

# Missing Data Techniques for Structural Equation Modeling

Paul D. Allison  
University of Pennsylvania

As with other statistical methods, missing data often create major problems for the estimation of structural equation models (SEMs). Conventional methods such as listwise or pairwise deletion generally do a poor job of using all the available information. However, structural equation modelers are fortunate that many programs for estimating SEMs now have maximum likelihood methods for handling missing data in an optimal fashion. In addition to maximum likelihood, this article also discusses multiple imputation. This method has statistical properties that are almost as good as those for maximum likelihood and can be applied to a much wider array of models and estimation methods.

Virtually all methods of statistical analysis are plagued by problems with missing data, and structural equation modeling is no exception. It is well known that the use of inappropriate methods for handling missing data can lead to bias in parameter estimates (Jones, 1996), bias in standard errors and test statistics (Glasser, 1964), and inefficient use of the data (Afifi & Elashoff, 1966). This article surveys various methods that are available for handling missing data in the estimation of structural equation models (SEMs). After reviewing such conventional methods as listwise deletion, pairwise deletion, and regression imputation, I focus on the implementation of two newer methods, maximum likelihood and multiple imputation. These methods have much better statistical properties than conventional methods have under considerably weaker assumptions, a rare combination for new statistical methods.

Before discussing the methods, it is essential to clarify the meaning of certain assumptions that are often invoked in justifying one method or another (Rubin, 1976). To keep things simple, suppose that a data set contains only two variables,  $X$  and  $Y$ . We observe  $X$  for all cases, but data are missing on  $Y$  for, say, 20% of the cases. We say that data on  $Y$  are missing completely at random (MCAR) if the probability that data are missing on  $Y$  depends on neither  $Y$  nor  $X$ . Formally, we have  $\Pr(Y \text{ is missing} \mid X, Y) = \Pr(Y \text{ is missing})$ . If  $Y$  is a measure of delinquency and  $X$  is years of schooling, MCAR would mean that the probability of missing data on delinquency is unrelated to either delinquency or schooling. Many missing data techniques are valid only if the MCAR assumption holds.

A weaker assumption is that the data are missing at random (MAR). This assumption holds that the probability that data are missing on  $Y$  may depend on the value of  $X$ , but does not depend on the value of  $Y$ , holding  $X$  constant. Formally, we have  $\Pr(Y \text{ is missing} \mid X, Y) = \Pr(Y \text{ is missing} \mid X)$ . Again, if  $Y$  is a measure of

delinquency and  $X$  is years of schooling, the MAR assumption would be satisfied if the probability that delinquency is missing depends on years of schooling, but within each level of schooling the probability of missing delinquency does not depend on delinquency. In essence, MAR allows missingness to depend on things that are observed, but not on things that are not observed. Clearly, if the data are missing completely at random, they are also missing at random.

It is straightforward to test whether the data are missing completely at random. For example, one could compare men and women to test whether they differ in the proportion of cases with missing data on particular variables. Any such difference would be a violation of MCAR. However, it is impossible to test whether the data are missing at random. For obvious reasons, one cannot tell whether delinquent children are more likely than nondelinquent children are to have missing data on delinquency.

Missing data are said to be *ignorable* if the data are MAR and, in addition, the parameters governing the missing data mechanism are completely distinct from the parameters of the model to be estimated. This somewhat technical condition is unlikely to be violated in the real world. Even if it were, methods that assume ignorability would still perform very well if the data are merely MAR. It is just that one could do even better by modeling the missing data mechanism as part of the estimation process. So, in practice, missing at random and ignorability are often used interchangeably.

The reader may have noticed that I have not said anything about *why* the data are missing. Obviously there are many different possible reasons, from a subject's refusal to answer a particular question to the inability to locate the subject on follow-up. Although one might think that different kinds of missing data ought to be treated differently, all the methods I consider here are completely agnostic about the reasons for the missing data (except to make one or more of the assumptions discussed above). These methods do presume, however, that when a datum is missing, there is some real value that could, in principle, be measured. For those cases in which a variable simply does not apply to a particular subject (e.g., a measure of

---

Correspondence concerning this article should be addressed to Paul D. Allison, Sociology Department, University of Pennsylvania, 3718 Locust Walk, Philadelphia, Pennsylvania 19104-6299. E-mail: allison@ssc.upenn.edu

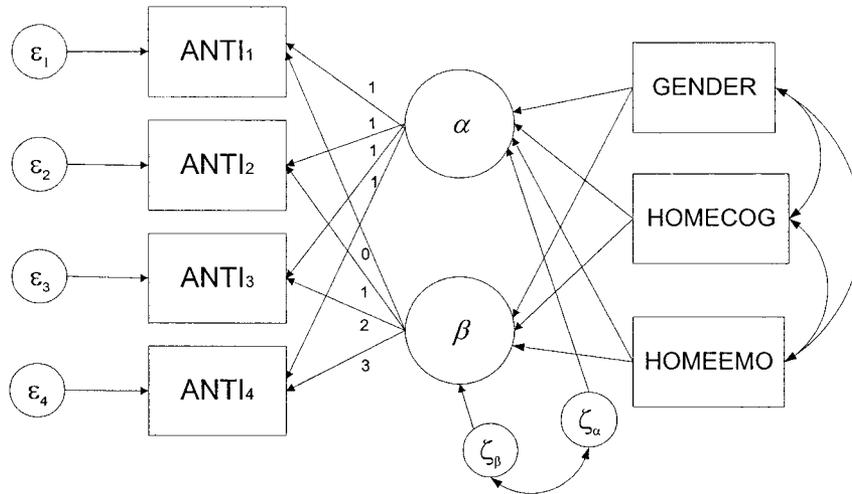


Figure 1. Path diagram for latent growth model. ANTI = a measure of antisocial behavior; HOMECOG = a measure of the cognitive stimulation provided at home; HOMEEMO = a measure of the emotional support provided at home.

marital satisfaction for unmarried persons), the methods discussed here may not be appropriate.<sup>1</sup>

### Example Data Set

To make things more concrete, I consider the various methods in the context of an example. I shall use a set of data that was analyzed by participants in a 1997 symposium sponsored by the Society for Research on Child Development (SRCD; Curran, 1997). The data set consisted of 405 children and their mothers who were interviewed in 1986, 1988, 1990, and 1992. The variables used in the analysis are as follows:

- ANTI1–ANTI4 — A measure of antisocial behavior at each of the four time points, ranging from 0 to 10
- GENDER — 1 = male, 0 = female
- HOMECOG — A measure of cognitive stimulation provided to the child at home, ranging from 1 to 13
- HOMEEMO — A measure of emotional support provided in the home, ranging from 0 to 14
- KIDAGE — Child’s age at time 1, ranging from 6 to 8
- MOMAGE — Mother’s age at time 1, ranging from 21 to 29

In this data set, there are no missing data for the variables GENDER through MOMAGE. But there was considerable attrition in the measure of antisocial behavior. For this variable, all 405 children were measured at the first interview, 374 at the second interview, 297 at the third interview, and 294 at the fourth interview. Only 261 children were measured at all four interviews.

