Mathematics and Statistics 11(1): 134-143, 2023
DOI: 10.13189/ms.2023.110115 http://www.hrpub.org

On the Performance of Full Information Maximum Likelihood in SEM Missing Data

Amal HMIMOU[1,*], M'barek IAOUSSE[2], Soumaia HMIMOU[3], Hanaa HACHIMI[4], Youssfi EL KETTANI[1]

[1]Laboratory of Partial Differential Equations, Spectral Algebra and Geometry, Department of Mathematics, Ibn Tofail University, Kenitra, Morocco

[2]C3S Laboratory, Hassan II University of Casablanca, Morocco

[3]Department of Biology, Ibn Tofail University, Kenitra, Morocco

[4]Department of Mathematics, Sultan Moulay Slimane University, Beni Mellal, Morocco

Received August 30, 2022; Revised December 19, 2022; Accepted January 16, 2023

Cite This Paper in the following Citation Styles

***(a)**: [1] Amal HMIMOU, M'barek IAOUSSE, Soumaia HMIMOU, Hanaa HACHIMI, Youssfi EL KETTANI, "On the Performance of Full Information Maximum Likelihood In SEM Missing Data," Mathematics and Statistics, Vol.11, No.1, pp. 134-143, 2023. DOI: 10.13189/ms.2023.110115*

***(b)**: Amal HMIMOU, M'barek IAOUSSE, Soumaia HMIMOU, Hanaa HACHIMI, Youssfi EL KETTANI, (2023). On the Performance of Full Information Maximum Likelihood In SEM Missing Data. Mathematics and Statistics, 11(1), 134-143. DOI: 10.13189/ms.2023.110115*

Abstract Missing data is a real problem in all statistical modeling fields, particularly, in structural equation modeling which is a set of statistical techniques used to estimate models with latent concepts. In this research paper, an investigation of the techniques used to handle missing data in structural equation models is elaborated. To clarify this, a presentation of the mechanisms of missing data is made based on the probability distribution. This presentation recognizes three mechanisms: missing completely at random, missing at random, and missing not at random. Ignoring missing data in the statistical analysis may mislead the estimation and generates biased estimates. Many techniques are used to remedy this problem. In the present paper, we have presented three of them, namely, listwise deletion, pairwise deletion, and full information maximum likelihood. To investigate the power of each of these methods while using structural equation models a simulation study is launched. Furthermore, an examination of the correlation between the exogenous latent variables is done to extend the previous studies. We simulated a three latent variable structural model each with three observed variables. Three sample sizes (700, 1000, 1500) are examined accordingly to three missing rates for two specified mechanisms (2%, 10%, 15%). In addition, for each sample, hundred other samples were generated and investigated using the same case design. The criteria of examination are a parameter bias calculated for each case design. The results illustrate as theoretically expected the following: (1) the non-convergence of pairwise deletion, (2) a huge loss of information when using listwise deletion, and (3) a relative performance for the full information maximum likelihood compared to list-wise deletion when using the parameters bias as a criterion, particularly, for the correlation between the exogenous latent variables. This performance is revealed, chiefly, for larger sample sizes where the multivariate normal distribution occurs.

Keywords Structural Equation Modeling, MCAR, MAR, Missing Data, Parameter Bias

1 Introduction

Seen as one of the most redundant problems in statistics, missing data is the fact of having one or more missed values in a given data set.

The mechanisms of missing data theory were introduced in [1]. As a point of fact, they are relevant to the research that follows and are related to the missing data in data analysis when the rate of missing data in a given data set is higher or follows a specific mechanism. The fact of not considering it may mislead the analysis and generate biased estimates. In this sense, analyzing and handling the missing data is a preparatory step that should be inspected before starting any statistical modeling. Based on this, practitioners of SEM models should note the importance of handling missing data.

In this research paper, we will give a detailed presentation of the missing data theory inspired by the pioneer [1], in the first section 2. To explain [1] has linked the theory of probability

to the process of missing data. In this context, we will relate the different patterns recognized for missing data. Then, we will show the differences between a missing data pattern and a missing data mechanism. These mechanisms will be formalized mathematically using a probability distribution denoted R.

In the second section 3 a general summary of the techniques used to fix the missing data will be presented. The difference between the present study and the previous ones is that we added all the possible parameters (including the correlations between the exogenous latent variables) to assess the efficiency of each of the following methods for fixing missing data in SEM : (1) The listwise deletion (LD), the pairwise deletion (PD), the missing values replacements, and the full information maximum likelihood (FIML).

The last section suggests a simulated case study to test the performance of each of these techniques in the context of SEM models, the criterion selected reflects the percentage of bias generated while using each of the methods for the three latent variables structural model. In addition, a Monte Carlo simulation is used for the models to give more generalized findings. A general discussion to test the conformity of the findings with the previous studies existing in the literature is made.

A structural equation model is generally defined as follows: as a reminder, the notations and hypotheses used for this paper are the same as denoted in [2-6].

The latent variable model considered is the following :

$$\eta = B\eta + \Gamma\xi + \zeta \tag{1}$$

The dimension of the latent endogenous variables vector is $m \times l$;

For the vector of latent exogenous variables, it is $n \times l$;

B is the coefficient matrix with dimension $m \times m$; it reflects the correlation between the latent random endogenous variables;

Γ is the $m \times n$ coefficient matrix for the effects of ξ on η.

$(I - B)$ is a nonsingular matrix.

ζ is the disturbance vector; with $E(\zeta = 0)$ and which is uncorrelated with ξ.

The measurement model is stated as follows:

$$y = \Lambda_y\eta + \epsilon \tag{2}$$
$$x = \Lambda_x\xi + \delta \tag{3}$$

The y $(p \times l)$ and the x $(q \times l)$ vectors are observed variables,

Δ_x $(p \times m)$ and Δ_y $(q \times n)$ are the coefficient matrices reflecting the correlation of y to η and x to ξ respectively,

and ϵ $(p \times 1)$ and δ $(q \times 1)$ are the errors of measurement, respectively for y and x.

