Biometrics 63, 1262–1268
December 2007
DOI: 10.1111/j.1541-0420.2007.00818.x

Just-Identified Versus Overidentified Two-Level Hierarchical Linear Models with Missing Data

Yongyun Shin

University of Michigan, 439 West Hall, 1085 South University,
Ann Arbor, Michigan 48109-1107, U.S.A.
email: choil@umich.edu

and

Stephen W. Raudenbush

University of Chicago, Department of Sociology,
Room 418, 1126 East Illinois 59th Street,
Chicago, Illinois 60637, U.S.A.
email: Sraudenb@uchicago.edu

Summary. The development of model-based methods for incomplete data has been a seminal contribution to statistical practice. Under the assumption of ignorable missingness, one estimates the joint distribution of the complete data $f(y \mid \theta) = f(y_{\text{obs}}, y_{\text{mis}} \mid \theta)$ for $\theta \in \Theta$ from the incomplete or observed data y_{obs}. Many interesting models involve one-to-one transformations of θ. For example, with $y_i \sim N(\mu, \Sigma)$ for $i = 1, \ldots, n$ and $\theta = (\mu, \Sigma)$, an ordinary least squares (OLS) regression model is a one-to-one transformation of θ. Inferences based on such a transformation are equivalent to inferences based on OLS using data multiply imputed from $f(y_{mis} \mid y_{obs}, \theta)$ for missing y_{mis}. Thus, identification of θ from y_{obs} is equivalent to identification of the regression model. In this article, we consider a model for two-level data with continuous outcomes where the observations within each cluster are dependent. The parameters of the hierarchical linear model (HLM) of interest, however, lie in a subspace of Θ in general. This identification of the joint distribution overidentifies the HLM. We show how to characterize the joint distribution so that its parameters are a one-to-one transformation of the parameters of the HLM. This leads to efficient estimation of the HLM from incomplete data using either the transformation method or the method of multiple imputation. The approach allows outcomes and covariates to be missing at either of the two levels, and the HLM of interest can involve the regression of any subset of variables on a disjoint subset of variables conceived as covariates.

Key words: Hierarchical linear model; Ignorably missing data; Maximum likelihood; Multiple imputation; Overidentified; Random coefficients model.

1. Introduction

Missing data are ubiquitous in many domains of inquiry. Until quite recently, most analysts facing missing data either discarded cases having missing values or applied ad hoc methods of imputation. Such strategies are, in general, subject to biases in point estimation, uncertainty estimation, or both. Seminal work in recent years has placed the analysis of such data on a principled basis (Orchard and Woodbury, 1972; Rubin, 1976, 1987, 1996; Dempster, Laird, and Rubin, 1977; Schafer, 1997; Liu, Taylor, and Belin, 2000; Little and Rubin, 2002; Schafer and Yucel, 2002). These model-based approaches are based on assumptions that are comparatively mild in many applications: that the data are missing at random (MAR) and that the parameters generating the complete data are distinct from the parameters governing the missing data process (Rubin, 1976). Under MAR, likelihood-based methods for missing information (Orchard and Woodbury, 1972), notably the expectation–maximization (EM) algorithm (Dempster et al., 1977; Wu, 1993), provide efficient estimation of the parameters of complete data based solely on analysis of the observed data.

Framed generally and following Little and Rubin (2002), we may denote the complete-data vector Y as generated by a model $Y \sim f(y \mid \theta)$. The object of our inquiry is to make inferences about θ. However, the data are subject to missingness. Let $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ for observed Y_{obs} and missing Y_{mis}. We assume that Y_{obs} is selected from Y by a stochastic process governed by parameters ϕ, where θ and ϕ are distinct. Let M be a missing-value indicator vector such that the kth element is 1 if the kth element of Y is missing and 0 otherwise. The key assumption is that $M \sim g(m \mid y, \phi)$ is conditionally independent of Y_{mis} given Y_{obs}, that is $g(m \mid y, \phi) = g(m \mid y_{\text{obs}}, \phi)$. This is the MAR assumption, which requires that any association between M and Y is explained by Y_{obs}.

