Published in final edited form as:

Biometrics. 2011 June ; 67(2): 629–635. doi:10.1111/j.1541-0420.2010.01483.x.

Likelihood methods for binary responses of present components in a cluster

Xiaoyun Li^{1,*}, Dipankar Bandyopadhyay^{2,**}, Stuart Lipsitz^{3,***}, and Debajyoti Sinha^{1,****}

¹Department of Statistics, Florida State University, Tallahassee, FL, U.S.A.

²Division of Biostatistics and Epidemiology, Medical University of South Carolina, Charleston, SC, U.S.A.

³Harvard Medical School, Boston, MA, U.S.A.

SUMMARY

In some biomedical studies involving clustered binary responses (say, disease status) the cluster sizes can vary because some components of the cluster can be absent. When both the presence of a cluster component as well as the binary disease status of a present component are treated as responses of interest, we propose a novel two-stage random effects logistic regression framework. For the ease of interpretation of regression effects, both the marginal probability of presence/absence of a component as well as the conditional probability of disease status of a present component, preserve the approximate logistic regression forms. We present a maximum likelihood method of estimation implementable using standard statistical software. We compare our models and the physical interpretation of regression effects with competing methods from literature. We also present a simulation study to assess the robustness of our procedure to wrong specification of the random effects distribution and to compare finite sample performances of estimates with existing methods. The methodology is illustrated via analyzing a study of the periodontal health status in a diabetic Gullah population.

Keywords

Bridge density; Clustered data; Logistic link; Random effects

1. Introduction

Clustered binary response data are encountered in a multitude of biomedical settings including studies of developmental toxicology, family-based genetic traits, as well as longitudinal and spatial studies. Data analysis methods need to account for the effect of within cluster association to obtain correct estimates of the regression parameters and their corresponding standard errors (Aerts et al., 2002). In addition to the correlation within clusters, it is quite often that the size of cluster may be random and would also influence the outcomes of interest. This phenomenon has been earlier considered in the literature as an ‘informative cluster size’ scenario (Williamson et al., 2003; Williamson et al., 2008). For example, in dental studies, people often have varying number of teeth and fewer teeth may be associated with poor dental health of existing teeth.

*xli@stat.fsu.edu

**bandyopd@muscd.edu

***slipsitz@partners.org

****sinhad@stat.fsu.edu

We consider the study on Gullah speaking African-American Type-2 diabetics (Fernandes et al., 2009) with data collected on the periodontal health status of each tooth (component) within a patient (cluster). Here we only consider the molar and premolar teeth at 16 locations of the jaws. For each tooth (corresponding to a location in jaw), the three states of the outcome variable of primary interest were absent tooth, tooth present with disease and tooth present with no-disease. The primary objective of the study was to evaluate the influence of various patient-level as well as tooth-level explanatory variables (covariates) on these three categories of periodontal health status. This study design provides some interesting challenges for analysis. Most of the methods traditionally used for analyzing clustered binary data, including the generalized estimating equations (GEE; Liang and Zeger, 1986) and the mixed-effects conditional logistic regression methods (CLR; Breslow and Day, 1980, Lin et.al., 2009), do not treat the cluster size as random. However, the latent factor that causes the loss of a tooth might also lead to the poor dental health of the rest of the teeth. Thus, if we omit the first level binary response (i.e., the presence/ absence of the tooth), the interpretation and evaluation of the covariate effects on the disease status become questionable.

An appropriate GEE based method (Williamson et al., 2003) should accommodate random (informative) cluster-sizes when we are interested in estimating the effects of covariates on a randomly selected tooth from a population of ‘exchangeable’ teeth. These extensions of GEE methods assume that all teeth are ‘at risk’ of disease, except, we may have a biased sample because the disease status for some teeth are ‘missing from observation’. For estimating effects of covariates on the risk of disease for a typical tooth of a randomly selected patient, Hoffman et al. (2001) introduced the ‘within-cluster resampling’ (WCR) technique and Williamson et al.(2003) proposed the simpler and asymptotically equivalent ‘clustered weighted GEE’ (CWGEE) method. The CWGEE and WCR aim to estimate the effect on a typical cluster member from a random cluster via properly weighing the observations based on their cluster sizes. The model assumes ‘independence’ working correlation, which is unlikely to hold in practice, and consequently may lead to inefficient estimates (Panageas et al., 2007). At the same time, the rate of periodontal decay is expected to vary with tooth location and modeling the cluster size as a whole using the above methods may not be appropriate. Under a Bayesian paradigm, Dunson et al. (2003) used a continuation ratio probit model for the informative cluster size and underlying normal linear mixed models for the categorical and continuous sub-unit responses. However, this method does not address the expression and estimation of the interpretable marginal effects of covariates on the risk of a tooth being present and on the risk of disease. Here, we focus on estimating the effects of both cluster- and tooth-specific covariates on whether a particular tooth (at a location) from a randomly selected patient is present, and the effects of these covariates on the risk of disease when that particular tooth is found to be present. In Section 2, we propose a two-stage model with useful practical and physical interpretations of the marginal as well as conditional (given cluster-specific random effects) regression functions to express the covariate effects on the probability of a present tooth and on the conditional probability of disease given the tooth being present. We use logit link with the convenient log-odds interpretation (Agresti, 2006) for the conditional regression function for the binary response in each of two stages. In the usual binary mixed model with logit link and Gaussian random effects, the logistic structure is no longer preserved marginally after integrating out the random effects. We use a bivariate extension of the bridge density of Wang and Louis (2003) instead of the traditional Gaussian density for the random effects in both stages to facilitate the marginal probability of each binary response to approximately preserve the logistic structure. In Section 3, we apply a maximum likelihood (ML) estimation tool, readily amenable to available statistical softwares (viz. SAS) using routine nonadaptive importance sampling, to analyze the motivating data example. Section 4 provides simulation studies to compare our method’s performances with competing methods, and to evaluate the