The errors of measurement are supposed to be uncorrelated with ξ and ζ and also with each other.

2 Missing Data in SEM

Missing data is the fact of having unobserved values in a set of data. This problem is a frequent complication of any real-world study. These unobserved values would be meaningful for analysis if observed indeed. That is, a missing value hides a meaningful value [7]. Rubin[1] establishes an early universal classification system for missing data problems. This classification has been entitled the mechanisms of missing data and it describes the relationships between measured variables and the probability of missing data. In other words, mechanisms of missing data are the assumptions for missing data analysis. It is necessary to recognize the difference between missing data patterns and mechanisms. They have different definitions but researchers sometimes use them in a commutable way which is wrong [7].

In what follows, we shall give the definitions of missing data patterns and missing data mechanisms to avoid any confusion between the mechanisms and the patterns of missing data.

2.1 Missingness patterns

A researcher should differentiate between the missing data patterns and the missing data mechanisms. We generally define a pattern of missing data as follows:

Definition 1. *Missing data pattern*

Missing data pattern refers to the configuration of observed and missing values in a data set. That is to say that the missing data patterns describe the location of the holes in the data.

As an evident example of a missing data pattern, we may assume that a relevant relationship occurs in a survey design between the educational level of the interviewed person and the plenty of missing data.

Enders[7] summarizes from the literature six prototypical missing data patterns.

- *The univariate pattern*
- *The unit nonresponse pattern*
- *The monotone missing data pattern*
- *The general missing data pattern*
- *The planned missing data pattern*
- *The latent variable pattern*

2.2 Mechanisms of Missing Data in SEM

Similar to missing data patterns, the missing data mechanism does not offer a causal explanation for the missing data. Rubin[1] is the pioneer of the missing data classification system, which is widely used in the literature today. Related to this classification we distinguish three mechanisms : (1) missing at random, (2) missing completely at random, and (3) missing not at random.

To illustrate these three mechanisms. We will be adopting notations in [7] for what follows. Despite that, if a data set contains missing values then the complete data must be a hypothetical entity and Y is an observed variable. For the need of analyzing the missing values, the variable Y will have the below structure in the data set.

- Y_{com}: The hypothetical complete scores.

- Y_{obs}: The observed scores in the data.
- Y_{mis}: The hypothetical scores that are missing.

We should also consider R as denoted in [7] as the indicator of missing values in a given data set.

And Φ: is a parameter or a set of parameters that describe the relationship between R and the data.

Definition 2. *Missing data mechanism :*

The missing data mechanism describes the possible relationships between measured variables and the probability of missing data. It gives a generic mathematical relationship between the data and missingness.

The distribution of missingness mechanism depends of the following variables: Y_{com}, Y_{obs}, Y_{mis}, R, and Φ.

In what follows, we define each mechanism as introduced by [1] and detailed by [7]. These definitions use formal expressions in terms of the probability distribution of each score Y and are harmonized with the mechanism definition.

Definition 3. *Missing at Random (MAR)*

When the probability of missing data on a variable Y is related to some other measured variable (or variables) in the analysis model but not to the values of Y itself. MAR means that there is a systematic relationship between one or more measured variables and the probability of missing data.

The distribution of missing data is expressed as : $p(R|Y_{obs}, \Phi)$

Definition 4. *Missing completely at Random (MCAR)*

It is purely haphazard missingness. Otherwise, the probability of missing data on a variable is unrelated to both other measured variables and the values of Y itself.

The distribution of missing data is expressed as $p(R|\Phi)$

Definition 5. *Missing not at Random (MNAR)*

When the probability of missing data on a variable Y is related to the values of Y itself, even after controlling for other variables.

A mathematical expression of the distribution of these missing data is given as : $p(R|Y_{obs}, Y_{mis}, \Phi)$

Formal expressions for these definitions can be given using the probability distribution of each with X designing variables in data for which the missing data may depend.

- For **MAR** :

$$P(Y_{mis}|Y_{com}, X) = P(Y_{mis}|X), \forall X \tag{4}$$

- For **MCAR** :

$$P(Y_{mis}|Y_{com}, X) = P(Y_{mis}|Y_{com}, X), \forall X \tag{5}$$

- For **MNAR** :

$$P(Y_{mis}|Y_{com}, X) = P(Y_{mis}), \forall X \tag{6}$$

The MNAR mechanism requires strict and unverifiable assumptions, for further details see[8]. For further information about missing data in SEM see [9] for MNAR and see [10] for MAR and MCAR.

3 Techniques for Handling Missing Data in SEM

To exceed this problem in SEM analysis, statistical methods have been actively pursued in recent years, as well as the imputation, likelihood, and weighting approach. Each approach depends on many variables such as the pattern of missing values and their mechanisms. In this section, we will present the four most used techniques for handling missing data in SEM. As a reminder, this is not an exhaustive list of all the techniques for missing data. The relevance of the next four treated missing values techniques is due to their performance in parameter estimations. To emphasize this, the method of mean imputation is not cited here since [11] has proven that it generates biased estimates with both mechanisms MCAR and MAR.

3.1 Listwise Deletion

Listwise deletion is about deleting the list of all the units that have at least one missed value. Then, the data used in the analysis contains only those cases that are complete on all observations.

Despite its benefits such as the easy implementation and comparability of univariate statistics, it generates a huge waste of information for the complete data. This may affect the results and brings out inefficient parameter estimates [12]. Likewise, this technique is available only for the missing data due to the MCAR mechanism. If not (i.e using the listwise deletion strategy for missing data due to MAR or MNAR mechanisms) the estimates would be biased.

In structural equation models, adopting the listwise is a two-step approach [13]: the researcher should, firstly, delete all the observations with one missing value or more and, secondly, calculate the new covariance matrix generated with the complete data. To put it differently, the second step corresponds to using a covariance matrix said S_L that contains all the remaining observations after deleting the ones with at least one missing value. This matrix is the one to be used for the analysis. Under those circumstances, the number of remaining cases (i.e the sample size) in S_L denoted N_L is less than the size of the complete sample said N. Regardless, estimations from the S_L are less efficient than they would be with the full sample S.

