CHAPTER 12

Multiple Imputation With PAN

Joseph L. Schafer

Missing values are a nuisance in many research efforts but especially so in the collection and analysis of longitudinal data. Multiple occasions bring greater opportunities for missed measurements. Fortunately, missing data is one area where statisticians have made substantial progress in recent years. In this chapter, I present a strategy for analyzing incomplete longitudinal data by multiple imputation (Rubin, 1987; Schafer, 1997a).

Missing data pose a difficulty because the overwhelming majority of paradigms and software for statistical analysis assume that the input data are complete. For this reason, the quickest and most convenient method for handling incomplete observations is case deletion, that is, ignoring participants with missing information. Case deletion suffers from a number of serious drawbacks, which have been well documented (e.g., Little & Rubin, 1987). For multivariate analyses involving a large number of items case deletion can be very inefficient, discarding an unacceptably high proportion of participants; even if the per-item rates of missingness are low, few participants may have complete data for all items. Moreover, case deletion leads to valid inferences in general only when missing data are missing completely at random (MCAR), in the sense that the discarded cases are like a random subsample of all cases. If the discarded cases differ systematically from the rest, then the resulting estimates may have potentially serious bias.

A natural alternative to case deletion is *imputation*, the practice of replacing missing data with plausible values. Various forms of imputation have been applied in federal surveys and censuses for decades (Madow, Nisselson, & Olkin, 1983). Imputation has been the survey statistician's method of choice for handling *item nonresponse*, situations in which a participant provides some infor-

This research was supported by Grant 1-P50-DA10075 from the National Institute on Drug Abuse and by Grant 2R44CA65147-02 from the National Cancer Institute. I extend special thanks to John Graham for providing data from the Adolescent Alcohol Prevention Trial and advice on their analysis.

mation but fails to respond to one or more individual items on a questionnaire. Imputation is attractive because it apparently solves the missing-data problem at the outset; once the missing values have been imputed, the data set can be summarized and analyzed by familiar complete-data methods. Another attractive feature of imputation is its efficiency: Unlike case deletion, imputation allows one to make full use of the data at hand.

Methods of imputation range from simple procedures, such as mean substitution—replacing each missing value with the observed mean for that variable—to elaborate hot-deck algorithms that jointly replace missing items with data obtained from donor cases chosen to match the original on selected items (e.g., Bailey, Chapman, & Kasprzyk, 1985). In longitudinal data sets with substantial participant-to-participant variation, analysts have sometimes filled in missed measurements by linear interpolation, extrapolation, or "last value carried forward." Unless great care is taken, these ad hoc imputation procedures may seriously distort important aspects of the distribution of a variable or its relationships with other variables. In general, it is desirable for the distribution of imputed values to resemble the distribution of the observed values, particularly with respect to intervariable relationships.

Even if an imputation method successfully preserves important aspects of the data distributions, a potentially serious problem remains: Imputation adds fictitious information to a data set. If imputed values are treated the same way as observed values in subsequent analyses, then the resulting inferences will be artificially precise, because the imputed values are imperfect proxies for the data they represent. With single imputation, there is no simple way to reflect uncertainty in the imputed values. In response, Rubin (1987, 1996) proposed the method of multiple imputation, by which each missing value is represented by a set of $m > 1$ simulated values. Let $Y = (Y_{obs}, Y_{mis})$ denote a generic data set, in which Y_{obs} is the observed part and Y_{mis} is the missing part. Multiple imputation replaces Y_{mis} with a set of simulated draws $Y_{mis}^{(1)}, Y_{mis}^{(2)}, \ldots, Y_{mis}^{(m)}$ from a predictive probability distribution $P(Y_{mis}|Y_{obs})$ arising from a model. After multiple imputation, one has m simulated complete data sets, $Y^{(j)} = (Y_{obs}, Y_{mis}^{(j)})$, $j = 1, 2, \ldots, m$, which are analyzed with standard complete-data methods. The results are then combined, using simple arithmetic rules, to produce overall estimates and standard errors that account for missing-data uncertainty. I reviewed these rules (Schafer, 1997a) and demonstrate them in the example near the end of this chapter.

The key idea of multiple imputation is that it treats missing data as an explicit source of random variability over which to be averaged. The process of creating imputations, analyzing the imputed data sets, and combining the results is a Monte Carlo version of averaging the statistical results over the predictive distribution $P(Y_{mis}|Y_{obs})$. In practice, a large number of multiple impu-

tations are not required; sufficiently accurate results can often be obtained with $m \leq 10$.

Carrying out multiple imputation requires two sets of assumptions. First, one must propose a model for the distribution of Y. This data model should be plausible and should bear some relation to the type of analysis to be performed. For example, one could assume that the variables in the data set are jointly normally distributed. In the case of longitudinal analyses the model should be capable of preserving the correlation structure and time trends within individuals. The second set of assumptions pertains to the manner in which the missing values became missing. It is most common to assume that the missing data are missing at random (MAR) in the technical sense defined by Rubin (1976), which means that the probabilities of missingness may depend on the observed values Y_{obs} but not on the missing data Y_{mis}. The MAR assumption is primarily a mathematical convenience that allows one to perform imputation without explicitly modeling the missing-data mechanism. In practice, MAR is essentially untestable; it cannot be verified or contradicted by examination of the observed data. If the assumption seems *prima facie* implausible, then alternative procedures can be developed by modeling the probabilities of missingness. General techniques and software for creating multiple imputations under non-MAR models have not yet been developed; this is an important area for future research. Further discussion on the plausibility and ramifications of MAR was given by Little and Rubin (1987); Graham, Hofer, and Piccinin (1994); and Schafer (1997a).

