MISSING DATA IN LINEAR MODELING

Verleye G.[*]

Abstract

In this paper, the results of a performance study that compares five missing data solutions in the context of linear modeling are presented. By means of a five factor simulation approach with both numerical and graphical evaluations, six research hypotheses are tested. A new and easy applicable method to handle multiple imputed data sets is also presented.

Key words: SEM, multiple imputation, performance study

Acknowledgements:

We thank A. Boomsma , J. Arbuckle, K. Bollen, D. Rubin, L. Chuanhai and P. Allison for the useful comments and assistance. This research was supported by Vrije Universiteit Brussel OZR grants.

1. INTRODUCTION

After decades of interest in estimation methods, new models, performance & robustness analyses and topics such as meta analysis, missing data issues boomed during the nineties. Indeed, data quality and especially absence of data may seriously affect the results from any statistical analysis based on the data at hand.

[*] *State University Gent, Dept. Social & Political Sciences, Universiteitsstraat 8, 9000 Gent Belgium, email: Gino.Verleye@rug.ac.be*

In statistics, the linear model is definitely a reference model. In this paper we will compare several techniques to estimate linear models in the presence of missing data. The model used in this paper is a generalization of the multivariate linear model enabling the presence of latent variables : the structural equation model (SEM). Over the last 20 years, SEM became very popular in social sciences because the concepts used in for instance psychology and sociology can often be represented as latent SEM variables where the observed data (e.g. test results and questionnaire items) are the manifest SEM variables. Such social sciences data often contain missing data. The information may be absent for many reasons: refusal to answer particular questionnaire items, lost data, non understood questions, attrition. In more recent survey technology, the data collection schedule can be designed so that several parts of the data are missing by design.

The actual question this paper deals with is how to handle the missing at random data in SEM applications. By means of a performance study relying on Monte Carlo simulation technology, we take a closer look at five solutions for the missing data problem in SEM. Two of them are simple, frequently used methods, while the remaining three techniques use ML estimation.

In the next section, we briefly summarize existing literature. In section three, we look at five particular techniques for handling missing data in SEM and explain our performance analysis approach. Hypotheses and results are presented in section four. Section five includes the discussion and suggestions for future research.

2. RESULTS FROM PREVIOUS PERFORMANCE STUDIES OF MISSING DATA TREATMENT FOR LINEAR MODELS

After analysis of previous research (Glasser (1964), Afifi and Elashoff (1966, 1967), Haitovsky (1968), Timm (1970), Beale and Little (1975), Gleason and Staelin (1975), Kim and Curry (1977), Finkbeiner (1979), Brown (1983), Malhotra (1987), Brown (1994) and Arbuckle (1995a, 1995b)) a number of conclusions can be formulated. Because the different studies use a mixture of several evaluative criteria combined with different designs, care should

be taken if the results are compared. Some findings however are present in multiple studies and can therefore be generalised as conclusions so far:

- Complete cases analysis (listwise deletion) remains a valid option when there are few MCAR missing data. This method is very wasteful as from moderate levels of missing data on.
- The available information method, also known as pairwise deletion is generally applicable for MCAR missing data problems. Research indicates that this method can be used if the correlations between the variables are rather low. In this case regression methods and ML techniques cannot outperform the pairwise method because of the lower redundancy in the data. The absence of a common sample size is a drawback.
- Assigning means to missing values is poor in comparison with pairwise and listwise deletion, regression methods, principal components solutions and ML techniques.
- A most interesting feature of recent ML methods is that they work under the weaker MAR condition. ML approaches are the efficient under the broad range of design factors. Their overall superiority holds even in the small sample case.
- In general, imputation methods such as mean substitution and hot-deck imputation do not yield efficient estimates. However, hot-deck imputation is in any case better than mean substitution. Multiple imputation seems to perform much better.
- Direct estimation of the model parameters, in contrast to indirect procedures (imputation, alternative covariance matrices that serve as input for the modeling), seems to be a valid option.

3. METHODS AND RESEARCH DESIGN

In the next section five different approaches are presented that will be used to handle MAR/MCAR data.

