

A Tutorial on Structural Equation Modeling With Incomplete Observations: Multiple Imputation and FIML Methods Using SAS†

Wei Zhang and Yiu-Fai Yung
SAS Institute Inc.

R & D, Advanced Analytics Division
SAS Campus Drive, Cary, NC 27513
USA

† Presented at the International Meeting of Psychometric Society, Tai Po, Hong Kong, on July 21, 2011. Some materials of this presentation are based on the talk by Yung and Zhang (2011) at the SAS Global Forum, Las Vegas, NV, on April 2011. The authors would like to thank Werner Wothke for suggesting the multiple-group analysis of the planned missingness example in this presentation. Correspondence should be sent to Wei.Zhang@sas.com or Yiu-Fai.Yung@sas.com. The program used in the presentation is available upon request.

Abstract

This presentation demonstrates the use of the MI, MIANALYSIS, and CALIS procedures of SAS/STAT (version 9.22 or later) to fit structural equation models with incomplete observations (or missing data). The multiple imputation method and the full information maximum likelihood (FIML) method are two statistically proven methods for analyzing structural equation models with incomplete observations. This presentation illustrates the steps required to carry out these two methods in the SAS system. Practical data examples are used throughout the presentation. Although the multiple imputation method and the FIML method yield similar estimation results, with its availability in the CALIS procedure the FIML method is more convenient to use. To help understand the source of missingness, the CALIS procedure provides detailed analysis of the missing patterns, including proportion coverage of sample moments and the mean profiles of dominant missing patterns. When the number of missing patterns is small, a multiple-group setup with the CALIS procedure can be used. This is also illustrated with a real data example.

