Paper 333-2011

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

Yiu-Fai Yung and Wei Zhang, SAS Institute Inc., Cary NC

ABSTRACT

Data are precious. Missing values are unavoidable. Instead of throwing away incomplete observations or imputing missing values ad hoc, what would be a better way to analyze data with random missing values? The full information maximum likelihood (FIML) method of the CALIS procedure in SAS/STAT® 9.22 and later enables you to use all the available information in your data to estimate your structural equation models. This paper illustrates how you can use PROC CALIS to deal with random missing values in the following data-analytic situations: (1) estimating means and covariances, (2) regression analysis, and (3) structural equation or path modeling.

Other methods for treating incomplete observations are described in a systematic manner. The FIML method is superior to the ad hoc methods for analysis with incomplete observations. With PROC CALIS, the FIML method is more convenient than the multiple imputation (MI) method for fitting path or structural equation models with incomplete observations. This paper also illustrates some new features of PROC CALIS for analyzing missing patterns and data coverages.

INTRODUCTION

Missing data are common in practical research studies. For example, a patient might forget a follow-up appointment with the research laboratory that takes his blood pressure measurements after a treatment. A respondent to a job satisfaction questionnaire might skip some items (or questions) in a survey. The responses from the patient and the respondent in the survey are referred to as incomplete observations. That is, an incomplete observation is a research unit (for example, the patient or the respondent in the preceding examples) that has at least one missing value (but not all missing) in the set of observed variables. In contrast, a complete observation refers to a research unit that has no missing values.

Most statistical analyses and software are traditionally developed for complete observations only. Incomplete observations with missing values cannot be handled and therefore are routinely excluded from statistical analyses, even though some incomplete observations have only a small portion of missing values. This results in the loss of valuable information that could have led to more reliable statistical estimation. This problem becomes more serious when the total number of observations in the data set is not very large and the researcher needs to use every bit of precious information from the data.

The full information maximum likelihood (FIML) method presented in this paper is an estimation method that uses the information from both the incomplete and the complete observations. The FIML estimation method for treating random missing data or incomplete observations is not limited to structural equation modeling (SEM); other types of analyses can also use the FIML estimation for treating incomplete observations. However, this paper focuses on the structural equation models that are analyzed by PROC CALIS, which introduced FIML estimation for handling incomplete observations data in SAS/STAT 9.22 and adds features for analyzing data coverages and missing patterns in SAS/STAT 9.3. Because SEM subsumes measurement error models, path analyses, regression models, LISREL models, and factor analyses as special cases, the FIML method of the CALIS procedure has a wide range of applications.

The remainder of this paper is organized as follows: First, some traditional and modern methods for treating incomplete observations are described and compared. All these methods are illustrated with the estimation of the population means and covariances. Second, the FIML method is shown to be superior to the traditional methods for treating incomplete observations. Because of its availability in PROC CALIS, the FIML method is also more convenient to use than the multiple imputation method, which has statistical properties similar to those of the FIML method. Third, the FIML estimation is illustrated by applying PROC CALIS to a regression problem with incomplete observations. Output from PROC CALIS is described and interpreted. Estimation results from PROC CALIS are compared with the results obtained from PROC REG, which uses listwise deletion to simply ignore the incomplete observations. Fourth, the FIML method of PROC CALIS is applied to a structural equation model. In addition to the FIML estimation of the model parameters, this application also illustrates how you can use PROC CALIS to study the data coverages of the sample moments and to analyze the missing patterns. Fifth, the options and controls related to the FIML method of PROC CALIS are described in detail. Finally, some limitations of the FIML method are described, followed by some concluding comments.

METHODS FOR TREATING INCOMPLETE OBSERVATIONS

Various methods have been proposed for treating incomplete observations in statistical analysis. This section first describes conventional ad hoc missing data methods and then describes the more principled methods such as the FIML and the multiple imputation methods. A comprehensive review of all existing incomplete data methods is not intended.

The following fictitious data, which are read in the SAS DATA step, are used to illustrate various incomplete data methods:

```
data miss1;
   input ID x1 x2 x3 x4 x5;
   datalines;
1   2   3   6   7   .
2   .   3   5   2   6
3   6   5   4   1   5
4   7   6   .   1   7
5   6   7   7   2   6
6   2   3   1   .   7
7   1   1   2   6   9
8   3   5   6   2   1
9   5   .   .   1   8
10  4   4   4   1   2
11  2   2   5   5   7
12  3   4   2   6   8
;
```

Missing data values are represented by the dots in the data set. Observations (as indicated by the ID variable) 1, 2, 4, 6, and 9 have missing values for one or more variables. These five observations are incomplete observations. The others are complete observations. Suppose you want to estimate the population means and covariances of all the x-variables. If the data set contained all complete data, the estimates would have been the sample means and covariances based on all observations. With the presence of incomplete data, there could be different possibilities.

Pairwise Deletion (PD) Method

Without thinking much about how to treat the incomplete observations, you might simply submit this data set to the CORR procedure to obtain the means and covariances, just as you do with complete data:

```
proc corr data=miss1 cov;
   var x1-x5;
   run;
```

Output 1 shows the estimates of the covariances and the means. The second table of Output 1 shows that the sample sizes for computing the means are different for the variables. For example, the mean for x3 is computed by averaging N=10 nonmissing values, while all other variable means are computing based on N=11 nonmissing values. Similarly, the first table of Output 1 shows that the covariances are computed based on different degrees of freedom (that is, different variance divisors), which range from 8 to 10. For example, the covariance between x1 and x3 is computed based on N=9 observations (hence with DF=8). In computing this covariance, observations 2, 4, and 9 have been excluded because they have at least one missing value in the x1-x3 variable-pair. The covariance between x1 and x5 is computed based on N=10 observations (hence with DF=9) because incomplete observations 1 and 2 are excluded. The mean and covariance computations here exemplify the pairwise deletion (PD) method for treating the incomplete observations. The estimates of means and covariances are computed based on the corresponding available cases (observations) with nonmissing values. For this reason, the PD method is sometimes called the available-case analysis. In this sense, the PD method uses all the available information from the data. With the PD method, different estimates are computed based on different sets of observations. As a result, this method might lead to a covariance matrix that is not positive-definite.

Output 1 Covariances and Means Computed from the Pairwise Deletion (PD) Method

```
                          Variances and Covariances
               Covariance / Row Var Variance / Col Var Variance / DF

                 x1             x2             x3             x4             x5

x1      4.018181818    3.333333333    1.597222222   -4.088888889   -0.777777778
        4.018181818    4.266666667    3.194444444    4.100000000    4.100000000
        4.018181818    3.333333333    4.361111111    6.177777778    6.888888889
                 10              9              8              9              9

x2      3.333333333    3.090909091    1.733333333   -2.888888889   -1.777777778
        3.333333333    3.090909091    2.900000000    3.333333333    3.333333333
        4.266666667    3.090909091    3.955555556    5.788888889    6.400000000
                  9             10              9              9              9

x3      1.597222222    1.733333333    3.955555556   -1.347222222   -2.750000000
        4.361111111    3.955555556    3.955555556    3.027777778    4.000000000
        3.194444444    2.900000000    3.955555556    5.777777778    7.000000000
                  8              9              9              8              8

x4     -4.088888889   -2.888888889   -1.347222222    5.690909091    2.855555556
        6.177777778    5.788888889    5.777777778    5.690909091    4.455555556
        4.100000000    3.333333333    3.027777778    5.690909091    6.766666667
                  9              9              8             10              9

x5     -0.777777778   -1.777777778   -2.750000000    2.855555556    6.200000000
        6.888888889    6.400000000    7.000000000    6.766666667    6.200000000
        4.100000000    3.333333333    4.000000000    4.455555556    6.200000000
                  9              9              8              9             10

                              Simple Statistics

Variable           N        Mean     Std Dev         Sum     Minimum     Maximum

x1                11     3.72727     2.00454    41.00000     1.00000     7.00000
x2                11     3.90909     1.75810    43.00000     1.00000     7.00000
x3                10     4.20000     1.98886    42.00000     1.00000     7.00000
x4                11     3.09091     2.38556    34.00000     1.00000     7.00000
x5                11     6.00000     2.48998    66.00000     1.00000     9.00000
```

Listwise Deletion (LD) Method

Unlike the PD method that uses all available information from the incomplete observations, the listwise deletion (LD) method (also called the complete case analysis) estimates the means and covariances based on the complete observations only. In other words, the LD method excludes all incomplete observations from the analysis. The LD method corresponds to the use of the NOMISS option in the PROC CORR statement, as shown in the following statements:

```
proc corr data=miss1 cov nomiss;
   var x1-x5;
   run;
```

The NOMISS option excludes incomplete observations 1, 2, 4, 6, and 9 from the computations of the mean and covariance estimates. Output 2 shows the estimates of covariances and the means with the LD method.