robustness of our estimates when our modeling assumptions are not valid. Finally, necessary discussions and concluding comments are in Section 5.

2. Model for presence and disease status of a present tooth

Let Y_{1ij} be the binary response of the presence of the tooth at location $j = 1, \dots, K$ (sub-unit) for the patient $i = 1, \dots, n$ (cluster), given by

$$Y_{1ij} = \begin{cases} 1 & \text{when tooth at location } j \text{ of patient } i \text{ is present,} \\ 0 & \text{when tooth at location } j \text{ of patient } i \text{ is absent,} \end{cases}$$

where $Y_{1i} = \sum_{j=1}^K Y_{1ij}$ is the total number of teeth present in patient i (cluster-size) and can take values from 0 to K , where K is the number of locations. Also, let Y_{2ij} be the binary response as to whether the tooth at location j for patient i is healthy or diseased given $Y_{1ij} = 1$, i.e., given the tooth is present. In particular, Y_{2ij} is given by

$$Y_{2ij} = \begin{cases} 1 & \text{the tooth at location } j \text{ for patient } i \text{ is healthy (not diseased) given } Y_{1ij}=1, \\ 0 & \text{the tooth at location } j \text{ for patient } i \text{ is diseased given } Y_{1ij}=1. \end{cases}$$

Additionally, each tooth has a $d \times 1$ covariate vector $\mathbf{X}_{ij}$, which can be both patient-specific (like age, gender, body mass index, etc), or location-specific (like tooth position indicator with respect to upper/lower jaw). We assume that the distribution of the Stage 1 clustered binary response Y_{1ij} (the indicator of tooth presence), given the patient-specific random effects $\mathbf{B}_i = (B_{1i}, B_{2i})$ to be Bernoulli with success probability p_{1ij} , given by

$$\text{logit}[E(Y_{1ij}|\mathbf{B}_i, \mathbf{X}_{ij})] = \text{logit}(p_{1ij}) = \beta_0 + \beta_1^T \mathbf{X}_{ij} + B_{1i}. \quad (1)$$

We assume that the first-stage random cluster-effects B_{1i} follows a bridge distribution (Wang and Louis, 2003) with unknown parameter $0 < \phi_1 < 1$ and density

$$f(B_{1i}|\phi_1) = \frac{1}{2\pi} \frac{\sin(\phi_1\pi)}{\cosh(\phi_1 B_{1i}) + \cos(\phi_1\pi)}, \quad (2)$$

$$\text{where } \cosh(x) = \frac{1}{2}(e^x + e^{-x}).$$

We assume the parameter ϕ_1 to be common for all clusters to assure the exchangeability of cluster (random) effects on the binary response of the presence of each tooth. Using the Bridge density of (2) instead of, say, the normal density for cluster-effects B_{1i} , allows the marginal probability of the tooth being present at each location j to have logit link with regression parameter being proportional to the conditional regression parameter β_1 , i.e.,

$$\text{logit}[E(Y_{1ij}|\mathbf{X}_{ij})] = \text{logit}(p_{1ij}^*) = \phi_1\beta_0 + \phi_1\beta_1^T \mathbf{X}_{ij}, \quad (3)$$

where $0 < \phi_1 < 1$ measures the attenuation of the marginal regression effect $\phi_1\beta_1$ due to heterogeneity of clusters (patients).

When $Y_{1ij} = 1$ (tooth is present), we also consider a random effects model for the indicator of the present tooth being healthy. We assume that the conditional distribution Y_{2ij} given $Y_{1ij} = 1$ and patient specific $\mathbf{B}_i = (B_{1i}, B_{2i})$ to be Bernoulli with success probability p_{2ij} , given by:

$$\text{logit}[E(Y_{2ij}|Y_{1ij}=1, \mathbf{X}_{ij}, \mathbf{B}_i)] = \text{logit}(p_{2ij}) = \alpha_0 + \alpha_1^T \mathbf{X}_{ij} + B_{2i}, \quad (4)$$

where the second-stage random effect B_{2i} also follows a bridge density $f(B_{2i}|\phi_2)$ of (2) with unknown parameter $0 < \phi_2 < 1$. Clearly, when $Y_{1ij} = 0$, the random variable Y_{2ij} is not available with probability 1.