To demonstrate this, if the rate of missing values for each variable is less than 10% then the percentage of information loss using the listwise may attain $(p+q)10\%$ which is a huge loss. Moreover, the values of non-missing variables for the observations with one missed value dropped out and cannot be used [14].

At the same time, the estimator of θ generated with listwise; using the minimization functions, namely, F_{ML}, F_{ULS}, and F_{GLS} is recognized as the fitting functions used in the stage of estimation for a structural equation model; and is consistent whenever $N_L \to \infty$ and $N \to \infty$. It is important to realize that while $N_L > p + q$, the covariance matrix with listwise correction; S_L; is positive-definite, and provided Σ [15,14]. Another property of S that maintains for S_L is the multinormal distribution assumption. To clarify the replacement of S by S_L and N by N_L does not affect the use of usual test statistics.

3.2 Pairwise Deletion

Pairwise deletion is used in a way that all of the available data is utilized by eliminating cases on a variable by variable. This is doable by including in the covariance matrix the elements with calculated values and the bivariate pairs with missing values dropping out of the matrix. The generated matrix using pairwise deletion is denoted S_P. Consequently, S_P will be based on different sets of observations (i.e cases). The main advantage of pairwise deletion is the fact that it utilizes all of the available information in calculating the S_P, notably it does not discard the whole cases.

Like listwise deletion, this technique requires the MCAR assumption. The estimates under MAR are misleading and output biased estimates for the parameter. This fact was proved using empirical studies [9,13,16]. In contrast, pairwise deletion has its issues. In the first place, the S_P can not be positive-definite, likewise the minimization functions do not need to have a minimum of zero [17,14].

The abovementioned fact of the undetermined dimension of S_P generates a second issue: the options for the N_P which is the size of the used sample cannot be defined simply, since the elements included in S_P depend on different numbers of cases. Importantly, the researcher has the possibility of choosing the N_P. In view of using this parameter to estimate the chi-square tests and the asymptotic standard errors of each parameter estimate.

Reference[17] gives the approximate formula for the asymptotic covariance of the cases s_{ij} and s_{gh}; when handling missing data with pairwise deletion, and having in mind the assumption that the complete data variables are multinormally distributed is as follows :

$$ACOV(s_{ij}, s_{gh}) = \frac{N_{ijgh}}{N_{ij}N_{gh}}(\sigma_{ij}\sigma_{jh} + \sigma_{ih}\sigma_{jg}) \qquad (7)$$

Assuming :

N_{ijgh} is the size of observations with complete data on the variables i, j , g and h.

N_{gh} is the number of observations with complete data on the variables g and h.

The main difference between this asymptotic covariance (7) and usual asymptotic covariance is the fact that the coefficient $\frac{1}{N}$ for F_{ML} and F_{GLS} is replaced by $\frac{N_{ijgh}}{N_{ij}N_{gh}}$.

Equally important, the use of the results of chi-square and standard errors generated with the S_P may be inappropriate since the distribution assumptions for the minimization functions are violated.

3.3 Missing Values replacements

The missing values replacement is a technique used for handling missing data while analyzing a model. The basic principle of this technique is each missing case has a designated method to use to estimate a new value for replacing the missing one. In the context of SEM, those estimated values are directly used to calculate the covariance matrix of the model denoted S_M for the N available observations. Many methods are available for estimating the missed values such as the mean of the observed variable, a simple regression estimate of its value, or the similar pattern introduced by [18] or by some other numbers.

In this study, we didn't investigate mean imputation with SEM models, since other studies have explored this technique and demonstrated the biased estimates generated while using it [11,16].

The main advantage of these techniques is that the N observations will be used to calculate the S_M. Yet, some weak points are observed while using this process. To illustrate this fact, we shall use an example of a simple single variable:

$$x_{com} = \lambda\xi + \delta \qquad (8)$$

x_{com} represents the ideal situation where all the indicators are not missing. And we do have the two following assumptions:

1. $COV(\xi, \delta) = 0$;
2. $E(\delta) = 0$

In case we are having missing values in $x*$ then we should estimate it using the below model :

$$x = x_{com} + \epsilon \qquad (9)$$

Considering ϵ as the difference between the estimated value x and the missing value x_{com}.

An evidence is that when $\epsilon = 0$ then the x_{com} is nonmissing. We also assume that :

1. $COV(x_{com}, \epsilon) = 0$;
2. $E(\epsilon) = 0$

To show the effect of this imputation on the variance of the indicator x. Let's calculate the $VAR(x)$ when $x = x_{com}$ (i.e there is no missing value in the data) and the $VAR(x)$ when x satisfy equation (9).

- for the first case, we will have :

$$VAR(x) = VAR(\lambda\xi + \delta)$$
$$VAR(x) = \lambda^2\phi + VAR(\delta) \qquad (10)$$

- for the second case, we will have :

$$VAR(x) = VAR(\lambda\xi + \delta + \epsilon)$$
$$VAR(x) = \lambda^2\phi + VAR(\delta) + VAR(\epsilon) \qquad (11)$$

The variance $VAR(x)$ in the second case equation (11) is higher than the one in the first case equation (10). Ignoring this discrepancy will generate heteroscedasticity for the error in the equation for x since the error variance is greater for estimated values. Moreover, the distribution for x in (9) is unlikely to be normal [14].

Similar Response Pattern Imputation

Jöreskog[19] presented a response pattern imputation. In a few words, the technique aims to impute values of a variable X from another case with similar observed values (i.e similar in the sense of having the same values for the specified observed variables that define the similarity). In a statistical sense, it is about minimizing a criterion on a set of matching variables. If no matching observation exists that has complete data on the set of the abovementioned variables, then the imputation cannot unroll. Nonetheless, if a case satisfies the listed conditions, then the value corresponding to the variable missed X in case i will be assigned to the missed value.

