MULTIPLE IMPUTATION OF MISSING DATA IN SUSTAINABLE DEVELOPMENT MODELLING

Roberto Benedetti*, Rita Lima#, Alessandro Pandimiglio^

JEL classification: C15, C51, O10.
Keywords: LISREL, Markov Chain Monte Carlo, Multiple Imputation, Sustainable Development.

1. Introduction

Since the 1992 Rio declaration on environment and development, sustainable development is commonly accepted as the development which meets the needs of the present without compromising the ability of future generations to meet their own needs. From a statistical point of view sustainable development (SD) is a multidimensional phenomenon that is not directly measurable; for this reason it would be of some interest to understand the relationship between such a complex variable and its several indicators obtained from the core observed variables.

Sustainability indicators are mostly environmental indicators such as climate changes, air quality, waste generations, etc. Development indicators are based on variables such as gdp per capita growth, labour productivity, human rights, quality of life, gender inequality, etc. Often, indicators are only a partial reflections of reality because of uncertainty and measurement errors.

Structural equation modelling is a very general multivariate analysis technique that hypothesizes casual relationships among variables and tests the casual models with a linear equation system (LISREL); casual models can involve manifest variables, latent variables or both.

What happens if we try to measure a latent variable, such as SD, using indicators that very often have a missing data problem? What is the effect of missing data on the parameters of a sustainable development model?

The purpose of this paper is to test empirically the effect of missing data on the parameters of a LISREL SD model. For this purpose we will have two data subsets. The first dataset simply takes the data from the SD dataset. We
will call this the ‘actual data.’ The actual data imply a realistic missing data problem.

The second dataset, ‘actual plus data,’ takes the actual data and adds imputed missing data.

The most commonly practiced methods of handling missing data in structural equation modelling are:

a) listwise deletion (LD): whenever a record has missing data for any one variable used in SD analysis this method omits that entire record from the analysis.

b) pairwise deletion (PD), a couple of variables are included in the analysis if and only if are both not missing to estimate correlations and/or covariances.

c) mean imputation (MEI), take the mean of observed values of the variable Y.

d) regression methods (RM), which develop a regression equation based on complete case data for a given variable, treating it as the outcome and using all other relevant variables as predictors. Then, for cases where Y is missing, plug the available data into the regression equation as predictors and substitute the equation’s predicted Y value into the database for use in other analyses. An improvement to this method involves adding uncertainty to the imputation of Y so that the mean response value is not always imputed.

e) hot deck imputation (HDI), which identify the most similar case to the case with a missing value and substitute the most similar case’s Y value for the missing case’s Y value.

f) expectation maximization (EM) approach, An iterative procedure that proceeds in two discrete steps. First, in the expectation (E) step you compute the expected value of the complete data log likelihood. In the maximization (M) step you substitute the expected values for the missing data obtained from the E step and then maximize the likelihood function as if no data were missing to obtain new parameter estimates. The procedure iterates through these two steps until convergence is obtained.

g) raw maximum likelihood methods (RLM), which use all available data to generate maximum likelihood-based sufficient statistics. Usually these consist of a covariance matrix of the variables and a vector of means. This technique is also known as Full Information Maximum Likelihood (FIML).
The performance of these methods under different types of missing data processes is summarized by Roth (1994) that concludes\(^1\) that PD and LD estimates are consistent, although not efficient. MEI is consistent in the first moments, but yields biased and/or inefficient estimates. In addition, PD does not provide standard errors of parameters estimates or tests of model fit, while MEI lead to standard errors that are too small and p-values that are artificially low. Listwise, pairwise, and mean substitution missing data handling methods are inferior when compared with maximum likelihood based methods such as raw maximum likelihood. Regression methods are somewhat better, but not as good as hot deck imputation or maximum likelihood approaches. The EM method falls somewhere in between: It is generally superior to listwise, pairwise, and mean substitution approaches, but it lacks the uncertainty component contained in the raw maximum likelihood and multiple imputation methods.