Adopting a bivariate density for (B_{1i}, B_{2i}) with bridge density for each marginal is the key to preserve the marginal logistic regression structures for Y_{1ij} in (3) as well as for Y_{2ij} . This bivariate bridge distribution is modeled using an inverse probability integral transformation

$$B_{ki} = F_{bk}^{-1}(\Phi(Z_{ki})), k=1, 2, \quad (5)$$

where (Z_{1i}, Z_{2i}) has a bivariate standard normal distribution with mean 0, variance 1 and correlation ρ , $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal density and $F_{bk}^{-1}(\cdot)$, $k=1, 2$ is the inverse cumulative distribution

$$F_{bk}^{-1}(u) = \frac{1}{\phi_k} \log \left\{ \frac{\sin(\phi_k \pi u)}{\sin[\phi_k \pi (1-u)]} \right\}$$

of the marginal bridge distribution for $0 < u < 1$. The transformation of (5) assures that the marginal densities of B_{1i} and B_{2i} are Bridge with c.d.f given by

$$F_{bk}(B_{ki}) = 1 - \frac{1}{\pi \phi_k} \left\{ \frac{\pi}{2} - \arctan \left[\frac{\exp(\phi_k B_{ki}) + \cos(\phi_k \pi)}{\sin(\phi_k \pi)} \right] \right\}.$$

The correlation parameter ρ induces the within patient association between B_{1i} and B_{2i} . When $\rho = 0$ (independence of B_{1i} and B_{2i}), the corresponding marginal model of Y_{2ij} given $Y_{1ij} = 1$ (integrating over the subject-specific random effect B_{2i}) is given by

$$\text{logit}[E(Y_{2ij}|\mathbf{X}_{ij}, Y_{1ij}=1)] = \text{logit}(p_{2ij}^*) = \phi_2 \alpha_0 + \phi_2 \alpha_1^T \mathbf{X}_{ij}, \quad (6)$$

where $\alpha_0^* = \phi_2 \alpha_0$ and $\alpha_1^* = \phi_2 \alpha_1$.

When B_{1i} and B_{2i} are not independent, (i.e., $\rho \neq 0$), $E(Y_{2ij}|Y_{1ij}=1, \mathbf{X}_{ij})$ integrating over $\mathbf{B}_i$ does not have closed form. We show the detailed derivation of $E(Y_{2ij}|Y_{1ij}=1, \mathbf{X}_{ij})$ in the Appendix that why this marginal regression function has the logistic form of (6) only when B_{1i} and B_{2i} are dependent. However, we can show numerically that $E(Y_{2ij}|Y_{1ij}=1, \mathbf{X}_{ij})$ still has an approximately logistic form given by $\text{logit}[E(Y_{2ij}|Y_{1ij}=1, \mathbf{X}_{ij})] \simeq \alpha_0^* + \alpha_1^{*T} \mathbf{X}_{ij}$ for a wide range of values of ϕ_1 and ϕ_2 , and moderate within cluster association ρ . Moreover, when the cluster-heterogeneities for B_{1i} and B_{2i} are moderate (i.e., $\phi_1 \geq 0.6$, $\phi_2 \geq 0.6$), this linear relationship is approximately the same as that in (6) with $\alpha_1^* \simeq \phi_2 \alpha_1$. We use Monte Carlo simulations to evaluate the linear form of $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ for $\alpha_0 = \beta_0 = 1$, $\alpha_1 = \beta_1 = -1$ and different values of ϕ_1 , ϕ_2 , ρ . To simplify the simulations, we assume $\phi_1 = \phi_2 = \phi$. We

choose different combinations of ρ and ϕ for different simulations, where $\rho \in \{0.2, 0.4, 0.6, 0.8\}$ and $\phi \in \{0.4, 0.6, 0.8\}$. We consider only one cluster-level covariate X ranging from -3 to 3 (corresponding to ± 3 standard deviation range for a standardized continuous covariate) with increments of 0.1 . We use 1000 simulated clusters for each X value and the total number of components within each cluster is 16. Gaussian copula transformation of (5) is used to simulate random effects (B_1, B_2) with marginal bridge density. The responses (Y_1, Y_2) for each value of $X = x \in [-3, 3]$ are generated using simulated (B_1, B_2) and simulations from binary distributions with $\text{logit}[E(Y_1|B_1, B_2; x)] = 1 - x + B_1$ and $\text{logit}[E(Y_2|Y_1 = 1, B_1, B_2; x)] = 1 - x + B_2$ with $\alpha_0 = \beta_0 = 1$ and $\alpha_1 = \beta_1 = -1$.

