10

Multiple Imputation of Multilevel Data

Stef van Buuren
TNO Quality of Life, Department of Statistics, Leiden
and University of Utrecht, The Netherlands

Au: Please check the Author details.

10.1 INTRODUCTION

In the early days of multilevel analysis, Goldstein wrote: "We shall require and assume that all the necessary data at each level are available" (Goldstein, 1987). Despite the many conceptual and computational advances that have been made over the last two decennia, Goldstein's requirement is still dominant today. To illustrate this, consider how modern software for fitting multilevel models deals with missing data. Dedicated packages like MLwiN (Rasbash, Steel, Browne, & Prosser, 2005) and HLM (Raudenbush, Bryk, & Congdon, 2008) remove all level-1 units with missing values on any level-1 variable. If level-2 explanatory variables have missing values, the associated level-2 units are deleted, including all level-1 data. Thus, if the age of the teacher is unknown, all data of all children within the class are removed prior to analysis. Multilevel procedures in general purpose statistical software, like SAS PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 1996), SPSS MIXED (SPSS Inc., 2008), STATA xtmixed (StataCorp LP, 2008), S-PLUS library `nlme3` and the R package `nlme` (Pinheiro & Bates, 2000), and the R package `arm` (Gelman & Hill, 2007) use a similar approach. Deletion is not only wasteful of costly collected data, but it may also bias the estimates of interest (Little, 1992; Little & Rubin, 2002).

Alternative approaches have been tried. In older versions of HLM it was possible to perform pairwise deletion, a method to calculate the covariance matrix where each element is based on the full number of complete cases for that pair of variables. However, this approach causes estimation problems due to the possibility of nonpositive definite covariance matrices. Also, model comparisons in terms of the log-likelihood are debatable since there is no clear-cut way to calculate the degrees of freedom. Version 6 of HLM therefore dropped this feature.

M*plus* (Muthén & Muthén, 2007) uses full information maximum likelihood. This approach specifically deals with the case of multiple outcome

variables. If one or more outcomes are missing, the values of the remaining dependent variables are still used. In this way, there is no need to delete the whole level-1 unit. When there are missing data in any covariates however, M*plus* resorts to listwise deletion.

Some general purpose programs offer modules to impute missing data (e.g., SAS PROC MI and the new Multiple Imputation procedure in SPSS V17.0). These approaches generally ignore the clustering structure in hierarchical data. Not much is known how imputation by such procedures affects the complete data analysis.

This chapter discusses critical issues associated with imputation of multilevel data. Section 10.2 introduces the notation used and outlines how two formulations of the same model are related. Section 10.3 dissects the multilevel missing data problem into five main questions that need to be addressed. Section 10.4 outlines six different strategies for dealing with the missing data problem. Section 10.5 describes a multilevel imputation method for univariate data, and discusses its properties. Section 10.6 describes a method to apply the univariate method iteratively to multivariate missing data. Finally, Section 10.7 sums up the major points and provides directions for future research.

10.2 TWO FORMULATIONS OF THE LINEAR MULTILEVEL MODEL

Let y_j denote the $n_j \times 1$ vector containing observed outcomes on units i $(i = 1,\ldots,n_j)$ within class j $(j = 1, \ldots, J)$. The univariate linear mixed-effects model (Laird & Ware, 1982) is written as

$$y_j = X_j\beta + Z_j u_j + e_j \qquad (10.1)$$

where X_j is a known $n_j \times p$ design matrix in class j associated with the common $p \times 1$ fixed effects vector β, and where Z_j is a known $n_j \times q$ design matrix in class j associated with the $q \times 1$ random effect vectors u_j. The random effects u_j are independently and interchangeably normally distributed as $u_j \sim N(0, \Omega)$. The number of random effects q is typically smaller than the number of fixed effects p. Symbol e_j denotes the $n_j \times 1$ vector of residuals, which are independently normally distributed as $e_j \sim N(0, \sigma_j^2 I(n_j))$ for $j = 1, \ldots, J$. It is often assumed that the residual variance is equal for all classes: $\sigma_j^2 = \sigma^2$. In addition, e_j and u_j are uncorrelated so $\text{cov}(e_j, u_j) = {}_{nj}0_q$, an $n_j \times q$ matrix of zeroes. Model formulation of Equation 10.1 clearly separates fixed from random effects.

AU: Please review: Should that inferior italic n be full size italic n?

It is also convenient to conceptualize Equation 10.1 as constructed from a set of different levels. To see how this works, write the two-level linear model as

$$y_j = Z_j\beta_j + e_j \quad \text{level-1 equation} \qquad (10.2a)$$

where β_j is a $q \times 1$ vector of regression coefficients that vary between the J classes. At level-2, we model β_j by the linear regression model

$$\beta_j = W_j\beta + u_j \quad \text{level-2 equation} \qquad (10.2b)$$

where W_j is a $q \times p$ matrix of a special structure (see below), and where u_j can be interpreted as the $q \times 1$ vector of level-2 residuals. Equations 2a and 2b are sometimes collectively called the slopes-as-outcome model (Bryk & Raudenbush, 1992). Note that the regression coefficient β is identical in all level-2 classes. Substituting Equation 2b into Equation 2a yields

$$y_j = Z_j W_j\beta + Z_j u_j + e_j, \qquad (10.3)$$

which is a special case of the linear mixed model (Equation 10.1) with $X_j = Z_j W_j$.

Matrix W_j has a special structure for the linear multilevel model. Suppose the model contains $q = 2$ random effects (an intercept and a slope) and a level-2 predictor whose values are denoted by w_j ($j = 1, \ldots, J$). The structure of W_j is then

$$W_j = \begin{bmatrix} 1 & 0 & w_j & 0 \\ 0 & 1 & 0 & w_j \end{bmatrix}. \qquad (10.4)$$

The first two columns of W_j correspond to the random intercept and random slope terms, respectively. In the expression $X_j = Z_j W_j$, this part effectively copies Z_j into X_j. Multiplication of Z_j by the third column W_j replicates w_j as n_j elements in class j, thus forming a covariate associated with the main (fixed) effect in matrix X_j. Multiplication by the fourth column adds the interaction between the random slope and the fixed level-2 predictor, also known as the cross-level interaction term. In applications where this term is not needed, one may simply drop the fourth column of W_j. It is easy to extend Equation 10.4 to multiple level-2 predictors by padding additional columns with the same structure. Note that Equation 10.2 implicitly assumes that all level-1 variables are treated as random effects. It is straightforward to exclude the random part for the lth ($l = 1, \ldots, q$) variable by requiring $u_{1l} = \ldots = u_{jl} = \ldots = u_{Jl} = 0$, or equivalently, by setting the corresponding diagonal element in Ω to zero. In the sequel, we assume that all level-1 data are collected into Z_j.

Equation 10.1 separates the fixed and random effects, but the same covariates may appear in both X_j and Z_j. This complicates imputation of those covariates. To make matters more complex, X_j can also contain interactions between covariates at level 1 and level 2. Equation 10.2 distinguishes the level-1 from the level-2 predictors. There is no overlap between W_j and Z_j. This is a convenient parameterization if we are trying to understand the missing data processes that operate on different levels of the data collection.

10.3 CLASSIFICATION OF MULTILEVEL INCOMPLETE DATA PROBLEMS

This section provides a typology of incomplete data problems that can appear in a multilevel context. There are five major factors to consider: the role of the variables in the model, the pattern of the missingness, the missing data mechanism, the distribution of the variable, the design of the study. In order to be able to provide an adequate treatment to the missing data we need answers on the following questions:

- Role: In which variables do the missing data occur?
- Pattern: Do the missing data form a pattern in the data?
- Mechanism: How is the probability to be missing related to the data?
- Scale: What is the scale of the incomplete variables?
- Design: What is the design of the study (e.g., random, clustered, longitudinal)?

This section classifies problems in incomplete multilevel data into five subproblems: role, pattern, mechanism, scale, and design. We briefly indicate the major difficulties

and consequences of missing data in each case. The typology can be used to characterize particular data analytic problems. In addition, the typology provides insight into what fields are well covered in the literature and those less covered. Different combinations of the five factors correspond to different analytic situations and may thus require specialized approaches.

