BIOMETRICS 54, 638–645
June 1998

Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data

J. M. Neuhaus*

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

and

J. D. Kalbfleisch

Department of Statistics and Actuarial Science, University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada

SUMMARY

Standard methods for the regression analysis of clustered data postulate models relating covariates to the response without regard to between- and within-cluster covariate effects. Implicit in these analyses is the assumption that these effects are identical. Example data show that this is frequently not the case and that analyses that ignore differential between- and within-cluster covariate effects can be misleading. Consideration of between- and within-cluster effects also helps to explain observed and theoretical differences between mixture model analyses and those based on conditional likelihood methods. In particular, we show that conditional likelihood methods estimate purely within-cluster covariate effects, whereas mixture model approaches estimate a weighted average of between- and within-cluster covariate effects.

1. Introduction

Clustered samples arise frequently in practice. This clustering may be due to gathering repeated measurements on experimental units as in longitudinal studies or may be due to subsampling the primary sampling units. The latter type of design is common in fields such as ophthalmology, where two eyes form natural clusters, and teratology, where one gathers data on all members of a litter. The data consist of an outcome variable Y_{ij} together with a p-dimensional vector of covariates X_{ij}. The data are gathered in clusters or groups, and $i = 1, \ldots, m$ indexes clusters while $j = 1, \ldots, n_i$ indexes units within clusters. Cluster data tend to exhibit intracluster correlation, which the analysis must address in order to obtain valid inferences.

With clustered data, it is useful to mention two special types of covariates. The first type, a cluster-constant or cluster-level covariate, has the same value for all the units in the cluster, i.e., $X_{ij} = \bar{X}_i$ for all j. Examples include covariates measured on the pregnant female in teratologic studies and on the individual in ophthalmologic studies. The second covariate type, a designed within-cluster covariate, varies with identical distribution across the units within each cluster. Such covariates arise in matched pair studies and more generally in randomized block experiments where each unit in a cluster receives a different treatment and the set of treatments is the same in all clusters. In general, however, the covariate assumes a different value for each unit in the cluster and the covariate pattern and cluster mean of the covariate, $\bar{X}_i$, vary between clusters. Examples of such covariates include nondesigned time-dependent covariates in longitudinal studies and the age of each family member in family studies.

In general, therefore, a covariate has both a between-cluster component, which we might summarize in terms of $\bar{X}_i$, the cluster mean, and a within-cluster component $X_{ij} - \bar{X}_i$; it is important

[*] *Corresponding author's email address:* john@biostat.ucsf.edu

Key words: Bias; Conditional likelihood; Misspecified models.

to assess how each component affects response. For example, in longitudinal studies, we can distinguish between (i) an overall effect of age on response, as measured by the association of the mean age $\bar{X}_i$ with the response, and (ii) the effects of deviations from the average age $X_{ij} - \bar{X}_i$ on the series of responses within the cluster. Typically, authors do not distinguish between- and within-cluster covariate effects in generalized linear mixed models and so implicitly assume that these effects are the same. Section 2 of this paper presents some examples to illustrate that between- and within-cluster covariate effects can be very different; as a consequence, models that incorrectly assume common effects can lead to very misleading assessments of the association of covariates with response. Section 3 derives expressions for the bias resulting from incorrectly assuming common between- and within-cluster covariate effects and points out the role of conditional likelihood methods in the estimation of within-cluster effects. Section 4 presents a discussion of our findings.

Generalized linear mixed models (see, e.g., Breslow and Clayton, 1993) provide a useful approach to assess covariate effects. Within the ith cluster, the responses Y_{ij} are independent and follow a generalized linear model with parameters that can vary between clusters. Thus, given a vector of parameters specific to the ith cluster, a_i, the conditional density of Y_{ij} is of the form $f(y_{ij} \mid a_i) = \exp[\{y_{ij}\theta_{ij}(a_i) - b[\theta_{ij}(a_i)]\}\phi + c(y_{ij}, \phi)]$, where b and c are functions of known form and $\mathrm{E}(Y_{ij} \mid a_i) = b'[\theta_{ij}(a_i)]$. In addition, we assume that

$$\mathrm{E}(Y_{ij} \mid a_i) = g^{-1}(a_i + \beta X_{ij}).$$

The function g links the linear predictor to the expected response and is assumed to be strictly monotone and differentiable. Interest focuses on the parameter β, which measures the change in the conditional expectation within the ith cluster corresponding to a unit increase in the covariate, i.e.,

$$\beta = g[\mathrm{E}(Y \mid X + 1, a)] - g[\mathrm{E}(Y \mid X, a)].$$

The model further assumes that the random effects a follow a distribution G, typically multivariate normal with unknown mean vector and covariance matrix. One estimates model parameters by maximizing the likelihood.

