Analyzing Structural Equation Models With Missing Data

Craig Enders*
Arizona State University
cenders@asu.edu

based on

Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 313-342). Greenwich, CT: Information Age Publishing.

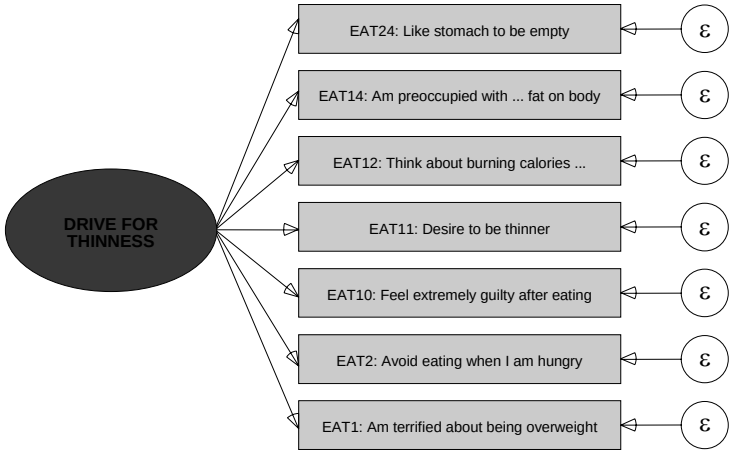
Overview

- Missing data theory, and assumptions pertaining to missing data analyses
- Full information maximum likelihood (FIML), and multiple imputation (MI)
- Incorporating information from auxiliary variables

Running Example

- Concepts illustrated using single factor CFA model with seven manifest indicators
- The indicators are a subset of items from the Eating Attitudes Test (EAT), a widely used inventory for assessing eating disorder risk
- The data also include body mass index (BMI) scores (BMI = weight(kg)/[height(m)]^2)
- The data are available upon request from Dr. Enders

Running Example



Missing Data Mechanisms
Overview

- Rubin (1976) provided a taxonomy for missing data problems
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)
- Missing data mechanisms describe how the missing values are related to the data, if at all
- Rubin's mechanisms can be viewed as assumptions that dictate the performance of a particular missing data technique

Missing Data Mechanisms
Missing Completely At Random (MCAR)

- Missing values on *Y* are unrelated to other variables, as well as to the unobserved values of *Y*
- The observed data are a random sample of the hypothetically complete data
- For example, the EAT was not administered due to a clerical error or scheduling difficulties; or individuals who did respond randomly failed to make dark enough marks in places on response sheet
- This is an unusually strict assumption in most cases
- FIML and MI are unbiased, as are most traditional deletion methods (listwise; pairwise)

Missing Data Mechanisms
Missing At Random (MAR)

- Missing values on *Y* can depend on other variables, but not on the unobserved values of *Y*
- For example, missing EAT item responses are related to BMI, such that individuals with very low BMI scores are more likely to skip certain items
- FIML and MI assume MAR, and are unbiased
- Traditional deletion (pairwise; listwise) methods are biased
- *Warning*: MAR is defined relative to the variables in the analysis. If the "cause" of the missing data is a measured variable (e.g., BMI) that does not appear in the analysis model, MAR does not hold

Missing Data Mechanisms
Missing Not At Random (MNAR)

- Missing values on *Y* depend on the unobserved values of *Y*
- For example, missing EAT scores are related to the respondent's level of eating disorder risk, such that individuals who are highly preoccupied with food tend to skip those items
- FIML and MI will yield biased estimates under MNAR, although less biased than traditional methods

The Multivariate Normal Probability Density Function (PDF)

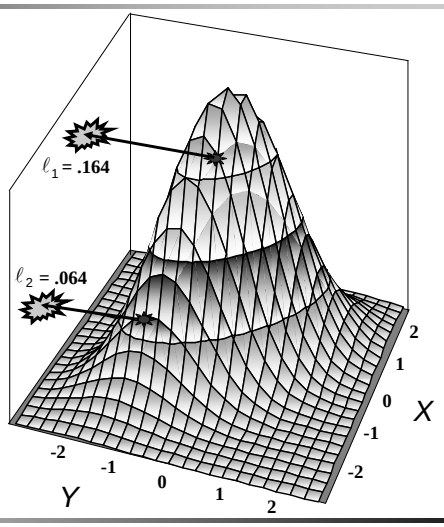
- The shape of the multivariate normal curve is defined by the familiar equation

$$\ell_i = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})/2}$$

- Plugging in a person's raw data values gives us the *likelihood* (i.e., the relative probability) of that set of scores, given the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Bivariate Normal Distribution

