to the outcome. Under independent working correlation, we consider estimating equation for unit j in cluster i being $U_{ij}(Y_{ij}, \mathbf{X}_{ij}, \beta) = h(\mathbf{X}_{ij}, \beta) \{Y_{ij} - g^{-1}(\mathbf{X}_{ij}^T \beta)\}$, where $h(\mathbf{X}_{ij}, \beta)$ is a function of $\mathbf{X}_{ij}$ and β . Let I be a randomly selected cluster out of the m clusters in the sample. The parameter β we are interested in is a solution to the marginal model estimating equation

$$E\{U_{IK_I}(Y_{IK_I}, \mathbf{X}_{IK_I}, \beta)\} = 0, \quad (2)$$

where K_i indicates a unit selected with discrete sampling probability $1/(2n_{i0})$ for those with $X = 0$ and selected with probability $1/(2n_{i1})$ for those with $X = 1$ in cluster i , with n_{i0} and n_{i1} being the number of units in cluster i with $X = 0$ and $X = 1$ respectively.

Consider the simple setting where X is the only independent predictor, assuming $g(\mu_X) = \beta_0 + \beta_1 X$. Then the population parameters of interest are $\beta_0 = g(\mu_0)$ and $\beta_1 = g(\mu_1) - g(\mu_0)$, where $\mu_X = E_I E(Y_{ij} | X_{ij} = x, I = i)$ for $x = 0, 1$ and E_I indicates taking expectation over clusters.

According to our definition of the population of interest, it is intuitive to see that a solution to (2) can be obtained by applying a resampling scheme analogous to Hoffman et al. (2001). That is, we sample a unit from each cluster according to the distribution of K_j . Specifically, we flip a coin to decide the exposure level and then randomly select a unit with this exposure level in the selected cluster. We can then apply independent-observation methods to this sample dataset and get one parameter estimate. Do this multiple times and combine the resulting estimates to form an overall estimate.

The within-cluster resampling method, of course, is computationally intensive and subject to resampling variability. For a given study dataset, the value of the estimator varies with the different units resampled, and its variability depends on the amount of resampling performed. Instead, a natural extension of the weighting method in Williamson et al. (2003)

can be implemented. To estimate β , we inversely weight the contribution to the estimating equations from a unit within a cluster by the number of observations in the cluster with the same exposure value. That is, we estimate β as the solution to

$$U(\hat{\beta}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{1}{n_{ix_{ij}}} U_{ij}(\hat{\beta}) = 0 \quad (3)$$

where $n_{ix_{ij}}$ is the number of observations in cluster i with the same exposure value as unit j . We call this estimating equation the type-1 doubly weighted generalized estimating equation (DWGEE1).

The asymptotic equivalence between solution of DWGEE1 and that of the within cluster resampling follows from similar arguments as in Williamson et al. (2003). Variance of this estimator can be estimated by a sandwich type estimator $\hat{A}^{-1} \hat{B} \hat{A}$ from standard GEE software, where

$$\hat{A} = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_{ix_{ij}}} \sum_{j=1}^{n_i} \left. \frac{\partial U_{ij}(\beta)}{\partial \beta} \right|_{\beta=\hat{\beta}}$$

$$\hat{B} = \frac{1}{m} \sum_{i=1}^m \left\{ \sum_{j=1}^{n_i} \frac{1}{n_{ix_{ij}}} U_{ij}(\hat{\beta}) \right\} \left\{ \sum_{j=1}^{n_i} \frac{1}{n_{ix_{ij}}} U_{ij}(\hat{\beta}) \right\}^T.$$

2.1.2 Clusters Possibly with One Exposure Level Not Present—For scenarios described in Section 2.1.1, the DWGEE1 estimator offers a simple and relatively robust way to account for informative cluster size without making extra assumptions about the mechanism of the informativeness. In some applications, however, some clusters in the sample might have one exposure level not present. When this happens, there are two strategies we could consider depending on our definition of the population. On the one hand, we can restrict the application of DWGEE1 to clusters with both types of exposure only, supposing we are not interested in contributions from clusters with only one type of exposure. Oftentimes, however, we would like to account for the fact that a cluster with only one exposure level could have had the underlying outcome for the other absent exposure level that we might be interested in. Applying the DWGEE1 estimator excluding those clusters could be misleading if the exposure effect tends to be different between clusters with one and both types of exposure. Here the concept of exposure effect involves extrapolation. For those clusters with only one level of exposure, we want to account for the absent exposure in the inference. For example, a patient in a dental study who has lost all their molar teeth may be different in important ways from other patients. Either exclusion of the patient or unweighted inclusion of the patient's (non-molar) teeth in the analysis may lead to bias.