We would like to explore whether $\text{logit}[E(Y_2|Y_1 = 1; x)]$ is linear in x , that is, $\text{logit}[P(Y_2=1|Y_1=1; x)] - \text{logit}[P(Y_2=1|Y_1=1; x=0)] = \alpha_1^* x$; We would also like to compare the real attenuation α_1^*/α_1 with estimated ϕ_2 from equation (6). We plot the Monte Carlo approximation of $\text{logit}[E(Y_2|Y_1 = 1, X = x)] - \text{logit}[E(Y_2|Y_1, X = 0)]$ versus $\phi_2 \alpha_1 x$ for different values of $\phi = \phi_1 = \phi_2$ and ρ in Figure 1.

The solid straight line corresponds to the equation in (6) when B_1 and B_2 are independent (i.e., $\rho = 0$). From Figure 1, we can see that when the heterogeneity level ϕ is moderate ($\phi \geq 0.6$), the marginal model $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ is still approximately linear in x for different values of correlation ρ . Also, $\text{logit}[P(Y_2 = 1|Y_1 = 1; x)]$ will be well approximated by equation (6) in this case. When the cluster heterogeneity level is high, corresponding to $\phi \leq 0.4$ in Figure 1, we can still approximate the marginal relationship

$\text{logit}[P(Y_2=1|Y_1=1, x)] \simeq \alpha_0^* + \alpha_1^{*T} x$, preserving the logistic form with desirable physical interpretation, except that in this case $\alpha_1^* \neq \phi_2 \alpha_1$ and the actual attenuation α_1^*/α_1 is more than ϕ_2 , the attenuation expected from (6).

The contribution to the likelihood for the patient i (cluster) is given by

$$\int_{\mathbf{B}_i} \left[\prod_{j=1}^K \left\{ \overbrace{P(Y_{1ij}=y_{1ij}|\mathbf{B}_i, \mathbf{X}_{ij})}^a \left[\overbrace{P(Y_{2ij}=y_{2ij}|Y_{1ij}=1, \mathbf{B}_i, \mathbf{X}_{ij})}^b \right]^{y_{1ij}} \right\} \right] f_b(\mathbf{B}_i) d\mathbf{B}_i, \quad (7)$$

where $f_b(\mathbf{B}_i)$ is the joint density of (B_{1i}, B_{2i}) given as

$$f_Z\{\Phi^{-1}[F_{b1}(b_{1i})], \Phi^{-1}[F_{b2}(b_{2i})]\} \frac{f_{b1}(b_{1i}) \times f_{b2}(b_{2i})}{f_N\{\Phi^{-1}[F_{b1}(b_{1i})]\} \times f_N\{\Phi^{-1}[F_{b2}(b_{2i})]\}} \quad (8)$$

with $f_Z(\cdot, \cdot)$ is the bivariate normal density with mean 0, variance 1 and correlation ρ , $f_{bj}(\cdot)$ is the density of B_j for $j = 1, 2$ and $f_N(\cdot)$ is the standard normal density function. Note that (b) contributes to the likelihood in (7) only when $Y_{1ij} = 1$, i.e. the tooth j of patient i is present. The integral in (7) does not have a closed form, so ML estimation can be implemented using numerical approximation through routine nonadaptive importance sampling techniques available in PROC NLMIXED of SAS (V9.1).

3. Application: Periodontal Data

In this section, we apply our methods to analyze data on the periodontal health study of Gullah-speaking African-American Type-2 diabetic patients (Fernandes et al., 2009), where the health status for each tooth (excluding the 3rd molars) was assessed by hygienists using a periodontal probe (Darby and Walsh, 1995) for six pre-specified sites. We considered only the molars and pre-molars in this analysis which gives us a maximum count of 16 teeth (8 in

each jaw) within each patient. The current data analysis uses 113 patients with full covariate information and with at least one molar/pre-molar teeth present.

In our model, two binary responses are (1) presence of a tooth and (2) binary health status when the tooth is present. The health-status of a present tooth is based on whether the mean clinical attachment loss (mean over the 6 sites) of the tooth is ≥ 3 mm. This is an indicator of moderate to severe periodontitis (Armitage, 1999). Our patient-specific covariates for both stages 1 and 2 include age (in years), gender (1 = female, 0 = male), body mass index (BMI), binary glycemic level (1 = high, 0 = controlled), binary poverty indicator (1 = poor, 0 = otherwise), binary brush-floss indicator (1 = if the patient brushed twice and flossed once every day, 0 = otherwise), smoker (1 = present or past smoker, 0 = never a smoker) and the only tooth-specific covariate is the molar location (1 = molar/pre-molar tooth in upper jaw, 0 = otherwise). Our objective is to assess the effects of both patient- and tooth-specific covariates on the binary responses defined in equation (1) and (4). We fit the proposed two-stage model as described in Section 2 and estimate both conditional as well as marginal regression parameters with bivariate Bridge random effects. The estimated ϕ_1 and ϕ_2 also allows us to estimate the marginal regression effect $\phi_1\beta_1$ of (3) and the approximate marginal regression effect $\alpha_1^* = \phi_2\alpha_1$ of $\text{logit}[P(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X}_{ij})]$ in (6).

