2011 International Conference on Information and Electronics Engineering
IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore

Comparison of a Data Imputation Structural Equation Modeling Accuracy Estimation Between Constrained and Unconstrained Approaches

Narong Phothi [1] and Somchai Prakancharoen [2 +]

[1] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand.
[2] Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Thailand.

Abstract. This study aimed to comparison of a data imputation structural equation modeling (SEM) accuracy estimation between constrained and unconstrained approaches. The measurement accuracy of the model based on the mean magnitude of relative error (MMRE) model. The model is developed by using the online database from University of California, Irvine (UCI) which is a data set on waveform generators. Indicators 21 (1,200 sets) methods were as follows: 1) Data set was divided into two groups (experimental group of 1,000 sets and test group of 200 sets); 2) The experimental group was analyzed by three main factors (F1, F2, F3); 3) Create a SEM; 4) The remaining indicators from the results in section 3 were used to create new factors with the constrained approach. All the indicators are related and used to construct a SEM to estimate the equation missing. The test data was substituted in the equation to find the accuracy which was 43.00% (MMRE was 57.00%); 5) the remaining indicators from the results in section 3 were used to create a new factor with the unconstrained approach. The test data was also substituted in the equation to find the accuracy which was 65.71% (MMRE was 34.29%). Thus, comparing estimates of missing data showed that using the SEM with the unconstrained approach employed indicators, which were related for more accuracy while MMRE declined using the constrained approach with related indicators.

Keywords: Data Imputation Estimation, Structural Equation Modeling (SEM), Constrained Approaches, Unconstrained Approaches.

1. Introduction

1.1. Background

General research information in this area is required to complete the analysis in order to achieve the most accurate and precise results. However, some data may be missing or incomplete. Therefore, in order to bring a data set that is complete and ready to use. Some data will be missing. This would result in the records becoming redundant or obsolete, thus analysis and forecast of data is needed. If some of the data set were missing in large amounts, data that is needed should avoid deviation, which would lead to error of the results and need to be obtained through processing. The estimation of missing data will help in preparing data to replace the missing research data sets. From the research on estimation of loss of data, such as researcher Prakancharoen [1] using SEM to estimate the time to develop application software oriented network, also researchers Phothi and Prakancharoen [2] using SEM between with discriminant analysis and without discriminant analysis for accuracy comparison of imputation methods, and researcher Rufus [3] using solutions for missing data in SEM for new data all use similar approaches to solve this issue.

In this study, researchers comparison of a data imputation SEM accuracy estimation between constrained and unconstrained approaches. The research information is taken from the online database, UCI Machine Learning Repository is a collection of waveform database generator data sets (1,200 sets). The measurement accuracy of the estimated loss from the MMRE was found to be highly accurate.

1.2. The purpose of the research

- To estimate missing data by using product indicator approaches of SEM with the constrained approach.

+ Corresponding author. Tel.: +66-2913-2500; fax: +66-2912-2019.
E-mail address: narong@sueksa.go.th, spk@kmutnb.ac.th

- To estimate missing data by using product indicator approaches of SEM with the unconstrained approach.
- To compare the accuracy of estimates the missing data by SEM, product indicator approaches (PIA) of SEM with the constrained approach and PIA of SEM with the unconstrained approach.

1.3. Scope of research

- The data used in this operation was a waveform database generator data set from the online database, and UCI Machine Learning Repository as a data type with 1,200 sets, which were divided into two groups: experimental group (1000 sets) and test group (200 sets).
- The data set has 21 indicators, namely, V1-V21 and C1 classes for the description of each indicator V., which can be viewed at http://archive.ics.uci.edu/ml/datasets.html determined that the fifth indicators (V5) in the equation of the test group were missing valuable data used to compare the accuracy of the estimation method. Missing value due to a measure of this needs to find the best relationship associated with other indicators in a waveform database generator data set.

2. Theory and Methodology

2.1. Factor Analysis

Factor Analysis [4] is a technique used to extract the factors (component) from a group of indicators that are related to each factor. This will be used instead of a group of indicators that have the same group. This is a technique that reduces the number of dimensions or manifest variable and considers the suitability of the extracted factors. By checking the statistics Kaiser-Meyer-Olkin: KMO (KMO>0.60) factors obtained will only validate the considered values. Able to explain the variability of all the factors together (total variance explained) with the inverse of each variable with no apparent extraction factor would greatly benefit this approach. If the value of a high percentage (cumulative explained variance) showed that the factors can represent a good indicator, this can be formulated as follows.

$$F_j = w_{j1}x_1 + w_{j2}x_2 + ... + w_{jp}x_p + e \tag{1}$$

where F =factor, w =coefficient of variable x, x =manifest variable and e =margin of error.