3.1 The complete cases method

This method, often called "listwise deletion", uses the N_L cases where all K variables are observed. Under the MCAR assumption, the complete cases are a random sub-sample of the

original cases and discarding cases does not bias estimates. If the MCAR condition is satisfied, this approach has many advantages. Standard complete data analysis methods can be applied without modifications (Little and Rubin, 1987, p.40). Univariate statistics can be compared since all such parameters are computed on the same number of cases (Little and Rubin, 1987, p.40). The loss of cases and information can be severe.

3.2 The available information method

Next to the complete cases method, a second procedure called "available information", or "pairwise deletion" is often used. In the case of pairwise calculation of the covariance matrix, the measures of the covariance for X and Y are based on the N_P cases where N_P stands for the number of cases for which we have values for X and Y at the same time. Although alternative computational versions exist (see Little and Rubin, 1987), the covariance estimate is obtained as:

$$S_{XY}^{N_P} = \frac{1}{N_P} \sum_{i=1}^{N_P} (x_i - \bar{x}^{N_P})(y_i - \bar{y}^{N_P}).$$

If the missing data process is MCAR, then this estimate is consistent Although this approach uses all information available, its practical utility is limited for at least two reasons. First of all, if the pairwise deletion method is used to estimate a covariance matrix, one cannot provide the sample size for the entire matrix since this sample size can be different for each pair of variables. In SEM, the chi-square tests of model fit and the estimated asymptotic standard errors are sensitive to the choice of N (see Bollen, 1987). A second problem is that if the number of variables K increases, the resulting covariance matrix S could not be positive-definite.

3.3 The Expectation Maximization (EM) covariance matrix estimator under the normal model

This approach is an indirect way to deal with missing data problems because it results in an alternative (but efficient) estimate of the covariance matrix that serves as input to a SEM program. The algorithm handles missing data in an iterative way. Each iteration consists of two steps. If ω^t is the estimate of ω at time t, the E step (expectation step) finds the conditional expectation of functions of the missing data appearing in the complete-data loglikelihood, given the observed data and the current value of the estimated parameter ω^t. Once the current estimates of the functions of the missing data are substituted, the M step (maximization step) finds the ML estimate of ω as if there were no missings. This goes on until the process converges. One of the advantages of EM is the fact that under general conditions, the loglikelihood is increased in each iteration. In other words, it converges reliably. On the other hand, the convergence of EM can be slow if a lot of data are missing. A method that yields good starting values is using the available information for the univariate parameters and setting the covariances to zero (see Little and Rubin, 1987, p.143). The FORTRAN 77 routine that computes ML estimates (under the normal model) of the covariances with EM, is made available for this study by Donald B. Rubin and written by Chuanhai Liu.

3.4 Full information estimation in SEM

Arbuckle (1995a, 1995b) generalised the ML estimation in confirmatory factor analysis with incomplete data by Finkbeiner (1979).

Let μ_i and Σ_i be the population mean and covariance matrix for the variables that are observed for case i. Each μ_i can be obtained by deleting elements of μ, and each Σ_i can be obtained by deleting rows and columns of Σ.

If we further assume multivariate normality, the loglikelihood of the i -th case is

$$\ln L_i = A_i - \tfrac{1}{2}\ln|\Sigma_i| - \tfrac{1}{2}(x_i - \mu_i)\Sigma_i^{-1}(x_i - \mu_i)'$$

where A_i is a constant that depends only on K_i. The loglikelihood of the total sample is then

$$\ln L(\mu, \Sigma) = \sum_{i=1}^{N} \ln L_i .$$

Given a structural equation model that specifies $\mu = \mu(\omega)$ and $\Sigma = \Sigma(\omega)$ as a function of some parameter vector ω, ML estimates of ω are obtained by maximizing

$$\ln L(\mu(\omega), \Sigma(\omega)),$$

or by minimizing

$$C(\omega) = -2\ln L(\mu(\omega), \Sigma(\omega)) + 2\sum_{i=1}^{N} A_i = \sum_{i=1}^{N} \ln|\Sigma_i| + \sum_{i=1}^{N} (x_i - \mu_i)\Sigma_i^{-1}(x_i - \mu_i)' .$$

The structural equation programs AMOS (Arbuckle, 1993) and Mx (Neale, 1994) use this approach to handle missing values. The full information method is a direct approach because the model parameters are estimated in the presence of missing data. More details and examples can be found in Finkbeiner (1979) and Arbuckle (1995a, 1995b).