Output 2 Covariances and Means Computed from the Listwise Deletion (LD) Method

```
                           Covariance Matrix, DF = 6

                 x1             x2             x3             x4             x5

x1      3.619047619    3.333333333    1.809523810   -3.357142857   -1.952380952
x2      3.333333333    4.000000000    2.500000000   -3.166666667   -2.833333333
x3      1.809523810    2.500000000    3.571428571   -2.595238095   -2.976190476
x4     -3.357142857   -3.166666667   -2.595238095    5.238095238    5.523809524
x5     -1.952380952   -2.833333333   -2.976190476    5.523809524    8.952380952

                              Simple Statistics

Variable           N        Mean     Std Dev         Sum     Minimum     Maximum

x1                 7     3.57143     1.90238    25.00000     1.00000     6.00000
x2                 7     4.00000     2.00000    28.00000     1.00000     7.00000
x3                 7     4.28571     1.88982    30.00000     2.00000     7.00000
x4                 7     3.28571     2.28869    23.00000     1.00000     6.00000
x5                 7     5.42857     2.99205    38.00000     1.00000     9.00000
```

The second table of Output 2 shows that all the means are computed based on the seven complete observations. This set of seven complete observations is also used for computing the covariance matrix in the first table of Output 2. Hence, the LD method uses much less information in estimation than the PD method, which uses at least nine data values in estimating each sample mean or covariance.

Mean Imputation Method

The mean imputation method replaces all the missing values with their corresponding variable means (from the available cases) and then carries out the estimation as if there were no incomplete observations. For example, you can impute those means obtained from the PD method (available-case analysis) shown in Output 1 for the corresponding missing values in the original data set to form a pseudo-complete data set. The following statements do such mean imputations on the miss1 data set and then obtain the means and covariances by running the CORR procedure with the pseudo-complete data set miss2:

```
data miss2;
   set miss1;
   if (x1 = .) then x1 = 3.72727;
   if (x2 = .) then x2 = 3.90909;
   if (x3 = .) then x3 = 4.20000;
   if (x4 = .) then x4 = 3.09091;
   if (x5 = .) then x5 = 6.00000;
run;

proc corr data=miss2 cov;
   var x1-x5;
   run;
```

Output 3 shows the estimates of covariances and means from the mean imputation method.

Output 3 Covariances and Means Computed from the Mean Imputation Method

```
                           Covariance Matrix, DF = 11

                 x1             x2             x3             x4             x5

x1      3.652892562    2.716754440    1.198346909   -3.328324440   -0.636363636
x2      2.716754440    2.809917355    1.418181818   -2.346356026   -1.471074545
x3      1.198346909    1.418181818    3.236363636   -0.844628364   -1.945454545
x4     -3.328324440   -2.346356026   -0.844628364    5.173553719    2.371900909
x5     -0.636363636   -1.471074545   -1.945454545    2.371900909    5.636363636

                              Simple Statistics

Variable           N        Mean     Std Dev          Sum     Minimum     Maximum

x1                12     3.72727     1.91125     44.72727     1.00000     7.00000
x2                12     3.90909     1.67628     46.90909     1.00000     7.00000
x3                12     4.20000     1.79899     50.40000     1.00000     7.00000
x4                12     3.09091     2.27454     37.09091     1.00000     7.00000
x5                12     6.00000     2.37410     72.00000     1.00000     9.00000
```

The sample size used in the computing the means and covariances is 12, which is the same as the total number of observations in the original data set. This sample size has been inflated due to the use of the pseudo-complete data set miss2, which imputes the means for the missing values as if they were actually observed. In this sense, the mean imputation method overuses the information from the original data. As a result, it overstates the certainty of the imputed values and the estimates. The same fact can also be illustrated by the following observation. The mean imputation method shows the same set of mean estimates in Output 3 as that of the PD method, as shown in Output 1. However, all the variance and covariance estimates in Output 3 for the mean imputation method are smaller than the corresponding PD estimates in Output 1. Therefore, even though the mean imputation method does not alter the means from the available-case analysis, it does make the data distribution look more centralized (that is, with observations more alike) than it is. Hence, the mean imputation method underestimates the true variability of the variables and the sampling errors of the estimates.

Summary of Various Methods for Treating Incomplete Observations

So far, the three conventional methods for treating incomplete observations are ad hoc methods because they do not propose a probability model to account for the missingness. These methods "work" simply because their implementations enable statistical software to generate output as usual. Whether the estimates obtained from these ad hoc methods have desirable statistical properties is another story. Nonetheless, the "ad hoc" label for these methods does not imply that these methods must be bad in all situations. In fact, under certain (restrictive) assumptions about the nature of the missingness, some of these ad hoc methods might possess some desirable statistical properties. The section "Advantages and Disadvantages of the FIML Method Compared with the MI Method" on page 10 discusses these relative advantages and disadvantages in more detail.

The ad hoc methods discussed so far are summarized in Table 1, according to how they deal with the missing values in the incomplete observations:

Table 1 Treatment of Incomplete Observations by Various Methods

	Ad Hoc Methods			More Principled Methods	
Incomplete Observations Treatment	LD	PD	Mean Imputation	FIML	Multiple Imputation
Not analyzed	x				
Analyzed without imputation		x		x	
Analyzed with imputation			x		x

The LD method does not analyze any incomplete observations, so it necessarily entails a loss of information. On the other hand, the mean imputation method imputes variable means for the missing values and then treats the data set as if it contained only complete observations. This entails an inflation of the certainty of the (imputed) information. As a result, standard errors of the estimates are usually underestimated by the mean imputation methods. The PD method might use just about the right amount of information from the incomplete observations. It does not abandon the available information in the incomplete observations nor does it create pseudo-values for the missing data to inflate the amount of information.

Note that the mean imputation method represents a *single* imputation method, to be distinguished from the multiple imputation (MI) method, which is also included in Table 1 for comparison. There are other variants of the single imputation method, including the similar response pattern imputation (SRPI) method (see, for example, Jöreskog and Sörbom 1993), and imputation methods based on regression or sampling from other responses in the data (that is, hot-deck imputation). All these methods replace the missing values by some imputed values so that the statistical analysis can be carried out as usual.

In contrast with the ad hoc methods, two principled methods, the full information maximum likelihood (FIML) and the multiple imputation (MI) methods, are also shown in Table 1. The FIML method can be viewed as a more theory-driven counterpart of the PD method. Both the FIML and PD methods analyze the incomplete observations without imputations, but the FIML method analyzes the model under the likelihood principle. The MI method can be viewed as a more advanced counterpart of the single imputation method, as represented by the mean imputation method. Both the MI and mean imputation methods analyze data with imputations for the missing values, but the MI method analyzes many imputation samples so as to take the uncertainty of the imputed values into account properly. The next two sections describe the two principled methods in more detail.

Not all methods for treating incomplete observations are included here, but this summary provides a framework for understanding various types of missing data methods in practical research. See, for example, Gold, Bentler, and Kim (2003) for a more complete overview of various techniques for treating incomplete observations in the field of structural equation modeling.