2.2. Structural Equation Modeling (SEM)

SEM [1], [5] is a technique used to analyze the relationship of factors from the survey (exploratory) with a key and then extract a model of the relationship of various factors, which is the main theory or hypothesis of this study. From the statistics of 1) Chi-square (χ^2) should be a non-significance (P>0.05) 2) Goodness of Fit Index (GFI>0.90) 3) Root Mean Square Error of Approximation (RMSEA<0.06) and 4) Hoelter's N, the value (Hoelter's. N>75) is used to check the adequacy and appropriateness of sample size (case) in SEM.

2.3. Product Indicator Approaches of Structural Equation Modeling (PIA of SEM)

PIA of SEM [6] is a technique used to estimate the stability of the equation appears in the relationship between variables. The equation is made up of indicators that are related, formulated as follows.

$$X_iZ_j = \lambda_i\lambda_j\xi_1\xi_2 + \lambda_i\xi_1\delta_j + \lambda_j\xi_2\delta_i + \delta_i\delta_j \tag{2}$$

where X_i=predictor of variable, Z_j=moderator of variable, ξ=factor, λ=factor loading and δ=margin of error. To build PIA with two different techniques. 1) Constrained approach [7] creates a new factor X_iZ_j by bringing a measure of the factor X_i by match multiplied with all indicators of factor Z_j and repeated until completed. 2) Unconstrained approach [8] creates a new factor X_iZ_j by bringing a measure of the main X_i to multiply indicator 1:1 match with a measure of factor Z_j.

2.4. Accuracy Evaluation Criterion

Accuracy Evaluation Criterion [1] of a new data set, which must be precisely compatible (model best fit) by applying a set of new data (predicted missing) derived from the estimation of missing data to verify the real data set (actual missing) and then calculate the Magnitude of Relative Error (MRE) according to the formula.

$$MRE = \frac{\left|ActualMissing - PredictedMissing\right|}{ActualMissing} \quad (3)$$

The missing data (i = 1, 2, ..., n) must be used for calculating the Mean Magnitude of Relative Error (MMRE). If it is found that the results of MMRE have small values, the results should be precise or very close to the real data as formulated below.

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|ActualMissing_i - PredictedMissing_i\right|}{ActualMissing_i} x100 \quad (4)$$

$$Accuracy = 100 - MMRE \quad (5)$$

3. Methodology

3.1. Classification of data sets for the research

Classification or divided data set of waveform database generator 1,200 sets into two groups: the experimental group was 1,000 data sets and the test group was 200 data sets.

3.2. The factor analysis of experimental group

The experimental group focused on the factor analysis method by principal component analysis to provide a measure that is relevant to the factors in the same way as rotation varimax to reduce the number of points. This should measure the weight of each factor to as low as possible. Results from the analysis of new factors with KMO were 0.961, and new factors from extraction consist of three main factors F1, F2 and F3 are shown in Table 1.

Table 1: Results of main factors and indicators

Factor	Indicator of Factor
F1	V17 , V9 , V10 , V16 , V15 , V18 , V8 , V19 , V20
F2	V5 , V13 , V12 , V6 , V7 , V4 , V14 , V11 , V3 , V2
F3	V21, V1

3.3. Structural Equation Modeling

The main factors F1, F2 and F3 of building a SEM are shown in Fig. 1. The model appropriate to review the statistics of the compatibility of the model to goodness of fit: RMSEA, GFI and Hoelter's N which are the adequacy of the sample cases. The results in Table 2 and the new SEM are shown in Fig. 2.