Brown[11]'s study is a reference for all researchers that look for the performance of missing data techniques with SEM. Brown's results proved the efficiency of this method under MCAR even though a little bias was generated for the structural parameters.

3.4 Full Information Maximum Likelihood

The Maximum Likelihood (ML) approaches for treating missing data is a widely used technique. Owing to the required less restrictive assumption, unbiased estimates will result in the MCAR and MAR. Additionally, the ML technique output more efficient parameter estimates than the two mentioned techniques (i.e listwise deletion and pairwise deletion).

Assuming multivariate normality, the principle of full information maximum likelihood (FIML) is to calculate a case-wise likelihood of the observed data for the case i:

$$logL_i = K_i - \frac{1}{2}log|\Sigma_i| - \frac{1}{2}(x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i) \quad (12)$$

Taking into account that :

K_i: is a constant depending on the size of complete data for the case i;

x_i: is the observed data for the case i;

μ_i: includes the parameter estimates for the mean vector of the complete variables for the case i;

and Σ_i: is the covariance matrix containing parameter estimates for the complete variables for the case i;

For all the cases i the variables included in the sample are accumulated and the general function of maximization is given as follows :

$$logL(\mu, \Sigma) = \sum_{i=1}^{N} logL_i \quad (13)$$

As a matter of fact, with the full maximum likelihood estimation, all of the variables in the data are used in the parameter estimation. For the case denoted i it is obvious that it contributes to estimating all the model parameters for which exists complete data. This case inclusion also known as the inclusion of partially recorded cases is a basic stage in the theoretically demonstrated advantages of the likelihood inference.

In the MCAR mechanism, the fact of including partially recorded cases helps to ameliorate the efficiency of maximum likelihood parameter estimates when compared to listwise deletion. The random property of this mechanism does not allow the bias to be affected.

Few studies have already examined the performance of this technique (FIML) with the ad-hoc other techniques. Reference[13] for example, shows a CFA model and is investigated for this purpose. Enders[11] has also used two models to test this performance: the first model was a CFA and the second one was a structural model with latent variables equivalent to the CFA one. Both have proven the relative performance of FIML when compared to ad-hoc methods, which is compatible and consistent with the theoretical investigations. In what follows, we have decided to extend the study to different sample sizes and to focus on the impact of these missing data techniques in the estimations of different type of parameters (e.g path coefficients between latent variables, factor loadings and the correlation between the exogenous latent variables) using a simulated study and investigating the ad-hoc techniques and the full information maximum likelihood strategy.

4 Illustration : Simulated case study

To define the performance of each of the above-cited techniques of handling missing data in SEM, we have adopted a similar simulation study made by [9,11,13,12] with minor changes. The main purpose of this study is to use relevant criteria for investigating the performance of each technique (i.e: Listwise deletion (LD), pairwise deletion (PD), and full information maximum likelihood (FIML)), using Monte Carlo simulation which offers the possibility of generating from the same sample many different sub-samples. In addition, the software used is R programming language, particularly the `package LAVAAN`.

4.1 Study Methodology

In this section, we will bring the different variables that define a case design for our study into sharp, by exploring the data generation process, and the simulation design for the missing data as far as the purpose behind the selected criteria.

Data generation

We used a SEM model (see Figure 1) containing three latent variables (ξ_1, ξ_2 and η_1), each latent variable is associated with three observed variables. Following the mentioned studies, we set the loading to one of the following values: .4, .6, and .8.

However, in order to extend the study to explore the effect of those missing data techniques on the correlation between the exogenous latent variables (ϕ_{12} : the correlation between ξ_1 and ξ_2), we parameter it to have one of the previous values (i.e .4, .6, .8). In addition to that, three levels of sample size are investigated during the simulation of the study design (700, 1000, and 1500) and three levels of missing data (2%, 10%, and 15%). Indeed, for each sample size, the value of the parameters, and rate of missing values we generated 100 data of the 9 observed variables computed as a function of the three latent variables and a standard normal noise.

The latent exogenous variables are drowned for a multivariate standard normal distribution with the correlation set as one

of the values of the parameters above-mentioned. The `mvnorm` function from `MASS` package in R is used for this purpose. It generates two independent normal distributions and then uses Cholesky factorization to obtain two linearly correlated distributions.

It should be noted that the `rnorm` function is used to generate the random noise. In the case of the endogenous latent variable, they are computed as a function of the two exogenous variables.

Simulation Design

In terms of the missing data, we have simulated the MCAR mechanism by sampling uniformly (with equal probabilities and without replacement) from the vector starting from 1 to the rate of missing data times the sample size.

We estimated the model using a maximum likelihood estimator with a listwise deletion method, pairwise deletion method, and full information maximum likelihood method. However, for unexpected reasons, the pairwise deletion method has failed to converge in the model. To clarify this, whenever we use the pairwise deletion an error message was displayed reporting that the covariance matrix was not positive definite.

These findings are recognized in theory [14] because of the size of the covariance matrix generated by using the pairwise deletion. As cited in the definition of pairwise deletion this problem may appear to offend the estimation to be made. Hence, we are limited to the other two methods (LD and FIML).

The criterion designated to test the performance of each of the three techniques: *Parameter estimate bias*. This bias is computed as the following percentage:

$$\%bias = \frac{\widehat{\theta_{ij}} - \theta_i}{\theta_i} \tag{14}$$

where $\widehat{\theta_{ij}}, i = 1, \ldots, 12; j = 1, \ldots, 100$ is the estimation of the model parameter θ_i in j^{th} simulation. Then the mean across all the simulations is then computed for each parameter.

These parameters are as follows :

- The loading associated with the exogenous latent variables.
- The loadings associated with the endogenous latent variable.
- The structural coefficients (path coefficients).
- The correlation between the exogenous latent variables.