Full Information Maximum Likelihood (FIML) Method

The full information maximum likelihood (FIML) method is a principled estimation method that treats complete and incomplete observations in an integrated manner. Basically, the FIML estimation maximizes the sum of the log-likelihood functions for individual observations, including both complete and incomplete observations, as shown in the expression

$$F = \frac{1}{N} \sum_{j=1}^{N} (ln(|\mathbf{\Sigma}_j|) + (\boldsymbol{x}_j - \boldsymbol{\mu}_j)' \mathbf{\Sigma}_j^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_j) + K_j)$$

where $\boldsymbol{x}_j$ is a data vector for observation j, and K_j is a constant term independent of the model parameters $\mathbf{\Theta}$. Individual observations $\boldsymbol{x}_j$'s are not required to have the same dimensions. For example, $\boldsymbol{x}_1$ could be a complete vector with the presence of all p variables while $\boldsymbol{x}_2$ is a $(p-1)\times 1$ vector with one missing value that has been excluded from the original $p \times 1$ data vector. As a consequence, subscript j is also used in $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ to denote the subvector and submatrix, respectively, that are extracted from the entire $p \times 1$ structured mean vector $\boldsymbol{\mu}$ ($\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\Theta})$) and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\Theta})$). In other words, in the current formulation, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ do not mean that each observation is fitted by distinct mean and covariance structures (although theoretically it is possible to formulate FIML in such a way). The notation simply signifies that the dimensions of $\boldsymbol{x}_j$ and the associated mean and covariance structures could vary from observation to observation.

Because the log-likelihood function for individual observations contains only the nonmissing variables, you can view the FIML method as a principled version of the PD method (available-case analysis). Like the PD method, the FIML method does not impute missing values in the estimation. Unlike the PD method, the FIML method ties to the model of interest directly. Model parameters are estimated directly by maximizing the likelihood function of the model parameters, which might not be as simple as population means and covariances. In contrast, the PD method first computes sample statistics such as the covariance matrix and the mean vector. Then it estimates the model parameters through these sample statistics.

To continue the previous data example, you can use the following statements to perform FIML estimation of the population means and covariances with the miss1 data set:

```
proc calis data=miss1 method=fiml;
   mstruct var=x1-x5;
   run;
```

Although PROC CALIS is designed as a procedure for general structural equation modeling, it supports a wide variety of models, including regression, factor analysis, measurement error models, and so on. You can also use PROC CALIS to fit an unstructured model such as the one shown here. The preceding statements specify an MSTRUCT (matrix structure) model with all the x-variables in the miss1 data set. By default, the means and covariances of such an MSTRUCT model are unstructured, which means that all elements in the population covariance matrix and mean vector are mathematically independent parameters. You specify METHOD=FIML so that PROC CALIS carries out a full information maximum estimation of the population means and covariances.

Output 4 shows some basic modeling information about the current data set. PROC CALIS identifies seven complete observations and five incomplete observations in the data set. All these observations are used in the estimation.

Output 4 Modeling Information about the Incomplete Data

```
                    Modeling Information

               Data Set                WORK.MISS1
               N Records Read          12
               N Complete Records      7
               N Incomplete Records    5
               N Complete Obs          7
               N Incomplete Obs        5
               Model Type              MSTRUCT
               Analysis                Means and Covariances
```

Output 5 shows the estimates of population means and covariances by the FIML method of PROC CALIS.

Output 5 Means and Covariances Estimated by the FIML Method

```
                    MSTRUCT _Mean_ Vector

                                  Standard
           Variable       Estimate       Error        t Value

           x1              3.75216     0.52975        7.08289
           x2              3.79040     0.48112        7.87834
           x3              4.46354     0.58507        7.62911
           x4              3.32587     0.66907        4.97089
           x5              6.36549     0.74780        8.51226
```

Output 5 *continued*

```
                  MSTRUCT _COV_ Matrix: Estimate/StdErr/t-value

                x1            x2            x3            x4            x5

x1          3.3611        2.3196        1.7248       -3.4898       -1.2238
            1.3748        1.1079        1.1839        1.5880        1.4172
            2.4448        2.0937        1.4568       -2.1976       -0.8635

x2          2.3196        2.7530        1.4872       -2.0884       -1.8747
            1.1079        1.1334        1.0669        1.2678        1.3585
            2.0937        2.4289        1.3940       -1.6472       -1.3800

x3          1.7248        1.4872        3.6784       -2.1391       -0.8334
            1.1839        1.0669        1.6184        1.4897        1.5354
            1.4568        1.3940        2.2728       -1.4359       -0.5428

x4         -3.4898       -2.0884       -2.1391        5.3601        3.7521
            1.5880        1.2678        1.4897        2.1930        2.0435
           -2.1976       -1.6472       -1.4359        2.4442        1.8361

x5         -1.2238       -1.8747       -0.8334        3.7521        6.6711
            1.4172        1.3585        1.5354        2.0435        2.7390
           -0.8635       -1.3800       -0.5428        1.8361        2.4356
```

In addition to the estimates, PROC CALIS also computes the standard errors and t values for these estimates. Apparently, PROC CALIS produces a set of estimates that is different from those of the ad hoc methods. Why should you trust the FIML estimates more than the ad hoc methods? The section “Advantages and Disadvantages of the FIML Method Compared with the MI Method” on page 10 justifies the FIML estimation.

FIML or ML? A Clarification

This section clarifies the “full information” terminology as used in the FIML method. Because this section is not essential for understanding the FIML methodology, you might choose to skip this section.

The FIML method for treating incomplete observations is sometimes called the direct maximum likelihood method because the FIML method is basically an ML estimation applied *directly* on individual observations. In fact, some statistical literature simply refers this method to as the ML method for incomplete data analysis. The “full information” qualifier is dropped. So, why add the “full information” qualifier when the ML label is already a clear terminology that describes the principle behind the missing data methodology?

In the literature of structural equation modeling (SEM), the ML estimation has been designated as an estimation method that fits the model covariance (and mean) structures to the sample covariance matrix (and mean vector) under the multivariate normal theory. Individual observations are not directly used in such an ML estimation. However, this does not mean that the ML method in SEM is not a “full information” method. When the data set contains only complete observations, under the multivariate normal distribution assumption the ML method in SEM is in fact a full information ML method, which means that analyzing the sample covariance matrix (and the sample mean vector) does not lose any information as compared with the analysis based on individual observations.

Therefore, the “full information” qualifier in FIML reflects more about the incomplete data treatment than about the underlying estimation principle. With the “full information” qualifier, the incomplete observations are included in the estimation. Without the “full information” qualifier, the incomplete observations are all excluded in the estimation. This is also how PROC CALIS distinguishes between the METHOD=FIML and METHOD=ML options. These two options essentially lead to the same estimation results if the data set contains only complete observations. However, if your data set contains missing values in the analysis variables, you must use the METHOD=FIML option to enjoy the advantages of the maximum likelihood estimation with incomplete observations.

Multiple Imputation (MI) Method

The multiple imputation (MI) method is considered to be an advanced version of the single imputation method (such as the mean imputation method) but with the uncertainty of the imputed values taken into account. See, for example, Sinharay, Stern, and Russell (2001) or Schafer and Olsen (1998) for a nontechnical overview of the MI method. By assuming a statistical model for the data, the MI method generates random values from certain conditional distributions to replace the missing values in the original sample. Such an imputed sample is then analyzed by the target statistical model as if it contained all complete observations. The MI method then repeats the process several times (usually 5–10 times) by drawing multiple imputed samples for model fitting and estimation. The analytic results from all the imputed samples are then combined in such a way that the uncertainty of the imputed values is taken into account properly.

The following statements apply the MI method to estimate the population means and covariances of the miss1 data set:

```
/*------ Stage 1: Use PROC MI to generate 20 imputation samples ------*/
proc mi data=miss1 nimpute=20 seed=135782 out=ImputedSamples;
   var x1-x5;
run;

/*---- Stage 2: Use PROC CALIS to analyze the 20 imputed samples ----*/
proc calis data=ImputedSamples
   outmodel=Est(rename=(_name_=parm _estim_=estimate _stderr_=stderr));
   mstruct var=x1-x5;
   matrix _cov_  = s11
                   s12 s22
                   s13 s23 s33
                   s14 s24 s34 s44
                   s15 s25 s35 s45 s55;
   matrix _mean_ = m1-m5;
   by _Imputation_;
run;

/*---- Stage 3: Use PROC MIANALYZE to combine the estimation results ----*/
proc mianalyze parms=Est;
   modeleffects s11 s12 s22 s13 s23 s33 s14 s24 s34 s44 s15 s25 s35 s45 s55
                m1 m2 m3 m4 m5;
run;
```

As indicated by the preceding specification, the MI method has three stages:

- The first stage generates the imputed samples by using the MI procedure. In the PROC MI statement, the DATA= option specifies the original data set that might contain 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 called _Imputation_ to index the 20 imputed data sets within the OUT= data set.