For the important case of generalized linear models with canonical links and random intercepts (a_i a scalar), one can estimate the regression parameters corresponding to within-cluster covariates using a conditional likelihood that eliminates the random intercepts. One obtains these conditional likelihoods by computing the probability of a cluster response conditional on the cluster sum $S_i = \Sigma_{j=1}^{n_i} y_{ij}$, which is a sufficient statistic for a_i. The conditional likelihood is the product of ratios of density functions of the form

$$f(y_{i1}, \ldots, y_{in_i} \mid S_i, X_{i1}, \ldots, X_{in_i}) = \frac{\prod_{j=1}^{n_i} f(y_{ij} \mid a_i, X_{ij})}{f(S_i \mid a_i, X_{i1}, \ldots, X_{in_i})}. \tag{1}$$

Note that (1) does not depend on a_i by the sufficiency of S_i. In addition to removing the cluster-specific effects a_i, the conditioning in (1) also removes the effects of cluster-level covariates from the likelihood.

For Gaussian responses, the canonical link is the identity and the conditional densities $f(y_{ij} \mid a_i, X_{ij})$ of (1) are each Gaussian with mean $a_i + \beta X_{ij}$ and variance $\sigma_w{}^2$. The conditional likelihood (1) associated with the linear mixed-effects model is the product of terms of the form

$$(n_i)^{1/2}\left[(2\pi)^{(n_i-1)/2}\sigma_w{}^{n_i-1}\right]^{-1}\exp\left\{-\left[2\sigma_w{}^2\right]^{-1}\sum_{j=1}^{n_i}[y_{ij} - \bar{y}_i - \beta(X_{ij} - \bar{X}_i)]^2\right\}, \tag{2}$$

where $\bar{y}_i$ and $\bar{X}_i$ are the means of the response and covariate, respectively, in the ith cluster. Note that (2) depends on the covariates and responses only through the differences $(X_{ij} - \bar{X}_i)$ and $(y_{ij} - \bar{y}_i)$ and that the ordinary least squares estimator is unbiased for β. Under the mixed model, the differences $(Y_{ij} - \bar{Y}_i)$ are correlated within clusters so that standard error calculations need to address this dependence using standard likelihood methods applied to (2) or using cluster variance estimation methods such as those in Liang and Zeger (1986). Regression methods that assess the association of changes in the response with changes in covariates are useful approaches for longitudinal data (see, e.g., Louis et al., 1986; Louis, 1988; Ware et al., 1990) and (2) shows that this approach is essentially the conditional likelihood approach for Gaussian responses.

The joint probability of $(Y_{i1}, \ldots, Y_{in_i})$ under the mixed-effects logistic model with a random intercept is

$$\text{pr}(Y_{i1} = y_{i1}, \ldots, Y_{in_i} = y_{in_i} \mid X_{i1}, \ldots, X_{in_i}) = \int \prod_j \{p_{ij}{}^{y_{ij}} q_{ij}{}^{1-y_{ij}}\} dG(a),$$

where $\text{logit}\, p_{ij} = \beta_0 + a + \beta X_{ij}$, $q_{ij} = 1 - p_{ij}$, and G is the distribution of the random effect. The conditional likelihood (1) associated with the mixed-effects logistic model is the product of terms of the form

$$\frac{\exp \sum_{j=1}^{n_i} y_{ij} X_{ij} \beta}{\sum_{\text{L}} \exp \sum_{k \in \text{L}} X_{ik} \beta}, \tag{3}$$

where L ranges over all the $\binom{n_i}{S_i}$ possible subsets of S_i elements from n_i. It is clear that (3) depends on the covariates only through the differences $(X_{ij} - X_{ik})$ and that the set of differences is equivalent to the set of deviations of each covariate from the cluster mean $(X_{ij} - \bar{X}_i)$.