Our basic model is a linear latent growth model. Let  $y_{it}$  be the measure of antisocial behavior for child  $i$  at time  $t$ . We assume that for each child, antisocial behavior changes as a linear function of time, but the slope and intercept of that linear regression varies from child to child. Thus

$$y_{it} = \alpha_i + \beta_i t + \epsilon_{it}$$

Furthermore, the intercept and slope are each functions of exogenous<sup>2</sup> variables:

$$\alpha_i = \mu_\alpha + \gamma_{\alpha 1} x_{i1} + \dots + \gamma_{\alpha k} x_{ik} + s_{\alpha i}$$

$$\beta_i = \mu_\beta + \gamma_{\beta 1} x_{i1} + \dots + \gamma_{\beta k} x_{ik} + s_{\beta i}$$

The error terms in these last two equations are assumed to be uncorrelated with  $\epsilon_{it}$  but are allowed to be correlated with each other. In the models to be estimated here, the  $x$  variables are GENDER, HOMECOG, and HOMEEMO, but not MOMAGE and KIDAGE. Principal interest centers on the  $\gamma$  parameters, which will be estimated using several different approaches to handling missing data.

A path diagram of this model is shown in Figure 1. Variables enclosed by rectangles are directly observed, and variables enclosed by circles are not directly observed. Note that  $\alpha$  and  $\beta$  are treated as latent variables with constraints imposed on their coefficients: For  $\alpha$ , all the coefficients are 1.0; for  $\beta$ , the coefficients go up linearly with time.

How might the various assumptions regarding missingness apply to these data? MCAR would mean that missingness on ANTI2–ANTI4 would be completely independent of any of the variables in the model, including the variables with missing data. If the data were MAR, then missingness on the antisocial measures could depend on other observed variables. In particular, missingness on ANTI2 could depend on ANTI1, but could not depend on ANTI2 (after controlling for all other observed variables).

<sup>1</sup> Schafer and Graham (2000) argued that, in some cases, treating non-existent data points as missing values may be conceptually useful or, at the least, computationally convenient.

<sup>2</sup> Exogenous variables are those that are not affected by any other variables in the system; that is, they do not appear as dependent variables in any equation.

Table 1  
*Estimates of  $\gamma$  Parameters Under Different Treatments of Missing Data*

Method	$\alpha$				$\beta$			
	Intercept	GENDER	HOMEEOG	HOMEEMO	Intercept	GENDER	HOMEEOG	HOMEEMO
Listwise deletion								
Par. Est.	3.208	.625	-.024	-.184	.348	.115	-.038	.009
SE	.430	.179**	.038	.041**	.194	.080	.017*	.019
Pairwise deletion								
Par. Est.	3.103	.854	-.042	-.154	.348	-.010	-.024	.004
SE (N = 275)	.423	.179**	.037	.042**	.189	.079	.017	.019
SE (N = 405)	.348	.147**	.031	.034**	.155	.066	.014	.015
Expectation-maximization algorithm (2-step)								
Par. Est.	3.080	.822	-.038	-.156	.368	.029	-.031	.004
SE (N = 275)	.419	.177**	.037	.041**	.190	.080	.017	.019
SE (N = 405)	.345	.146**	.030	.034**	.156	.066	.013*	.015
Direct maximum likelihood								
Par. Est.	3.078	.819	-.037	-.157	.376	.037	-.035	.006
SE	.376	.148**	.031	.034**	.177	.075	.016*	.017
Multiple imputation								
Par. Est.	3.059	.830	-.039	-.157	.380	.030	-.033	.004
SE	.356	.150**	.031	.034**	.176	.073	.015*	.017

Note. In the latent growth equation for ANTI,  $\alpha$  is the intercept,  $\beta$  is the slope. Par. Est. = parameter estimate; N = assumed sample size; Homecog = a measure of the cognitive stimulation provided at home; Homeemo = a measure of the emotional support provided at home.  
 \*  $p < .05$ . \*\*  $p < .01$ .

## Conventional Methods

### Listwise Deletion

The simplest approach to missing data, and the one that is the default in virtually all statistical packages (except for Amos), is the method known to statisticians as complete case analysis but more commonly known among social scientists as *listwise deletion*. In this method, cases are deleted from the sample if they have missing data on any of the variables in the analysis to be conducted. Because this produces a working sample with no missing data, any statistical method may then be applied.

In addition to its simplicity and generality, listwise deletion has two important statistical properties. First, if the data are missing completely at random, listwise deletion will not introduce any bias into the parameter estimates: If the data are MCAR, the subsample with complete data is effectively a simple random sample from the original sample. It is well known that simple random sampling does not introduce any bias. Second (and for the same reason), the standard error estimates under listwise deletion should be approximately unbiased estimates of the true standard errors. This is particularly important because standard error estimates under most other conventional methods are flawed for one reason or another.

If data are missing at random, but not *completely* at random, listwise deletion may yield biased parameter estimates (Wothke, 2000). For example, if men are more likely than are women to be missing ANTI3, and if men also tend to have more antisocial behavior than women have, then the mean of ANTI3 will be biased downward under listwise deletion.<sup>3</sup> On the other hand, listwise deletion can tolerate certain kinds of MAR dependencies in regression analysis. In general, if the probability of missing data on any variable in a regression model does *not* depend on the dependent variable (controlling for other variables), estimates of regression coefficients will not be biased by listwise deletion.

Irrespective of bias, the big problem with listwise deletion is that a large fraction of the original sample may be deleted, especially when the model to be estimated has many variables. In our example, 36% of the cases are lost under this method. Consequently, even though the standard errors produced by listwise deletion are approximately unbiased estimates of the true standard errors, those standard errors may be substantially larger than they would be under methods that preserved more of the available data. As a result, confidence intervals will be wider and hypothesis tests will have lower power than those produced by more efficient methods.

Table 1 first shows estimates of the  $\gamma$  parameters and their standard errors for the latent growth model under listwise deletion. These estimates were produced by the CALIS procedure in SAS, which does maximum likelihood estimation under the assumption of multivariate normality. The left side of the table (labeled  $\alpha$ ) gives the effect of each variable on the intercept of the growth line, and the right side (labeled  $\beta$ ) gives the effect of each variable on the slope. Both GENDER and HOMEEMO have highly significant effects on the intercept, both in the expected direction. For the slope, only HOMEEOG has a statistically significant effect, although it just barely makes the critical value for the .05 level.

### Pairwise Deletion

A simple alternative to listwise deletion that apparently preserves more of the data is *pairwise deletion*, also known as available case analysis. This method relies on the fact that esti-

<sup>3</sup> In fact, there is little evidence in this data set of any violations of the MCAR assumption. Logistic regressions for missingness in each of the three measures of antisocial behavior as a function of the observed variables show no significant coefficients, either individually or as a group.

mates for many linear models—including the general SEM estimated in such programs as LISREL, Amos, and EQS—are functions of the first and second moments (i.e., means, variances, and covariances). Under pairwise deletion, each of these moments is estimated using all cases that have data present for each variable or each pair of variables. The resulting moment estimates are then used as input for standard linear modeling software.