Table 1 provides the ML point estimate of regression parameters (conditional on random effects) in terms of log odds-ratio (OR), corresponding standard error and p-values for our model. We omit the covariates BMI, binary glycemic level, poverty and brush-floss indicators because they have been found to be statistically very insignificant for both responses. For the Stage 1 regression of equation (1), gender is the only covariate found to have a significant effect ($p = 0.037$), indicating that females are likely to have more molars/pre-molars (log OR = 0.451, 95% confidence interval (C.I.) = [0.027, 0.875]) present as compared to the males. For the Stage 2 regression of equation (4), the effect of gender is found to be statistically significant, indicating that under the same latent risk group (cluster), after adjusting for other covariates, the odds of a healthy present molar/pre-molar for females is about 2.10 (= $\exp(0.743)$) times higher than the odds for males. The binary indicator ‘molar location’ is also significant revealing that a molar/pre-molar present in the lower jaw is more likely (log OR = -1.187, 95% C.I. = [-1.501, -0.874]) to be healthy in comparison to that present in the upper jaw. The estimates (standard error) of the attenuation parameters ϕ_1 and ϕ_2 are 0.857(0.02) and 0.852(0.03) respectively. The estimated value of ϕ_1 confirms that heterogeneity induced by cluster effect causes moderate attenuation of the marginal regression parameter $\phi_1\beta_1$ of (3). The estimated value of ϕ_2 also imply that we can use the approximation of $\text{logit}[P(Y_{2ij} = 1|Y_{1ij} = 1, \mathbf{X}_{ij})]$ with $\alpha_1^* = \phi_2\alpha_1$ in (6) to estimate the effects of covariates on the risk of disease for a present tooth.

Table 2 compares our ML estimates of fixed-effects parameters in terms of log odds-ratio (OR), corresponding standard errors and p-values for the approximate marginal regression parameter $\alpha_1^* = \phi_2\alpha_1$ of (6) with estimates of comparable quantities from the CWGEE method of Williamson et al. (2003). Since the WCR method is asymptotically equivalent to the CWGEE, we didn't include it in Table 2. The marginal effect (after integrating out cluster effects) of gender and molar location on ‘risk of a tooth being healthy’ is statistically significant in both models. We like to emphasize that for CWGEE, the ‘risk of tooth being healthy’ means the probability of being healthy for a ‘typical tooth’ of a random person. On the contrary, our model's regression functions pertains to the conditional probability of tooth being healthy given that tooth is present, as expressed in (6). We can see from the Table 2 that our proposed model have slightly larger estimated log OR for the significant covariates than the CWGEE method, however, these are estimates of log OR of two different events.

To evaluate model diagnostics, we calculate the Pearson residuals of both response variables in our proposed model. The Pearson residuals r_{1ij} for the first response variable (the

presence/absence of the tooth), are calculated using the formula: $r_{1ij} = \frac{Y_{1ij} - \hat{p}_{1ij}}{\sqrt{\hat{p}_{1ij}(1 - \hat{p}_{1ij})}}$, and plotted against $\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{X}_{ij}$ in Figure 2. Similarly, we calculate the Pearson residuals

$r_{2ij} = \frac{Y_{2ij} - \hat{p}_{2ij}}{\sqrt{\hat{p}_{2ij}(1 - \hat{p}_{2ij})}}$ for the tooth health status when the tooth is present. Figure 3 gives the Pearson residual plot of r_{2ij} versus $\hat{\alpha}_0 + \hat{\alpha}_1^T \mathbf{X}_{ij}$. In both figures, LOWESS smoothing curves for residuals are indistinguishable to the line of zero residuals (parallel to the x-axis), indicating no trend and residual bias over different values of fitted regression effects for both the responses. This suggests that no further (non-linear) transformations of covariates are necessary.

4. Simulation Studies

To assess the performance of our proposed ML estimators based on the two-stage model, we conduct two simulation studies to (a) compare the performance with the CWGEE and the ML method for single-stage generalized linear mixed model, and (b) investigate the robustness of our ML estimates when the true distribution of the random effects is different from bridge density.

Each simulated data in our studies has $N = 50$ clusters with one binary cluster-specific covariate X_i indicating cluster exposure status ($X_i = 1$ for exposed, $X_i = 0$ for unexposed). We choose equal number of clusters in both the exposed and unexposed clusters. The maximum possible cluster size (total number of components) for each cluster is 9. We generated two binary responses: (i) Y_{1ij} being the presence of the component j in cluster i and (ii) Y_{2ij} being the health status of component j only when $Y_{1ij} = 1$, using

$$\text{logit}[\Pr(Y_{1ij}=1|\mathbf{X}_{ij}, B_{1i}, B_{2i})] = 1 - 2x_i + B_{1i} \quad \text{and} \quad (9)$$

$$\text{logit}[\Pr(Y_{2ij}=1|Y_{1ij}=1, \mathbf{X}_{ij}, B_{1i}, B_{2i})] = 1 - 2x_i + B_{2i}, \quad (10)$$

with $\alpha_0 = \beta_0 = 1$ and $\alpha_1 = \beta_1 = -2$. Using equation (5), the random effects $\mathbf{B}_i = (B_{1i}, B_{2i})$ in (9) and (10) are simulated via standard bivariate normal variables (Z_{1i}, Z_{2i}) with correlation ρ . We set $\phi_1 = \phi_2 = 0.8$ for the Bridge random effects (B_{1i}, B_{2i}) , to induce moderate variability within clusters in the simulated data.