To infer the underlying 'exposure' effect requires modeling the numbers of units with given values of the exposure on the basis of other covariates available. We make the assumption that for each type of exposure, the number of units in a cluster depends only on some observable cluster-level characteristics. Under this assumption, the mechanism of not observing a cluster member with a particular type of exposure is missing at random. To get valid inference for solution of (2), we modify (3) so that, for each exposure level, the contribution of a unit with this exposure is inversely weighted by the *expected* (rather than observed) number of units with this exposure level in the cluster that the unit belongs to.

That is, we propose the type-2 Doubly Weighted Generalized Estimating Equations (DWGEE2) estimator which solves for,

$$\sum_{i=1}^m \sum_{j=1}^{n_i} \frac{U_i(Y_{ij}, \mathbf{X}_{ij}, \hat{\beta})}{\pi_{ix_{ij}}} = 0,$$

where $\pi_{ix_{ij}}$ is the expected number of units included in cluster i for exposure level equal to X_{ij} . A heuristic proof that this yields unbiased estimators is provided in Appendix A1 based on a survey framework.

We estimate π_{ix} for exposure $x = 0, 1$ separately based on observed cluster-level covariates. For $x = 0$ or 1 , suppose that there is a natural maximum value for the number of N_x units in each cluster (which does exist in dental applications), whereas there are n_{ix} units present in the i^{th} cluster in the sample. Then we can estimate π_{ix} by modeling the distribution of n_{ix} out of N_x as a binomial variable with mean π_{ix} being a function of the cluster-level covariates. Alternatively, if there does not exist a natural maximum size (N_x), then n_{ix} can be modeled as a Poisson random variable with mean π_{ix} being a function of the cluster-level covariates.

2.2 Manipulatable Exposure

In this section, we consider the scenario where the binary exposure of interest is a treatment applied after the sample is collected. Consider a hypothetical example where we study effect of a treatment based on patients outcomes from several clinics. A patient's outcome can be related to clinic size as well as the probability a patient gets the active treatment in a clinic. For example, the probability a patient receives treatment could depend on the patient's insurance status, whereas patients with insurance may tend to have better outcomes. Here we use slightly different notations from before to accommodate the potential outcomes. Again let Y be the outcome of interest and X be the treatment indicator. We use Z to indicate other covariates to be adjusted in the marginal model. For $x = 0, 1$, let $D(x)$ be the binary indicator of receiving treatment x , and let $Y(x)$ be the potential outcome of Y given treatment x . Let W be a vector of pre-treatment variables, such that a weak unconfoundedness assumption holds: $D(x) \perp Y(x) | W$ (Imbens, 2000). In addition, assume Z is a subset of W and let $\beta = (\beta^X, \beta^Z)$ be a vector of parameters that relates values of $Y(x)$ to x and additional covariates Z through a GLM model, $g[E\{Y(x)|x, Z\}] = x\beta^X + Z^T\beta^Z$.

Let $U_{ij}(Y_{ij}, X_{ij}, Z_{ij}; \beta) = h(X_{ij}, Z_{ij}; \beta) \{Y_{ij} - g^{-1}(X_{ij}\beta^X + Z_{ij}^T\beta^Z)\}$ be an estimating function for unit j in cluster i . Let $U_{ij}^{(x)}(Z_{ij}; \beta) = h(x, Z_{ij}; \beta) \{Y_{ij}(x) - g^{-1}(x\beta^X + Z_{ij}^T\beta^Z)\}$ for $x = 0, 1$. We are interested in evaluating the potential treatment effect on the outcome adjusting for other covariates for a *typical* unit within a *typical* cluster. Note here the definition of a *typical* unit is different from that in Section 2.1, which was conditional on exposure status. Here each unit has the potential of receiving each type of treatment, so a *typical* unit within a cluster is randomly selected from all units. The population parameters we are interested in solve for

$$EU_{IJ}^{(x)}(Y_{IJ}(x), x, Z_{IJ}, \beta) = 0, \quad (4)$$

for $x = 0, 1$, where I indicates a random cluster and, for $I = i$, J_i is a random variable with a discrete uniform distribution on the integers $1, \dots, n_i$.

Let $r(x, w) = P(X = x|W = w)$, we propose the type-3 Doubly Weighted Generalized Estimating Equations (DWGEE3) estimator to solve the following estimating equation where the contribution from each unit is inversely weighted by the product of cluster size and $r(X, W)$, the probability it receives the current treatment,

$$\sum_{i=1}^m \sum_{j=1}^{n_i} \frac{U_{ij}(\hat{\beta})}{n_i r(X_{ij}, W_{ij})} = 0.$$

The unbiasedness of this equation for estimating the parameters defined in (4) is demonstrated in Appendix A2.