If the data are MCAR, pairwise deletion is known to produce parameter estimates that are consistent<sup>4</sup> and, hence, approximately unbiased (Glasser, 1964). However, it also suffers from a number of potential problems. First, the pairwise deleted correlation matrix may not be positive definite, implying that the parameters for many linear models cannot be estimated at all. Second, under some conditions (e.g., regression analysis with high correlations among the explanatory variables), estimates produced by pairwise deletion may actually have greater sampling variance than those produced by listwise deletion (Haitovsky, 1968; Kim & Curry, 1977). Third, and perhaps most important, the standard error estimates obtained under pairwise deletion are not consistent estimates of the true standard errors, calling into question the validity of confidence intervals and hypothesis tests (Glasser, 1964).

Symptomatic of the standard error problem is that a sample size must be specified in order to use a correlation or covariance matrix as input to linear modeling software. But what sample size should be used for pairwise deletion? The size of the complete-case subsample is too small, which leads to overestimates of the standard errors. But the original sample size is too large, leading to underestimates of the standard errors. Unfortunately, there is no single sample size that produces valid estimates of all the standard errors.

Table 1 gives estimates of the gamma parameters in the linear growth model based on pairwise deletion. For the standard errors, I tried two different sample sizes:  $N = 275$  is the number of nonmissing cases for the variable with the most missing data (ANTI4), and  $N = 405$  is the number of cases in the full sample. The former is probably too conservative, whereas the latter is probably too liberal. A few of the parameter estimates are noticeably different from those produced by listwise deletion. For example, the effect of HOME COG on  $\beta$  declines from  $-.038$  to  $-.024$  so that this coefficient is no longer statistically significant, even with the smaller standard error obtained with a sample size of 405.

### *Dummy Variable Adjustment*

Cohen and Cohen (1983) popularized a simple method for handling data that are missing on explanatory variables in a regression model. If variable  $X$  has missing data, create a dummy variable  $D = 1$  if  $X$  is missing and  $D = 0$  if  $X$  is not missing. If  $X$  is missing, set  $X = c$ , where  $c$  is some constant. For example,  $c$  could be 0 or  $c$  could be the mean of  $X$  for cases with data present: It does not make much difference which constant is chosen. Then include both  $X$  and  $D$  as explanatory variables in the regression model. Even though this method has some intuitive appeal, Jones (1996) showed that it typically produces biased estimates of the regression coefficients.

### *Imputation*

Another general approach to missing data is to make some reasonable guesses for the values of the missing data and then proceed to a conventional analysis of the real and imputed data. Of the many methods of imputation, the simplest is unconditional mean imputation: For each variable with missing data, calculate the mean for the nonmissing cases and substitute that for the missing data. Unfortunately, this method yields biased estimates of many parameters (Little & Rubin, 1987). Much better is conditional mean imputation. This can be accomplished by regressing the variable with missing data on other variables in the data set for those cases with complete data. The estimated regression equation is then used to generate predicted values for the cases with missing data. One version of this approach has been shown to produce approximately unbiased estimates of regression coefficients when there are missing data on explanatory variables, under the MCAR assumption (Little, 1992). There is, however, a general tendency for conditional mean imputation to produce underestimates of variances and overestimates of correlations (Little & Rubin, 1987).

Even if one can avoid bias in parameter estimates, all conventional methods of imputation lead to underestimates of standard errors (Little & Rubin, 1987). The reason is quite simple. Standard methods of analysis presume that all the data are real. If some data are imputed, the imputation process introduces additional sampling variability that is not adequately accounted for.

### Maximum Likelihood

The most widely used method for estimating SEMs is maximum likelihood (ML) under the assumption of multivariate normality. When this assumption is met, ML produces parameter estimates that have optimal large-sample properties: consistency, asymptotic efficiency, and asymptotic normality.<sup>5</sup> It turns out that ML can be readily extended to handle missing data under the MAR assumption, and this method is now available in several popular software packages.

Although ML for missing data can be implemented under a variety of distributional assumptions, I shall focus here on methods that are based on multivariate normality. Under this assumption, the parameters of the multivariate distribution are the vector of means ( $\mu$ ) and the covariance matrix ( $\Sigma$ ). In an SEM, these parameters are typically functions of a smaller set of parameters. There are three available approaches to ML estimation of these parameters when some data are missing: factoring the likelihood, the EM algorithm, and direct ML. The method of factoring the likelihood (Marini, Olsen, & Rubin, 1979) is easily implemented,

<sup>4</sup> Consistency means that the estimates converge in probability to the true parameter values as the sample size gets larger. This implies that the estimator is approximately unbiased in large samples, with the magnitude of the bias going to zero with increasing sample size.

<sup>5</sup> Consistency is explained in footnote 4. Efficiency means that an estimator has a sampling variance that is at least as small as that of any other estimator. To say that an estimator is asymptotically efficient means that it is approximately efficient, with the approximation improving with increasing sample size. Similarly, asymptotic normality means that the sampling distribution of the estimator is approximately normal, and the approximation improves with increasing sample size.

but it only works under a rather restrictive pattern of missing data (monotone missing) and does not produce standard errors. For these reasons, I will not discuss it further. The other two methods deserve more detailed consideration.

### *EM Algorithm*

This algorithm is simply a computational device for getting ML estimates of the means and the covariance matrix (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 1997). Once those estimates are produced, they may be used as input to SEM software although, as we shall see, that second step is problematic in a couple of ways.

The algorithm is called EM because it consists of two steps: an expectation step (E-step) and a maximization step (M-step). In general terms, the E-step consists of finding the expected value of the log-likelihood given the current parameter values. (The expectation is taken over the possible values of the missing data.) The M-step consists of maximizing the expected log-likelihood to produce new estimates of the parameters. This step is usually accomplished by conventional methods. These two steps are iterated until there is no change in the parameter estimates from one iteration to the next.

Under multivariate normality and the MAR assumption, the EM algorithm reduces to a kind of iterated linear regression imputation. Here are the steps:

1. Choose starting values for means and covariance matrix.
2. If data are missing on a variable  $X$ , use current values of parameters to calculate the linear regression of  $X$  on other variables. This is done separately for each pattern of missing data (i.e., each set of variables present and variables missing).
3. Use the estimated linear regression equations to impute values of  $x$ .
4. After all data have been imputed, recalculate means and covariance matrix, with corrections for variances and covariances.
5. Repeat Steps 2–4 until convergence.

The corrections in Step 4 are designed to avoid the usual underestimation of variance that arises in imputation. Suppose that for some subset of the observations,  $X$  was imputed using variables  $W$  and  $Z$ . Let  $s_{x.wz}^2$  be the residual variance from that regression. Then, in calculating the variance for  $X$ , wherever you would ordinarily use  $x_i^2$ , substitute  $x_i^2 + s_{x.wz}^2$ . There is a similar correction for covariances that is used whenever data are missing on two variables for the same subject.

Most comprehensive statistical packages have a routine for the EM algorithm under multivariate normality, including SPSS, BMDP, SAS, S-PLUS, and SYSTAT. Not surprisingly, many SEM analysts have used the means and covariance matrix produced by the EM algorithm as input to SEM software. However, this two-step approach is less than ideal, for two reasons. When the SEM to be estimated is just-identified (i.e., the model implies no restrictions on the covariance matrix), then the resulting parameter

estimates are true ML estimates. But in the more usual case when the SEM is overidentified,<sup>6</sup> the resulting estimates are not true ML estimates and are generally less efficient (although the loss of efficiency is likely to be small).

