Statistica Sinica **18**(2008), 731-748

BETWEEN- AND WITHIN-CLUSTER COVARIATE EFFECTS AND MODEL MISSPECIFICATION IN THE ANALYSIS OF CLUSTERED DATA

Lei Shen[1], Jun Shao[2], Soomin Park[1] and Mari Palta[2]

[1]*Eli Lilly and Company and* [2]*University of Wisconsin-Madison*

Abstract: We consider the analysis of clustered data using linear mixed effects models and generalized estimating equations, where covariates can be decomposed into between- and within-cluster components. Under the false assumption of equal between- and within-cluster covariate effects, we simultaneously study the asymptotic behavior of the estimators for regression coefficients, intra-cluster correlation and residual error variance. This provides a more complete assessment of the effect of such model misspecification than is currently available in the literature. We then apply the results to gain insights into the effects of confounding and measurement error. Key findings include the structure of bias when both cohort and period effect confounding are present, quantification of the attenuation effect of measurement error, effects of measurement error of some covariates on the estimation of coefficients of error-free covariates, and consistent estimation in the presence of measurement error. The results are extended to allow different cluster sizes, and three longitudinal data sets are used for illustrative purposes.

Key words and phrases: Attenuation, confounding, cohort effects, generalized estimating equations, linear mixed effects models, measurement error, period effects.

1. Introduction

Clustered data arise in many fields and in various forms. Although our work is motivated by longitudinal data from health-related studies, the results in this paper apply equally to, for example, panel data in econometrics, clustered data in clinical trials, and survey data collected by cluster sampling. In these types of data, a number of outcomes together with some covariates are available for each cluster or subject. Covariates can then be decomposed into between- and within-cluster components. For example, consider a survey completed by each subject at multiple time points. When age is considered as a covariate, the average of ages over all time points represents the between-cluster component, while the deviation between the age at a specific time point and the average age constitutes the within-cluster component. In this paper we concentrate on the situation where the outcome variable is continuous and a linear model between the outcome variable and the between- and within-cluster covariates is appropriate.

Quite often, one fails to distinguish between- and within-cluster components of covariates, which is unwise. We demonstrate unequal covariate effects in three data sets. The first concerns the effect of aging on body mass index. The second arose from a study of formaldehyde emission in mobile homes. The third study investigates the relationship between perceived sleepiness and variables including sleep latency, the length of time it takes a person to fall asleep at night. These data sets are analyzed in Section 5 to illustrate the usefulness of findings in this paper.

Assuming equal between- and within-cluster covariate effects leads to estimates which are misleading when the assumption is wrong. Scott and Holt (1982) obtained important results in a sample survey setting. They showed that the generalized least squares estimate of the regression coefficient, under the erroneous assumption, is a weighted average of the ordinary least squares estimates from the cross-sectional and within-cluster regressions. Neuhaus and Kalbfleisch (1998) derived extensions for linear and generalized linear mixed models. Historically, an equivalent problem was considered by Wishart (1938) who formulated it as dependence of the distribution of the random subject effects on covariates, and it was further addressed by Maddala (1971), Mundlak (1978) and Hausman (1978). These authors showed that such dependence shows up as unequal covariate effects.

In practice, intra-cluster correlation is typically estimated, rather than known or fixed as assumed by Scott and Holt (1982). We often fit linear mixed effects models to longitudinal data with continuous outcomes (Laird and Ware (1982)) or use generalized estimating equations (Liang and Zeger (1986)). The intra-cluster correlation is then estimated simultaneously with the regression coefficients and the residual error variance. In Section 2, we investigate the estimators under the false assumption of equal covariate effects. By simultaneously studying these estimators, we are able to better evaluate the consequence of wrongly assuming equal covariate effects. It is shown that the regression coefficient estimator under the misspecified model converges to a linear combination of two covariate effects, with coefficients related to the estimated, rather than true, intra-cluster correlation. The estimated residual error variance under the misspecified model over-estimates the true residual error variance. Important extensions are made to allow multiple covariates and unequal cluster sizes, which commonly occur in practice. Based on our results for multiple covariates, one can identify situations where different between- and within-cluster effects of some covariates not only affect the estimates for the effects of these covariates, but also cause biases in the estimation of effects of other covariates that have equal between- and within-cluster effects.

Even if one correctly specifies a model with different covariate effects, interpretation of the between- and within-cluster coefficients is often difficult without

explaining how the difference may have arisen. Thus, it is important to investigate possible causes. Louis et al. (1986) showed that different covariate effects can result from non-linearity. Ware et al. (1990) proposed cohort and period effects as possible causes. Palta and Yao (1991) formulated a formal model for these types of confounding and showed that they lead to discrepancy between the two covariate effects. Chao, Palta and Young (1997) derived similar results for binary data. Using the results in Section 2, we study in Section 3 the effect of confounding under a model which extends that of Palta and Yao (1991). Some results on the structure of bias in parameter estimation are derived. In Section 4, we extend some results in the literature on measurement error in linear regression, such as attenuation caused by measurement error and the impact of measurement error of one covariate on the estimation of effects of other covariates. The relationship between the amount of bias in the slope estimator and cluster size, covariate structure and measurement error variance is investigated. Furthermore, we are able to identify some situations where consistent estimators can be derived for parameters of interest without requiring supplemental data.

Our study sheds light on the nature of different covariate effects, and on the bias in parameter estimation when omitted confounders or measurement error are not properly taken into account in the analysis of clustered data. Some of our findings, such as parameter estimation in the presence of measurement error without supplemental data and the relationship between the bias of estimators and the cluster size, choice of working correlation in generalized estimating equations, and the structure of the covariates, have important implications for the design and analysis of studies with clustered data.