Multiple imputation is not the only principled method for handling missing data. For parametric models, a main competitor is the technique of direct maximum likelihood, sometimes called *raw* or *full-information* maximum likelihood, which maximizes a likelihood function on the basis of the observed data Y_{obs} alone. This likelihood function may be written as

$$L(\theta|Y_{obs}) = \int L(\theta|Y_{obs}, Y_{mis})\, dY_{mis}, \qquad \textbf{(12.1)}$$

where θ represents the unknown parameters of the data model, and $L(\theta|Y_{obs}, Y_{mis})$ denotes the likelihood function that one would use if no data were missing. The integration in Equation 12.1 eliminates the dependence on Y_{mis}, broadening the likelihood function to reflect the additional uncertainty due to the fact that Y_{mis} is unknown. In effect, this integration is nearly the same as the averaging over $P(Y_{mis}|Y_{obs})$ that takes place in multiple imputation. Except in very simple problems, the likelihood function Equation 12.1 tends to be complicated, often requiring complicated numerical techniques or approximations. When carried out properly, direct maximum likelihood can be statistically more efficient than multiple imputation because it is a deterministic procedure; no simulation is

involved, so no extra variability is introduced into summary statistics. (In most cases, this extra randomness introduced by multiple imputation is quite minor.) In large samples, estimates and standard errors obtained by direct maximum likelihood and by multiple imputation tend to be very similar.

Applications of direct maximum likelihood are now common in longitudinal analyses. Modern algorithms for growth modeling as implemented in hierarchical linear modeling (HLM; Bryk, Raudenbush, & Congdon, 1996), Proc Mixed in SAS (Littell, Milliken, Stroup, & Wolfinger, 1996), and similar packages are designed for unbalanced data, where measurements on each participant may be taken at a different set of time points. Responses that are missing, either unintentionally or by design, are removed from the likelihood by integration as in Equation 12.1. An important limitation of these packages is that the missing values must be confined to the response variable; missing values on predictors are not allowed. If the individuals in the study have been assessed at a common set of occasions, models equivalent to those fit by HLM and Proc Mixed can be formulated using latent growth curves (McArdle, 1988; Meredith & Tisak, 1990; Willett & Sayer, 1994) and structural equations software. Two recent programs for structural equations, Mx (Neale, 1994) and Amos (Arbuckle, 1995), perform direct maximum likelihood from a raw data set with missing values. Missing data can be accommodated in other structural equations software by using the technique of multiple groups (Allison, 1987; Duncan & Duncan, 1994; Muthén, Kaplan, & Hollis, 1987). An advantage of the latent growth curve approach is that missing values may occur on predictors as well as the response; however, the measurements must be taken at a relatively small number of common time points.

When a direct maximum-likelihood procedure is available for a particular analysis, it may indeed be the most convenient and attractive method. Despite the increasing popularity of direct maximum likelihood, however, multiple imputation still offers some unique advantages for data analysts. First, it allows them to use their favorite models and software; an imputed data set may be analyzed by virtually any method that would be appropriate if the data were complete. As computing environments and statistical models grow increasingly complex, the value of using familiar methods and software should not be underestimated. Second, there are still many classes of problems for which no direct maximum-likelihood procedure is available. For example, in longitudinal analyses there is no direct maximum-likelihood method for incomplete covariates when occasions of measurement vary by individual.

A third reason why multiple imputation can be more attractive than direct maximum likelihood is that the separation of the imputation phase from the analysis phase lends a greater flexibility to the entire process. With multiple imputation the imputer is free to use additional variables that may be helpful for imputation but that are not of direct interest for the analysis. For example,

consider a covariate that helps to explain reasons for nonresponse. Using this variable in the imputation procedure tends to reduce bias in subsequent analyses, even in analyses that do not involve that variable.

Finally, an important advantage of multiple imputation over direct maximum likelihood is that it singles out missing data as a source of random variation distinct from ordinary sampling variability. The likelihood function Equation 12.1 lumps these two types of variability together; summary statistics (e.g., standard errors) derived from direct maximum likelihood do not reveal two sources. With multiple imputation, however, the overall uncertainty is formally partitioned into sampling variability and missing-data uncertainty. This partition immediately yields an estimated rate of missing information, which can be quite helpful for assessing the impact of missing data on inferences for any parameter of interest.

The purpose of this chapter is not to criticize direct maximum likelihood in favor of multiple imputation; rather, it is my hope that more analysts will recognize the important advantages offered by both of these modern missing-data methods and begin to use them instead of case deletion or other ad hoc procedures. In most real-life applications, missing data are not the main focus of scientific inquiry but an unpleasant nuisance. Missing data should be handled quickly and effectively but without compromising the integrity of the analytic results. Multiple imputation might not be the optimal choice for every analysis, but it is a handy statistical tool and a valuable addition to a researcher's methodological toolkit.

In the remainder of this chapter, I describe a method for creating multiple imputations in longitudinal databases. Previous algorithms and software for multiple imputation, as described in Schafer (1997a), have focused on missing data in general multivariate settings. In response to the specific need for longitudinal analyses, a library of algorithms called *PAN* has been developed for imputing multivariate panel data, where a group of variables is measured for individuals at multiple time points. Alternatively, PAN may be applied to clustered data where variables are measured at a single point for participants nested within some larger unit (e.g., students within classrooms). Future versions of the software will be able to handle repeated measures and clustering simultaneously.

PAN is at present available as a library of functions for the statistical programming language S-PLUS (MathSoft, Inc., 1997).[1] Current efforts are focused on developing a version of PAN that operates as a stand-alone program in the Windows 95/98/NT environment.

[1]This can be downloaded free of charge from http://www.stat.psu.edu/~jls/misoftwa.html.