10.3.1 Role of the Variable In the Model

Missing data can occur in y_j, Z_j, W_j, and j. The consequences of incompleteness of a variable depend on the role the variable plays in the multilevel model.

10.3.1.1 Missing Data in y_j

Many classical statistical techniques for experimental designs require balanced data with equal group sizes (Cochran & Cox, 1957). The experimental factors are under control of the experimenter and the missing data typically occur in y_j. The problem of missing data in y_j is that they may destroy the balance present in the original design. In the days of Fisher, this used to be a major setback since the calculations required for the analysis of unbalanced data are much more demanding than those for the balanced case. In a similar vein, the classic approach to analyzing change relies on repeated measurements of the same subject on a fixed number of occasions (de Leeuw & Meijer, 2008). Missing data that occur in repeated measures result in incompleteness of the subject's response vector, which leads to severe complications in MANOVA. Many techniques have been proposed to circumvent and deal with problems of missing outcomes in experiments (Dodge, 1985).

The advent of multilevel modeling opened up new ways of analyzing data with missing y_j. Modern likelihood-based methods have been developed in which missing data in y_j no longer present a problem. Snijders and Bosker (1999, p. 52) write that the model can be applied "even if some groups have sample size $n_j = 1$, as long as other groups have greater sizes." We add that this statement will only go as far as the assumptions of the model are met: data in y_j are missing at random and the model is correctly specified. Section 10.4.5 discusses the likelihood-based approach in more detail.

The problem of missing data in y_j has received vast attention. There is an extensive literature, which often concentrates on the longitudinal case (Daniels & Hogan, 2008; Molenberghs & Verbeke, 2005; Verbeke & Molenberghs, 2000). For more details, see the overview of the state-of-the-art including direct likelihood approaches, Generalized Estimating Equations (GEE), Weighted GEE, and others (Beunckens, Molenberghs, Thijs, & Verbeke, 2007).

10.3.1.2 Missing Data in Z_j

Missing data can also occur in the level-1 predictors Z_j. In applications where pupils are nested within classes, missing data in Z_j occur at the child level: age of the pupil, occupational status of the father, ethnic background, and so on. In longitudinal applications where time is nested within persons, missing data in Z_j may occur on time-varying covariates. Examples include breast-feeding status and stage of pubertal development at a particular age.

The effect of missing data in Z_j is that the estimators become undefined. The usual solution is simply to remove the

incomplete cases before analysis. This is not only wasteful, but may also bias estimates of the regression weights (Little, 1992). Some authors suggest that data missing at the micro units may not need to be replaced or imputed if the data are to be aggregated and the analysis is to be done at the macro level (McKnight, McKnight, Sidani, & Figueredo, 2007). While easy to perform, this advice is only sound under the restrictive assumption that the process that caused the missing data is missing completely at random.

Several solutions for handling missing data in Z_j have been offered. Goldstein proposed to extend the multilevel model with one extra level that contains a dummy variable for each incomplete variable (1987). Petrin implemented this suggestion, and noted that the procedure is "susceptible to producing biased parameters estimates." The procedure requires reorganization of the data and, according to Petrin, is "very tedious" (2006). Schafer noted that missing values in Z_j are problematic since they require a probability model on the covariates (1997). Handling this in general "would require us to incorporate random effects into the imputation model, which remains an open problem." Longford observed that drawing imputations using random effects models is hard because the relevant parameter distributions depend on the within–between classes variance ratio, which is often not estimated with high precision (Longford, 2005).

Schafer and Yucel (2002) suggested transferring incomplete variables in Z_j to the other side of the equation, and impute the missing data in the multivariate outcomes under a joint multivariate model (Yucel, 2008). This approach has been implemented in their PAN package. There is a macro for MLwiN that implements this approach (Carpenter & Goldstein, 2004). Multiple imputation of multilevel data is possible using the chained equations approach (Jacobusse, 2005). This method is implemented in the WinMICE computer program, which can be downloaded from www.multiple-imputation.com. Similar research was done by Yucel, Schenker, and Raghunathan (2006), who called their approach SHRIMP. Longford (2008) proposed an EM-algorithm to estimate the parameters in the multilevel model in case of missing Z_j. In its generality, this approach requires substantial programming effort and, according to Longford, would only be practical if few missing data patterns arise.

10.3.1.3 Missing Data in W_j

The problem of missing data in W_j has received little attention. Missing data in the level-2 predictors W_j occur if, for example, it is not known whether a school is public or private. In a longitudinal setting, missing data in fixed person characteristics, like sex or education, lead to incomplete W_j.

Missing entries in W_j complicate the estimation of group-level effects. The typical fix is to delete all records in the class. For example, suppose that the model contains the professional qualification of the teacher (e.g., teacher school, university, PhD). If the qualification is missing, the data of all pupils in the class are removed before the analysis. Again, this strategy is not only wasteful, but may also lead to selection effects at level 2.

Some have studied the use of (inappropriate) flat-file imputation methods that ignore the hierarchical group structure in multilevel data. Standard errors are underestimated, leading to confidence intervals

that are too short (Cheung, 2007; Gibson & Olejnik, 2003; Roudsari, Field, & Caetano, 2008). Zhang (2005) reports however that flat multiple imputation worked well with multilevel data, and advises that future researchers should feel confident applying the procedure with a missing data level up to 30%. There is no consensus yet on this issue, and some more work is needed to clear things up.

Imputation methods for level-2 predictors should assign the same imputed value to all members within the same class. Some authors suggest creating two data sets, one with only individual-level data, and one with group-level data, and do separate imputations within each data set while using the results from one in the other (Gelman & Hill, 2007; Petrin, 2006). Note that the steps can also be iterated.

10.3.1.4 Missing Data in j

It is also possible that the group identification is unknown. For example, some pupils may have failed to fill in their class number on the form. The result is that the investigator cannot allocate the pupil to a group. Though one might envisage applications of imputing class memberships, we will not deal with the case of missing data in *j*. For now, the only action one could do is to eliminate the record from the data.

10.3.2 Missing Data Pattern

For both theoretical and practical reasons, it is useful to distinguish between monotone and nonmonotone missing data patterns, and between univariate and multivariate missing data patterns. A pattern is monotone if the variables can be ordered such that, for each person, all earlier variables are observed if all subsequent variables are observed. Monotone patterns often occur as a result of *drop out* in a longitudinal study. It is often useful to sort variables and cases to approach a monotone pattern.

Little and Rubin (2002) graphically demonstrate the univariate/multivariate and the monotone/nonmonotone distinctions for flat files. Things become more complicated in the context of multilevel data. Figure 10.1 demonstrates several possibilities. Figure 10.1a is the case where all missing data are confined to the outcome y_j, and where a person is lost once dropped out. Figure 101b depicts the situation where the person only misses one or more visits, but does not completely drop out. This leads to missing data that are *intermittent*. Note that the difference between 10.1a and 10.1b only makes sense for longitudinal data (i.e., when Z_j can be interpreted as time).

If Z_j attains identical values in each group (i.e., if the data are repeated measures at fixed time points), we can reorder the file into a broad matrix where each cluster occupies one record, and where a set of columns represent the time points. It is then easy to see that drop out leads to a monotone missing data problem, whereas intermittent missing data result in a nonmonotone pattern. The practical usefulness of a monotone pattern is that it opens up the possibility to solve the missing data problem by a sequence of simple steps without the need to iterate (Little & Rubin, 2002).