Simulation 1

For each of four different values of $\rho = 0, 0.3, 0.5$ and 0.8 , we simulate 100 replicates of the dataset. We fit our proposed two-stage model as in Section 2 using ML method with $\text{logit}[\Pr(Y_{1ij}=1|\mathbf{X}_{ij}, \mathbf{B}_i)] = \beta_0 + \beta_1 \mathbf{X}_{ij} + B_{1i}$ and $\text{logit}[\Pr(Y_{2ij}=1|Y_{1ij}=1, \mathbf{X}_{ij}, \mathbf{B}_i)] = \alpha_0 + \alpha_1 \mathbf{X}_{ij} + B_{2i}$, with $\mathbf{B}_i$ as distributed in (8). For comparison, we also obtain (i) the CWGEE estimate for the marginal logistic regression model $\text{logit}[\Pr(Y_{2ij}=1|\mathbf{X}_{ij})] = \alpha_0 + \alpha_1 x_{ij}$, and (ii) the one-stage ML estimate using $\text{logit}[\Pr(Y_{2ij}=1|Y_{1ij}=1, \mathbf{X}_{ij}, B_{2i})] = \alpha_0 + \alpha_1 \mathbf{X}_{ij} + B_{2i}$ based only on the responses Y_{2ij} for cases when $Y_{1ij}=1$. Note that $\mathbf{X}_{ij} = X_i$, the binary exposure status.

Table 3 provides a comparison of our proposed two-stage ML method with (i) and (ii) as described above, for the regression function of Y_2 . The one-stage ML method treats the

cluster-size $(\sum_{j=1}^K Y_{1ij})$ as non-random (given) and uses a single bridge random cluster effect B_{2i} (instead of bivariate $\mathbf{B}_i$) with equation (4) for the likelihood contribution. The comparisons are based on the estimate with respect to the average of regression parameter estimates, bias and coverage probability of the 95% confidence interval (CI) over the replicates. When the two random effects are independent, (i.e., $\rho = 0$), the one-stage ML method performs as well as two-stage ML method. However, when ρ is even moderately different from 0, both one-stage ML and CWGEE give biased estimates and the bias increase as $|\rho|$ increases. Our two-stage ML method produces robust estimates for all ρ as far as the relative bias and 95% coverage probabilities are concerned.

Simulation 2

Here, we investigate the robustness of our ML estimators when the true distribution of the random effects (B_{1i}, B_{2i}) has multiple modes unlike the unimodal shape of the normal density and bridge density of (2). The simulated data were generated using the same scheme as that in Simulation 1, except that (B_{1i}, B_{2i}) are generated from a mixture of Gaussian

densities $(\frac{1}{2}N(-1, 1) + \frac{1}{2}N(1, 1))$ with two modes, however with finite mean zero. Simulations are conducted for various choices of clusters ($N = 50, 100$ and 500). Table 4 shows that when total number of clusters is $N = 50$, the ML regression estimates based on the wrongly specified random effects density of (2) are still quite close to the true value with a small relative bias (-0.003 for β_1 and 0.065 for α_1). The coverage probabilities of 95% interval estimates of ML method are 0.88 and 0.91 respectively for α_1 and β_1 . The coverage probabilities are a little low, which suggests that the estimated standard errors from the inverse of the information matrix are too small. In this case, limited simulations we have performed show that a robust sandwich variance estimate (White, 1982) tends to give less biased estimates of the standard errors, and thus coverage probabilities closer to the nominal 95% coverage probability. As the total number of clusters N increases, the relative biases remain relatively small (less than 10%). Thus, we can see that even when the underlying (true) distribution of random effects is very different from unimodal bridge density, our ML method performs with satisfactory robustness as far as the estimation of regression parameters are concerned.

5. Conclusions

Motivated by a periodontal study where a patient's tooth at a particular location of the jaw can be either absent or diseased or healthy, we develop a two-stage random effects regression model for clustered binary responses with random cluster sizes. In our model, the Stage 1 logistic regression function expresses the effects of covariates and random cluster effects on the presence of a component (tooth) within a cluster (patient). The Stage 2 logistic regression function explains the effects of covariates and cluster effects on the binary disease status of a tooth given that the tooth is present. The effects of clustering is efficiently handled through cluster-specific random effects at both stages and allowing a within cluster association among these two random effects. We also show numerically that when the correlation between two random effects is not high ($\rho \leq 0.8$) and the cluster-heterogeneity level is at least moderate, the marginal regression parameter α_1^* for the second-stage response is approximately proportional to the conditional regression parameter α_1 , as in (6).

