Biometrika (2011), **98**, 1, *pp.* 147–162
© 2011 Biometrika Trust
Printed in Great Britain

doi: 10.1093/biomet/asq066
Advance Access publication 31 January 2011

Estimation of covariate effects in generalized linear mixed models with informative cluster sizes

By JOHN M. NEUHAUS and CHARLES E. McCULLOCH

Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94143-0560, U.S.A.

john@biostat.ucsf.edu chuck@biostat.ucsf.edu

Summary

In standard regression analyses of clustered data, one typically assumes that the expected value of the response is independent of cluster size. However, this is often false. For example, in studies of surgical interventions, investigators have frequently found surgery volume and outcomes to be related to the skill level of the surgeons. This paper examines the effect of ignoring response-dependent, informative, cluster sizes on standard analytical methods such as mixed-effects models and conditional likelihood methods using analytic calculations, simulation studies and an example from a study of periodontal disease. We consider the case in which cluster sizes and responses share random effects which we assume to be independent of the covariates. Our focus is on maximum likelihood methods that ignore informative cluster sizes, and we show that they exhibit little bias in estimating covariate effects that are uncorrelated with the random effects associated with cluster sizes. However, estimation of covariate effects that are associated with the random effects can be biased. In particular, for models with random intercepts only, ignoring informative cluster sizes can yield biased estimators of the intercept but little bias in estimation of all covariate effects.

Some key words: Conditional likelihood; Generalized linear mixed model; Misspecified mixing distribution; Random slope.

1. Introduction

In the regression analysis of clustered data using methods such as generalized linear mixed models (McCulloch et al., 2008) and generalized estimating equations (Diggle et al., 2002), data analysts typically assume that the expected value of the response is independent of cluster size, although this is frequently not the case in practice. For example, Gansky et al. (1999) conducted a study of periodontal disease where investigators gathered indicators of disease and covariate information at multiple tooth sites within each subject's mouth. In this analysis, the prevalence of diseased sites within mouths was associated with the number of teeth in the mouth, perhaps reflecting an association of each variable with the overall health status of the mouth. We will analyse these data further below. As another example, several studies have demonstrated inverse associations between the annual number of coronary artery bypass graft surgeries a hospital performs and its mortality rate (Dudley et al., 2002). As a final example, longitudinal studies typically plan to measure all subjects the same number of times but frequently end up with differing numbers of responses per subject. If the probability of missingness depends on the responses, either observed or unobserved, then there will be an association between the expected value of the response and the number of repeated measures for a subject. We call cluster sizes that are related to components of the model for the response, for example, the expected value of the response or the random effects, informative cluster sizes.

In longitudinal studies where the probability of missingness depends on unobserved responses, ignoring the association of cluster size with the expected value of the response may lead to bias. Little & Rubin (2002) classify such settings as exhibiting informative missingness and show that approaches that fail to address it can yield biased estimates of the associations of covariates with responses.

Previous work (Hoffman et al., 2001; Williamson et al., 2003) has examined the effects of response-cluster size correlations on marginal analyses of clustered data such as those based on generalized estimating equations. These papers generally show that ignoring these correlations can yield highly biased estimates of means and intercept terms of regression models but also make statements that imply that there are important effects on regression coefficients as well, e.g., '. . . the usual generalized estimating equation approach resulted in severely biased estimates of both the marginal regression and association parameters' (Williamson et al., 2003, p. 39). However, a closer examination of their results shows that the effect on covariate coefficients is much weaker; ignoring response-cluster size correlations leads to little bias in covariate effects. For example, while the simulation results in Table 4 of Hoffman et al. (2001) indicate relative biases of up to 37% in estimates of the intercept using marginal model approaches that ignore response-dependent clusters sizes, the table also indicates essentially no bias in estimates of the slope coefficient: marginal model working independence and exchangeable approaches that ignored response-dependent clusters sizes yielded average slope parameter estimates of 0·4991 and 0·4796, respectively, compared with the true value 0·4796. The simulation results in Table 1 of Williamson et al. (2003) indicate similar results.

Like Hoffman et al. (2001) and Williamson et al. (2003), Dunson et al. (2003) claimed that ignoring informative cluster sizes can lead to biased estimates of regression coefficients but provide no direct evidence that this is so. Indeed, the careful reanalysis of the data of Dunson et al. (2003) by Gueorguieva (2005) showed that ignoring informative cluster sizes leads to essentially no bias in regression coefficient estimates.

This paper evaluates the performance of standard cluster-specific methods such as generalized linear mixed models and conditional maximum likelihood estimation for analysing clustered data in cases where the distribution of the response is associated with cluster sizes. Cluster-specific approaches provide more meaningful covariate effect estimates than marginal or population-averaged ones when scientific interest focuses on the associations of covariates that vary within clusters with the response (Neuhaus et al., 1991; Neuhaus, 2001; McCulloch & Neuhaus, 2005). We focus on the case where random effects are uncorrelated with the distribution of the covariates, since papers such as Neuhaus & McCulloch (2006) show that ignoring such correlations can produce biased estimates of regression coefficients.

We show that the problem of assessing the effects of ignoring response-dependent cluster sizes with cluster-specific methods is a form of misspecification of the shape of the random effects distribution, and results from that area apply here. In particular, we give theoretical arguments that explain why ignoring response-dependent, informative cluster sizes can yield highly biased estimates of intercept parameters but very little bias in estimates of regression/slope parameters. We also provide simulation results for cluster-specific methods that corroborate the previous results for marginal approaches. This work indicates little bias in covariate effect estimates, but we acknowledge that misspecifications involving the covariates such as ignoring correlations of covariates and random effects (Neuhaus & McCulloch, 2006) or ignoring the dependence of random effects variances on covariates (Heagerty & Kurland, 2001) can produce substantial bias in regression coefficient estimates. However, we also note that Neuhaus & McCulloch (2006) showed that conditional likelihood methods and analogous covariate partitioning methods provide consistent estimates of regression parameters in settings with correlations between covariates and random effects.

2. Cluster-specific models

2·1. *Model specification*

This paper assumes that we have clustered or longitudinal responses Y_{ij} with p-dimensional covariates X_{ij}, where i indexes clusters and j units within clusters, and that we want to fit a generalized linear mixed model to assess the association of within-cluster changes in X with a known function of $E(Y)$. That is, given a vector b_i of parameters specific to the ith cluster, for the jth unit, the conditional density of Y_{ij} is of the form

$$f_Y(y_{ij} \mid b_i, x_{ij}) = \exp[\{y_{ij}\theta_{ij} - c(\theta_{ij})\}\phi + d(y_{ij}, \phi)] \tag{1}$$

where c and d are functions of known form, ϕ is a scale parameter and θ_{ij} depends on covariates X_{ij}. In addition, one assumes that

$$\mu_{ij} = E(Y_{ij} \mid b_i, z_{ij}, x_{ij}) = g^{-1}(z_{ij}^{\mathrm{T}} b_i + x_{ij}^{\mathrm{T}}\beta), \tag{2}$$

where x_{ij}^{T} and z_{ij}^{T} are the specified covariate row vectors relating the fixed and random effects, respectively, to the observations, g is a link function and μ_{ij} is a function of θ_{ij}. Without loss of generality, we assume that $E(X) = 0$. Given b_i, we assume that the responses $Y_{i1}, \ldots, Y_{in_i}$ are independent.