3.5 The multiple imputation approach

Imputation methods replace each missing value by a value making the data matrix free of missing data. In multiple imputation we replace each missing value by M values. With multiple imputation the uncertainty about the real values of the missing data are considered by using M values instead of one. In practice we can use $M = 3$ (Rubin, 1987). After multiple imputing a data set containing missing data, we have M complete data sets. In this study an advanced version of data augmentation (Tanner and Wong, 1987) programmed by Liu Chuanhai (1992a, 1992b) was used. According to Rubin, the multiply imputed data should be analyzed separately and the results combined. In this case however, the multiple imputed data

sets are simultaneously analysed with SEM programming using the so-called multiple group approach. Each imputed data set is then assigned to a group and the parameters are estimated with equality constraints over the three groups (see Jöreskog and Sörbom, 1989). The advantage of this particular approach is the easy and correct handling of the multiple imputed data sets.

3.6 Study Design

This study can be considered a factorial design performance study with five factors. The first factor is the percentage of missing data for each variable. This factor has three levels : 5%, 15% and 25%. Factor two is the inclusion of two SEM models : a measurement model versus a full model. The measurement model is the 4CM model used by Boomsma (1983) in his robustness study. It is a two correlated factors model with four indicators for each latent variable and medium sized factor loadings. The full SEM model used in this study is the Peer influence on aspiration model by Duncan, Haller and Portes (1968). Design factor three is the missing data process with two levels : MCAR and MAR. In the MCAR case, the data values are randomly erased while in the MAR case, we applied a procedure described by Rubin (1976, p.583). A simple MAR process univariate example: if the sum of the first Nc values from a random starting point exceeds some predefined value, then all values that come after Nc are made missing. Since we worked with both multivariate normal distributed data sets and non normal sets, this is factor four. In the non-normal data sets, each variable is χ^2_3 distributed. The five missing data solutions already mentioned are factor five. In each cell of the design 300 data sets with a fixed covariance matrix that goes with the model of interest and distribution type are generated with the SIMCHI procedure (Verleye, 1996). Each data file contains 1000 cases to avoid small sample issues in SEM.

To evaluate the performance of the five techniques two sets of criteria are applied:
1. numerical indicators of (1) non-convergence and improper solutions, (2) bias of parameter estimates, (3) bias of estimates for standard errors, (4) confidence intervals for parameter estimates, (5) confidence intervals for the mean of standardized parameter estimates, (6) the

chi-square statistic for goodness of fit, (7) dependencies between parameter estimates and their corresponding standard error, (8) normality tests for the standardized parameter estimates.
2. graphical analysis of the standardized parameter estimates and the goodness of fit statistic: compare the theoretical sampling distribution with the empirical sampling distribution.

4. RESEARCH HYPOTHESES AND RESULTS.

In this section a number of hypotheses are presented. These statements are motivated by means of the special features that characterize the five different methods tested in this study. More detailed results and tables can be found in Verleye (1996).

1. Given the redundancy (due to the non-zero correlations) in the two covariance matrices that are used in this analysis, *the EM maximum likelihood solution and the multiple imputation procedure should be more efficient compared to listwise and pairwise deletion*. One can verify that pairwise and listwise deletion do not use the redundancy in the data while the two ML methods estimate parameters using the information in the non-zero correlations between the variables. From our analysis of the results it is clear that this hypothesis is confirmed. The two ML methods always yield convergence and there are no improper solutions. The performance of the listwise deletion method is clearly worse. However, the pairwise deletion method leads neither to such convergence nor to improper solution problems. The bias of the parameter estimates after treating the missing data problem with an indirect ML method is smaller than the bias obtained with the two quick methods. Pairwise deletion shows better results (smaller bias for the parameter estimates) than the listwise deletion method. However, the pairwise method leads too frequently to rejection of a correct model. This could be due to the fact that is the case of pairwise deletion N was fixed at 1000.

