Werner Wothke, SmallWaters Corp.

Longitudinal and multi-group modeling with missing data

Reprinted with permission from

T.D. Little, K.U. Schnabel and J. Baumert [Eds.] (2000)

Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.

Mahwah, NJ: Lawrence Erlbaum Associates

Summary

Missing data are almost always a problem in longitudinal research. Item non-response, differential attrition, failure to obtain measurements at equal time intervals, and unbalanced panel designs used to be difficult to analyze at best and remain a threat to the validity of a study. A related technical problem, customarily given little importance but nevertheless strongly related to validity threats, is that most multivariate methods require complete data.

Incomplete data are often dealt with by listwise or pairwise deletion methods, which omit entire records, or pairs of variables, with missing values. Sometimes a researcher will substitute sample means for the missing values. All three approaches aim to fix up the data so that they can be analyzed by methods designed for complete data but are ad hoc and have little theoretical justification.

The method of *full-information maximum likelihood* (FIML), in contrast, has long been known as a theory-based approach to the treatment of missing data. FIML assumes multivariate normality, and maximizes the likelihood of the model given the observed data. The theoretical advantages of this full-information method are widely recognized, and it is now implemented in the Amos and Mx structural equation modeling programs.

Unfortunately, theory has not had much influence on practice in the treatment of missing data. In part, the under-utilization of maximum likelihood estimation in the presence of missing data may be due to the unavailability of the method as a standard option in packaged data- analysis programs. There may also exist a (mistaken) belief that the benefits of using maximum likelihood (ML) estimation rather than conventional missing-data techniques will in practice be small.

This paper presents several examples of time-structured and multi-group problems demonstrating the ease of FIML and its greater statistical efficiency when compared to mean-imputation and listwise or pairwise deletion methods. Model specifications for these problems and Visual Basic code used in simulations are available on the World Wide Web at the locations
http://www.mpib-berlin.mpg.de/research_resources/index.html and
http://www.smallwaters.com/books/mpi_modeling_code.html.

Current practice in the treatment of missing data

The most commonly practiced methods for structural equation modeling (SEM) with missing data apply complete-data ML estimation to covariance matrices that have been somehow corrected. Such corrections can be

(a) *listwise deletion* (LD), which excludes from the calculations all records with missing values on any of the variables,

(b) *pairwise deletion* (PD), by which each sample covariance between two variables is computed from pairwise-complete data, excluding cases with missing values on one or both of the variables, or

(c) *mean-imputation* (MI) which replaces the missing values of a variable by the mean of its observed values.

Brown (1983) studied LD, PD, MI and FIML methods by Monte-Carlo simulation in the factor analysis context, Brown (1994) studied the performance of LD, PD and MI by Monte-Carlo simulation in the context of structural equation modeling, and Little and Rubin (1987) reviewed all four methods in the general multivariate case. All three studies are critical of mean-imputation and listwise and pairwise deletion methods, citing biased and/or inefficient estimates as well as the increased potential of obtaining indefinite sample covariance matrices. Brown (1983) qualifies his comments about LD, PD, and MI with respect to frequency and type of the missing data.

Model-based imputation of missing values is well known in the statistical literature but rarely used in structural equation modeling (Kim and Curry, 1977; Roth, 1994). In particular, the EM algorithm (Dempster, Laird and Rubin, 1977), which implements the FIML approach by repeated imputation-estimation cycles, has recently been discussed as a method for estimating means and covariance matrices from incomplete data (Graham, *et al.*, 1997; Graham and Hofer, 2000; Rovine, 1994; Verleye, 1996). However, the EM algorithm has, to my knowledge, not been incorporated in a generally available computer program for structural equation modeling.

Maximum likelihood estimation with incomplete data

The principles of ML estimation with incomplete data are well known (Hartley and Hocking, 1971; Dempster, Laird and Rubin, 1977; Little and Rubin, 1987, 1989; Rubin, 1976; Wilks, 1932). Allison (1987) and Muthén, Kaplan and Hollis (1987) show how the method applies to structural equation modeling. Unfortunately, their approaches are only practical when the data have just a few distinct patterns of missing data. They also require an exceptionally high level of technical expertise in the use of particular SEM programs. At present, ML estimation with missing data is a standard option in at least two structural equation modeling programs, Amos (Arbuckle, 1995) and Mx (Neale, 1994). Both maximize the case wise likelihood of the observed data, computed by minimizing the function

$$C(\gamma) = \sum_{i=1}^{N} \log\left|\boldsymbol{\Sigma}_{i,mm}\right| + \sum_{i=1}^{N} (\mathbf{y}_{i,m} - \boldsymbol{\mu}_{i,m})' \boldsymbol{\Sigma}_{i,mm}^{-1} (\mathbf{y}_{i,m} - \boldsymbol{\mu}_{i,m}), \qquad (1)$$

where $\mathbf{y}_{i,m}$ is the observed (or *m*easured) portion of the data vector for case *i*, and $\boldsymbol{\mu}_{i,m}$ and $\boldsymbol{\Sigma}_{i,mm}$ are the mean vector and covariance matrix parameters but with only the rows and columns corresponding to the observed portions of the data vector for case *i*. Thus, the Amos and Mx programs are not limited by the number of missing-data patterns, and do not require the user to take elaborate steps to accommodate missing data.

Numeric example

Consider the data set:

Case	V1	V2	V3
1	13	23	21
2	14	22	17
3	15	—	11
4	16	18	—
5	17	17	12
6	—	20	8
7	—	20	15

There are three variables (V1 - V3) and seven cases (1-7). Four of the possible 21 observations are missing, as indicated by a '—' symbol. There are four different missingness patterns.

Possible alternatives for estimating means, variances and covariances from the incomplete data set are

- **Listwise deletion:**

  All cases with missing observations are dropped from the computations. The complete-data formulae are then applied to the complete cases (here: cases 1,2, and 5). The estimates are:

cov	V1	V2	V3
V1	4.33		
V2	-6.67	10.33	
V3	-9.17	13.83	20.33
mean	14.67	20.67	16.67

  In this example, the LD method discards the records of four of the seven cases from calculations. Obviously, LD does not make efficient use of the observed data.

- **Pairwise deletion:**

  For each variable, PD computes mean and variance estimates from the univariate complete data. For each pair of variables, PD calculates the covariance estimates from all cases with complete observations on both variables; for instance, the covariance estimate for variables V1 and V2 would be based on cases 1, 2, 4 and 5:

cov	V1	V2	V3
V1	2.50		
V2	-5.33	5.20	
V3	-6.58	7.95	21.60
mean	15.00	20.00	14.00

  PD apparently uses more information from the data and should thus provide more efficient method than LD. On the other hand, analysis of PD covariance matrices presents some known statistical problems that are often overlooked. For once, each entry of such a matrix can be based on a different sample size, and this possibility imposes considerable complications on deriving the joint statistical distribution of the entries of the covariance matrix. In particular, the joint distribution of the elements of a PD covariance matrix cannot usually be considered Wishart[1], even when the matrix is computed from multinormal data. As a conseqence, it is not clear how the fit of a model to a PD covariance matrix can be statistically evaluated . A second, often more obvious issue is that the elements of the

covariance matrix are estimated not just from different sample sizes but more generally from different portions of the data set, and this can lead to inconsistencies. For instance, the value of -5.33 for cov(V1,V2) corresponds to a correlation of r = -1.48, which is an inadmissible value. PD-based sample covariance matrices are a common source of indefiniteness problems in structural equation modeling (Wothke, 1993).

- **Mean imputation:**

  Each missing value is replaced with the mean observed value of the same variable. In other words, MI is an attempt to make the raw data matrix complete. Afterwards means and covariances can be calculated as if from complete data:

Cov	V1	V2	V3
V1	1.67		
V2	-2.67	4.33	
V3	-3.50	5.50	18.00
Mean	15.00	20.00	14.00

  MI yields at the same sample means as PD. Since MI's missing-data replacements happen to be the PD means, this should hardly be surprising. The variance estimates under MI are clearly smaller than those obtained under PD. This is a function of the MI algorithm: Brown (1994) and Little and Rubin (1987) point out that variance estimates under MI are generally negatively biased. The covariance estimates are also different from either LD or PD. Depending on the pattern of missing data, MI covariance estimates may be systematically larger or systematically smaller than those obtained by LD or PD.