The PAN Model

Suppose that a group of time-varying continuous variables $Y_1, Y_2, \ldots, Y_r$ is measured for individuals $i = 1, 2, \ldots, N$ at multiple occasions. The responses for participant i may be arranged as a matrix with one column for each variable and one row for each occasion,

$$y_i = \begin{bmatrix} y_{i11} & y_{i12} & \cdots & y_{i1r} \\ y_{i21} & y_{i22} & \cdots & y_{i2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{in_i1} & y_{in_i2} & \cdots & y_{in_ir} \end{bmatrix}, \tag{12.2}$$

where y_{ijk} denotes the value of variable Y_k at occasion j. The number of occasions n_i and their temporal spacing may vary by participant. I assume that missing values occur throughout the matrices $y_1, y_2, \ldots, y_m$ and that these missing values are MAR. The immediate goal is to multiply impute the missing values so that the data can be analyzed in a straightforward manner. Ultimately, the analyst may choose to regard one column of Equation 12.2 as a response and the other columns as potential predictors in a conventional growth model. For the moment, however, I regard all r columns of y_i as random responses and model them jointly for the purpose of imputation. I construct a multivariate growth model to describe the joint distribution of the variables $Y_1, Y_2, \ldots, Y_r$, possibly given other time-varying or static covariates that are fully observed and require no imputation.

The model used by PAN was designed to preserve the following relationships: (a) relationships among the variables $Y_1, Y_2, \ldots, Y_r$ within an individual at each time point. These are reflected by the covariances among the elements of any row of y_i. (b) Growth or change in any variable Y_j within an individual across time points. This growth is reflected by trends within the columns of y_i. (c) Relationships between the response variables $Y_1, Y_2, \ldots, Y_r$ and any additional participant-level (non-time-varying) covariates included in the model. The participant-level covariates may be continuous or categorical, but they must be fully observed; missing values on these non-time-varying variables are allowed in the current version. Missing values in time-varying covariates are allowed and will be imputed, provided that they are included among $Y_1, Y_2, \ldots, Y_r$.

PAN relies on a multivariate extension of a linear mixed-effects model that has been popular for nearly 20 years. The model is

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{12.3}$$

where $X_i(n_i \times p)$ and $Z_i(n_i \times q)$ are known covariate matrices, β contains regression coefficients common to all units, and b_i contains coefficients specific to unit i. Note that Equation 12.3 is a multivariate regression; β and b_i are

matrices with r columns, one column for predicting each of the variables Y_1, Y_2, . . . , Y_r, and $\boldsymbol{\varepsilon}_i$ is also a matrix with the same dimensions as $y_i(\eta_i \times r)$. The univariate ($r = 1$) version, which was proposed by Hartley and Rao (1967) and later popularized by Laird and Ware (1982), Jennrich and Schluchter (1986), Bryk and Raudenbush (1992), and others, is the basis for many of the linear growth models in use today. The coefficients $\boldsymbol{\beta}$ and b_i are often called "fixed effects" and "random effects," respectively.

With univariate versions of this model, it is common to assume that the random effects and residuals are independently drawn from normal populations, $b_i \sim N(0, \psi)$ and $\boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 I)$, $i = 1, 2, \ldots, N$, where ψ is a $q \times q$ covariance matrix and I is the identity matrix ($n_i \times n_i$). For the multivariate case, one generalizes these assumptions to

$$\text{vec}(b_i) \sim N(0, \boldsymbol{\Psi}) \tag{12.4}$$

$$\text{vec}(\boldsymbol{\varepsilon}_i) \sim N[0, (\boldsymbol{\Sigma} \otimes I)], \tag{12.5}$$

where *vec* denotes the vectorization of a matrix by stacking its columns. The covariance matrix $\boldsymbol{\Psi}$ in Equation 12.4 has dimension $qr \times qr$, and the Kronecker product notation in Equation 12.5 indicates that the rows of $\boldsymbol{\varepsilon}_i$ are independently distributed as $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is $r \times r$.

In typical applications, the times of measurement are incorporated into X_i, and perhaps Z_i, as linear, quadratic, or higher order polynomials, and Z_i is a subset of the columns of X_i. For example, suppose that the first two columns of X_i are $(1, 1, \ldots, 1)^T$ and $(t_1, t_2, \ldots, t_{n_i})^T$, respectively, where $t_1, t_2, \ldots, t_{n_i}$ are the times of measurement for participant i; beyond these, X_i may have additional columns containing static or time-varying covariates for participant i. Setting Z_i equal to the first column of X_i produces a model of linear growth with intercepts randomly varying by individuals; setting Z_i equal to the first two columns of X_i produces random intercepts and slopes. Centering the distribution of b_i at zero causes $\boldsymbol{\beta}$ to become the population-averaged regression coefficients and the random effects $b_1, \ldots, b_m$ become perturbations due to interparticipant variation.

Note that in this multivariate model all of the covariates in X_i and Z_i appear as predictors for each of the columns of y_i. As a result, the same group of predictors and the same type of trend over time (e.g., linear mean growth with varying slopes and intercepts) are used to describe each of the response variables $Y_1, Y_2, \ldots, Y_r$. The actual coefficients for the response variables, as contained in the r columns of $\boldsymbol{\beta}$ and b_i, vary, but the same group of predictors is applied to each response. At first glance, this may appear to be a serious limitation of the model; in many scientific contexts there is no reason to believe that $Y_1, Y_2, \ldots, Y_r$ should depend on precisely the same set of covariates. One must remember, however, that the purpose of PAN is not to construct a theoretically

meaningful model but to impute missing responses in such a way that important relations are preserved. If a covariate appears in subsequent analyses as a potential predictor of one or more of the response variables $Y_1, Y_2, \ldots, Y_r$, then that covariate should be included in the imputation model, even though its effects on some of the responses may be irrelevant or null. No biases incur by using an imputation model that is larger or more general than necessary for any given analysis. For more discussion on the purpose of imputation modeling and the interplay between the imputer's and analyst's assumptions, see Meng (1994), Rubin (1996), and Schafer (1997a, chapter 4).