Let $e(W) = P(X = 1|W)$ be the propensity score given W such that $r(1, W) = e(W)$ and $r(0, W) = 1 - e(W)$. In practice $e(W)$ can be estimated using a logistic model (Rosenbaum, 1987). Note that in the scenario of Section 2.1 where X is an inherent characteristic of the unit under study, we had to model the relative frequencies of exposure levels based on observable cluster-level covariates. Here when exposure is a treatment, however, covariates W for an individual unit can be observed even if one of the exposure levels is not present. As a result, missing at random is ensured by the unconfoundedness assumption.

In the special case where the experiment is designed such that both types of treatment are always present in each cluster and in addition the probability of receiving a treatment depends on cluster-level covariates only, then the DWGEE1 estimating equation (3) defined in Section 2.1.1 can be applied since $n_i X_{ij}$ would be a consistent estimate of $n_i r(X_{ij}, W_{ij})$.

In summary, depending on the nature of the exposure we propose different methods for modeling the exposure frequency. Specifically, for a manipulable treatment, presence or absence of a potential outcome given $X = x$ corresponds to the absence or presence of the other potential outcome given $X = 1 - x$, so only one model is needed; on the other hand, for nonmanipulable exposure, we allow separate modeling for frequencies of exposed and unexposed units, assuming conditional independence between different exposure levels after conditioning on some cluster-level covariates.

3. Simulation study

We evaluate performance of the proposed estimators in the presence of informativeness for a binary sub-cluster level exposure X , and a binary disease outcome Y . We are interested in making inference for a typical unit j within a typical cluster i based on the marginal model

$$\text{logit}\{P(Y_{ij}=1|X_{ij})\}=\beta_0+\beta_1 X_{ij},$$

where a typical unit is defined unconditional or conditional on the value of X for manipulable or nonmanipulable X respectively.

Three mechanisms of informativeness were examined. The exposure is not manipulable in the first two but is manipulable in the third setting. For each setting, we assume there is an observable cluster-level covariate $W_i \sim N(0, 1)$, which is related to cluster size and outcome Y , and which we use for modeling the exposure frequency distribution.

In the first setting (Table 1), for $x = 0$ and 1 separately, the number of units with value $X = x$ in cluster i follows a *binomial*(40, p_{ix}) distribution with logit of p_{ix} linearly increasing in W_i ,

and we allow the intercept and slope to differ between x . In the second setting (Table 2), for $x = 0$ and 1 separately, number of units with value $X = x$ in cluster i follows a $\text{poisson}(40\lambda_x)$ distribution with log of λ_{ix} linearly increasing in W_i . In both settings, Y_{ij} is generated with logit mean equal to $\beta_0 + \beta_1 X_{ij} + W_i$.

For the third setting (Tables 3 and 4) we assume X is a manipulable treatment and we allow the total number of units and the probability of active treatment within each cluster to depend on different cluster-level covariates. Specifically, total sample size n_i in cluster i follows a $\text{binomial}(40, q_i)$ distribution with logit of q_i a linear function of an *unobservable* cluster-level covariate V_i , while conditional on n_i , the logit of the probability of $X = 1$ in cluster i is linearly increasing in W_i . In our simulations, the unobservable V_i is generated from $N(0, 1)$ independent of W_i and is associated with the outcome as well. Two scenarios were considered depending on whether V_i affects treatment effect within a cluster. In the first scenario, within-cluster treatment effect is the same across clusters: Y_{ij} is generated with logit mean equal to $W_i - 2X_{ij} + V_i$. In the second scenario, treatment effect is allowed to vary with V_i : logit of mean of Y_{ij} equals to $W_i - 2X_{ij} + V_i X_{ij}$.

For each setting, two scenarios were considered. The chance that any cluster is missing one exposure level is zero in the first scenario (Tables 1(a), 2(a), 3(a), 4(a)) and nonignorable (8%–12%) in the second one (Tables 1(b), 2(b), 3(b), 4(b)). Using 5000 Monte-Carlo simulations, bias and coverage of 95% Wald confidence intervals (based on sandwich variance estimates) were compared between the proposed estimators and several other estimators for estimating β_0 and β_1 . The correct model for exposure levels based on W_i is assumed in each setting with parameters estimated from the data to obtain weights for DWGEE2 and DWGEE3. Besides GEE, CWGEE, DWGEE1, DWGEE2 (for non-manipulable exposure) and DWGEE3 (for manipulable treatment), we also examine a few other estimators commonly used in practice, including GEEN, where a regular GEE is fitted with n_i (the size of each cluster) added as an additional covariate to adjust for; SBW, the GEE estimator where the between and within-cluster exposure effects are separated; CWSBW, the weighted SBW estimator with each observation inversely weighted by the size of the cluster it belongs to; and SBWN, the extended SBW estimator adjusting for n_i . In the third setting where X is a treatment, the established IPTW estimator is also examined. An independent working correlation matrix is employed for all the estimators.