In this article, we consider inference for the two-level hierarchical linear model (HLM) when outcomes and covariates are MAR. Our approach extends the logic of inference for the normal theory ordinary least squares (OLS) regression under MAR. The OLS parameters are, of course, one-to-one transformations of the parameters of the multivariate normal distribution $Y_i \sim N(\beta, \Sigma)$, where Y_i is a complete data vector of a response variable and covariates for unit $i = 1, 2, \ldots, n$. Let O_i denote an observed-value indicator matrix, each row of which contains a single 1 with all other elements in that row equal to 0, such that $Y_{\text{iobs}} = O_i Y_i \sim N(O_i\beta, O_i\Sigma O_i^T)$. Inference about the parameters $\theta = (\beta, \Sigma)$ can be obtained from Y_{iobs} via maximum likelihood (ML) using, for example, the EM algorithm or Fisher scoring. Under OLS, we partition the data such that $Y_i = [R_i\ W_i^T]^T$ where R_i is a scalar response variable and W_i is a vector of covariates. We focus on the parameters of the conditional distribution $Y_i \mid W_i \sim N(\gamma_0 + W_i^T\gamma_1, \sigma^2)$ as well as the marginal distribution $W_i \sim N(\beta_w, \Sigma_{ww})$. The parameters $\theta^* = (\gamma_0, \gamma_1, \sigma^2, \beta_w, \Sigma_{ww})$ represent a one-to-one transformation of θ (Little and Rubin, 2002, Chapter 6). Under MAR, inference for the OLS may proceed in one of the two ways. First, one may compute the ML estimates (MLE) $\hat{\theta}$ of θ, then transform to θ^*; we call this "*MLE on* Y_{obs}." Second, one may generate multiply imputed $Y^{mi} = (R^{mi}, W^{mi})$ from $f(y \mid y_{\text{obs}}, \hat{\theta})$ and then compute the usual OLS estimates of θ^* from each of the multiply imputed data sets, combining the estimates as specified by Rubin (1987); we call this "*MLE on* Y^{mi}."

We extend these two methods to the two-level HLM. The extension, however, is not straightforward because, unlike in the OLS case, the parameters of the joint distribution of the response variables and covariates are not generally one-to-one transformations of the parameters of the HLM. We focus on the problem of aligning this joint distribution with the HLM of interest. If this problem is ignored, substantially biased inferences may result.

Related work by Liu et al. (2000) considered Bayes inference to longitudinal designs having a fixed within-subject design with repeated measures. This is a special case of a two-level design where level-1 units are occasions nested within persons at level 2, where the level-1 design is invariant across level-2 units, and where the data are ignorably missing at both levels. The level-1 covariance matrix was diagonal. Schafer and Yucel (2002) developed Bayes and ML inference for a broader class of two-level HLMs in which the level-1 design matrix could vary across level-2 units. This flexibility incorporates longitudinal designs in which the timing of repeated measures varies arbitrarily across subjects. It also extends to two-level cross-sectional designs, for example, in which students are nested within schools or workers within firms. The approach allows level-1 data to be MAR.

This article builds on the past work. First, our primary aim is to consider the overidentification problem that arises in an HLM. Unless considerable care is taken in specifying the joint distribution of the complete data, the HLM will be overfit and that may cause substantially biased inferences due to the imputation model being uncongenial to subsequent analysis of the HLM (Meng, 1994; Rubin, 1996), and this issue has not yet commanded attention. Second, we generalize the application by allowing data to be missing at either level and by allowing a flexible multivariate approach in which any subset of variables are regressed on a disjoint subset of variables conceived as covariates. We shall consider models in which regressors having random coefficients are completely observed. The more general case with such regressors partially observed can be handled within a Bayesian framework, but we shall avoid that option to maintain our focus on the identification problem. Following Liu et al. (2000) and Schafer and Yucel (2002), we restrict our attention to two-level multivariate normal data using likelihood-based inference (Dempster, Rubin, and Tsutakawa, 1981; Laird and Ware, 1982; Longford, 1993; Goldstein, 1995; Pinheiro and Bates, 2000; Raudenbush and Bryk, 2002) and leave useful extensions to a broader range of distributions, and to three or more levels to future work.

The ML approach provides fast computation and is most appropriate when the number of level-2 units is moderately large. In this context, we explore some issues of data analysis and interpretation that arise in an HLM, accommodating a general missing pattern.

Sections 2 and 3 describe the model and estimation including special cases that yield familiar methods. Sections 4 and 5 illustrate the approach using data sets from two large-scale surveys. The discussion section follows.

2. Model

Our general strategy is to model the joint distribution of a response variable and covariates subject to missingness. Having accomplished this objective, we have the option of either estimating an HLM using the *MLE on* Y_{obs} or the *MLE on* Y^{mi}. In pursuing this strategy, however, the general form of the joint distribution identifies more parameters than are typically of interest in subsequent analysis. We illustrate the problem of over-identification and propose a reasonably general modeling framework for managing this problem.

To illustrate the problem of over-identification, consider a simple HLM

$$R_{ij} = \gamma_0 + \gamma_1 W_{ij} + u_j + e_{ij} \sim N\big(\gamma_0 + \gamma_1 W_{ij}, \tau + \sigma^2\big), \quad (1)$$

where $u_j \sim N(0, \tau)$ and $e_{ij} \sim N(0, \sigma^2)$ for level-1 $i = 1, \ldots, n_j$ nested within level-2 $j = 1, \ldots, J$. A joint model of $Y_{ij} = [R_{ij}\ W_{ij}]^T$ is