  On the positive side, MI does not share the indefiniteness problems encountered under PD: Covariance matrices computed under MI must be positive definite or semi-definite.

- **Full-information maximum likelihood:**

  The FIML estimates of the means and covariances are obtained by maximizing (1) with respect to first and second moments:

cov	V1	V2	V3
V1	1.44		
V2	-2.29	3.73	
V3	-3.59	6.14	19.48
mean	14.98	19.98	13.31

  This FIML estimate uses all the information of the observed data, including information about the mean and variance of missing portions of a variable, given the observed portion(s) of other variables. Even though the indefiniteness problem observed with the PD estimate may also occur with FIML estimation, it does not seem to be as frequent a problem. In the present case, the FIML covariance matrix estimate is positive definite.

Obviously, the four methods of computing means and covariance matrices from incomplete data can produce radically different solutions, even when the exact same data are used. These method differences depend on several factors, including the proportion of data missing and the type of process(es) causing the incompleteness of the data.

Missing-data mechanisms

In order to state the advantages of ML estimation over MI, PD, and LD, it is necessary to consider the mechanisms by which missing data can arise. Rubin (1976) and Little and Rubin (1987) distinguish the processes that generate the missing data with respect to the information they provide about the unobserved data. Missing values of a random variable *Y* can be *missing completely at random* (MCAR), *missing at random* (MAR), or *nonignorable*. Under an MCAR process, the fact that a variable's data are observed or missing is not thought to affect its distribution, i.e.,

$$P(Y|\, y \text{ missing}) = P(Y|\, y \text{ observed})$$

In this paper, MCAR is the most restrictive assumption considered for missing-data processes. MCAR can sometimes be established in behavioral and social surveys by randomly assigning test booklets or blocks of survey questions to different respondents.

MAR is a more relaxed condition, assuming only that missing and observed distributions of *Y* are identical, conditional on a set of predictor or stratifying variables **X**, *i.e.*,

$$P(Y|\, y \text{ missing}, \mathbf{X}) = P(Y|\, y \text{ observed}, \mathbf{X})$$

One way to establish MAR processes is to include completely observed variables **X** that are highly predictive of *Y*. For instance, inasmuch as past behavior is an effective predictor of future behavior, initial (complete) measurement(s) in longitudinal designs can be a good choice of **X**.

The performance of the four methods under different types of missing data processes is summarized by Little and Schenker (1995). For data that are missing completely at random, PD and LD estimates are consistent, although not efficient. MI is consistent in the first moments, but yields biased variance and covariance estimates. If the data are only MAR, then PD and LD estimates also may yield biased results. ML estimates, on the other hand, are already both consistent and efficient when the data are only MAR. In addition, some authors have suggested that ML estimates will tend to show less bias than estimates based on MI, LD or PD, even when the data deviate from MAR (Little and Rubin, 1989; Muthén, Kaplan and Hollis, 1987). As final shortcomings, PD does not provide standard errors of parameter estimates or tests of model fit, while MI can produce standard error estimates and fit statistics that are far too optimistic.

Application: Growth curve modeling

Simulation 1: MCAR data

To demonstrate the efficiency of ML estimation relative to MI, LD, and PD for a single, fairly typical estimation problem with MCAR data, a small Monte-Carlo simulation was undertaken. The variable names and parameter values are taken from a reanalysis of STEP science data collected by Hilton and Beaton (1971, *pp.* 343–344). To keep things simple, the Monte-Carlo simulation uses a multivariate normal distribution, with structural parameters provided by the path model shown in Figure 1. Suppose that the STEP science test was administered to the same students on four occasions—in 1961, 1963, 1965, and 1967. The substantive interest is to gauge the growth of the science scores over the four test occasions. However, this task is somewhat complicated by the measurement error in the test scores.

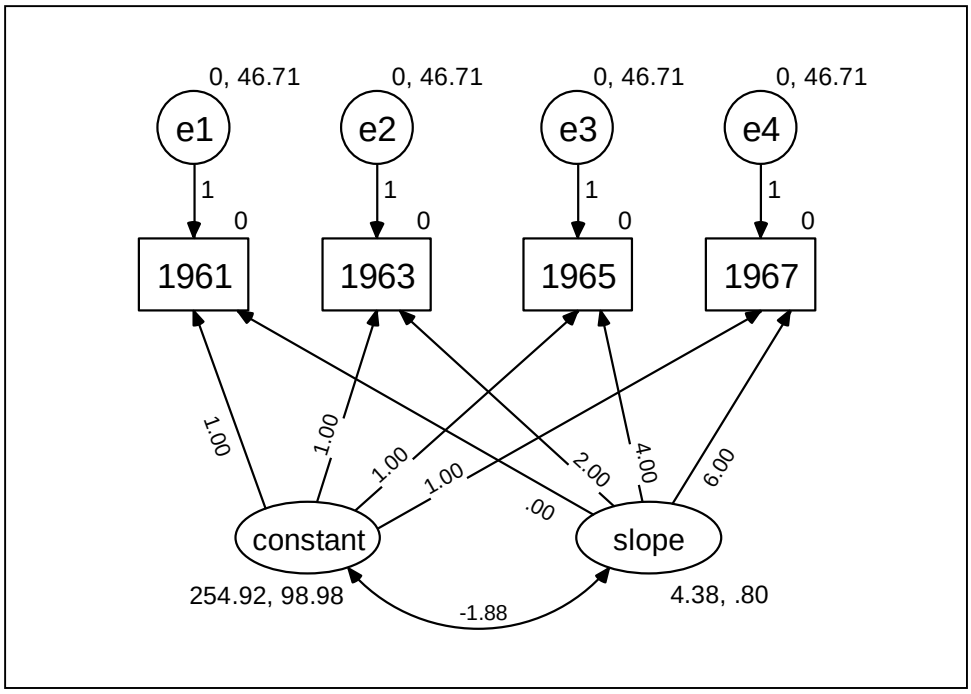

Figure 1: Parameters of a linear growth model (from STEP Science Test, students with high school educated fathers)

The model of Figure 1 is essentially the MANOVA approach to growth-curve modeling (Bock, 1975). Each test score is composed of a constant term, a slope, and a residual. The constant and slope terms are modeled as correlated random components, summarizing the individual differences of both initial level and subsequent improvement of the students' science knowledge. The constant term, with mean 254.92 and variance 98.98, is connected to the observed variables with fixed weights of unity. In addition, the constant term makes the *only* systematic contribution to the 1961 science knowledge test scores. Thus, the mean (254.92) of the constant term gives the mean of the 1961 scores in science knowledge, while its variance (98.98) describes the systematic dispersion of 1961 science knowledge among the group of students. The residual variance of 46.71 provides an estimate of the measurement error of the STEP science test.

The four paths pointing from the slope term to the four observed variables have their coefficients fixed to a linear trend (0, 2, 4, 6). These fixed coefficients reflect the number of years since the first measurement (in 1961). Hence, the slope mean of 4.38 shows the average growth in science scores per year, and its variance of 0.80 quantifies the inter-individual variation of the yearly slope. Constant and slope have somewhat of a negative correlation (r = -0.21), implying that students who start out with high 1961 scores have smaller subsequent gains than students with lower initial scores.

The Monte-Carlo simulation employed 400 random data sets of sample size 500, generated from a multivariate normal distribution with the parameters of Figure 1. The four variables had some of their values deleted completely at random. Missing data probabilities were 0% (none) for 1961, 10% for 1963, 20% for 1965, and 30% for 1967. Multivariate normal quasi-random numbers were generated by the RANLIB.C routines (Brown and Lovato, 1994) implementing the algorithms by L'Ecuyer and Côté (1991) and Ahrens and Dieter (1973).