The current version of PAN allows two types of assumptions about Ψ, the covariance matrix for the participant-level random effects $b_1, b_2, \ldots, b_N$. One allows the Ψ matrix to be either (a) an unstructured or arbitrary covariance matrix or (b) a block diagonal covariance matrix of the form

$$\Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Psi_r \end{bmatrix}, \tag{12.6}$$

where the nonzero blocks $\Psi_j, j = 1, \ldots, r$ are covariance matrices of size $q \times q$. The unstructured Ψ allows the random effects for any two responses Y_j and Y_k to be correlated, whereas the block-diagonal form assumes that the random effects for each response are independent of those for any other response.

The choice between these two depends on both theoretical and practical considerations. Suppose that $Y_1, Y_2 \ldots, Y_r$ represent achievement scores (mathematics, reading comprehension, etc.) recorded for schoolchildren over time, and one applies a model of linear growth with intercepts and slopes that vary by individual. If there is reason to believe that growth patterns for the various achievement scores are related—for example, that participants with high rates of increase for mathematics may also tend to have high rates of increase for reading comprehension—then it would be wise to use an unstructured Ψ. As the number of response variables grows, however, it often becomes impractical to estimate covariances among all of their random effects unless the number of participants is very large; to obtain a stable estimate for Ψ one may need to specify a block-diagonal structure. Unless the correlations among the random effects for some pairs of responses are unusually strong, the potential biases incurred by using a block-diagonal Ψ rather than an unstructured Ψ tend to be minor.

The basic strategy for specifying a PAN model can be summarized as follows. First, any time-varying covariates with missing values should be placed in the columns of y_i, regardless of whether they are treated as "responses" or "predictors" in later analyses. If a variable is to be imputed, then it must be included among the variables $Y_1, Y_2, \ldots, Y_r$. Second, other covariates of interest

should be included in the columns of X_i and, possibly, Z_i. These include (a) variables that may be related to $Y_1, Y_2, \ldots, Y_r$ and (b) variables that may explain missingness on $Y_1, Y_2, \ldots, Y_r$. Placing a covariate in X_i allows it to influence the distribution of any or all of the variables $Y_1, Y_2, \ldots, Y_r$ in the population. Placing a time-varying covariate in both X_i and Z_i allows its degree of influence on $Y_1, Y_2, \ldots, Y_r$ to vary across individuals. Note that static or non-time-varying covariates (e.g., gender or pretest measures) should not be included in Z_i because it is impossible to estimate participant-specific effects for such variables. Finally, polynomial terms such as *1*, *time*, *time*^{2}, and so on, may be appended to X_i and Z_i as desired, to allow the mean levels of $Y_1, Y_2, \ldots, Y_r$ and the trends in these variables over time to vary across individuals. The choice of which terms to include will depend on what types of effects are believed to exist and what effects will be investigated in subsequent analyses.

Computational Algorithms

The computational engine of PAN is a Markov chain Monte Carlo (MCMC) algorithm called a *Gibbs sampler*. MCMC is a relatively new class of simulation techniques that are especially useful in Bayesian statistical analyses. A review of MCMC is beyond the scope of this chapter, but a gentle introduction is given by Casella and George (1992) and Schafer (1997a, chapters 3–4); more comprehensive references are the volume edited by Gilks, Richardson, and Spiegelhalter (1996) and the article by Gelfand and Smith (1990). Specific details and formulas for the computations used in PAN have been provided by me (Schafer, 1997b; Yucel & Schafer, 1998).

The MCMC algorithm in PAN is based on the observation that the model specified by Equations 12.3–12.5 has the following unknown components: the missing values in $y_1, y_2, \ldots, y_N$, the random effects $b_1, b_2, \ldots, b_N$, the fixed effects β, and the covariance matrices Σ and Ψ. For the purpose of imputation, I am interested only in simulating the missing data in $y_1, y_2, \ldots, y_N$; the other unknown quantities are merely a nuisance. To simulate the missing data properly, however, one must take into account the uncertainty in these other quantities and how it contributes to missing-data uncertainty. Expressing this uncertainty through mathematical formulas is difficult, so one accounts for the interdependence among the unknown quantities through a process of iterative simulation.

PAN simulates the unknown quantities in a three-step cycle.

1. Draw random values of $b_1, b_2, \ldots, b_N$ on the basis of some plausible assumed values for the missing data and the parameters β, Σ, and Ψ.
2. Draw new random values of the unknown parameters β, Σ, and

Ψ on the basis of the assumed values for the missing data and the values of $b_1, b_2, \ldots, b_N$ obtained in Step 1.

3. Draw new random values for the missing data given the values of $b_1, b_2, \ldots, b_N$ obtained in Step 1 and the parameters obtained in Step 2.

At the end of this cycle the parameters and missing data from Steps 2 and 3 become the values assumed in Step 1 at the start of the next cycle. Repeating Steps 1, 2, and 3 in turn defines a Markov chain, a sequence in which the distribution of the unknown quantities at any cycle depends on their simulated values at the previous cycle. The state of the process at Cycle 2 may be strongly correlated with its state at Cycle 1, but at subsequent Cycles 3, 4, 5, and so on, the relationship to the original state weakens. When a sufficient number of cycles has been taken to make the resulting state essentially independent of the original state, then the process is said to have *converged* or *achieved stationarity*. On convergence, the final simulated values for the missing data have in fact come from the distribution from which multiple imputations should be drawn.

This algorithm may be used to create m multiple imputations in the following way. Starting with some plausible initial values, run the Gibbs sampler for k cycles where k is large enough to ensure convergence, and take the final simulated version of the missing data as the first imputation; then return to the original starting values, run the Gibbs sampler for another k cycles, and take the final simulated version of the missing data as the second imputation; and so on. This method requires m runs of length k cycles each. Another and perhaps more convenient way is to perform one long run of mk cycles, saving the simulated values of the missing data after cycle k, $2k$, $\ldots$, mk as the m imputations. The latter method differs from the former only in that the final values from each subchain of length k become the starting values for the next subchain of length k.