- The second stage analyzes the sample means and covariances of the 20 imputed samples by using the CALIS procedure. You could have used some simpler procedures such as PROC CORR to compute the means and covariances. PROC CALIS is used here because the code is comparable to what has been done previously with the FIML estimation. In addition, PROC CALIS provides a handy way to name all the parameters in the analysis model. In the PROC CALIS statement, you use the DATA= option to specify ImputedSamples as the data set that contains the 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. PROC CALIS carries out separate model estimation for these 20 imputed samples as BY groups. The OUTMODEL= option in the PROC CALIS statement stores all 20 estimation results in the SAS data set named Est. Some of the variable names in this data set are renamed so that they are compatible with the MIANALYZE procedure in the next stage. Next, the MSTRUCT statement specifies the set of analysis variables x1–x5 in the VAR= option. Like the PROC CALIS specification for the FIML method in a preceding section, the MSTRUCT statement here specifies an unstructured model in which all variable means and covariances are distinct parameters of interest. The additional specifications with the two MATRIX statements label the parameters in the covariance matrix and the mean vector. That is, parameters with the “s” prefix are for naming the covariances, and parameters m1–m5 are for naming the variable means.

- The last stage of the MI method combines the estimation results of the 20 imputed samples. In the PROC MIANALYZE statement, the PARMS= option specifies that the SAS data set Est contains the estimates of the 20 imputed samples. The MODELEFFECTS statement specifies that the parameters of interest are those covariance and mean parameters that are named by the CALIS procedure in the previous stage. PROC MIANALYZE combines the 20 estimation results and shows the final estimation results in Output 6.

Output 6 Means and Covariances Estimated by the MI Method

```
                         Parameter Estimates

              Parameter        Estimate      Std Error

              s11              3.670674       1.565353
              s12              2.542071       1.265561
              s22              3.007762       1.295771
              s13              2.055875       1.757924
              s23              1.780442       1.401083
              s33              4.995708       3.061973
              s14             -3.767178       1.796987
              s24             -2.279206       1.444503
              s34             -2.388182       2.051419
              s44              5.771349       2.482014
              s15             -1.365900       1.655267
              s25             -2.047802       1.572461
              s35             -0.858875       1.943426
              s45              4.146591       2.404187
              s55              7.540717       3.453811
              m1               3.749265       0.578620
              m2               3.796816       0.526994
              m3               4.510968       0.765780
              m4               3.307360       0.726599
              m5               6.382312       0.836285
```

Overall, the MI estimates are similar to those obtained from the FIML method and are shown in Output 5. Table 2 summarizes the mean estimates by various methods for the miss1 data set.

Table 2 Mean Estimates by Various Missing Data Methods for the miss1 Data Set

Variable	LD	PD	Mean Imputation	FIML	MI
x1	3.57143	3.72727	3.72727	3.75216	3.749265
x2	4.00000	3.90909	3.90909	3.79040	3.796816
x3	4.28571	4.20000	4.20000	4.46354	4.510968
x4	3.28571	3.09091	3.09091	3.32587	3.307360
x5	5.42857	6.00000	6.00000	6.36549	6.382312

As discussed previously, the PD and the mean imputation methods have the same set of mean estimates—an artifact created by imputing means for the missing values in the latter method. The FIML and the MI methods produce similar sets of mean estimates. The sets of estimates produced by the ad hoc methods are somewhat different from those of the FIML and the MI methods. For example, the FIML and the MI mean estimates for x5 match quite well, while all the ad hoc methods produce much smaller mean estimates for x5.

In summary, this section illustrates different methods for treating incomplete observations. Different methods can produce different estimation results. Which sets of results should you trust? The answer is the FIML and the MI estimates. The next section provides the justifications of these methods.

ADVANTAGES OF THE FIML METHOD

The following subsections describe the advantages of the FIML method, as compared with other methods for treating incomplete observations. The FIML method is clearly superior to the ad hoc methods. The FIML and the MI methods have similar statistical properties, but the availability of the FIML method in PROC CALIS makes it more convenient to use.

Advantages of the FIML Method Compared with the Ad Hoc Methods

The advantages of the FIML method, compared with the ad hoc methods, are described under the following two conditions: the kind of missing mechanisms and the underlying distribution of the data.

Following Rubin (1976), three kinds of missing mechanisms are distinguished. When the missingness of a variable does not depend on the values of any variables (including the variable itself), it is called missing completely at random (MCAR). This is the strictest criterion for random missingness. This definition also matches the common-sense definition for random missingness. When the missingness of a variable depends on the values of other variables but does not depend on its own values, it is called missing at random (MAR). Hence, when the data are MCAR, they are automatically MAR—a less stringent criterion for random missingness. When the missingness of variable does depend on its own values, it is called missing not at random (MNAR). Consider an example in which blood pressure of patients is measured. Patients in different age groups might have different degrees of missingness because they might have different levels of health awareness. The younger groups might tend to miss more appointments for measuring blood pressure than the older groups. In this case, the missingness of the blood pressure measurements depends on the age. This missingness is still MAR if the missingness does not depend on the actual blood pressure of the patients. However, if patients who have normal blood pressure levels tend to ignore their blood pressure appointments more than the patients with either high or low blood pressure, then the missingness of the blood pressure measurement depends on the blood pressure level itself. Hence, the data is MNAR in this situation. Finally, if the missingness of the blood pressure measurement does not depend on the blood pressure value itself or any other variable values, it is MCAR.

Assume that the data are multivariate-normally distributed. Under MCAR and MAR, the FIML method for treating incomplete observations is consistent and efficient. Consistency refers to the fact that the FIML estimates converge to the true values probabilistically when the sample size becomes large. Efficiency refers to the fact that the FIML estimates attain the minimum variance property asymptotically. In contrast, the PD and LD estimates are consistent but not efficient under MCAR. Under MAR, the PD and LD estimates are not consistent and can be biased (Arbuckle 1996). The mean imputation method does not yield consistent estimates, and its standard error estimates are negatively biased. Simulation studies illustrate these results. For example, Enders and Bandalos (2001) conduct a simulation study about the relative performance of the FIML estimation in structural equation models. They compare the statistical behavior of the FIML method with the LD, PD, and SRPI (which applies a single imputation) methods. Their simulation results indicate that the FIML estimation is superior to all those ad hoc methods for treating incomplete observations in all conditions studied. The FIML estimates are unbiased and more efficient than the ad hoc methods. In addition, the FIML estimation produces the lowest proportion of convergence failures during optimization.

What happens if the data are not multivariate normal? Is the FIML method still better than the ad hoc methods? To answer this, Enders (2001) carries out another simulation study to compare the statistical behavior of the FIML method with the LD, PD, and mean imputation methods under nonnormal situations. His simulation results indicate that under MCAR and MAR, the FIML estimates involve less bias and are generally more efficient than those of the ad hoc methods.

Therefore, the relative statistical advantages of using the FIML method, compared with the ad hoc methods such as the LD, PD, and mean imputation methods, are widely recognized under the multivariate normal distribution. Even under nonnormal distributions, the FIML method has been shown to be superior to the ad hoc methods for treating incomplete observations.