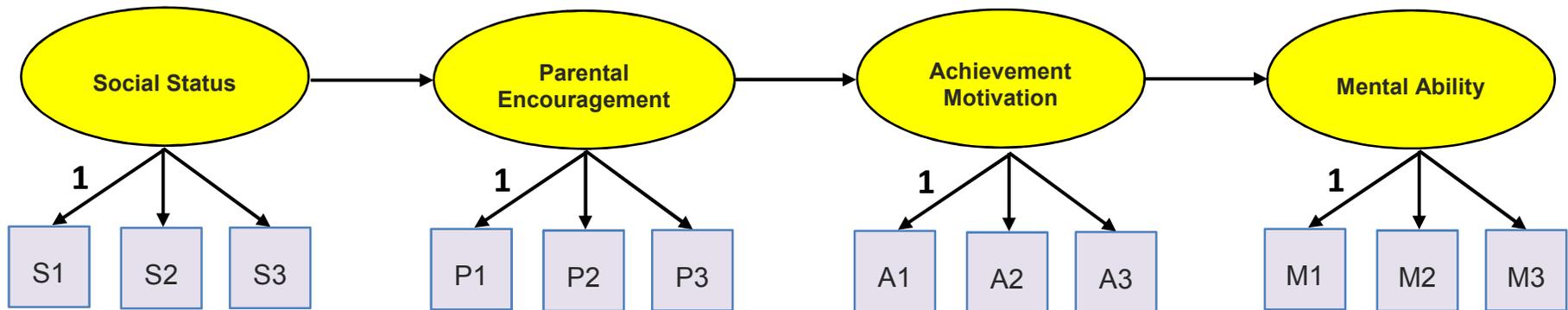
Introduction

Incomplete observations are those observations that have at least one (but not all) missing values in the statistical analysis. Regular maximum likelihood (ML) estimation in structural equation modeling (SEM) software excludes the incomplete observations from analysis. If the sample size is very large and the number of incomplete observations is relatively small, the estimation of the structural equation model might not suffer too much from ignoring the information of the incomplete observations. Unfortunately, in a typical social and behavioral research, the sample size is usually not that large and it is desirable to use an estimation method that could utilize the information from incomplete observations. The multiple imputation (MI) method and the full information maximum likelihood (FIML) method are two statistically proven methods that provide sound statistical treatments of the incomplete observations in the estimation and model fitting. This paper illustrates these two methods by using some recently developed features in the CALIS procedure of the SAS/STAT software (version 9.22 or later). A numerical example based on simulated data is used to illustrate the steps required to carry out the multiple imputation method by the MI and MIZNALYZE procedures and the FIML method by the CALIS procedure. The same numerical example can also be found in a recent paper by Yung and Zhang (2011), who provide more details about various methodologies for treating incomplete observations, including ad hoc methods such as pairwise deletion, listwise deletion, and mean imputation, and the more principled methods such as MI and FIML. See Yung and Zhang (2011) and the references therein for the discussion of the advantages of the MI and FIML methods over the ad hoc methods.

The Model and the Simulated Data

This example is inspired by an example in Marjoribanks (1974). However, the structural equation model used here is purely hypothetical and the data are simulated. Except for the similarity in the names of the hypothetical constructs, no part of the current analysis represents the original research.

The following path diagram represents the structural equation model for predicting mental ability:



The measured variables are represented by squares and the latent variables are represented by ovals. A data set (“miss3”) of $N=200$ observations was generated from a multivariate normal distribution for the 12 observed variables. Half of these observations have random missing values. In other words, there were 100 complete observations and 100 incomplete observations.

The Full Information Maximum Likelihood Estimation by PROC CALIS

To estimate the model parameters by the FIML method, the following PROC CALIS code is used:

```
PROC CALIS DATA=miss3 METHOD=FIML;  
  PATH  
    S1-S3 <--- SocialStatus          = 1,  
    P1-P3 <--- ParentalEncouragement = 1,  
    A1-A3 <--- AchievementMotivation = 1,  
    M1-M3 <--- MentalAbility         = 1,  
    SocialStatus          ---> ParentalEncouragement,  
    ParentalEncouragement ---> AchievementMotivation,  
    AchievementMotivation ---> MentalAbility;  
RUN;
```

The METHOD=FIML option in the PROC CALIS statement invokes the full information maximum likelihood method to estimate the model parameters. The DATA= option specifies the name of data set. Notice that raw data must be provided for the FIML estimation. The PATH statement transcribes the path diagram information into path entries. Multiple-path specification syntax is used for the first four entries of the PATH statement. For example, the first entry specifies three paths from SocialStatus to S1, S2, and S3, respectively. The fixed constant 1, however, is applied only to the first path coefficient for S1 (or effect on S1). Path coefficients for S2 and S3 from SocialStatus are unrestricted free parameters. No specifications for variance or covariance parameters are necessary for the current model because they are set by default in PROC CALIS.

Setting Up the Multiple Imputation Method

To fit the same model by the multiple imputation method, the following three stages of analyses are used:

```
/*----- Stage 1: Use PROC MI to generate 20 imputation samples -----*/
PROC MI DATA=miss3 NIMPUTE=20 SEED=12345 OUT=ImputedSamples;
  VAR S1-S3 P1-P3 A1-A3 M1-M3;
RUN;

/*---- Stage 2: Use PROC CALIS to analyze the 20 imputed samples ----*/
PROC CALIS DATA=ImputedSamples OUTEST=Est NOPRINT;
  PATH
    S1-S3 <--- SocialStatus          = 1,
    P1-P3 <--- ParentalEncouragement = 1,
    A1-A3 <--- AchievementMotivation = 1,
    M1-M3 <--- MentalAbility         = 1,
    SocialStatus          ---> ParentalEncouragement,
    ParentalEncouragement ---> AchievementMotivation,
    AchievementMotivation ---> MentalAbility;
  BY _Imputation_;
RUN;

/*---- Stage 3: Use PROC MIANALYZE to combine the estimation results ----*/
PROC MIANALYZE DATA=Est;
  MODELEFFECTS  _Parm01 _Parm02 _Parm03 _Parm04 _Parm05
                _Parm06 _Parm07 _Parm08 _Parm09 _Parm10 _Parm11;
RUN;
```