It is important to note that convergence of an MCMC procedure means convergence to a probability distribution rather than convergence to a set of fixed values. To say that the algorithm has converged by k cycles actually means that the random state of the process at cycle $t + k$ is statistically independent of its state at cycle t for $t = 1, 2, \ldots$. After running the Gibbs sampler, one can examine the output stream over many cycles to see how many are needed to achieve this independence. Suppose that one collects and stores the simulated values for one parameter θ (a particular element of β, Ψ, or Σ) over a large number C of consecutive cycles. These values $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(C)}$ can be regarded as a time series. The lag-k autocorrelation, which is the correlation between pairs $\theta^{(t)}$ and $\theta^{(t+k)}$ ($t = 1, 2, \ldots, C - k$), can be calculated for various values of k to determine how large k must be for the correlations to die down. In principle, one should examine autocorrelations for each parameter in the model

and identify a value of k large enough to guarantee that the lag-k autocorrelations for all parameters are effectively zero. In my experiences with real data, however, I have found that the greatest levels of serial dependence are almost always seen in variance and covariance parameters, and in particular within the elements of Ψ. It is usually sufficient to monitor the behavior of the elements of Ψ because it is with respect to these parameters that the algorithm tends to converge the most slowly. For more discussion on monitoring the convergence of MCMC algorithms, see Schafer (1997a, chapter 4).

The rate of convergence of this Gibbs sampler is influenced by a combination of factors pertaining to the data and the model. First, it is affected by the amounts and patterns of missing data in the matrices $y_1, y_2, \ldots, y_N$; greater rates of missing information lead to slower convergence. It is also affected by one's ability to estimate the individual random effects $b_1, b_2, \ldots, b_N$; if estimates of random effects are highly variable, then convergence is slowed. Finally, convergence behavior is also influenced by the number of participants (N). As the sample size grows, the distribution of the random Ψ matrix at each cycle becomes more tightly concentrated around the sample covariance matrix of $b_1, b_2, \ldots, b_N$ from the previous cycle. As this distribution becomes tighter, the elements of Ψ are less free to wander away from their values at the previous cycle, producing higher correlations from one cycle to the next. It is somewhat ironic that the algorithm converges more slowly as one's ability to estimate the parameters increases. With a large number of participants and a small number of occasions per participant, it is not uncommon for the Gibbs sampler to require several hundred or even 1,000 cycles to converge. Slow convergence is not necessarily a problem, however, because in most cases only a few imputations are necessary. If $k = 1{,}000$ cycles are needed to achieve stationarity, then five imputations can be produced in 5,000 cycles, which even for a large data set requires no more than a few hours on a personal computer.

In addition to deciding how many cycles are needed, the user must also specify Bayesian prior distributions for the covariance matrices Ψ and Σ. Bayesian procedures, which are becoming increasingly popular in many areas of statistical analyses, treat unknown parameters as random variables and assign prior probability distributions to them to reflect one's knowledge of or belief about the parameters before the data are seen. An excellent introduction to the Bayesian statistical paradigm was given by Novick and Jackson (1974); for a modern overview of Bayesian modeling and computation, see Gelman, Rubin, Carlin, and Stern (1995). Some statisticians tend to prefer Bayesian procedures on principle, whereas others avoid them on principle. I hold a pragmatic view, accepting the prior distribution simply as a mathematical device that allows one to generate the imputations in a principled fashion. In applications, I like to use prior distributions that are weak or highly dispersed, reflecting a state of relative ignorance about model parameters. Weak priors tend to minimize

the subjective influence of the prior, allowing the observed data to speak for themselves.

The prior distribution most commonly applied to a covariance matrix is the inverted Wishart distribution. The Wishart, a natural generalization of the chi-square to random matrices, is discussed in standard texts on multivariate analysis (e.g., Anderson, 1984; Johnson & Wichern, 1992). The prior distribution for Σ is

$$\Sigma^{-1} \sim W(a, B), \tag{12.7}$$

where $W(a, B)$ denotes a Wishart with a degrees of freedom and scale B. The scale is a symmetric, positive definite matrix with the same dimensions ($r \times r$) as Σ. The degrees of freedom, which should be greater than or equal to r, govern the spread or variability; lower values of a make the distribution more dispersed. The user of PAN must provide numeric values for a and B^{-1}. Our usual practice is to set $a = r$ to make the prior as dispersed as possible and then to set $B^{-1} = a\hat{\Sigma}$, where $\hat{\Sigma}$ is a reasonable prior guess or estimate of Σ. If a guess for Σ is unavailable, the data themselves may be used to obtain one. Yucel and Schafer (1998) recently developed a new expectation–maximization algorithm for calculating maximum-likelihood estimates of the parameters β, Ψ, and Σ from the incomplete data. Running this EM algorithm before the Gibbs sampler is an excellent way to obtain a reasonable guess for Σ.

In a similar fashion, I also use inverted Wishart prior distributions for the between-subjects covariance matrix Ψ. If Ψ is unstructured, one assumes $\Psi^{-1} \sim W(c, D)$ where D is a $qr \times qr$ matrix and $c > qr$. My usual practice is to set $c = qr$ and $D^{-1} = c\hat{\Psi}$, where $\hat{\Psi}$ is a prior guess or estimate of Ψ. If Ψ is taken to be block diagonal as in Equation 12.6, then independent inverted Wishart prior distributions are applied to the nonzero blocks, $\Psi_j^{-1} \sim W(c_j, D_j)$, $j = 1, \ldots, r$, where $c_j \geq q$. To make the priors weak, one sets $c_j = q$ and $D_j^{-1} = c_j\hat{\Psi}_j$ where $\hat{\Psi}_j$ is an estimate or guess for ψ_j. The EM algorithm described by Yucel and Schafer (1998) provides a maximum-likelihood estimate for an unstructured ψ or estimates of the submatrices $\Psi_1, \ldots, \Psi_r$ when Ψ is block diagonal.