- Suppose that $\mu = 0$, $\sigma^2 = 1$, and $r = .60$
- Two cases:
  - $X_1 = -.5$, $Y_1 = 0$
  - $X_2 = -1.5$, $Y_2 = -1$
- Case 1 is closer to the parameter values, and thus has a higher likelihood)



The Log Likelihood

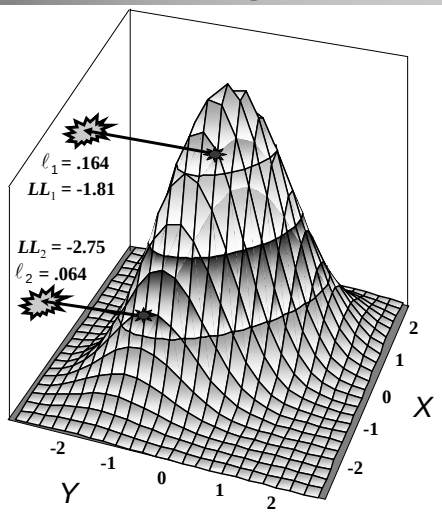
- Likelihoods tend to be very small numbers
- Taking the natural log of the likelihood makes the math a bit more tractable

$$\log L_i = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

- The log likelihood still quantifies the same thing – the relative probability of a person's scores, given a normal distribution with a particular $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Casewise Log Likelihoods

- Two cases:
  - $X_1 = -.5$, $Y_1 = 0$
  - $X_2 = -1.5$, $Y_2 = -1$
- Case 1 has a higher log likelihood, given these parameters



Log Likelihood Functions

- With complete data, each case's contribution to the log likelihood is

$$\log L_i = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

- In the missing data context, each *i*th case's contribution to the log likelihood is

$$\log L_i = -\frac{p_i}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)$$

The Missing Data Log Likelihood

- The sample log likelihood is obtained by summing over the *N* cases

$$\log L = K - \frac{1}{2}\sum_{i=1}^{N}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)$$

- In the current CFA example, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are functions of the model parameters:
  - $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$
  - $\boldsymbol{\mu} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}$

How Does ML Differ With Missing Data?

- The only difference between the complete data and missing data log likelihood is the *i* subscript
- This allows the size and content of the data and parameter arrays to vary across cases
- The log likelihood is based only on those variables (and parameters) for which case *i* has complete data

Example (1)

- Consider a case with three variables, Y_1, Y_2, and Y_3
- The contribution to the log likelihood for cases who have complete data is

$$\log L_i = K_i - \frac{1}{2}\log\left|\begin{matrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{matrix}\right| - \frac{1}{2}\left(\begin{bmatrix} y_1\\ y_2\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right)' \begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{bmatrix}^{-1} \left(\begin{bmatrix} y_1\\ y_2\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}\right)$$

Σ $\mathbf{y}$ μ Σ $\mathbf{y}$ μ

Example (2)

- The contribution to the log likelihood for cases who are missing Y_2 is

$$\log L_i = K_i - \frac{1}{2}\log\left|\begin{matrix}\sigma_{11} & \sigma_{13}\\ \sigma_{31} & \sigma_{33}\end{matrix}\right| - \frac{1}{2}\left(\begin{bmatrix}y_1\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_3\end{bmatrix}\right)' \begin{bmatrix}\sigma_{11} & \sigma_{13}\\ \sigma_{31} & \sigma_{33}\end{bmatrix}^{-1} \left(\begin{bmatrix}y_1\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_3\end{bmatrix}\right)$$

- The rows and columns for Y_2 were removed

$$\log L_i = K_i - \frac{1}{2}\log\left|\begin{matrix}\sigma_{11} & \sigma_{13}\\ \sigma_{31} & \sigma_{33}\end{matrix}\right| - \frac{1}{2}\left(\begin{bmatrix}y_1\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_3\end{bmatrix}\right)' \begin{bmatrix}\sigma_{11} & \sigma_{13}\\ \sigma_{31} & \sigma_{33}\end{bmatrix}^{-1} \left(\begin{bmatrix}y_1\\ y_3\end{bmatrix} - \begin{bmatrix}\mu_1\\ \mu_3\end{bmatrix}\right)$$