Realized missing data rates varied somewhat within samples, because the MCAR process was executed independently on each individual observed value. The growth model of Figure 1 was fitted to each Monte-Carlo sample using MI, ML, LD, and PD methods. Altogether, the simulation comprised 1600 attempts to fit the model[2]. This task was automated by calling the open programming AmosEngine interface from a Visual Basic routine.

The performance of a single estimation method (say, ML) was assessed in the following way: First, the method was applied to estimate the model for each of 400 Monte-Carlo samples, then the accuracy of the estimates was judged by comparing them to the bootstrap population parameters in Figure 1. The question is which, if any, of the four methods reproduces the parameter values most closely.

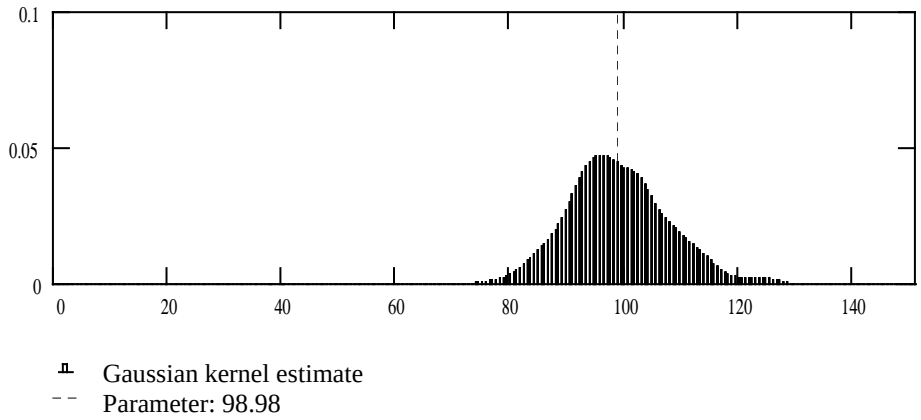

Figure 2: Distribution of v(constant) estimates, MCAR with FIML. Var = 74.13

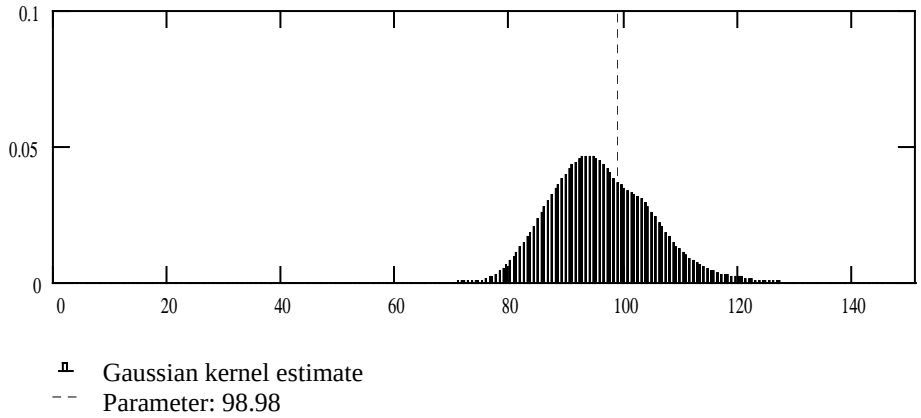

Figure 3: Distribution of v(constant) estimates, MCAR with MI. Var = 73.62

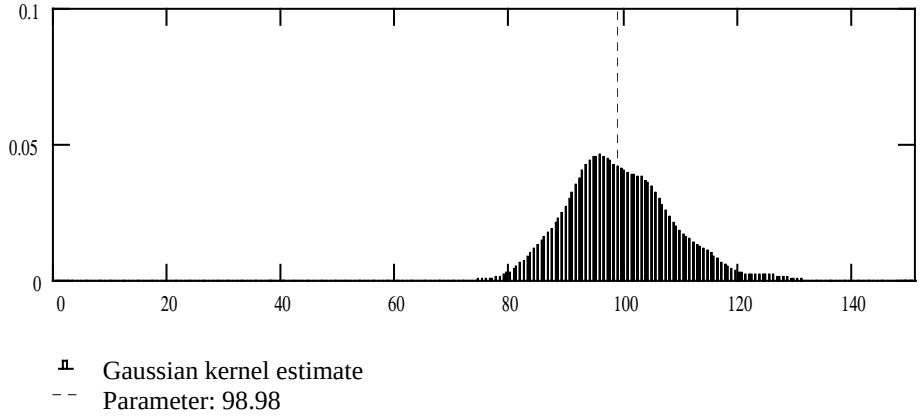

Figure 4: Distribution of v(constant) estimates, MCAR with PD. Var = 78.67

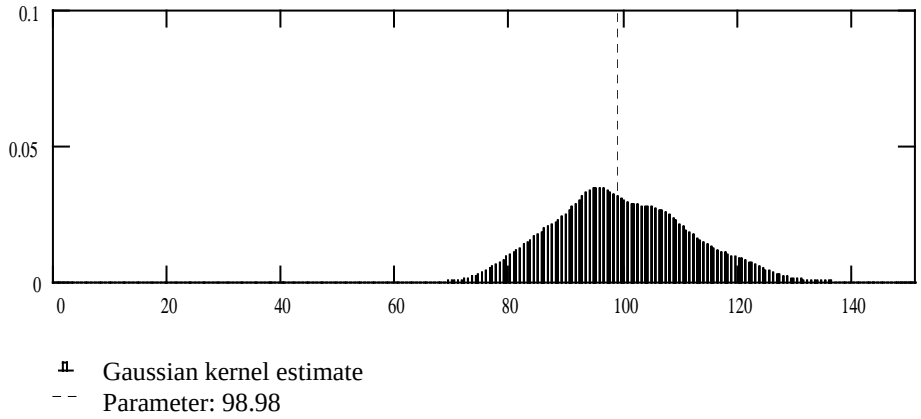

Figure 5: Distribution of v(constant) estimates, MCAR with LD. Var = 133.17

Figures 2–5 display the distributions of the variance estimates for the constant term, with the parameter value of 98.98 indicated by a vertical dotted line. With the exception of the MI estimate, which is biased downward by three percent, the distributions are centered on the parameter value within the margins of

sampling error. Estimation bias for this parameter thus appears to be negligible under the FIML, PD and LD methods, although the data indicate some difference in precision (or efficiency) of estimation. The relative sampling variance can be used to estimate relative gains in efficiency. Under asymptotic theory, the sampling variances of means, regression coefficients and variances are inversely related to sample size Kendal and Stuart, 1977, *p.* 258). Thus, when estimating the variance of the constant term, switching from FIML to LD nearly doubles its sampling variance (1.80 = 133.17/74.13). According to this asymptotic rule, the sample size for LD would have to be increased by approximately 80% (*i.e.*, to N = 900) in order to achieve the degree of precision provided by FIML at N=500.

Table 1: Model estimates under simulated MCAR

Estimation Type	Statistic	constant			slope		
		mean	var	cov	mean	var	ve
FIML	Mean	254.95	98.99	-1.91	4.37	.80	46.56
	s.e.	*.51*	*8.61*	*1.21*	*.09*	*.28*	*2.47*
MI	Mean	254.95	96.00	-5.74	4.37	.73	51.43
	s.e.	*.51*	*8.58*	*1.26*	*.10*	*.24*	*2.36*
PD	Mean	254.95	99.24	-1.93	4.37	.80	46.71
	s.e.	*.51*	*8.87*	*1.30*	*.10*	*.28*	*2.55*
LD	Mean	254.94	99.53	-2.00	4.37	.81	46.59
	s.e.	*.76*	*11.54*	*1.54*	*.11*	*.32*	*2.90*
Parameter value		254.92	98.98	-1.88	4.38	.80	46.71

It has to be stressed that this relative figure of 80% is specific to a single parameter, a single sample size, and a particular choice of missing-data rates. Table 1 shows means and empirical standard errors for all six parameters in the model. These statistics are affected by an undetermined amount of sampling error. Nevertheless, some broad trends are apparent. As expected with MCAR, the FIML, PD, and LD estimates are all unbiased. The MI method, while unbiased in means, produces notably biased variance and covariance estimates. Among the unbiased estimation methods, FIML yields the most efficient estimates—standard errors generally increase as one moves from FIML estimation to PD, and then to LD. The size of the overall change in precision is difficult to gauge, however, as it depends on characteristics of the model, complete-case population mean and covariance structure, sample size, and frequency of missing data. Simulation work by Arbuckle (1996), Graham *et al*. (2000) and Verleye (1996) suggests that the relative efficiency of FIML increases as the missing data rate increases.