The existing methods like GEE, CWGEE or Bayesian methods either estimate the marginal or the conditional (given random effects) regression parameters of the response at Stage 2. The attenuation of the marginal regression effect (after integrating random patient-effect) is not captured in these methods. Instead, our method provides conditional (on the random intercepts) as well as marginal inference for model parameters with approximate log-odds interpretation of the covariate effects at both stages. Appropriately, our method is useful when the cluster size is also considered as an important response by itself, as in the dental study. Unlike the CWGEE method, our method evaluates the probability for a present tooth and the conditional probability that the tooth is healthy given being present. For some studies, particularly when the maximum number of components are fixed and the components are not exactly exchangeable (e.g., teeth from different locations within the mouth, patients from different geographical areas, etc), it may not make sense to talk about the risk of a ‘typical’ cluster member. Our approach which evaluates the risk on a cluster component when it is present is more appropriate. Because the event of interest of our approach is very different from the event of interest of a CWGEE or some other missing data approach, there is no direct relationship between the regression parameters of these approaches. The comparison of regression estimates of the simulation study of Section 4 to some extent demonstrates how much these regression parameters from different models can differ from each other, rather than relative merit of one estimator over the other.

Acknowledgments

The authors are grateful for the support provided by grants R01CA69222, R01A160373, P20RR017696-06, U10DA013727-09 and 1R01LM009153 from the National Institutes of Health. The authors thank Dr. Elizabeth Slate for helpful insights, and the Center for Oral Health Research (COHR) at the Medical University of South Carolina for providing the motivating data and the context for this work.

References

- Aerts, M.; Geys, H.; Molenberghs, G.; Ryan, LM. Topics in Modelling of Clustered Data. Chapman and Hall; 2002.
- Agresti, A. Categorical Data Analysis. 2nd Edition. Wiley; 2006.
- Armitage GC. Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology*. 1999; 4:1–6. [PubMed: 10863370]
- Breslow, NE.; Day, NE. Statistical Methods in Cancer Research, Volume 1: The analysis of case-control studies. Lyon: International Agency for Research on Cancer; 1980.
- Darby, ML.; Walsh, MM. Dental Hygiene: Theory and Practice. 1st edition. USA: W. B. Saunders Company; 1995.
- Dunson DB, Chen Z, Harry J. A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics*. 2003; 59:521–530. [PubMed: 14601753]
- Fernandez JK, Wiegand RE, Salinas CF, Grossi SG, Sanders JJ, Lopes-Virella MF, Slate EH. Periodontal disease status in Gullah African Americans with Type 2 diabetics living in South Carolina. *Journal of Periodontology*. 2009; 80:1062–1068. [PubMed: 19563285]
- Hoffman EB, Sen PK, Weinberg CR. Within-cluster resampling. *Biometrika*. 2001; 88:1121–1134.
- Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Lin L, Bandyopadhyay D, Lipsitz SR, Sinha D. Association models for clustered data with binary and continuous responses. *Biometrics*. 2009; 66:287–293. [PubMed: 19432772]
- Panageas KS, Schrag D, Localio AR, Venkatraman ES, Begg CB. Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine*. 2007; 26:2017–2035. [PubMed: 17016864]
- Wang Z, Louis TA. Matching conditional and marginal shapes in binary mixed-effects models using a bridge distribution. *Biometrika*. 2003; 90:765–775.
- White H. Maximum likelihood estimation under mis-specified models. *Econometrica*. 1982; 50:1–26.

Williamson JM, Datta S, Satten GA. Marginal analysis of clustered data when the cluster size is informative. *Biometrics*. 2003; 59:36–42. [PubMed: 12762439]

Williamson JM, Kim H-Y, Manatunga A, Addiss DG. Modeling survival data with informative cluster size. *Statistics in Medicine*. 2008; 27:543–555. [PubMed: 17640035]

Appendix

Below, we show that $E[Y_{2ij} | Y_{1ij} = 1, \mathbf{X}_{ij}]$ has logistic form of (6) only when B_{1i} and B_{2i} are independent

$$\begin{aligned} E(Y_{2ij} | Y_{1ij} = 1, \mathbf{X}_{ij}) &= E[E(Y_{2ij} | Y_{1ij} = 1, \mathbf{X}_{ij}, B_{2i})] \\ &= \int_{B_{2i}} E(Y_{2ij} | Y_{1ij} = 1, \mathbf{X}_{ij}, B_{2i}) f(B_{2i} | Y_{1ij} = 1, \mathbf{X}_{ij}) dB_{2i}, \text{ (using (4))}, \end{aligned}$$

where

$$\begin{aligned} f(B_{2i} | Y_{1ij} = 1, \mathbf{X}_{ij}) &= \frac{\int_{B_{1i}} f(B_{1i}, B_{2i}) P(Y_{1ij} = 1 | B_{1i}, \mathbf{X}_{ij}) dB_{1i}}{P(Y_{1ij} = 1)} \\ &= f(B_{2i}) \frac{\int_{B_{1i}} f(B_{1i} | B_{2i}) P(Y_{1ij} = 1 | B_{1i}, \mathbf{X}_{ij}) dB_{1i}}{P(Y_{1ij} = 1 | \mathbf{X}_{ij})}. \end{aligned}$$

Hence $f(B_{2i} | Y_{1ij} = 1, \mathbf{X}_{ij}) = f(B_{2i})$ if and only if

$$\int_{B_{1i}} f(B_{1i} | B_{2i}) P(Y_{1ij} = 1 | B_{1i}, \mathbf{X}_{ij}) dB_{1i} = P(Y_{1ij} = 1 | \mathbf{X}_{ij}), \text{ for all } B_{2i} \in (-\infty, +\infty).$$

This equation above holds if and only if $f(B_{1i} | B_{2i}) = f(B_{1i})$.