2. Estimators under the Misspecified Model

We consider a clustered data set with n subjects (clusters). For the ith subject, we observe outcomes Y_{ij} and a p dimensional covariate vector $\boldsymbol{X}_{ij}$, $j = 1, \ldots, k_i$, where k_i is the cluster size and $i = 1, \ldots, n$. We first focus on the balanced case of $k_i = k$ for all i. Postulate the model

$$\boldsymbol{Y}_i = \alpha^* \mathbf{1} + \boldsymbol{X}_i \boldsymbol{\beta}^* + \boldsymbol{e}_i^* = \tilde{\boldsymbol{X}}_i \boldsymbol{\gamma}^* + \boldsymbol{e}_i^*, \tag{1}$$

where $\boldsymbol{Y}_i$ is the k-vector whose jth component is Y_{ij}, $\mathbf{1}$ is the k-vector of ones, $\boldsymbol{X}_i$ is the $k \times p$ matrix whose jth row is $\boldsymbol{X}_{ij}$, α^* is an unknown parameter, $\boldsymbol{\beta}^*$ is a p-vector of unknown parameters, $\tilde{\boldsymbol{X}}_i = (\mathbf{1}, \boldsymbol{X}_i)$, $\boldsymbol{\gamma}^* = (\alpha^*, \boldsymbol{\beta}^{*'})'$, the $\boldsymbol{e}_i^*$ are independent and identically distributed with mean zero and $k \times k$ covariance matrix $\sigma_{e^*}^2 \boldsymbol{V}_{\rho^*}$, $\boldsymbol{V}_a$ denotes the matrix consisting of 1 for every diagonal entry and a for every off diagonal entry, and ρ^* and $\sigma_{e^*}^2$ are the unknown intra-cluster correlation and the residual error variance, respectively. In the linear mixed effects model, the $\boldsymbol{e}_i^*$ are assumed to be normally distributed, and parameters $\boldsymbol{\gamma}^*$, ρ^* and $\sigma_{e^*}^2$

can be estimated by maximum likelihood (Lindstrom and Bates (1988)). These estimates coincide with solutions to the following generalized estimating equations (Liang and Zeger (1986)) with a compound-symmetry working correlation matrix:

$$\boldsymbol{\gamma}^* = \Big(\sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i{}' \boldsymbol{V}_{\rho^*}^{-1} \tilde{\boldsymbol{X}}_i \Big)^{-1} \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i{}' \boldsymbol{V}_{\rho^*}^{-1} \boldsymbol{Y}_i \tag{2}$$

$$\sigma_{e^*}^2 = \sum_{i=1}^{n} \frac{(\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\gamma}^*)' \boldsymbol{V}_{\rho^*}^{-1} (\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\gamma}^*)}{nk} \tag{3}$$

$$[1 + (k-1)\rho^*]\sigma_{e^*}^2 = \sum_{i=1}^{n} \frac{(\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\gamma}^*)' \mathbf{1}\mathbf{1}' (\boldsymbol{Y}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\gamma}^*)}{nk}. \tag{4}$$

Now, assume that the true underlying data-generating model is not (1), but

$$\boldsymbol{Y}_i = \alpha \mathbf{1} + (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i)\boldsymbol{\beta}_{wc} + \bar{\boldsymbol{X}}_i \boldsymbol{\beta}_{bc} + \boldsymbol{e}_i, \tag{5}$$

where $\bar{\boldsymbol{X}}_i = (\bar{X}_{1i}\mathbf{1}, \ldots, \bar{X}_{pi}\mathbf{1})'$, $\bar{X}_{ai}$ is the mean of the ath covariate over the ith cluster for $a = 1, \ldots, p$, $\boldsymbol{\beta}_{bc}$ and $\boldsymbol{\beta}_{wc}$ correspond to the between- and within-cluster covariate effects, respectively, and the $\boldsymbol{e}_i$ are independent and identically distributed with mean zero and $k \times k$ covariance matrix $\sigma_e^2 \boldsymbol{V}_\rho$.

The asymptotic properties of the estimators under the misspecified Model (1) can be obtained by studying equations (2)-(4). Since the number of observations for each subject is typically not large in a longitudinal study and the number of subjects can be large, we focus on the asymptotics when $n \to \infty$ and k is fixed. The proof of the following result can be found in the Appendix given in the webside www.stat.wisc.edu/~shao.

Theorem 1. *Assume that $\{\boldsymbol{Y}_i, \boldsymbol{X}_i\}$ are independent and identically distributed with finite second moments. The solutions of the equations (2)−(4), denoted by $\hat{\alpha}^*$, $\hat{\boldsymbol{\beta}}^*$, $\hat{\rho}^*$, and $\hat{\sigma}_{e^*}^2$, converge a.s. to some values α_∞^*, $\boldsymbol{\beta}_\infty^*$, ρ_∞^* and $\sigma_{e^*,\infty}^2$, respectively. There exists a $p \times p$ matrix $\boldsymbol{\Lambda}$ such that*

$$\boldsymbol{\beta}_\infty^* = (\boldsymbol{I} - \boldsymbol{\Lambda})\boldsymbol{\beta}_{wc} + \boldsymbol{\Lambda}\boldsymbol{\beta}_{bc}, \tag{6}$$

where $\boldsymbol{I}$ is the identity matrix of order p. Moreover, $\sigma_{e^*,\infty}^2 \geq \sigma_e^2$.

Hence, $\sigma_{e^*,\infty}^2$ is an inflated estimator of the residual error variance. The estimator of $\boldsymbol{\beta}^*$ converges to a linear combination of $\boldsymbol{\beta}_{bc}$ and $\boldsymbol{\beta}_{wc}$. In many cases, the parameter of interest is not any of $\boldsymbol{\beta}_{bc}$, $\boldsymbol{\beta}_{wc}$ and $\boldsymbol{\beta}_\infty^*$. This will be further explored in Sections 3-4.

The form of $\boldsymbol{\Lambda}$ and more results are given for the following specific situations.

Case 1. Single Covariate

Consider the situation where there is a single covariate ($p = 1$). That is, $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{ik})'$. Often, the covariates within clusters are not independent. Rather, cluster means of the covariate can be assumed to follow a certain distribution while a specific covariate value deviates from the corresponding cluster mean. Therefore, we consider the covariate as the sum of between- and within-cluster components. Furthermore, in longitudinal studies, the covariates are often time-dependent. Thus, we consider the following model for the covariate:

$$x_{ij} = \xi_{x,j} + m_{x_i} + \epsilon_{x_{ij}}, \tag{7}$$

where $(\xi_{x,1}, \ldots, \xi_{x,k})'$ is a vector of parameters, $\{m_{x_i}\}$ and $\{\epsilon_{x_{ij}}\}$ are independent of each other and independently and identically distributed with means zero and variances $\sigma^2_{m_x}$ and $\sigma^2_{\epsilon_x}$, respectively. We denote $\sigma^2_{m_x}/(\sigma^2_{m_x}+\sigma^2_{\epsilon_x})$ by ρ_x, $\sum_j \xi_{x,j}/k$ by $\bar{\xi}_x$, and $\sum_j (\xi_{x,j} - \bar{\xi}_x)^2/(k\sigma^2_{m_x} + k\sigma^2_{\epsilon_x})$ by S^2_ξ. Then, the result in Theorem 1 applies with

$$\alpha^*_\infty = \alpha + (1-\lambda)(\beta_{bc} - \beta_{wc})\bar{\xi}_x \tag{8}$$

$$\beta^*_\infty = (1-\lambda)\beta_{wc} + \lambda\beta_{bc}, \tag{9}$$

where

$$\lambda = \lambda(\rho^*_\infty), \quad \lambda(\rho) = \frac{(1-\rho)\{1+(k-1)\rho_x\}k^{-1}}{1+\{k-2-(k-1)\rho_x\}\rho + \{1+(k-1)\rho\}S^2_\xi} \tag{10}$$

is between 0 and 1 (see the derivations in the Appendix). Moreover,

$$\rho^*_\infty \geq \rho \quad \text{if} \quad \rho_x \leq 1 - \frac{(1-\rho) - (k-1)\{1+(k-1)\rho\}S^2_\xi}{(k-1)\{1+(k-2)\rho\}}. \tag{11}$$