More seriously, regardless of whether the model is just-identified or overidentified, the standard errors reported by SEM software using this two-step method will not be consistent estimates of the true standard errors. This problem is similar to the one discussed above for pairwise deletion. When using the EM covariance matrix as input to an SEM program, a sample size must be specified. But there is no single number that will give correct standard errors for all parameters. Consequently, confidence intervals and  $p$ -values that are based on those standard errors should be treated with some caution. Yuan and Bentler (2000) have proposed methods for getting consistent standard error estimates that are also robust to departures from normality, and these methods are expected to be introduced into EQS 6.0.

For the SRCD data, I used the EM algorithm to produce ML estimates of the means and the covariance matrix using the MI procedure in SAS. As with most EM programs, the only necessary input from the analyst is to specify the set of variables to include in the estimation process. This should obviously include all the variables in the desired SEM model. But it may also be helpful to include other variables that are either related to the variables with missing data or related to the probability that those variables will be missing. The inclusion of such “auxiliary variables” can yield lower standard error estimates for parameters of interest and may also reduce bias (Collins, Schafer, & Kam, 2001). For the example at hand, I included MOMAGE and KIDAGE even though those variables were not in the SEM to be estimated.

After running EM, I used the means and covariances as input to the CALIS procedure in SAS to estimate the latent growth model described above. The parameter estimates are shown in Table 1. As with pairwise deletion, standard errors were estimated under two different sample size specifications,  $N = 275$  (the number of nonmissing cases for the variable with the most missing data) and  $N = 405$  (the number of cases with data present for at least one variable). The coefficient estimates are very similar to those produced by pairwise deletion. The standard errors are also similar under the two different sample size specifications. The only noteworthy differences between these two methods is that the effect of HOME COG on  $\beta$  using the EM method is closer to the listwise deletion estimate than to the pairwise deletion estimate. This coefficient shows up as statistically significant at the .05 level under the more optimistic standard error estimate.

### *Direct ML*

Also known as “raw” ML (because it requires raw data as input) or “full information” ML, direct ML solves the problems that arise in the two-step EM method. When there are no missing data, the likelihood function for independent observations from a multivariate normal distribution can be written as

<sup>6</sup> Overidentified models have fewer parameters than the number of means, variances, and covariances, implying restrictions on those moments.

$$L(\mu, \Sigma) = \prod_i f(x_i | \mu, \Sigma)$$

where  $x_i$  is a vector of observed values for individual  $i$ ,  $f(\cdot)$  is the multivariate normal density function,  $\mu$  is a vector of means and  $\Sigma$  is the covariance matrix. When some data are missing, the likelihood is modified to read

$$L(\mu, \Sigma) = \prod_i f(x_i | \mu_i, \Sigma_i)$$

where  $x_i$  may now be a smaller vector that simply deletes the missing elements from  $x$ ,  $\mu_i$  is the subvector of  $\mu$  that deletes the corresponding elements that are missing from  $x_i$ , and  $\Sigma_i$  is a submatrix of  $\Sigma$  formed by deleting the rows and column corresponding to missing elements of  $x$ . ML estimates are simply the values of  $\mu$  and  $\Sigma$  that maximize this likelihood. In practice, the maximization may be carried out by conventional numerical methods like the Newton-Raphson algorithm. Standard errors are obtained in the usual way by inverting the information matrix.

Direct ML for SEMs was introduced by Allison (1987) and Muthén, Kaplan, & Hollis (1987), who showed that it could be accomplished with standard SEM programs like EQS or LISREL using multiple group options. But this method is only practical when the missing data fall into a small number of patterns, each with a sizable number of cases. More recently, Arbuckle (1996) proposed the use of direct ML for general missing data patterns and implemented the method in the Amos program. Since then, other SEM programs have also introduced direct ML for missing data, including LISREL, M-PLUS, and MX.

I used Amos to estimate the latent growth curve model using direct ML. In Version 4 of Amos, the default is to use direct ML whenever the data set has missing data. In Amos Graphics (the point-and-click interface), the model is specified just as if there were no missing data. In Amos Basic (the text interface), the use of direct ML requires that one explicitly model the means and intercepts, something that would usually be done anyway for a latent growth model. The inclusion of auxiliary variables (e.g., KIDAGE and MOMAGE) requires some additional complications in the model specification. Essentially, the model must allow for correlations of each auxiliary variable with (1) all measured exogenous variables and (2) the error terms for each measured endogenous variable.

Amos Basic code for the latent growth model is displayed in the Appendix. Results are shown in Table 1. Not surprisingly, the parameter estimates are very close to those from the two-step EM method (both are consistent; direct ML estimates are efficient). The standard error estimates for direct ML all lie between the optimistic and pessimistic estimates produced by the EM method.

Given its many advantages, direct ML appears to be the best method for handling missing data for most SEM applications (Duncan, Duncan, & Li, 1998). It is important to keep its limitations in mind, however. First, standard implementations assume that the missing data mechanism is ignorable. Even though a considerably weaker assumption than MCAR (which is required by many conventional methods), it will still be implausible in many situations. Maximum likelihood can also be used in situations where the data are *not* missing at random, but such applications are fraught with difficulties. I will have more to say about that near the end of this article.

Second, the method assumes that the data have a multivariate normal distribution. Of course, that assumption is commonly made

for SEM estimation even without missing data, so it is no great leap to use it for missing data estimation. Nevertheless, most data sets are demonstrably nonnormal, and one should be concerned about the robustness of the methods. Yuan and Bentler (2000) have proposed standard error estimates and test statistics for direct ML (with missing data) that are robust to departures from normality, and these have been incorporated into recent versions of Mplus.

## Multiple Imputation

Although I generally prefer direct ML for SEM estimation, another method worth serious consideration is multiple imputation (MI). MI has statistical properties that closely approach the optimality of ML. The principal advantage of MI is that it can be used in almost any situation, whereas ML is much more restricted in its applications. For example, if you do not have access to a program that does direct ML, you can still use MI to handle the missing data, or if you want to estimate your SEM using some method other than ML, such as two-stage least squares (Bollen, 1995), then MI would be a good choice.

This flexibility comes at some cost, however. First, because random variation is deliberately introduced into the imputation process, MI does not yield a determinate result: Every time you use it, you get different answers. Second, unlike ML, there are many different ways to implement MI for a particular application. Deciding among the various approaches can be a daunting task, especially for the novice user. Still, it is worth understanding this method and having it in your tool kit for those situations where ML just will not do.

Multiple imputation was first proposed by Rubin (1978) and greatly elaborated in Rubin (1987). The method did not see much use, however, until the advent of computationally efficient and user-friendly software (Schafer, 1997). As with ML, the most widely used method for doing multiple imputation is based on the assumptions of ignorability and multivariate normality. Like EM under multivariate normality, the imputation method is based on iterated linear regressions in which each variable with missing data is regressed on other observed variables. The key difference is that random variation is introduced at two different steps in the iterative process.