Figure 1.

Plot of the Monte Carlo approximation of $P[Y_2 = 1|Y_1 = 1, X] - P[Y_2|Y_1 = 1, X = 0]$ (vertical axis) versus $\phi\alpha_1X$ (horizontal axis) for different values of ϕ and ρ

Figure 2. Pearson residual plot for Stage-1 regression (presence/absence status of a tooth) versus the estimated linear regression $\hat{\beta}_0 + \hat{\beta}_1^T \mathbf{X}_{ij}$.

Figure 3. Pearson residual plot for Stage-2 regression (health status of present tooth) versus the estimated linear regression $\widehat{a}_0 + \widehat{a}_1^T \mathbf{X}_{ij}$.

Table 1

Estimates of regression parameters (conditional on random effects) using our proposed ML method for the periodontal data. OR indicates 'odds-ratio' and SE denotes the corresponding standard errors of the estimate.

Parameter	log OR (SE)	P-value
Presence of tooth: Y_1 (Stage 1 response)		
Age	-0.010 (0.008)	0.258
Gender	0.451 (0.214)	0.037
Smoker	-0.059 (0.221)	0.790
Molar location	0.0569 (0.102)	0.577
Health status for a present tooth: Y_2 (Stage 2 response)		
Age	0.005 (0.010)	0.619
Gender	0.743 (0.269)	0.007
Smoker	0.154 (0.270)	0.569
Molar location	-1.187 (0.158)	<.0001
Other Parameters		
ϕ_1	0.857 (0.024)	<.0001
ϕ_2	0.852 (0.030)	<.0001
ρ	0.218 (0.175)	0.215

Table 2

Estimates of the marginal regression parameters using ML and CWGEE methods for the periodontal data. Estimates in the proposed ML model corresponds to Stage 2 regression. OR indicates 'odds-ratio' and SE denotes the corresponding standard errors of the estimate.

	Proposed ML model		CWGEE	
	log OR (SE)	P-value	log OR (SE)	P-value
Age	0.004 (0.009)	0.6194	0.003(0.009)	0.739
Gender	0.633 (0.223)	0.007	0.510(0.217)	0.019
Smoker	0.132 (0.230)	0.569	0.167(0.223)	0.454
Molar location	-1.012 (0.132)	<.0001	-0.936(0.152)	<.0001

Results of the simulation study to compare two-stage ML method vs. the CWGEE method: 100 replicated datasets with $N = 50$ clusters and maximum cluster-size $J = 9$ were simulated from the two-stage model of equation (1) and (4) with bridge density as in equation (8); $\phi_1 = \phi_2 = 0.8$, $\beta_1 = \alpha_1 = -2$ and $\rho = 0, 0.3, 0.5, 0.8$.

Table 3

	True value	Mean estimate	Relative Bias	Coverage probability
$\rho = 0$				
Proposed model	α_1	-1.992	0.004	92%
One-stage ML	α_1	-2.004	-0.002	93%
CWGEE	$\tilde{\alpha}_1$	-1.612	0.194	86%
$\rho = 0.3$				
Proposed model	α_1	-1.960	0.020	95%
One-stage ML	α_1	-1.866	0.068	93%
CWGEE	$\tilde{\alpha}_1$	-1.546	-0.227	80%
$\rho = 0.5$				
Proposed model	α_1	-1.985	0.008	97%
One-stage ML	α_1	-1.858	0.710	93%
CWGEE	$\tilde{\alpha}_1$	-1.528	0.236	80%
$\rho = 0.8$				
Proposed model	α_1	-1.960	0.02	95%
One-stage ML	α_1	-1.784	0.108	86%
CWGEE	$\tilde{\alpha}_1$	-1.541	0.230	80%

Table 4

Results of simulation study of misspecification of our proposed model using 100 replicated datasets with maximum cluster-size $J = 9$ random effects simulated from a mixture of Normal densities:

$$\frac{1}{2}N(-1, 1) + \frac{1}{2}N(1, 1).$$

	True value	Mean estimate	Relative Bias	Coverage probability
N=50				
β_1	-2	-1.994	0.003	88%
α_1	-2	-1.870	0.065	91%
N=100				
β_1	-2	-1.918	0.041	91%
α_1	-2	-1.836	0.082	86%
N=500				
β_1	-2	-1.897	0.0514	84%
α_1	-2	-1.865	0.068	80%