We consider two popular approaches for estimating the parameters β of (2). The first assumes a distribution for the b_i, with parameter Σ_b, and maximizes the likelihood obtained by integrating over b_i. The second treats the b_i as fixed constants and eliminates them from the likelihood using conditioning methods.

2·2. *Maximum likelihood estimation*

Maximum likelihood approaches assume that the random effects b follow a distribution F_b. Integrating over b, the likelihood for generalized linear mixed models fited to m independent clusters, with the ith cluster containing n_i units, is

$$\mathrm{L}(\beta, F_b) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_Y(y_{ij} \mid b,\ x_{ij})\, dF_b(b) \tag{3}$$

where $Y_{ij} \mid b$ follows a generalized linear model as in (1). This approach estimates the model parameters by maximizing (3).

2·3. *Conditional likelihood methods*

With canonical link models whose only random effects are random intercepts, we can estimate the effects of within-cluster covariates of (2) using a conditional likelihood that eliminates the random effects from the model. Rather than integrating the random effects out of the model as in (3), the approach works with the response distribution conditional on sufficient statistics for the cluster-specific intercepts. For example, for canonical link models such as identity and logistic link models and random intercepts only, the sufficient statistics are $s_i = \sum_{j=1}^{n_i} y_{ij}$ and the conditional likelihood has terms

$$f_Y\left(y_i \mid x_i, n_i, b_i, \sum_{j=1}^{n_i} y_{ij}\right) \tag{4}$$

where $y_i = (y_{i1}, \ldots, y_{in_i})$ and $x_i = (x_{i1}, \ldots, x_{in_i})$. Conditional likelihood methods are difficult to develop for models featuring more complicated random effects structures such as random slopes. This paper considers only the case of fitting conditional likelihood methods that assume random intercepts.

2·4. *Associations of cluster size with the response*

When the cluster size is associated with the outcome, we can view the observed data for the ith cluster as $y_{i1}, \ldots, y_{in_i}, n_i, x_{i1}, \ldots, x_{in_i}$. The corresponding joint likelihood is the product, over i, of terms

$$f_{Y,N,X}(y_i, n_i, x_i) = \int_b \prod_{j=1}^{n_i} f_Y(y_{ij} \mid n_i, x_{ij}, b) f_N(n_i \mid x_{ij}, b) f_X(x_{ij} \mid b)\, dF_b(b) \tag{5}$$

$$= \int_b \prod_{j=1}^{n_i} f_Y(y_{ij} \mid n_i, x_{ij}, b) f_N(n_i \mid x_{ij}, b) f_X(x_{ij})\, dF_b(b) \tag{6}$$

$$= \int_b \prod_{j=1}^{n_i} f_Y(y_{ij} \mid x_{ij}, b) f_N(n_i \mid x_{ij}, b) f_X(x_{ij})\, dF_b(b). \tag{7}$$

Note that b may be a vector to accommodate, for example, random intercepts and slopes. The equality (6) follows because we are assuming that X_{ij} is independent of b_i. The equality (7) follows because we assume that once the cluster size is determined, which is a function of x, β and/or b, e.g. through the conditional mean, the data-generating mechanism for each observation does not depend on n_i.

2·5. *Cluster size associations are misspecified mixing distributions*

We assume that the parameters in the joint distribution of N_i and X_{ij} are not functionally related to the parameters of interest, namely those in the distribution of outcomes or random effects. By Bayes' theorem

$$f_b(b_i \mid n_i, \mathbf{x}_i) \propto f_N(n_i \mid \mathbf{x}_i, b_i) f_X(\mathbf{x}_i) f_b(b). \tag{8}$$

Then using (8) in (7), as a function of β, we have

$$f_{Y,N,X}(y_i, n_i, x_i) \propto \int_b \left\{ \prod_{j=1}^{n_i} f_Y(y_{ij} \mid n_i, x_{ij}, b) \right\} dF_{b|n,x}(b \mid n_i, x_{ij}), \tag{9}$$

$$= f_Y(y_i \mid n_i, x_i).$$

This is a useful representation of (7) since analysts would typically model the distribution of the responses conditional on covariates and sample size. In particular, ignoring the association of cluster size with response, one would base inference on the incorrect likelihood built up of terms

$$f_Y^*(y_{i1}, \ldots, y_{in_i} \mid x_{i1}, \ldots, x_{in_i}, n_i) = \int_b \prod_{j=1}^{n_i} f_Y(y_{ij} \mid x_{ij}, b)\, dF_b^*(b), \qquad (10)$$

where the asterisk denotes a fitted distribution.

Comparing equations (10) and (9), we see that (10) is the same as (9) but with a misspecified random effects distribution, namely the conditional distribution of b given n_i and x_i is incorrectly specified as $f_b^*(b)$. This is important because of the extensive literature on misspecification of the shape of the mixing distribution and its generally inconsequential effects. For example, Neuhaus et al. (1992) examined shape misspecification for mixed-effects logistic models and found that it led to bias in the intercept and random effects variance but little bias in the estimates of regression coefficients. Thus, these results indicate that we should find similar effects of ignoring associations of cluster sizes and responses.

3. Asymptotic values of estimators under a cluster size association

3·1. *Inference under misspecification*

Given models for the cluster sizes and covariates as functions of b, as well as a specification for $dF_b(b)$, one could maximize the likelihood built up of terms (7) to obtain maximum likelihood estimates for all model parameters of interest.

Following Akaike (1973) and White (1982), we evaluate the asymptotic expectation of the maximum likelihood estimator, $\hat{\xi}^*$, of ξ^* obtained from the model that ignores the association of cluster sizes and random effects. These authors show that $\hat{\xi}^*$ converges to the value ξ^* which minimizes the Kullback–Leibler divergence (Kullback, 1959) between the true and misspecified models. That is, ξ^* minimizes

$$E_X\left(E_{Y,N|X}\left[\log\{f(Y \mid \xi, X)/f^*(Y \mid \xi^*, X)\}\right]\right), \qquad (11)$$

where ξ is the true parameter, f and f^* denote the true and fitted response densities, respectively, and we calculate the expectation with respect to the true model.

3·2. *Minimizing Kullback–Leibler equations for generalized linear mixed models*

Consistent with §2·4, we assume that the association between the response and cluster sizes arises through the random effect b. Specifically, we assume that cluster sizes are random variables that depend on b and denote these random variables by $N_i(b_i)$.