In the first stage, the MI procedure generates 20 imputed samples. The DATA= option specifies the original data set that contains missing values. The NIMPUTE= option requests 20 imputed samples in which the missing values are replaced with the generated random values. The SEED= option specifies the seed for random number generation in the imputation process. This option is not required but is used here to enable future replications of the results. Otherwise, PROC MI uses different seeds for generating imputed samples each time, making replications impossible. The OUT= option stores the 20 imputed data sets in a single SAS data set called `ImputedSamples`. PROC MI uses a variable named `_Imputation_` to index the 20 imputed data sets within the OUT= data set.

In the second stage, the CALIS procedure fits the structural equation model specified by the PATH statement to the 20 imputed samples. The BY statement specifies that the variable `_Imputation_` is the BY-group variable for indexing the 20 imputed samples (or 20 BY-groups) in the input data set **`ImputedSamples`**. PROC CALIS carries out separate model estimation for these 20 imputed samples as BY-groups. Because each imputed data set now contains only complete observations, the default maximum likelihood method (METHOD=ML) is used. The NOPRINT option in the PROC CALIS statement suppresses the displayed output. Instead, the OUTEST= option stores all the estimates and their standard errors from the 20 sets of ML estimations in the SAS dataset named `Est`, which is then analyzed by the MIANALYZE procedure in the next stage.

In the third stage, the MIANALYZE procedure combines the results from the 20 ML estimation. The DATA= option specifies that the SAS data set `Est` contains the estimates from the 20 imputed samples. The MODELEFFECTS statement specifies the parameters of interest. `_Parm01–_Parm11` is parameter names generated from PROC CALIS for the path coefficients (or effects). The MIANALYZE procedure produces an output for the parameters estimates and their standard error estimates.

Comparing the Estimates from FIML and MI

The following table shows the estimates of path coefficients or effects and their estimated standard errors from the FIML estimation and from the multiple imputation method.

-----Path-----	Parameter	FIML with CALIS	Multiple Imputation
S1	<--- SocialStatus	1.00000	1.00000
S2	<--- SocialStatus _Parm01	0.97633 (0.05143)	0.973775 (0.051477)
S3	<--- SocialStatus _Parm02	0.93340 (0.05879)	0.936031 (0.058573)
P1	<--- ParentalEncouragement	1.00000	1.00000
P2	<--- ParentalEncouragement _Parm03	1.02929 (0.08125)	1.037531 (0.079291)
P3	<--- ParentalEncouragement _Parm04	1.01757 (0.08056)	1.041038 (0.077584)
A1	<--- AchievementMotivation	1.00000	1.00000
A2	<--- AchievementMotivation _Parm05	1.06612 (0.07252)	1.065004 (0.071658)
A3	<--- AchievementMotivation _Parm06	1.04898 (0.06795)	1.045694 (0.070222)
M1	<--- MentalAbility	1.00000	1.00000
M2	<--- MentalAbility _Parm07	1.02426 (0.06116)	1.025313 (0.060573)
M3	<--- MentalAbility _Parm08	1.04717 (0.06530)	1.046538 (0.067612)
SocialStatus	---> ParentalEncouragement _Parm09	0.70886 (0.06736)	0.703038 (0.064195)
ParentalEncouragement	---> AchievementMotivation _Parm10	0.77686 (0.07893)	0.780869 (0.080593)
AchievementMotivation	---> MentalAbility _Parm11	0.86009 (0.07966)	0.860276 (0.079848)

Overall, the two sets of estimates from the two methods are very close to each other. Even the standard error estimates match well. This is actually not surprising because both methods are derived from the likelihood principle under multivariate normality and certain random missing assumptions. In fact, the multiple imputation method is supposed to approximate the full information maximum likelihood estimation with the presence of incomplete observations. Therefore, it is quite convenient to do full information maximum likelihood estimation directly with a single run of PROC CALIS, rather than doing multiple imputations and then combining estimation results.