To understand MI, it is helpful to consider a very simple situation. Suppose we want to estimate the correlation between two variables,  $X$  and  $Y$ , but data on  $X$  are missing (at random) for half the cases in the sample. The missing values could be imputed by regressing  $X$  on  $Y$  for the complete cases and then using the estimated regression line ( $a + bY$ ) to generate predicted values of  $X$ . Together with the real data, these imputed values could be used to calculate the correlation coefficient.

The problem with this method is that it leads to overestimates of the correlation. The reason is simple: For cases with imputed values,  $X$  is a perfect linear function of  $Y$ . So for those cases, the correlation is 1.0. As we saw earlier, the EM algorithm solves this problem by introducing a correction into the estimated variance of  $X$ . With multiple imputation, the solution is to introduce random variation into the imputation. Specifically,  $X$  is imputed from the equation  $a + bY + S_{x,y}E$ , where  $E$  is a random draw from a standard normal distribution, and  $S_{x,y}$  is the estimated standard deviation of the error term in the regression (the root mean squared error). This random variation dampens the correlation between  $Y$

and the imputed values of  $X$ , leading to approximately unbiased estimates of the correlation.

Although random imputation can eliminate the biases that would otherwise result from deterministic imputation, it suffers from two other problems. First, the parameter estimates are not fully efficient because of the added random variation. Second, conventional standard error estimates are too low because they do not take account of the additional variability. The solution to both problems is to repeat the process multiple times, producing  $M$  complete data sets, each with slightly different imputed values. With small to moderate amounts of missing data,  $M = 5$  is usually sufficient. Using standard software, the parameters and their standard errors are estimated on each data set, and the results are combined into a single set of estimates and standard errors.

### Combining the Estimates

To combine the parameter estimates, one simply averages the estimates across the  $M$  replications, thereby greatly reducing their sampling variability. The standard errors are obtained by combining the estimated variability within replications and the variability across replications. The steps are:

1. Square the estimated standard errors (producing variance estimates) and average them across the  $M$  replications.
2. Calculate the variance of the parameter estimates across the replications.
3. Add the results of Steps 1 and 2 (applying a small correction factor to the variance) and take the square root.

Rubin's (1987) formula for accomplishing this is

$$\sqrt{\frac{1}{M} \sum_{k=1}^M s_k^2 + \left(1 + \frac{1}{M}\right) \left(\frac{1}{M-1}\right) \sum_{k=1}^M (b_k - \bar{b})^2},$$

where  $b_k$  is the parameter estimate in replication  $k$ , and  $s_k$  is the standard error of  $b_k$ . This formula is used for virtually all applications of multiple imputation. The resulting standard errors may be used to construct confidence intervals and test statistics.

### Random Draws for Regression Parameters

The method just described is pretty good, but it still produces confidence intervals that are a little too large and hypothesis tests whose true error rates are somewhat greater than the nominal error rates. The reason is that in generating the imputed values, estimated values of the regression coefficients were used rather than the true population values, but the estimated standard errors do not reflect this fact. The solution is to use different regression coefficients for each replication. In the method I use here, the regression coefficients used in the imputations are random draws from normal distributions whose means are the estimated regression coefficients and whose standard deviations are the estimated standard errors of these coefficients. This additional variability produces more variability

in the parameter estimates across replications, leading to larger (and more correct) standard errors.

### Markov Chain Monte Carlo Algorithm

The method I have just described for a single variable with missing data is quite sufficient to produce good estimates of parameters and standard errors. Unfortunately, this method usually runs into difficulty when there is more than one variable with missing data. The basic problem is this: Optimal imputations require optimal estimates of the regression coefficients. But with complex missing data patterns, estimates of the regression parameters can be improved by using imputed data. So there is something of a chicken-and-egg problem.

One solution to this problem is an iterative method known as the Markov chain Monte Carlo (MCMC) algorithm. In recent years, MCMC has become extremely popular for doing Bayesian inference (Gilks, Richardson, & Spiegelhalter, 1996). The algorithm is also known as data augmentation when it is specifically applied to missing data problems (Schafer, 1997; Tanner & Wong, 1987). Under the multivariate normal model, the MCMC algorithm takes the following general form:

1. Choose a set of starting values for the parameters to be estimated, namely, the means and the covariance matrix. Starting values are often obtained with the EM algorithm.
2. For each missing data pattern, use the current parameter values to calculate the linear regression of variables with missing data on variables with observed data.
3. Use the regression coefficients to generate predicted values of the missing data, augmenting these values with random draws from the estimated distribution of residuals.
4. When all the data have been imputed, recalculate the means and the covariance matrix.
5. Make random draws of the parameters (means and covariances) from the posterior distribution (a normal, inverted Wishart distribution).
6. Repeat Steps 2–5 for many iterations. The imputed values from Step 3 for every  $k$ th iteration are set aside for later analysis (with  $k$  chosen by the user).

Software for accomplishing these tasks is now readily available in the stand-alone NORM package (Schafer, 1999). This software has also been incorporated directly into S-PLUS. Other commercial software packages that implement MCMC imputation include LISREL 8.5 and the MI procedure in SAS (release 8.2 and later).

### Example of Multiple Imputation

I used PROC MI in SAS to generate multiple imputations for the SRCED example. Besides specifying an input and output data set, the only information needed from the user is a list of variables to

be used in the imputation process. Thus, the program code needed to run PROC MI is

```
PROC MI DATA=SRCD OUT=SCRDIMP;
VAR ANTI1-ANTI4 GENDER HOMECOG HOMEEMO KIDAGE
MOMAGE;
RUN;
```

As with the EM algorithm, the variable list should include all variables in the model to be estimated, plus any variables that are either associated with the variables that have missing data or with the probability that those variables will have missing data. In this case, I have included KIDAGE and MOMAGE as auxiliary variables.

The default in PROC MI is to produce five completed data sets. However, these are stacked into one SAS data set, which I have named SRCDIMP. Thus, there were 405 cases in the input data set but 2,025 cases in the output data set. The five component data sets are distinguished by the variable `_IMPUTATION_`, which has values 1–5.

The next step is to estimate the latent growth model on each of the five data sets, which I did with the CALIS procedure in SAS. Because the replicate data sets are contained in a single SAS data set, this is easily accomplished with the same CALIS program used above (see Appendix for program code), except for the inclusion of the statement

```
BY _IMPUTATION_;
```

which requests a separate analysis for each value of the variable `_IMPUTATION_`. In addition, I requested that the parameter estimates and their covariance matrix be written to SAS data sets (named A and B) for further analysis.

The last step is to combine the replicate parameter estimates into a single set of estimates using a companion SAS procedure, MIANALYZE. All that is necessary is to specify the name of the data sets produced by PROC CALIS and the names of the parameters to be combined:

```
PROC MIANALYZE PARMS=A COVB=B;
VAR AL1 GAMMA1 GAMMA2 GAMMA3 AL2 GAMMA4 GAMMA5
GAMMA6;
RUN;
```

Results from MIANALYZE are shown in Table 1. Both the parameter estimates and their standard errors are quite similar to those produced by direct ML.