To test the overall bias we will calculate a criterion as the mean percentage of bias for the model parameters. Each one of the above-mentioned parameters is estimated for different sample size, value and the missing data technique used.

4.2 Findings and Results

In what follows, we have summarized the results generated from this study. For the listwise deletion techniques Table 1, shows the results of the calculated percentage of bias for the loadings associated with the exogenous latent variables, by sample size and estimated value for each missing data rate.

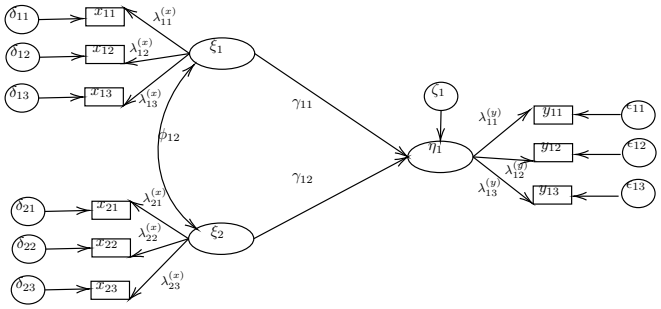

Figure 1. A SEM model with two exogenous latent variables and one endogenous latent variable each associated with three observed variables

Whatever the missing data rate and the magnitudes of loadings are, there is relatively a higher bias observed in samples of small size.

For table 2 reporting the percentage of bias for the loadings associated with the exogenous latent variables, the main result is shown independently from the sample size and the missing rate. The method has generated a negligible bias (ranging from 0% to 1%). In contrast, with the listwise deletion where the bias for some parameter has exceeded the 1.62%.

Otherwise, the parameter estimates generated for the loading of the exogenous latent variables are relatively better for the FIML than those estimated while using LD.

To illustrate the effect of the loadings associated with the endogenous latent variables, we have originally, produced the calculations for the bias for both techniques (FIML and LD). The table 3 shows those results for the listwise deletion method. Fewer percentages of bias are observed for samples with the smallest sizes. In fact, for the sample size of 700, In fact, for the sample size of 700, the larger the magnitude is the higher the bias is, too for the loadings associated with the latent variables. Moreover, the higher the rate of missing data is, the higher the percentage of bias generated is. Similar results were shown in table 4 with full information maximum likelihood technique.

For the correlation between the exogenous latent variables (ϕ_{12}) the bias generated is smaller for the FIML than the listwise deletion, particularly for large sample size.

A mean of the abovementioned percentage of bias was generated for both methods (LD and FIML) to illustrate clearly the effect of the rate of missing data and the magnitude in factor loadings are presented in table 5.

For samples with 700 as a size, independently from the magnitude of factors, the listwise deletion is a better technique if the missing data rate is equal to two percent. For samples with over 1000 and 1500, the FIML appears to bring better estimation when compared to the listwise deletion.

4.3 Discussion

The results generated with this study are not surprising since, we have theoretically, discussed the performance of the full information maximum likelihood technique for handling missing

Table 1. Percentage of bias for the loadings associated the exogenous latent variables, by sample size, value, and missing data using listwise method

sample_size	parameter	missing_data	x_11	x_12	x_13	x_21	x_22	x_23
700	0.4	2%	0.88	0.81	0.7	0.32	0.45	0.1
700	0.4	10%	-0.29	-0.35	-0.33	-0.15	0.03	0.1
700	0.4	15%	-0.85	-1.44	-1.39	0.13	-0.58	-0.19
700	0.6	2%	-0.07	0	-0.19	-0.07	-0.32	-0.16
700	0.6	10%	-0.26	-0.43	-0.38	0.05	-0.26	0.02
700	0.6	15%	-0.2	-0.38	-0.43	-0.15	-0.42	-0.09
700	0.8	2%	-0.09	0.28	0.02	0.07	0.01	0.15
700	0.8	10%	-1.13	-1.2	-0.93	-1.09	-0.85	-1.04
700	0.8	15%	-0.19	0.23	0.59	-0.36	0.41	0.38
1000	0.4	2%	-0.05	0.05	0.07	-0.16	0.04	-0.14
1000	0.4	10%	-0.32	-0.1	0.15	-0.19	-0.5	-0.32
1000	0.4	15%	0.61	1.03	1.11	0.18	-0.06	-0.12
1000	0.6	2%	0.23	0.32	0.33	-0.08	0.07	0
1000	0.6	10%	-0.2	-0.01	0.18	0.21	0.55	0.7
1000	0.6	15%	-1.62	-1.17	-1.15	-0.56	0.03	-0.5
1000	0.8	2%	-0.33	-0.42	-0.57	-0.81	-0.56	-0.6
1000	0.8	10%	-0.26	-0.34	-0.17	0.16	-0.06	-0.02
1000	0.8	15%	-0.94	-1.09	-0.92	-1.18	-1.06	-1.11
1500	0.4	2%	-0.14	-0.18	-0.19	0.03	-0.15	-0.1
1500	0.4	10%	-0.11	-0.16	0.05	-0.02	0.41	0.41
1500	0.4	15%	0.24	0.01	-0.17	-0.1	-0.51	-0.1
1500	0.6	2%	0.07	-0.13	-0.01	-0.07	-0.08	-0.07
1500	0.6	10%	-0.12	0.02	-0.12	0.54	0.46	0.19
1500	0.6	15%	0.11	0.25	0.25	0.31	0.49	0.3
1500	0.8	2%	0.22	0.34	0.07	0.34	0.3	0.21
1500	0.8	10%	-0.03	-0.42	-0.29	-0.14	-0.31	-0.49
1500	0.8	15%	-0.23	-0.67	-0.01	-0.58	-0.8	-0.53