$$Y_{ij} = \beta + b_j + \epsilon_{ij} \sim N(\beta, \Psi + \Sigma), \quad (2)$$

$$\text{for } \beta = \begin{bmatrix} \beta_r \\ \beta_w \end{bmatrix}, \Psi = \begin{bmatrix} \psi_r & \psi_{rw} \\ \psi_{rw} & \psi_{ww} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \sigma_{rr} & \sigma_{rw} \\ \sigma_{rw} & \sigma_{ww} \end{bmatrix},$$

$$\text{where } b_j = [b_{rj} b_{wj}]^T \sim N(0, \Psi),$$

and $\epsilon_{ij} \sim N(0, \Sigma)$. Relating model (1) to model (2) results in $\gamma_0 = \beta_r - \frac{\psi_{rw} + \sigma_{rw}}{\psi_{ww} + \sigma_{ww}}\beta_w, \gamma_1 = \frac{\psi_{rw} + \sigma_{rw}}{\psi_{ww} + \sigma_{ww}}, \tau = \psi_{rr}$, and $\sigma^2 = \sigma_{rr} - \frac{(\psi_{rw} + \sigma_{rw})^2}{\psi_{ww} + \sigma_{ww}}$. Model (2), which contains eight parameters, is over-parameterized in representing model (1), which involves seven parameters including $W_{ij} \sim N(\beta_w, \psi_{ww} + \sigma_{ww})$. Let us

express model (2) such that it recognizes the latent random effect b_{wj} of W_{ij} in cluster j

$$\begin{bmatrix} R_{ij} \\ W_{ij} - b_{wj} \\ b_{wj} \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_r \\ \beta_w \\ 0 \end{bmatrix}, \begin{bmatrix} \psi_{rr} + \sigma_{rr} & \sigma_{rw} & \psi_{rw} \\ \sigma_{rw} & \sigma_{ww} & 0 \\ \psi_{rw} & 0 & \psi_{ww} \end{bmatrix} \right). \quad (3)$$

Then, a regression of R_{ij} on other variables leads to

$$\begin{aligned} &R_{ij} \,|\, W_{ij}, b_{wj} \\ &\sim N\left(\left(\beta_r - \frac{\sigma_{rw}}{\sigma_{ww}}\beta_w \right) + \frac{\sigma_{rw}}{\sigma_{ww}} W_{ij} + \left(\frac{\psi_{rw}}{\psi_{ww}} - \frac{\sigma_{rw}}{\sigma_{ww}} \right) b_{wj}, \right. \\ &\qquad \left. \left(\psi_{rr} - \frac{\psi_{rw}^2}{\psi_{ww}} \right) + \left(\sigma_{rr} - \frac{\sigma_{rw}^2}{\sigma_{ww}} \right) \right). \end{aligned} \quad (4)$$

Model (4) implies model (1) if $b_{wj} = 0$. The transformation of model (2) for $b_{wj} = 0$ identifies model (1) with $\gamma_0 = \beta_r - \frac{\sigma_{rw}}{\sigma_{ww}}\beta_w, \gamma_1 = \frac{\sigma_{rw}}{\sigma_{ww}}, \tau = \psi_{rr}$, and $\sigma^2 = \sigma_{rr} - \frac{\sigma_{rw}^2}{\sigma_{ww}}$. Model (2) with $b_{wj} = 0$, however, has a strong assumption that W_{ij} does not vary across level-2 units. Violation of the assumption leads to underestimation of the standard errors of effects.

Model (4) also implies model (1) if $\alpha = \frac{\psi_{rw}}{\psi_{ww}} = \frac{\sigma_{rw}}{\sigma_{ww}}$. Model (2) under the "$\alpha$" constraint implies $\gamma_0 = \beta_r - \alpha\beta_w$, $\gamma_1 = \alpha$, $\tau = \psi_{rr} - \psi_{rw}\ \alpha$, and $\sigma^2 = \sigma_{rr} - \sigma_{rw}\ \alpha$. Thus, the constrained joint model (2) identifies model (1). The standard errors in model (1) will be correctly estimated taking uncertainty at both levels into account.

To extend these ideas, we now propose a reasonably general HLM

$$R_{ij} = C_{ij}^T\gamma + D_{ij}^T u_j + e_{ij} \sim N\big(C_{ij}^T\gamma,\ D_{ij}^T\tau D_{ij} + \sigma^2\big), \quad (5)$$

where R_{ij} is a scalar response variable, C_{ij} is a vector of covariates having fixed effects γ, D_{ij} is a vector of completely observed covariates having random effects $u_j \sim N(0,\ \tau)$, and $e_{ij} \sim N(0,\ \sigma^2)$ for $i = 1, \ldots, n_j$ nested within $j = 1, \ldots, J$. Although our method does not require the presence of an intercept in D_{ij}, many applications do. Thus, we let $D_{ij} = [1\ \ X_{dij}^T]^T$ for covariates X_{dij} having random slopes where the subscript "d" denotes covariates in D_{ij}. Raudenbush and Bryk (2002, "RB" hereafter) review statistical inference in case of complete data using either ML or Bayes methods.