Given the cluster sum S_i, the conditional probability (3) of a cluster of concordant outcomes does not depend on β, so the conditional likelihood approach only uses data from clusters with discordant outcomes.

This paper focuses on two commonly used generalized linear mixed models, the mixed-effects linear (Laird and Ware, 1982) and logistic models (Stiratelli, Laird, and Ware, 1984) for the case of random intercepts. The mixed-effects linear model assumes that

$$\text{E}(Y_{ij} \mid a_i, X_{ij}) = a_i + \beta X_{ij}, \tag{4}$$

with $a \sim$ G. The mixed-effects logistic model assumes that

$$\text{logit}\, \text{pr}(Y_{ij} = 1 \mid a_i, X_{ij}) = a_i + \beta X_{ij}, \tag{5}$$

with $a \sim$ G.

2. Between- and Within-Cluster Covariate Effects

In this section, we fit models that allow separate between- and within-cluster covariate effects to some examples to illustrate that these effects can be very different. The examples also illustrate that an incorrect assumption of common between- and within-cluster covariate effects can give quite misleading estimates of the effect of the covariate on response. We decompose the covariate X_{ij} into between- $(\bar{X}_i = n_i{}^{-1}\Sigma_{j=1}^{n_i} X_{ij})$ and within-cluster $(X_{ij} - \bar{X}_i)$ components and extend models (4) and (5) to allow different effects of the two components on response, i.e., we replace βX_{ij} in (4) and (5) by $\beta_B \bar{X}_i + \beta_W (X_{ij} - \bar{X}_i)$.

The first data set comes from a Centers for Disease Control study of birth weights in Georgia and the sample consists of 880 women, each of whom had 5 children. The outcome of interest is birth weight, and we consider two response variables, a continuous measure of birth weight (grams) and a binary measure where $Y = 1$ indicates birth weight $\leq$2500 grams and $Y = 0$ indicates birth weight >2500 grams. The covariate of interest X_{ij} is the mother's age at each birth, which we decompose into between- and within-cluster components.

To the continuous response data, we fit a mixed-effects linear model of the form

$$\text{E}(Y_{ij} \mid a_i, X_{ij}) = a_i + \beta_B \bar{X}_i + \beta_W (X_{ij} - \bar{X}_i), \tag{6}$$

with $a \sim \text{N}(\beta_0, \sigma_b{}^2)$ using the 5V program in BMDP. Table 1 presents the results for model (6) as well as for model (4), which assumes that $\beta_B = \beta_W$. Table 1 also presents the conditional likelihood estimate from (2). We obtained the conditional likelihood estimate, using the 5V program, as the generalized least squares estimator of the slope of the regression of the differences $(Y_{ij} - \bar{Y}_i)$ on $(X_{ij} - \bar{X}_i)$ with no intercept. Table 1 shows that the effect of mother's average age is very different from the effect of the deviation of her age at each birth from her mean. One can easily perform significance tests of the hypothesis of common between- and within-cluster covariate effects using likelihood ratio or Wald methods and the observed difference in Table 1 is highly significant. From model (6), we estimate that average birth weight will differ by 30.35 grams between two women whose average age at birth differs by one year. The coefficient of $\overline{\text{Age}}_i$ measures differences in birth weights between women who had children at different periods in their lives. The increase in birth weight associated with increasing mean age may relate to the fact that the women in this sample began having children at a relatively young age; the median age of these women at the birth of their first child was 17 years. On the other hand, the coefficient of Age $-$ $\overline{\text{Age}}_i$ suggests that, for a

Table 1

Comparison of mixed-effects linear model parameter estimates from models assuming common (4) *and different* (6) *between- and within-cluster age effects and the conditional likelihood approach* (2) *for the continuous Georgia birth-weight data* (*CDC*)

	$\hat{\beta}$ (S.E.)		
Var	$\beta_B = \beta_W$ (4)	$\beta_B \neq \beta_W$ (6)	Cond like (2)
Age	17.14 (1.98)		
$\overline{\text{Age}}_i$		30.35 (3.67)	
Age $-$ $\overline{\text{Age}}_i$		11.83 (2.34)	11.85 (2.35)
σ_b	354.6	351.3	
σ_w	434.2	433.9	

given woman, the birth weights of her children increase by an average of about 11.8 grams for each year that she ages. Table 1 also illustrates that the model (4), with common between- and within-cluster effects, estimates neither within- nor between-cluster effects and that the within-cluster age effect from (6) is nearly identical to the conditional likelihood estimate from (2).