An Example: Expectancies and Alcohol Use in the Adolescent Alcohol Prevention Trial

The Adolescent Alcohol Prevention Trial (AAPT) was a longitudinal school-based intervention study of substance use carried out in the Los Angeles area (Hansen & Graham, 1991). In one panel of AAPT, attitudes and behaviors pertaining to the use of alcohol, tobacco, and marijuana were measured by self-report questionnaires administered yearly in Grades 5–10. The data exhibit

typical rates of uncontrolled nonresponse due to absenteeism, attrition, and so on, which I assume to be MAR. This assumption has been given careful consideration by the researchers and appears to be plausible; for example, much of the attrition is due to students moving to other schools or districts, which is at most only weakly associated with substance use patterns (Graham et al., 1994).

In addition to this uncontrolled nonresponse, large amounts of truly MAR missing data (MCAR, in fact) arose by design. The AAPT study made use of an innovative three-form design in which each student received only a subset of the items in any year, as described in chapter 11 of this volume, by Graham, Taylor, and Cumsille. In some years, certain items were omitted entirely. For the present analysis, I examine a cohort of m = 3,574 children and focus attention on three variables: "drinking," a composite measure of self-reported alcohol use; POSCON, a measure of the degree to which the student perceives that alcohol use has positive consequences; and NEGCON, a measure of the perceived negative consequences of use. Drinking appeared on the questionnaire every year, where POSCON was omitted in Grade 8 and NEGCON was omitted in Grades 8–10. Missingness rates for the three variables by grade are shown in Table 12.1; observed means and standard deviations appear in Table 12.2.

My analysis will focus on the possible influences of POSCON and NEGCON on drinking. Without missing data, it would be straightforward to build a growth model for drinking that includes the expectancy measures POSCON and NEGCON as time-varying covariates. Current software for multilevel models cannot accommodate missing values on covariates, however, so I first use PAN to jointly impute the missing values for drinking, POSCON, and NEGCON.

Notice in Table 12.2 that both the average level of drinking and its variation increase dramatically over time. This is somewhat problematic, because standard growth models—and the multivariate model used by PAN—assume constant variance in a response over time. To make the assumption of constant

TABLE 12.1

Missingness Rates (%) for Three Variables by Grade

	GRADE					
VARIABLE	**5**	**6**	**7**	**8**	**9**	**10**
Drinking	2	24	24	33	35	44
POSCON	47	55	62	100	66	63
NEGCON	48	56	62	100	100	100

Note. POSCON = positive consequences; NEGCON = negative consequences.

TABLE 12.2

Means and Standard Deviations of Observed Variables by Grade

	GRADE											
	5		6		7		8		9		10	
VARIABLE	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Drinking	−1.43	1.33	−1.12	1.96	−0.57	2.73	0.09	3.47	1.29	4.40	1.97	4.78
POSCON	1.30	0.61	1.34	0.62	1.48	0.74			1.84	0.89	1.96	0.91
NEGCON	2.94	0.76	3.05	0.75	3.07	0.77						

Note. POSCON = positive consequences; NEGCON = negative consequences.

variance more plausible, I transformed drinking by taking its logarithm (after adding a small constant to ensure that all values were positive). After this transformation, the increase in variation became much less noticeable. The log-transformed version of drinking was used both in the imputation procedure and in subsequent analysis described below, because the transformed version more closely fit the assumptions of both the imputation procedure and the analysis. With multiple imputation, however, it is not necessary for variables to be imputed and analyzed on the same scale. Applying transformations at the imputation phase can be a highly effective tool for preserving important distributional features of nonnormal variables, regardless of how the variables are later analyzed (Schafer & Olsen, 1998).

To set up the data for PAN, one first arranges the responses for each individual in the form of a matrix y_i of dimension 6×3, with the rows corresponding to occasions (Grades 5, . . . , 10) and columns for drinking, POSCON, and NEGCON. In devising the imputation model the primary concern is to preserve growth in the variable drinking and its potential relationships to the expectancy measures. With only six time points, the model for growth must be rather simple, so let us posit a linear model with intercepts and slopes randomly varying across individuals. That is, we create a model in which drinking, POSCON, and NEGCON are each described by a linear trend with a random intercept and a random slope, for a total of six random effects in each b_i. Random intercepts and slopes are specified by placing $(1, 1, 1, 1, 1, 1)^T$ and $(1, 2, 3, 4, 5, 6)^T$ into the columns of X_i and Z_i. Finally, to incorporate potential gender differences, I allow the population average slopes and intercepts for boys and girls to vary by adding two additional columns to each X_i matrix: $\text{sex}_i \times (1, 1, 1, 1, 1, 1)^T$ and $\text{sex}_i \times (1, 2, 3, 4, 5, 6)^T$, where sex_i is a dummy indicator for participant i's gender (0 for girl, 1 for boy).

In defining a PAN model, there is no particular importance attached to the specific coding scheme used to create the design matrices X_i and Z_i. For example, the linear effect of time could have been expressed as $(-5, -3, -1, 1, 3, 5)^T$ or any other set of equally spaced scores, and the gender effect sex_i could have been coded as any two values (e.g., -1 and $+1$) rather than as 0 and 1. The particulars of the coding scheme affect the precise meaning of the parameters in $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}$, but these parameters are not of inherent interest—the goal at this stage is not to interpret parameters but to impute the missing values in y_i. Changing the coding scheme in X_i and Z_i does not change the distribution of imputed values, provided that the linear space spanned by the columns of these design matrices does not change.