### Complications With Multiple Imputation

For estimating the parameters of SEMs, the program defaults of multiple imputation software should generally produce satisfactory results. However, there are several issues and potential complications that sometimes need to be addressed. For the SRCD example, one such issue is the skewness of the variables being imputed. PROC MI imputes under the assumption that the variables are normally distributed, but the measures of antisocial behavior in the SRCD data set are manifestly nonnormal. For example, here's the frequency distribution of ANTI2:

anti2	Frequency
0	98
1	89
2	71
3	44
4	23
5	16
6	20
7	5
8	5
9	2
10	1
Frequency Missing = 31	

When a variable like this is imputed under a multivariate normal model, the imputed values will tend to be more symmetrically distributed than the original distribution. One consequence is that, in this case, 21% of the imputed values were less than 0.

This is not necessarily a fatal defect. Despite the apparent incongruity between the imputed and the real values, parameter estimates based on these imputations might still be approximately unbiased. Still, it is attractive to consider imputation methods that remove this incongruity. One possibility is to transform the variable in some way to remove the lower bound of 0 before doing the imputations, then reverse the transformation after doing the imputations. For example, letting  $y$  be one of the measures of antisocial behavior, I computed  $x = \ln(y + .5)$  before imputing, then  $y = \exp(x) - .5$  after imputing. Results were very similar to the multiple imputation results in Table 1, except that the effect of HOMECOG on  $\beta$  had a larger standard error and was not statistically significant.

Another approach is to directly impose lower bounds on imputed variables in the imputation process, an option that is available in some software packages that do multiple imputation. The method for accomplishing this is quite simple: Whenever an imputed value is outside a user-specified range, the value is rejected and another is drawn instead. This process may be repeated until an acceptable imputed value is produced. I tried this with the SRCD data, and again the results were very similar to those in Table 1. However, again the effect of HOMECOG on  $\beta$  was not statistically significant.

What about missing data for dummy variables (dichotomous variables with values of 0 or 1)? Under multivariate normality, imputed values can be any real numbers and will sometimes lie outside the (0, 1) interval. Again, this is not necessarily a problem for SEMs because the parameter estimates may still have good properties. But many analysts are uncomfortable with such "impossible" imputations, and some statistical methods require that a variable be a true dichotomy. A simple, but ad hoc solution is to round the imputations so that any value greater than (or equal to) .5 is recoded as 1 and anything less than .5 is recoded as 0. Simulation results (Schafer, 1997) suggest that this method performs well under a variety of circumstances.

In many applications, categorical variables with more than two categories are represented by a set of dummy variables, one fewer than the number of categories. If such variables have missing data, they can be imputed in the usual way. However, the simple rounding rule for single dummy variables may create awkward inconsistencies. For example, suppose that marital status has three categories: never married, currently married, and formerly married. Suppose we create a dummy variable  $N$  for never married,

and a dummy variable  $F$  for formerly married, with currently married as the reference category. For those people with missing data on marital status, it is possible for a person to get imputed values for  $N$  and  $F$  that are both greater than .5. Rounding these up to 1 would mean that the person would be classified as both never married and formerly married. My ad hoc solution for dealing with this problem is this: Impute as usual. Calculate  $1 - N - F$ , to produce an imputed value for  $C$ , the currently married category. Checking the three imputed values, assign the person to the category with the highest imputed value. If this is category  $C$ , both  $N$  and  $F$  would be coded as 0. The extension to four or more categories should be straightforward.

### How Many Imputations?

As previously mentioned, the default in PROC MI is to produce five imputed data sets, and this is widely regarded as sufficient for small to moderate amounts of missing data. But more is always better, both in terms of getting more stable parameter estimates and better estimates of standard errors and test statistics. Table 2 shows how the efficiency of estimates produced by MI depends on the number of imputations and the fraction of missing information. As a standard of comparison, direct ML estimates should be 100% efficient. Even when half of the information is missing, five imputations yields better than 90% efficiency.

There are other considerations besides efficiency of the parameter estimates, however. We also want good estimates of the standard errors and valid confidence intervals that are as narrow as possible. Although there has been little discussion of these issues in the multiple imputation literature, my own informal experience suggests that achieving optimal confidence intervals and hypothesis tests may require substantially more imputations.<sup>7</sup> Another advantage of doing more imputations is that the parameter estimates become more stable from one application of MI to another. Given the potential benefits, if you can afford the computation time, why not do 10, 20, or even more? In SAS, at least, the program code is no more complicated for larger numbers of data sets.

### How Many Iterations?

The default in PROC MI is to do 200 “burn-in” iterations of the MCMC algorithm before producing the first imputed data set.

Table 2  
Percent Efficiency of MI Estimation by Number of Imputations  
 $M$  and Fraction of Missing Information

$M$	Fraction of missing information				
	.10	.30	.50	.70	.90
3	97	91	86	81	77
5	98	94	91	88	85
10	99	97	95	93	92
20	100	99	98	97	96

Note. From “Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst’s Perspective,” by J. L. Schafer and M. K. Olsen, 1998, *Multivariate Behavioral Research*, 33, p. 548. Copyright 1998 by Lawrence Erlbaum Associates. Reprinted with permission.

Thereafter, each imputed data set is separated by 100 iterations. There are two reasons for requiring so many iterations: (1) to ensure that the algorithm has converged to the correct distribution and (2) to avoid any statistical dependence between the observations in one data set and another. Unfortunately, not much is known about just how many iterations are needed to achieve these aims. Graphical diagnostics are available in some computer programs, but these are by no means definitive. I am increasingly of the opinion that the default values in SAS are more than enough for the vast majority of data sets (although more may be needed with large fractions of missing information). If you have computing time to burn, it may be better to go for more data sets than for more iterations between data sets.

### Estimates of Other Parameters

Table 1 reports estimates of the principal parameters of the specified SEM. What about other statistics like squared multiple correlations or one of the many goodness-of-fit measures that are reported by most SEM software? If you are not interested in doing hypothesis tests or confidence intervals on these statistics, then it is only necessary to average them across the replicate data sets. If you want to do statistical inference, then you need both a parameter estimate and a standard error from each completed data set in order to apply Rubin’s formula for the standard error. If the statistic of interest is itself a test statistic—like the chi-square goodness-of-fit statistic—simple averaging will not suffice. Then you need a method for multivariate inference.

### Multivariate Inference

The MI methods just described work fine for statistical tests regarding a single parameter, where all that is needed is an estimate and its standard error. But what if you want to test whether two parameters are equal, or, more generally, what if you want to test whether one model is significantly better than another, more restricted model, a common practice in SEM modeling? There are three available methods for doing this in the context of multiple imputation (Schafer, 1997). First, one can use covariance matrices of the parameter estimates to produce Wald tests. This method is currently available in PROC MIANALYZE for testing whether two or more parameters are all 0. (Later versions will test any linear hypothesis about a set of parameters.) Unfortunately, this method rests on the somewhat unrealistic assumption that the percentage of missing information is the same for all parameter estimates.