These results indicate that the weights used to form the weighted average of β_{bc} and β_{wc} in (6) depend on the cluster size, the estimated intra-cluster correlation, and the nature of the covariate. Figures 1−2 illustrate the relationship between the estimated regression coefficients and these parameters. Both plots are generated under $\xi_{x,j} = \bar{\xi}_x$ for all j, where $\delta = (\beta_{wc} - \beta_{bc})^2(\sigma^2_{m_x} + \sigma^2_{\epsilon_x})/\sigma^2_e$ is a measure of the discrepancy between β_{wc} and β_{bc}. In general, larger cluster size, higher intra-cluster correlation, and lower correlation among the covariate values correspond to more longitudinal information and less cross-sectional information. Thus, it is not surprising that the estimated covariate effect under the misspecified model is closer to the within-cluster covariate effect β_{wc} when k is larger, when ρ_x is smaller (Figure 1), and when ρ is larger (Figure 2).

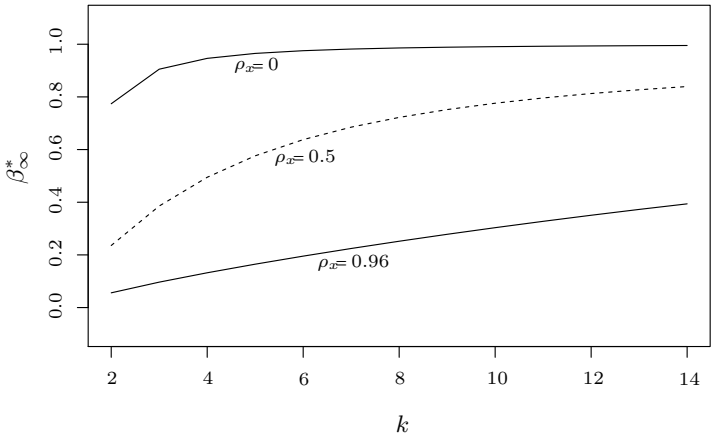

Figure 1. β^*_∞ in (9), as a function of cluster size k or ρ_x, when $\beta_{wc} = 1$, $\beta_{bc} = 0$, $\rho = 0.5$, and $\delta = 1$.

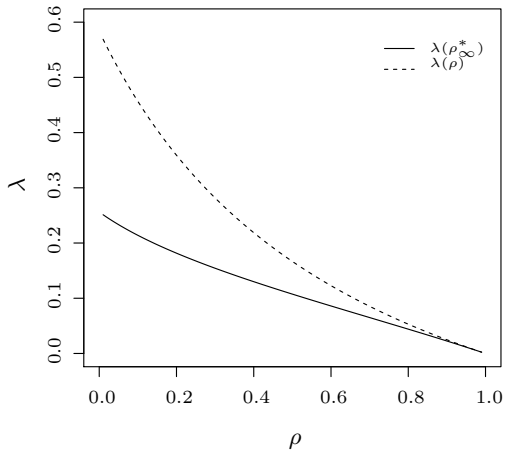

Figure 2. $\lambda(\rho^*_\infty)$ and $\lambda(\rho)$ in (10) when $k = 5$, $\rho_x = 0.5$ and $\delta = 1.7$.

The result also indicates that the estimation of the intercept parameter α is affected by the different between- and within-cluster covariate effects, unless the overall mean of the covariate is zero.

Scott and Holt (1982) considered a similar problem and derived an asymptotic limit of the generalized least squares estimator $\hat{\beta}^*$, which is given by (9) with λ replaced by $\lambda(\rho)$, assuming that the correlation ρ is pre-determined. Neuhaus and Kalbfleisch (1998) directly applied this result to maximum likelihood estimates, which can also be interpreted as generalized least squares estimates, as indicated by the score equation (2). However, we note that in the

latter case, the correlation used in forming the generalized least squares is $\hat{\rho}^*$, the estimate of ρ^*. This correlation has complex behavior since it is obtained under the misspecified model. Figure 2 illustrates the difference between $\lambda(\rho)$ and $\lambda(\rho^*_\infty)$, which affects the calculation of β^*_∞ according to (9). The formula of Scott and Holt (1982) leads to $\lambda(\rho)$ corresponding to the higher curve, generally about twice as large as $\lambda = \lambda(\rho^*_\infty)$ corresponding to the lower curve.

When GEE are used to estimate the regression parameters, we have the option of fixing the intra-cluster correlation at a certain value. Formula (10) indicates that λ monotonously decreases as ρ^*_∞ increases. Therefore, choosing a high correlation when applying GEE leads to a β^*_∞ close to the within-cluster slope β_{wc}, while the choice of a lower correlation results in a β^*_∞ close to the between-cluster slope β_{bc}.

Furthermore, (11) implies that the intra-cluster correlation estimated under the wrong model is higher than in the true model if the covariate structure is more longitudinal than cross-sectional, that is, ρ_x is small or S^2_ξ is large. In fact, (11) is guaranteed to hold if $\rho_x \le (k-2)/(k-1)$ or $S^2_\xi \ge (k-1)(1-\rho)/\{1+(k-1)\rho\}$.