Table 12.1 indicates that NEGCON is entirely missing for the last 3 years of the study. It may seem unusual to impute a variable that is entirely missing. Under this model the likely values of NEGCON for Grades 8–10 are being inferred from two sources: extrapolation from Grades 5–7 on the basis of the

assumption of linear growth, and the residual covariances among the three response variables in $\boldsymbol{\Sigma}$, which are assumed to be constant across time. Neither of these assumptions can be effectively tested with the data at hand, so inferences pertaining to NEGCON are heavily model based. In retrospect, it would have been very helpful to collect NEGCON in the final year (Grade 10) to provide more stable estimates of this variable's growth.

Before running the Gibbs sampler, I first obtained initial estimates of the unknown parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Psi}$ by running the EM algorithm. This EM procedure, which assumed an unstructured form for $\boldsymbol{\Psi}$, converged in 134 iterations and took less than 1 h on a 400 MHz Pentium II computer. The resulting maximum-likelihood estimates for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ were then used to formulate weak prior distributions as described in the Computational Algorithms section.

Because of the high rates of missing information, I anticipated that the Gibbs sampler would converge slowly. To assess convergence, I ran it for an initial 2,000 cycles and examined time series plots and sample autocorrelations for a variety of parameters. As anticipated, the elements of $\boldsymbol{\Psi}$ pertaining to the slopes and intercepts of NEGCON were among the slowest to converge because of the extreme sensitivity of these parameters to missing data. On the basis of this exploratory run, it appeared that several hundred cycles might be sufficient to achieve approximate stationarity. The Gibbs sampler was then run for an additional 9,000 cycles, with the simulated value of Y_{mis} stored at cycles 2,000, 3,000, . . . , 11,000. Autocorrelations estimated from cycles 1,001–11,000 verified that the dependence in all components of θ had indeed died down by lag 200, so the 10 stored imputations could be reasonably regarded as independent draws from $P(Y_{mis}|Y_{obs})$. The entire imputation procedure took less than 2 hr with a 400 MHz Pentium II.

After imputation, the data were analyzed by a conventional linear growth-curve model for the logarithmically transformed drinking. The model was similar to the one used for imputation, except that POSCON and NEGCON now appear as time-varying covariates rather than responses. The model included an intercept and fixed effects for gender, grade, gender × grade, POSCON, and NEGCON, plus random intercepts and slopes for grade. Time was coded as $(1, 2, 3, 4, 5, 6)^T$, and gender was expressed as a dummy indicator (0 for girls, 1 for boys). Parameter estimates were computed for each imputed data set using a procedure equivalent to that used by standard packages such as HLM.

Finally, the 10 sets of fixed-effects estimates and their standard errors were then combined using Rubin's (1987) rules for multiple-imputation inference for scalar estimands. These rules are summarized as follows. Let Q denote the quantity to be estimated, in this case a regression coefficient. Let $\hat{Q}^{(j)}$ denote the estimate of Q from the jth imputed data set, and U_j its squared standard error ($j = 1, 2, \ldots, m$). The overall estimate of Q is simply the average

$$\bar{Q} = m^{-1}\Sigma\hat{Q}^{(l)}. \tag{12.8}$$

To obtain a standard error for $\bar{Q}$, one calculates the between-imputation variance $B = (m - 1)^{-1}\Sigma(\hat{Q}^{(l)} - \bar{Q})^2$ and the within-imputation variance $\bar{U} = m^{-1}\Sigma U^{(l)}$. The estimated total variance is

$$T = (1 + m^{-1})B + \bar{U}, \tag{12.9}$$

and tests and confidence intervals are based on a Student's t approximation

$$(\bar{Q} - Q)/\sqrt{T} \sim t_\nu, \tag{12.10}$$

with degrees of freedom

$$\nu = (m - 1)\left[1 + \frac{\bar{U}}{(1 + m^{-1})B}\right]^2.$$

The ratio $r = (1 + m^{-1})B/\bar{U}$ measures the relative increase in variance due to missing data, and the rate of missing information in the system is approximately $\lambda = r/(1 + r)$. A more refined estimate of this rate is

$$\lambda = \frac{r + 2/(\nu + 3)}{1 + r}. \tag{12.11}$$

The results of this procedure are summarized in Table 12.3, which shows the overall estimates, standard errors, degrees of freedom for the t approximation, and estimated percentage rates of missing information. All coefficients are highly statistically significant. The high rates of missing information indicate that the inferences for all coefficients (except sex) may be highly dependent on the form of the imputation model and the MAR assumption. The latter assumption is not particularly troubling for these data because the majority of

TABLE 12.3

Estimated Coefficients (Est.), Standard Errors, Degrees of Freedom, and Percentage Missing Information From Multiply Imputed Growth-Curve Analysis

VARIABLE	EST.	SE	df	% MISSING
Intercept	−2.572	0.084	19	71
Grade (1 = 5th, . . . , 6 = 10th)	0.386	0.011	35	53
Sex (0 = female, 1 = male)	0.370	0.046	324	17
Sex × grade	−0.105	0.013	88	33
POSCON	0.549	0.023	17	76
NEGCON	−0.090	0.023	15	80

Note. POSCON = positive consequences; NEGCON = negative consequences.

missing values are missing by design. Certain assumptions of the imputation model, however—in particular, the assumed linear growth for NEGCON and constancy of the residual covariances across time—are not really testable from the observed data, so results from this analysis should be interpreted with caution.

Despite these caveats, the estimates in Table 12.3 provide some intriguing and plausible interpretations about the behavior of this cohort. The positive coefficient for sex indicates that boys reported higher average rates of alcohol use than girls in the initial years of the study. The negative effect of sex × grade, however, shows that girls exhibit higher rates of increase than boys, so that the girls' average overtakes the boys' by Grade 8. The large positive effect of POSCON indicates that increasing perceptions about the positive consequences of alcohol use are highly associated with increasing levels of reported use. The negative coefficient for NEGCON suggests that increasing beliefs about negative consequences do tend to reduce level of use, but the effect is much smaller than that of POSCON. These results are consistent with those of previous studies (e.g., MacKinnon et al., 1991) that demonstrate that perceived positive consequences may be influential determinants of substance use behavior, but beliefs about negative consequences have little or no discernible effect.