Notes on FIML

- FIML is not much different from complete-data ML estimation, but incorporates a mean structure
- Missing values are not imputed!
- Parameters and standard errors are estimated directly using all the observed data
- Including the partially complete cases serves to "steer" the estimation algorithm toward a more accurate set of parameters, via the relations among the variables

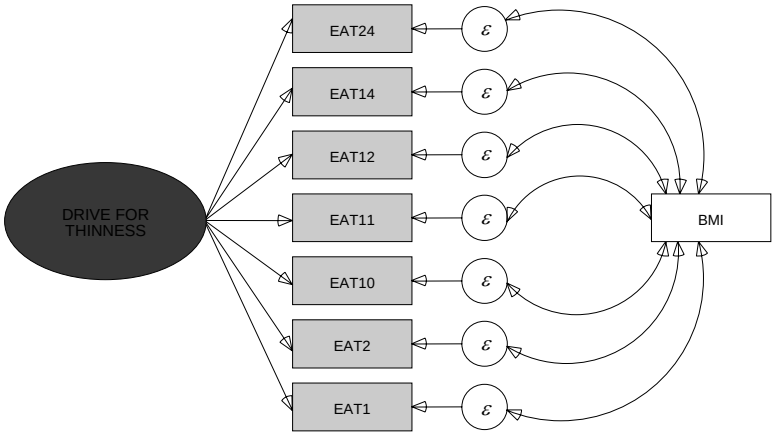
Incorporating Auxiliary Variables

- In order for MAR to hold, the "cause" of the missing data must appear in the analysis model
- An auxiliary variable (AV) such as BMI is ancillary to one's substantive hypotheses
- Adopting an "inclusive" analysis strategy that incorporates AVs can make MAR more plausible
- A useful AV is either a potential cause or correlate of missingness, or a correlate of the variable that is missing

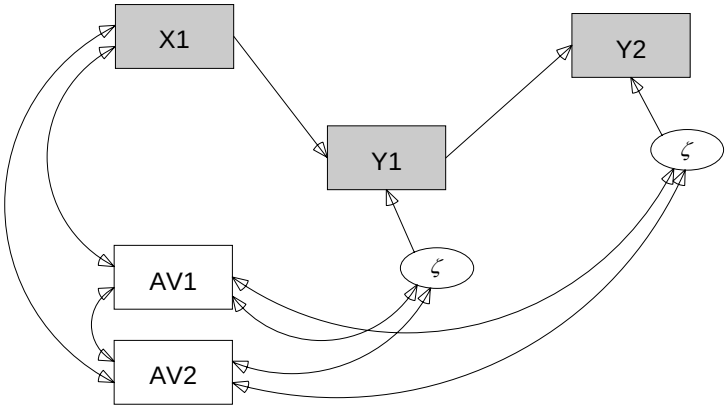
The Saturated Correlates Model

- Graham (2003) outlined a "saturated correlates" model for including AVs in an SEM analysis
- Three rules apply:
  - AVs should be correlated with observed (not latent) predictors in the model
  - AVs should be correlated with the residual terms from observed (not latent) dependent variables
  - AVs should be correlated with one another

The Saturated Correlates Model
EAT Analysis CFA Model

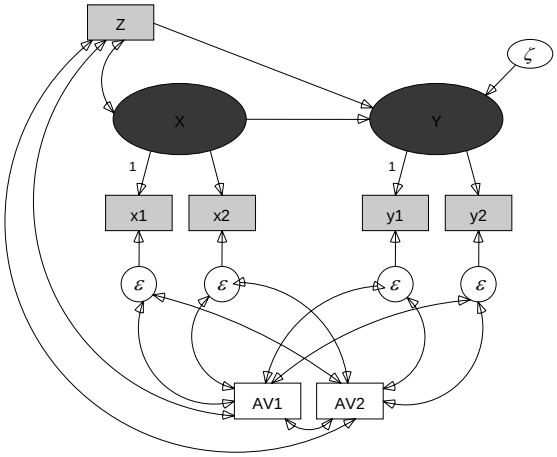


The Saturated Correlates Model
Path Model Example



The Saturated Correlates Model

General Structural Model Example



FIML Miscellanea

- When given the choice, compute standard errors using the observed information matrix
  - The observed information assumes MAR, while the expected information requires MCAR
- Rescaled test statistics for nonnormal data are available in M*plus* and EQS (see Yuan and Bentler, 2000)
- Sandwich estimator (i.e., "robust") standard errors are also available, but assume MCAR data