Case 2. Multiple Covariates

Consider a multiple covariates following

$$\boldsymbol{X}_{ij} = \boldsymbol{\mu}_x + \boldsymbol{m}_{x_i} + \boldsymbol{\epsilon}_{x_{ij}}, \tag{12}$$

where $\boldsymbol{\mu}_x$ is a p-vector of parameters, $\{\boldsymbol{m}_{x_i}\}$ are independent and identically distributed with mean zero and $p \times p$ covariance matrix $\boldsymbol{\Sigma}_{m_x}$, $\{\boldsymbol{\epsilon}_{x_{ij}}\}$ are independent and identically distributed with mean zero and $p \times p$ covariance matrix $\boldsymbol{\Sigma}_{\epsilon_x}$, and $\{\boldsymbol{m}_{x_i}\}$ are independent of $\{\boldsymbol{\epsilon}_{x_{ij}}\}$. Under (12), Theorem 1 holds with

$$\boldsymbol{\Lambda} = \left(k\boldsymbol{\Sigma}_{m_x} + k\Big[\frac{1+(k-2)\rho^*_\infty}{1-\rho^*_\infty}\Big]\boldsymbol{\Sigma}_{\epsilon_x} \right)^{-1} (k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{\epsilon_x}), \tag{13}$$

$$\alpha^*_\infty = \alpha + \boldsymbol{\mu}'_x(\boldsymbol{I} - \boldsymbol{\Lambda})(\boldsymbol{\beta}_{bc} - \boldsymbol{\beta}_{wc}) \tag{14}$$

(see the Appendix). The structure of the matrix $\boldsymbol{\Lambda}$ is important in applications. When there are p covariates, some of them may have different between- and within-cluster covariate effects, whereas the others do not. If $\boldsymbol{\Lambda}$ is diagonal, then any difference between the two effects of one covariate does not affect the consistency of the estimator of any other regression parameter under the misspecified model. This can be illustrated more precisely as follows. Assume that the between- and within-cluster covariate effects are equal for the last $p-r$ covariates but not for the first r covariates. We partition $\boldsymbol{\beta}^*_\infty$ into the first r elements and the last $(p-r)$ so that $\boldsymbol{\beta}^*_\infty = (\boldsymbol{\beta}^{*'}_{\infty,1}, \boldsymbol{\beta}^{*'}_{\infty,2})'$, and similarly $\boldsymbol{\beta}_{bc} = (\boldsymbol{\beta}'_{bc,1}, \boldsymbol{\beta}'_{bc,2})'$, $\boldsymbol{\beta}_{wc} = (\boldsymbol{\beta}'_{wc,1}, \boldsymbol{\beta}'_{wc,2})'$, and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix},$$

where $\mathbf{\Lambda}_{11}$ is a $r \times r$ matrix and $\mathbf{\Lambda}_{22}$ is a $(p-r) \times (p-r)$ matrix. From (6) and the assumption that $\boldsymbol{\beta}_{bc,2} = \boldsymbol{\beta}_{wc,2} = \boldsymbol{\beta}_2$, we have

$$\boldsymbol{\beta}^*_{\infty,1} = (\boldsymbol{I} - \mathbf{\Lambda}_{11})\boldsymbol{\beta}_{wc,1} + \mathbf{\Lambda}_{11}\boldsymbol{\beta}_{bc,1} \tag{15}$$

$$\boldsymbol{\beta}^*_{\infty,2} = \boldsymbol{\beta}_2 + \mathbf{\Lambda}_{21}(\boldsymbol{\beta}_{bc,1} - \boldsymbol{\beta}_{wc,1}). \tag{16}$$

Equation (15) indicates that asymptotically, estimators for the first r regression parameters estimate a linear combination of the corresponding between- and within-cluster covariate effects, while (16) implies that estimators for the remaining $(p-r)$ regression parameters may still be biased even though the between- and within-cluster covariate effects for these covariates are equal ($\boldsymbol{\beta}_{bc,2} = \boldsymbol{\beta}_{wc,2}$), unless $\mathbf{\Lambda}_{21} = \mathbf{0}$.

The following can be shown for the matrix $\mathbf{\Lambda}$ given by (13):

- $\mathbf{\Lambda}$ is a diagonal matrix if both $\mathbf{\Sigma}_{m_x}$ and $\mathbf{\Sigma}_{e_x}$ are diagonal. That is, the p covariates are uncorrelated with each other.
- $\mathbf{\Lambda}$ is a diagonal matrix if $\mathbf{\Sigma}_{m_x}$ and $\mathbf{\Sigma}_{e_x}$ differ only by a multiplicative constant. That is, the correlation structures of the between-cluster component and the within-cluster component of the covariates are the same.
- $\mathbf{\Lambda}$ is a diagonal matrix if either $\mathbf{\Sigma}_{m_x}$ or $\mathbf{\Sigma}_{e_x}$ is $\mathbf{0}$, i.e., the covariates vary only within or between the clusters.
- $\mathbf{\Lambda}_{21}$ is $\mathbf{0}$ when the last $(p-r)$ covariates are uncorrelated with the first r covariates.

Finally, we extend Theorem 1 to the case of unequal cluster sizes k_i with k_i ranging from 2 to some integer L, which can be due to an imbalanced design or missing data. We assume that as $n \to \infty$, the proportion of clusters with l observations converges to a fixed constant p_l for $l = 2, \ldots, L$. We consider the following underlying model for the data:

$$\boldsymbol{Y}_i = \alpha_{k_i}\mathbf{1} + (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i)\boldsymbol{\beta}_{wc,k_i} + \bar{\boldsymbol{X}}_i\boldsymbol{\beta}_{bc,k_i} + \boldsymbol{e}_i. \tag{17}$$

Note that we allow parameters α, $\boldsymbol{\beta}_{bc}$ and $\boldsymbol{\beta}_{wc}$ to depend on the cluster size. Model (17) is natural when the difference in cluster sizes is due to an imbalanced design. When the difference in cluster sizes is caused by missing data in responses, (17) can be derived from (5) under the assumptions that the probabilities of missing responses depend only on the covariates and that, given the covariates associated with observed responses, the conditional expected values of covariates associated with missing responses are linear functions of covariates associated with observed responses (see the Appendix). Under Model (17), the result in Theorem 1 still holds (see the Appendix) with

$$\begin{pmatrix} \alpha^*_\infty \\ \boldsymbol{\beta}^*_\infty \end{pmatrix} = \sum_{l=2}^{L} \boldsymbol{P}_l \left\{ (\boldsymbol{I} - \mathbf{\Lambda}_l) \begin{pmatrix} 0 \\ \boldsymbol{\beta}_{wc,l} \end{pmatrix} + \mathbf{\Lambda}_l \begin{pmatrix} \alpha_l \\ \boldsymbol{\beta}_{bc,l} \end{pmatrix} \right\}, \tag{18}$$

where $\boldsymbol{P}_2, \ldots, \boldsymbol{P}_L$ are $p \times p$ positive definite matrices whose sum is the identity matrix and $\boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_L$ are some $p \times p$ matrices. In particular, when $\alpha_l = \alpha$, $\boldsymbol{\beta}_{wc,l} = \boldsymbol{\beta}_{wc}$, $\boldsymbol{\beta}_{bc,l} = \boldsymbol{\beta}_{bc}$ for all l, and Model (12) holds, (6) and (14) hold with

$$\boldsymbol{\Lambda} = \left[\sum_{l=2}^{L} p_l \left\{ \frac{l\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x}}{1+(l-1)\rho^*_\infty} + \frac{(l-1)\boldsymbol{\Sigma}_{e_x}}{1-\rho^*_\infty} \right\}\right]^{-1} \sum_{l=2}^{L} p_l \frac{l\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x}}{1+(l-1)\rho^*_\infty}.$$

