

NORM software review: handling missing values with multiple imputation methods¹

I Gusti Ngurah Darmawan

Evaluation studies often lack sophistication in their statistical analyses, particularly where there are small data sets or missing data. Until recently, the methods used for analysing incomplete data focused on removing the missing values, either by deleting records with incomplete information or by substituting the missing values with estimated mean scores. These methods, though simple to implement, are problematic. However, recent advances in theoretical and computational statistics have led to more flexible techniques with sound statistical bases. These procedures involve multiple imputation (MI), a technique in which the missing values are replaced by $m > 1$ estimated values, where m is typically small (e.g. 3–10). Each of the resultant m data sets is then analysed by standard methods, and the results are combined to produce estimates and confidence intervals that incorporate missing data uncertainty. This paper reviews the key ideas of multiple imputation, discusses the currently available software programs relevant to evaluation studies, and demonstrates their use with data from a study of the adoption and implementation of information technology in Bali, Indonesia.

Introduction

Missing observations occur in many areas of research and evaluation (Kline 1998). When data are collected by surveys, questionnaire responses may be incomplete because some respondents refuse to answer certain questions. In longitudinal studies, subjects may drop out early or be unavailable during one or more data collection periods. These types of missing data are unintended and uncontrolled by the researcher but the overall result is that useful data collected from a survey cannot be analysed in detail because of the extent of missing data. This paper introduces and discusses the three ad hoc methods for dealing with missing data and then focuses on the software developed to process missing data by the multiple imputation method.

Taxonomy of missing data methods

Little and Rubin (1987, 1990) contend that, with standard statistical techniques, there are basically three methods to handle multivariate data with missing values: (1) complete case analysis (listwise deletion), (2) available case methods (pairwise deletion), and (3) filling in the missing values with estimated scores (imputation).

Some advantages of the complete case approach are: (a) simplicity, since standard analysis can be applied without modification, and (b) comparability of univariate statistics,

I Gusti Ngurah Darmawan is a PhD student at Flinders Institute of Public Policy and Management, Flinders University of South Australia.

FIGURE 1: TAXONOMY OF MISSING DATA METHODS^(§)

TYPE OF METHOD	AD HOC	LIKELIHOOD		MULTIPLE IMPUTATION
	└ 1970	└ EM └ 1980	└ MI └ 1990	└ 2000
Applied widely	Yes	Yes		No
Advantages	Easy	Efficient		Flexible Good standard errors
Disadvantages	Inefficient Biased	More difficult Model specific		Relatively unknown

(§) Adapted from Schafer, JL 1997b, *Introduction to multiple imputations for missing data problems*, viewed 6 May 2002, <www.stat.psu.edu/~jls/asa97/slide7.html>.

since these are all calculated with a common sample base of cases. However, there are disadvantages particularly due to the potential loss of information in discarding incomplete cases.

Pairwise deletion uses all cases where the variable of interest is present. This technique has the advantage of being simple and increases the sample size. However, its disadvantage is that the sample base changes from variable to variable according to the pattern of missing data.

Imputation methods may also be problematic. The method of mean substitution, where the imputed average on a variable-by-variable basis is used to fill in the missing data, preserves the observed sample means, but it distorts the covariance structure, biasing estimated variance and covariance toward zero. Regression substitution, imputing predicted values from regression models, on the other hand, tends to inflate observed correlations, biasing them away from zero (Schafer 1997a).

Over the last two decades, there has been substantial progress in developing software to handle statistical procedures for missing data. In the late 1970s, Dempster, Laird and Rubin (1977) (cited in Schafer & Olsen 1998) formalised the Expectation and Maximisation (EM) algorithm, a computational method for efficient estimation from

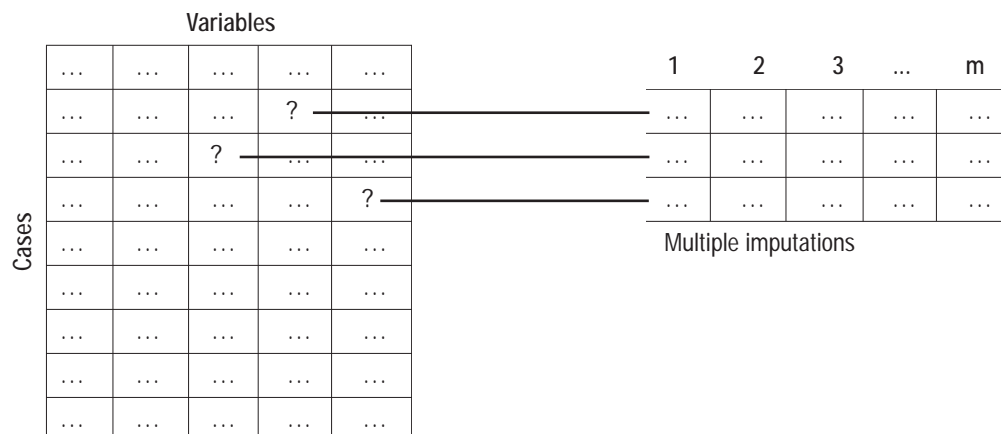
incomplete data. More recently, Rubin, arguing that an important limitation of single imputation methods is that 'standard variance formulas applied to the filled-in data systematically underestimated the variance of estimates' (Little & Rubin 1987, p.61), has proposed a procedure for multiple imputation (MI) (Rubin 1987). Rubin's MI methods allow valid estimates of the variance to be calculated using standard complete data procedures. A taxonomy of missing data methods is given in Figure 1.