Other advantages of using the FIML method instead of the MI method are: (1) The estimation results from the MI method depends on the simulated imputed values. These values could be different if a different seed for generating random numbers is used or if a different random number generation algorithm is used. The FIML method, however, would give you the same set of estimates every time the optimized solution is found. (2) The MI method does not have an easy way to combine the model fit test statistics and other fit indices from the imputed samples. However, the FIML method yields the model fit statistics and other fit indices just as straightforward as any other estimation methods for complete data.

The following table shows some fit statistics obtained from the FIML estimation for the simulated data. The model fit chi-square tests statistic is not significant, supporting a good model for the data. The RMSEA and standardized root-mean-squared residuals (SRMSR) indicate very good model fit. The CFI shows superior model fit.

Fit Summary of the FIML Estimation	
Chi-Square	58.6765
Chi-Square DF	51
Pr > Chi-Square	0.2147
Standardized RMSR (SRMSR)	0.0403
RMSEA Estimate	0.0274
Bentler Comparative Fit Index	0.9958

What If You Do a Regular Maximum Likelihood (ML) Estimation That Ignores the Incomplete Observations?

If the METHOD=FIML option in the PROC CALIS statement is not specified, a regular ML estimation is used by default. All the incomplete observations are ignored in the default ML estimation. This following shows the code for a regular ML estimation:

```
PROC CALIS DATA=miss3;
  PATH
  S1-S3 <--- SocialStatus          = 1,
  P1-P3 <--- ParentalEncouragement = 1,
  A1-A3 <--- AchievementMotivation = 1,
  M1-M3 <--- MentalAbility          = 1,
  SocialStatus          ---> ParentalEncouragement,
  ParentalEncouragement ---> AchievementMotivation,
  AchievementMotivation ---> MentalAbility;
RUN;
```

The following table shows the fit summary of the regular ML estimation for the simulated data. The model fit chi-square test statistic is now significant. The theoretical model is rejected at the .05 α -level. The SRMSR and RMSEA indicate only marginally acceptable fit, even though the CFI is still showing good fit. Clearly, the regular ML estimation loses some precious supporting information from the incomplete observations that would have led to better indications of model fit, as shown by the previous FIML fit summary.

Fit Summary of the ML Estimation

Chi-Square	70.0624
Chi-Square DF	51
Pr > Chi-Square	0.0394
Standardized RMSR (SRMSR)	0.0647
RMSEA Estimate	0.0614
Bentler Comparative Fit Index	0.9831

Ranking the Coverages. To locate the problematic coverages (or serious missingness), PROC CALIS ranks the coverages from smallest to largest in the following output. The first table shows that variable P1 (or the mean of P1) had only 53.5% of data coverage, while all other variables have at least 89% of data coverage. The second table shows that the covariance coverages that involve P1 and many other variables fall to about 50%. Therefore, you might want to take a closer look at variable P1 to understand why it has such a high proportion of missing values.

Rank Order of the 6 Smallest Variable (Mean) Coverages		
Variable	Coverage	
P1	0.5350	
M1	0.8900	
A2	0.9000	
S3	0.9100	
A3	0.9100	
M2	0.9150	

Rank Order of the 10 Smallest Covariance Coverages		
Var1	Var2	Coverage
P1	S3	0.5000
P1	M1	0.5050
P1	S2	0.5100
P1	M2	0.5100
P1	M3	0.5100
P1	A2	0.5100
P3	P1	0.5100
P1	S1	0.5150
P1	A1	0.5150
P1	A3	0.5150

Analysis of Missing Patterns. The following output shows the dominant missing patterns in the data set and their means.