Table 2. Percentage of bias for the loadings associated the exogenous latent variables, by sample Size, value, and missing data using FIML method

sample_size	parameter	missing_data	lambdax_11	x_12	x_13	x_21	x_22	x_23
700	0.4	2%	0.77	0.7	0.57	0.24	0.36	0.03
700	0.4	10%	0.16	0.02	0.03	-0.21	0.26	-0.14
700	0.4	15%	-0.21	-0.29	-0.05	0.25	-0.02	0.01
700	0.6	2%	-0.03	0.03	-0.04	0.03	-0.12	-0.15
700	0.6	10%	-0.13	-0.05	-0.05	-0.1	-0.09	0
700	0.6	15%	-0.23	-0.15	-0.04	-0.25	-0.11	-0.11
700	0.8	2%	-0.21	0.11	-0.08	0.08	-0.02	0.06
700	0.8	10%	-0.14	-0.33	-0.19	-0.22	-0.03	-0.27
700	0.8	15%	0.39	0.33	0.47	0.14	0.3	0.35
1000	0.4	2%	0.15	0.26	0.32	0.01	0.12	-0.01
1000	0.4	10%	-0.02	0.17	0.28	-0.45	-0.42	-0.46
1000	0.4	15%	-0.02	-0.04	0.17	-0.12	-0.09	-0.23
1000	0.6	2%	0.19	0.23	0.22	-0.23	-0.12	-0.13
1000	0.6	10%	0.35	0.31	0.37	0.02	0.16	0.26
1000	0.6	15%	-0.21	0.03	0.02	-0.03	0.32	-0.15
1000	0.8	2%	0	-0.21	-0.39	-0.58	-0.29	-0.31
1000	0.8	10%	-0.18	-0.11	-0.35	-0.14	-0.17	0.05
1000	0.8	15%	-0.12	-0.1	-0.29	-0.26	-0.18	-0.34
1500	0.4	2%	-0.17	-0.29	-0.31	0.06	-0.07	-0.04
1500	0.4	10%	-0.09	-0.15	-0.1	-0.02	0.34	0.29
1500	0.4	15%	0.07	-0.05	-0.23	0.08	-0.15	0.02
1500	0.6	2%	-0.02	-0.24	-0.06	-0.06	-0.03	-0.04
1500	0.6	10%	-0.11	-0.09	0.07	0.14	0.04	-0.19
1500	0.6	15%	0.08	0.11	0.1	0.09	0.17	0.2
1500	0.8	2%	0.16	0.33	0.03	0.4	0.29	0.28
1500	0.8	10%	0.25	0.06	0.09	0.15	0.11	0.1
1500	0.8	15%	-0.12	-0.32	-0.17	-0.22	-0.21	-0.21

Table 3. Percentage of bias for the loadings associated the endogenous latent variable and the structural coefficients, by sample Size, value, and missing data using listwise method

sample_size	parameter	missing_data	y_11	y_12	y_13	gamma_11	gamma_12	phi_12
700	0.4	2%	-40.49	-40.39	-40.37	103.68	99.24	1.22
700	0.4	10%	-40.97	-40.84	-40.66	99.48	105.55	2.92
700	0.4	15%	-41.32	-41.19	-41.18	105.88	102.35	-1.24
700	0.6	2%	-57.71	-57.61	-57.63	100.03	98.44	0.02
700	0.6	10%	-57.65	-57.66	-57.7	98.57	100.34	-0.2
700	0.6	15%	-58.92	-58.84	-58.89	103.59	107.9	0.66
700	0.8	2%	-69.14	-69.15	-69.19	104.73	102.67	0.14
700	0.8	10%	-68.8	-68.81	-68.74	99.06	101.6	-0.37
700	0.8	15%	-69.24	-69.31	-69.32	103.43	106.31	0.46
1000	0.4	2%	-40.42	-40.34	-40.35	100.93	100.54	0.2
1000	0.4	10%	-39.72	-40.07	-39.94	96.68	99.5	0.05
1000	0.4	15%	-40.6	-40.31	-40.44	102.56	102.64	-1.18
1000	0.6	2%	-57.91	-57.84	-57.85	100.47	102.31	-0.23
1000	0.6	10%	-58.41	-58.58	-58.38	104.09	103.4	0.53
1000	0.6	15%	-58.08	-58.18	-58.2	101.13	102.89	-0.78
1000	0.8	2%	-68.67	-68.63	-68.72	99.74	98.94	-0.31
1000	0.8	10%	-68.81	-68.86	-68.89	99.27	104.43	-0.05
1000	0.8	15%	-69.05	-69.05	-69.09	102.15	102.11	-0.87
1500	0.4	2%	-40.11	-40.11	-40.02	99.8	99.53	-0.04
1500	0.4	10%	-40.33	-40.24	-40.14	100.99	99.9	-0.3
1500	0.4	15%	-40.66	-40.6	-40.52	102.23	102.01	1.45
1500	0.6	2%	-57.81	-57.85	-57.86	101.29	100.19	-0.12
1500	0.6	10%	-58.25	-58.27	-58.33	102.82	103.75	0.78
1500	0.6	15%	-58.01	-58.09	-58.04	100.98	101.81	0.61
1500	0.8	2%	-68.66	-68.74	-68.65	99.03	100.55	0.28
1500	0.8	10%	-68.68	-68.76	-68.77	99.7	98.98	0.15
1500	0.8	15%	-68.78	-68.69	-68.78	103.4	97.45	0.16