A second method, based on likelihood ratio tests, does not make this assumption and is particularly well suited for comparing two nested models. However, it is considerably more complicated to

<sup>7</sup> A potentially useful diagnostic for the quality of confidence intervals and test statistics is the estimate of the degrees of freedom associated with each parameter estimate. This statistic is produced by most multiple imputation software and is used for evaluating a  $t$  distribution or as the denominator degrees of freedom for the  $F$  distribution. It is a function of both the fraction of missing information and the number of imputations and can always be increased by raising the latter. If the degrees of freedom fall appreciably below 100, it is an indication that inference can be improved by adding more imputations.

implement, requiring that the two models be estimated twice on each data set, once in the usual way and again with all parameters constrained at the means for the previous estimates. For details, see Schafer (1997). A third method requires only a chi-square statistic for testing the desired hypothesis in each completed data set. These are then combined in a formula (Allison, 2001; Schafer, 1997) that produces a single  $F$ -statistic. This formula has been incorporated into a SAS macro called COMBCHI (Allison, 2000). For the latent growth model fit to the SRCD data, the chi-squares for comparing the fitted model to the saturated model in the five completed data sets were 36.1, 48.9, 42.0, 47.3, and 45.2, each with 14 degrees of freedom. When these values are plugged into the formula, the result is an  $F$ -statistic of 2.55 with 14 and 35 degrees of freedom, yielding a  $p$ -value of .0125. Although the method of combining chi-squares is very easy to use, it may not be as accurate as the other two methods (Li, Meng, Raghunathan, & Rubin, 1991).

### *Interactions and Nonlinearities*

Under the assumption of multivariate normality, the MCMC algorithm produces imputations that are linear combinations of observed variables, plus random error. Consequently, the method usually does a good job of reproducing the linear relationships among the set of variables used in the imputation process. But what if the model of interest involves nonlinearities or interactions or both?

Suppose, for example, that HOMEEMO is hypothesized to have a different effect on the  $\beta$  coefficient for boys and for girls. That could create a problem because, under the current imputation model, the imputed values for ANTI reflect additive rather than interactive effects of GENDER and HOMEEMO. In fact, when I estimated an SEM model that included the product of these two variables, using the previously imputed data set, the interaction had a  $t$ -value of  $-1.54$ , not statistically significant. Here is a better way. I reimputed the data with the inclusion of the product of GENDER and HOMEEMO as one of the variables in the imputation model. When I reestimated the SEM model, the interaction had a  $t$ -value of  $-2.13$ , significant at the .05 level. An even better way to allow for an interaction is to run the imputation algorithm separately for boys and girls. After combining the imputed data sets for boys and girls, I created the product term for GENDER and HOMEEMO, and reestimated the SEM model. When I did that, the interaction had a  $t$ -value of  $-1.87$ .

The general point is this: The model used to generate imputations should reflect, as much as possible, the model of substantive interest. If the latter has interactions and nonlinearities and interactions, then so should the former. Of course, the importance of this rule depends on the proportion of cases with missing data. If that proportion is small, it is not so critical that the imputation model closely track the model of interest.

### *Other Approaches to Multiple Imputation*

The multiple imputation methods I have described, based on iterated linear regression imputation, are the most widely available and the most easily implemented for a wide variety of applications and data structures. Nevertheless, there are many other approaches, currently available or under development, that may be attractive in some situations. Here is a brief survey of some of those methods.

One of the most glaring limitations of imputation methods based on linear regression is their apparent inappropriateness for imputing categorical variables. As we have seen, ad hoc adjustments often produce satisfactory results. But it would be preferable to have imputation methods that do not assume linearity and normality. For data in which all variables are categorical, Schafer (1997) discussed an imputation method based on the multinomial distribution, possibly with loglinear constraints. To implement this method, he has developed a freeware package called CAT which runs as a function in S-PLUS. For data containing both categorical and quantitative variables, Schafer considered an imputation method that assumes a multinomial distribution for the categorical variables and a multivariate normal distribution for the quantitative variables within each cell of the contingency table (allowing different means but a constant covariance matrix). The method can be implemented with his MIX package (again, a function in S-PLUS).

A rather different approach to multiple imputation with nonnormal data is *sequential generalized regression* (Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). In this method, a different regression model is specified for each variable with missing data. For categorical data, this could be a logistic regression model, either binary or multinomial. For count data, a Poisson regression model could be specified. For quantitative variables, a conventional linear regression model suffices. Imputation (with a random component) is done on an equation-by-equation basis, using imputed values from earlier rounds for the predictor variables. As with linear regression imputation, the process may be iterated many times. A similar method was proposed by Van Buuren and Oudshoorn (1999) who coined the term multiple imputation using chained equations (MICE). Even though this general approach seems to have a lot of promise, there is no theory guaranteeing that the iterations will eventually converge to a correct result.

All the imputation methods discussed so far are parametric in that they make explicit assumptions about the probability distributions of the variables with missing data. There are also a number of nonparametric and partially parametric methods available. The *hot deck* method is an extension of methods long used by the U.S. Census Bureau to impute missing data. This method is best suited to imputation of a single variable  $Y$  using categorical, nonmissing predictors. The data are arrayed in a contingency table based on the predictor variables. If there are cases with missing  $Y$  values within a particular cell of the table, we take one or more of the nonmissing cases in the same cell and use their  $Y$  values to impute the missing  $Y$  values. To avoid bias, the nonmissing cases should obviously be drawn by some kind of probability sampling method, and this naturally lends itself to multiple imputation. However, in order to produce the appropriate amount of sampling variability, it is essential that the cases be drawn by the *approximate Bayesian bootstrap*, an algorithm proposed by Rubin (1987).

Among partially parametric methods, one that has been implemented in several software packages is *predictive mean matching* (Little, 1988). This method begins with a linear regression of the variable with missing data  $Y$  on other variables that do not have missing data. Based on this regression, predicted values of  $Y$  are calculated for all cases, both missing and nonmissing. For each case with missing data, we find a set of cases with complete data whose predicted values of  $Y$  are "close" to the predicted value for

the case with missing data. From this set of cases, we randomly draw one whose  $Y$  value is “donated” to the missing case.

Both the hot deck and predicted mean matching methods can work well when there is only a single variable with missing data or when the missing data follow a simple “monotonic” structure (Little & Rubin, 1987). However, they are awkward to implement in typical data sets that have numerous, overlapping patterns of missing data.

### Nonignorable Missing Data

For both maximum likelihood and multiple imputation, all the methods discussed so far are based on the assumption that the missing data mechanism is ignorable and, hence, that the data are missing at random. Unfortunately, there are often good reasons for believing that data are *not* missing at random. For the SRCD data, for example, it is plausible that those children with higher levels of antisocial behavior are more likely to drop out of the study than are children with lower levels. This is okay if the probability of drop out depends on earlier, observed values of antisocial behavior, but not the later unobserved values (controlling for the earlier values). As I mentioned earlier, however, there is no way to test whether the MAR assumption is violated.

Both maximum likelihood and multiple imputation can be extended to situations where the data are not missing at random. However, such extensions are considerably more difficult to implement and highly vulnerable to misspecification. Maximum likelihood requires that one simultaneously estimate the SEM of interest and a model for the process determining whether data are missing. Such models are prone to underidentification, and software is very hard to come by. Furthermore, a variety of plausible models may fit the data equally well and yet give very different results. Similar problems arise in the use of multiple imputation for nonignorable missing data mechanisms. See Little and Rubin (1987) for more details.