Results are presented in Tables 1–4. Whether or not the exposure is manipulable, the regular GEE estimator is usually biased for both β_0 and β_1 in presence of informative cluster size. Inverse weighting by cluster size or simply adjusting for total cluster size does not help with the performance. Without accounting for informativeness in total cluster size, the SBW estimator does not provide unbiased estimates of within-cluster exposure effect, and further inverse weighting by n_i or adjusting for n_i in the regression model does not help much either. Note that when the exposure is a binary treatment, average exposure within a cluster is an estimate of the propensity score (Rosenbaum, 1987; Rosenbaum and Rubin, 1983). Hence the SBW estimator separating between- and within- cluster effects is essentially adjusting for the propensity score. Separating between- and within- cluster covariate effects combined with accounting for informative total cluster size either through inverse weighting (CWSBW) or through adjustment for n_i (SBWN) attempt to adjust for informativeness both at overall cluster size and at sub-cluster level. However, they are substantially biased when a linear term in average exposure is not enough to characterize the relationship with the outcome.

Where a binary treatment is concerned (Tables 3 and 4), with a treatment effect that is invariant between clusters, the IPTW estimator is satisfactory for estimating β_1 in our simulation setting but is biased where β_0 is concerned (Table 3). When the treatment effect

is allowed to vary between clusters, IPTW estimator becomes biased for estimating β_1 as well (Table 4).

In contrast, the DWGEE2 estimator is unbiased for both β_0 and β_1 with good coverage probability for non-manipulated exposure (Tables 1 and 2), and the DWGEE3 estimator is unbiased for β_0 and β_1 with good coverage for manipulable exposure (Tables 3 and 4), whether or not there exists clusters with only one level of exposure. When it is unlikely that any cluster has only one exposure level, the more ‘nonparametric’ DWGEE1 estimator is also unbiased for β_0 and β_1 . When clusters with one exposure level are involved, the magnitude of the bias for DWGEE1 is relatively small compared to other estimators in our simulation studies.

4. Illustration

We illustrate our methods using a real example studying dental caries in young children. Data come from a study designed to collect pilot data for purpose of planning a caries prevention trial. We use year 1 followup data of this study as the analysis set. Here every child is a cluster and each tooth is a sub-cluster unit. We evaluate the effect of having incipient lesions (yes vs no) at baseline on tooth decay at the same surface as indicated by the presence or absence of caries on buccal surface at year 1 followup. Only teeth with no caries on buccal surface at baseline are included. Additional risk factors we are interested in assessing include gender and tooth-level covariates that indicate whether a tooth is upper molar, lower molar, upper central incisor, or lower central incisor. Our analysis set is comprised of 118 children 3 to 5 years old, 45% of which are girls. The total number of teeth each child has ranges from 10 to 20 (mean 18.6, SD 2.5). Overall 55 of the children have both teeth with and without baseline incipient lesions on buccal surface, whereas 63 have teeth all without baseline incipient lesions on buccal surface.

First we apply generalized estimating equations for the binary outcome using a logit link with independent working correlation to the data, adjusting for the presence of baseline incipient lesions on buccal surface, gender and indicators of tooth location. There appears to be a significant effect of baseline incipient lesions at the 0.05 level of significance (GEE in Table 5). After including the cluster size n_i as an additional covariate to adjust for, a significant effect of n_i is also found (Table 5, GEEN), suggesting a deterioration in tooth health in children with smaller number of teeth. Moreover, in a GEE model separating between- and within- cluster effect of having baseline incipient lesions, the difference in magnitude of the two components suggest a possible informativeness with respect to this exposure (Table 5, SBW). This is not surprising considering that the proportion of teeth with incipient lesions at baseline is related to dental health, which in turn relates to the extent of tooth decay at the year 1 followup.

To account for possible informativeness of cluster size and the proportion of teeth with baseline incipient lesions, we generate DWGEE1 and DWGEE2 estimators using the indicator of having baseline incipient lesions as the binary exposure of interest. For the latter, the number of teeth with or without incipient lesions on buccal surface for a child at baseline is fitted using a Poisson model with log-transformed mean modeled as a linear function of the child’s gender and total number of teeth. Table 5 presents log odds ratio estimates using GEE, CWGEE, DWGEE1, and DWGEE2. Different methods differ in magnitude of coefficient estimates as well as their significance level. For example, although estimates of the effect for baseline incipient lesions are significant for all models at 0.05 level, the magnitude of the log odds ratio estimate decreases by 10% in CWGEE, 34% in DWGEE1, and 31% in DWGEE2. Compared to GEE, the magnitude of the effect of being a lower central tooth decreases by 5% in CWGEE, 35% in DWGEE1, and 27% in DWGEE2.