3. Omitted Confounders

Having derived results for how a difference in between- and within-cluster effects affect parameter estimates, we now investigate how such difference may arise from a subject matter point of view to aid in model interpretation. A confounder is a factor, the control of which changes the relationship between the primary factor under study and the outcome Rothman and Greenland (1998, p.59). In multiple regression this occurs when the factor is correlated with both the outcome and covariates under investigation. The investigation and control of confounding is a major emphasis in the analysis of observational studies.

The nature of clustered data allows the relationship between an omitted confounder and the covariate of interest to be different within versus across clusters. Well-known examples of such confounders are the so-called cohort and period effects in longitudinal studies of aging. The former arise as cross-sectional analyses estimate a combination of the effects of true aging and covariates associated with birth cohort. The latter arise in longitudinal data collection as the progression of time leads to changes in measurement technique or general health status of the population.

Let $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ be vectors of values of a measured covariate and an omitted confounder for the ith subject, respectively. Conditional on both $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, the vector of outcomes $\boldsymbol{Y}_i$ satisfies a multivariate normal regression model:

$$\boldsymbol{Y}_i | \boldsymbol{x}_i, \boldsymbol{z}_i \sim N(\beta_0 \mathbf{1} + \beta_x \boldsymbol{x}_i + \beta_z \boldsymbol{z}_i, \sigma^2_{\tilde{e}} \boldsymbol{V}_{\tilde{\rho}}). \tag{19}$$

We assume that the covariate and confounder are sums of two components:

$$\boldsymbol{x}_i = \boldsymbol{\xi}_x + m_{x_i}\mathbf{1} + \boldsymbol{e}_{x_i} \tag{20}$$

$$\boldsymbol{z}_i = m_{z_i}\mathbf{1} + \boldsymbol{e}_{z_i}, \tag{21}$$

where $\boldsymbol{\xi}_x$ is a vector of parameters, $\{(m_{x_i}, m_{z_i})\}$ are independent and normally distributed with mean zero and 2×2 covariance matrix

$$\sigma^2_{m_x} \begin{pmatrix} 1 & r_m\sqrt{c_m} \\ r_m\sqrt{c_m} & c_m \end{pmatrix},$$

$\{\boldsymbol{e}_{x_i}\}$ are independent and normally distributed with mean zero and covariance matrix $\sigma^2_{e_x}\boldsymbol{I}$, $\{\boldsymbol{e}_{z_i}\}$ are independent and normally distributed, conditional on $\{\boldsymbol{e}_{x_i}\}$, with mean vectors $\{\gamma_0\mathbf{1} + \gamma\boldsymbol{e}_{x_i}\}$ and covariance matrix $\sigma^2_{e_z}\boldsymbol{I}$, and $\{(m_{x_i}, m_{z_i})\}$ are independent of $\{(\boldsymbol{e}_{x_i}, \boldsymbol{e}_{z_i})\}$. By allowing systematic time dependence, (20)-(21) extend the model of Palta and Yao (1991). The $\{m_{z_i}\}$ and $\{\boldsymbol{e}_{z_i}\}$, correspond to cohort effects and period effects, respectively.

Using properties of the multivariate normal distribution, we can derive (see the Appendix)

$$\boldsymbol{Y}_i|\boldsymbol{x}_i \sim N(\alpha\mathbf{1} + \beta_{wc}(\boldsymbol{x}_i - \bar{x}_i\mathbf{1}) + \beta_{bc}\bar{x}_i\mathbf{1}, \sigma^2_e\boldsymbol{V}_\rho), \tag{22}$$

where $\beta_{wc} = \beta_x + \gamma\beta_z$, $\beta_{bc} = \beta_x + \eta_x r_m\sqrt{c_m}\beta_z + (1-\eta_x)\gamma\beta_z$, and α, σ^2_e, ρ are functions of β_0, β_x, β_z, $\sigma^2_{\tilde{e}}$, $\tilde{\rho}$, $\sigma^2_{m_x}$, r_m, c_m, $\bar{\xi}_x$, γ_0, γ, $\sigma^2_{e_x}$, $\sigma^2_{e_z}$ and k. Here $\eta_x = \sigma^2_{m_x}/(\sigma^2_{m_x} + \sigma^2_{e_x}/k)$ indicates the structure of the covariate.

As shown by (22), omitted confounders potentially lead to different between- and within-cluster covariate effects in the marginal model relating the outcome to the measured covariate. Assuming equal between- and within-cluster covariate effects in this case is the same as ignoring the omitted confounder. Thus, we can apply the results in Section 2 to investigate the effects of cohort and period effects. Under the above model, we have the following conclusions on $\hat{\beta}^*$, the estimator for the regression coefficient under the assumption of equal covariate effects:

- When only a cohort effect is present ($\gamma = 0$), β_{wc} is equal to β_x, whereas β_{bc} is not. The bias of $\hat{\beta}^*$ is $\lambda\eta_x r_m\sqrt{c_m}\beta_z$, which can be reduced if the study design is more longitudinal (for example larger S^2_ξ).
- When only a period effect is present ($r_m = 0$), β_{wc} differs from β_x more than β_{bc} does. The bias of $\hat{\beta}^*$ is $(1-\lambda\eta_x)\gamma\beta_z$, which can be reduced if the study design is more cross-sectional (for example smaller S^2_ξ).
- When the between- and within-cluster correlations between the covariate and the confounder are both positive (negative), $\hat{\beta}^*$ over-estimates (under-estimates) β_x.
- In general, the bias of $\hat{\beta}^*$ as an estimator of β_x is $\{(1-\phi)\gamma + \phi r_m\sqrt{c_m}\}\beta_z$, where $\phi = \lambda\eta_x \in [0,1]$.
- The estimator of $\sigma^2_{\tilde{e}^*}$ is an inflated estimator of the original residual error variance $\sigma^2_{\tilde{e}}$.