To the binary response data we fit an analogous mixed-effects logistic model of the form

$$\operatorname{logit} \operatorname{pr}(Y_{ij} = 1 \mid X_{ij} a_i) = a_i + \beta_B \bar{X}_i + \beta_W (X_{ij} - \bar{X}_i), \tag{7}$$

with $a \sim \mathrm{N}(\beta_0, {\sigma_b}^2)$ using a routine in the EGRET package. Table 2 presents the estimates from model (7), those from model (5) with $\beta_B = \beta_W$, and the conditional likelihood estimate from (3) and a routine in the EGRET package. Again, the estimated effect of the mother's average age is very different from the effect of the deviation of her age at each birth from her mean. Analogous to model (6), the coefficient of $\overline{\text{Age}}_i$ in Table 2 measures differences in the logit of the probability of a low birth weight between women who differ in the mean age at which they had their children. In general, the regression coefficients associated with cluster-level covariates in nonlinear mixed-effects models are difficult to interpret (Neuhaus, Kalbfleisch, and Hauck, 1991). Since model (7) measures covariate effects conditional on the random effects a_i, the coefficient of $\overline{\text{Age}}_i$ actually measures differences on a logit scale among women who share the same random effect. This interpretation is conceptually simpler, perhaps, for random effects with discrete distributions such as in latent class settings. The coefficient of Age $-$ $\overline{\text{Age}}_i$ suggests that there is no evidence that the probability of a low birth weight baby changes for a given mother as she ages. The model that assumes common between- and within-cluster age effects measures neither of these effects, and the within-cluster age effect from (7) is nearly identical to the conditional likelihood estimate from (3).

The second data set comes from a study of 31 patients with periodontal disease described by TenHave, Landis, and Weaver (1995). The study examined the association of a method for identifying tooth sites at high risk of periodontal disease progression, a visual elastase kit (X), with an indicator of periodontal disease progression over 26 weeks (Y). The study investigators gathered ordinal kit scores with a range of 0 to 4 and indicators of disease progression from up to five tooth sites for each patient. The total sample consisted of 131 tooth sites. Table 3 presents estimates from model (7), from model (5) with $\beta_B = \beta_W$, and the conditional likelihood estimate from (3). The results here are similar to those of the birth weight data. The effect across patients of average

Table 2

Comparison of mixed-effects logistic model parameter estimates from models assuming common (5) *and different* (7) *between- and within-cluster age effects and the conditional likelihood approach* (3) *for the binary Georgia birth-weight data* (*CDC*)

	$\hat{\beta}$ (S.E.)		
Var	$\beta_B = \beta_W$ (5)	$\beta_B \neq \beta_W$ (7)	Cond like (3)
Age (10 year)	-0.34 (0.14)		
$\overline{\text{Age}}_i$		-0.81 (0.22)	
Age $-$ $\overline{\text{Age}}_i$		0.05 (0.20)	0.05 (0.20)
σ_b	1.29 (0.10)	1.30 (0.10)	

Table 3

Comparison of mixed-effects logistic model parameter estimates from models assuming common (5) *and different* (7) *between- and within-cluster kit score effects and the conditional likelihood approach* (3) *for the periodontal data* (*TenHave et al.,* 1995)

	$\hat{\beta}$ (S.E.)		
Var	$\beta_B = \beta_W$ (5)	$\beta_B \neq \beta_W$ (7)	Cond like (3)
Kitscore	0.72 (0.20)		
$\overline{\text{Kit}}_i$		1.46 (0.42)	
Kit $- \overline{\text{Kit}}_i$		0.44 (0.23)	0.46 (0.24)
σ_b	0.97 (0.39)	0.91 (0.38)	

kit score is very different from the within-patient effect of the deviation of the score at each tooth site from the patient's average. The coefficient of $\overline{\text{Kit}}_i$ measures differences between patients who have different average kit scores. On the other hand, the coefficient of Kit $- \overline{\text{Kit}}_i$ suggests that, for a given patient, the logit will differ by 0.44 units between two sites that differ by one unit on the kit score. Model (5) measures neither between- nor within-cluster covariate effects, and the within-mouth kit score effect from (7) is nearly identical to the conditional likelihood estimate from (3).