2. *The higher efficiency of ML solutions should be more pronounced in the MAR case than in the MCAR case*. ML methods should still be efficient under the weaker MAR condition, while listwise and pairwise deletion require MCAR data. From the analysis it is clear that ML methods work equally well under both MCAR and MAR conditions. This is not the case for the listwise method: this procedure performs worse under MAR conditions than under MCAR

conditions. In terms of bias for the parameter estimates, the pairwise deletion method yields results that are similar for MCAR and MAR. The only performance indicator for the pairwise deletion method that is influenced by the nature of the missing data process is the chi-square goodness of fit. The pairwise deletion method shows less model rejections under the MCAR condition than under MAR condition.

3. *The full information approach implemented in AMOS should be superior to EM estimation of the covariance matrix and the multiple imputation method.* One reason for this is that AMOS uses a direct estimation procedure. A more fundamental reason is that the EM covariance estimates and the multiple imputation estimates of the missing values do not take into account the identification of both SEM models. In fact both models are overidentified. According to Allison (1987, p.79), "in order to have efficient parameter estimates in the presence of missing data, the overidentifying restrictions should be incorporated in the ML estimation procedure". A number of findings support this hypothesis. In the case of the measurement model the bias of the parameter estimates (for the 5% missing values condition) is better with AMOS compared to the four other missing data methods. Although AMOS slightly overestimates the standard errors when the fraction of missing data is moderate or high, the results from the analysis of the confidence intervals for the parameter estimates consistently indicate that direct ML yields the best estimates. This full information method shows smaller correlations between the parameter estimates and the standard errors compared to the two indirect ML methods. As should be the case according to theory, the standardized parameter estimates computed with AMOS are standard normal distributed. With the two indirect ML methods, the variances of these standardized parameter estimates are larger than 1.

4. *The results obtained for the measurement model should be equal to those obtained for the full structural model.* None of the five techniques dealing with missing data problems yields consistently better or worse results for one of the two models. Analysis of the numerical and graphical output confirms this hypothesis

5. *Because the three ML techniques are developed under the normal model, we expect better results with multivariate-normal data sets.* In contrast to that, no impact of the distribution of the data is expected for the two quick methods because these techniques do not require distributional assumptions. As for the three ML methods, no improvement in parameter estimation due to normality is observed. The only result is that the differences between the parameter estimates for the ML methods become smaller when the data are normally distributed. No single best method exists however. The bias for the standard errors is not influenced by the presence of non-normality for the three ML methods. In the case of listwise deletion, the bias in the parameter estimates in the normal data case is smaller compared to this bias in the presence of skewed data. This may be due to LISREL which is known to yield better parameter estimates under normality conditions in small samples (see Boomsma, 1983). No relationship between distributional characteristics and the quality of the parameter estimates is noticed for the pairwise deletion procedure. For the two quick methods, the bias of the standard errors is not influenced by the skewness of the data.

6. *In the presence of few missing data (5%), smaller differences are expected between the performance of the two quick methods and the three ML methods compared to the situations with moderate (15%) and higher (25%) fractions of missing data.* This effect should be noticeable because the two quick methods are both characterised by an absence of effort to do something about the missing data problem. This hypothesis is not rejected and the effect is especially observable for the bias in the parameter estimates. Large differences are present for the bias of the parameter estimates with 25% missing data. The differences are smaller with 5% missing data.

A general picture of the results can be seen in figures 1 and 2. The data are the lambda parameter values (regression coefficients of observed on latent variables) from the measurement model. The three levels of missing data are represented as 5% MV, 15% MV and 25% MV. Furthermore LW=listwise deletion, PW=pairwise deletion and MI=multiple imputation.