Simulation 2: A Case of MAR data

A second Monte-Carlo simulation[3] was performed to illustrate the benefits of ML with data that are MAR but not MCAR. The structural model of Figure 1 was used again for this simulation. The sample size was 500, except that the MAR process was set up to simulate a selective drop-out mechanism in three stages: First, 80% of the cases with a 1961 knowledge score of less than 246 had their 1963-1967 measurements deleted. Second, of the remaining cases, 80% of those with a 1963 score of less than 255 had both their 1965 and 1967 scores set to missing. Finally, 80% of the remaining cases with 1965 scores of less than 263 had their 1967 values set to missing. The realized missing data rates were 0% in 1961, 16–20% in 1963, 25–33% in 1965, and 30–40% in 1967, with differences in these proportions due to sampling variation.

This type of MAR process emulates the situation in which a person participates in a study for some time and then drops out after showing a low score and encountering other, presumably random, conditions. It is particularly easy to see how MI and LD would lead to biased estimates in this situation, as selectively removing records with low scores or substituting the means of the remaining higher scores would affect both means and covariances of the remaining sample. The effect of the missing-data pattern on PD is not so clear-cut. We have already observed the superior efficiency of FIML estimates in the MCAR simulation where PD and LD estimates were known to be unbiased. By contrast, because the present data are only MAR, estimation bias is now of central concern with all estimates.

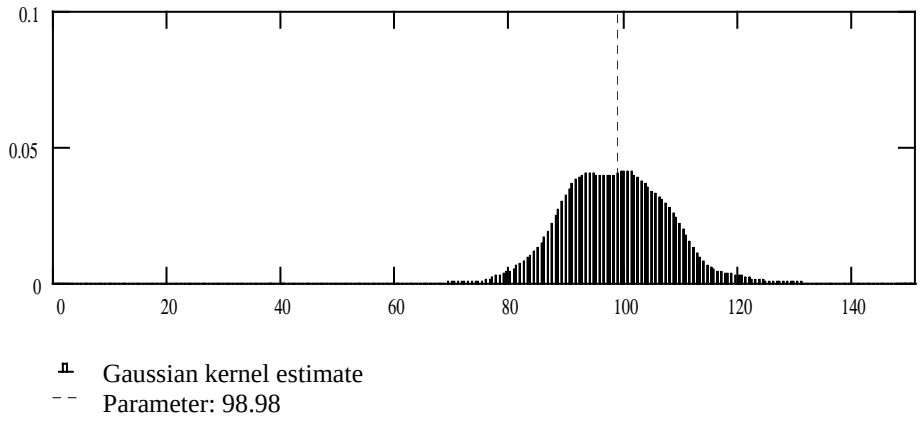

Figure 6: Distribution of v(constant) estimates, MAR with FIML. Var = 77.62

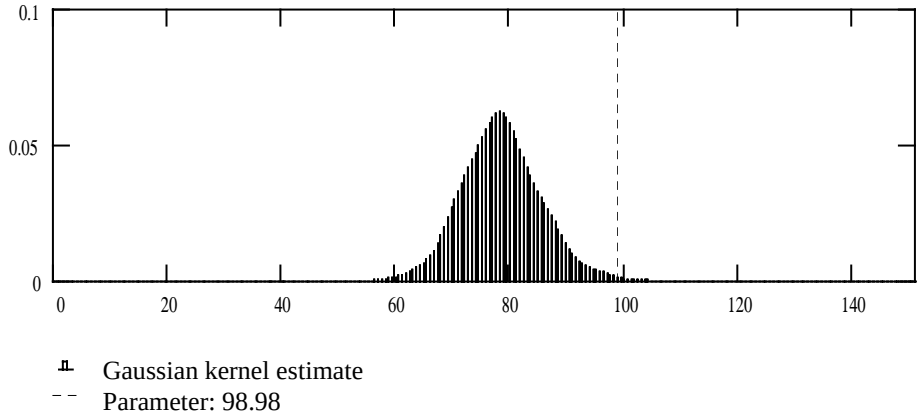

Figure 7: Distribution of v(constant) estimates, MAR with MI. Var = 46.24

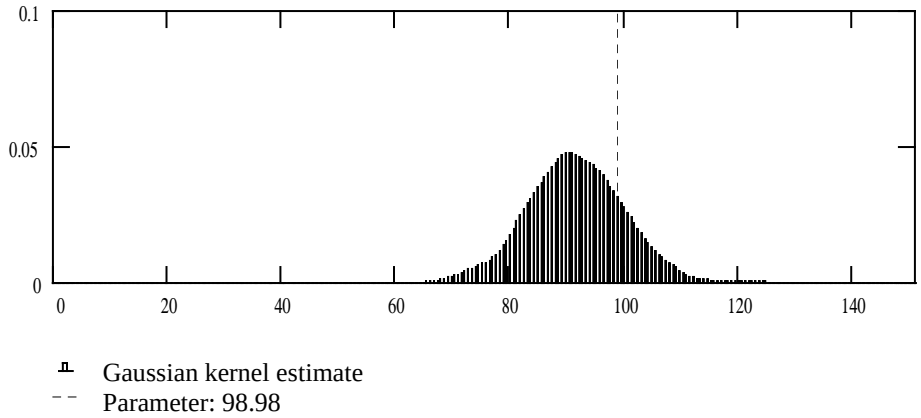

Figure 8: Distribution of v(constant) estimates, MAR with PD. Var = 70.22

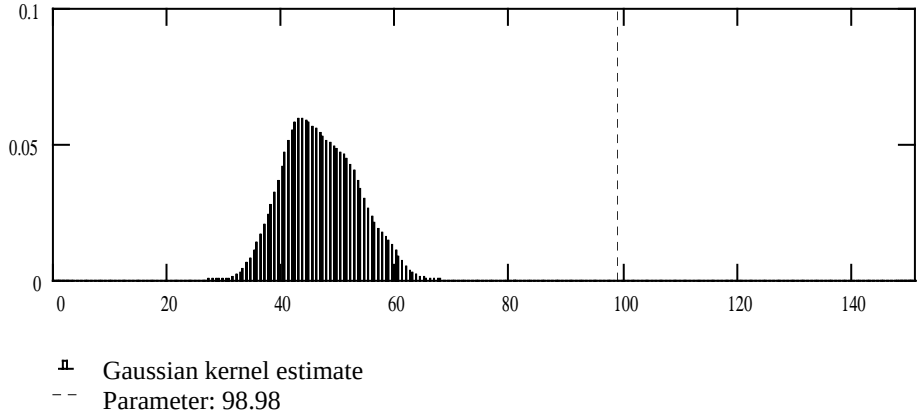

Figure 9: Distribution of v(constant) estimates, MAR with LD. Var = 39.69

Examples of method-specific estimation bias with MAR data are shown in Figures 6–9. While the FIML variance estimate of the constant is neatly centered at the parameter value—although perhaps with a somewhat large sampling variance—the MI, PD, and LD estimates all show negative bias. PD estimates are biased downwards by a moderate degree (approximately 7.5%), but the sampling distributions of MI and LD estimates do not even appear to include the parameter value. Particularly, the LD estimates are all located below 70. Note that the sampling variance under MI and LD is at least 40% smaller than under FIML. One might summarize that MI and LD yield very precise estimates of exactly the wrong parameter.