5. Concluding Remarks

In this paper, we studied the population parameters of interest for marginal regression analysis in the presence of “informative” cluster size for binary within-cluster level covariates. Extensions of Williamson et al.’s (2003) CWGEE method for cluster-level covariates to within-cluster level covariates were studied and compared with other methods for dealing with informativeness with respect to covariate distribution in addition to informativeness with respect to total cluster size. The weighting method under informative cluster size is closely related to the inverse probability weighted GEE for dealing with missing data problem if we think of “informativeness” as a special type of ‘missingness’. On the other hand, the two are different conceptually, as they have been developed to answer different types of questions. Inverse probability weighted GEE were developed for the classical missing data problem where the actual observations exist but failed to be collected, whereas in “informative cluster size” problem, the ‘missingness’ is with respect to a pseudo-population where the inclusion probabilities of each type of units are specified. Consider the cluster-level exposure as an example. As pointed out by Williamson et al. (2003), when there exists a maximum number of units in each cluster, the weighting by cluster size is in fact equivalent to the inverse probability weighted method. This is also true for sub-cluster level covariates if there exists a maximum number of units in each cluster with any particular exposure level and weighting is performed for each exposure level separately. However, when there does not exist a maximum number of units in each cluster, the inverse probability weighted technique does not directly apply.

We propose different weighted GEE methods to handle the informativeness, for the scenario where the exposure is an inherent characteristic of the study unit and also for the scenario where the exposure can be manipulated. Under the correctly specified mechanism of informativeness, our proposed estimators that model the frequency distribution of the exposure parametrically (DWGEE2 and DWGEE3) are consistent and have satisfactory performances in all the simulation settings examined, when other GEE methods could be biased. The validity of these estimators, of course, relies on proper assumptions about the exposure frequency distribution. In the setting when the probability of one exposure not being present in any cluster is small, a more nonparametric weighting method (DWGEE1) can be applied and might be preferred since it does not rely on extra assumptions about the exposure frequency distribution considering that a good set of baseline covariates may not always be available in practice. Future work is needed to examine the trade-off between ignoring clusters with only one covariate level with DWGEE1 and the need to model the exposure frequency distribution with DWGEE2 or DWGEE3.

The methodology introduced in this manuscript based on a binary exposure can be easily extended to the multinomial exposure status (as shown in the Web Supplementary Appendix A). The extension to continuous exposure is non-trivial and is currently under investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge funding from the National Institute of Dental and Craniofacial Research grant #DE15651.

References

Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001; 88:1121–1134.

- Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000; 87(3):706–710.
- Manc LA, Leroux BG, DeRouen TA. Between-subject and within-subject statistical information in dental research. *J Dent Res*. 2000; 79(10):1778–1781. [PubMed: 11077994]
- Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *J R Statist Soc B*. 2006; 5:859–872.
- Neuhaus JM, McCulloch CE. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*. 1998; 54:638–645. [PubMed: 9629647]
- Rieger RH, Weinberg CR. Analysis of clustered binary outcomes using within-cluster paired resampling. *Biometrics*. 2002; 58:332–341. [PubMed: 12071406]
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11(5):550–560. [PubMed: 10955408]
- Rosenbaum PR. Model-based direct adjustment. *JASA*. 1987; 82:387–394.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983; 70:41–56.
- Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics*. 2003; 59:36–42. [PubMed: 12762439]

A1. Unbiasedness of DWGEE2

Suppose we have a finite population of clusters, $i = 1, \dots, M$, and cluster i has $N_{i1} > 0$ exposed and $N_{i0} > 0$ unexposed observations. In our sample, we have m clusters and n_{i1} and n_{i0} observations exposed and unexposed respectively in the i^{th} cluster. Note here either n_{i1} or n_{i0} could be equal to zero. Let $U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta)$ and $U_{i1j}(Y_{ij}, \mathbf{X}_{ij}, \beta)$ be estimating equations to be used in the i^{th} cluster based on a unit with exposure level $X = 0$ or $X = 1$ respectively. First, considering $U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta)$, let R_{i0j} indicate the event that unit j in cluster i with

exposure 0 is sampled such that $\pi_{i0} = \sum_{j=1}^{N_{i0}} ER_{i0j} \times M/m$, then