Figure 10.1c represents the situation where there are also missing data in level-1 predictors Z_j. For example, Z_j could contain body height and y_j could be body weight. Multilevel multivariate missing data usually correspond to a missing data pattern that is nonmonotone. Figure 10.1d depicts the

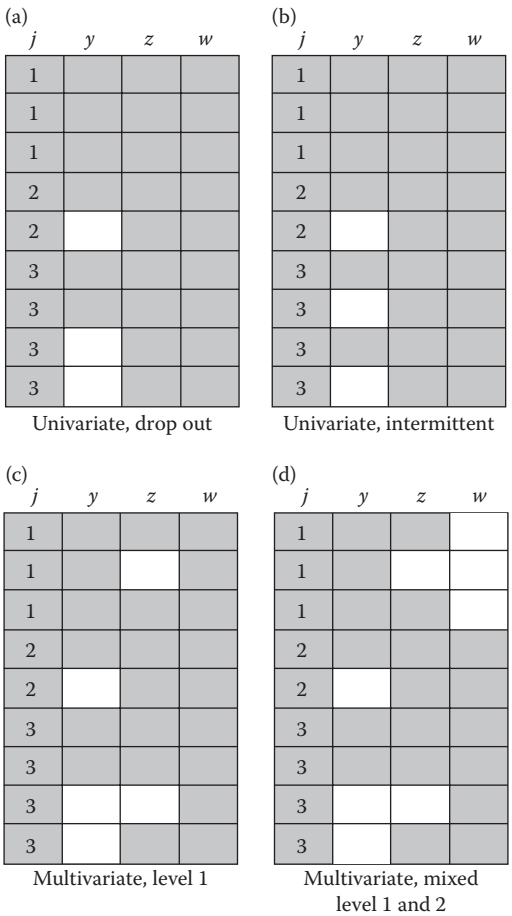

FIGURE 10.1
Four typical missing data patterns in the multilevel data with two levels and three groups. The grey parts represent observed data, whereas the transparent cells indicate the missing data.

one most general situation where missing data occur in level-2 predictors W_j, level-1 predictors Z_j and level-1 outcomes y_j. Note that all level-1 units have missing level-2 predictors if W_j is missing. This is perhaps the most complex case, but also a case that occurs frequently.

10.3.3 Missing Data Mechanism

The process that created the missing data influences the way the data should be analyzed. Except in artificial cases, the precise form of the missingness process is generally unknown, so one has to make assumptions. If the probability to be missing is independent of both unobserved and observed data, then the data are said to be Missing Completely at Random (MCAR; Rubin, 1976). If, conditional on the observed data, the probability to be missing does not depend on the unobserved data, then the data are said to be Missing at Random (MAR). Note that MCAR is a special case of MAR. A mechanism that is neither MCAR nor MAR is called Missing Not at Random (MNAR).

It is possible to test between MCAR and MAR. For data missing due to drop out, Diggle (1988) proposed a test for the hypothesis that the probability a unit drops out at time t_j is independent of the measurement on that unit up to time t_{j-1}. An alternative for general monotone data was developed by

Little (1988). It is not possible to test MNAR versus MAR since the data needed for such a test are, by definition, missing.

A closely related concept is ignorability of the missing data process. If the data are MAR and if the parameters of the complete data model are independent of those of the missing data mechanism, then likelihood inference of the observed data can ignore the missing data process. Suppose that the random variable $R = 1$ indicates that Y is observed, whereas $R = 0$ for missing Y. The information about Y that is present in X, Z, and R is summarized by the conditional distribution $P(Y|X, Z, R)$. Cases with missing Y; that is, with $R = 0$, do not provide any information about $P(Y|X, Z, R)$, and so we have only information to fit models for $P(Y|X, Z, R = 1)$. However, we need the distribution $P(Y|X, Z, R = 0)$ to model the missing Ys. Assuming that the missing data mechanism is ignorable corresponds to equating $P(Y|X, Z, R = 0) = P(Y|X, Z, R = 1)$ (Rubin, 1987).

The assumption of ignorability generally provides a natural starting point for analysis. If the assumption is clearly not reasonable (e.g., when data are censored), we may use other forms for $P(Y|X, Z, R = 0)$. The fact that $R = 0$ allows for the possibility that the $P(Y|X, Z, R = 1) \neq P(Y|X, Z, R = 0$; cf. Rubin, 1987, p. 205), so nonignorable nonresponse can be modeled by specifying $P(Y|X, Z, R = 0)$ different from $P(Y|X, Z, R = 1)$. The difference can be just a simple shift in the mean of the distribution (Van Buuren, Boshuizen, & Knook, 1999), but it may also consist of highly customized (selection, pattern mixture, shared parameter) models that mimic the nonresponse mechanism (Daniels & Hogan, 2008; Demirtas & Schafer, 2003; Little & Rubin, 2002). Daniels and Hogan (2008) suggest viewing the effects of alternative missing data assumption in terms of departures from MAR. A key requirement is that the assumed nonignorable model should be more reasonable and sensible than the model implied by the assumption of ignorability.

A somewhat different strategy to bypass the assumption of ignorability is to construct double robust estimators. An estimator is double robust if it remains consistent when either (but not necessarily both) a model for the missing data mechanism or a model for the distribution of the complete data is correctly specified (Bang & Robins, 2005; Scharfstein, Rotnitzky, & Robins, 1999). The approach uses inverse probability weighting, and its pros and cons with respect to multiple imputation have been the subject of debate (Kang & Schafer, 2007). The literature is now moving toward using the best of both worlds from inverse probability weighting and multiple imputation (Beunckens, Sotto, & Molenberghs, 2008; Carpenter, Kenward, & Vansteelandt, 2006).

10.3.4 Scale

Data can be measured on many types of scales: continuous (but are usually rounded to whole units), ordered categorical, unordered categorical, binary, semicontinuous (i.e., a mixture of a binary and a continuous variable), counts, censored (with known or unknown censoring points), truncated (with known or unknown truncation points), below the detection limit, bracketed response (e.g., obtained by a format that zooms in by posing successively more detailed questions), constrained by other data (e.g., a sum score or interaction term), and so on. In addition the data can take almost any distribution, including bimodal, skewed, and kurtotic shapes. Moreover, the relations can be highly nonlinear.

All these factors can occur in conjunction with multilevel data. The most advanced methods for dealing with missing data in a multilevel context invariably assume that variables follow a multivariate normal distribution. Though multiple imputation is generally robust to violations of the multivariate normality assumption (Schafer, 1997), advances could be made that respect the scale, the distribution, and nonlinear relations of the data.

10.3.5 Study Design

The study design determines the class of incomplete data models that can be usefully applied to the data. Popular designs that lead to hierarchical data include:

Multistage sample: A design where sampling progresses in a number of stages, for example, first sample from school, then sample classes within schools, and then sample pupils within classes. Missing data can occur at any stage of sampling, but usually only missing data in the level-1 outcomes are explicitly considered as missing data. This is a common design in the social sciences.

Longitudinal study with fixed occasions: A design where data are collected according to a number of planned visits. Missing data may result from missed visits (intermittent missing data) or panel attrition (drop out). This design is common in the biomedical field.

Longitudinal study, varying occasions: A design where the data are ordered according to time and nested within individuals. There is no such thing as a complete data vector. The number of observations per individuals may vary widely, can be as low as one, and can occur anywhere in time (Snijders & Bosker, 1999).

Planned missing data: A design where intentional missing data occur in the data as a consequence of the administration procedures. For example, the investigator could use matrix-sampling to minimize the number of questions posed to a student (Thomas & Gan, 1997). Missing data are an automatic part of the data. The percentage of missing data is typically large, sometimes over 75%.

File matching: A post-hoc procedure for combining two or more data sets measured on the same units. Missing data occur in the rows and in the columns since different data sources can measure different units on different attributes (Rässler, 2002; Rubin, 1986).

Relational databases: A common way for storing information on different types of units (e.g., customers, products, stores) as a set of linked tables. Missing data result from partial tables and imperfect links.

10.4 STRATEGIES TO DEAL WITH INCOMPLETE DATA

10.4.1 Prevention

The best solution to the missing data problem is not to have any. Consequently, the best strategy is to deal with unintentional missing data and to minimize their number. There are many factors that influence the response rate in social and medical studies: design of the study (number of variables collected, number and spacing of time repeated measures, follow-up time, missing data retrieval strategy), data collection method (mode of collection, intrusive measures, sensitivity of information collected, incentives, match of the interviewer and the respondent), measures (clarity, layout),

treatment burden (intensity of the intervention) and data entry coding errors. For more information, we refer to the appropriate literature (De Leeuw, Hox, & Dillman, 2008; McKnight et al., 2007; Stoop, 2005). When carefully planned and executed, prevention of missing data may substantially increase the completeness of the information.