For missing data that are completely random, multiple imputation (MI) is a modern missing data technique, developed by Rubin (1987), which carries out the averaging via simulation. In MI each missing value is replaced by a set of \(m \geq 1\) plausible values drawn from their predictive distribution. The variation among the \(m\) imputation reflects the uncertainty with which the missing values can be predicted from the observed ones. MI often is attractive because it can be highly efficient even for small values of \(m\).

In this paper we will apply this methodology to a SD dataset showing how MI technique could give excellent results.

The plan of the paper is as follows. In section 2, it is introduced the main idea of MI. To reduce the missing data effect, Markov chain Monte Carlo (MCMC) method is proposed in order to generate imputation for the missing value (Schafer, 1997). In section 3, it is described the LISREL model used to measure SD. Finally, in section 4 there is a comparison between estimates achieved by ignoring missing items and estimates obtained by substituting plausible values for missing data with MI methodology.

2. Handling missing data with multiple imputation methodology

In a multivariate dataset, where missing values occur on more than one variable, the incomplete cases could be a substantial portion of the entire dataset, therefore a simple method of case deletion may be inefficient because large amount of information may be discarded. In this case MI is more efficient because uses all the available information in the dataset.

To create imputations we can use the:

\(^1\) See also Little & Rubin (1987) and Wothke (1998).
a) \textit{parametric method:}
- specify a model for complete data;
- for each missing data point:
  \begin{itemize}
  \item estimate predictive distribution of the missing data;
  \item impute with a random value from this distribution.
  \end{itemize}

b) \textit{non parametric method:}
- group similar cases into adjustment cells;
- for each missing data point:
  \begin{itemize}
  \item collect non-missing cases from adjustment cell;
  \item impute with value from randomly selected non-missing case.
  \end{itemize}

In the parametric approach, MI data sets are simulated draws from a predictive distribution of the missing data. It requires a model for the complete data with uncertainty about missing values and parameters of predictive distribution. Complex computation use the Markov Chain Monte Carlo (MCMC) method (Schafer 1997).

In the non parametric approach, a bootstrapped regression equation in total dataset is used to predict nonresponse on incomplete variable; dividing the dataset into imputation classes based on predicted nonresponse probability (propensity score), observed valued is randomly imputated from imputation class.

Another classification of MI’s method depends on the patterns of missingness in the data. For data sets with monotone missing patterns, either a parametric regression method (Rubin 1987) that assumes multivariate normality or a nonparametric method that uses propensity scores (Rubin 1987; Dawson and Shera 1995) is appropriate. Suppose Y is the nxp matrix of complete data, in which Y_{obs} is the observed part of Y and Y_{mis} is the missing part of Y. Suppose that R is the nxp matrix of response indicators whose elements are zero or one depending on whether the Y’s elements are missing or observed. The missing data are missing at random (MAR) if
\[ p(Y_{obs}|Y_{mis}) = p(Y_{obs}) \]
That is, missing data values carry no information about probabilities of missingness. This assumption is mathematically convenient because it allows to eschew an explicit probability model for missingness.

When data sets have a MAR pattern and are multivariate normally distributed, MCMC is used to generate pseudorandom draws from multidimensional probability distributions using simulations from a Bayesian prediction distribution for normal data. In Bayesian inference, information about unknown parameters \( \theta \) is expressed in the form of a posterior probability distribution, which is computed using Bayes theorem. These random draws are then used for MI inference.
Through MCMC it is possible to simulate the entire joint posterior distribution of the unknown parameters and obtain simulated estimates of posterior parameters that are of interest. In many incomplete data problems, the posterior distribution $p(\theta | Y_{\text{obs}})$ is much more intractable and cannot easily be simulated than the posterior distribution $p(\theta | Y_{\text{obs}}, Y_{\text{mis}})$ when $Y_{\text{mis}}$ is augmented by an estimated/simulated value of $Y_{\text{mis}}$. Data augmentation can be applied to Bayesian inference with missing data starting with a given mean vector $\mu$ and covariance matrix $\Sigma$ and drawing values for the missing data from the conditional distribution $Y_{\text{mis}} | Y_{\text{obs}}$. The posterior population mean vector $\mu$ and covariance matrix $\Sigma$ are simulated from prior information for $\mu$ and $\Sigma$ and the complete sample estimates. Without prior information about $\mu$ and $\Sigma$, a non-informative prior can be used. In other words, data augmentation can be obtained by repeating the following steps:

1. **The imputation I-step**: with the estimated mean vector and covariance matrix, the missing values are simulated for each observation independently. That is, if you denote the variables with missing values for observation $i$ by $Y_{i(\text{mis})}$ and the variables with observed values by $Y_{i(\text{obs})}$, then the I-step draws values for $Y_{i(\text{mis})}$ from the conditional distribution of $Y_{i(\text{mis})} | Y_{i(\text{obs})}$;

2. **The posterior P-step**: the posterior population mean vector and covariance matrix are simulated from the complete sample estimates. These new estimates are then used in the I-step. Without prior information about parameters, a non-informative prior is used.

The two-steps are iterated long enough for the results to be reliable for a multiply imputed data set (Schafer, 1997). Often, few imputations are adequate in multiple imputation (Rubin, 1996).

The goal is to have the iterates converge to their stationary distribution and then to simulate an approximately independent draw of the missing values. That is, with a current parameter estimate $\theta^{(t)}$ at $t^{\text{th}}$ iteration, the I-step draws $Y_{\text{mis}}^{(t+1)}$ from $p(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(t)})$ and the P-step draws $\theta^{(t+1)}$ from $p(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(t+1)})$.

This creates a Markov chain $(Y_{\text{mis}}^{(1)}, \theta^{(1)}), (Y_{\text{mis}}^{(2)}, \theta^{(2)}), \ldots$, which converges in distribution to $p(Y_{\text{mis}}, \theta | Y_{\text{obs}})$.

Rate of convergence is related to fractions of missing information. It can be used either a single chain for all imputation (single chain) or a separate chain for each imputation (multiple chain).
Our procedure uses a single chain to simulate draws from the posterior distribution of \( p(Y_{mis}, Y_{obs}) \) because there is not be noticed a significant difference between single and multiple chain.

The ML estimates from EM algorithm provides a good starting values of parameters for the chain. With \( m \) imputation, it can be possible to compute \( m \) different sets of mean and variance estimates for a parameter \( \theta \).

The variability among the results from these \( m \) repeated analyses provides a measure of the uncertainty due to missing data. The relative efficiency of the small \( m \) imputation estimator is high for cases with little missing information.

Combining this between imputation variations with the ordinary within imputation sample variation provides statistical inference for \( \theta \). Rubin (1987) presented some rules for combining results from a data analysis performed \( m \) times, once for each of \( m \) imputed data sets, to obtain a single set of results.

After obtaining \( m \) imputations of \( Y_{mis} \), suppose that \( \hat{Q}_j \) is the complete-data point estimate obtained from data set \( j \) and \( U_j \) is the complete-data standard error associated with \( \hat{Q}_j \).

The overall estimate is the average of the individual estimates \( \bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j \). The overall standard error is the square root of \( T = \bar{U} + \left( 1 + \frac{1}{m} \right) B \) where \( \bar{U} = \frac{1}{m} \sum_{j=1}^{m} U_j \) is the within-imputation variance and \( B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})^2 \) is the between-imputation variance (Rules for scalar estimands; Rubin, 1987).

3. Structural equation modelling: model, data and hypotheses

The most frequently quoted definition of sustainable development comes from the Report Our Common Future - also known as the Brundtland Report - (World Commission on Environment and Development, 1987):

"Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

Sustainable development contains within it two key concepts:
1) the concept of “need”, with particular attention to the essential needs of the present and future poor people, to which overriding priority should be given;

2) the idea of “limitations” imposed by the state of technology and social organization on the environment’s ability to meet present and future needs.

Sustainable development aims to improve the quality of life for all of the Earth's citizens without increasing the use of natural resources beyond the capacity of the environment to supply them indefinitely. To be realized SD requires a sharp change in the way of thinking of people, and the role of institutional structure is crucial in order to influence and change individual behaviour. It is about taking action, changing policy and practice at all levels, from the individual to the international.