To facilitate statistical inference with incomplete data, we reparameterize model (5) in terms of the joint distribution of the response and all covariates subject to missingness conditional on all completely observed covariates. Let p_1-vector X_{1ij} and p_2-vector X_{2j} be completely observed level-1 and level-2 covariates, respectively, in C_{ij}. The covariate vectors subject to missingness are q_1-vector W_{1ij} and q_2-vector W_{2j} at levels 1 and 2, respectively. Thus, $C_{ij}^T = [X_{1ij}^T\, X_{2j}^T\, W_{1ij}^T\, W_{2j}^T]$ and $\gamma = [\gamma_{x1}^T\, \gamma_{x2}^T\, \gamma_{w1}^T\, \gamma_{w2}^T]^T$. For n a positive integer, let I_n denote an n by n identity matrix. The joint distribution of R_{ij}, W_{1ij}, $W_{2j}\,|\,X_{1ij}$, X_{2j}, X_{dij} is

$$\begin{aligned} \begin{bmatrix} R_{ij} \\ W_{1ij} \\ W_{2j} \end{bmatrix} &= \begin{bmatrix} X_{1ij}^T & X_{2j}^T & 0 & 0 \\ 0 & 0 & I_{q_1} \otimes \big[X_{1ij}^T\ X_{2j}^T\big] & 0 \\ 0 & 0 & 0 & I_{q_2} \otimes X_{2j}^T \end{bmatrix} \begin{bmatrix} \beta_{r1} \\ \beta_{r2} \\ \beta_{w1} \\ \beta_{w2} \end{bmatrix} \\ &+ \begin{bmatrix} 1 & X_{dij}^T & 0 & 0 \\ 0 & 0 & I_{q_1} & 0 \\ 0 & 0 & 0 & I_{q_2} \end{bmatrix} \begin{bmatrix} b_{r0j} \\ b_{r1j} \\ b_{w1j} \\ b_{w2j} \end{bmatrix} + \begin{bmatrix} \epsilon_{rij} \\ \epsilon_{w1ij} \\ 0 \end{bmatrix}, \end{aligned} \quad (6)$$

where

$$\begin{bmatrix} b_{r0j} \\ b_{r1j} \\ b_{w1j} \\ b_{w2j} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \psi_{r0r0} & \psi_{r0r1} & \psi_{r0w1} & \psi_{r0w2} \\ \psi_{r1r0} & \psi_{r1r1} & 0 & 0 \\ \psi_{w1r0} & 0 & \psi_{w1w1} & \psi_{w1w2} \\ \psi_{w2r0} & 0 & \psi_{w2w1} & \psi_{w2w2} \end{bmatrix} \right),$$

and $\begin{bmatrix} \epsilon_{rij} \\ \epsilon_{w1ij} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_{rr} & \Sigma_{rw1} \\ \Sigma_{w1r} & \Sigma_{w1w1} \end{bmatrix} \right)$.

We assume $\mathrm{Cov}(b_{r1j},\ b_{w1j}) = \mathrm{Cov}(b_{r1j},\ b_{w2j}) = 0$. The nonzero covariances could be estimated. However, they introduce extraneous quadratic effects between D_{ij} and C_{ij} in model (5), leading to interpretational difficulties. Let $\alpha^T = \Sigma_{rw1}\Sigma_{w1w1}^{-1} = (\psi_{r0w1} - \psi_{r0w2}\psi_{w2w2}^{-1}\psi_{w2w1}) \times (\psi_{w1w1} - \psi_{w1w2}\psi_{w2w2}^{-1}\psi_{w2w1})^{-1}$ and $\pi^T = (\psi_{r0w2} - \alpha^T\ \psi_{w1w2})\psi_{w2w2}^{-1}$. Represent $\beta_{w1} = [\beta_{w11}^T\, \beta_{w12}^T \ldots \beta_{w1q_1}^T]^T$ for a vector β_{w1k} of $(p_1 + p_2)$ fixed effects for the kth covariate in W_{1ij} such that $\beta_{w1}^* = [\beta_{w11}\, \beta_{w12} \ldots \beta_{w1q_1}]$. Likewise, let $\beta_{w2} = [\beta_{w21}^T\, \beta_{w22}^T \ldots \beta_{w2q_2}^T]^T$ for a p_2-vector β_{w2l} for the lth covariate in W_{2j} such that $\beta_{w2}^* = [\beta_{w21}\, \beta_{w22} \ldots \beta_{w2q_2}]$. Model (6) is a reparameterization of model (5), where $[\gamma_{x1}^T\ \ \gamma_{x2}^T]^T = [\beta_{r1}^T\ (\beta_{r2} - \beta_{w2}^*\,\pi)^T]^T - (\beta_{w1}^*\,\alpha)$, $\gamma_{w1} = \alpha$, $\gamma_{w2} = \pi$, $\sigma^2 = \Sigma_{rr} - \Sigma_{rw1}\,\alpha$ and

$$\tau = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \begin{bmatrix} \psi_{r00} - \psi_{r0w1}\alpha - \psi_{r0w2}\pi & \psi_{r01} \\ \psi_{r10} & \psi_{r11} \end{bmatrix}.$$