10.4.2 Listwise Deletion

Listwise deletion (or complete case analysis) is the simplest and most popular way of dealing with missing data. Listwise deletion simply eliminates any incomplete record from the analysis. This is potentially a very wasteful strategy because valuable data are thrown away, especially when variables at the higher levels have missing data. If the missing data are confined to y_j and if the missing data mechanism is MAR, then listwise deletion followed by the appropriate likelihood-based analysis is unbiased. Note that any covariates that predict the missingness in y_j should be included into the model, even if they are of no scientific interest to the researcher. For missing data in W_j or Z_j, analysis of the complete cases will generally bias parameter estimates, even under MCAR (Little, 1992).

10.4.3 Last Observation Carried Forward

Last Observation Carries Forward (LOCF) is a technique applicable only to longitudinal data with drop out. The LOCF substitutes any missing y_j after drop out by the last observation. LOCF is popular for clinical trials in order to be able to perform an "intention to treat" analysis; that is, an analysis of the subject as randomized, irrespective of treatment compliance. However, LOCF makes the strong and often unrealistic assumptions that the response profile of the subject remains constant after dropping out of the study. The LOCF does not even work under MCAR (Molenberghs & Kenward, 2007). The magnitude and direction of this bias depend on the true but unknown treatment effects. In contrast to the widespread belief that LOCF leads to conservative tests, it is entirely possible that LOCF induces effects where none exist. Furthermore, because there is no distinction between the observed and the imputed data, LOCF artificially increases the amount of information in the data. This results in confidence intervals that are too short. All in all, the use of LOCF is discouraged (Lavori, 1992; Little & Yau, 1996).

10.4.4 Class Mean Imputation

Class mean imputation replaces each missing value with the class or cluster mean. The method is applicable to both longitudinal and nonlongitudinal data. Thus, class mean imputation substitutes the missing grade of a pupil by the average of the known grades of all pupils in the class. Just like LOCF, the method is unconditional on any other information from the pupil, so the method may distort relations between variables. Unless special methods are used to analyze the imputed data, the variability may be severely underestimated (Little & Rubin, 2002; Schafer & Schenker, 2000). All in all, class mean imputation can be as damaging as LOCF and should generally not be used.

10.4.5 Likelihood-Based Methods

Likelihood-based methods attempt to analyze the entire data without systematically biasing the conclusions of the subject

matter question. The method maximizes the likelihood function derived from the underlying model. If there are missing data, the likelihood function is restricted to the observed data only. If the missing data mechanism is ignorable, we may write the likelihood of the observed data $L(\theta|Y_{obs})$ as

$$L(\theta|Y_{obs}) = \int L(\theta|Y_{obs}, Y_{mis})dY_{mis} \quad (10.5)$$

where θ are the parameters of interest, and where $L(\theta|Y_{obs}, Y_{mis})$ is the likelihood of the hypothetically complete data. The observed data likelihood averages over the distribution of the missing data. The Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) maximizes $L(\theta|Y_{obs})$ by filling in the complete data sufficient statistics.

The linear mixed-effects model (Equation 10.1) subsumes repeated-measures ANOVA and growth curve models for longitudinal data as special cases. The model parameters can be estimated efficiently via likelihood-based methods. Laird and Ware developed an EM algorithm that can be used to fit the mixed linear model to longitudinal data (1982). Jennrich and Schluchter (1986) improved the speed of the method by Fisher scoring and Newton-Raphson. Currently, full-information maximum likelihood (FIML) is widely used to estimate the model parameters. Restricted maximum likelihood estimation (REML) is a closely related alternative that is less sensitive to small-sample bias of maximum likelihood (Fitzmaurice, Laird, & Ware, 2004; Verbeke & Molenberghs, 2000).

Software for fitting mixed models has the ability to handle unbalanced longitudinal data, where the response data y_j are observed at arbitrary time points for each subject. Missing data in y_j are ignored by the maximum likelihood and REML methods along with their values on W_j and Z_j. An advantage of the multilevel model for the analysis of longitudinal data is its ability to handle arbitrary time points. Missing values in W_j and Z_j are however problematic (Longford, 2008; Schafer, 1997). No generally applicable likelihood-based approach has yet been developed for the case of missing values in W_j and Z_j.

Despite the attractive properties of the multilevel model, likelihood-based methods should be used with some care when data are incomplete. First, the standard multilevel model implicitly assumes is that the missing data in the outcomes are MAR. This assumption can be suspect in some settings. For example, patients who drop out early from a trial often have slopes that differ from patients who stay in the trial. Another assumption is that the individual patient slopes have a common normal distribution. This assumption may not be realistic if drop out occurs. There is an active statistical literature on the problem of estimating the linear mixed model under MNAR situations (Daniels & Hogan, 2008).

In the case that the MAR assumption is correct, the factors that govern the probability of the missing data must be included into the multilevel model, for example, as covariates. Failing to do so may introduce biases in the estimate of the treatment effect. Note that this requirement complicates the interpretation of the complete-data model, and may lead to models that are impossible to estimate and more complex to interpret. Also, missing data problems may actually worsen if the additional covariate(s) contain missing values themselves.

Third, the missing data may increase the sensitivity of inferences to misspecification of the model for the complete data. Incorrectly assuming a linear relationship between an outcome and a covariate may lead to more serious bias when missingness depends on the value of the covariate than when it does not (Little, 2008). Zaidman-Zait and Zumbo (2005) performed simulations where the missing data mechanism depended on a person factor. Theoretically, including the person factor into the model should adequately deal with the missing data. However, they found bias in the MAR case and attribute that to the incorrect specification of the level-1 model.

Fourth, it is generally more difficult to derive appropriate standard errors if there are missing data. For example, the occurrence of missing data may destroy the block-diagonal structure of the information matrix in many repeated measure designs. Hence, the full matrix needs to be inverted, which can be time consuming (Little, 2008).

In summary, likelihood-based methods are the preferred approach to missing data if all of the following hold:

1. The missing data are confined to y_j,
2. The MAR assumption is plausible,
3. Any factors in the MAR mechanism are included into the multilevel model,
4. The multilevel model for the complete data is correctly specific.

If one or more of these conditions are not met, using likelihood methods for incomplete data could be problematic. Not much is yet known about the relative importance of each factor.

10.4.6 Multiple Imputation

The likelihood-based approach attempts to solve both the missing data and complete data problems in one step. An alternative strategy is to attack the problem in two steps: First solve the missing data problem by imputing the missing data, and then fit the complete data analysis on the imputed data. Such a modular approach breaks down the model complexity in each step. It is well known that the precision of the complete-data estimates is overestimated if no distinction is made between observed and imputed data. The solution to this problem is to use multiple imputation (MI), which can produce correct estimates of the sampling variance of the estimates of interest (Rubin, 1987, 1996).

10.5 IMPUTATION OF UNIVARIATE MISSING DATA IN y_j

10.5.1 Multilevel Imputation Algorithm

The linear mixed model formulation of the multilevel model is given by Equation 10.1: $y_j = X_j\beta + Z_j u_j + e_j$ with $u_j \sim N(0, \Omega)$ and $e_j \sim N(0, \sigma^2 I(n_j))$. In order to derive imputations under this model, we adopt a Bayesian approach. For complete data, the distribution of the parameters can be simulated by Markov chain Monte Carlo (MCMC) methods (Schafer & Yucel, 2002; Zeger & Karim, 1991). The main steps are:

1. Sample β from $p(\beta \mid y, u, \sigma^2)$
2. Sample u_j from $p(u \mid y, \beta, \Omega, \sigma^2)$
3. Sample Ω from $p(\Omega \mid u)$ (10.6)
4. Sample σ^2 from $p(\sigma^2 \mid y, \beta, u)$
5. Repeat step 1-4 until convergence

The rate of convergence of this Gibbs sampler depends on the magnitude of the correlation between the steps. Many variations on the above scheme have been proposed (Chib & Carlin, 1999; Cowles, 2002; Gelman, Carlin, Stern, & Rubin, 2004; Gelman, Van Dyk, Huang, & Boscardin, 2008).