$$E \sum_{i=1}^m \sum_{j=1}^{n_{i0}} \frac{U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta)}{\pi_{i0}} = \frac{m}{M} E \sum_{i=1}^M \sum_{j=1}^{N_{i0}} \frac{U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta) R_{i0j}}{E \sum_{i=1}^M R_{i0j}} = \frac{m}{M} \sum_{i=1}^M \sum_{j=1}^{N_{i0}} \frac{U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta) ER_{i0j}}{N_{i0} ER_{i0j}} = \frac{m}{M} \sum_{i=1}^M \sum_{j=1}^{N_{i0}} \frac{U_{i0j}(Y_{ij}, \mathbf{X}_{ij}, \beta)}{N_{i0}},$$

similarly

$$E \sum_{i=1}^m \sum_{j=1}^{n_{i1}} \frac{U_{i1j}(Y_{ij}, \mathbf{X}_{ij}, \beta)}{\pi_{i1}} = \frac{m}{M} \sum_{i=1}^M \sum_{j=1}^{N_{i1}} \frac{U_{i1j}(Y_{ij}, \mathbf{X}_{ij}, \beta)}{N_{i1}}.$$

Consequently we have

$$E \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{U_i(Y_{ij}, \mathbf{X}_{ij}, \beta)}{\pi_{ix_{ij}}} = \frac{m}{M} \sum_{i=1}^M \sum_{j=1}^{N_i} \frac{U_{ij}(Y_{ij}, \mathbf{X}_{ij}, \beta)}{N_{ix_{ij}}},$$

which is the population version of (2).

A2. Unbiasedness of DWGEE3

Let $r(x, w) = P(X = x|W = w)$, then we have $D(x) \perp Y(x)|r(X, W)$ and $D(x) \perp Z|r(X, W)$. To see the unbiasedness of this estimating equation, let $e(W) = P(X = 1|W)$ be the propensity score given W , we have

$$\begin{aligned}
 E \sum_{i=1}^m \frac{1}{m} \sum_{j=1}^{n_i} \frac{U_{ij}(\beta)}{n_i r(X_{ij}, W_{ij})} &= E \sum_{i=1}^m \sum_{j=1}^{n_i} \left[\frac{U_{ij}^{(1)}(\beta) D_{ij}(1)}{n_i e(W_{ij})} + \frac{U_{ij}^{(0)}(\beta) D_{ij}(0)}{n_i \{1 - e(W_{ij})\}} \right] \\
 &= \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \left[E \left\{ \frac{U_{ij}^{(1)}(\beta) D_{ij}(1)}{e(W_{ij})} \right\} + E \left\{ \frac{U_{ij}^{(0)}(\beta) D_{ij}(0)}{1 - e(W_{ij})} \right\} \right] \right) \\
 &= \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \left(E \left[E \left\{ \frac{U_{ij}^{(1)}(\beta) D_{ij}(1)}{e(W_{ij})} \middle| e(W_{ij}) \right\} \right] + E \left[E \left\{ \frac{U_{ij}^{(0)}(\beta) D_{ij}(1)}{1 - e(W_{ij})} \middle| 1 - e(W_{ij}) \right\} \right] \right) \right\} \\
 &= \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \left(E \left[\frac{1}{e(W_{ij})} E \left\{ U_{ij}^{(1)}(\beta) | e(W_{ij}) \right\} E \{ D_{ij}(1) | e(W_{ij}) \} \right] + E \left[\frac{1}{1 - e(W_{ij})} E \left\{ U_{ij}^{(1)}(\beta) | 1 - e(W_{ij}) \right\} E \{ D_{ij}(0) | 1 - e(W_{ij}) \} \right] \right) \right\} \\
 &= \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{n_i} \sum_{j=1}^{n_i} \left(E \left[E \left\{ U_{ij}^{(1)}(\beta) | e(W_{ij}) \right\} \right] + E \left[E \left\{ U_{ij}^{(0)}(\beta) | 1 - e(W_{ij}) \right\} \right] \right) \right\} \\
 &= \frac{1}{m} \sum_{i=1}^m \left[E \left\{ U_{ij}^{(1)}(\beta) \right\} + E \left\{ U_{ij}^{(0)}(\beta) \right\} \right] \\
 &= E \{ U_{ij}^{(1)}(Y_{ij}(1), 1, Z_{ij}, \beta) \} + E \{ U_{ij}^{(0)}(Y_{ij}(0), 0, Z_{ij}, \beta) \} = 0.
 \end{aligned}$$

Table 1

Performance of different estimators of β_0 , β_1 for a non-manipulable binary exposure. Here the number of observations in a cluster with $X = x$ follows a binomial distribution.