Unfortunately there is a single indicator of SD, there are a lots of variables that must be taken into account in order to measure the state of the world development. Starting from the human development and economic growth indicators, SD requires the knowledge of many others environmental indicators such as the quality of air, land and water. Very often all those variables are only partially available for some countries therefore missing data are inevitable in SD dataset and the measure of such important phenomenon is relatively poor. Measures and indicators of sustainability combine social, economic and environmental trends and illustrate the links between and among systems. Designing good sustainability indicators is a formidable task. Separate indicators of social, economic and environmental progress have been developed only in the 20th century. And comprehensive indicators of sustainability began to emerge only during the last decade, but they are advancing rapidly. From indicators developed for rural communities to indicators required by the United Nations, hundreds of sustainability indicator sets have been created. Still, they have far to go before they can claim to be widely used. Most sustainability indicators come as large, unwieldy reports, crammed with complex charts and graphs. Although useful to policy professionals and academics, most indicator sets are not practical for the media and public.

The growing ranks of indicator projects and professionals worldwide face two challenges that seemingly contradict each other:

- **Growing complexity.** As our understanding of the complexity of sustainability grows, how do we manage the huge quantity of data required to monitor it?
The demand for simplicity. Since public education and resulting political action are seen increasingly and urgently as the purpose for creating indicators, how do we present them in ways that are simple, elegant and effective, without compromising the underlying complexity?

Complex problems of SD require integrated or interlinked sets of indicators, or aggregations of indicators into indices. High-level decision-makers (government ministers, foundation executives, heads of corporations) routinely ask for a small number of indices that are easy to understand and use in decision-making. A sustainable development single indicator would be desirable in order to compete with the enormous political power of the Gross Domestic Product, a number that provides information about the total market value of production and services in a country in a single number. But many are skeptical that a single number could assess something as complex as sustainable development.

It is hard to think that one number, that roughly measures the overall national well-being, could have any real functional value as a policy tool. But many also acknowledge that the attempt to create an index of sustainable development may be useful because it would force a concerted effort to deal with the complexity of sustainable development.

A structural equation model (LISREL) allow us to present a handful of aggregated indices that could useful to measure SD. Determining the appropriate output variables (indicators of SD) is extremely important for the LISREL model. In general, there are two basic types of variables in a LISREL: the latent variables and the observable variables. Latent variables are represented or measured by observable variables. In the model it is also assumed that there is a casual structure among latent variables and observed variables.

A LISREL model consists of two parts:

a. the measurement model specifies how latent variables, or hypothetical constructs, depend upon or are indicated by the observed variables, furthermore it describes the measurement properties of the observed variables;

b. the structural equation model specifies the casual relationships among the latent variables and describes the casual effects.

In a well-known framework the general LISREL can be represented by the following equations:

\[
\eta_{(m \times 1)} = R_{(m \times m)} \cdot \eta_{(m \times 1)} + \Gamma_{(m \times n)} \cdot \xi_{(n \times 1)} + \zeta_{(m \times 1)}
\]
2. \[ y_{(p \times 1)} = \Lambda_{y(p \times m)} \ast \eta_{(m \times 1)} + \epsilon_{(p \times 1)} \]
3. \[ x_{(q \times 1)} = \Lambda_{x(q \times n)} \ast \xi_{(n \times 1)} + \delta_{(q \times 1)} \]

where:
\[ y = (y_1, \ldots, y_p)' \] is the dependent (observable) variables’ vector of \((p \times 1)\) elements;
\[ x = (x_1, \ldots, x_q)' \] is the independent (observable) variables’ vector of \((q \times 1)\) elements;
\( \Lambda_y \) is the \((p \times m)\) coefficients’ matrix related to \( y \) and \( \eta \);
\( \Lambda_x \) is the \((q \times n)\) coefficients’ matrix related to \( y \) and \( \xi \);
\( \epsilon \) is the \((p \times 1)\) vector of error terms in the second matrix equation;
\( \delta \) is the \((p \times 1)\) vector of error terms in the third matrix equation.