In my judgment, methods for nonignorable missing data should be used with great caution and only if one has a good understanding of the mechanism by which data are missing. In any event, such methods should always be accompanied by a sensitivity analysis exploring the consequences of different modeling assumptions (Little & Rubin, 1987).

### Discussion

Conventional methods for handling missing data leave much to be desired. They may produce biased estimates of parameters or their standard errors (or both). They typically make inefficient use of the data. To the extent that they avoid any of these problems, they rest on the strong assumption that the data are missing completely at random. By contrast, ML methods for handling missing data have nearly optimal statistical properties under the weaker assumption that the data are missing at random, which allows the missingness to depend on observed data.

Structural equation modelers are in a particularly advantageous position for using ML for handling missing data. They commonly use ML under the assumption of multivariate normality to estimate their models, and this is the easiest setting in which to implement ML methods for missing data. Several SEM programs already

implement direct ML estimation for missing data, and I suspect that all the others will soon follow suit.

Although the MAR assumption may be troubling to some, I believe that any general-purpose method for handling missing data will have to invoke this assumption. The alternative is to directly model the missing data process, and this requires extremely good prior knowledge of that mechanism. Furthermore, as of this writing, there is no readily available software to do this kind of modeling, except for very specific kinds of models (e.g., Heckman's, 1976, model for selection bias in linear regression). Even when the data are not missing at random, methods that assume MAR can often produce good results, at least much better than conventional approaches can (Sinharay, Stern, & Russell, 2001).

Multiple imputation has statistical properties that are nearly as good as those of ML. The attraction of MI is that it can easily be implemented with any kind of model or estimation method. Its chief disadvantage is that, unlike ML, it does not produce a determinate result. Furthermore, the great flexibility of the MI principle has allowed for a proliferation of different approaches and algorithms. This can lead to considerable uncertainty and confusion about how best to implement MI in any particular application.

### References

- Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics: I. Review of the literature. *Journal of the American Statistical Association*, *61*, 595–604.
- Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology 1987* (pp. 71–103). Washington, DC: American Sociological Association.
- Allison, P. D. (2000). *COMBCHI*. Retrieved August 7, 2003, from <http://www.ssc.upenn.edu/~allison/combchi.sas>
- Allison, P. D. (2001). *Missing data*. Thousand Oaks, CA: Sage.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.
- Bollen, K. A. (1995). Structural equation models that are nonlinear in latent variables: A least squares estimator. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 223–251). Boston: Blackwell Publishers.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression and correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351.
- Curran, P. J. (1997). *A fully worked example of a growth curve analysis*. Retrieved August 7, 2003, from <http://www.unc.edu/~curran/example.html>
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.
- Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analysis with partial missingness. *Structural Equation Modeling*, *5*, 1–21.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.
- Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, *59*, 834–844.
- Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, Series B*, *30*, 67–82.

- Heckman, J. J. (1976). The common structure of statistical models of truncated, sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement*, 5, 475–492.
- Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91, 222–230.
- Kim, J.-O., & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6, 215–240.
- Li, K. H., Meng, X. L., Raghunathan, T. E., & Rubin, D. B. (1991). Significance levels from repeated  $p$ -values and multiply imputed data. *Statistica Sinica*, 1, 65–92.
- Little, R. J. A. (1988). Missing data in large surveys (with discussion). *Journal of Business and Economic Statistics*, 6, 287–301.
- 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., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Marini, M. M., Olsen, A. R., & Rubin, D. (1979). Maximum likelihood estimation in panel studies with missing data. In K. F. Schuessler (Ed.), *Sociological methodology 1980* (pp. 314–357). San Francisco: Jossey-Bass.
- McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York: Wiley.
- Muthén, B., Kaplan, K., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 42, 431–462.
- 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.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Rubin, D. B. (1978). Multiple imputation in sample surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20–34.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L. (1999). *Software for multiple imputation*. Retrieved August 7, 2003, from <http://www.stat.psu.edu/~jls/misoftwa.html>
- Schafer, J. L., & Graham, J. W. (2000). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33, 545–571.
- Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, 6, 317–329.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528–550.
- Van Buuren, S., & Oudshoorn, K. (1999). *Flexible multiple imputation by MICE*. Retrieved October 22, 2003, from <http://www.multiple-imputation.com/>
- Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp. 1–24). Mahwah, NJ: Erlbaum.
- Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. P. Becker (Ed.), *Sociological methodology 2000* (pp. 165–200). Boston: Blackwell Publishers.

## Appendix

## Amos Program Code for Maximum Likelihood Estimation

```

Sub Main
Dim Sem As New AmosEngine
Sem.TableOutput
Sem.Smc
Sem.ModelMeansAndIntercepts
Sem.BeginGroup "c:\allison\msfiles\rcd.sav"
Sem.Mean "homecog"
Sem.Mean "homeemo"
Sem.Mean "gender"
Sem.Mean "kidage"
Sem.Mean "momage"
Sem.Structure "anti1 = (1)f1 + (0)f2 + (1)e1"
Sem.Structure "anti2 = (1)f1 + (1)f2 + (1)e2"
Sem.Structure "anti3 = (1)f1 + (2)f2 + (1)e3"
Sem.Structure "anti4 = (1)f1 + (3)f2 + (1)e4"
Sem.Structure "f1 = () + gender + homecog+homeemo+(1)d1"
Sem.Structure "f2 = () + gender + homecog+homeemo+(1)d2"
Sem.Structure "e1 (th)"
Sem.Structure "e2 (th)"
Sem.Structure "e3 (th)"
Sem.Structure "e4 (th)"
Sem.Structure "d1<>d2"
Sem.Structure "kidage<>e1"
Sem.Structure "kidage<>e2"
Sem.Structure "kidage<>e3"
Sem.Structure "kidage<>e4"
Sem.Structure "momage<>e1"
Sem.Structure "momage<>e2"
Sem.Structure "momage<>e3"
Sem.Structure "momage<>e4"
Sem.FitModel
End Sub

```

## SAS Program Code for Multiple Imputation

```

proc mi data=my.srcd out=srcondimp;
var anti1-anti4 gender homecog homeemo kidage momage;
run;

proc calis data=srcondimp ucov aug privec pcoves;
lineqs
  anti1 = 1f1 + 0f2 + e1,
  anti2 = 1f1 + 1f2 + e2,
  anti3 = 1f1 + 2f2 + e3,
  anti4 = 1f1 + 3f2 + e4,
  f1 = all intercept + gamma1 gender + gamma2 homecog + gamma5 homeemo + d1,
  f2 = a12 intercept + gamma3 gender + gamma4 homecog + gamma6 homeemo + d2;
std
  e1-e4 = th th th th,
  d1-d2 = ph11 ph22;
cov
  d1 & d2 = ph21;
by _imputation_;
ods output Estimates=a covmat=b;
run;

proc mianalyze parms=a covb=b;
var all gamma1 gamma2 gamma5 a12 gamma3 gamma4 gamma6;
run;

```