4. Measurement Error

It is well-known that measurement error in covariates leads to biased estimators of regression parameters. Measurement error is often assumed to be additive and independent of the true covariate. Then, the regression coefficient estimator

is known to be subject to attenuation in linear regression. In this section, we study the effects of measurement error on the analysis of clustered data. Since between- and within-cluster effects of the error-prone covariate are shown to be different, results in Section 2 are useful. Our results in this section are extensions of those of Wang et al. (1998) to multiple covariates.

Let $\boldsymbol{T}_{ij}$ be the p-vector of covariate values associated with Y_{ij}, and $\boldsymbol{T}_i$ be the $k \times p$ matrix whose jth row is $\boldsymbol{T}_{ij}$, $j = 1, \ldots, k$, $i = 1, \ldots, n$. We consider the normal regression model

$$\boldsymbol{Y}_i|\boldsymbol{T}_i \sim N(\tilde{\alpha}\mathbf{1} + \boldsymbol{T}_i\tilde{\boldsymbol{\beta}}, V(\boldsymbol{Y}_i|\boldsymbol{T}_i)), \tag{23}$$

where $V(\boldsymbol{Y}_i|\boldsymbol{T}_i)$ is the conditional covariance matrix of $\boldsymbol{Y}_i$ given $\boldsymbol{T}_i$. The covariate $\boldsymbol{T}_{ij}$ is assumed to follow Model (12).

Instead of $\boldsymbol{T}_{ij}$, we observe an error-prone covariate

$$\boldsymbol{X}_{ij} = \boldsymbol{T}_{ij} + \boldsymbol{U}_{ij}, \tag{24}$$

where $\{\boldsymbol{U}_{ij}\}$ are independent and normally distributed with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}_u$, and are independent of $\{\boldsymbol{T}_{ij}\}$ and $\{Y_{ij}\}$. This model is often referred to as the classical measurement error model. As in the situation of confounding, we can derive the model that relates the outcome to the observed covariate $\boldsymbol{X}_{ij}$ (derivations are in the Appendix), which involves different covariate effects:

$$\boldsymbol{Y}_i|\boldsymbol{X}_i \sim N(\alpha\mathbf{1} + (\boldsymbol{X}_i - \bar{\boldsymbol{X}}_i)\boldsymbol{\beta}_{wc} + \bar{\boldsymbol{X}}_i\boldsymbol{\beta}_{bc}, V(\boldsymbol{Y}_i|\boldsymbol{T}_i) + \boldsymbol{\Sigma}_\Delta),$$

where

$$\alpha = \tilde{\alpha} + \tilde{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_u(k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{\mu}_x, \tag{25}$$

$$\boldsymbol{\beta}_{wc} = (\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{\Sigma}_{e_x}\tilde{\boldsymbol{\beta}}, \tag{26}$$

$$\boldsymbol{\beta}_{bc} = (k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}(k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x})\tilde{\boldsymbol{\beta}}, \tag{27}$$

and $\boldsymbol{\Sigma}_\Delta$ is a compound symmetric matrix with diagonal entries

$$\tilde{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_u(k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}[\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_{m_x}(\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}(k\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)]\tilde{\boldsymbol{\beta}}$$

and off-diagonal entries

$$\tilde{\boldsymbol{\beta}}'\boldsymbol{\Sigma}_u(k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{\Sigma}_{m_x}(\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)^{-1}\boldsymbol{\Sigma}_u\tilde{\boldsymbol{\beta}}.$$

In this case, assuming equal between- and within-cluster covariate effects is the same as ignoring measurement error.

In multiple linear regression, measurement error following model (24) causes attenuation in parameter estimation. Applying the results in Section 2 shows

attenuation by measurement error also in the analysis of clustered data. In this case, (6) holds with

$$\boldsymbol{\Lambda} = [k\boldsymbol{\Sigma}_{m_x} + k\{1+(k-2)\rho_\infty\}(1-\rho_\infty)^{-1}(\boldsymbol{\Sigma}_{e_x}+\boldsymbol{\Sigma}_u)]^{-1}(k\boldsymbol{\Sigma}_{m_x}+\boldsymbol{\Sigma}_{e_x}+\boldsymbol{\Sigma}_u),$$

where ρ_∞ is the asymptotic limit of the estimated working correlation. Using Theorem 1 and (25)-(27), we can show that the estimators of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$ with measurement error ignored, i.e., those assuming equal between- and within-cluster covariate effects, converge a.s. to

$$\boldsymbol{\Lambda}^*\tilde{\boldsymbol{\beta}} \quad \text{and} \quad \tilde{\alpha} - \tilde{\boldsymbol{\beta}}'(\boldsymbol{I} - \boldsymbol{\Lambda}^*)\boldsymbol{\mu}_x, \tag{28}$$

respectively, where

$$\boldsymbol{\Lambda}^* = \boldsymbol{I} - [\boldsymbol{\Sigma}_{m_x}\{1 + (k-2)\rho_\infty\}^{-1}(1-\rho_\infty) + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u]^{-1}\boldsymbol{\Sigma}_u. \tag{29}$$

From the form of $\boldsymbol{\Lambda}^*$ in (29) we conclude that, when there are multiple covariates, one of which is measured with error, in general the regression coefficients of other covariates are also biased, unless the covariate with measurement error is uncorrelated with other covariates.

If $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$ are the parameters of interest, then consistent estimators can be derived using (25)-(29). If we have supplemental data (e.g., validation or replicate data) to estimate $\boldsymbol{\Sigma}_u$, which is necessary for cross-sectional data problems, then consistent estimators of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$ can be obtained using (28)-(29) with $\boldsymbol{\Sigma}_{m_x}$ and $\boldsymbol{\Sigma}_{e_x}$ estimated by the between- and within-cluster variabilities of the observed $\{\boldsymbol{X}_{ij}\}$. For clustered data, supplemental data are not necessary, since data within each cluster can be regarded as partial replicates. It follows from (25)-(27) that

$$\tilde{\boldsymbol{\beta}} = (k\boldsymbol{\Sigma}_{m_x})^{-1}[(k\boldsymbol{\Sigma}_{m_x} + \boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)\boldsymbol{\beta}_{bc} - (\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u)\boldsymbol{\beta}_{wc}], \tag{30}$$

$$\tilde{\alpha} = \alpha + \boldsymbol{\mu}_x'(\boldsymbol{\beta}_{bc} - \tilde{\boldsymbol{\beta}}). \tag{31}$$