Advantages and Disadvantages of the FIML Method Compared with the MI Method

The FIML method has the following advantages over the MI method:

- The FIML method involves only a single run of the analysis with the original sample. It is computationally more efficient than the MI method, which must fit multiple imputed samples.
- Given the data set, the FIML method always produces the same estimation results for fitting the same model. But the MI method might yield different estimation results, which depend on the random number generation algorithm and the seed of the random number sequence that is used in generating imputed values.
- In practice, the FIML estimates are slightly more efficient (that is, with smaller variance) than the estimates obtained from the MI method. The MI method loses efficiency because it has to generate random imputed values for the missing data (Schafer and Olsen 1998). However, with reasonably large sample sizes and large enough multiple imputations in MI, the FIML and MI methods should lead to essentially the same results (Schafer and Olsen 1998).
- For testing model fit, the FIML method can perform the likelihood ratio test easily and directly. Testing model fit with the MI method is much more complicated.

The MI method has the following advantages over the FIML method:

- The MI method is more general than the FIML method. Once the MI method is implemented, you can apply it to any data set, regardless of the models being fitted later on. The FIML method is model-specific, and it must be implemented separately for different types of models.

- The MI method can improve estimation by adding variables that are relevant to the missing data in the imputation stage (see Collins, Schafer, and Kam (2001) and the references therein). Adding variables that are relevant to the missing data in the FIML method is more difficult to achieve in general. Graham (2003) proposes a strategy in the field of SEM.

Because of the relative advantages of the FIML method and its availability in the CALIS procedure, the FIML method should be the method of choice for treating incomplete observations in structural equation and related models. The FIML method is clearly superior to all ad hoc methods. It also has the same or better statistical properties than the MI method. Although the FIML method is not as general a statistical tool as the MI method, the FIML method implemented in PROC CALIS already supports quite a large class of statistical models. Finally, with respect to testing model fit, the FIML method might be the only handy choice for practical users.

MULTIPLE REGRESSION EXAMPLE

This section illustrates the use of the FIML method of PROC CALIS for estimating parameters in multiple regression analysis. Again, the data set miss1 is used.

Suppose you want to regress variable x1 on variables x3–x5. You can use the following statements to specify the regression model:

```
proc reg data=miss1;
   model x1=x3-x5;
   run;
```

Output 7 shows the number of observations read and used in the regression analysis. There are five observations with missing values, and only seven observations are used in the analysis. Seven is only slightly more than a half of the original sample size. So, you can expect that the listwise deletion (LD) in the regression analysis loses a lot of valuable information in the incomplete observations.

Output 7 Number of Observations Read and Used in PROC REG

```
             Number of Observations Read                        12
             Number of Observations Used                         7
             Number of Observations with Missing Values          5
```

Output 7 shows the least squares estimates of PROC REG. These least squares estimates are also maximum likelihood (ML) estimates for the regression model. If you look at the t values of these parameter estimates, none of them are significant at the 0.05 α-level.

Output 8 Parameter Estimates by the REG Procedure

```
                                Parameter Estimates

                                Parameter       Standard
       Variable     DF           Estimate          Error    t Value    Pr > |t|

       Intercept     1            3.96408        2.06750       1.92      0.1510
       x3            1            0.11810        0.30423       0.39      0.7238
       x4            1           -1.12738        0.36139      -3.12      0.0525
       x5            1            0.51679        0.26009       1.99      0.1411
```

To exploit the information of the incomplete observations, you can use the FIML method of PROC CALIS to specify the same regression model, as shown in the following statements:

```
proc calis data=miss1 method=fiml;
   path x1 <--- x3-x5;
   run;
```

PROC CALIS does not use the same syntax as PROC REG uses to specify regression models. In fact, PROC CALIS has many different modeling languages for specifying various types of models, including path, factor-analysis, and regression models. The PATH statement in the current example specifies that there are three paths from variables x3–x5 to variable x1. This essentially states that x3–x5 predict x1, which is equivalent to the regression model specified with the REG procedure. What is different in the PROC CALIS specification is the use of METHOD=FIML, which ensures that all incomplete observations are analyzed together with the complete observations. The regression coefficient estimates by PROC CALIS are shown in Output 9.

Output 9 Parameter Estimates by the FIML Method of the CALIS Procedure

```
                                    PATH List

                                                             Standard
--------Path--------      Parameter       Estimate              Error          t Value

x1       <---     x3      _Parm1           0.28263            0.16034          1.76272
x1       <---     x4      _Parm2          -0.83912            0.14458         -5.80382
x1       <---     x5      _Parm3           0.36316            0.14565          2.49345
```

Output 9 shows that the estimates by the FIML method are different from those of the regression analysis with the listwise deletion (LD) of the incomplete observations. For example, the regression coefficient of x3 on x1 is 0.118 with the regression analysis, as shown in Output 8. However, with the incomplete observations included in the FIML method, this estimate is 0.283, as shown in Output 9. Because the FIML method has been shown to be superior to the LD method, this example illustrates that ignoring the incomplete observations (as with the LD treatment of PROC REG) might skew the estimation.

Another notable difference between the results in Output 8 and Output 9 is the magnitude of the estimated standard errors. The FIML method shows considerably smaller estimated standard errors for all estimates. In fact, it has been well-documented that the FIML estimates are more efficient than the ad hoc methods for treating incomplete observations. This example illustrates this point numerically. The statistical consequence is that the FIML estimates have more precise (narrower) confidence intervals and more statistical power in significance tests. In the field of SEM, the significance of a t value is referenced to the z-distribution. Therefore, an estimate is significantly different from zero at the 0.05 α-level if its corresponding absolute t value is greater than 1.96. Output 9 shows that the effects of x3 and x4 on x1 are statistically significant.

Output 10 shows the intercept estimate (for x1) by the FIML method of PROC CALIS, along with the estimates of the variable means for x3–x5. The intercept estimate by PROC CALIS, 3.132, is quite different from that of PROC REG, which is 3.964. Also, the FIML estimation has a much smaller estimated standard error (1.230) for the intercept estimate than that of PROC REG, which is 2.068.

Output 10 Intercept and Mean Estimates by the FIML Method of the CALIS Procedure

```
                               Means and Intercepts

                                                              Standard
Type          Variable      Parameter        Estimate            Error         t Value

Intercept     x1            _Add08            3.13153          1.23029         2.54536
Mean          x3            _Add09            4.11466          0.58428         7.04227
              x4            _Add10            3.20836          0.64712         4.95788
              x5            _Add11            6.15386          0.69481         8.85690
```

In summary, this example shows that estimation results in regression analysis could be quite different if the incomplete observations are not included in the analysis. Although PROC CALIS is not designed mainly for regression analysis, you can still use it to fit regression models with the FIML estimation, which treats the incomplete observations much more efficiently than the listwise deletion method applied automatically with the regression analysis.

DATA COVERAGES AND ANALYSIS OF MISSING PATTERNS

In addition to the FIML estimation with the specification of the METHOD=FIML (or METHOD=LSFIML) option, PROC CALIS outputs two groups of descriptive analyses about the incomplete observations. These descriptive analyses are available only in SAS/STAT 9.3 (or later).

- The first group of descriptive analyses shows the proportion coverage statistics for the means and covariances. The proportion coverage statistic is similar to the "available-case" concept in the pairwise deletion (PD) method. It measures the proportion of observations available to compute a particular sample moment (mean or covariance). For a sample mean, the proportion coverage shows the proportion of observations with nonmissing values in the corresponding variable. For a sample covariance, the proportion coverage shows the proportion of observations with nonmissing values in both of the variables involved. The proportion coverage statistics are useful for gauging the relative severity of missingness in the sample moments. They help locate the variables or sample moments that have the most serious missing problems or have the least information.