Although model (6) is useful in revealing the relationship between the joint model and model (5), an alternative representation that combines data within level-2 j is essential in deriving estimators. Thus, we define

$$Y_{1j} = \begin{bmatrix} Y_{11j} \\ Y_{12j} \\ \vdots \\ Y_{1n_jj} \end{bmatrix}, X_{1j} = \begin{bmatrix} I_{q_1+1} \otimes \big[X_{11j}^T\ X_{2j}^T\big] \\ I_{q_1+1} \otimes \big[X_{12j}^T\ X_{2j}^T\big] \\ \vdots \\ I_{q_1+1} \otimes \big[X_{1n_jj}^T\ X_{2j}^T\big] \end{bmatrix},$$

$$Z_{1j} = \begin{bmatrix} \mathrm{diag}\big\{D_{1j}^T, I_{q_1}\big\} \\ \mathrm{diag}\big\{D_{2j}^T, I_{q_1}\big\} \\ \vdots \\ \mathrm{diag}\big\{D_{n_jj}^T, I_{q_1}\big\} \end{bmatrix}, \epsilon_{1j} = \begin{bmatrix} \epsilon_{11j} \\ \epsilon_{12j} \\ \vdots \\ \epsilon_{1n_jj} \end{bmatrix},$$

where $Y_{1ij} = [R_{ij}\ W_{1ij}^T]^T, \text{diag}\{D_{ij}^T, I_{q_1}\}$ is a diagonal matrix with diagonal elements D_{ij}^T and $I_{q_1}, \beta_1 = [\beta_{r1}^T\ \beta_{r2}^T\ \beta_{w1}^T]^T, b_{1j} = [b_{r0j}\ b_{r1j}^T\ b_{w1j}^T]^T$, and $\epsilon_{1ij} = [\epsilon_{rij}\ \ \epsilon_{w1ij}^T]^T \sim N(0, \Sigma)$. Adopting the modeling notation of Schafer and Yucel (2002), we express model (6) at level-2 j

$$\begin{aligned} Y_j &= X_j\beta + Z_j b_j + \epsilon_j,\ \ b_j \sim N(0, \Psi), \\ \epsilon_j &\sim N\big(0, \text{diag}\big\{I_{n_j} \otimes \Sigma, 0\big\}\big), \end{aligned} \tag{7}$$

which is equivalent to

$$\begin{bmatrix} Y_{1j} \\ W_{2j} \end{bmatrix} = \begin{bmatrix} X_{1j} & 0 \\ 0 & I_{q_2} \otimes X_{2j}^T \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_{w2} \end{bmatrix} + \begin{bmatrix} Z_{1j} & 0 \\ 0 & I_{q_2} \end{bmatrix} \begin{bmatrix} b_{1j} \\ b_{w2j} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ 0 \end{bmatrix}. \tag{8}$$

Let O_j be an observed-value indicator matrix for Y_j such that $Y_{\text{jobs}} = O_j X_j\beta + O_j Z_j b_j + O_j\epsilon_j \sim N\left(O_j X_j\beta, V_j\right)$ for $V_j = O_j(Z_j\Psi Z_j^T + \text{diag}\{I_{n_j} \otimes \Sigma, 0\})O_j^T$. An application to a multivariate HLM (5) with missing data is straightforward by a vector R_{ij} and matrices C_{ij} and D_{ij}.

3. Estimation

We maximize the observed data likelihood via a combination of the EM algorithm for (Ψ, Σ) and Fisher scoring for β in the joint model (7) to hasten the slow convergence of the conventional EM (Laird and Ware, 1982; Schafer and Yucel, 2002). We sketch the estimation of parameters and standard errors. See Web Appendices A, B, and C for details (Magnus and Neudecker, 1998).

The complete data are $Y = (Y_{\text{obs}}, Y_{\text{mis}})$ and b_1. We view (Y_{mis}, b_1) missing (Dempster et al., 1981). The presence of α in both Ψ and Σ makes the estimation difficult. Instead, we use model (5) to estimate α and translate it back into model (7). For the M-step, the complete data MLE are $(\hat\gamma, \hat\Sigma, \hat\Psi, \hat\beta)$, which are then translated to $(\hat\Sigma_{rw1}, \hat\psi_{r0w1}, \hat\psi_{r0w2}, \hat\beta_{r1}, \hat\beta_{r2})$. For the E-step, the estimates of the complete data sufficient statistics ϵ_{1j}, $\epsilon_{1j}\,\epsilon_{1j}^T$, $\epsilon_{1j}\,b_j^T$, b_j and $b_j b_j^T$ are obtained from ϵ_{1j}, $b_j \mid Y_{obs}$, Ψ, Σ, β. Let $d_j = O_j(Y_j - X_j\beta)$. The score and expected Hessian matrices are $S = \sum_{j=1}^J X_j^T O_j^T V_j^{-1} d_j$ and $EH = -\sum_{j=1}^J X_j^T O_j^T V_j^{-1} O_j X_j = H$ (RB). The Fisher scoring update is $\hat\beta = \beta - (EH)^{-1}S$.