Table 2: Model estimates under simulated MAR

Estimation Type	Statistic	Constant			slope		
		mean	var	cov	mean	var	ve
FIML	mean	254.95	98.53	-1.93	4.38	.80	46.53
	s.e.	*.53*	*8.81*	*1.49*	*.11*	*.30*	*2.54*
MI	mean	255.33	78.57	-8.62	5.13	1.59	51.24
	s.e.	*.51*	*6.80*	*1.19*	*.11*	*.24*	*2.17*
PD	mean	255.33	91.58	-5.87	5.13	1.22	50.75
	s.e.	*.51*	*8.38*	*1.47*	*.11*	*.34*	*2.69*
LD	mean	260.57	46.93	-.79	4.24	.80	42.02
	s.e.	*.47*	*6.30*	*1.00*	*.09*	*.26*	*2.38*
Parameter value		254.92	98.98	-1.88	4.38	.80	46.71

Table 2 shows the means and standard errors of the six parameter estimates computed by the four estimation methods averaged across 400 samples of size 500. For almost every parameter, FIML provides the estimate with the least bias. Due to the missing data process being MAR instead of MCAR, the MI, PD and LD methods are not only biased in variances and covariances but also in the mean parameters of the constant and slope terms. For several parameters, estimation is dramatically better with FIML than with PD and LD.

Summary

It is impossible to put a single figure on the gain in accuracy of estimation to be had by abandoning MI, PD, and LD in favor of FIML. It is hard to imagine a situation, though, in which FIML would yield worse results than MI, PD, or LD. The advantage of FIML depends on the missing-data rate, the covariance structure of the data and size of the sample, and it differs from one parameter to another. Nevertheless, the two simulations demonstrate that FIML can be superior to PD, and superior to MI and LD by a wide margin.

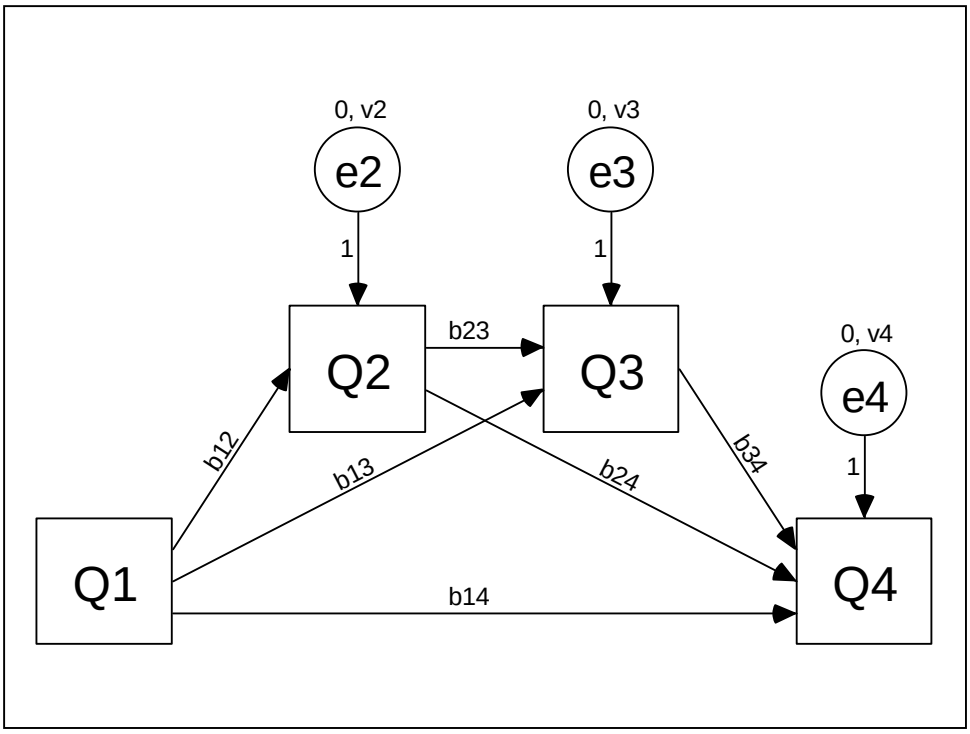

Figure 10: Time-dependent process—model specification

Application: Autoregressive Process

A series of time-dependent autoregression and Markov models (Jöreskog, 1977) demonstrates the ease of analyzing incomplete longitudinal data, testing all models simultaneously against an artificial sample. James Arbuckle first presented this example at the 1996 meetings of the American Educational Research Association. The Amos Graphics specification of the most general model appears in Figure 10. There are four time-dependent variables, Q1, Q2, Q3, and Q4, which can be thought of as four consecutive measurements of the same variable or quantity. Each observed variables is modeled as a linear function of the earlier variables, plus a random shock or residual term.

The three-equation model is just-identified. It is has zero degrees of freedom and cannot be rejected by a global test of fit. Note that the path diagram shows nine parameters with the distinct labels b12, b13, b14, b23, b24, b34, v2, v3, and v4. These labels may be used in constraints defining submodels that *can* be tested against the data. Three submodels might be considered interesting in this type of longitudinal application:

1. **Saturated model**. This is the model of Figure 10 without any constraints. It is a descriptive account of a four-occasion longitudinal design. The measurement at a given occasion is a linear function of the preceding measurements. While the model itself cannot be tested, the parameter estimates and their approximate standard errors may be of exploratory value.
2. **Markov model**. In a Markov model, the values of a time-dependent variable are dependent only on the values of the previous occasion. In other words, there are neither lag-2 nor lag-3 effects. A (linear) Markov model can be defined by the three constraints:

   ```
   b13 = b14 = b24 = 0.
   ```

   Assuming normality, model fit can be assessed by a χ^2 test with three degrees of freedom.

3. **Stationary Markov model**. A Markov model with time-invariant prediction equations is called stationary. Ignoring intercept terms, two equality constraints are required to make the regression weights stationary:

```
b12 = b23 = b34
```

and two additional constraints are needed to render the residual variance terms stationary as well:

```
v2 = v3 = v4.
```

The stationary Markov model can be tested under normality by the χ^2 fit-statistic with seven degrees of freedom. Incremental change of fit from the general Markov model can be assessed by the likelihood ratio χ^2 test with four degrees of freedom.

Table 3: Input data for autoregressive process

```
Case, Q1, Q2, Q3, Q4
 1, 19, 14, 15, 17
 2, 19, 16, 17,
 3, 18, 20, 18, 17

  ... many more similar records ...

37,   ,  7,   ,
38, 12, 12, 18, 17
39,   , 19, 15, 15
```

The input data for estimating the three models may look like the file fragment shown in Table 3. Amos handles several popular data formats with different conventions for coding missing values. The comma-delimited format of Table 3 begins with a line of *p* variable names, followed by *N* lines of *p* data entries each, separated by *p-1* commas. Missing data appear as blank entries. For instance, case 3 is missing the value of Q4, case 37 is missing Q1, Q3 and Q4, and case 39 is missing Q1.

When Amos encounters missing values among the modeled data, the program switches automatically from its default moment-based maximum likelihood algorithm to the case-wise formula (1). The Amos user must request estimation of means and intercepts because, with incomplete data, their contribution to the likelihood is no longer independent of variance and covariance terms. The most general model is specified as the input-path diagram of Figure 10. The three submodels are defined within the Amos 4 Model Manager.

Table 4: Model fit

Summary of models

Model	NPAR	CMIN	DF	P	CMIN/DF
Saturated	14	0.000	0		
Markov	11	3.822	3	0.281	1.274
Stationary Markov	7	31.364	7	0.000	4.481
Saturated model	14	0.000	0		
Null model	8	424.405	6	0.000	70.734

...