Let us first consider the case where y contains missing data. Let y^{obs} represent the observed data and let y^{mis} be the missing data, so that $y = [y^{obs}, y^{mis}]$. If the MAR assumption is plausible, we can replace y by y^{obs} in the above steps, and simulate the parameter distribution using only the complete records. At the end, we append an additional step to generate imputations for the missing data:

6. Sample y^{mis} from $p(y^{mis} \mid y^{obs}, \beta, u, \Omega, \sigma^2)$. (10.7)

Under model Equation 10.1, we calculate imputations by drawing

$$e_j^* \sim N(0, \sigma^2) \tag{10.8}$$

$$y_j^* = X_j\beta + Z_j u_j + e_j^* \tag{10.9}$$

where all parameters that appear on the right are replaced by their values drawn under the Gibbs sampler.

The classic algorithm outlined above will not produce good imputations for incomplete predictors. A considerable advance in imputation quality is possible by using a slightly more general version of model Equation 10.1, where the within-cluster variance σ_j^2 is allowed to vary over the clusters. Kasim and Raudenbush (1998) proposed a Gibbs sampler for this heterogeneous model. They specify

$$p(\sigma_j^2 \mid \sigma_0^2, \phi) \sim \frac{\sigma_0^2 \chi^2_{1/\phi}}{\phi} \tag{10.10}$$

where σ_0^2 and ϕ are hyperparameters. The hyperparameter σ_0^2 describes the location of prior belief about residual variance σ_j^2 in the conjugate prior distribution for σ_j^2. The hyperparameter ϕ is a measure of variability of the variances σ_j^2. Both σ_0^2 and ϕ are also updated within the Gibbs sampler. The algorithm was implemented in R by Roel de Jong, where $\sigma_j^2 = 1$ and $\phi = 1$ are used as starting parameters. Below, we will refer to this method as multilevel imputation (ML).

10.5.2 Simulation Study

Data with a multilevel structure were generated according to the model $y_{ij} = 0.5 z_{ij} + u_j + e_{ij}$ with $e_j \sim N(0, \sigma^2)$ and $u_j \sim N(0, \Omega)$. This model is a special case of Equation 10.1 and 10.2), where $X_j = Z_j = (1, z_{ij})$ with $i = 1,\ldots,n_j$ is the $n_j \times 2$ data matrix of class j, where $\Omega = \text{diag}(\omega^2,0)$, where $\beta = (0,0.5)^T$ is a 2×1 vector of fixed parameters, and where W_j is the identity matrix. We varied the variance parameters (σ^2, ω^2) in pairs as {(0.75,0.00), (0.65,0.10), (0.45,0.30), (0.25, 0.50)}. Since variable z_{ij} was drawn as $z_{ij} \sim N(0,1)$, the intraclass correlation coefficient (ICC) under the stated model equals ω^2, so the ICC effectively varies between 0.0 and 0.5. We fixed the total number of respondents to 1,200. The number of classes was chosen 12, 24, and 60, yielding 100, 50, and 20 respondents per class, respectively.

Two missing data mechanisms were specified: Y and Z. Mechanism Y generates 50% missing data in y_{ij} under MAR. For values of $z_{ij} < 0$, the nonresponse probability in y_{ij} is 10%. For $z_{ij} \geq 0$, this probability is 90%. Vice versa, mechanism Z generates 50% missing data in z_{ij} under MAR given y_{ij}. For values of $y_{ij} < 0$, the nonresponse probability is 10%. For $y_{ij} \geq 0$, the probability is 90%.

The following methods for handling the missing data were used:

- Complete Case Analysis (CC). This method removes any incomplete records before analysis, also known as listwise deletion.
- Multiple Imputation Flat File (FF). This method multiple imputes missing data while ignoring any clustering structure in the data by standard linear regression imputation.
- Multiple Imputation Separate Classes (SC). This method multiple imputes missing data by treating the cluster allocation as a fixed factor, so that differences in intercepts between classes are modeled.
- Multiple Imputation Multilevel Imputation (ML). This method applies the Gibbs sampler as described above to generate multiple imputations from posterior of the missing data given the observed data.

The number of multiple imputation was fixed to 5. Parameter estimates are pooled using Rubin's rules (Rubin, 1987; Rubin, 1996). The complete-data model was fitted by the `lmer()` function in R package `lme4` (Pinheiro & Bates, 2000).

10.5.3 Results

Table 10.1 contains results of the simulations. When missing data are confined to y_{ij}, then CC is unbiased for both the fixed and random parameters, as expected. Method FF is unbiased in the fixed parameters, but severely biased in the random parameters for clustered data (i.e., when $\omega^2 > 0$). Method SC produces unbiased estimates of both the fixed and random parameters. Note that this is related to the fact that the model that generated the data included only random intercepts and no random slopes. Also, method ML is unbiased in both the fixed and random parameters.

If missing data occur in z_{ij}, the results are drastically different. The estimates under CC are severely biased, both for the fixed and random parameters. Thus even under MAR, the standard practice of eliminating incomplete records can produce estimates that are plainly wrong. Of the three imputation methods, SC and ML yield estimates that are close to population values, FF is generally less successful. Method SC had computational problems for small cluster sizes ($n_j = 20$) because the number of observations in the cluster that remain after missing data were created could become too low (≤3). The FF and ML methods are insensitive to this problem since they combine information across clusters.

Table 10.2 contains estimates of the coverage of the 95% confidence interval for the fixed parameters. The number of replications used is equal to 100, so the simulation standard error is $\sqrt{(0.95(1-0.95)/100)} = 2.2\%$. For missing data in y_{ij}, CC has appropriate coverage. However, coverage for missing data in z_{ij} is dismal, so statistical inferences are unwarranted under incomplete z_{ij}. The FF is generally not well calibrated, and may achieve both under- or overcoverage depending on the amount of clustering. The SC has appropriate coverage of β_0, but coverage is suboptimal for β_x. The ML has appropriate coverage for larger cluster sizes for both β_0 and β_x. Coverage for small cluster sizes is however less than ideal, though still reasonable.

This section addressed the properties of four methods for dealing with univariate