Let φ contain distinct elements in (Ψ, Σ). The observed information for $\theta = (\varphi, \beta)$ is

$$I_O = \begin{bmatrix} I_{\varphi\varphi} & I_{\varphi\beta} \\ I_{\beta\varphi} & I_{\beta\beta} \end{bmatrix}$$

for $I_{\varphi\varphi} = \frac{\partial^2 l}{\partial\varphi^T\partial\varphi}, I_{\varphi\beta} = \frac{\partial^2 l}{\partial\varphi^T\partial\beta} = I_{\beta\varphi}^T$, and $I_{\beta\beta} = \frac{\partial^2 l}{\partial\beta^T\partial\beta}$. The Fisher information is $E(I_O)$. The asymptotic variance of $\hat\theta$ is $[E(I_O)]^{-1}$ or I_O^{-1}.

Two special cases yield familiar methods. The first is $Y = Y_{\text{obs}}$. The missing data are b_1. This is an HLM whose level-1 estimation is the full ML method in Dempster et al. (1981); see also Lindley and Smith (1972). The second special case involves no random effects $b = 0$. The joint model is $Y_{1j} \sim N(X_{1j}\beta_1, I_{n_j} \otimes \Sigma)$. The MLE $\hat\beta_1$ is the generalized least-squares estimator with missing data (Beale and Little, 1975; Dempster et al., 1977; Schafer, 1997; Little and Rubin, 2002).

In the next two sections, we illustrate the over-identification problem and the method. The *MLE on* Y_{obs}, both unconstrained and α-constrained, and the *MLE on* Y^{mi} are carried out by `C` programs written by the authors. A random number generating library of `C` routines, `RANDLIB 1.3`, by Barry W. Brown, James Lovato, Kathy Russell, and John Venier is used. `HLM5` (Raudenbush et al., 2002) fits model (5) on Y and Y^{mi} following Rubin's rule (Rubin, 1987). Fisher information is used to provide an objective comparison of standard errors with those of HLM5 based on Fisher information. Starting values for θ are the least-squares estimates. The convergence criterion is the difference in observed loglikelihoods between two consecutive iterations, which is taken to be $<10^{-6}$.

4. Illustrative Example I: Over-Identification Problem

We illustrate the over-identification problem with a subset of the High School and Beyond Study of 1980 that does not contain missing data (RB). The data has 7185 students within 160 schools. Each school has 14 to 67 students surveyed. The variables are described in Table 1. Model (5) of interest has $R_{ij} = \text{MATHACH}_{\ ij}$, $C_{ij} = [1\ \text{SES}_{ij}\ \text{DISCLIM}_j\ \text{LOGSIZE}_j]$ and $D_{ij} = 1$. The estimates appear under HLM in Table 2. The unconstrained and α-constrained joint models (7) of $[R_{ij}\ C_{ij}^T]^T$ are estimated and translated to $R_{ij} \mid C_{ij}$. These estimates follow those under HLM in Table 2. The inaccuracy of estimates in the over-identified model (5) is apparent under the *MLE on* Y unconstrained. There are discrepancies in all estimates and their standard errors against those under HLM. On the contrary, the estimates in the last column closely match those of HLM. It is interesting that there exists so much sensitivity in the estimated HLM (5) by the unconstrained joint model (7).

5. Illustrative Example II: Missing Data

We illustrate estimation of an HLM subject to missingness using the Chicago Community Adult Health Study (CCAHS,

Table 1
Variables used in analysis of the High School and Beyond Study

Level	Variable	Description	Mean (S.D.)
I	MATHACH	Math achievement score	12.75 (6.88)
	SES	Standardized socioeconomic score	0 (0.78)
II	DISCLIM	Measure of disciplinary climate, the higher the worse	−0.02 (0.98)
	LOGSIZE	Log(school enrollment)	6.79 (0.73)

Table 2
Regression of MATHACH on SES, DISCLIM, and LOGSIZE

		Estimate (S.E.)		
	Predictor	HLM	*MLE on Y* Unconstrained	*MLE on Y* Constrained
γ	Intercept	10.083 (1.642)	10.345 (1.534)	10.083 (1.623)
	SES	2.378 (0.105)	2.975 (0.140)	2.378 (0.105)
	DISCLIM	−1.203 (0.179)	−1.097 (0.168)	−1.202 (0.177)
	LOGSIZE	0.373 (0.240)	0.336 (0.225)	0.373 (0.238)
τ		3.505 (0.490)	6.348 (1.055)	3.505 (0.490)
σ^2		37.024 (0.625)	33.970 (0.800)	37.024 (0.625)