Model	AIC	BCC	BIC	CAIC
Saturated	28.000	32.242		
Markov	25.822	29.155		
Stationary Markov	45.364	47.485		
Saturated model	28.000	32.242		
Null model	440.405	442.829		

Table 4 presents a short list of the more than twenty[4] fit statistics provided by Amos 4 with incomplete data. To evaluate the fit of a working model in the missing-data case, Amos must fit both working and saturated models. Assuming both models converge to global solutions[5], the χ^2 fit-statistic is obtained as the difference in function of the log likelihood (1) values between the working and saturated models. The (positive) difference in number of parameters between models gives the associated degrees of freedom. The χ^2 fit-statistics appear in the CMIN column of Table 4. The first three lines display the fit of the specified Saturated, Markov and Stationary Markov submodels. With our artificial data, the fit of the Markov model would appear quite reasonable (χ^2 =3.8, df=3; p=0.28), while the large χ^2 of the Stationary Markov model (χ^2 =31.4, df=7; p=0.00) would indicate misspecification of that model. The bottom panel of Table 4 shows the Markov model with the smallest AIC and BCC statistics—smaller even than the Saturated model. According to Akaike (1987), the model with the smallest AIC has the best fit. Aikaike's rule would pick the Markov model over either Stationary Markov or Saturated models.

Table 5: Incremental fit statistics

Model Comparisons

Assuming model Saturated to be correct:

	DF	CMIN	P	NFI Delta-1	IFI Delta-2	RFI rho-1	TLI rho-2
Markov	3	3.822	0.281	0.009	0.009		
Stationary Markov	7	31.364	0.000	0.074	0.074		

Assuming model Markov to be correct:

	DF	CMIN	P	NFI Delta-1	IFI Delta-2	RFI rho-1	TLI rho-2
Stationary Markov	4	27.542	0.000	0.065	0.065	0.045	0.046

Because the three models are hierarchically nested, their relative discrepancies can be tested by the likelihood ratio chi-square statistic. Table 5 summarizes these incremental fit statistics. Comparing the Markov and Stationary Markov models is particularly interesting. The large χ^2 of 27.542 (df = 4) is a

strong rejection of the stationarity assumption: The residual variances and lag-1 regression weights do vary over time.

Table 6: Parameter estimates of the Markov model

```
Regression Weights:                    Estimate     S.E.      C.R.     Label
-------------------                    --------   -------   -------   -------
                        Q2 <---- Q1       0.857     0.137     6.261    b12
                        Q3 <---- Q2       0.452     0.152     2.972    b23
                        Q4 <---- Q3       0.167     0.072     2.311    b34

Means:                                 Estimate     S.E.      C.R.     Label
------                                 --------   -------   -------   -------
                                  Q1     15.059     0.636    23.692

Intercepts:                            Estimate     S.E.      C.R.     Label
-----------                            --------   -------   -------   -------
                                  Q2      2.475     2.126     1.164
                                  Q3      7.714     2.433     3.170
                                  Q4     13.584     1.100    12.349

Variances:                             Estimate     S.E.      C.R.     Label
----------                             --------   -------   -------   -------
                                  Q1     14.576     3.478     4.191
                                  e2      9.279     2.287     4.057    v2
                                  e3     14.128     3.632     3.890    v3
                                  e4      2.229     0.637     3.498    v4

Implied (for all variables) Covariances

          Q1       Q2       Q3       Q4
          -------- -------- -------- --------
Q1          14.576
Q2          12.490   19.982
Q3           5.639    9.022   18.202
Q4           0.941    1.505    3.037    2.736

        ...

Implied (for all variables) Means

          Q1       Q2       Q3       Q4
          -------- -------- -------- --------
            15.059   15.380   14.658   16.030
```

Parameter estimates and approximate standard errors of the general Markov model appear in Table 6. The standard errors are an implicit and convenient by-product of the FIML algorithm employed by Amos. Their primary use is to gauge the likely ranges of the parameter estimates under replication. A common rule of thumb (based on the asymptotic normality of the estimates) assumes a 95% confidence interval at ±2 standard errors from the estimate. According to this rule, the size of the lag-1 autoregression weights appears to decline over time. In addition, the residual variances at occasions 2, 3 and 4 come out heterogeneous. Both findings corroborate the earlier decision against the stationary Markov model.

Note the implied covariance matrix and mean vector at near the end of Table 6. These first and second moments derive from the estimated parameters of the Markov model and are thus a function of the observed data as well as the working model.

Summary

Model specification and estimation with missing data follows the same strategies as in the complete data scenario. Except for specifying a missing data code, the FIML implementation of the Amos and Mx programs does not complicate the model setup. In return, FIML delivers parameter estimates unbiased under MAR and standard errors estimates based on asymptotic normal theory. For model testing, fit χ^2

statistics are usually available whenever the saturated model has a FIML solution. In addition, when competing models are hierarchically nested, the likelihood ratio chi-square test provides a powerful tool for detecting sources of misfit.

Application: Multiple groups with missing data

The first practical maximum likelihood implementation of incomplete data modeling in the SEM framework used a multiple group approach (Allison, 1987). The example in this section is based on one presented in Allison's original paper but has been modified to show how latent variables are conceptually the same as missing data (*cf*., Dempster, Laird and Rubin, 1977).

Bielby, Hauser and Featherman (1977) studied the relationship between indicators of occupational status and educational attainment in a sample of 2020 African-American fathers. Using a single indicator for each construct, Bielby *et al*. estimated the correlation between occupational status and educational attainment as $r = 0.43$ for the entire sample. Realizing that measurement error and temporal instability of their single indicators would likely attenuate the correlation estimate, Bielby *et al.* re-interviewed a random subsample of 348 study participants approximately three weeks after the first interview. For this sub-group, they obtained a second set of occupational status and educational attainment indicators that can be used for estimating the size of the measurement error. The data were reported (Allison, 1987) as two subsamples:

a) Bielby *et al.* (1977) complete data ($N = 348$):

cov	FAOC_t1	FAOC_t2	FAED_t1	FAED_t2
FAOC_t1	180.90			
FAOC_t2	126.77	217.56		
FAED_t1	23.96	30.20	16.24	
FAED_t2	22.86	30.47	14.36	15.13
mean	16.62	17.39	6.65	6.75

b) Bielby *et al.* (1977) incomplete data ($N = 1672$); unobserved means and (co-)variances indicated by dashes:

cov	FAOC_t1	FAOC_t2	FAED_t1	FAED_t2
FAOC_t1	217.27			
FAOC_t2	—	—		
FAED_t1	25.57	—	16.16	
FAED_t2	—	—	—	—
mean	16.98	—	6.83	—

The two-group factor model of Figures 11 and 12 proposes a simple way to separate stable (or systematic) measurement components from measurement error and estimate the disattenuated correlation. Figure 11 shows the confirmatory factor model for the complete-data subsample. Father's occupational status has two observed indicators, FAOC_t1 and FAOC_t2, with independent error terms e1 and e2. Father's educational attainment has two observed indicators, FAED_t1 and FAED_t2, again with independent error terms e3 and e4. The hypothetically error-free occupational status and educational attainment variables are correlated. The model has 13 free parameters, all labeled: One factor covariance (cov_oe), two factor variances (v_faoc and v_faed), two regression weights (b2 and b4), four intercepts (i1, i2, i3, and i4), and four specific variance terms (v1, v2, v3, and v4).