3. Bias Under Model Misspecification

We assume that, under the true model, the responses Y_{ij} follow a generalized linear mixed model with different between- and within-cluster covariate effects, i.e., conditional on the cluster-specific parameter a_i, the true linear predictor is

$$\eta = a_i + \beta_B \bar{X}_i + \beta_W (X_{ij} - \bar{X}_i) \tag{8}$$

and $a \sim \mathrm{N}(\beta_0, {\sigma_b}^2)$. Suppose, however, that we fit a generalized linear mixed model that assumes common between- and within-cluster covariate effects, i.e., we fit a model with the linear predictor

$$\eta^* = +{a^*}_i + \beta^* X_{ij} \tag{9}$$

and $a^* \sim \mathrm{N}({\beta_0}^*, {\sigma^*{}_b}^2)$. We compare β^* to β_B and β_W.

Scott and Holt (1982) examined the behavior of ordinary and generalized least squares estimators obtained from clustered data, and we can see that their results apply here. They considered the case in which the true model is a linear regression model whose errors have a compound symmetric correlation structure. Such a case arises, e.g., when the true model is of form (6), with $\beta_B \neq \beta_W$ and $n_i = n$. They showed that, if one fits a model of form (4), then the maximum likelihood estimate $\hat{\beta}^*$ will satisfy

$$\hat{\beta}^* = \hat{\beta}_W + \lambda(\rho)(\hat{\beta}_B - \hat{\beta}_W), \tag{10}$$

where $\rho = \mathrm{corr}(Y_{ij}, Y_{ik})$,

$$\lambda(\rho) = \frac{(1-\rho)\hat{\gamma}}{(1-\rho) + n\rho(1-\hat{\gamma})},$$

$\hat{\beta}_B$ and $\hat{\beta}_W$ are maximum likelihood estimates under (6) and

$$\hat{\gamma} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X})(\bar{X}_i - \bar{X})}{\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2},$$

the slope of the regression of $\bar{X}_i$ on X_{ij}. Equation (10) shows that $\hat{\beta}^* - \hat{\beta}_W$ increases as $\hat{\beta}_B - \hat{\beta}_W$ or $\lambda(\rho)$ increases. The factor $\lambda(\rho)$ increases as $\hat{\gamma}$ increases or ρ decreases. If $\bar{X}_i = \bar{X}$, as happens, e.g., with designed within-cluster covariates, then $\hat{\gamma} = 0$ and $\lambda(\rho) = 0$. Thus, for such covariates, $\hat{\beta}^*$ consistently estimates within-cluster covariate effects. For cluster-level covariates, $X_{ij} = \bar{X}_i$ so that $\hat{\gamma} = 1$ and $\lambda(\rho) = 1$. Thus, for such covariates, $\hat{\beta}^*$ consistently estimates between-cluster covariate effects.

We can view (10) as a special case of equation (9) of Neuhaus and Jewell (1993), which evaluates bias due to omitted covariates, and we can cast the problem of incorrectly assuming common between- and within-cluster covariate effects as an omitted covariate problem. To see this, we write the true linear predictor (8) as

$$\eta = a_i + (\beta_B - \beta_W)\bar{X}_i + \beta_W X_{ij} \tag{11}$$

and note that the fitted linear predictor (9) omits the term $(\beta_B - \beta_W)\bar{X}_i$. The work of Neuhaus and Jewell (1993) applies to independent observations and relates the covariate effects measured by the model with omitted covariates to those of the true model. The present problem requires a model for the change induced in the omitted covariate $\bar{X}_i$ when the included covariate X_{ij} increases by one unit, i.e., $\gamma = \mathrm{E}(\bar{X}_i \mid X_{ij} = s+1) - \mathrm{E}(\bar{X}_i \mid X_{ij} = s)$, which we assume is independent of i, j, and s. However, we can easily adapt this approach to incorporate more complicated relationships between $\bar{X}_i$ and X_{ij}. Note that, in the present problem, γ represents the slope of the (linear) regression of $\bar{X}_i$ on X_{ij}. The work of Neuhaus and Jewell (1993) compares the parameters of generalized linear models with linear predictor $\eta^{**} = \beta_0{}^{**} + \beta^{**} X_{ij}$ to those with linear predictor $\eta_M = \delta_0 + \delta_B \bar{X}_i + \delta_W (X_{ij} - \bar{X}_i)$ and shows that