Table 4. Percentage of bias for the loadings associated the endogenous latent variable and the structural coefficients, by sample Size, value, and missing data using FIML method

sample_size	parameter	missing_data	y_11	y_12	y_13	gamma_11	gamma_12	phi_12
700	0.4	2%	-40.54	-40.34	-40.36	103.86	98.69	1.04
700	0.4	10%	-40.47	-40.35	-40.25	98.68	101.92	1.05
700	0.4	15%	-40.26	-40.19	-40.21	100.75	99.97	-0.67
700	0.6	2%	-57.79	-57.7	-57.68	100.43	98.79	0.15
700	0.6	10%	-57.69	-57.72	-57.74	99.58	99.48	0.33
700	0.6	15%	-58.07	-58.05	-58.07	100.81	103.07	0.1
700	0.8	2%	-69.08	-69.07	-69.08	104.22	101.32	0.04
700	0.8	10%	-68.78	-68.73	-68.73	98.67	102.4	-0.18
700	0.8	15%	-69	-69.09	-69.06	101.76	103.03	0.34
1000	0.4	2%	-40.46	-40.36	-40.35	101.72	100.1	0.4
1000	0.4	10%	-39.79	-39.99	-40.01	99.53	97.97	-0.02
1000	0.4	15%	-39.96	-39.87	-39.96	99.08	98.73	-0.29
1000	0.6	2%	-57.94	-57.86	-57.91	100.37	101.89	-0.45
1000	0.6	10%	-58.15	-58.21	-58.1	103.72	100.53	0.73
1000	0.6	15%	-57.88	-57.93	-57.94	101.43	100.65	-0.18
1000	0.8	2%	-68.69	-68.66	-68.74	100.38	99.64	-0.16
1000	0.8	10%	-68.75	-68.79	-68.79	99.27	101.94	-0.13
1000	0.8	15%	-68.79	-68.78	-68.77	100.9	100.96	-0.24
1500	0.4	2%	-40.02	-40.03	-39.98	99.53	98.99	-0.14
1500	0.4	10%	-40.17	-40.19	-40.15	100.59	99.83	-0.31
1500	0.4	15%	-40.34	-40.45	-40.33	101.46	100.83	0.13
1500	0.6	2%	-57.76	-57.77	-57.79	100.34	99.98	-0.09
1500	0.6	10%	-58.02	-57.99	-58	100.52	102.19	0.44
1500	0.6	15%	-57.8	-57.79	-57.79	99.15	100.75	0.07
1500	0.8	2%	-68.7	-68.79	-68.69	99.45	100.96	0.21
1500	0.8	10%	-68.59	-68.68	-68.67	99.46	99.62	0.14
1500	0.8	15%	-68.78	-68.76	-68.81	102.22	98.97	-0.07

Table 5. Mean percentage of bias for the model parameters by sample Size, value, and missing data using LD and FIML method,

sample_size	parameter	missing_data	mean_listwise	mean_fiml
700	0.4	2%	7.18	7.09
700	0.4	10%	7.04	6.72
700	0.4	15%	6.58	6.59
700	0.6	2%	2.06	2.16
700	0.6	10%	2.04	2.15
700	0.6	15%	2.82	2.41
700	0.8	2%	0.04	-0.14
700	0.8	10%	-1.02	-0.54
700	0.8	15%	0.28	0
1000	0.4	2%	6.7	6.83
1000	0.4	10%	6.27	6.4
1000	0.4	15%	7.12	6.45
1000	0.6	2%	2.49	2.36
1000	0.6	10%	2.84	2.67
1000	0.6	15%	1.98	2.34
1000	0.8	2%	-0.91	-0.67
1000	0.8	10%	-0.3	-0.51
1000	0.8	15%	-0.84	-0.5
1500	0.4	2%	6.53	6.46
1500	0.4	10%	6.71	6.66
1500	0.4	15%	6.94	6.75
1500	0.6	2%	2.3	2.2
1500	0.6	10%	2.79	2.42
1500	0.6	15%	2.58	2.28
1500	0.8	2%	-0.39	-0.34
1500	0.8	10%	-0.75	-0.5
1500	0.8	15%	-0.67	-0.54

data in SEM, indeed, this study allows us to put in sharper the points that offer this preferability to FIML when compared to other methods. The pairwise deletion has failed to generate estimation in R programming using `LAVAAN PACKAGE`. The error message provided claims that the covariance matrix generated by utilizing the pairwise deletion to handle data having missing data is non-positive definite. This finding is theoretically true. Since as discussed in [14] the pairwise deletion has such an issue.

The listwise deletion and full information maximum likelihood are then included in the simulation design to investigate their performance by settling the criteria of percentage of bias generated from their estimations.

Likewise, analysing the results of each method shows that the higher the missing data rate is, the higher the percentage of bias is. For samples with size 700, the listwise has a small bias chiefly for the loading exogenous factors.

Similar results were shown for the FIML.

Those results are explicitly explained by the fact that for huge sizes using the listwise deletion is equivalent to the loss of huge information. In this sense, more bias should be recognized.

In contrast, the full information maximum likelihood with a huge size of samples seems to generate less bias relative to listwise deletion which is theoretically admissible since the huge sample sizes (over 1500) guarantee the maximum likelihood estimation that the assumptions about the multinormal distribution are true. Another obvious finding is that, independently from the method used, the higher the rate of missing data is the higher the bias generated for each parameter is too.

The most relevant finding in our study is originally taking into account the correlation between the exogenous latent variables and investigating the effect of each missing data technique in a SEM model. The results shows that the full information maximum likelihood performed better in handling missing data compared to the others ad-hoc techniques. This result joined the previous studies and extended their findings that were, chiefly, concerned with the other types of parameters. Moreover, another relevant finding for our study is the fact of examining the structural coefficients. The results were aligned with the other parameters' findings: The relative performance of the FIML is revealed.

5 Conclusion

The present work explores the missing data in SEM models. A conceptual section was reserved for the theory of missing data mechanisms, each mechanism was properly presented and defined and the probabilities were written as presented by the pioneer Rubin[1]. In addition for the second section 3, the main techniques used for structural equation models were presented and the properties and limits of the generated estimators were shown also. (i.e the listwise deletion, the pairwise deletion, the missing values replacements, and the full information maximum likelihood).

This research paper investigates the power of each of the aforementioned techniques in treating the missing data on SEM by calculating a defined criterion named percentage bias for three types of parameters: (a)loadings associated with the exogenous latent variables and loadings between the endogenous latent variables, (b) the structural coefficients, and (c) the correlation between exogenous latent variables) and defined as the relative difference between the estimated parameter, using LD or FIML and for a specified rate of missing values, and the true value of the parameter.