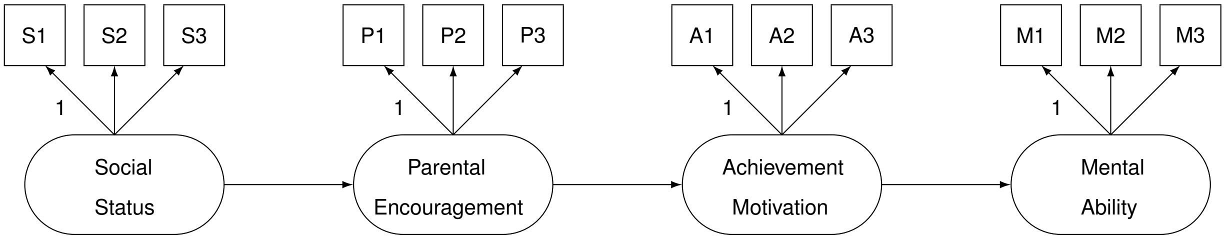
- The second group of descriptive analyses shows the dominant missing patterns and their means. The dominant missing patterns are those missing patterns that account for the majority of the incomplete observations in the data set. PROC CALIS ranks these dominant missing patterns in the output. The means of these dominant patterns are computed so that you can compare them with the means of the complete observations. PROC CALIS does not show all missing patterns in the output because the number of distinct missing patterns could be quite large in practical applications. However, you can control the amount of missing pattern output by some options. The section "PROC CALIS OPTIONS RELATED TO FIML ESTIMATION" on page 17 covers the details of these options and controls.

The next section illustrates the uses of these descriptive analyses in the FIML estimation of a structural equation model.

STRUCTURAL EQUATION MODEL EXAMPLE WITH INCOMPLETE OBSERVATIONS

This section applies the full information maximum likelihood (FIML) estimation to a structural equation model with latent variables. The path diagram shown in Output 1 is inspired by Marjoribanks (1974), but it is not intended to represent any theoretical model of Marjoribanks. The data in the current illustration are fictitious and contain incomplete observations, so no part of the current analysis is meant to be comparable to the results of Marjoribanks. The purpose here is to illustrate the FIML estimation by PROC CALIS in an interesting substantive context.

Figure 1 Factors Affecting Mental Abilities: Path Diagram



As expressed as a path diagram in Figure 1, a researcher proposes a theory about how Social Status predicts Mental Ability in a linear "causal" sequence of effects: Social Status predicts Parental Encouragement, which predicts Achievement Motivation, which predicts Mental Ability. All four variables are latent variables in the model. They are theoretical constructs that are not directly observed, but are reflected by some observed variables (or indicators). For example, S1–S3 are observed indicators for Social Status. S1 could be a variable that indicates the annual family salary class of an individual, with 7 being the highest salary class and 1 being the lowest. M1–M3 could be rating scales of an individual on three different mental tests. Again, seven-point scales are used for these observed variables. (This example assumes the seven-point rating scales for all observed variables only for the simplicity in presentation. The observed variables could be measured on ordinal, interval, or ratio scales.) The whole path diagram shown in Figure 1 forms the structural equation model.

In Figure 1, some paths are labeled with 1. This means that the corresponding path coefficients or effects are fixed to 1, which would not be changed during the estimation. For example, the effect of Social Status on S1 is fixed to 1. Such a fixed coefficient is necessary to identify the scale of the latent variable Social Status. Similarly, each latent construct in Figure 1 has one fixed path with one of its observed indicators. Due to the limited space, this paper cannot introduce the structural equation modeling more thoroughly. For more information, see the classic textbook by Bollen (1989) about the basics of structural equation modeling.

You can easily transcribe the path diagram in Figure 1 into the following specification by using the PATH modeling language of the CALIS procedure, as shown in the following statements:

```
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;
```

In the PROC CALIS statement, you specify the METHOD=FIML option so that the incomplete observations are analyzed together with the complete data. In the PATH statement, you specify the paths shown in Figure 1. The first four entries specify the measurement model for the theoretical constructs. You use the multiple-path specification syntax for the measurement model. For example, the first path entry means that SocialStatus has three paths to S1, S2, and S3. The first path to S1 has a fixed coefficient 1, while the other two paths have coefficients or effects to be freely estimated. The last three entries in the PATH statement represent the structural model, which describes the functional relationships between the latent variables.

Output 11 shows some general modeling information obtained from the PROC CALIS output. The first table shows that there are 200 data records in the miss3 data set, but only 100 are complete observations. With METHOD=FIML, the 100 incomplete observations are also analyzed together with the complete observations. The second table classifies the variables according to whether they are endogenous or exogenous and whether they are manifest (observed) or latent (unobserved).

Output 11 Basic Modeling Information with the FIML Estimation for the Miss3 Data Set

```
                              Modeling Information

                  Data Set                  WORK.MISS3
                  N Records Read            200
                  N Complete Records        100
                  N Incomplete Records      100
                  N Complete Obs            100
                  N Incomplete Obs          100
                  Model Type                PATH
                  Analysis                  Means and Covariances

                              Variables in the Model

       Endogenous     Manifest     A1  A2  A3  M1  M2  M3  P1  P2  P3  S1  S2
                                   S3
                      Latent       AchievementMotivation  MentalAbility
                                   ParentalEncouragement
       Exogenous      Manifest
                      Latent       SocialStatus

                           Number of Endogenous Variables = 15
                            Number of Exogenous Variables  = 1
```

Output 12 shows the proportions of data present for computing the sample means and covariances. The entries in the table are called proportion coverages. The diagonal of the matrix shows the proportion coverages of the means or the proportion coverages of the variables. The off-diagonal of the matrix shows the proportion coverages of the covariances or the joint proportion coverages of the variable-pairs. For example, the (1,1) entry shows that 92.5% of observations have nonmissing values in variable S1 or shows that 92.5% of observations are used for computing the sample mean of S1. The (2,1) entry shows that 89% of observations have nonmissing values in variables S1 and S2 or shows that 89% of observations are used for computing the covariance between S1 and S2. The average proportion coverages of means and covariances are shown at the bottom of Output 12. They are 88.4% and 82.4%, respectively. Overall, the missing data problem does not seem to be very serious.

Output 12 Proportion Coverages

```
                  Proportions of Data Present for Means (Diagonal) and Covariances (Off-Diagonal)

                S1        S2        S3        M1        M2        M3        A1        A2        A3        P1        P2        P3

      S1    0.9250
      S2    0.8900    0.9250
      S3    0.8800    0.8900    0.9100
      M1    0.8750    0.8800    0.8800    0.8900
      M2    0.8950    0.8950    0.8800    0.8750    0.9150
      M3    0.9000    0.8950    0.8850    0.8850    0.8900    0.9200
      A1    0.8900    0.8850    0.8850    0.8800    0.8900    0.8850    0.9200
      A2    0.8950    0.8800    0.8750    0.8750    0.8850    0.8850    0.8850    0.9000
      A3    0.8900    0.8900    0.8800    0.8800    0.8800    0.8800    0.8850    0.8800    0.9100
      P1    0.5150    0.5100    0.5000    0.5050    0.5100    0.5100    0.5150    0.5100    0.5150    0.5350
      P2    0.8900    0.9100    0.8900    0.8800    0.9050    0.8850    0.9000    0.8900    0.8950    0.5150    0.9350
      P3    0.8850    0.8850    0.8900    0.8800    0.8800    0.8900    0.8900    0.8850    0.8950    0.5100    0.8900    0.9200


                              Average Proportion Coverage of Means            0.883750
                              Average Proportion Coverage of Covariances      0.824015
```

In order to locate the smallest coverages of the sample moments, Output 13 orders the smallest coverages for the means and for the covariances. The first table shows that variable 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.

Output 13 Smallest Coverages

```
             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
```

Output 14 shows the dominant missing patterns in the data set and their means. The first table in Output 14 shows that the most dominant missing pattern “xxxxxxxxx.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 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 current data set would not add to the understanding of the current results. See the next section for the details about how to control the amount of output for displaying missing patterns.