Rank Order of the 5 Most Frequent Missing Patterns					
Total Number of Distinct Patterns with Missing Values = 26					
	Pattern	NVar Miss	Freq	Proportion	Cumulative
1	xxxxxxxxxx.xx	1	75	0.3750	0.3750
2	x...xx..x..x	7	1	0.0050	0.3800
3	.x.x....xxx.	7	1	0.0050	0.3850
4	...x.xx.....	9	1	0.0050	0.3900
5	..xx.x.....x	8	1	0.0050	0.3950

NOTE: Nonmissing Pattern Proportion = 0.5000 (N=100)

Means of the Nonmissing and the Most Frequent Missing Patterns						
Variable	Nonmissing (N=100)	-----Missing Pattern-----				
		1 (N=75)	2 (N=1)	3 (N=1)	4 (N=1)	5 (N=1)
S1	4.04000	3.58667	4.00000	.	.	.
S2	3.91000	3.65333	.	3.00000	.	.
S3	4.00000	3.52000	.	.	.	7.00000
M1	4.02000	3.56000	.	1.00000	2.00000	6.00000
M2	4.04000	3.56000	6.00000	.	.	.
M3	4.01000	3.42667	3.00000	.	4.00000	6.00000
A1	4.18000	3.66667	.	.	4.00000	.
A2	4.29000	3.50667
A3	4.30000	3.46667	4.00000	2.00000	.	.
P1	4.08000	.	.	3.00000	.	.
P2	4.15000	3.73333	.	3.00000	.	.
P3	4.06000	3.70667	6.00000	.	.	6.00000

The first table in the output shows that the most dominant missing pattern “xxxxxxxx.xx” has one missing variable, which is denoted by a dot in the pattern. About 38% of the observations (N=75) have this missing pattern. All other missing patterns shown in the table are relatively trivial because each has only one observation. Notice that for the current simulated data, the total number of distinct missing patterns is 26, as shown in the title of the first table. However, PROC CALIS shows only the five most dominant missing patterns. Showing more missing patterns (each with a frequency of 1) for the data set would not add to the understanding of the current missing pattern results. PROC CALIS uses a set of reasonable rules to determine the number of missing patterns to be displayed. Through the MAXMISSPAT=, NOMISSPAT, and TMISSPAT= options, the user can control the parameters of these rules to increase or decrease the number of missing patterns to be displayed. See Yung and Zhang (2011) for details.

The second table of the output shows the means of the most dominant missing patterns and the complete data (that is, the nonmissing pattern). You can use this table to locate the missing variables in the missing patterns. For example, P1 is a missing variable in the most dominant missing pattern because its mean is represented by a dot. Recall that P1 has also been identified as the most troublesome variable in the data coverage analysis. What could have happened to this variable? Are there any implications for practical research that you can draw from these results for data coverage and missing patterns? The answer is yes, but it depends on the substantive context of your research. See Yung and Zhang (2011) for a hypothetical explanation for the simulated data set.

Another notable observation from the missing pattern analysis is that all the variable means for the dominant missing pattern are consistently lower than those of the complete observations of the nonmissing pattern. This suggests that the observations with such a dominant missing pattern might have some peculiarity that warrants a separate analysis. Such a possibility might be explored in practical research.

Multiple-Group Analysis of Planned Missingness

When the number of missing patterns is small, a computationally more efficient multiple-group analysis can be carried out instead of the full information maximum likelihood estimation. This section illustrates the use of multiple-group analysis to treat certain planned missingness problems.

Bielby, Hauser, and Featherman (1977) collected $N=2020$ data to study the relationship between African-American fathers' occupational status (FAOC_t1) and educational attainment (FAED_t1). The correlation between these two variables was 0.43.

To address the possibility of measurement error, 348 participants were re-interviewed approximately three weeks later. The occupational status (FAOC_t2) and the education attainment (FAED_t2) were measured again. Bielby et al. (1977) would like to use the information from the second interview to assess the measurement errors so that a more accurate correlation between occupational status and educational attainment could be obtained. Wothke (2000) proposed a solution to this measurement error problem based on a multiple-group analysis of structural equation models. This section follows Wothke's (2000) suggestion and demonstrates the multiple-group analysis of PROC CALIS for treating the planned missingness problem.