Table 3
Variables used in analysis of CCAHS

Level	Variable	Description	Mean (S.D.)	Missing (%)
I	BMI	Body mass index	28.55 (6.96)	40 (1.3)
	EDUC	The number of years educated	12.71 (3.51)	5 (0.2)
	INCOME	In $10K's, incomes >20 set to 20	4.50 (3.66)	577 (18.6)
	AGE	Age in years	42.50 (16.46)	0 (0.0)
	FEMALE	1 if female	0.60 (0.49)	0 (0.0)
II	SDISO	Measure of social disorder	−7.17 (0.91)	263 (76.7)

Morenoff et al. 2006), a survey of 3105 adults living in 343 Chicago neighborhoods. The general aim of this study is to investigate social disparities in health arising from neighborhood and person-level risk factors. Here we consider neighborhood social disorder (SDISO) and individual-level covariates education (EDUC), household income (INCOME), age (AGE), and female indicator (FEMALE) as predictors of body mass index (BMI; see Table 3). SDISO is a scale of seven items indicating loitering adults, public drinking, youth displaying gang indicators, adults fighting or arguing hostilely, and the presence of drug sales or prostitutes on the street (see Raudenbush and Sampson, 1999, for details). We reason that healthful food would be relatively less available in more highly disordered neighborhoods, and that lifestyles in such neighborhoods would be unconducive to good nutrition. We therefore expect that residence in neighborhoods characterized by high levels of social disorder will be associated with elevated BMI even after adjustment for education, income, age, and gender. Measurement of social disorder is expensive, requiring the videotaping of each of several hundred "block faces" (sides of a city block) in each neighborhood, followed by coding of the videotapes. As a result, the investigators decided to study SDISO in a random sample of 80 of the 343 neighborhoods from which the subjects were drawn. The key explanatory variable at the neighborhood level is thus missing completely at random (MCAR, Rubin, 1976) in 77% of the neighborhoods. In addition, INCOME is missing on 19% of the 3105 persons, a common result in large-scale survey research. Our analysis uses all available data under the assumption of MAR.

5.1 *Random Intercept Model with Missing Data*

In this section, we illustrate the *MLE on* Y_{obs} and the *MLE on* Y^{mi}. Let $R_{ij} = \text{BMI}_{ij}$, $X_{1ij} = [\text{AGE}_{ij}\, \text{FEMALE}_{ij}]^T$, $X_{2j} = 1$, $W_{1ij} = [\text{EDUC}_{ij}\, \text{INCOME}_{ij}]^T$, $W_{2j} = \text{SDISO}_j$ and $D_{ij} = 1$ in HLM (5). It took 122 iterations for the constrained joint model (7) to converge. We imputed five complete data sets (Rubin, 1987; Schafer, 1997). To propagate the uncertainty in estimation, we generated θ from its sampling distribution estimated by ML and then generated missing data given the θ. Let v_{ii} and v_{ij} be variances of $\log(\psi_{ii})$ and $\log\frac{1+\rho_{ij}}{1-\rho_{ij}}$ for $i \neq j$, where $\rho_{ij} = \frac{\psi_{ij}}{\sqrt{\psi_{ii}\psi_{jj}}}$. We generated $\beta \sim N(\beta,\, I_\beta^{-1})$, diagonal $\psi_{ii} \sim N(\log\,(\psi_{ii}),\, v_{ii})$, and off-diagonal $\psi_{ij} \sim N(\log\frac{1+\rho_{ij}}{1-\rho_{ij}}, v_{ij})$. With many level-1 units, Σ was estimated accurately enough to be fixed at ML $\hat{\Sigma}$. Both methods suggest, as expected based on available literature, that education is negatively related to BMI while age and female gender are positively related to BMI in Table 4. Net these associations, both methods show a positive association between neighborhood social disorder and BMI. Moderate differences show up in the effects of social disorder and the τ estimates across the *MLE on* Y_{obs} and the *MLE on* Y^{mi}. Other estimates

Table 4
Random intercept model of BMI on covariates in Table 3

		Estimate (S.E.)	
	Predictor	*MLE on* Y_{obs}	*MLE on* Y^{mi}
γ	Intercept	27.980 (0.220)	28.010 (0.216)
	EDUC	−0.158 (0.041)	−0.178 (0.040)
	INCOME	0.058 (0.042)	0.047 (0.039)
	AGE	0.032 (0.008)	0.031 (0.008)
	FEMALE	1.022 (0.254)	1.054 (0.259)
	SDISO	0.839 (0.250)	0.659 (0.265)
τ		2.256 (0.589)	1.996 (0.575)
σ^2		44.668 (1.207)	44.996 (1.205)

Table 5
Random coefficient model with AGE having a random coefficient on BMI

	Predictor	Estimate (S.E.) *MLE on* Y_{obs}
γ	Intercept	27.996 (0.221)
	EDUC	−0.160 (0.043)
	INCOME	0.057 (0.043)
	AGE	0.034 (0.008)
	FEMALE	1.017 (0.254)
	SDISO	0.825 (0.085)
τ	τ_{00}	2.322 (0.597)
	τ_{01}	0.020 (0.020)
	τ_{11}	0.001 (0.001)
σ^2		44.279 (1.249)

and standard errors appear to be similar between the two methods.