Output 14 Missing Patterns and Their Mean Profiles

```
               Rank Order of the 5 Most Frequent Missing Patterns
           Total Number of Distinct Patterns with Missing Values = 26

                                NVar
                Pattern         Miss    Freq    Proportion    Cumulative

           1    xxxxxxxxx.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

                      ----------------------Missing Pattern---------------------
          Nonmissing           1           2           3           4           5
Variable     (N=100)      (N=75)       (N=1)       (N=1)       (N=1)       (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 second table of Output 14 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.

To complete the argument, suppose P1 is a rating variable that indicates whether an individual receives consistent parental encouragement. The item might be phrased as: “Both of my parents set the same goals for me to achieve.” Respondents who live with a single parent might not give responses to this item. Although the missingness of this variable seems to be determined by some other variables, it does not mean that the nature of the missingness of this variable must be MNAR. If these respondents had both parents living with them, would their missingness in this variable be dependent on their potential responses to the item? If yes, then the missingness is MNAR. Otherwise, it is not MNAR. Unfortunately, there seems to be no way to answer such a hypothetical question. In any case, the dominant missing pattern here exposes the problematic item, which you might want to revise in future research so as to eliminate the ambiguity.

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.

Output 15 shows the fit summary table. PROC CALIS shows all the available fit indices by default. Although it is not done here, you can control the amount of fit indices to display in this table by using the FITINDEX statement. To test whether the proposed structural equation model is significantly worse than the saturated model for the data, you can use the chi-square model fit statistic, which is defined as the -2 multiple of the difference between the log-likelihood values under the theoretical model and the saturated model. As shown in Output 15, the chi-square model fit statistic is 58.6765 ($df = 51$, $p = 0.22$). Hence, the theoretical model fits as well as the saturated model and is not rejected on the ground of the likelihood ratio test. In SEM, model fit is often judged also by other fit indices. Some of the recommended indices in the field of SEM are examined here. The root mean square error of approximation (RMSEA) is 0.0274, and the standardized root mean square residual (SRMSR) is 0.0403. Both of these fit indices show very good model fit. The CFI (Bentler comparative fit index) is 0.9958, which also indicates an excellent model fit.

Output 15 Fit Summary with the FIML Estimation for the miss3 Data Set

```
                                 Fit Summary

      Modeling Info        N Complete Observations                    100
                           N Incomplete Observations                  100
                           N Variables                                 12
                           N Moments                                   90
                           N Parameters                                39
                           N Active Constraints                         0
                           Saturated Model Estimation                FIML
                           Saturated Model Function Value         35.5034
                           Saturated Model -2 Log-Likelihood    7100.6750
                           Baseline Model Estimation            Converged
                           Baseline Model Function Value          44.8799
                           Baseline Model -2 Log-Likelihood     8975.9700
                           Baseline Model Chi-Square            1875.2950
                           Baseline Model Chi-Square DF                66
                           Pr > Baseline Model Chi-Square          <.0001
      Absolute Index       Fit Function                           35.7968
                           -2 Log-Likelihood                    7159.3515
                           Chi-Square                             58.6765
                           Chi-Square DF                               51
                           Pr > Chi-Square                         0.2147
                           Z-Test of Wilson & Hilferty             0.7909
                           Hoelter Critical N                         234
                           Root Mean Square Residual (RMSR)        0.1622
                           Standardized RMSR (SRMSR)               0.0403
                           Goodness of Fit Index (GFI)             0.9619
      Parsimony Index      Adjusted GFI (AGFI)                     0.9328
                           Parsimonious GFI                        0.7433
                           RMSEA Estimate                          0.0274
                           RMSEA Lower 90% Confidence Limit        0.0000
                           RMSEA Upper 90% Confidence Limit        0.0550
                           Probability of Close Fit                0.9019
                           Akaike Information Criterion         7237.3515
                           Bozdogan CAIC                        7404.9859
                           Schwarz Bayesian Criterion           7365.9859
                           McDonald Centrality                     0.9810
      Incremental Index    Bentler Comparative Fit Index           0.9958
                           Bentler-Bonett NFI                      0.2024
                           Bentler-Bonett Non-normed Index         0.9945
                           Bollen Normed Index Rho1                0.9595
                           Bollen Non-normed Index Delta2          0.9958
                           James et al. Parsimonious NFI           0.7485
```

Output 16 shows the estimates of the path effects and their significance. All estimates are significant at the 0.05 α-level because all their corresponding t values are greater than 1.96 (the critical value based on the z-distribution).

Output 16 Estimated Path Coefficients with the FIML Estimation for the miss3 Data Set

```
                                        PATH List

                                                                              Standard
----------------------Path---------------------- Parameter   Estimate      Error     t Value

S1                     <--- SocialStatus                      1.00000
S2                     <--- SocialStatus          _Parm01      0.97633    0.05143    18.98353
S3                     <--- SocialStatus          _Parm02      0.93340    0.05879    15.87763
P1                     <--- ParentalEncouragement             1.00000
P2                     <--- ParentalEncouragement _Parm03      1.02929    0.08125    12.66804
P3                     <--- ParentalEncouragement _Parm04      1.01757    0.08056    12.63135
A1                     <--- AchievementMotivation             1.00000
A2                     <--- AchievementMotivation _Parm05      1.06612    0.07252    14.70005
A3                     <--- AchievementMotivation _Parm06      1.04898    0.06795    15.43673
M1                     <--- MentalAbility                     1.00000
M2                     <--- MentalAbility         _Parm07      1.02426    0.06116    16.74674
M3                     <--- MentalAbility         _Parm08      1.04717    0.06530    16.03538
SocialStatus           ---> ParentalEncouragement _Parm09      0.70886    0.06736    10.52383
ParentalEncouragement ---> AchievementMotivation _Parm10      0.77686    0.07893     9.84259
AchievementMotivation ---> MentalAbility         _Parm11      0.86009    0.07966    10.79700
```

What results would you get if you use instead the regular ML method with listwise deletion of the incomplete observations? You can obtain the regular ML results by replacing the METHOD=FIML option with the METHOD=ML option in the preceding PROC CALIS specification. Output 17 shows some selected fit summary results with the METHOD=ML option.

Output 17 Fit Summary for the ML Estimation with Listwise Deletion for the miss3 Data Set