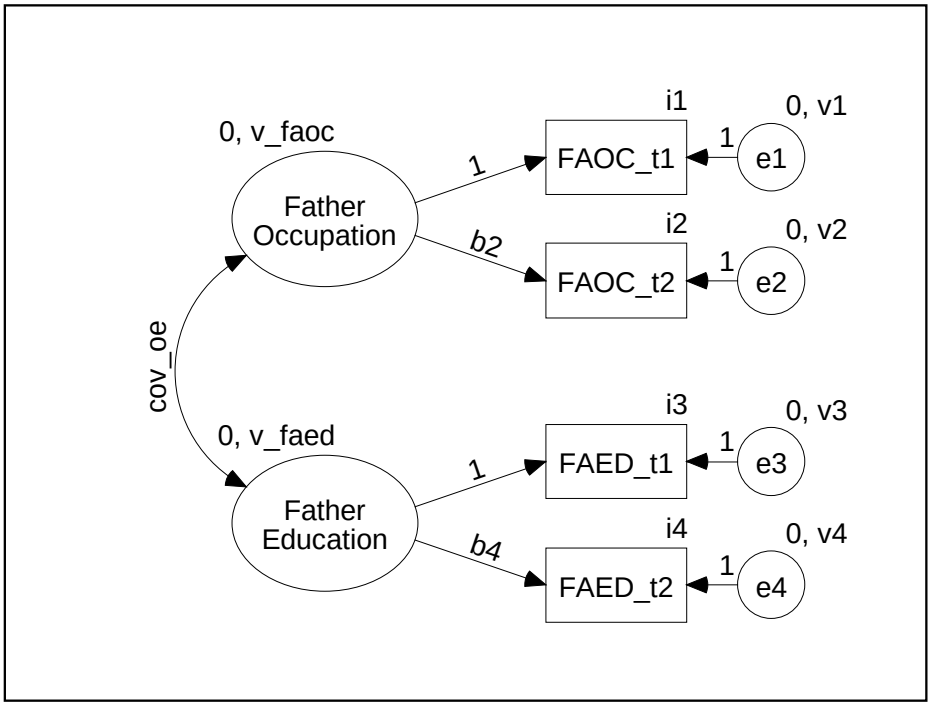

Figure 11: Measurement model for the complete subsample

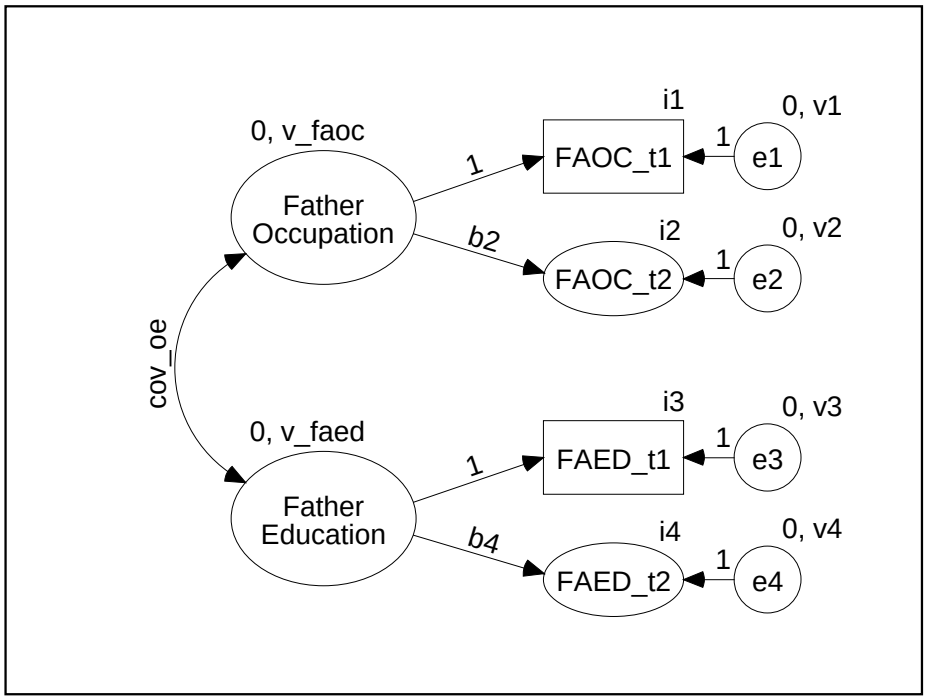

Figure 12: Measurement model for the incomplete subsample

Figure 12 uses the same measurement model for the incomplete subsample, including parameter labels. In Amos's notation, two parameters that share the same label are automatically equal-valued. In other words, using the same label for a parameter throughout all groups makes that parameter group-invariant. All

corresponding free parameters in Figures 11 and 12 are shown with group-invariant labels, and all fixed parameters have the same value in both groups. This means the entire model is group-invariant, and implies that the missing data process is MCAR. The only difference is that the model in Figure 12 accommodates the missing variables. Placing FAOC_t2 and FAED_t2 in ellipses in the second group declares these variables as latent or unobserved—in other words, missing.

The analysis estimates the disattenuated correlation between occupational status and educational attainment as $r = 0.623$ (s.e. = 0.029). The overall fit of the two-group model is quite acceptable ($\chi^2 = 7.8$, df = 6), supporting the implied assumption that the data are MCAR[6]. This estimate is quite different from the correlation of 0.43 between the observed variables FAOC_t1 and FAED_t1.

Summary

Allison's (1987) multi-group implementation can also be accommodated within Amos. The setup is similar to other contemporary SEM programs, except that Amos permits additional simplifications. The multi-group setup for missing data analysis involves these steps:

1. Draw a model for complete data and label all free parameters. Select mean-level analysis.
2. Turn the single group specification into a multi-group analysis, declaring one "group" for each pattern of missing data. Issue Amos's *Heterogeneous Groups* command. Connect each group with its data file. The *Heterogeneous Groups* command removes the necessity of having the same model and the same variables in all groups; it also allows different groups to use different numbers of variables.
3. In each group's path diagram, use the *Toggle Observed/Unobserved* tool to mark the missing variables as latent (or unobserved).

Allison's multi-group approach may be regarded as an alternative to Amos's case wise (default) FIML estimation with incomplete data. There are some obvious tradeoffs to using Allison's multi-group setups. As mentioned previously, model specification by the multi-group approach is both more laborious and only feasible with a small number of missing data patterns. On the other hand, all fit statistics, modification indices and residual analyses normally available with complete data are provided when the multi-group setup is employed. In addition, the group-specific means and variances of the exogenous variables are under the modeler's control, permitting detailed tests of MCAR, MAR and other assumptions.

Discussion

Maximum likelihood (FIML) estimation with incomplete data is a feasible method, now available in the Amos and Mx structural equation modeling programs. FIML is more efficient and less biased than listwise and pairwise deletion and mean-imputation methods. In the case of MAR data, FIML can be dramatically less biased than listwise deletion and mean-imputation methods. This is why FIML should be the preferred method of treating missing data when the alternative is pairwise or listwise deletion or mean-imputation.

Maximum likelihood's lack of reliance on the MCAR requirement is a feature that remains to be fully exploited. Unbiasedness under MAR and higher efficiency under MCAR make maximum likelihood the method of choice in situation with incomplete multinormal data.

Feasible alternatives to the FIML approach of Amos and Mx are the EMCOV and NORM approaches (Graham and Hofer, 2000) which use EM and data augmentation methods based on the saturated model for imputing values of the missing data. The completed data matrices would subsequently be analyzed by traditional SEM methods.

With the FIML approach of Amos and Mx, in contrast, it is not necessary either to impute values for missing data or to estimate the population moments as a prerequisite to model fitting by ML. These are

optional steps which—if performed at all—are best done after the model is fitted (see Appendix), not before. Most structural modeling programs report estimates of population means, variances and covariances, calculated from parameter estimates under the assumption of a correct model.

It should not be overlooked that structural modeling with the Amos program can also be used to solve missing-data problems that arise in conventional analyses, such as regression with observed variables or the simple estimation of means and variances.