$\beta_0 = 0$						$\beta_1 = -1.69$					
	Bias	SE	ECP★	Bias	SE	ECP★		Bias	SE	ECP★	
(a): 0% clusters with only one exposure level present											
GEE	0.157	0.149	79.6	-0.155	0.127	78.0					
CWGEE	0.099	0.147	88.8	-0.164	0.128	75.9					
GEEN	-4.697	1.136	1.1	-0.150	0.132	79.3					
SBW				-0.104	0.132	88.0					
CWSBW				-0.105	0.132	87.9					
SBWN				-0.131	0.133	83.7					
DWGEE1	-0.001	0.148	94.1	-0.007	0.133	94.8					
DWGEE2	-0.001	0.146	94.3	-0.007	0.131	95.1					
(b): 9% clusters with only one exposure level present											
GEE	0.625	0.167	4.4	-0.443	0.164	22.1					
CWGEE	0.361	0.166	40.3	-0.559	0.184	13.6					
GEEN	-1.679	0.338	0.1	-0.330	0.175	51.8					
SBW				-0.190	0.171	79.8					
CWSBW				-0.189	0.185	82.7					
SBWN				-0.282	0.177	63.0					
DWGEE1	0.125	0.195	88.5	-0.134	0.209	90.9					
DWGEE2	0.009	0.213	93.9	-0.018	0.216	94.9					

★ coverage of 95% Wald confidence interval based on sandwich variance estimate
Here $n_{xi} \sim \text{binomial}(20, p_{xi})$, $x = 0, 1$. In (a), $\text{logit}(p0i) = 0.4W_i$, $\text{logit}(p1i) = 2+0.1W_i$; in (b), $\text{logit}(p0i) = -1+1.5W_i$, $\text{logit}(p1i) = 0.5W_i$.

Table 2

Performance of different estimators of β_0 , β_1 for a non-manipulable binary exposure. Here the number of observations in a cluster conditional on $X = x$ follows a Poisson distribution.

$\beta_0 = 0$			$\beta_1 = 1.69$		
Bias	SE	ECP★	Bias	SE	ECP★
(a) 0% clusters with only one exposure level present					
GEE	0.645	0.173	1.7	-0.399	0.090
CWGEE	0.094	0.121	87.4	-0.432	0.109
GEEN	-0.728	0.234	7.8	-0.334	0.094
SBW			-0.237	0.089	23.0
CWSBW			-0.228	0.108	42.7
SBWN			-0.289	0.094	12.2
DWGEE1	-0.001	0.124	94.6	-0.008	0.104
DWGEE2	-0.001	0.124	94.6	-0.009	0.104
(b) 12% clusters with only one exposure level present					
GEE	0.810	0.233	3.1	-0.396	0.141
CWGEE	0.322	0.183	56.4	-0.421	0.165
GEEN	-0.668	0.314	18.3	-0.326	0.145
SBW			-0.169	0.141	76.2
CWSBW			-0.140	0.162	85.5
SBWN			-0.281	0.148	49.9
DWGEE1	0.110	0.200	91.0	-0.119	0.196
DWGEE2	0.005	0.202	95.0	-0.013	0.191

★ coverage of 95% Wald confidence interval based on sandwich variance estimate

Here $n_{xi} \sim \text{poisson}(40x_i)$, $x = 0, 1$. In (a), $\log(\lambda_{0i}) = 1.2 + 0.8W_i$, $\log(\lambda_{1i}) = 0.3W_i$; in (b), $\log(\lambda_{0i}) = -1.5 + W_i$, $\log(\lambda_{1i}) = 0.5W_i$.

Table 3

Performance of different estimators of β_0 , β_1 for a manipulable binary treatment when treatment effect is invariant between clusters. Here the number of observations in a cluster follows a binomial distribution; and conditional on the cluster size, the number of observations treated in each cluster follows a binomial distribution.