We display the derivatives of (11) with respect to the intercept, β_0^*, and regression parameters, β_l^* $(l = 1, \ldots, p)$, and exploit the fact that the fitted conditional densities are from generalized linear models to yield the simultaneous equations

$$E_X\left(E_{Y,N|X}\left[\sum_{i=1}^{m} \frac{\int_u \left\{\sum_{j=1}^{N_i(b_i)} r_{ij}^*\right\}\left\{\prod_{j=1}^{N_i(b_i)} f_Y^*(Y_{ij} \mid X_{ij}, u)\right\} dF_b^*(u)}{\int_u \prod_{j=1}^{N_i(b_i)} f_Y^*(Y_{ij} \mid X_{ij}, u)\big\}\, dF_b^*(u)}\right]\right) = 0, \qquad (12)$$

$$E_X\left(E_{Y,N|X}\left[\sum_{i=1}^{m} \frac{\int_u \left\{ \sum_{j=1}^{N_i(b_i)} X_{ijl}\, r^*_{ij} \right\} \left\{ \prod_{j=1}^{N_i(b_i)} f^*_Y(Y_{ij} \mid X_{ij}, u) \right\} dF^*_b(u)}{\int_u \prod_{j=1}^{N_i(b_i)} f^*_Y(Y_{ij} \mid X_{ij}, u) \Big\} \, dF^*_b(u)} \right] \right) = 0 \qquad (13)$$

$$(l = 1, \ldots, p),$$

where $r^*_{ij} = \{(Y_{ij} - \mu^*_{ij}) \text{ of } \phi V(\mu^*_{ij}) g'(\mu^*_{ij})\}$, $V(\cdot)$ and $f^*(\cdot)$ are the variance function and conditional density (1), respectively, of the generalized linear model of interest and u is the variable of integration representing the fitted random effect. Note that $r^*_{ij} = (Y_{ij} - \mu^*_{ij})/\phi$ for canonical link generalized linear models. Solving the system (12)–(13), along with analogous equations from derivatives with respect to Σ^*_b, yields the limiting values $(\beta^*_0, \beta^*_1, \ldots, \beta^*_p)$ of the estimator $(\hat{\beta}^*_0, \hat{\beta}^*_1, \ldots, \hat{\beta}^*_p)$ based on a likelihood that ignores informative cluster sizes. Thus, minimizing the Kullback–Leibler divergence is equivalent to calculating the expected scores obtained from the misspecified likelihood with respect to the true model and determining $(\beta^*_0, \beta^*_1, , \ldots, \beta^*_p)$ so that the expected scores are exactly zero.

We examine several special cases below.

3·3. *Linear mixed models with random intercepts*

We first consider fitting the linear mixed effects model

$$Y_{ij} = \beta^*_0 + b^*_i + \beta^*_1 X_{ij} + e^*_{ij}, \qquad (14)$$

with $X_{ij} \sim (0, \sigma^{*2}_x)$, $b^*_i \sim N(0, \sigma^{*2}_b)$, $e^*_{ij} \sim N(0, \sigma^{*2}_e)$, $b^*_i \perp\!\!\!\perp e^*_{ij}$, and $X_{ij} \perp\!\!\!\perp e^*_{ij}$, $(i = 1, \ldots, m;$ $j = 1, \ldots, N_i(b_i))$. Under this canonical link model, the scale factor $\phi = \sigma^2_e$ and the link function g is the identity, so that $\mu_{ij} = \beta_0 + b_i + \beta_1 X_{ij}$. Under the true model

$$Y_{ij} = \beta_0 + b_i + \beta_1 X_{ij} + e_{ij},$$

so that $r^*_{ij} = \{(\beta_0 - \beta^*_0) + (\beta_1 - \beta^*_1) X_{ij} + (b_i - b^*_i) + e_{ij}\}/\sigma^{*2}_e$ and

$$f^*_Y(y_{ij} \mid x_{ij}, b_i) = \{(2\pi)^{1/2}\, \sigma^*_e\}^{-1}\, e^{-(y_{ij} - \mu^*_{ij})^2/(2\sigma^{*2}_e)} = \{(2\pi)^{1/2}\, \sigma^*_e\}^{-1}\, e^{-r^{*2}_{ij} \sigma^{*2}_e/2}.$$

When $\beta^*_1 = \beta_1$, both r^*_{ij} and $f^*_Y(y_{ij} \mid x_{ij}, b_i)$ are free of x_{ij} so that the left-hand side of (13) is a linear function of X_{ij}. Since $E(X) = 0$, it follows that the left-hand side of (13) is zero so that $\beta^*_1 = \beta_1$ solves (13) for all values of β^*_0, σ^{*2}_b and σ^{*2}_e. This result of consistent estimation extends to the case of multiple covariates. Thus, for the linear mixed effects model with random intercepts, ignoring informative cluster sizes, as in (10), yields consistent estimates of slope parameters.

Using the fact that $\beta^*_1 = \beta_1$, we next argue that $\hat{\beta}^*_0$, does not consistently estimate β_0, because the left-hand side of (12) is not a linear function of X_{ij}, unlike (13). Using (12) and focussing on the ith term, and dropping the subscript i, we see that the densities in the numerator and denominator form

$$\frac{f_Y(y \mid x, b) f^*_b(b)}{\int_b f_Y(y \mid x, b) f^*_b(b)} = \frac{f^*_Y(y, b \mid x)}{f^*_Y(y \mid x)} = f^*_b(b \mid y, x).$$

Therefore, the integral in the numerator of (12) is the expectation with respect to b^* conditional on y and x. Denote that expectation by E^*.

The left-hand side of (12), the equation associated with the intercept, therefore becomes

$$E_X\left[E_{N,Y|X}\left\{\sum_i\sum_{j=1}^{N(b_i)}E^*(r^*_{ij}\mid Y,X)\right\}\right]$$

$$=E_XE_{N,Y|X}\sum_i\sum_{j=1}^{N(b_i)}\left\{Y_{ij}-E^*(b^*_i\mid Y,X)-\beta^*_0-\beta^*_1X_{ij}\right\}/\sigma^{*2}_e. \qquad (15)$$

For model (14)

$$E^*(b^*_i\mid Y,X)=\frac{\sigma^{*2}_b}{\sigma^{*2}_b+\sigma^{*2}_e/N(b_i)}(\bar{Y}_{i\cdot}-\beta^*_0-\beta^*_1\bar{X}_{i\cdot})\equiv\lambda(b_i)(\bar{Y}_{i\cdot}-\beta^*_0-\beta^*_1\bar{X}_{i\cdot}).$$

Thus, (15) implies

$$\sigma^{*2}_e\frac{\partial\log\mathrm{L}}{\partial\beta_0}=E_X\left(E_{N,Y|X}\left[\sum_i N(b_i)\{1-\lambda(b_i)\}\left(\bar{Y}_{i\cdot}-\beta^*_0-\beta^*_1\bar{X}_{i\cdot}\right)\right]\right).$$

Under the situation in which $\beta^*_j=\beta_j$, this simplifies to

$$E_X\left(E_{N,Y|X}\left[\sum_i N(b_i)\{1-\lambda(b_i)\}\left(b_i+\bar{e}_{i\cdot}\right)\right]\right).$$

This is basically a covariance term between b_i and $N(b_i)$, which will not be zero under an informative cluster size model. That is, $\hat{\beta}^*_0$ does not consistently estimate β_0 when one ignores informative cluster sizes and this work indicates why intercept estimators behave differently than covariate effect estimators. Benhin et al. (2005) obtained equivalent results for intercept and slope estimators using a different approach.