TABLE 10.1

MAR Missing Data in Either y_{ij} or z_{ij}

	J	n_j	β_0	CC	FF	SC	ML	β_x	CC	FF	SC	ML	σ^2	CC	FF	SC	ML	ω^2	CC	FF	SC	ML
Y																						
A	12	100	0.00	0.00	0.00	0.00	0.01	0.50	0.51	0.50	0.50	0.50	0.75	0.75	0.75	0.75	0.76	0.00	0.00	0.00	0.02	0.02
B	12	100	0.00	0.00	0.01	−0.02	0.01	0.50	0.50	0.49	0.50	0.50	0.65	0.65	0.71	0.65	0.65	0.10	0.10	0.03	0.12	0.11
C	12	100	0.00	−0.01	0.00	0.01	0.00	0.50	0.50	0.50	0.50	0.50	0.45	0.45	0.63	0.45	0.45	0.30	0.30	0.08	0.33	0.31
D	12	100	0.00	0.03	−0.01	0.00	0.02	0.50	0.50	0.49	0.50	0.50	0.25	0.25	0.55	0.25	0.25	0.50	0.49	0.13	0.51	0.51
E	24	50	0.00	0.00	0.00	0.00	0.00	0.50	0.49	0.50	0.50	0.50	0.75	0.74	0.74	0.75	0.75	0.00	0.01	0.00	0.03	0.02
F	24	50	0.00	0.02	0.00	0.00	0.00	0.50	0.51	0.50	0.50	0.50	0.65	0.65	0.71	0.65	0.66	0.10	0.11	0.03	0.12	0.12
G	24	50	0.00	0.01	0.00	0.00	−0.01	0.50	0.50	0.51	0.50	0.50	0.45	0.44	0.62	0.45	0.45	0.30	0.30	0.07	0.32	0.31
H	24	50	0.00	−0.02	0.00	−0.01	−0.02	0.50	0.50	0.50	0.51	0.51	0.25	0.25	0.57	0.25	0.25	0.50	0.48	0.13	0.48	0.50
I	60	20	0.00	0.00	−0.01	0.00	−0.01	0.50	0.49	0.50	0.50	0.50	0.75	0.74	0.74	0.74	0.74	0.00	0.01	0.00	0.08	0.03
J	60	20	0.00	−0.01	0.01	0.00	0.00	0.50	0.50	0.51	0.50	0.50	0.65	0.65	0.71	0.65	0.65	0.10	0.10	0.03	0.17	0.12
K	60	20	0.00	0.00	−0.01	0.00	0.02	0.50	0.50	0.50	0.50	0.50	0.45	0.45	0.64	0.45	0.45	0.30	0.29	0.07	0.36	0.31
L	60	20	0.00	−0.01	0.01	0.00	−0.01	0.50	0.50	0.50	0.49	0.49	0.25	0.25	0.57	0.25	0.25	0.50	0.49	0.13	0.53	0.49
Z																						
A	12	100	0.00	−0.53	0.00	0.00	0.00	0.50	0.32	0.49	0.49	0.48	0.75	0.49	0.75	0.75	0.74	0.00	0.00	0.00	0.00	0.00
B	12	100	0.00	−0.49	0.00	0.00	−0.01	0.50	0.34	0.48	0.49	0.48	0.65	0.44	0.66	0.65	0.66	0.10	0.05	0.08	0.11	0.10
C	12	100	0.00	−0.36	0.01	0.01	0.01	0.50	0.40	0.45	0.50	0.49	0.45	0.34	0.50	0.45	0.46	0.30	0.20	0.23	0.31	0.30
D	12	100	0.00	−0.22	−0.01	−0.01	−0.01	0.50	0.43	0.40	0.50	0.50	0.25	0.21	0.34	0.25	0.25	0.50	0.39	0.42	0.48	0.52
E	24	50	0.00	−0.53	0.00	0.00	0.00	0.50	0.33	0.50	0.48	0.48	0.75	0.49	0.75	0.75	0.74	0.00	0.00	0.00	0.01	0.01
F	24	50	0.00	−0.49	0.00	0.00	−0.01	0.50	0.35	0.48	0.50	0.47	0.65	0.45	0.67	0.65	0.66	0.10	0.06	0.07	0.10	0.10
G	24	50	0.00	−0.39	−0.01	−0.01	0.01	0.50	0.39	0.44	0.50	0.49	0.45	0.33	0.51	0.45	0.46	0.30	0.20	0.23	0.30	0.29
H	24	50	0.00	−0.23	−0.02	0.00	0.00	0.50	0.43	0.40	0.50	0.49	0.25	0.21	0.35	0.25	0.25	0.50	0.41	0.39	0.50	0.50

(*Continued*)

TABLE 10.1

MAR Missing Data in Either y_{ij} or z_{ij} (*Continued*)

	J	n_j	β_0	CC	FF	SC	ML	β_x	CC	FF	SC	ML	σ^2	CC	FF	SC	ML	ω^2	CC	FF	SC	ML
I	60	20	0.00	−0.53	0.00	−0.01	−0.01	0.50	0.33	0.50	0.47	0.48	0.75	0.49	0.74	0.74	0.73	0.00	0.00	0.00	0.02	0.01
J	60	20	0.00	−0.50	0.00	0.00	−0.01	0.50	0.34	0.49	0.49	0.48	0.65	0.44	0.66	0.65	0.65	0.10	0.05	0.08	0.12	0.09
K	60	20	0.00	−0.41	−0.01	#	−0.01	0.50	0.38	0.44	#	0.47	0.45	0.33	0.52	#	0.47	0.30	0.18	0.25	#	0.27
L	60	20	0.00	−0.26	−0.01	#	0.00	0.50	0.42	0.40	#	0.49	0.25	0.20	0.35	#	0.27	0.50	0.39	0.41	#	0.46

Notes: Average estimates of fixed (β_0, β_x) and random variance (σ^2, ω^2) parameters in four methods for handling missing data (CC = complete case analysis, FF = MI flat file, SC = MI separate group, ML = MI multilevel).

solution could not be calculated due to almost empty classes.

TABLE 10.2

Coverage (in Percentage) of the True Values by the 95% Confidence Interval for Fixed Parameter Estimates Under Four Methods for Treating Missing Data in *Y* or *Z*, Respectively

	J	n_j	β_0	CC	FF	SC	ML	β_x	CC	FF	SC	ML
Y												
A	12	100	95	96	72	90	90	95	96	73	72	90
B	12	100	95	89	69	96	87	95	96	82	76	91
C	12	100	95	94	71	94	91	95	97	98	70	93
D	12	100	95	94	68	94	97	95	94	100	78	91
E	24	50	95	95	71	91	87	95	97	66	68	88
F	24	50	95	96	73	90	89	95	97	76	63	87
G	24	50	95	92	63	93	88	95	96	90	66	94
H	24	50	95	91	73	94	95	95	96	95	72	87
I	60	20	95	98	66	92	84	95	98	73	69	90
J	60	20	95	99	64	88	88	95	93	71	68	89
K	60	20	95	97	67	88	98	95	97	79	76	86
L	60	20	95	92	66	96	88	95	97	89	73	87
Z												
A	12	100	95	0	88	92	95	95	0	84	84	93
B	12	100	95	0	84	94	87	95	2	83	85	94
C	12	100	95	25	82	90	94	95	23	49	86	94
D	12	100	95	75	91	91	92	95	39	5	87	95
E	24	50	95	0	88	93	90	95	0	94	80	87
F	24	50	95	0	88	99	95	95	1	78	84	87
G	24	50	95	5	96	96	95	95	11	25	94	91
H	24	50	95	54	91	94	94	95	29	1	94	94
I	60	20	95	0	91	92	89	95	0	77	78	85
J	60	20	95	0	87	95	98	95	1	83	86	83
K	60	20	95	0	90	#	96	95	2	35	#	79
L	60	20	95	17	88	#	91	95	16	1	#	85

Notes: CC = complete case analysis, FF = MI flat file, SC = MI separate group, ML = MI multilevel.
solution could not be calculated due to almost empty classes.

missing data within a multilevel context. The CC method is easy and works well under MAR when missing data are restricted to y_{ij}. However, the performance CC with z_{ij} missing at random is bad. We therefore recommend against CC if many z_{ij} are missing. An alternative is to apply multiple imputation. Three such methods were studied. The overall best performance was obtained by the ML Gibbs sampling method.

10.6 MULTIVARIATE MISSING DATA IN y_j AND z_j

10.6.1 General Approach

Missing data may also occur in y_{ij} and z_{ij} simultaneously. The present section deals with the case where both y_{ij} and z_{ij} are incomplete. There are two general approaches to impute multivariate missing data: Joint

Modeling (JM) and Fully Conditional Specification (FCS).

Joint modeling partitions the observations into groups of identical missing data patterns, and imputes the missing entries within each pattern according to a joint model for all variables. The first such model was developed for the multivariate normal model (Rubin & Schafer, 1990). Schafer (1997) extended this line and developed sophisticated JM methods for generating multivariate imputations under the multivariate normal, the log-linear, and the general location model. This work was extended to include multilevel models (Schafer & Yucel, 2002; Yucel, 2008).

The fully conditional specification imputes data on a variable-by-variable basis by specifying an imputation model per variable. The FCS is an attempt to specify the full multivariate distribution of the variables by a set of conditional densities for each incomplete variable. This set of densities is used to impute each variable by iteration, where we start from simple initial guesses. Though convergence can only be proved in some special cases, the method has been found to work well in practice (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; Van Buuren et al., 1999; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006). The R `mice` package (Van Buuren & Groothuis-Oudshoorn, 2000) enjoys a growing popularity. Van Buuren (2007) provides an overview of the similarities and contrasts of JM and FCS.