$\beta_0 = 0$						$\beta_1 = -1.49$					
	Bias	SE	ECP*	Bias	SE	ECP*	Bias	SE	ECP*	Bias	SE
(a) 0% clusters with only one exposure level present											
GEE	0.030	0.174	94.0	0.285	0.182	63.5					
CWGEE	-0.139	0.169	86.6	0.271	0.179	65.5					
GEEN	-2.706	0.576	0.3	0.119	0.179	88.6					
SBW				-0.097	0.145	89.3					
CWSBW				-0.109	0.148	88.8					
SBWN				-0.308	0.152	45.7					
IPTW	0.168	0.172	83.8	0.008	0.144	94.8					
DWGEE1	-0.000	0.166	94.9	-0.010	0.140	94.0					
DWGEE3	-0.000	0.167	94.7	-0.009	0.144	95.0					
(b) 12% clusters with only one exposure level present											
GEE	-0.407	0.204	46.7	0.851	0.262	8.0					
CWGEE	-0.587	0.198	15.4	0.847	0.251	6.5					
GEEN	-3.272	0.514	0.0	0.743	0.242	10.3					
SBW				-0.190	0.187	82.6					
CWSBW				-0.189	0.190	83.5					
SBWN				-0.428	0.202	42.2					
IPTW	0.156	0.244	91.0	0.020	0.220	95.3					
DWGEE1	-0.114	0.212	91.2	0.119	0.206	91.2					
DWGEE3	-0.014	0.249	93.9	0.005	0.230	95.4					

Here $n_i \sim \text{binomial}(40, q_i)$, and $n_1 | n_i \sim \text{binomial}(n_i, p_i)$. In (a), $\text{logit}(q_i) = 1.8V_i$, $\text{logit}(p_i) = 1 + 1.8W_i$.

Table 4

Performance of different estimators of β_0 , β_1 for a manipulable binary treatment when treatment effect varies between clusters. Here the number of observations in a cluster follows a binomial distribution; and conditional on the cluster size, the number of observations treated in each cluster follows a binomial distribution.

$\beta_0 = 0$						$\beta_1 = 1.49$					
	Bias	SE	ECP*	Bias	SE	ECP*		Bias	SE	ECP*	
(a) 0% clusters with only one exposure level present											
GEE	-0.157	0.144	80.1	0.471	0.204	34.2					
CWGEE	-0.157	0.141	79.7	0.289	0.202	68.0					
GEEN	-1.303	0.555	30.7	0.436	0.201	38.5					
SBW				0.079	0.196	91.4					
CWSBW				-0.115	0.192	91.3					
SBWN				0.044	0.196	97.6					
IPTW	0.001	0.142	95.3	0.174	0.181	82.5					
DWGEE1	0.001	0.141	94.9	-0.011	0.179	94.1					
DWGEE3	0.001	0.139	95.3	-0.010	0.179	95.1					
(b) 8% clusters with only one exposure level present											
GEE	-0.672	0.159	1.3	1.117	0.252	0.7					
CWGEE	-0.674	0.157	1.1	0.934	0.247	3.1					
GEEN	-2.416	0.526	0.7	1.073	0.263	1.3					
SBW				0.029	0.246	93.5					
CWSBW				-0.170	0.237	89.0					
SBWN				-0.050	0.264	94.7					
IPTW	-0.012	0.228	93.1	0.188	0.250	86.9					
DWGEE1	-0.140	0.190	87.0	0.145	0.235	89.8					
DWGEE3	-0.013	0.231	92.9	0.004	0.252	94.9					

Here $n_i \sim \text{binomial}(40, q_i)$, and $n1 \sim \text{binomial}(n_i, p_i)$. In (a), $\text{logit}(q_i) = 1.8V_i$, $\text{logit}(p_i) = 1 + 1.8W_i$.

Table 5

Log odds ratio estimates and their standard errors for the caries data.

	GEE	GEEN	SBW	CWGEE	DWGEE1	DWGEE2
Intercept	-3.79(0.48)	-1.32(0.88)	-3.83(0.46)	-3.35(0.57)	-2.35(0.65)	-2.57(0.70)
Baseline incipient lesion (Y vs N) ^a	1.86(0.27)	1.12(0.27)	1.13(0.3)	1.67(0.3)	1.22(0.32)	1.29(0.35)
Female vs Male	0.27(0.26)	0.16(0.25)	0.17(0.26)	0.07(0.33)	-0.06(0.37)	-0.16(0.41)
Upper molar (Y vs N)	0.96(0.37)	0.9(0.38)	0.9(0.38)	1.4(0.46)	0.92(0.49)	1.47(0.52)
Lower molar (Y vs N)	0.94(0.27)	0.98(0.28)	0.98(0.28)	0.88(0.31)	0.43(0.49)	1.33(0.32)
Upper central incisor (Y vs N)	1.01(0.27)	1.12(0.27)	1.09(0.27)	1.01(0.26)	0.91(0.54)	0.73(0.48)
Lower central incisor (Y vs N)	2.6(0.29)	2.58(0.3)	2.6(0.3)	2.46(0.27)	1.68(0.39)	1.91(0.49)
Cluster size		-0.13(0.04)				
Baseline incipient lesion (between) ^b			3.96(1.07)			

^athe within-child effect of baseline incipient for SBW, the main effect of baseline incipient for other estimators

^bthe between-child effect of baseline incipient