As far as the latent variables are concerned, the following assumptions are made:

i. \( M(\eta) = M(\xi) = M(\varsigma) = 0 \);
ii. \( \text{cov}(\xi, \varsigma) = 0 \);
iii. \( \text{diag}(B) = 0 \);
iv. \( B^* = (I - B) \) not singular matrix ;
v. \( M(\epsilon) = M(\delta) = 0 \).

\( \Phi = M(\xi \xi') \), \( \Psi = M(\varsigma \varsigma') \), \( \Theta_\varepsilon = M(\epsilon \epsilon') \) and \( \Theta_\delta = M(\delta \delta') \) are the covariance’s matrixes between latent exogenous variables and casual errors.

All these matrixes could have any form. However, if we assume that some elements should play or not an important role, they could have a particular configuration. For example, a LISREL for which \( B \) have elements close to zero in the highest triangular part is said partial recursive model. A partial recursive model for which \( \Psi \) is a diagonal matrix, is said recursive model. The SD model is specified in the form of three relationships.

The first two relationships specify how the \( y \)-variables, i.e. the two observed variables on the right side of equation, depend on the latent variable SD.

1. \( \text{HDI} = \lambda_1 SD + \varepsilon_1 \)
2. \( \text{GS} = \lambda_2 SD + \varepsilon_2 \)

The last relationship specifies how the latent variable SD depends on the \( x \)-variables, the exogenous variables.
3. \[ SD = \gamma_1 LIFEXPE + \gamma_2 ADULITE + \gamma_3 COMENROL + \gamma_4 PPP + \gamma_5 NDS + \gamma_6 EDUC + \gamma_7 DIOXI + \gamma_8 FORES + \gamma_9 MINER + \gamma_{10} ENER + \varsigma \]

The disturbance terms, which are random variables, consist of errors in variables or measurement errors \( \epsilon_i \) (i=1,2) and error in equations, \( \varsigma \). These random components in each relationship are assumed uncorrelated with all the independent constructs.

To illustrate the meaning of this LISREL, let us represent graphically the SD structural equations model in a path diagram (figure 3.1). Path diagram plays a fundamental role in structural modelling. It is like a flowchart in which variables are interconnected with lines used to indicate casual flow.

**Fig. 3.1 - Path diagram for SD**

As a rule, variable in a path diagram may be grouped in two classes: those that do not receive inputs from any other variable in the diagram (the exogenous or independent variables), and those that receive one or more such casual inputs (the endogenous or dependent variables). Straight arrows are said to represent casual relationships between variables. The essential feature for the use of a casual arrow in the diagram is the assumption that a change in the variable at the tail of the arrow will result in a change in the variable at the head of the arrow, any other situation being equal.

The numerical solution of the path diagram gives the path coefficients that are \( \gamma \)s and \( \lambda \)s. These coefficients are as the standard multiple regression coefficients (Bollen, 1989). In the path diagram they tell us to what extent a change on the variable at the tail of the arrow is transmitted to the variable at the head of the arrow. Boxes indicate observed variables, cycle latent variable.
The non-recursiveness of the SD hypothetical diagram clearly shows the direct effect of observed variables on the latent variable and how the latent variable directly influences two of the most popular SD indicators that are the genuine saving (GS) and the human development index (HDI). In particular GS is the true saving rate in a country after accounting for investments in human capital, depreciation of produced assets, and the depletion and degradation of the environment. HDI is a (fairly complex) way of measuring development by combining indicators of life expectancy, educational attainment and PPP per capita income.

The model’s exogenous variables are:

a) *net domestic savings* (NDS) is the traditional gross saving, including foreign savings and excluding produced asset depreciation;

b) *education expenditure* (EDUC) is the public current expenditure on education as a percentage of GNP;

c) *energy* (ENER), *mineral* (MINER) and *net forest depletion* (FOREST) are assumed to be equal to total resource rents;

d) *carbon dioxide damage* (DIOXI) is a measure of the carbon dioxide emissions those stemming from the burnings of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid liquid and gas fuels and gas flaring (per capita metric tons);

e) *life expectancy at birth* (LIFEXPE) is the average number of years a new-born infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life;

f) *adult literacy rate* (ADULITE) is the percentage of population aged 15 years and over who are literate;

g) *combined gross enrolment ratio* (COENROL) is the total enrolment in primary/secondary/tertiary education, regardless of age, divided by the population of the age group which officially corresponds to primary/secondary/tertiary schooling;

h) *real GDP per capita* (PPP) is the gross domestic product per capita in current US dollars.