10.6.2 Simulation Study

Using the same complete-data model as before, we created missing data in both x_{ij} and y_{ij} by applying mechanisms Y and Z each to a random split of the data. For missing z_{ij} the procedure is identical to that given before. For missing y_{ij}, the procedure is reversed. For values of $z_{ij} < 0$, the nonresponse probability in y_{ij} is 90%. For $z_{ij} \geq 0$, this probability is 10%. Thus, many high z_{ij} and many low y_{ij} will be missing.

We created five multiple imputed data sets with `mice` using the three imputation methods. The number of iterations in `mice` was fixed to 20.

10.6.3 Results

Table 10.3 contains the parameter estimates averaged over 100 simulations. Complete case (CC) analysis severely biases the estimates of the intercept term β_0 and the within-group variance σ^2, especially when the clustering is weak. Methods FF and SC have a somewhat better performance for the fixed effects, and behave differently for the variance estimates. The best overall method is ML, but note that ML is not yet ideal since β_0 is biased slightly upward while β_x is biased slightly downward. No systematic bias appears to be present in the variance estimates, so ML seems to recover the multilevel structure present in the original data quite well.

Table 10.4 contains the accompanying coverage percentages. The best method is ML, but none of the methods is really satisfactory. Trouble cases include A, E, and I, where $\omega^2 = 0$. The Gibbs sampler can get stuck if there is no between-cluster variation (Gelman et al., 2008), so this might be a reason for the low coverage. It also appears to be difficult to get appropriate coverage for small cluster sizes.

The simulations suggest that FCS is a promising option for imputing incomplete

TABLE 10.3

MAR Missing Data in both y_{ij} and z_{ij}. Average Estimates of Fixed (β_0, β_x) and Random Variance (σ^2, ω^2) Parameters in Four Methods for Handling Missing Data

	J	n_j	β_0	CC	FF	SC	ML	β_x	CC	FF	SC	ML	σ^2	CC	FF	SC	ML	ω^2	CC	FF	SC	ML
YZ																						
A	12	100	0.00	−0.46	−0.16	−0.16	0.09	0.50	0.40	0.39	0.38	0.46	0.75	0.55	0.77	0.77	0.75	0.00	0.01	0.00	0.05	0.01
B	12	100	0.00	−0.42	−0.16	−0.18	0.09	0.50	0.41	0.39	0.38	0.44	0.65	0.49	0.75	0.69	0.68	0.10	0.06	0.01	0.16	0.10
C	12	100	0.00	−0.30	−0.15	−0.15	0.08	0.50	0.44	0.41	0.41	0.46	0.45	0.36	0.69	0.47	0.48	0.30	0.24	0.02	0.34	0.28
D	12	100	0.00	−0.14	−0.13	−0.17	0.04	0.50	0.48	0.40	0.41	0.47	0.25	0.22	0.70	0.28	0.27	0.50	0.47	0.03	0.54	0.49
E	24	50	0.00	−0.46	−0.16	−0.16	0.08	0.50	0.40	0.39	0.38	0.44	0.75	0.55	0.77	0.76	0.76	0.00	0.01	0.00	0.11	0.01
F	24	50	0.00	−0.42	−0.17	−0.16	0.08	0.50	0.41	0.40	0.39	0.45	0.65	0.48	0.77	0.67	0.68	0.10	0.07	0.01	0.22	0.10
G	24	50	0.00	−0.32	−0.18	−0.15	0.07	0.50	0.45	0.38	0.39	0.44	0.45	0.35	0.73	0.48	0.49	0.30	0.20	0.02	0.38	0.29
H	24	50	0.00	−0.20	−0.15	−0.15	0.09	0.50	0.47	0.37	0.39	0.47	0.25	0.21	0.72	0.28	0.28	0.50	0.47	0.03	0.59	0.52
I	60	20	0.00	−0.47	−0.17	#	0.08	0.50	0.39	0.37	#	0.44	0.75	0.52	0.78	#	0.73	0.00	0.01	0.00	#	0.02
J	60	20	0.00	−0.44	−0.14	#	0.09	0.50	0.41	0.39	#	0.44	0.65	0.49	0.76	#	0.67	0.10	0.07	0.01	#	0.09
K	60	20	0.00	−0.36	−0.17	#	0.09	0.50	0.42	0.38	#	0.43	0.45	0.35	0.75	#	0.50	0.30	0.19	0.02	#	0.28
L	60	20	0.00	−0.23	−0.17	#	0.08	0.50	0.46	0.40	#	0.44	0.25	0.21	0.70	#	0.30	0.50	0.40	0.03	#	0.47

Notes: CC = complete case analysis, FF = MI flat file, SC = MI separate group, ML = MI multilevel.
solution could not be calculated due to almost empty classes.

TABLE 10.4

Coverage (in Percentage) of the True Values by the 95% Confidence Interval for Fixed Parameter Estimates Under Four Methods for Treating Missing Data in Both *Y* and *X*

	J	n_j	β_0	CC	FF	SC	ML	β_x	CC	FF	SC	ML
YZ												
A	12	100	95	0	5	42	37	95	46	29	27	85
B	12	100	95	2	18	64	81	95	55	23	22	77
C	12	100	95	45	25	83	89	95	71	32	26	76
D	12	100	95	83	38	85	90	95	88	29	17	82
E	24	50	95	0	6	39	37	95	48	28	30	64
F	24	50	95	0	9	56	79	95	56	30	27	67
G	24	50	95	16	21	76	84	95	75	25	15	55
H	24	50	95	69	28	81	87	95	82	28	13	72
I	60	20	95	0	1	#	34	95	42	19	#	55
J	60	20	95	0	13	#	50	95	53	24	#	57
K	60	20	95	1	12	#	73	95	52	22	#	42
L	60	20	95	28	17	#	82	95	76	27	#	43

Notes: CC = complete case analysis, FF = MI flat file, SC = MI separate group, ML = MI multilevel.
solution could not be calculated due to almost empty classes.

multilevel data. The FCS used in conjunction with multiple multilevel imputation is a considerable improvement over standard practice. The methodology is not yet ideal however, and further optimization and tuning is needed.

10.7 CONCLUSIONS

Multilevel data can be missing at different levels. Variables in which missing data occur can have different roles in the analysis. The optimal way to deal with missing data depends on both the level and the role of the variable in the analysis.

Multilevel models are often presented in the form of the linear mixed model Equation 10.1. This formulation complicates conceptualization of the missing data problem because the same variable can appear at different places. It is useful to write the multilevel model as a slopes-as-outcomes model Equation 10.2, which clearly separates the variables at the different levels. Section 10.2 describes how Equations 10.1 and 10.2 are related.

Missing data can occur in y_j (level-1 outcomes), Z_j (level-1 predictors) are W_j (level-2 predictors) and j (class variable). The problem of missing data in y_j has received considerable attention. The linear multilevel model provides an efficient solution to this problem if the data are missing at random and if the model fits the data. There is a large literature on what can be done if the MAR assumption is suspect, or when models for other outcome distributions are needed. By comparison, the problem of missing data in Z_j, W_j and j received only scant attention. The usual solution is to remove any incomplete records, which is wasteful and could bias the estimates of interest. Several fixes

have been proposed, but none of these have yet gained wide use.

Other questions that need to be addressed are less particular to the multilevel setting: the missing data pattern, the missing data mechanism, the measurement scales used, and the study design. A successful attack on a given incomplete data problem depends on our capability to address these factors for the application at hand.

Section 10.3 outlines six strategies. Quick fixes like listwise deletion, last observation carried forward and class mean imputation will only work in a limited set of circumstances and are generally discouraged. Prevention, likelihood-based methods, and multiple imputation are methodologically sound approaches based on explicit assumptions about the missing data process.

Multiple imputation is a general statistical technique for handling incomplete data problems. Some work on MI in multilevel setting has been done, but many open issues remain. We performed a simulation study with missing data in y_{ij} or z_{ij}, and compared complete case analysis with three MI techniques: flat file (FF) imputation that ignores the multilevel structure, separate clusters (SC) imputation that includes a group factor, and multilevel (ML) imputation by means of the Gibbs sampler. Complete case analysis was found to be a bad strategy with missing data in z_{ij}. The best imputation technique was ML. A second simulation addressed the question of how the methods behave when missing data occur simultaneously in y_{ij} or z_{ij}. Though its performance is not yet ideal, multiple imputation by ML within the FCS framework considerably improves upon standard practice.