5.2 *Random Slope Model with Missing Data*

Consider the same HLM (5) as in Table 4 except $D_{ij} = [1\,\text{AGE}_{ij}]^T$. The constrained joint model (7) converged in 1345 iterations. The *MLE on* Y_{obs} appears in Table 5. The results are very close to those for the random intercept model. This is not surprising in that the slope for age varies at most modestly with a 95% confidence interval of (0.000, 0.012).

6. Discussion

The joint model (6) of the variables subject to missingness in HLM (5) over-identifies the HLM. Consequently, considerably biased inferences may result. The approach in this article established a one-to-one transformation between the two models by constraining model (6).

Based on existing methods in estimating an HLM with missing data, analysts will prefer the *MLE on* Y^{mi} to the *MLE on* Y_{obs} (Schafer and Yucel, 2002). The reason is that the joint model assumptions of covariates subject to missingness are nontrivial but affect missing data only under the *MLE on* Y^{mi}. Under the approach in this article, the *MLE on* Y^{mi} and the *MLE on* Y_{obs} are based on the same joint model and hence are equivalent as in single-level data analysis (Collins, Schafer, and Kam, 2001).

Despite benefits of multiple imputation when imputers and analysts are different (Collins et al., 2001; Schafer, 2003), special care needs to be taken with analysis of multilevel data. Because modeling covariates subject to missingness becomes of concern in analysis of an HLM, imputers should concern themselves far more carefully with a model of interest than they would in single-level data. As a result, a more dynamic relationship between imputer and analyst arises in a multilevel data analysis with missing data.

The unconstrained model (6) adds one more parameter for each level-1 covariate subject to missingness in HLM (5) than are of interest in subsequent analysis. This one parameter represents the coefficient for each of the latent random effects b_{w1j} in model (6), which, under the unconstrained model, are assumed distinct from each of the coefficients for (W_{1ij}, W_{2j}). Because the latent means are likely to be highly correlated in many applications, the unconstrained model is likely to become overfit as the number of level-1 covariates W_{1ij} increases. The problem worsens with multivariate R_{ij}. On the contrary, the constrained model (6) achieves parsimony in such cases. One would still economize on parameters by modeling completely observed covariates in X_j because they do not affect the constraints among variables subject to missingness. One other possibility is to constrain some of level-1 variables subject to missingness while others may be left unconstrained.

It is helpful to use extra variables in model (7) not of direct interest but having high correlations with the variables subject to missingness as Collins et al. (2001) and Schafer and Graham (2002) have done in a single-level context. Model (7) can accommodate an arbitrary number of extra variables in both completely observed X_j and missing Y_j.

In model (6), we have encountered discrete covariates subject to missingness under normal distribution at both levels and found no difficulty in both real and simulated data so far. A similar experience has been observed in a single-level context (Schafer, 1997).

Finally, our methodology, implemented in `C`, automatically handles the constraints in model (6) and generates the reduced number, due to the constraints, of variance estimates and their standard errors. It also handles a general missing pattern across multiple levels distinct from existing software.

7. Supplementary Materials

Web Appendices referenced in Section 3 are available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

We thank the reviewers and the associate editor for the careful and helpful comments. The first author thanks his lifetime mentor Kibum Shin.

References

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B* **37,** 129–145.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. *Psychological Methods* **6,** 330–351.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B* **76,** 1–38.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of American Statistical Association* **76,** 341–353.

Goldstein, H. (1995). *Multilevel Statistical Models.* London: Edward Arnold.

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

Lindley, D. and Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B* **34,** 1–41.

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

Liu, M., Taylor, J., and Belin, T. (2000). Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* **56,** 1157–1163.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford: Oxford University Press.

Magnus, J. R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: Wiley.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9,** 538–558.

Morenoff, J. D., House, J. S., Hansen, B. B., Williams D. R., Kaplan, G. A., and Hunte, H. E. (2006). Understanding social disparities in hypertension prevalence, awareness, treatment, and control: The role of neighborhood context. Accepted for publication by *Social Science and Medicine*.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, **1,** 697–715.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-plus*. New York: Springer.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong Y. F., and Congdon R. (2002). *HLM5 Hierarchical Linear and Nonlinear Modeling*. Scientific Software International, Inc.

Raudenbush, S. W. and Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology* **29,** 1–41.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association* **91,** 473–489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when imputation and analysis models differ. *Statistica Neerlandica* **57,** 19–35.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods* **7,** 147–177.

Schafer, J. L. and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* **11,** 437–457.

Wu, C. F. J. (1993). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11,** 95–103.

Received December 2005. *Revised February* 2007.
Accepted February 2007.