$$\beta^{**} \approx [\delta_W + (\delta_B - \delta_W)\gamma] H_X{}'(0), \tag{12}$$

where $H_X{}'(0) = g'[\mathrm{E}[g^{-1}\{(\delta_B - \delta_W)\bar{X}_i\}]]\mathrm{E}[1/g'[g^{-1}\{(\delta_B - \delta_W)\bar{X}_i\}]]$ and one calculates the expectations in $H_X{}'(0)$ with respect to the conditional distribution of $\bar{X}_i$ given X_{ij}. For link functions such as the logit and probit for which $1/g'(t)$ is concave, we can see that $0 \leq H_X{}'(0) \leq 1$. For the identity and log links, $H_X{}'(0) = 1$.

The results of Neuhaus and Jewell (1993) also show that the covariate effects measured by the marginal models are simple multiples of those measured by mixed-effects models. For example,

$$\delta_W \approx \beta_W H'(0), \tag{13}$$

where $H'(0)$ is of the same form as $H_X{}'(0)$ but one calculates expectations with respect to the random effects a. The approximation that follows assumes that the factors H' that relate the parameters of the marginal and mixed-effects models in the common (9) and separate (8) between- and within-cluster covariate settings are approximately equal. This will be the case if the random effects variances are similar in the two settings. Putting the approximations (13) relating marginal and mixed-effects models into (12) yields the approximation

$$\beta^{*} \approx [\beta_W + (\beta_B - \beta_W)\gamma] H_X{}'(0). \tag{14}$$

Equation (14) shows that $\beta^* - \beta_W$ increases with $\beta_B - \beta_W$ and the association of X_{ij} with $\bar{X}_i$. The factor $H_X{}'(0)$ depends on the conditional variance of $\bar{X}_i$ given X_{ij} and equals one when this conditional variance is zero. The conditional variance is zero for designed within-cluster covariates and cluster-level covariates.

We used the model fits from the two example data sets to examine the quality of the approximation (14). We estimated the slope $\hat{\gamma}$ to be 0.64 for the Georgia birth-weight data using ordinary least squares regression of $\bar{X}_i$ on X_{ij}. The fit of model (6) to the Georgia birth-weight data yielded an intracluster correlation estimate of $\hat{\rho} = 0.40$. Using the between- and within-cluster age effects given in Table 1, (14) yields a predicted age effect of 23.60, the ordinary least squares estimate from the model that assumed common between- and within-cluster effects. Equation (10) yields a predicted age effect of 17.14, which is the maximum likelihood estimate under (4). The estimates from marginal logistic regression models fit to the binary Georgia birth-weight data and the moments of mean age $\bar{X}_i$ yielded $H_X{}'(0) = 0.996$. Using the between- and within-cluster age effects given in Table 2, (14) yields a predicted age effect of -0.50, which corresponds fairly closely to the estimate from the model that assumed common between- and within-cluster effects. We estimated the slope $\hat{\gamma}$ to be 0.44 for the periodontal data using ordinary least squares regression of $\bar{X}_i$ on X_{ij}. Using the birth-weight data approach with the periodontal data yielded $H_X{}'(0) = 0.95$ and a prediction of 0.85 from (14), which closely corresponded to the observed value in Table 3.

4. Discussion

4.1 *Comparison Between Conditional Likelihood and Mixture Models*

Conditional likelihood methods have been suggested as an alternative for estimating the regression coefficients of generalized linear mixed models (see, e.g., Diggle, Liang, and Zeger, 1994, pp. 175–176). It is clear from (2) and (3) that conditional likelihood methods actually measure purely within-cluster covariate effects, i.e., the associations of $(X_{ij} - \bar{X}_i)$ with the response. The example data and approximations of Section 3 show that mixed-effects models that incorrectly assume $\beta_B = \beta_W$ do not measure these effects and that, in practice, mixed-effects and conditional likelihood estimates may not agree. Studies such as randomized block experiments in which the covariate pattern is the same for all clusters typically yield nearly identical mixed-effects and conditional likelihood estimates (Lindsay, Clogg, and Grego, 1991; Neuhaus, Kalbfleisch, and Hauck, 1994).