The data used to estimate the model’s parameters are collected from 180 countries and reflect 1997 experience on the manifest variables (UNESCO, 1998; the Human Development Report, 1998; World Bank, 1997). In figure 2 it is possible to notice some correlation between manifest variables.
Fig. 3.2 - Scatterplot of observable variables; Pearson Correlation Coefficient (R) and number of observation (N) available

<table>
<thead>
<tr>
<th>R</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6991</td>
<td>121</td>
</tr>
<tr>
<td>0.8818</td>
<td>171</td>
</tr>
<tr>
<td>0.7541</td>
<td>171</td>
</tr>
<tr>
<td>0.9318</td>
<td>166</td>
</tr>
</tbody>
</table>

Once the relationships in the theoretical model have been translated into a statistical model for a set of linear stochastic equations among random observable variables (the indicators) and the latent variable (the theoretical construct), the model can be estimated and tested by using the maximum likelihood (ML) method. In practice, the usual procedure is to solve the three equations simultaneously because, doing so, one brings to bear all information available about each path. The parameters of LISREL are obtained by minimising ML function of the discrepancy between the model
predicted variances and covariances of the observable variables $\Sigma$ and the sample variance and covariances of those variables $S$, that is:

$$F_{ML} = \log|\Sigma| + \text{tr}(\Sigma^{-1}) - \log|S| - (p + q) \quad \text{(Bollen, 1989)}$$

where

$$\Sigma = \text{var}\left(\begin{bmatrix} y \\ x \end{bmatrix}\right) = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx} & \Sigma_{xx} \end{bmatrix}$$

and:

i. $\Sigma_{yy} = \Lambda_y \text{cov}(\eta) \Lambda_y' + \Theta_x$;
ii. $\Sigma_{yx} = \Lambda_y \Phi \Lambda_x' + \Theta_x$;
iii. $\Sigma_{yx} = \Lambda_x \text{cov}(\eta, \xi) \Lambda_x' = \Sigma_{yx}'$;
iv. $\text{cov}(\eta) = (I - B)^{-1}(\Gamma \Phi \Gamma + \Psi)(I - B)^{-1}$;
v. $\text{cov}(\eta, \xi) = (I - B)^{-1} \Gamma \Phi$.

4. Empirical results and interpretation

For this paper, estimation of LISREL is accomplished by using MI methodology to solve missing data problem. In fact, the primary interest of this paper lies in evaluating the effect of missing data on LISREL used to measure sustainable development. In other words, the principal goal of the empirical study is to test the correctness of multiple imputation method on the measure of the latent variable and to verify the hypothetical causality relations between observable variables. This is obtained making a comparison of LISREL between parameters estimates obtained by using the classical practice of ignoring records with incomplete information and the estimates obtained by using a multiple imputation procedure to substitute any missing item with values which can be considered plausible according to the multivariate distribution of the data set.

The variables included in the analysis and their rates of missingness are reported in table 1. Notice that EDUC, DIOXI, NDS and GS are missing for approximately one-third of the subjects. Most of other variables have a rate of missingness of about 8%.

All variables are thought to be strong predictors of SD missingness and are included in the imputation procedure. In order to generate imputations for the missing values, it is imposed that the SD complete data (observed and missing values) are: 1) missing at random (MAR) and 2) multivariate normally distributed.

Because the missingness for these variables is uncontrolled, there is no way to support or discredit this hypothesis directly. Therefore, we suppose that applying the MAR assumption to these missing values is not
unreasonable. That is, variables that are heavily skewed have a low rate of missingness (about 8%) and including them under normality assumption will produce a little distortion of their distributional shapes. Anyway MI tends to be quite forgiving of departures from normality assumption. So it makes sense to use a normal model to create MI even when the observed data are not normal (Schafer, 1997).

The MI’s steps that we have made are:

a) create imputations;
b) analyze completed data sets;
c) combine the results.