Simulation is not reality. The missing data mechanisms we have used in the simulation have a considerable amount of missing information, and are probably more extreme than those encountered in practice. The simulations are still useful though. Differences between methods in absolute terms may be smaller in practice, but the best methods will continue to dominate others in less extreme situations. All other things being equal, we therefore prefer to use imputation methods that performs best "asymptotically" in extreme situations.

Since ML requires more work than complete case analysis it would be useful to have clear-cut rules that say when doing ML is not worth the trouble. No such rules have yet been devised. This would be a useful area of further research. Another area for research would be to further optimize and tune the ML imputation method to the multivariate missing data problem. For example, taking alternative distributions for within-cluster residual variance σ_j^2 could improve performance. The current implementation of the method uses a full Gibbs sampler. Though the algorithm is robust, it is not particularly fast. Adding parameter expansion (Gelman et al., 2008) could be useful to prevent the Gibbs sampler from getting stuck at the border of the parameter space at $\omega^2 = 0$. Computations could be speeded up, for example by obtaining marginal maximum likelihood estimates of β and Ω using numerical integration via Gauss-Hermite (Pinheiro & Bates, 2000). Extensions toward higher level models are also possible (Yucel, 2008). Finally, we can classify missing data problems by combining the answers on the five questions posed in Section 10.3. Classification of the combinations opens up a whole research agenda with many white spots.

REFERENCES

Bang, K., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics, 61,* 962–972.

Beunckens, C., Molenberghs, G., Thijs, H., & Verbeke, G. (2007). Incomplete hierarchical data. *Statistical Methods in Medical Research, 16,* 457–492.

Beunckens, C., Sotto, C., & Molenberghs, G. (2008). A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics and Data Analysis, 52,* 1533–1548.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage Publications, Inc.

Carpenter, J., & Goldstein, H. (2004). Multiple imputation in MLwiN. MLwin newsletter [On-line]. Available: http://www.lshtm.ac.uk/msu/missingdata/papers/newsletterdec04.pdf

Carpenter, J., Kenward, M. G., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society. Series A: Statistics in Society, 169,* 571–584.

Cheung, M. W. L. (2007). Comparison of methods of handling missing time-invariant covariates in latent growth models under the assumption of missing completely at random. *Organizational Research Methods, 10,* 609–634.

Chib, S., & Carlin, B. P. (1999). On MCMC sampling in hierarchical longitudinal models. *Statistics and Computing, 9,* 26.

Cochran, W. G., & Cox, G. M. (1957). *Experimental designs*. New York, NY: John Wiley & Sons, Ltd.

Cowles, M. K. (2002). MCMC sampler convergence rates for hierarchical normal linear models: A simulation approach. *Statistics and Computing, 12,* 377–389.

Daniels, M. J., & Hogan, J. W. (2008). *Missing data in longitudinal studies. Strategies for Bayesian modeling and sensitivity analysis*. Boca Raton, FL: Chapman & Hall/CRC.

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). *International handbook of survey methodology*. New York, NY: Lawrence Erlbaum Associates.

de Leeuw, J., & Meijer, E. (2008). *Handbook of multilevel analysis*. New York, NY: Springer.

Demirtas, H., & Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine, 22,* 2553–2575.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology,* 1–38.

Diggle, P. J. (1989). Testing for random dropouts in repeated measurement data. *Biometrics, 45,* 1258.

Dodge, Y. (1985). *Analysis of experiments with missing data*. New York, NY: John Wiley & Sons, Ltd.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York, NY: John Wiley & Sons, Ltd.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, UK: Chapman and Hall.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel and hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Van Dyk, D. A., Huang, Z., & Boscardin, W. J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics, 17,* 95–122.

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. *Educational and Psychological Measurement, 63,* 204–238.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London, UK: Charles Griffin & Company Ltd.

Jacobusse, G. W. (2005). WinMICE user's manual [Computer software]. Leiden, The Netherlands: TNO Quality of Life.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 42,* 805–820.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science, 22,* 523–539.

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics, 23,* 93–116.

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

Lavori, P. W. (1992). Clinical trials in psychiatry: Should protocol deviation censor patient data? (with discussion). *Neuropsychopharmacology, 6,* 39–63.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83,* 1198–1202.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87,* 1227–1237.

Little, R. J. A. (2008). Selection and pattern-mixture models. In G. M. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis: A handbook of modern statistical methods* (pp. 409–431). New York, NY: Wiley.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data.* (2nd ed.). New York: Wiley.

Little, R. J. A., & Yau, L. (1996). Intent-to-treat analysis for longitudinal studies with drop-outs. *Biometrics, 52,* 1324–1333.

Longford, N. T. (2005). *Missing data and small-area estimation*. New York, NY: Springer.

Longford, N. T. (2008). Missing data. In J. De Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 377–399). New York, NY: Springer.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data. A gentle introduction*. New York, NY: Guilford Press.

Molenberghs, G., & Kenward, M. G. (2007). *Missing data in clinical studies*. Chichester, UK: John Wiley & Sons, Ltd.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. New York, NY: Springer.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (Version V5.1) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Petrin, R. A. (2006). Item nonresponse and multiple imputation for hierarchical linear models. Paper presented at the annual meeting of the American Sociological Association, Montreal Convention Center, Montreal, Quebec, Canada [On-line]. Available: http://www.allacademic.com/meta/p102126_index.html

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27,* 85–95.

Rasbash, J., Steel, F., Browne, W., & Prosser, B. (2005). A user's guide to MLwiN Version 2.0 [Computer software]. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Rässler, S. (2002). *Statistical matching. A frequentist theory, practical applications, and alternative Bayesian approaches*. New York, NY: Springer.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2008). *HLM 6* [Computer software]. Chicago, IL: SSI Software International.

Roudsari, B., Field, C., & Caetano, R. (2008). Clustered and missing data in the US National Trauma Data Bank: Implications for analysis. *Injury Prevention, 14,* 96–100.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–590.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business Economics and Statistics, 4,* 87–94.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

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

Rubin, D. B., & Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *1990 Proceedings of the Statistical Computing Section, American Statistical Association,* 83–88.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London, UK: Chapman & Hall.

Schafer, J. L., & Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 449,* 144–154.

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

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association, 94,* 1096–1120.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London, UK: Sage Publications Ltd.

SPSS Inc. (2008). *SPSS 17.0 base user's guide* [Computer software]. Chicago, IL: SPSS Inc.

StataCorp LP (2008). *STATA 10 user's guide* [Computer software]. College Station, TX: STATA Press.

Stoop, I. A. L. (2005). *The hunt for the last respondent: Nonresponse in sample surveys*. Rijswijk, The Netherlands: Sociaal en Cultureel Planbureau.

Thomas, N., & Gan, N. (1997). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics, 22,* 425–445.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16,* 219–242.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18,* 681–694.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76,* 1049–1064.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. (PG/VGZ/00.038 ed.) Leiden, The Netherlands: TNO Quality of Life.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Yucel, R. M. (2008). Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A, 366,* 2389–2403.

Yucel, R. M., Schenker, N., & Raghunathan, T. E. (2006). Multiple imputation for incomplete multilevel data with SHRIMP. Online citation [On-line]. Available: http://www.umass.edu/family/pdfs/talkyucel.pdf

Zaidman-Zait, A., & Zumbo, B. D. (2005). Multilevel (HLM) models for modeling change with incomplete data: Demonstrating the effects of missing data and level-1 model mis-specification. Paper presented at the Hierarchical Linear Modeling (SIG) of the American Educational Research Association conference April 2005 in Montreal, Quebec, Canada. [On-line].

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86,* 79–86.

Zhang, D. (2005). *A Monte Carlo investigation of robustness to nonnormal incomplete data of multilevel modeling.* College Station, TX: Texas A&M University.