3·4. *Linear mixed models with random intercepts and slopes*

We next consider fitting a linear mixed effects model with both random intercepts and slopes, with cluster sizes associated with both the random intercepts and slopes. As is common in practice, the model includes both a between-cluster covariate, X_B, and a within-cluster covariate, X_{W1}. We will see that consistency results depend on whether or not a model covariate X_l is correlated with a covariate Z_k from (2) where the corresponding random effect b_k is associated with cluster size. Therefore, we also include a second within-cluster covariate, X_{W2}, with no random slope term, in the model. We construct X_{W2} so that it is orthogonal to X_{W1} and X_B. Thus, we consider fitting the model

$$Y_{ij}=\beta^*_0+b^*_{0i}+(\beta^*_{W1}+b^*_{1i})X_{W1ij}+\beta^*_{W2}X_{W2ij}+\beta^*_BX_{Bi}+e^*_{ij}, \qquad (16)$$

with $X_{W1ij}\sim(0,\sigma^{*2}_{XW1})$, $X_{W2ij}\sim(0,\sigma^{*2}_{XW2})$, $X_{Bi}\sim(0,\sigma^{*2}_{XB})$,

$$\begin{pmatrix}b^*_{0i}\\ b^*_{1i}\end{pmatrix}\overset{\text{ind}}{\sim}N\left\{\begin{pmatrix}0\\0\end{pmatrix}\quad\begin{pmatrix}\sigma^{*2}_{b0}&\sigma^*_{12}\\ \sigma^*_{12}&\sigma^{*2}_{b1}\end{pmatrix}\right\},$$

$e^*_{ij} \sim N(0, \sigma_e^{*2})$, and the covariates $(X_{W1ij}, X_{W2ij}, X_{Bi})$ and random effects (b^*_{0i}, b^*_{1i}) are independent of e^*_{ij}. In addition, without loss of generality, we assume that $\sum_{j=1}^{N(b_i)} x_{W1ij} = \sum_{j=1}^{N(b_i)} x_{W2ij} = 0$, so that X_{W1ij} and X_{W2ij} are orthogonal to X_{Bi}. We can achieve this structure with standard decompositions of covariates into between- and within-cluster components (Neuhaus & McCulloch, 2006). As in §3·3, we have $\mu_{ij} = \beta_0 + b_{0i} + (\beta_{W1} + b_{1i})X_{W1ij} + \beta_{W2}X_{W2ij} + \beta_B X_{Bi}$. Under the true model

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_{W1} + b_{1i})X_{W1ij} + \beta_{W2}X_{W2ij} + \beta_B X_{Bi} + e_{ij},$$

so that

$$\begin{aligned} r^*_{ij} = \sigma_e^{*-2} \big\{ &(\beta_0 - \beta^*_0) + (b_{0i} - b^*_{0i}) + (\beta_{W1} - \beta^*_{W1} + b_{1i} - b^*_{1i})X_{W1ij} \\ &+ (\beta_{W2} - \beta^*_{W2})X_{W2ij} + (\beta_B - \beta^*_B)X_{Bi} + e_{ij} \big\} \end{aligned}$$

and $f^*_Y(y_{ij} \mid x_{ij}, b_i) = \left\{(2\pi)^{1/2}\, \sigma^*_e\right\}^{-1} e^{-(y_{ij}-\mu^*_{ij})^2/(2\sigma_e^{*2})} = \{(2\pi)^{1/2}\, \sigma^*_e\}^{-1} e^{-r_{ij}^{*2}\sigma_e^{*2}/2}$.

Again following §3·3, the integrals in the numerators of (12) and (13) are the expectations with respect to $b^* = (b^*_0, b^*_1)$ conditional on y and x. The left-hand sides of (12) and (13), the equations associated with the intercept and β_l, $l = B, W1, W2$, respectively, therefore become

$$\begin{aligned} &E_X E_{N,Y|X} \sum_i \sum_{j=1}^{N(b_i)} X_{lij} E^*(r^*_{ij} \mid Y, X) \\ &= E_X E_{N,Y|X} \sum_i \sum_{j=1}^{N(b_i)} X_{lij} \left[Y_{ij} - E^*(b^*_{0i} \mid Y, X) - \beta^*_0 - \left\{ \beta^*_{W1} + E^*(b^*_{1i} \mid Y, X) \right\} X_{W1ij} \right. \\ &\quad \left. - \beta^*_{W2} X_{W2ij} - \beta^*_B X_{Bij} \right] / \sigma_e^{*2}. \end{aligned} \tag{17}$$

The equation associated with β_0 is just (17) with $X_{lij} \equiv 1$. Use $\tilde{b}^*_{0i} = E^*(b^*_{0i} \mid Y, X)$ and $\tilde{b}^*_{1i} = E^*(b^*_{1i} \mid Y, X)$ to denote the best linear unbiased predictors under the assumed conditional density $f^*_b(b \mid y, x)$.

Using the fact that $(X_{W1ij}, X_{W2ij}, X_{Bi})$ are mutually orthogonal within clusters, the following argument shows that $\hat\beta^*_{W2}$ and $\hat\beta^*_B$ consistently estimate β_{W2} and β_B, respectively, when one ignores informative cluster sizes. Consider (17) with respect to β^*_{W2}. By orthogonality, all terms involving $X_{W2ij}X_{W1ij}$ and $X_{W2ij}X_{Bi}$ drop out of (17). When $\beta^*_{W2} = \beta_{W2}$, both r^*_{ij} and $f^*_Y(y_{ij} \mid x_{ij}, b_i)$ are free of x_{W2ij} so that the left-hand side of (13) is a linear function of X_{W2ij} and no other Xs. Since $E(X_{W2}) = 0$, it follows that $\beta^*_{W2} = \beta_{W2}$ solves (13) for all values of β^*_0, β^*_{W1}, β^*_B, σ^{*2}_{b0}, σ^{*2}_{b1}, σ^*_{12} and σ_e^{*2}. An analogous argument shows that $\hat\beta^*_B$ consistently estimates β_B when one ignores informative cluster sizes.

However, the above result on consistent estimation does not extend to $\hat\beta^*_{W1}$, the estimator associated with the covariate involved with the random slope. Even when $\beta^*_{W1} = \beta_{W1}$, r^*_{ij} depends on X_{W1ij}, unless $b^*_{1i} = b_{1i}$, so that the left-hand side of (13) will not be a linear function of X_{W1ij}. By orthogonality of the covariates, (17) with $l = W1$ is

$$E_X \left(E_{N,Y|X} \left[\sum_i \sum_{j=1}^{N(b_i)} X_{W1ij} \left\{ Y_{ij} - \tilde{b}^*_{0i} - \beta^*_0 - (\beta^*_{W1} + \tilde{b}^*_{1i}) X_{W1ij} \right\} / \sigma_e^{*2} \right] \right). \tag{18}$$

When $\beta^*_{W1} = \beta_{W1}$, (18) reduces to

$$E_X\left(E_{N,Y|X}\left[\sum_i\left(b_{1i} - \tilde{b}^*_{1i}\right)\sum_{j=1}^{N(b_i)} X^2_{W1ij}/\sigma_e^{*2}\right]\right), \tag{19}$$

which will typically not be zero. For example, we evaluate (19) using the covariate X_{W1} of the simulation studies of §4 to further study the magnitude of bias in $\hat{\beta}^*_{W1}$. In these simulations, X_{W1} takes on equally spaced values between -1 and 1 within clusters. For such a covariate, it is easy to show that $\sum_{j=1}^{n_i} x^2_{W1ij} = (n_i + 1)n_i/\{3(n_i - 1)\} \sim n_i/3$. The prediction $\tilde{b}^*_{1i}$ is a shrunken prediction of b_{1i} under the model ignoring informative cluster size and, as such, is a function of b_{1i}. Thus, (19) is essentially a covariance between $N(b_i)$ and a function of b_i, which is not zero.