Discussion

The multivariate mixed model (Equation 12.3) used by PAN is a natural extension of univariate growth models, which are popular in the analysis of longitudinal data. The imputation procedures described here are appropriate for longitudinal analyses with partially missing covariates. These methods are also appropriate for multivariate cross-sectional studies in which units are nested within naturally occurring groups (e.g., children within schools). The algorithm and software described in this chapter provide a principled solution to missing-data problems for this important and frequently occurring class of analyses.

The imputation model and Gibbs sampler can be extended in a number of important ways. One extension pertains to models with additional random effects due to higher levels of clustering; this would arise, for example, in multivariate studies in which individuals are grouped into larger units and multiple observations on individuals are taken over time. Another useful extension pertains to columns of y_i that are necessarily constant across the rows $1, \ldots, n_i$. In longitudinal studies, these columns would represent covariates that do not vary over time; in clustered applications, they would represent characteristics of the clusters rather than the units nested with them. If these covariates have no missing values, they can be handled under the current model by simply moving them to the matrix X_i. When missing values are present, however, they

must be explicitly modeled for purposes of imputation. If one imposes a simple parametric distribution on these covariates (e.g., multivariate normal), then it is straightforward to extend the Gibbs sampling procedure to impute these as well.

Another useful extension involves interactions among the columns of y_i. The multivariate normal model allows only simple linear associations among the variables $Y_1, \ldots, Y_r$, but in many studies one would like to preserve and detect certain nonlinear associations and interactions. In the AAPT example, it may have been useful to see whether the strong effect of POSCON on drinking may have been increasing or decreasing over time; the imputation model, however, imputed the missing values under an assumption of a constant POSCON $\times$ drinking association. Extensions of the multivariate model to allow more elaborate fixed associations, such as POSCON $\times$ drinking $\times$ grade, or random associations, such as POSCON $\times$ drinking $\times$ participant, are an important topic for future research.

In the current PAN model, the rows of y_i are assumed to be conditionally independent given b_i with common covariance matrix Σ. This assumption has been relaxed by Jennrich and Schluchter (1986), Lindstrom and Bates (1988), and others in the univariate case to allow a residual covariance matrix of the form $\sigma^2 V_i$, where V_i has a simple (e.g., autoregressive or banded) pattern dependent on one or more unknown parameters. Extensions of these patterned covariance structures to a multivariate setting tend to produce models and algorithms that are complex even apart from missing data. For example, the obvious extension of $\text{vec}(\varepsilon_i) \sim N[0, (\Sigma \otimes I)]$ to $\text{vec}(\varepsilon_i) \sim N[0, (\Sigma \otimes V_i)]$ seems too restrictive for many longitudinal data sets, because the response variables $Y_1, \ldots, Y_r$ are then required to have identical autocorrelations. Accounting for autocorrelated residuals in a sensible manner may prove to be a daunting task in the multivariate case. In practice, nonzero correlations among the rows of ε_i may arise because of a misspecified model for the mean structure over time. The problem may sometimes be reduced or eliminated by including additional (e.g., higher order polynomial) terms for time in the covariate matrices X_i or Z_i.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology, 17*, 71–103.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York: Wiley.

Arbuckle, J. L. (1995). *Amos users' guide*. Chicago: Small Waters.

Bailey, L., Chapman, D., & Kasprzyk, D. (1985). Nonresponse adjustment procedures

at the Census Bureau: A review. In *Proceedings of the annual research conference* (pp. 421–444). Washington, DC: U.S. Bureau of the Census.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician, 46*, 167–174.

Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research, 29*, 313–338.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Gelman, A., Rubin, D. B., Carlin, J., & Stern, H. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Graham, J. W., Hofer, S. M., & Piccinin, A. M. (1994). Analysis with missing data in drug prevention research. In L. Collins & L. Seitz (Eds.), *National Institute on Drug Abuse research monograph series* (Vol. 142, pp. 13–62). Washington, DC: National Institute on Drug Abuse.

Hansen, W. B., & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing consecutive norms. *Preventive Medicine, 20*, 414–430.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*, 93–108.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics, 38*, 967–974.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

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

Lindstrom, M. J., & Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83*, 1014–1022.

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., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

MacKinnon, D. P., Johnson, C. A., Pentz, M. A., Dwyer, J. H., Hansen, W. B., Flay, B. R., & Wang, E. Y. (1991). Mediating mechanisms in a school-based drug prevention program: First-year effects of the Midwestern Prevention Project. *Health Psychology, 10*, 164–172.

Madow, W. G., Nisselson, H., & Olkin, I. (1983). *Incomplete data in sample surveys, Vol. 1: Report and case studies*. New York: Academic Press.

MathSoft, Inc. (1997). *S-PLUS user's guide*. Seattle, WA: Author.

McArdle, J. (1988). Dynamic but structural modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 561–614). New York: Plenum Press.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science, 10*, 538–573.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107–122.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431–462.

Neale, M. C. (1994). *Mx: Statistical modeling* (2nd ed.). Richmond: Medical College of Virginia, Department of Psychiatry.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

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 (with discussion). *Journal of the American Statistical Association, 91*, 473–489.

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

Schafer, J. L. (1997b). *Imputation of missing covariates under a multivariate linear mixed model* (Tech. Rep. 97-10). University Park: Pennsylvania State University, The Methodology Center.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33*, 545–571.

Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Psychological Bulletin, 116*, 363–381.

Yucel, R., & Schafer, J. L. (1998). Fitting multivariate linear mixed models with incomplete data. In *Proceedings of the Statistical Computing Section of the American Statistical Association* (pp. 177–182). Alexandria, VA: American Statistical Association.