To create imputations we use the MCMC parametric approach here. The proposed methodology extends the LISREL by imputing 3000 values of SD dataset for each missing value. Following this approach, the LISREL parameters are derived separately from these 3000 imputations. These 3000 values are combined using simple computational rules (Rubin, 1987) to provide valid statistical inferences. Table 4.2 shows estimated coefficients and their standard errors obtained in two LISREL models: the first LISREL ignores subjects with incomplete information while second combines inferences from imputed data sets. The method of multiple imputation, analysis and diagnostics based on LISREL achieves the goal of improving the effectiveness of obtaining a SD measure.

It can be noticed that:

1. the size of estimated coefficients are highly different in the two models because of incomplete information (table 4.2);
2. the statistical signification of a coefficient is function of information lost by missing data (for example, the results that DIOXI and PPP appear to be not statistically significant in the first model come from a rate of missing of 33% for DIOXI and of 8% for PPP) (table 4.1);
3. missing values can be a source of variability for the sign of path coefficients (for example, in the first LISREL the negative sign of LIFEXPE coefficient is not acceptable in SD theory).

Estimates that are derived combing MI use the information from all 3000 replications and incorporate imputation variability. Results confirm that variables that have the greater effects on SD explanation are: adult literacy rate (ADULITE), carbon dioxide damage (DIOXI) and genuine saving (GS).
Tab. 4.1 - Missingness rates of variables that are used in the SD’s LISREL

<table>
<thead>
<tr>
<th>variable</th>
<th>Description</th>
<th>% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS</td>
<td>Genuine domestic saving</td>
<td>36,31579</td>
</tr>
<tr>
<td>NDS</td>
<td>Net domestic savings</td>
<td>35,78947</td>
</tr>
<tr>
<td>DIOXI</td>
<td>Carbon dioxide damage</td>
<td>33,15789</td>
</tr>
<tr>
<td>EDUC</td>
<td>Education expenditure</td>
<td>30,52632</td>
</tr>
<tr>
<td>ENER</td>
<td>Energy depletion</td>
<td>24,21053</td>
</tr>
<tr>
<td>MINER</td>
<td>Mineral depletion</td>
<td>24,21053</td>
</tr>
<tr>
<td>FOREST</td>
<td>Net forest depletion</td>
<td>24,21053</td>
</tr>
<tr>
<td>HDI</td>
<td>Human dev. index</td>
<td>8,42105</td>
</tr>
<tr>
<td>ADULITE</td>
<td>Adult literacy rate</td>
<td>8,42105</td>
</tr>
<tr>
<td>PPP</td>
<td>GDP per capita</td>
<td>8,42105</td>
</tr>
<tr>
<td>LIFEXPE</td>
<td>Life expectation at birth</td>
<td>7,89474</td>
</tr>
<tr>
<td>COENROL</td>
<td>Level gross enrolment ratio</td>
<td>7,89474</td>
</tr>
</tbody>
</table>

These effects are measured by \( \gamma \) parameters that are like to multiple regressions’ standardised coefficients *betas*. For example, if DIOXI increases one unit, while all variables are held fixed, the expected decrease of SD is 1,56 units on average.

The remaining variables have relevant impacts but less than expected (for example HDI in comparison to GS). MINER, FOREST and ENER are not statistically significant. The sign of all estimated coefficients in SD equations has the sign predicted by the SD theory except that LIFEXPE, EDUC, NDS and GDP. The negative sign of these estimated coefficients are more difficult to rationalize.

In particular, ADULITE and COENROL values have a significant positive impact on SD. However the most negative important impact on SD is PPP (-1.86).