For this reason, a simulation study was launched by generating data sets for a specified model with three latent variables as shown in figure 1. For every sample, a Monte Carlo simulation was adopted for each model. Note that three models with different sample sizes were simulated with sizes ranging from 700, 1000, and 1500. Then we have to proceed with the deletion of some values from the data with different rates of missing values (2%, 10%, and 15%). The main purpose was to estimate the model parameter for each design and to compare the bias criteria as it has been defined by comparing the estimated parameters of each case design with the true values (set on the simulation design).

The results for all designs have demonstrated a relative performance for the strategy of FIML for those models, particularly, with the sample size of 1500 where the bias percentages were relatively less than the ones observed for the same design in LD. For the sample size of 700, results have shown that the listwise deletion generates less bias. Those facts are compatible with the theoretical facts assessing that the maximum likelihood estimation method is based on assumptions of normal distribution for the variables, this is achievable with samples of larger size (over 1500). In contrast, for listwise

deletion and larger size, the loss of information generated by deleting a whole observation just because one case is missed is recognized to be a larger loss. In other words, larger bias.

Similar findings were given in some previous studies [9,11,13,12]. The main contribution of the presented study is making an extension to those studies by focusing on additional types of parameters which are the correlations between the exogenous latent variables and the structural coefficients. In addition to that, we use higher sample sizes (more than 700) and three thresholds for the rate of missing data (2, 10, and 15 percent).

In other words, the limitation of the previous research is that they didn't take into account the path coefficient and the correlation between the latent variables. Hence, to make sure that these parameters do not affect the power of FIML, we took them into account and the results were as theoretically expected.

The main advantages of our study are :

- It discussed a three latent variable structural model and the use of simulations. Moreover, we have discussed the theoretical properties of each estimator generated with the given methods.
- It extended the previous studies to include all possible parameters (including the path coefficients and the unanalyzed association between the exogenous latent variables).
- It explores different samples sizes and missing data rates.

In all of the above results, the FIML seems to be better than the other methods included in the study for handling missing data in structural equation models.

We also introduced the R programming for the simulation study and used the package `LAVAAN` which needs further development to cover all aspects of SEM. An R program is written to the seen methods and to produce the numerical results reported in the manuscript. The program is available upon request.

Acknowledgements

The authors gratefully acknowledge the support of editors during the publication process as well as the helpful feedback provided by the anonymous referees.

REFERENCES

[1] Rubin, D. B. "Inference and missing data". Biometrika, vol.63, no.3, pp. 581-592, 1976.

[2] K. G. Jöreskog. "Structural equation models in the social sciences: Specification, estimation and testing". In Applications of Statistics, R. Krishnaiah (Ed). Amsterdam: North-Holland, 1977, pp. 265–287.

[3] K. G. Jöreskog. "Structural analysis of covariance and correlation matrices". Psychometrika, vol. 43, no. 4, pp. 443–477, 1978. doi: 10.1007/BF02293808.

[4] Joreskog, K.G. and H. Wold "Systems under indirect observation: Causality, structure, prediction, Part I and Part II". Amsterdam: North Holland, 1982.

[5] Iaousse, M'barek et al. "A Modified Algorithm for the Computation of the Covariance Matrix Implied by a Structural Recursive Model with Latent Variables Using the Finite Iterative Method". Statistics, Optimization and; Information Computing, vol.8, no.2, pp. 359–373, 2020.

[6] Hmimou, Amal et al. "Treatment of Correlated Errors in Structural Equation Models". In: 7th International Conference on Optimization and Applications (ICOA), 2021, pp. 1–6. doi: 10.1109/ICOA51614.2021.9442624.

[7] Enders, Craig K. "Introduction to Missing data" in Applied Missing Data Analysis. Guilford Press. 2010. pp. 1-36.

[8] Enders, Craig K ."Missing Not at Random Processes" , in Applied missing data analysis. Second Edition. Guilford Publications, 2022. pp. 348-400.

[9] Bengt Muthén, David Kaplan and Michael Hollis. "On structural equation modeling with data that are not missing completely at random". Psychome- trika, vol. 52, no. 3, 431–462, 1987.

[10] Gold, Michael Steven and Peter M Bentler. "Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization". Structural Equation Modeling, vol. 7, no. 3, pp. 319–355, 2000.

[11] Brown, Roger L. "Efficacy of the indirect approach for estimating structural equation models with missing data: A comparison of five methods". Structural Equation Modeling: A Multidisciplinary Journal, vol.1, no.4, pp. 287–316, 1994.

[12] Craig K Enders and Deborah L Bandalos. "The relative performance of full information maximum likelihood estimation for missing data in structural equation models". Structural equation modeling, vol. 8, no. 3, pp. 430–457, 2001.

[13] Arbuckle, James L, George A Marcoulides, and Randall E Schumacker ."Full information estimation in the presence of incomplete data". in Advanced structural equation modeling: Issues and techniques, 1996,p. 243-277.

[14] Bollen, Kenneth A. "The General model, Part I: Latent variable and measurement models combined", in Structural equations with latent variables. John Wiley & Sons, 1989, pp. 319-394.

[15] Dijkstra, Taeke Klaas. “Latent variables in linear stochastic models: Reflections on “maximum likelihood”and “partial least squares”methods”. In: Groningen University, Groningen, a second edition was published in, 1985.

[16] Wothke, Werner. “Longitudinal and multigroup modeling with missing data”. In: Modeling longitudinal and multilevel data, 1st Edition, Psychology Press, 2000, pp. 205–224.

[17] Browne, Michael W. “Covariance structures”. In: Topics in applied multivariate analysis, 1982, pp. 72–141.

[18] Joreskog, K. G. and D. Sorbom. ¨ LISREL 8: Structural equation modeling with the SIMPLIS command language. Scientific Software international, 1993.

[19] Joreskog, K. G. and D. Sorbom. PRELIS 2: User Reference Guide, Scientific Software International, 1996.