An analogous argument shows that $\hat{\beta}^*_0$ will not consistently estimate β_0 even when the regression coefficients are consistently estimated. Again, this is true because the left-hand side of (12) is not a linear function of X_{lij} for $l = B, W1, W2$.

3·5. *Generalized linear mixed models with random intercepts at $\beta = 0$*

We first develop theory for the case where all covariates are uncorrelated with the outcome, that is, the case where $\beta_l = 0\ (l = 1, \ldots, p)$. Consistency results at $\beta_l = 0$ are useful since they are a prerequisite for tests of the hypothesis $H_0 : \beta_l = 0$ using estimators $\hat{\beta}^*_l$ and based on likelihoods that ignore informative cluster sizes to have the correct type I error rate. Focusing on single covariate models, when $\beta_1 = 0$, the canonical parameter θ_{ij} of (1) and the conditional mean μ_{ij} of (2) are both free of X_{ij} so that both r^*_{ij} and the $f^*_Y(y_{ij} \mid x_{ij}, b_i)$ in (12) and (13) are also free of x_{ij}. That is, when $\beta_1 = 0$, (13) is a linear function of X_{ij} and since $E(X) = 0$, it follows that the left-hand side of (13) is zero. Thus, when $\beta_1 = 0$, estimates $\hat{\beta}^*_1$ based on a likelihood that ignores informative cluster sizes consistently estimate zero and this result extends naturally to the case of multiple covariates. The above results hold for all generalized linear mixed models and thus for the mixed effects logistic and Poisson models we investigate further below.

As in §3·3, $\beta^*_0 = \beta_0$, $\beta^*_1 = \beta_1 = 0$ does not solve (12) so that $\hat{\beta}^*_0$ does not consistently estimate β_0. When $\beta^*_0 = \beta_0$ and $\beta^*_1 = \beta_1 = 0$, the left-hand side of (12) is not a linear function of X_{ij} and following an argument similar to that of §3·3 will not have expectation zero. Thus, $\hat{\beta}^*_0$ does not consistently estimate β_0 when one ignores informative cluster sizes even when $\beta_1 = 0$.

For $\beta_1 \neq 0$, analytical evaluation of the expectations of (12) and (13) is typically intractable. As an alternative, §4 presents the results of simulation studies that examine the performance of estimators from mixed effects binary and Poisson models that ignore associations between cluster sizes and outcomes.

3·6. *Generalized linear mixed models with random intercepts and slopes*

As in §3·4, we consider fitting generalized linear mixed models with random intercepts and slopes in settings where cluster sizes are associated with both the random intercepts and slopes. We assume the same covariate structure as in §3·4 with two within-cluster covariates, X_{W1} and X_{W2}, a single between-cluster covariate X_B and (X_{W1}, X_{W2}, X_B) mutually orthogonal. As in §3·5, we focus on the case where all covariates are unrelated to the response, $\beta_{W1} = \beta_{W2} = \beta_B = 0$. In this case, the canonical parameter θ_{ij} of (1) and the conditional mean μ_{ij} of (2) are both free of X_{W2ij} and X_{Bi} but not free of X_{W1ij} because of the random slope term $b_{1i}X_{W1ij}$. Thus, the consistency results for the random intercept case in §3·5 do not carry over to the case where cluster sizes are associated with both random slopes and intercepts.

However, we can obtain approximate results using Taylor expansions about $X_{W1ij} = 0$, its average value. For example, consider fitting canonical link generalized linear mixed models. For such models, $r^*_{ij} = (y_{ij} - \mu^*_{ij})/\phi$, where $\mu^*_{ij} = g^{-1}(\eta^*_{ij})$, g is a link function and η^*_{ij} is the linear predictor as in (16). Expanding r^*_{ij} in a Taylor series about $X_{W1ij} = 0$ yields the approximation

$$r^*_{ij}(X_{W1ij}) \approx r^*_{ij}(0) + X_{W1ij} r^{*\prime}_{ij}(0). \tag{20}$$

Consider (13) with respect to β^*_{W2}, replacing r^*_{ij} by its Taylor approximation (20). By orthogonality, all terms involving $X_{W2ij}X_{W1ij}$ and $X_{W2ij}X_{Bi}$ drop out of this approximation to (13). When $\beta^*_{W2} = \beta_{W2} = 0$, both r^*_{ij} and $f^*_Y(y_{ij} \mid x_{ij}, b_i)$ are free of x_{W2ij} so that the left-hand side of (13) is a linear function of X_{W2ij} and no other Xs. Since $E(X_{W2}) = 0$, it follows that $\beta^*_{W2} = \beta_{W2} = 0$ solves the approximation to (13) for all values of β^*_0, β^*_{W1}, β^*_B, σ^{*2}_{b0}, σ^{*2}_{b1}, σ^*_{12} and σ^{*2}_e. An analogous argument shows that $\beta^*_B = \beta_B = 0$ solves the approximation to (13) with respect to β^*_B. These approximations suggest that $\hat\beta^*_{W2}$ and $\hat\beta^*_B$ will exhibit little bias when one ignores informative cluster sizes.

Like the results of §3·5, the above results on consistent estimation do not extend to $\hat\beta^*_{W1}$, the estimator associated with the covariate involved with the random slope. Even when $\beta_{W1} = 0$ and using the Taylor approximation in (20), both θ_{ij} and μ_{ij} depend on X_{W1ij} through the random slope term $b_{1i}X_{W1ij}$. Thus, r^*_{ij} depends on X_{W1ij}, unless $b^*_{1i} = b_{1i}$, so that the left-hand side of (13) will not be a linear function of X_{W1ij}.

As in §3·3, $\beta^*_0 = \beta_0$, $\beta^*_1 = \beta_1 = 0$ does not solve (12) so that $\hat\beta^*_0$ does not consistently estimate β_0. When $\beta^*_0 = \beta_0$ and $\beta^*_1 = \beta_1 = 0$, the left-hand side of (12) is not a linear function of X_{ij} and following an argument similar to that of §3·3 will not have expectation zero. Thus, $\hat\beta^*_0$ does not consistently estimate β_0 when one ignores informative cluster sizes even when $\beta_1 = 0$.

For $\beta_l \neq 0$, analytical evaluation of the expections of (12)–(13) is typically intractable. As an alternative, §4 presents the results of simulation studies that examine the performance of estimators from mixed effects binary and Poisson models that ignore associations between cluster sizes and outcomes.