Appendix

Imputation of missing values

Let $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ be the population means and covariances of all variables in the model, both measured and unmeasured, and let $\hat{\boldsymbol{\mu}}^*$ and $\hat{\boldsymbol{\Sigma}}^*$ be their estimates assuming a correct model. For an individual case *i*, let the partitioned data vector $\mathbf{y}_i$ contain all unobserved and observed data of the model, with unobserved variables ordered first:

$$\mathbf{y}_i = \left(u_{i,1} \quad u_{i,2} \quad \cdots \quad u_{i,p} \,\middle|\, m_{i,1} \quad m_{i,2} \quad \cdots \quad m_{i,q} \right) = \left[\mathbf{y}_{i,u} \,\middle|\, \mathbf{y}_{i,m} \right]. \tag{1}$$

The subvectors $\mathbf{y}_{i,u}$ and $\mathbf{y}_{i,m}$ will be of different sizes for different missing-data patterns. Arranging the values in $\hat{\boldsymbol{\mu}}^*$ and $\hat{\boldsymbol{\Sigma}}^*$ in the same order as in $\mathbf{y}_i$ yields the partitioned implied mean vector:

$$\hat{\boldsymbol{\mu}}_i = \left(\hat{\boldsymbol{\mu}}_{i,u} \,\middle|\, \hat{\boldsymbol{\mu}}_{i,m} \right) \tag{2}$$

and covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_i = \left| \begin{array}{c|c} \hat{\boldsymbol{\Sigma}}_{i,uu} & \hat{\boldsymbol{\Sigma}}_{i,um} \\ \hline \hat{\boldsymbol{\Sigma}}_{i,mu} & \hat{\boldsymbol{\Sigma}}_{i,mm} \end{array} \right|. \tag{3}$$

Under normality, the expectation of the missing data, conditional on the observed values, is estimated as:

$$\overline{\mathrm{E}\left(\mathbf{y}_{i,u} \,\middle|\, \mathbf{y}_{i,m} \right)} = \hat{\boldsymbol{\mu}}_{i,u} + \hat{\boldsymbol{\Sigma}}_{i,um} \hat{\boldsymbol{\Sigma}}_{i,mm}^{-1} \left(\mathbf{y}_{i,m} - \hat{\boldsymbol{\mu}}_{i,m} \right), \tag{4}$$

and their conditional covariance matrix as:

$$\overline{\mathrm{Cov}(\mathbf{y}_{i,u} | \mathbf{y}_{i,m})} = \hat{\boldsymbol{\Sigma}}_{i,uu} - \hat{\boldsymbol{\Sigma}}_{i,um} \hat{\boldsymbol{\Sigma}}_{i,mm}^{-1} \hat{\boldsymbol{\Sigma}}_{i,mu}. \tag{5}$$

These statistics can be used to impute the model-based means and confidence intervals of the missing data, given the observed portion of the data. With complete data, using (4) produces the usual regression estimates of factor scores provided by many structural modeling programs. Stochastic regression imputation (Little and Rubin, 1989) and multiple imputation (Little and Rubin, 1987) are important variants of this imputation method.

References

Ahrens, J.H. and Dieter, U. (1973) Extensions of Forsythe's method for random sampling from the normal distribution. *Mathematics of Computation*, **27**(124), 927–937

Akaike, H. (1987) Factor analysis and AIC. *Psychometrika*, **52**(3), 317–332

Allison, P.D. (1987) Estimation of linear models with incomplete data. In C.C. Clogg [Ed.] *Sociological Methodology, 1987*. San Francisco: Jossey-Bass, 71–103

Arbuckle, J.L. (1995) *Amos for Windows. Analysis of moment structures. Version 3.5.* Chicago: SmallWaters Corp.

Arbuckle, J.L. (1996) Full information estimation in the presence of incomplete data. In G.A. Marcoulides and R.E. Schumacker [Eds.] *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Publishers

Bielby, W.T., Hauser, R.M. and Featherman, D.L. (1977) Response errors of black and nonblack males in models of the intergenerational transmission of socioeconomic status. *American Journal of Sociology*, **82**, 1242–1288

Bock, R.D. (1975) *Multivariate statistical methods in behavioral research*. New York: Wiley

Brown, B.W and Lovato, J. (1994) RANLIB.C: Library of C routines for random number generation. Houston, TX: The University of Texas, M.D. Anderson Cancer Center (available on the NETLIB internet archive)

Brown, C.H. (1983) Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika*, **48**(2), 269–291

Brown, R.L. (1994) Efficacy of the indirect approach for estimating structural equation model with missing data: A comparison of five methods. *Structural Equation Modeling: A Multidisciplinary Journal*, **1**, 287–316

Dempster, A.P., Laird, N.M. and 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

Graham, J.W., Hofer, S.M., Donaldson, S.I., MacKinnon, D.P., and Schafer, J.L. (1997) Analysis with missing data in prevention research. In K. Bryant, M. Windle & S. West [Eds.] *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 325-366). Washington, DC: American Psychological Association

Graham, J.W. and Hofer, S.M. (2000) Multiple imputation in multivariate research. In T.D. Little, K.U. Schnabel and J. Baumert [Eds.] *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.* Mahwah, NJ: Lawrence Erlbaum Associates

Hartley, H.O. and Hocking, R.R. (1971) The analysis of incomplete data. *Biometrics*, **27**, 783–823

Hilton, T.L. and Beaton, A.E. (1971) *Stability and instability in academic growth—a compilation of longitudinal data.* U.S. Dept. of Health, Education and Welfare, Office of Education, Bureau of Research.

Johnson, N.L. and Kotz, S. (1972) *Distributions in Statistics: Continuous multivariate distributions*. New York: Wiley

Jöreskog, K.G. (1977) Statistical models and methods for analysis of longitudinal data. In D.V. Aigner & A.S. Goldberger [Eds.} *Latent variable models in the social sciences.* Amsterdam: North Holland Publishing Co.

Kendall, M. and Stuart, A. (1977) *The advanced theory of statistics.* Volume 3. London, UK: Charles Griffin & Co.

Kim, J.-O. and Curry, J. (1977) The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, **6**, 215–240

L'Ecuyer, P. and Côté, S. (1991) Implementing a random number package with splitting facilities. *ACM Transactions on Mathematical Software*, **17**(1), 98–111

Little, R.J.A. and Rubin, D.B. (1987) *Statistical analysis with missing data*. New York: Wiley

Little, R.J.A. and Rubin, D.B. (1989) The analysis of social science data with missing values. *Sociological Methods and Research*, **18**, 292–326

Little, R.J.A. and Schenker, N. (1995) Missing data. In G. Arminger, C.C. Clogg and M.E. Sobel [Eds.] *Handbook of statistical modeling for the social and behavioral sciences.* New York: Plenum, 39–75

Muthén, B., Kaplan, D. and 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 Edition.* Box 710 MCV, Richmond, VA 23298: Department of Psychiatry

Roth, P.L. (1994) Missing data: A conceptual view for applied psychologists. *Personnel Psychology*, **47**, 537–560

Rovine, M.J. (1994) Latent variable models and missing data analysis. In A. von Eye and C.C. Clogg [Eds.] *Latent variable analysis: Applications for developmental research.* Thousand Oaks, CA: Sage Publications

Rubin, D.B. (1976) Inference and missing data. *Biometrika*, **61**, 581–592

Verleye, G. (1996) *Missing at random data problems in attitude measurement using maximum likelihood structural equation modeling.* Unpublished dissertation. Frije Universiteit Brussels, Department of Psychology

Wilks, S.S. (1932) Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, **3**, 163–195

Wothke, W. (1993) Nonpositive definite matrices in structural modeling. In K.A. Bollen and J.S. Long [Eds.] *Testing Structural Equation Models.* Newbury Park, CA: Sage Publications

Footnotes

[1] When complete data vectors are sampled from a multivariate normal population, the joint distribution of the elements in the resulting sample covariance matrix follows a Wishart distribution (Johnson and Kotz, 1972).

[2] In 25%-37% of these simulation runs, Amos indicated some convergence problems. However, this did not seem to make a difference in results. Solutions from "converged" and "non-converged" runs were statistically indistinguishable, thus all results from all simulations were included in the statistical reports.

[3] Convergence problems were also encountered with the MAR simulation, but again they did not appear to affect the overall results.

[4] Some fit indices, such as the GFI and RMR, are defined only for complete data. Other fit indices, including the BIC and CAIC, were devised only for single group analyses without mean structures. These statistics do not apply to incomplete data, at least not in their current formulations.

[5] Unless the number of observed variables becomes large relative to number of cases and observed data rates, both working and saturated models usually converge to global maxima. This χ^2 statistic can usually be computed whenever the saturated model has a global solution.

[6] The MAR assumption can be incorporated into the multi-group missing data approach by letting means and covariances of the exogenous observed variables vary free across groups. Such a MAR analysis is possible when missing data occur only among endogenous variables, but not exogenous ones.