Consistent estimators of α, $\boldsymbol{\beta}_{bc}$, and $\boldsymbol{\beta}_{wc}$ can be obtained by fitting Model (5). Although each of $\boldsymbol{\Sigma}_{e_x}$ and $\boldsymbol{\Sigma}_u$ may not be estimable, the sum $\boldsymbol{\Sigma}_{e_x} + \boldsymbol{\Sigma}_u$ has a consistent estimator $\boldsymbol{S}_{wc} = \sum_i \sum_j (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)(\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)'/n(k-1)$. Also, a consistent estimator of $\boldsymbol{\mu}_x$ is $\bar{\boldsymbol{X}}$, the average of all $\boldsymbol{X}_{ij}$, and a consistent estimator of $\boldsymbol{\Sigma}_{m_x}$ is $\boldsymbol{S}_{bc} - \boldsymbol{S}_{wc}/k$, where $\boldsymbol{S}_{bc} = \sum_i (\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{X}})'/n$. Therefore, under the above assumptions on covariate structure and measurement error, consistent estimators of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}$ can be obtained by using (30)-(31) with unknown parameters replaced by the previously constructed consistent estimators. Because this method first decomposes covariates into between- and within-cluster components, and then recombines estimated regression parameters according to (30), we refer to it as the Decomposition and Recombination (DeAR) method.

Table 1. Estimates of regression parameters in Example 1. The response variable is the logarithm of BMI and the predictor variable is age in decades. The numbers in the parentheses are standard errors.

	Model (1)	Model (5)	
Covariate		within	between
Intercept	3.00	3.16	
Age	0.065 (0.003)	0.088 (0.003)	0.029 (0.004)
$(\text{Age} - \mu_{\text{Age}})^2$	-0.01	-0.01	
correlation	0.928	0.927	

Tosteson, Buonaccorsi and Demidenko (1998) and Buonaccorsi, Demidenko, and Tosteson (2000) considered the similar problem of covariate measurement error for a linear mixed effects model. For the situation of measurement error in a single variable, they derived a consistent estimator for the regression coefficients without supplemental data by taking advantage of an assumption on the variance structure of the covariates.

5. Examples

Example 1 (Body Mass Index and Aging). We consider data from the Wisconsin Sleep Cohort Study (Young et al. (1993)). A total of 3211 employees age 30-60 at four State of Wisconsin agencies filled out a survey twice, approximately four years apart. We are interested in changes in body mass index (weight in kilograms divided by the square of height in meters) with age. The logarithm of body mass index (BMI) shows an increasing trend with age, and it appears reasonable to assume that the data are normally distributed. Even after slight non-linearity is removed by including a quadratic term, the estimated between- and within-cluster slopes are rather different. Table 1 shows the results of fitting both Model (5) and Model (1).

Note the large discrepancy between the within-cluster (0.088) and between-cluster (0.029) slope estimates. If one unsuspectingly fits the model which assumes common covariate effects, an estimate of 0.065 would be obtained, which lies between the between- and within-cluster slope estimates as our theory predicted. The estimated regression parameters represent the change in the logarithm of BMI per decade of age. For a more direct interpretation, the within-cluster and between-cluster estimates correspond to 9.2% and 2.9% change in BMI per decade of age, respectively, whereas the naive estimate implies a 6.7% change.

Since the effect of the quadratic term is very small and estimated to be the same in the two models, we may treat this problem as a single covariate problem. The covariate in this data follows Model (7) rather well. It can be estimated

that $(\xi_{x_1}, \xi_{x,2}) = (42.88, 46.52)$ and $\rho_x = 0.854$. Thus the $\lambda(\hat{\rho}^*)$ defined by (10) is calculated to be 0.384, leading to $\hat{\alpha}^* = 2.9987$ and $\hat{\beta}^* = 0.065$, both of which agree with the GEE estimates from Model (1).

In this example, a likely cohort effect can be attributed to the weight-gaining trend of the U.S. population. It is well known that the American (as well as other countries') population has continuously experienced substantial weight gain. In other words, the population cohort born in 1970, for example, is on the average heavier than the cohort born in 1960. This corresponds to a cohort effect negatively correlated with the primary covariate. The results of Section 3 imply that, if a cohort effect is the only cause of the different between- and within-cluster effects of aging, the estimate 0.088 is produced by an unbiased estimator of β_x whereas 0.029 is the value of a biased estimator. The direction of bias agrees with that given by Theorem 1, when the omitted confounder is negatively correlated with age.

Example 2 (Formaldehyde Emission in Mobile Homes). Hanrahan et al. (1985) examined the health effects and trends in formaldehyde levels in mobile homes, by collecting data in mobile homes during 1980-1981. The number of observations in each home ranges from 2 to 10, with 9 the modal value. For illustration, we consider two subsets of data: 49 homes with 6 observations and 51 homes with 9 observations. We study changes in formaldehyde level (in ppm) with home aging. Previous studies (Palta, Yao and Velu (1994)) have found that taking the logarithm of home age (in months) removes the non-linearity of the trend. We fit Models (5) and (1) to the first group, the second group, and the combined data. The results are summarized in Table 2. The estimated slope in Model (1) is negative, consistent with a decrease in formaldehyde emission as building materials age. However, in Model (5), the rate of decrease is rather different within and across the clusters. For each group of the data, the estimates from the misspecified model follow the results in Theorem 1. In the combined data set, the slope estimate (-0.27) in the misspecified model is a weighted combination of the two slope estimates (-0.25 and -0.31) from the two groups, as indicated by (18).

Previous analysis (Palta and Qu (1995)) of this data set provided evidence for the presence of period effect caused by the variation of temperature, as many homes entered the study in spring and were followed through fall and formaldehyde emission decreases with lower temperature and humidity. Based on the results in Section 3, this would suggest that -0.20 is a better estimate of β_x than -0.49.

Example 3 (Perceived Sleepiness). In the Wisconsin Sleep Cohort Study, subjects scored themselves at five levels from "never" to "almost always" on aspects of perceived sleepiness, such as "not feeling rested during the day, no matter

Table 2. Estimates of regression parameters in Example 2. The response variable is formaldehyde level in ppm and the predictor is the logarithm of home age in months. The numbers in the parentheses are standard errors.