3·7. *Conditional likelihood methods*

For canonical link generalized linear mixed models with only random intercepts, a conditional likelihood approach would treat the cluster-specific intercepts b_i in (2) as fixed constants and eliminate them from the likelihood by conditioning on their sufficient statistics $\sum_{j=1}^{n_i} y_{ij}$. That is, one would compute

$$f_{Y,N,X}\left(y_{i1}, \ldots, y_{in_i}, n_i, x_{i1}, \ldots, x_{in_i} \mid b_i, \sum_{j=1}^{n_i} y_{ij}\right).$$

From (9), we see that it is equivalent to compute

$$f_Y\left(y_{i1}, \ldots, y_{in_i} \mid n_i, x_{i1}, \ldots, x_{in_i}, b_i, \sum_{j=1}^{n_i} y_{ij}\right). \tag{21}$$

However, (21) is just the standard conditional likelihood one would compute from a generalized linear mixed model where there are no correlations between the cluster sizes and covariates, i.e., (4). Thus, the conditional likelihoods corresponding to (9) and to (10) will coincide and the approach will provide consistent estimates of the effects of within-cluster covariates when cluster sizes are associated with the covariates. The consistency of conditional likelihood estimates makes sense intuitively. The densities of cluster sizes of (9) given the cluster-specific intercept b_i, do not vary within clusters and the conditional likelihood approach eliminates such terms from the likelihood. The simulations studies in §4 illustrate the performance of the conditional likelihood approach for models with only random intercepts. Standard conditional likelihood methods will perform poorly in settings with subject-specific slopes as well as intercepts since conditioning on sufficient statistics for the intercepts will not appropriately accommodate the slopes.

4. Simulations

We carried out two sets of simulations to examine the effects of ignoring associations of cluster sizes with the distribution of the responses on the standard cluster-specific methods of mixed-effects models and conditional likelihood approaches. The simulations generated clustered binary or Poisson responses from generalized linear mixed models where the random effects b were correlated with cluster size:

$$y_{ij} \mid b_i, n_i, x_{ij} \overset{\text{ind}}{\sim} f_{y|b,n,x} \quad (i = 1, \ldots, m;\ j = 1, \ldots, n_i), \tag{22}$$

$$g\{\mathrm{E}(y_{ij}|b_i, z_{ij}, x_{ij})\} = z_{ij}^{\mathrm{T}} b_i + x_{ij}^{\mathrm{T}} \beta, \tag{23}$$

$$N_i \mid b_i \sim \mathrm{Po}\left(e^{\gamma_0 + \gamma_1 b_{0i} + \gamma_2 b_{1i}}\right) + N_{\min}, \tag{24}$$

$$b_i \overset{\text{ind}}{\sim} N(0, \Sigma_B), \tag{25}$$

and $N_{\min}$ is an offset to avoid cluster sizes that are too small for estimating within-cluster covariate effects. Specifically, the first set of simulations included a single covariate, x_{ij}, a Bernoulli (0·5) within-cluster covariate, a random intercept, $z_{ij} \equiv 1$, and no random slope terms. The second set of simulations followed the work of §3·4 and §3·6 and generated responses with both random intercepts and slopes and three mutually orthogonal covariates: a within-cluster covariate, x_{W1ij} that was equally spaced values between -1 and 1 within clusters; x_{W2ij}, a within-cluster covariate constructed to be orthogonal to x_{W1ij}; and a between-cluster covariate, x_{Bi}, distributed Bernoulli (0·5). The random slope was associated with x_{W1ij} and the model used $z_{ij}^{\mathrm{T}} = (1, x_{W1ij})^{\mathrm{T}}$. Simulation 1 set $N_{\min} = 2$ in equation (24) to preclude cluster sizes of 0 or 1, whereas simulation 2 set $N_{\min} = 4$. The simulation set 1 used $\gamma_0 = \gamma_1 = 1{\cdot}0$, whereas simulation set 2 used $\gamma_0 = 1$, $\gamma_1 = \gamma_2 = 3^{-1/2}$. Calculations show that the expected cluster size is $E(N) = \exp(\gamma_0 + 0{\cdot}5\, \sigma_b^2\, \gamma_1^2) + 2$ for simulation 1 and $E(N) = \exp\{\gamma_0 + 0{\cdot}5(\sigma_{b0}^2 \gamma_1^2 + 2\sigma_{12}\gamma_1\gamma_2 + \sigma_{b1}^2 \gamma_2^2)\} + 4$ for simulation 2.

The simulations generated responses from three different binary mixed-effects models using the logistic, probit and complementary log-log link functions, as well as from a mixed-effects Poisson model with a log link, for both settings described above. Each simulation generated 1000 datasets, each with 100 clusters. The mixed-effects model parameter values for the first set of simulations were $\beta_0 = -2{\cdot}5$, $\beta_1 = 1{\cdot}0$ and $\mathrm{var}(b_i) = 1{\cdot}0$, while they were $\beta_0 = -2{\cdot}5$, $\beta_{W1} = 1{\cdot}0$, $\beta_{W2} = 1{\cdot}0$, $\beta_B = 1{\cdot}0$, $\mathrm{var}(b_{0i}) = \mathrm{var}(b_{1i}) = 1{\cdot}0$ and $\mathrm{cov}(b_{0i}, b_{1i}) = 0{\cdot}5$ for simulation set 2. For the given inputs, the average cluster size was $E(N) = 6{\cdot}48$ for the first set of simulations and $E(N) = 8{\cdot}48$ for the second set.

We fit two approaches to each dataset: the joint model based on (22)–(25) that used correct specifications of $f_Y(y \mid x, n, b)$ and $f_N(n \mid b)$; and a standard binary mixed-effects or Poisson model that ignored cluster size associations, based on (22), (23) and (25), but not (24). For the first set of simulations involving mixed-effects logistic and Poisson models, we also fit the

Table 1. *Observed means and standard deviations, in parentheses, of the regression coefficients of several methods for fitting generalized linear mixed models with random intercepts to simulated clustered data where the random intercepts were associated with cluster sizes*

Distribution	Link	Model	β_0	β_1	σ_b
Binary	logit	Ignore n_i	−2·16 (0·23)	1·01 (0·14)	0·94 (0·20)
		CML		1·01 (0·15)	
		Joint	−2·52 (0·24)	1·01 (0·13)	1·00 (0·17)
Binary	probit	Ignore n_i	−2·16 (0·24)	1·00 (0·13)	0·89 (0·18)
		Joint	−2·53 (0·25)	1·01 (0·12)	1·00 (0·16)
Binary	comp log-log	Ignore n_i	−2·20 (0·21)	1·02 (0·11)	0·93 (0·16)
		Joint	−2·51 (0·22)	1·01 (0·11)	0·98 (0·17)
Poisson	log	Ignore n_i	−2·19 (0·19)	1·01 (0·07)	0·89 (0·14)
		CML		1·00 (0·07)	
		Joint	−2·50 (0·20)	1·00 (0·07)	0·98 (0·14)

Ignore n_i, standard generalized linear mixed models that ignored cluster size associations, based on (22), (23) and (25); Joint, joint models using (22)–(25) that used correct specifications of $f_Y(y \mid x, n, b)$ and $f_N(n \mid b)$; and for mixed-effects logistic and Poisson models, CML, conditional likelihood approaches (4). True values: $\beta_0 = -2{\cdot}5$, $\beta_1 = 1{\cdot}0$, $\sigma_b = 1{\cdot}0$.

conditional likelihood approach of (4). A conditional likelihood approach is not available for the mixed-effects probit or complementary log-log models since these are not canonical link models. We fit the mixed-effects and joint models using Proc NLMIXED in SAS and the conditional likelihood methods for the logistic link using Proc PHREG in SAS. Lancaster (2002) showed that one can obtain conditional maximum likelihood estimates for Poisson responses using a standard Poisson model that includes a fixed parameter b_i for each cluster. We fit the conditional likelihood methods for Poisson responses following such a fixed effects Poisson approach using Proc GENMOD in SAS.