Two sample covariance matrices are formed. One is for the 348 observations that were measured in both occasions *t1* and *t2* and the other is for the remaining 1672 observations that were measured only at *t1*. These two sample covariance matrices are represented by the following SAS data sets:

```

data complete(type=cov);
  input _type_ $ _name_ $ FAOC_t1 FAOC_t2 FAED_t1 FAED_t2;
  datalines;
COV    FAOC_t1  180.9      .      .      .
COV    FAOC_t2  126.77    217.56  .      .
COV    FAED_t1   23.96     30.20  16.24  .
COV    FAED_t2   22.86     30.47  14.36  15.13
MEAN   .         16.62     17.39   6.65   6.75
N      .         348      348    348    348
;

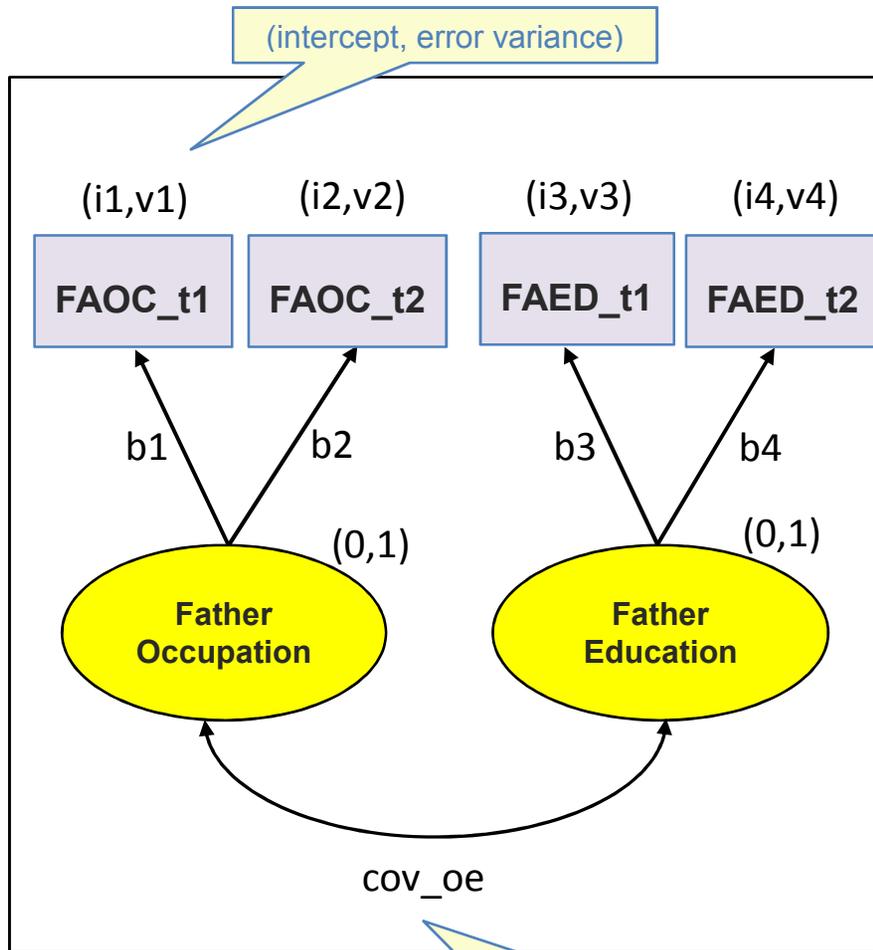
data incomplete(type=cov);
  input _type_ $ _name_ $ FAOC_t1 FAOC_t2 FAED_t1 FAED_t2;
  datalines;
COV    FAOC_t1  217.27      .      .      .
COV    FAOC_t2   .          .      .      .
COV    FAED_t1   25.57      .  16.16  .
COV    FAED_t2   .          .      .      .
MEAN   .         16.98      .   6.83  .
N      .         1672      .  1672  .
;