Of course it has to be careful for the interpretation of these path coefficients due to a multicollinearity problem.
Tab. 4.2 - Parameter’s estimates with and without multiple imputation (MI)

<table>
<thead>
<tr>
<th>parameter</th>
<th>no MI</th>
<th>combing MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(standard error)</td>
<td>(standard error)</td>
</tr>
<tr>
<td>( \lambda_1 )</td>
<td>0.00244 (0.00051)</td>
<td>0.00262 (0.00031)</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td>0.55913 (0.02869)</td>
<td>0.61965 (0.03673)</td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>-0.00017 (0.00011)</td>
<td>-0.00025 (0.00012)</td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>1.78862 (0.10599)</td>
<td>1.62382 (0.11135)</td>
</tr>
<tr>
<td>( \gamma_3 )</td>
<td>1.93897 (0.35393)</td>
<td>1.69791 (0.30358)</td>
</tr>
<tr>
<td>( \gamma_4 )</td>
<td>-0.89419 (0.70503)</td>
<td>-1.86481 (0.48112)</td>
</tr>
<tr>
<td>( \gamma_5 )</td>
<td>-1.92140 (0.37686)</td>
<td>-1.40226 (0.23848)</td>
</tr>
<tr>
<td>( \gamma_6 )</td>
<td>-1.93861 (0.24775)</td>
<td>-1.47581 (0.16937)</td>
</tr>
<tr>
<td>( \gamma_7 )</td>
<td>-1.90504 (0.12695)</td>
<td>-1.56045 (0.09884)</td>
</tr>
<tr>
<td>( \gamma_8 )</td>
<td>-0.01667 (0.08747)</td>
<td>0.06830 (0.05747)</td>
</tr>
<tr>
<td>( \gamma_9 )</td>
<td>-0.06717 (0.04995)</td>
<td>-0.02016 (0.03087)</td>
</tr>
<tr>
<td>( \gamma_{10} )</td>
<td>2.09042 (5.03913)</td>
<td>3.64727 (4.04977)</td>
</tr>
<tr>
<td>theta</td>
<td>0.01702 (0.00175)</td>
<td></td>
</tr>
<tr>
<td>psi</td>
<td>35.30247 (0.14724)</td>
<td></td>
</tr>
</tbody>
</table>

Some aspects of model adequacy can be addressed looking at the plots of imputed data in figure 4.1. For each plot, the parameters’ estimates without multiple imputations are compared to those that are obtained from all 3000 replications. It is very important to notice that a complete data matrix is systematically generated drawing from a multivariate normal distribution. Parameters obtained by excluding from the initial data set all that records with have at least a missing value are indicated by a vertical line. In all cases, MI reproduces the parameter values most correctly.

This is probably due to the high efficiency and more complete use of the information in the data set by imputing missing data with all the variables in the analysis. Estimation bias for all parameters thus appears to be not negligible under listwise deletion. Because the present data are MAR, estimation bias is of central concern with all estimates.

Although the distributions indicate that the MI estimates are more precise in the full data set analysed using MI, some differences in efficiency of estimation is consequence of a different amount of missing information (table 4.1). When the proportion of cases with imputed values is high, as for net domestic savings and genuine domestic saving, incorporating information on variability due to imputation is likely to have large effects relative to when
only a moderate proportion (i.e. energy depletion) or small proportion (i.e. life expectation at birth) of cases have imputed values.

These findings corroborate the utility of MI: LISREL appear more appropriate when the complete data sets (adjusted by missing values) is analysed than when the uncertainty due to missing data is ignored.

5. Conclusions

In this paper we have provided an empirical way to measure sustainable development when there is a missing data problem.

In order to do so, we used a multiple imputation technique with a model of structural equation.

The indicators that are commonly used to measure sustainable development contain a lot of missing data, for some indicators the percentage of missing values is up to 30% and this problem affect significantly the estimated coefficients.

In order to understand the influence of missing data we compared the estimation of the structural model with missing values with the estimation of the same model where the missing values are filled with a multiple imputation methodology.

The main result of our estimation is that multiple imputation method achieves the goal of improving the SD measure since it reduces the variability and improve the statistical signification of the estimated coefficients.
Fig. 4.1 - Kernel Density estimates of the distribution of the 3000 Multiple Imputed LISREL parameters
Fig. 4.1 continued
REFERENCES


EVERITT, B.S. (1993), An introduction to latent variable models, Chapman and Hall, N.Y.


UNITED NATIONS DEVELOPMENT PROGRAM, The Human Development Report, various years.


WORLD BANK (1997), Expanding the Measure of Wealth: Indicators of Environmentally Sustainable Development. Washington D.C.