Table 1 displays the average values of the parameter estimates along with standard deviations of these estimates from the first set of simulations that generated data from and fit random intercept models. The simulation findings in Table 1 closely correspond to the results of §3. Ignoring the correlation between cluster sizes and random effects produced no detectable bias in estimates of the regression coefficient β_1. However, the estimates of the intercept β_0 were biased. Also, there were small biases in estimating the variance component σ_b. The conditional likelihood approach yielded consistent estimates of β_1, but they were slightly less efficient than the mixed-effects model estimates. All parameter estimates from the correctly specified joint model, (22)–(25), closely corresponded to the true values and their simulation standard errors were slightly smaller than those of the other approaches. Thus, correctly modelling the association of random effects with cluster sizes when the covariates were independent of the random effects did not produce large efficiency gains.

The simulation results in Table 1 for the probit, complementary log–log and Poisson models paralleled those for the logistic model. When the random effects were associated with cluster sizes, Table 1 shows that ignoring cluster size associations produces no bias in estimates of the regression coefficient β_1. As with the logistic model, Table 1 shows that all parameter estimates from the correctly specified joint probit, complementary log-log and Poisson models (22)–(25) closely corresponded to the true values and their simulation standard errors were slightly smaller than those of the other approaches. Again, however, correctly modelling the association of random intercepts with cluster sizes did not produce large efficiency gains. As with the logistic model, Table 1 shows that the conditional likelihood approach for Poisson responses yielded

Table 2. *Observed means and standard deviations, in parentheses, of the regression coefficients of several methods for fitting generalized linear mixed models with random intercepts and slopes to simulated clustered data where the random intercepts and slopes were associated with cluster sizes*

Distribution	Link	Model	β_0	β_{W1}	β_{W2}	β_B
Binary	logit	Ignore n_i	−2·38(0·27)	1·10 (0·23)	0·97 (0·20)	1·01 (0·32)
		Joint	−2·55 (0·25)	1·01 (0·24)	1·01 (0·21)	1·02 (0·25)
Binary	probit	Ignore n_i	−2·42 (0·31)	1·03 (0·23)	0·96 (0·17)	1·00 (0·32)
		Joint	−2·55 (0·28)	1·00 (0·24)	1·01 (0·17)	1·02 (0·22)
Binary	comp log-log	Ignore n_i	−2·44 (0·26)	1·07 (0·21)	0·99 (0·16)	1·01 (0·30)
		Joint	−2·54 (0·23)	1·02 (0·21)	1·01 (0·16)	1·03 (0·21)
Poisson	log	Ignore n_i	−2·35 (0·21)	1·06 (0·18)	0·93 (0·11)	0·95 (0·25)
		Joint	−2·51 (0·17)	1·01 (0·18)	1·00 (0·11)	1·00 (0·10)

Distribution	Link	Model	$\log\sigma_{b0}$	$\log\sigma_{b1}$	σ_{12}
Binary	logit	Ignore n_i	−0·02 (0·19)	0·01 (0·28)	0·63 (0·29)
		Joint	−0·01 (0·18)	0·02 (0·05)	0·56 (0·29)
Binary	probit	Ignore n_i	−0·05 (0·19)	0·02 (0·23)	0·62 (0·23)
		Joint	−0·01 (0·17)	0·01 (0·20)	0·55 (0·24)
Binary	comp log-log	Ignore n_i	−0·01 (0·16)	0·07 (0·24)	0·62 (0·23)
		Joint	−0·01 (0·16)	0·00 (0·21)	0·55 (0·24)
Poisson	log	Ignore n_i	−0·09 (0·14)	−0·06 (0·15)	0·52 (0·18)
		Joint	−0·01 (0·12)	−0·02 (0·13)	0·52 (0·17)

Ignore n_i, standard generalized linear mixed models that ignored cluster size associations, based on (22), (23) and (25); Joint, joint models using (22)–(25) that used correct specifications of $f_Y(y \mid x, n, b)$ and $f_N(n \mid b)$. True values: $\beta_0 = -2{\cdot}5$, $\beta_{W1} = \beta_{W2} = \beta_B = 1{\cdot}0$, $\log\sigma_{b0} = \log\sigma_{b1} = 0$ and $\sigma_{12} = 0{\cdot}5$.

consistent estimates of β_1 but they were slightly less efficient than the mixed-effects model estimates. The convergence rates for the simulations of Table 1 were very high, exceeding 99.8% for most models. However, with the complementary log-log link the joint model, (22)–(25) failed to converge for 21 of 1000 datasets. Table 1 reports the results for the 979 datasets where both methods converged.

Table 2 displays the results from the second set of simulations that generated data from and fit models that included both random intercepts and slopes, as well as both within- and between-cluster covariates. The results in Table 2 closely follow the findings of §3·4 and §3·6. For all three binary link models, as well as for the Poisson, ignoring the associations of the random intercept and random slope for x_{W1ij} with cluster sizes produced no substantial bias in estimates of the regression coefficients β_{W2} or β_B. As in Table 1, correctly modelling the association of random intercepts and slopes with cluster sizes did not produce large efficiency gains in estimates of β_{W2} or β_B. Following the theory of §3·4 and §3·6 and analogous to the results of Table 1 for intercepts, ignoring the association of the random slope for x_{W1ij} with cluster sizes yielded biased estimates of β_{W1}. As in the simulations involving random intercept models, ignoring informative cluster sizes in settings with both random intercepts and slopes yielded biased estimates of the intercept, typically not the parameter of central interest. As in Table 1, convergence rates for the simulations of Table 1 were very high, exceeding 98% for all models.

We also conducted limited simulation studies not reported here which followed the design of the second set of simulations but generated cluster sizes dependent only on the random intercepts. Consistent with the work of §3, maximum likelihood methods ignoring informative cluster sizes

Table 3. *Parameter and standard error estimates, in parentheses, from three methods for fitting mixed-effects logistic models to the periodontal data (Gansky et al., 1999).*

	Ignore n_i	CML	Joint
β_0	$-1{\cdot}30$ $(0{\cdot}15)$		$-1{\cdot}09$ $(0{\cdot}16)$
Molar (β_1)	$0{\cdot}77$ $(0{\cdot}09)$	$0{\cdot}83$ $(0{\cdot}09)$	$0{\cdot}82$ $(0{\cdot}09)$
σ_b	$2{\cdot}67$ $(0{\cdot}16)$		$2{\cdot}77$ $(0{\cdot}16)$

Ignore n_i, standard mixed-effects logistic models that ignored cluster size associations, based on (22), (23) and (25); CML, conditional likelihood approaches (4); Joint, joint models using (22)–(25) that used correct specifications of $f_Y(y \mid x, n, b)$ and $f_N(n \mid b)$.

yielded little bias in estimated covariate effects. However, standard conditional likelihood methods, assuming random intercepts only, yielded biased covariate effects.