```

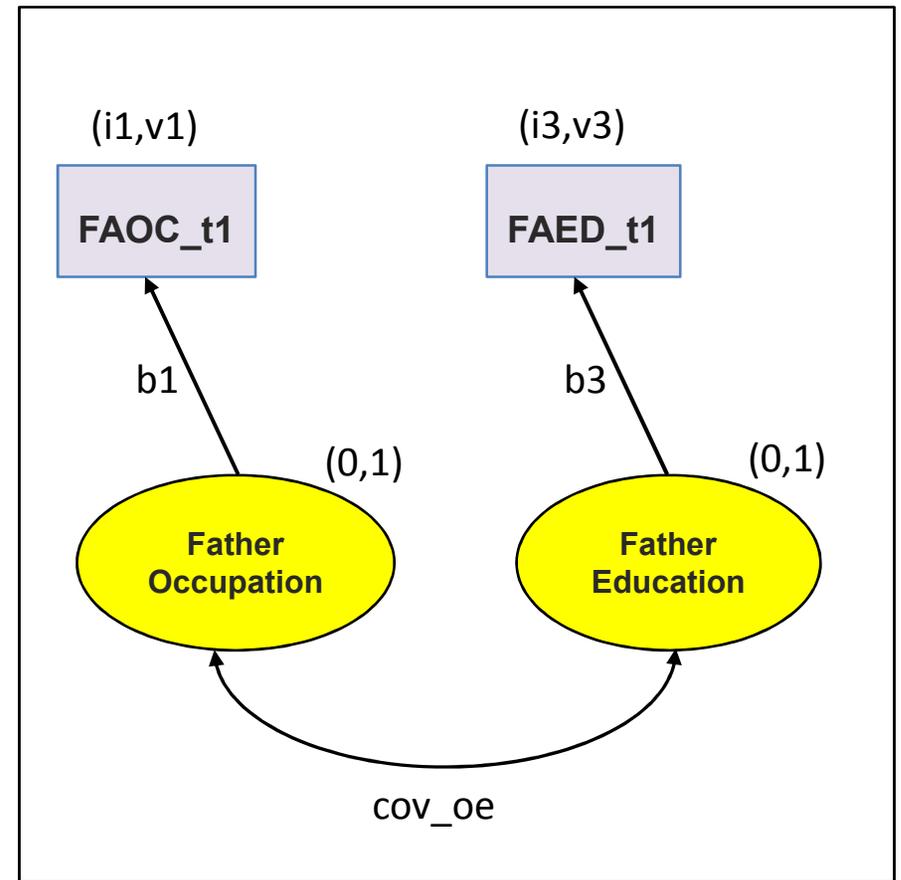
Clearly, there are two patterns of the data. One is the nonmissing pattern in the dataset “complete”. The other is the planned missing pattern in the dataset “incomplete”.

It would be useful to represent the measurement error models for the two covariance matrices by the following path diagrams:

Model for the Complete Data



Model for the Incomplete Data



Correlation with measurement errors filtered out

In the path diagram for the complete data, latent variables for fathers' occupation and education are created. These two latent variables represent "pure" or "true" measures without measurement errors. In contrast, the observed variables FAOC_t1, FAOC_t2, FAED_t1, and FAED_t2 merely reflect the true scores and are measured with errors. That is, each observed variable is predicted by its corresponding true latent measure with a path coefficient (b), an intercept term (i), and an error variance (v). The means and variances of the two latent variables are fixed to 0 and 1, respectively, for the identification of latent variable locations and scales. The most critical parameter in the diagram is cov_oe, which represents the correlation between fathers' occupation and education with the measurement errors filtered out.