Overview Of Multiple Imputation

- Multiple copies (e.g., $m = 10$) of the incomplete data set are created
- Each data set is imputed with different estimates of the missing values using multiple regression
- The SEM analysis is performed on each of the "filled-in" data sets
- A single point estimate and standard error is obtained by averaging over the *m* data sets

Imputation And Analysis Phase

- Unlike FIML, missing data are handled in an imputation phase that is distinct from the analysis
- The imputation phase is used to create the *m* imputed data sets
- Once the data sets are constructed, any number of different analyses can be performed using the same set of imputations

Imputation Phase: Imputation (I) Step

- Begin with initial estimate of μ and Σ
- Use μ and Σ to construct regression equations for each missing data pattern
- Impute missing values with predicted scores, and add a normally distributed residual to each to restore lost variability

Imputation Phase: Posterior (P) Step

- Compute new estimates of μ and Σ using the imputed data
- Conditional on the updated estimates of μ and Σ, randomly sample new values for μ and Σ from a simulated sampling distribution
- Carry the new values of μ and Σ forward to the next I step, and impute missing values using a new set of regression equations

EAT Example

Case	EAT1	$m = 1$	$m = 2$	$m = 3$	$m = 9$	$m = 10$
1	1	1	1	1	1	1
2	3	3	3	3	3	3
3	?	2	3	3	5	6
4	?	1	1	3	3	2
5	1	1	1	1	1	1

Analysis Phase: Combining Estimates

- After the imputations are created, the model of interest is fit to each of the *m* data sets
- A point estimate for any parameter of interest is obtained by averaging the estimates from the *m* analyses

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i$$

Analysis Phase: Combining S.E.s

- Between-imputation variance is the variance of the parameter across the *m* imputations

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{\theta}_i - \bar{\theta}\right)^2$$

- Within-imputation variance is the mean of the *m* squared S.E.s

$$\bar{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i$$

- The total variance (i.e., squared S.E.)

$$T = \bar{V} + \left(1 + \frac{1}{m}\right)B$$

Single Parameter Significance Tests

- Single parameter tests are obtained via a *t* statistic

$$t = \frac{\bar{\theta}}{\sqrt{T}}$$

- The degrees of freedom are complex, and depend on between- and within-imputation variance and *m*
- Use adjusted degrees of freedom (Barnard & Rubin, 1999) whenever possible

Incorporating Auxiliary Variables

- AVs are incorporated as additional predictor variables in the imputation phase
- The AVs need not be included in the subsequent analyses, because the imputed values have already been conditioned on the AVs
- As a general rule, the imputation phase should include all variables and design effects (e.g., interaction terms) that appear in the subsequent analysis model

MI Miscellanea

- M*plus* and LISREL have functions that automate the process of estimating and combining parameter estimates and standard errors
- χ^2 statistics can be combined, but this is more complex than a simple average (e-mail Dr. Enders if you want a SAS program to do this)
- It is unclear how to combine fit indices at this point
- Assessing MI convergence requires special graphical techniques (see the book chapter)

Selected Analysis Results

Loading	FIML	MI
EAT1	1.185 (.109)	1.164 (.107)
EAT2	.542 (.059)	.541 (.059)
EAT10	.745 (.065)	.737 (.066)
EAT11	1.096 (.077)	1.093 (.077)
EAT12	1.122 (.103)	1.099 (.102)
EAT14	.981 (.073)	.980 (.073)
EAT24	.721 (.077)	.707 (.074)

A Final Comparison of FIML and MI

- Both procedures assume MAR data and multivariate normality
- FIML and MI should yield very similar estimates if the same set of input variables are used
- MI is more complex to use, but the resulting analyses don't require specialized software
- If an estimation routine exists, it is probably more straightforward to use FIML and incorporate AVs into the analysis