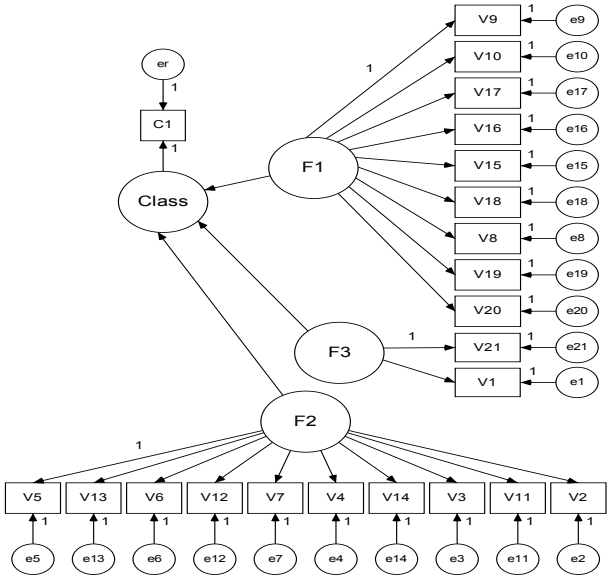

Fig. 1: Prototype of SEM

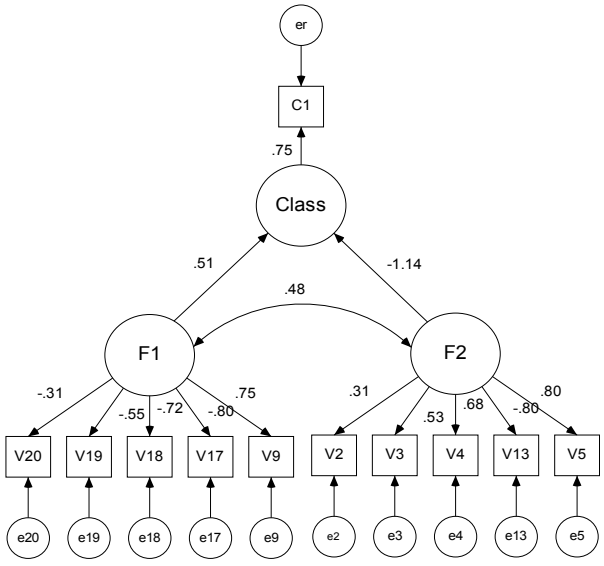

Fig. 2: SEM standardized type

Table 2: The statistic's compatibility of SEM

Model	χ^2	P	GFI	RMSEA	Hoelter's N
Default	27.700	0.956	0.995	0.000	2094/2385

3.4. Product indicator approaches of structural equation modeling with constrained approach

The only measure left over from the results of the SEM according to Fig. 2 is F1 = {V9, V17, V18, V19, V20} and F2 = {V2, V3, V4, V5, V13} to create new factors related. By bringing a measure of the factor F1 one by one to match multiplied with a measure of factor F2 all and repeats until all indicators of the factors F1. A result is F1F2 = {V17V2, V17V3, V17V4, V17V5, V17V13, V18V2, V18V3, V18V4, V18V5, V18V13, V19V2, V19V3, V19V4, V19V5, V19V13, V20V2, V20V3, V20V4, V20V5, V20V13} and then create SEM have the statistics of compatibility in Table 3, and the new SEM is depicted in Fig. 3 with the equation estimated by equation 6-9.

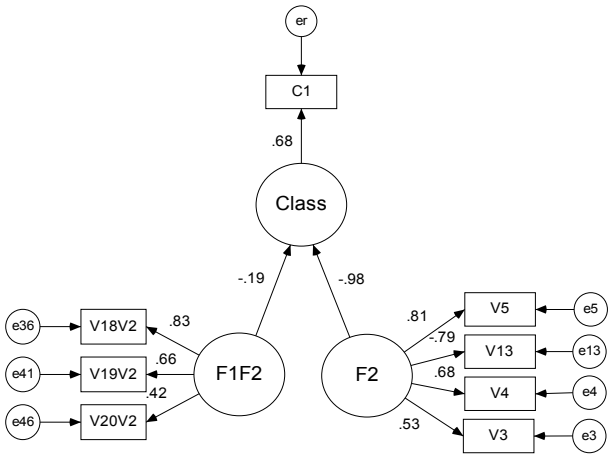

Fig. 3: The constrained approach model standardized type

Table 3: The statistic's compatibility of the constrained approach model

Model	χ^2	P	GFI	RMSEA	Hoelter's N
Default	26.900	0.107	0.993	0.020	1120/1344

$$\text{Class} = (0.68*\text{C1})+\text{er} \quad (6)$$

$$\text{F2} = (\text{Class}+(0.19*\text{F1F2}))/(-0.98) \quad (7)$$

$$\text{F1F2} = (0.83*\text{V18V2})+(0.66*\text{V19V2})+(0.42*\text{V20V2}) \quad (8)$$

$$\text{V5} = (\text{F2}-((0.53*\text{V3})+(0.68*\text{V4})-(0.79*\text{V13})))/(0.81) \quad (9)$$

3.5. Product indicator approaches of structural equation modeling with unconstrained approach

The only measure left over from the results of the SEM according to Fig. 2 is F1 = {V9, V17, V18, V19, V20} and F2 = {V2, V3, V4, V5, V13} to create new factors related. By applying a metric from the factor F1 to multiply 1:1 match with a measure of the factors F2. A result is F1F2 = {V9V5, V17V13, V18V4, V19V3, V20V2}, and then create SEM have the statistics together in Table 4, and the new SEM as Fig. 4 with the equation estimated by equation 10-13.