The path diagram for the incomplete data is formed by replicating the specifications that are related only to the two nonmissing variables FAOC_t1 and FAED_t1. In addition, the parameters in the model for the incomplete data are constrained with those in the model for the complete data.

The path diagram is readily transcribed into the following PROC CALIS specification:

```
PROC CALIS;
  GROUP 1 / DATA=complete;
  GROUP 2 / DATA=incomplete;
  MODEL 1 / GROUP=1;
    PATH
      FatherOccupation ---> FAOC_t1  FAOC_t2  = b1 b2,
      FatherEducation   ---> FAED_t1  FAED_t2  = b3 b4;
    PVAR
      FAOC_t1  FAOC_t2  FAED_t1  FAED_t2 = v1-v4,
      FatherOccupation FatherEducation  = 1. 1.;
    PCOV
      FatherOccupation FatherEducation  = cov_oe;
    MEAN
      FAOC_t1  FAOC_t2  FAED_t1  FAED_t2 = i1-i4,
      FatherOccupation FatherEducation  = 0. 0.;
  MODEL 2 / GROUP=2;
    PATH
      FatherOccupation ---> FAOC_t1 = b1,
      FatherEducation   ---> FAED_t1 = b3;
    PVAR
      FAOC_t1  FAED_t1 = v1 v3,
      FatherOccupation FatherEducation = 1. 1.;
    PCOV
      FatherOccupation FatherEducation = cov_oe;
    MEAN
      FAOC_t1  FAED_t1 = i1 i3,
      FatherOccupation FatherEducation  = 0. 0.;
  RUN;
```

The following fit summary table shows that the measurement error model provides a very good fit of the data.

Chi-Square	7.6917
Chi-Square DF	6
Pr > Chi-Square	0.2616
Standardized RMSR (SRMSR)	0.0367
RMSEA Estimate	0.0167
Bentler Comparative Fit Index	0.9987

The following output shows that the correlation between occupation and education estimated under the measurement error model is 0.62, which is much larger than the correlation 0.43 measured at *t1* by using the observed scores with measurement errors. This shows that if the measurement errors are not probably filtered out, the true correlation between fathers' occupation and education could be seriously underestimated.

Model 1. Covariances Among Exogenous Variables					
Var1	Var2	Parameter	Estimate	Standard Error	t Value
FatherOccupation	FatherEducation	cov_oe	0.61659	0.02622	23.51558

Conclusions

This presentation contends that the FIML method is the preferred method for treating incomplete observations, especially when the sample size is not large. Although the FIML and multiple imputation methods produce similar estimation results, the FIML method is more convenient to use. In PROC CALIS, one can simply use the METHOD=FIML option to do the FIML estimation with incomplete observations. But the multiple imputation method is accomplished in separate stages that create the imputed datasets, fit the model repeatedly for the imputed datasets, and combine the estimation results, respectively. In addition, the FIML method can produce the model fit chi-square statistic and other fit indices in a straightforward manner, which the MI method cannot do that easily.

To help locate the sources of problematic missingness, PROC CALIS produces outputs that show the data coverages of the sample moments, the ranking of the coverages, and the dominant missing patterns and their mean profile. To make the results more concise, PROC CALIS outputs only the most dominant missing patterns by a set of reasonable rules. But the users are also given options to control the amount of the output for missing patterns.

When the data contain only a very small number of missing patterns, a multiple-group analysis can be more efficient computationally. PROC CALIS also provides syntax that the users can specify the multiple-group analysis intuitively and efficiently.

References

- 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.
- Marjoribanks, K., ed. (1974), *Environments for Learning*, London: National Foundation for Educational Research Publications.
- Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.): *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Yung, Y. F., & Zhang, W. (2011). Making use of incomplete observations in the analysis of structural equation models: The CALIS procedure's full information maximum likelihood method in SAS/STAT® 9.3. Paper 333-2011, SAS Global Forum, Las Vegas, NV. (<http://support.sas.com/resources/papers/proceedings11/333-2011.pdf>)