Figure 1.
Plot of Means : MCAR case

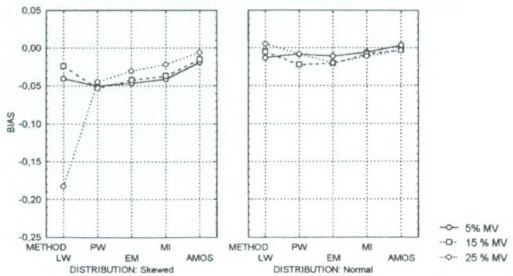
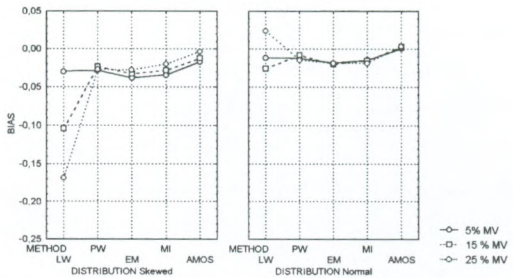


Figure 2.
Plot of Means : MAR case



5. DISCUSSION

Our results from the performance study clearly line up with findings from many other sources. Yet our study tried to test them in one overall research design. The next section contains the findings that match ours:

Haitovski (1968) concluded that in the presence of a lot of missing data, the complete cases method (listwise deletion) is worse than the available information method (pairwise deletion). According to Beale and Little (1975) ML outperforms the listwise deletion method. They also found that different ML approaches yield not very different results. Kim and Curry (1977) concluded that the pairwise method is better than the listwise procedure. Finkbeiner (1979) concluded that direct ML is best. Another finding by Finkbeiner is that the pairwise deletion

method yields estimates that are close to ML estimates while the listwise procedure is much worse. The listwise procedure was found to improve as less data are missing. Brown (1983) found the pairwise method more efficient than the listwise procedure. The EM technique outperforms the pairwise method. All ML procedures yield similar high quality estimates that are better than the pairwise deletion results. Malhotra (1987) concludes that the EM method outperforms the listwise deletion method. He also finds that as the fraction of missing data decreases, the differences between the methods decrease. In Brown (1994) listwise deletion yields overestimated standard errors. These are larger compared to other methods. The pairwise deletion method yields good estimates for the standard errors.
Arbuckle (1995a, 1995b) indicates that both the pairwise method and the direct ML procedure provide very good estimates of the parameter values under the MCAR condition. The ML estimator is more efficient and is normal in shape. The pairwise method yields results that are in between the results of listwise deletion and direct ML. Arbuckle concludes that under MCAR conditions, this ML method is superior to the pairwise and listwise results. Under MAR conditions, the direct ML method outperforms the pairwise and listwise method. With normal distributed data, the direct ML method performs better than under non-normal data conditions.

In general, for data that is to be modelled with SEM, it is our belief that the two quick methods (listwise and pairwise deletion) are better not applied. Since we prefer one approach that is generally applicable, the AMOS direct estimation procedure is to be preferred. Despite the lack of a chi-square fit test, this approach has many qualities. This is also reported by Duncan, Duncan and Li (1998).

Some questions do remain. Although the MAR method used in this study is a procedure that satisfies the criteria for MAR, other missing data processes that are MAR exist. The efficiency of the ML methods is independent of the kind of MAR process, as long as it is MAR. Maybe the two quick methods yield different results for different MAR processes. An other question is to what extent missing data solutions yield efficient results with even larger fractions of missing data. How much observed data do they need do work properly?

This Monte Carlo performance analysis compares five missing data techniques. Even if the inclusion of the present five techniques was motivated , the inclusion of other methods could have been interesting. A potential fruitful approach is presented by Steinberg D. and Colla P. (1995) and Breiman L. et al. (1984) and implemented in the CART program .

Next to the factors that are present in this performance analysis, other factors might be included in a future exercise. In a following step, it might be interesting to assess the efficiency of missing data techniques that are applied to models that are not correctly specified. Two ML methods used in this study (EM and multiple imputation) need redundancy in the variables (covariance), as previously treated. How large must the correlations minimally be so that these methods can yield results? Another line of future research is in the domain of systematic missing data where the missing data process is non ignorable.