```
                          Fit Summary

           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
```

The chi-square model fit statistic is 70.0624 ($df = 51$, $p = 0.0394$). The theoretical model is rejected at the 0.05 α-level. The root mean square error of approximation (RMSEA) is 0.0614, and the standardized root mean square residual (SRMSR) is 0.0647. Both indices indicate only a marginally acceptable model fit. The CFI is 0.9831, which still shows an excellent model fit, but not as good as the analysis with the FIML estimation.

In summary, the moral here is not that you get better model fit with the FIML estimation, but that with the FIML estimation you can use the most information from the incomplete observations to make a more sound statistical decision. It just happens in this example that the FIML estimation is able to obtain more supporting information from the incomplete observations to substantiate the theoretical model. In addition, analyses of the data coverage and the missing patterns, which accompany with the FIML estimation of PROC CALIS, offer useful insights about the data.

PROC CALIS OPTIONS RELATED TO FIML ESTIMATION

This section describes in detail the options and controls related to the FIML method of PROC CALIS and is provided as a reference to make this paper self-contained. You might skip this section in the first reading.

Invoking the FIML Estimation

You can use either the METHOD=FIML or METHOD=LSFIML option in the PROC CALIS statement to invoke the FIML estimation. With METHOD=LSFIML, PROC CALIS first conducts the unweighted least square (ULS) estimation based on the complete observations only. The ULS parameter estimates are then used as the initial estimates for the subsequent FIML estimation. With METHOD=FIML, PROC CALIS conducts the FIML estimation only. Data coverages for the sample moments and the missing pattern analysis are displayed automatically with the specification of either the METHOD=FIML or METHOD=LSFIML option.

If your data set does not contain any missing values in the analysis model, the FIML estimation is essentially an ML analysis on the complete data. In this case, using the FIML method would have the same estimation results as using the ML method with the VARDEF=N option. The VARDEF=N option is required to make the ML method use N as the variance divisor (instead of the default $N-1$), which matches the variance divisor of the FIML estimation. Even without the VARDEF=N option, using the FIML method would have almost the same estimation results as using the ML method for analysis with complete data only. Because the FIML estimation requires much more computing resource than the regular ML estimation, it is recommended that you do not use the FIML method when the data set contains only complete observations for analysis.

Proportion Coverages of the Variables and Sample Moments

The proportion data coverage of a particular sample moment (mean or covariance) is the proportion of the observations that have nonmissing values for computing that sample moment. PROC CALIS displays these proportion coverages in the ODS output table called MeanCovCoverage, and displays the average coverages of the means and covariances in the ODS output table called AveCoverage. You can use these results to gauge the severity of missingness in the sample moments. A small coverage for a moment means that it suffers from the relative lack of information.

To locate the sample moments that suffer the highest rates of missing values, PROC CALIS ranks the smallest coverages for the sample moments in two output tables. The RankVariableCoverage table shows which variables or variable means have the smallest data coverages (the largest proportions of missing values). The RankCovCoverage table shows you which variable pairs or their corresponding covariances have the smallest joint data coverages (the largest proportions of missing values in both variables). These two tables highlight the sample moments that have the most serious missing value problems.

Table 3 summarizes the ODS tables concerning the proportion data coverages of the sample moments.

Table 3 ODS Tables for Displaying the Proportion Data Coverages

ODS Table Name	Description
AveCoverage	Average proportion coverages of means (variances) and covariances
MeanCovCoverage	Proportions of data present for means (variances) and covariances
RankCovCoverage	Rank order of the covariance coverages
RankVariableCoverage	Rank order of the proportion coverages of the variables

Analysis of Missing Patterns

Some missing patterns in the data set might occur more frequently than the others. The dominant missing patterns are those missing patterns that account for the highest proportions of the incomplete observations. PROC CALIS outputs two tables that enable you to examine these dominant missing patterns. The RankMissPatterns table displays the most dominant missing patterns in order. The MissPatternsMeans table displays the means for the nonmissing pattern and for the dominant missing patterns. By default, PROC CALIS displays up to 10 of the most dominant missing patterns. You can change this default behavior by using some options, which are described in the next section.

Table 4 summarizes the ODS tables for the analysis of the missing patterns.

Table 4 ODS Tables for Displaying the Missing Patterns

ODS Table Name	Description
MissPatternsMeans	Means of the nonmissing and the most dominant missing patterns
RankMissPatterns	Rank order of the most dominant missing patterns

Controlling the Number of Missing Patterns to Display

The number of missing patterns can be quite large in data sets with a lot of variables. To make the analysis output for the missing patterns concise and interpretable, PROC CALIS prints only the results for the most dominant missing patterns in the output (although *all* incomplete observations are still used in model estimation).

The most dominant (frequent) missing patterns are those patterns that have the largest proportions of incomplete observations. PROC CALIS displays the k most dominant missing patterns in the analysis output for the missing patterns. The value of k is between a minimum number (min) and a maximum number (max). The maximum number max is 10 by default. The minimum number min is the smallest among 5, the actual number of missing patterns, and max, which can be overridden by the MAXMISSPAT= option (by using any value between 1 and 9,999). After the minimum number of the most dominant missing patterns are included in the output, PROC CALIS continues to include the next most dominant missing patterns if each of these missing patterns can account for at least 5% of the total number of observations, until the maximum number of missing patterns to display (max) is reached. The 5% required proportion for a missing pattern to be displayed is referred to as the proportion threshold, which is set to 0.05 by default and can be overridden by the TMISSPAT= option.

To summarize, you can specify the following options to control the number of missing patterns in the analysis output:

MAXMISSPAT=*max*

specifies the maximum number of missing patterns to be displayed, where *max* is between 1 and 9,999. The default for *max* is 10 or the actual number of missing patterns in the data, whichever is smaller.

NOMISSPAT

suppresses the display of the missing pattern analysis. By default, the missing pattern analysis is displayed.

TMISSPAT=*min*

specifies the data proportion threshold for displaying the missing patterns, where *min* is between 0 and 1. The default for *min* is 0.05. This is the criterion for including additional missing patterns in the analysis output *after* the minimum number of the most dominant missing patterns have been included.

With a combination of these options, you can easily control the amount of missing patterns in the output. For example, if you do not want to analyze any missing patterns, use the NOMISSPAT option. If you want to include a very large number of missing patterns (up to 9,999), you can use a large number for the MAXMISSPAT= option (for example, MAXMISSPAT=9999) and a very small number for the TMISSPAT= option (for example, TMISSPAT=0). Finally, if you want to include up to 25 missing patterns such that each additional missing pattern accounts for at least 4% of the total number of observations (after the minimum number of patterns have been included), you can set TMISSPAT=0.04 and MAXMISSPAT=25.

SOME LIMITATIONS OF THE FIML METHOD

Following are some limitations of the FIML method:

- The FIML method cannot treat observations with nonrandom missing values—it is not designed to treat all kinds of missingness. In general, nonrandom missingness can be treated only within the formulation of the statistical models themselves. The FIML method is a principled statistical method only for estimating the model parameters; it is not a statistical model itself.
- The FIML method is computationally much more expensive than the regular ML method with the listwise deletion (LD) method for structural equation modeling. This is because the FIML method must evaluate all the individual likelihood functions in each iteration during the optimization, while the ML method evaluates only one likelihood function based on the sample moments in an iteration. If you have a very large data set with a lot of variables (say 100,000 observations and 200 variables), the computing resources might become so strained that the FIML method cannot be completed. If the listwise deletion still leaves you with a relatively large number of observations (say, several thousands), the regular ML method could be a better choice.
- If the proportion of incomplete observations for fitting a structural equation model is large (for example, above 85%) and the number of complete observations is small (say, under 15), the authors' experience is that the FIML method tends to have problems obtaining convergent solutions in optimization. However, the convergence problem might also happen to other incomplete data methods. Whether this limitation is more serious in the FIML estimation than in other methods needs to be studied more systematically.

CONCLUSION

With the availability of the FIML method in PROC CALIS, incomplete observations with random missing values (MCAR or MAR) can now be treated appropriately in fitting structural equation and related models. The main advantages of the FIML method are the consistency and efficiency of the maximum likelihood estimates under the multivariate normal distribution. Even if the multivariate normal assumption is violated, the FIML method is still much better than the ad hoc methods for treating incomplete observations. The FIML method in PROC CALIS is also more convenient to use than the multiple imputation method for treating incomplete observations, even though the two methods have similar statistical properties. Finally, the data coverage and missing pattern analyses of the CALIS procedure provide useful tools for gauging the seriousness of missingness and locating the dominant missing patterns, which in turn help researchers obtain more insights about their data.

REFERENCES

Arbuckle, J. L. (1996), *Full Information Estimation in the Presence of Incomplete Data*, chapter 9, 243–277, Mahwah, NJ: Lawrence Erlbaum Associates.

Bollen, K. A. (1989), *Structural Equations with Latent Variables*, New York: John Wiley & Sons.

Collins, L. M., Schafer, J. L., and Kam, C.-M. (2001), "A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures," *Psychological Methods*, 6(4), 330–351.

Enders, C. K. (2001), "The Impact of Nonnormality on Full Information Maximum-Likelihood Estimation for Structural Equation Models With Missing Data," *Psychological Methods*, 6(4), 352–370.

Enders, C. K. and Bandalos, D. L. (2001), "The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models," *Structural Equation Modeling*, 8, 430–457.

Gold, M. S., Bentler, P. M., and Kim, K. H. (2003), "A Comparison of Maximum-Likelihood and Asymptotically Distribution-Free Methods of Treating Incomplete Nonnormal Data," *Structural Equation Modeling*, 10(1), 47–49.

Graham, J. A. (2003), "Adding Missing-Data-Relevant Variables to FIML-Based Structural Equation Models," *Structural Equation Modeling*, 10(1), 80–100.

Jöreskog, K. G. and Sörbom, D. (1993), *PRELIS 2 User's Reference Guide*, Chicago: Scientific Software International.

Marjoribanks, K., ed. (1974), *Environments for Learning*, London: National Foundation for Educational Research Publications.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Schafer, J. L. and Olsen, M. K. (1998), "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst Perspective," *Multivariate Behavioral Research*, 33, 545–571.

Sinharay, S., Stern, H. S., and Russell, D. (2001), "The Use of Multiple Imputation for the Analysis of Missing Data," *Psychological Methods*, 6(4), 317–329.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors:

Yiu-Fai Yung and Wei Zhang
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
Phone: 919-531-4032 or 919-531-4025
Email: yiu-fai.yung@sas.com or Wei.Zhang@sas.com