4.2 *Are Conditional Likelihoods Efficient or Inefficient?*

Investigators often object to conditional likelihood methods, particularly with binary data, because the methods only use data from clusters that are discordant on both the outcome and the covariates. With small clusters such as pairs, the probability of outcome or covariate concordance may be quite large and conditional likelihood methods may not use data from a large proportion of the clusters. Indeed, the asymptotic relative efficiency expressions in TenHave et al. (1995) and Neuhaus and Lesperance (1996) show that with binary data, small clusters, and high intracluster correlation of the covariate X, conditional likelihood estimators may be much less efficient than mixed-model estimators. The comparison of the estimates that assume common between- and within-cluster effects in Tables 2 and 3 to the corresponding conditional likelihood estimates support this. However, the estimates in Tables 2 and 3 that allow different between- and within-cluster effects show that this increased efficiency arises solely through the assumption of common between- and within-cluster effects, an assumption that the data in Tables 2 and 3 do not seem to satisfy. Mixed-effects models can only provide improved efficiency over conditional likelihood estimators when the marginal and conditional likelihoods measure common covariate effects.

4.3 *General Applicability of These Results*

Within-cluster covariates exhibit variability between and within clusters and the data presented here illustrate that the effects of these two components of variability on the response may be different. When between- and within-cluster covariate effects are different, models that assume that these effects are the same do not provide estimates of any substantive interest; the misspecified models measure neither between- nor within-cluster covariate effects. While we have focussed on parametric generalized linear mixed models, the problems associated with incorrectly assuming common between- and within-cluster covariate effects carry over to more general methods such as semiparametric mixture models (Lindsay et al., 1991) and models such as generalized additive models (Hastie and Tibshirani, 1990). As common practice, we recommend that analysts of clustered data examine whether the between- and within-cluster components of covariate variability exhibit common effects on response.

Acknowledgements

We thank Drs Melissa Adams, Paul Gargiullo, and Brian McCarthy of the Centers for Disease Control for providing access to the birth-weight data and two anonymous referees for their helpful comments.

Résumé

Dans les modèles standard de régression utilisés pour l'analyse de données groupées, on postule que les covariables influencent la réponse sans distinguer les effets inter- et intra-groupes. On suppose implicitement que ces effets sont identiques. Les données de l'exemple montrent que ce n'est souvent pas le cas et que les analyses qui ignorent les différences entre effets inter- et intra-groupes peuvent induire en erreur. La prise en compte des effets inter- et intra-groupes des covariables aide aussi à expliquer les différences, observées et théoriques, entre les analyses par modèles de mélanges et celles fondées sur la vraisemblance conditionnelle. Nous montrons en particulier que les méthodes conditionnelles estiment des effets intra-groupes purs alors que les modèles de mélange fournissent des moyennes pondérées des effets inter- et intra-groupes.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88,** 9–25.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Clarendon Press.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Laird, N. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38,** 963–974.

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

Lindsay, B., Clogg, C. C., and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86,** 96–107.

Louis, T. A. (1988). General methods for analyzing repeated measures. *Statistics in Medicine* **7,** 29–45.

Louis, T. A., Robins, J., Dockery, D. W., Spiro, A., and Ware, J. H. (1986). Explaining discrepancies between longitudinal and cross-sectional models. *Journal of Chronic Diseases* **39,** 831–839.

Neuhaus, J. M. and Jewell, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80,** 807–816.

Neuhaus, J. M. and Lesperance, M. L. (1996). Estimation efficiency in a binary mixed-effects model setting. *Biometrika* **83,** 441–446.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59,** 25–35.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched pairs data. *Canadian Journal of Statistics* **22,** 139–148.

Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77,** 848–854.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects model for serial observations with binary response. *Biometrics* **40,** 961–971.

TenHave, T. R., Landis, J. R., and Weaver, S. L. (1995). Association models for periodontal disease progression: A comparison of methods for clustered binary data. *Statistics in Medicine* **14,** 413–429.

Ware, J. H., Dockery, D. W., Louis, T. A., Xu, X., Ferris, B. G., and Speizer, F. (1990). Longitudinal and cross-sectional estimates of pulmonary function decline in never-smoking adults. *American Journal of Epidemiology* **132,** 685–700.

Received March 1996; *revised January* 1997; *accepted June* 1997.