		Model (1)	Model (5)	
	Covariate		within	between
$k = 6$	Intercept	1.24	1.12	
	log(age)	-0.25(0.04)	-0.47(0.10)	-0.21(0.04)
$k = 9$	Intercept	1.43	1.04	
	log(age)	-0.31(0.03)	-0.49(0.05)	-0.18(0.04)
Combined	Intercept	1.32	1.09	
	log(age)	-0.27(0.02)	-0.49(0.05)	-0.20(0.03)

how many hours of sleep you had", "feelings of excessive daytime sleepiness", and "need for coffee or other stimulants to stay awake during the day". A factor score based on answers to six questions has been found to be significantly related to general health status. It takes values between 0 and 100, with higher scores indicating more serious day-time sleepiness. Our interest is to assess the relationship between the factor score and sleep latency, which is the amount of time (in minutes) required to fall asleep at night, and some other covariates, including the amount of sleep during a workday night (in hours), body mass index (weight in kilograms divided by the square of height in meters), and age (in years). Preliminary analysis indicated that the log-transformation should be applied to sleep latency and body mass index (BMI).

Results from two different models are in Table 3. The estimated coefficient for sleep latency (2.932) under the misspecified model lies between the corresponding between- and within-cluster coefficients estimates (4.450 and 1.268). The estimated coefficients of sleep time and BMI do not significantly differ between the two models. The estimated longitudinal and cross-sectional correlation matrices indicate that these covariates are essentially uncorrelated with age and sleep latency. Therefore, according to Section 2, their effects can still be consistently estimated despite the false assumption of the common between- and within-cluster effects of age and sleep latency.

In this study, measurement error may have caused the discrepancy between two covariate effects of sleep latency. Conceivably, it is difficult to accurately self-assess how long it takes before one falls asleep, thus the self-reported sleep latency contains a fair amount of measurement error. This is confirmed by laboratory measurements on a subset of subjects. Estimates from the DeAR method developed in Section 4 are given in the last column of Table 3. The DeAR estimate of the coefficient for sleep latency is 5.212, with a standard error 0.669 estimated from 10,000 bootstrap samples. This is very different from the naive estimate, 2.932, obtained by fitting Model (1) with ignored measurement error

Table 3. Estimates of regression parameters in Example 3. The response variable is score for sleepiness (between 0 and 100) and the predictor variables are the logarithm of sleep latency in minutes, sleep time in hours, the logarithm of BMI in kg/m^2 and age in years. The numbers in the parentheses are standard errors.

	Model (1)	Model (5)		DeAR
Covariate		within	between	
log(Latency)	2.932(0.365)	4.450(0.507)	1.268(0.524)	5.212(0.669)
Sleep time	-2.451(0.294)	-2.654(0.434)	-2.212(0.399)	-2.741(0.616)
log(BMI)	7.652(1.789)	7.375(1.946)	6.938(4.543)	6.976(2.200)
Age	-0.305(0.047)	-0.296(0.048)	0.069(0.314)	-0.307(0.051)

in sleep latency. The DeAR estimates of other regression coefficients are not very different from the naive ones, because of the weak correlation among the covariates.

6. Conclusion

When the between- and within-cluster covariate effects are different, we have shown, for general situations such as multiple covariates and unequal cluster sizes, that fitting the naive Model (1) not only leads to misleading regression coefficient estimates, but also produces biased estimators for the variance components. Any meaningful interpretation of the coefficients in Model (5) requires an underlying framework which explains the difference between the between- and within-cluster covariate effects. We have studied omitted confounders and measurement error, and shown that different between- and within-cluster covariate effects arise from these two situations, and derived results on the bias of the naive estimators. Our study provides important information for the design of longitudinal studies when confounding or measurement error is of concern. We have shown that, for the measurement error models discussed in this paper, it is possible to consistently estimate regression parameters without using supplemental data.

Acknowledgements

This research was partially supported by the NCI grant R01-CA53786 and NSF grant DMS-0404535. The authors would like to thank two referees for their helpful comments.

References

Buonaccorsi, J., Demidenko, E. and Tosteson, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statist. Sinica* **10**, 885-903.

Chao, W-H., Palta, M. and Young, T. (1997). Effect of omitted confounders on the analysis of correlated binary data. *Biometrics* **53**, 678-689.

Hanrahan, L. P., Anderson, H. A., Dally, K. A., Eckman, A. D., and Kanarek, M. S. (1985). Formaldehyde concentrations in Wisconsin mobile homes. *J. the Air Pollution Control Assoc.* **35**, 1164-1167.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251-1271.

Laird, N. M. 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 generalized linear models. *Biometrika* **73**, 13-22.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effect models for repeated-measures data. *J. Amer. Statist. Assoc.* **83**, 1014-1022.

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

Maddala, G. S. (1971). The use of variance components models in pooling cross section and time series data. *Econometrica* **39**, 341-358.

Mundlak, Y. (1978). On the pooling of time series and cross-section data. *Econometrica* **46**, 69-85.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54**, 638-645.

Palta, M. and Yao, T-J. (1991). Analysis of longitudinal data with unmeasured confounders. *Biometrics* **47**, 1355-1369.

Palta, M., Yao, T.-J. and Velu, R. (1994). Testing for omitted variables and non-linearity in regression models for longitudinal data. *Statist. Medicine* **13**, 2219-2231.

Palta, M. and Qu, R. P. (1995). Testing lack of fit in mixed effects models for longitudinal data. *New Trends in Probability and Statistics. Vol.* 3. *TEV Vilnius, Lithuania*, 93-106.

Rothman, R. J. and Greenland, S. (1998). *Modern Epidemiology.* 2nd edition. Lippicott-Raven, Philadelphia, PA.

Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *J. Amer. Statist. Assoc.* **77**, 848-854.

Tosteson, T. D., Buonaccorsi, J. P. and Demidenko, E. (1998). Covariate measurement error and the estimation of random effect parameters in a mixed model for longitudinal data. *Statist. Medicine* **17**, 1959-71.

Wang, N., Lin, X., Gutierrez, R. G. and Carroll, R. J. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *J. Amer. Statist. Assoc.* **93**, 249-261.

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

Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika* **30**, 16-28

Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S. and Badr, S. (1993). The occurrence of sleep disordered breathing among middle-aged adults. *New England J. Medicine* **328**, 1230-1235.

Eli Lilly and Company, Lilly Corporate Center, Dropcode 0734, Indianapolis, IN 46285 U.S.A.
E-mail: shen_lei@lilly.com

Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.
E-mail: shao@stat.wisc.edu

Eli Lilly and Company, Lilly Corporate Center, Dropcode 0734, Indianapolis, IN 46285 U.S.A.
E-mail: park_soomin@lilly.com

Department of Public Health, University of Wisconsin-Madison, Madison, WI 53706, U.S.A.
E-mail: mpalta@wisc.edu

(Received April 2006; accepted November 2006)