Bibliography

Afifi A.A. and Elashoff R.M., 1966, Missing observations in multivariate statistics I, *Journal of the American Statistical Association*, 61, 595-604

Afifi A.A. and Elashoff R.M., 1967, Missing observations in multivariate statistics II, *Journal of the American Statistical Association*, 62, 10-29

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

Arbuckle J., 1993, ,AMOS: the Analysis of MOment Structures, computer program, dept. of psychology, Temple University, Pennsylvania

Arbuckle J., 1995, Full information estimation in the presence of incomplete data, working paper, dept. of psychology, Temple University, Pennsylvania

Arbuckle J., 1995, Advantages of model-based analysis of missing data over pairwise deletion, research paper, Temple University, Dept. of Psychology, Philadelphia, Pennsylvania 19122, USA

Beale E.M. and Little R.J.A., 1975, Missing values in multivariate analysis, *Journal of the Royal Statistical Society*, 37, 129-145

Bollen K.A., 1987, *Structural equations with latent variables*, John Wiley and Sons Inc., New York

Boomsma A., 1983, On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality, Unpublished doctoral dissertation, University of Groningen, Groningen, The Nederlands

Breiman L., Friedman J., Olshen R. and Stone C., 1984, Classification and regression trees. Pacific Grove: Wadsworth

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

Brown R.L., 1994, Efficacy of the indirect approach for estimating structural equation models with missing data: a comparison of five methods, *Structural equation modeling*, 4, 287-316

Chuanhai Liu, 1992a, Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data, Working Paper, Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Chuanhai Liu, 1992b, Efficiently drawing the posterior mean and covariance for monotone normal ignorable missing data, Technical Report, Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Duncan T.E., Duncan C.D. and Fuzhong Li, 1998, A comparison of model- and multiple imputation based appraoches to longitudinal analyses with partial missingness, *Structural Equation Modeling*, 5(1), 1-21

Duncan O.D., Haller A.O. and A. Portes, 1968, Peer influences on aspirations : a reinterpretation., *The American Journal of Sociology*, 74, 119-137

Finkbeiner C., 1979, Estimation for the multiple factor model when the data are missing, *Psychometrika*, 44, 409-420

Glasser M., 1964, Linear regression analysis with missing observations among the independent variables, *Journal of the American Statistical Association*, 59, 834-844

Gleason T.C. and Staelin R.A., 1975, A proposal for handling missing data, *Psychometrika*, 40, 229-252

Haitovsky Y., 1968, Missing data in regression analysis, *Journal of the Royal Statistical Society*, Series B, 30, 67-82

Jöreskog K.G. and Dag Sörbom, 1989, LISREL 7 User's reference Guide, Scientific Software Inc., Mooresville

Kim J. and Curry J., 1977, The treatment of missing data in multivariate analysis, *Sociological Methods and Research*, 6, 215-241

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

Malhotra N.K., 1987, Analyzing market research data with incomplete information on the dependent variable, *Journal of Marketing Research*, 14, 74-84

Neale M.C., 1994, Mx statistical modeling 2nd edition, Box 710 MCV, Richmond, VA 23298, Department of Psychiatry

Rubin D.B., 1976, Inference and missing data, *Biometrika*, 63, 581-592

Rubin D.B., 1987, Multiple imputation for nonresponse in surveys, John Wiley and Sons, New York

Steinberg D. and Colla P., 1995, CART: Tree-structured Non-Parametric Data Analysis. San Diego, CA: Salford Systems

Tanner M.A. and Wong W.H., 1987, The calculation of posterior distributions by data augmentation, *Journal of the American Statistical Association*, 82, 528-550

Timm N.H., 1970, The estimation of variance-covariance and correlation matrices from incomplete data, *Psychometrika*, 35, 417-437

Verleye G., 1996, Missing at random data problems in attitude measurement using maximum likelihood structural equation modeling, unpublished Ph.D. dissertation, Vrije Universiteit Brussel, Centrum voor Statistiek en Wiskunde.

Ontvangen: 22-01-1999
Geaccepteerd: 25-07-1999