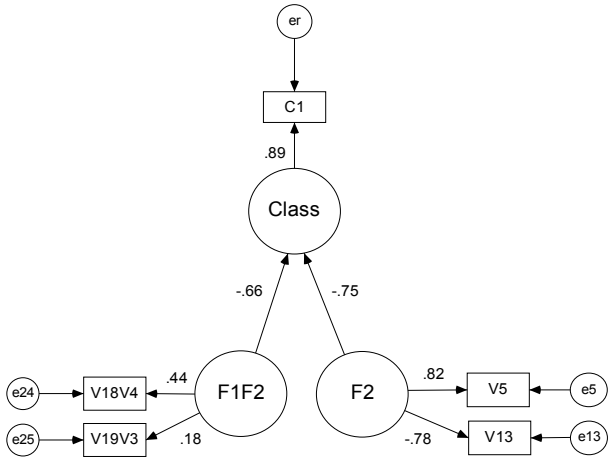

Fig. 4: The unconstrained approach model standardized type

Table 4: The statistic's compatibility of the unconstrained approach model

Model	χ^2	P	GFI	RMSEA	Hoelter's N
Default	3.632	0.458	0.999	0.000	2610/3652

$$Class = (0.89*C1)+er \quad (10)$$

$$F2 = (Class+(0.66*F1F2))/(-0.75) \quad (11)$$

$$F1F2 = (0.44*V18V4)+(0.18*V19V3) \quad (12)$$

$$V5 = (F2+(0.82*V13))/(0.82) \quad (13)$$

4. Results

The test group of 200 sets was assigned to find missing V5 and estimate the replacement value of missing data as follows: 1) the data imputation estimation methods using PIA of SEM with the constrained approach as equation 6-9, the result of accuracy was 43.00% (MMRE was 57.00%) and 2) the data imputation estimation methods using PIA of SEM with the unconstrained approach as equation 10-13, The result of accuracy was 65.71% (MMRE was 34.29%).

Thus, comparing estimates of missing data showed that using the SEM with the unconstrained approach and related indicators had high accuracy, while MMRE declined using the constrained approach with related indicators.

5. Conclusions

Data imputation estimation methods using PIA of SEM with a data set from the waveform database generator. Numeric indicators 21 of 1,200 sets of nonlinear type showed that the grouping of data sets or analysis of main factors for the indicators are related to factors in the same area. When estimating missing data, the results of MMRE errors were reduced. Making a new data from the missing estimation method is more accurate than the new values.

Suggestions about the data imputation estimation methods using PIA. The related indicators are used in the case of latent factors outside the relationship between the two directions only. If no such event, this method will not be able to be used.

6. References

[1] Prakancharoen S. *The estimated time to develop application software oriented network Using structural equation modeling*. Information Technology Journal. Year 4 Vol. 7. Bangkok: King Mongkut's University of Technology North Bangkok, 2008.

[2] Phothi N. and Prakancharoen S. "Accuracy Comparison of Imputation Methods Using Structural Equation Modeling Between With Discriminant Analysis and Without Discriminant Analysis". *Conference on Science and Technology No. 8*. Pathum Thani: Thammasat University Rangsit Campus, 2010.

[3] Rufus L. C. *Solutions for Missing Data in Structural Equation Modeling*. Research & Practice in Assessment Vol. 1, Issue 1 March 2006.

[4] Vanitbancha K. *Multivariate Data Analysis*. Vol. 2. Bangkok: Chulalongkorn University Book Center, 2007.

[5] Garson G. D. *Data Imputation for Missing Values*. North Carolina State University, USA, 2005.

[6] Karin S., Christina W., Helfried M. "Nonlinear Structural Equation Modeling: Is Partial Least Squares an Alternative?". *Meeting of the Working Group Structural Equation Modeling*. Berlin, Germany, February 26-27, 2009.

[7] Joreskog, K. G., & Yang, F. *Nonlinear structural equation models: The Kenny-Judd model with interaction effects*. In G. Marcoulides & R. Schumacker (Eds.), Advanced structural equation modeling (pp. 57-87). Mahwah, NJ: Lawrence Erlbaum Associates. 1996.

[8] Marsh, H. W., Wen, Z., & Hau, K. T. *Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction*. Psychological Methods, 9, 275–300. 2004.