5. EXAMPLE: PERIODONTAL DISEASE

Data from a study of periodontal disease (Gansky et al., 1999) motivated our investigations of the effects of informative cluster sizes. The dataset consists of tooth-specific observations on 407 subjects. Analyses of periodontal disease data often focus on pre-molar and molar teeth since the front teeth are less susceptible to gum disease. We thus restrict the analysis to the molar and pre-molar teeth so that a subject can provide a maximum of 16 observations. The outcome, Y, is a binary, tooth-specific indicator of periodontal disease and the single fitted covariate, x, is a binary indicator of whether the tooth is a molar versus pre-molar. The number of teeth per subject varied from 1 to 16 with an average of $11{\cdot}6$.

A logistic regression of presence of periodontal disease on the number of teeth using the generalized estimating equations method and independence working correlation (Diggle et al., 2002) yielded an estimated regression coefficient of $-0{\cdot}15$ for the number of teeth along with a robust standard error of $0{\cdot}015$, indicating a highly statistically significant association of cluster size and the expected value of the response.

Table 3 presents the results of three methods for fitting mixed-effects logistic models to the clustered periodontal disease data to assess the association of tooth-specific disease prevalence with tooth type. The three methods are the same as those in Table 1 and we used the same SAS procedures to implement them. That is, we fit the joint model (22)–(25), a standard mixed-effects logistic model that ignored cluster size associations (22), (23) and (25) and the conditional likelihood approach, (4). The joint model used a similar specification of $f_N(n \mid b)$ as in Table 1 simulations namely $N \sim \text{Po}\{\exp(\gamma_0 + \gamma_1 b)\} + 1$.

The results in Table 3 closely correspond to the findings in §4. In §1, we noted that the number of teeth per subject was associated with the distribution of the response, the binary indicator of periodontal disease. The estimates of the parameters of $f_N(n \mid b)$ were $\hat{\gamma}_0 = 2{\cdot}424$, $\text{se}(\hat{\gamma}_0) = 0{\cdot}019$ and $\hat{\gamma}_1 = -0{\cdot}244$, $\text{se}(\hat{\gamma}_1) = 0{\cdot}019$. Thus, there is a highly statistically significant relationship between N and b. To put the value of $\hat{\gamma}_1$ in context, for each standard deviation increase in

b, across a reasonable range of values for b, this is about a 40% decrease in the mean value of N. Table 3 also indicates that the conditional likelihood estimate and standard error are nearly identical to the estimated coefficient and standard error of the joint model approach. The estimate from the fit that ignored cluster size associations is also similar. Estimates of the intercept, β_0, are more discrepant.

6. Discussion

We studied linear mixed effects and generalized linear mixed models in cases where cluster sizes are associated with one or more random effects and focus on the performance of maximum likelihood methods when ignoring informative cluster sizes. In the case of linear mixed effects models, ignoring informative cluster sizes yields consistent estimation for the effects of covariates uncorrelated with the random effects that are associated with cluster size. However, estimation of covariate effects that are associated with the random effects can be biased. In particular, for models with random intercepts only, ignoring informative cluster sizes can yield biased estimators of the intercept but yields consistent estimators of all covariate effects.

Our theoretical results for the case of generalized linear mixed models are less comprehensive than for linear mixed effects models. For the random intercepts case, we show that ignoring informative cluster sizes yields consistent estimators of covariate effects when they are zero. For models that also include random slopes, approximations and simulation studies suggest little bias in estimating the effects of covariates uncorrelated with the random effects that are associated with cluster size, but bias otherwise.

Standard conditional maximum likelihood methods, which assume random intercepts only, give consistent estimates of covariate effects when the true model contains only random intercepts and the cluster size depends on those random intercepts. However, in the presence of random slopes, conditional maximum likelihood methods give biased estimation of covariate effects, even when the cluster size depends only on the random intercepts.

While we have assumed throughout the paper that cluster sizes do not depend on covariates, in practice they may. If this dependence arises through shared random effects as in (5), then the results of Neuhaus & McCulloch (2006) indicate that ignoring informative cluster sizes in this case can produce biased covariate effect estimators. However, Neuhaus & McCulloch (2006) also suggest that approaches that partition covariates into between- and within-cluster components, as well as conditional likelihood methods, provide consistent estimation in settings with covariate-dependent cluster sizes.

Acknowledgement

This research was supported by a grant from the U.S. National Institutes of Health. We thank Drs Stuart Gansky and Jane Weintraub of the University of California, San Francisco, for providing the periodontal study data and Dr Ross Boylan of the University of California, San Francisco, for assistance with computing.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second Int. Symp. Inform. Theory*, Ed. B. N. Petrov and F. Czáki, pp. 267–81. Budapest: Akademiai Kiadó.

Benhin, E., Rao, J. N. K. & Scott A.J. (2005) Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92**, 435–50.

Diggle, P. J., Heagerty, P. J., Liang, K. Y., & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford: Oxford University Press.

Dudley, R. A., Johansen, K. L., Brand, R. J., Rennie, D. J., & Milstein, A. (2002). Selective referral to high-volume hospitals: estimating potentially avoidable deaths. *J. Am. Med. Assoc.* **283**, 1191–3.

Dunson, D., Chen, Z. & Harry, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59**, 521–30.

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

Gueorguieva, R. (2005). Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics* **61**, 862–7.

Heagerty, P. J. & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika* **88**, 973–86.

Hoffman, E., Sen, P. & Weinberg, C. (2001). Within-cluster resampling. *Biometrika* **85**, 1121–34.

Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Rev. Econ. Studies* **69**, 647–66.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley.

McCulloch, C. E. & Neuhaus, J. M. (2005). Generalized linear mixed models. In *Encyclopedia of Biostatistics,* 2nd ed., Ed. P. Armitage and T. Colton, pp. 2085–9. Chicester: Wiley.

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M. (2008). *Generalized, Linear and Mixed Models*, 2nd ed. New York: Wiley.

Neuhaus, J. M. (2001). Assessing change with longitudinal and clustered binary data. *Ann. Rev. Public Health* **22**, 115–28.

Neuhaus, J. M., Hauck, W. W. & Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–62.

Neuhaus, J. M., Kalbfleisch, J. D. & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged models for analyzing correlated binary data. *Int. Statist. Rev.* **59**, 25–35.

Neuhaus, J. M. & McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects using conditional and partitioning methods. *J. R. Statist. Soc.* B **68**, 859–72.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Williamson, J., Datta, S. & Satten, G. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**, 36–42.

[*Received January* 2009. *Revised August* 2010]