Multiple imputation is a technique in which the missing values are replaced by $m > 1$ plausible values drawn from their predictive distribution. The variation among the m imputations reflects the uncertainty with which the missing values can be predicted from the observed ones. As a result, there are m complete data sets. In Rubin's method for multiple imputed inference, each of the simulated complete data sets is analysed by standard methods, and the results (estimates and standard errors) are combined to produce estimates and confidence intervals that incorporate missing data uncertainty.

In this approach, the first step is to specify one encompassing multivariate model for the entire data set. Four different classes of multivariate complete data models are available (Schafer & Olsen 1998):

- normal model, which performs multiple

FIGURE 2: MULTIPLE IMPUTATIONS FOR MISSING VALUES^(§)



(§) Adapted from *Statistical solutions* 2001, 'What is multiple imputation?', viewed 6 May 2002, <www.statsol.ie/solas/solas.htm>.

imputation under a multivariate normal distribution;

- loglinear model, which has been traditionally used by social scientists to describe associations among variables in cross-classified data;
- general location model, which combines a loglinear model for the categorical variables with a multivariate normal regression model for the continuous variables;
- two-level linear regression model, which is commonly applied to multi-level data.

It should be noted that an imputation model should be chosen to be compatible with the subsequent analyses. In particular, the model should be flexible enough to preserve the associations or relationships among variables that will be the focus of later investigation. Therefore, a flexible imputation model that preserves a large number of associations is desired because it may be used for a variety of post-implementation analyses.

In general, application of the multiple imputation technique requires three steps: *imputation*, *analysis* and *pooling*.

- **Imputation** – impute the missing entries of the incomplete data sets, not once, but m times. Imputed values are drawn for a distribution (that can be different for each missing entry). This step results in m complete data sets.
- **Analysis** – analyse each of the m completed data sets using procedures in SPSS, LISREL, AMOS (Arbuckle 1995), or virtually any other statistical package. This step results in m analyses (the repeated analysis step on the imputed data is actually somewhat simpler than the same analysis without imputation, since there is no need to bother with the missing data).
- **Pooling** – integrate the m analysis results into a final result. Simple rules exist for combining the m analyses to produce overall estimates and standard errors (Rubin 1987). This step consists of computing the mean over the m repeated analyses, its variance, and its confidence interval or P value.

Schafer (1997a) has written general purpose MI software for incomplete multivariate data. Each of the four software packages applies a different class of multivariate complete-data models. These packages are:

- **NORM**, which uses the multivariate normal distribution;
- **CAT**, for multivariate categorical data, which is based on the loglinear model;
- **MIX**, for a mixed data set containing both continuous and categorical variables, which relies on the general location model, a combination of a loglinear model for the categorical variables with multivariate normal regression for the continuous ones;
- **PAN**, for multivariate panel or clustered data, which uses a multivariate extension of a two-level linear regression model commonly applied to multi-level data.

These programs can be downloaded free of charge at the Multiple Imputation website (<http://stat.psu.edu/~jls/misoftwa.html>). The four software packages are available as functions in S-PLUS (www.insightful.com/products/default.asp) statistical software. A stand-alone version of NORM (Schafer 1999) suitable for PCs running Windows (95/98/NT) is also offered for free. Another program that can handle multiple imputation for missing values that is available on the market is SOLAS™ 3.0 (www.statsol.ie/solas/solas.htm), which is based on the work of Rubin (1987). The rest of this paper reviews NORM as it is the most readily available.

NORM and its application: information technology usage, user satisfaction, and user performance

The data in this example come from a study (Darmawan 2001) focusing on the adoption and implementation of information technology by local government in Bali, Indonesia. 153 agencies across all regions of Bali participated in this study. These agencies employed a total of 10,034 employees, of whom 1427 (approximately 14%) used information technology in their daily duties. Of these, 975 employees participated in this study. The goal of this study is to examine various potential factors that might impact the adoption and implementation processes of information technology in the context of Bali's local government.

This exercise is designed to demonstrate the use of NORM and reports only some of the relevant variables taken from the study. In particular, two issues were not addressed in this demonstration of NORM:

- The subjects in this study may not act independently of one another. The employees are grouped into agencies and an ideal analysis of these data should account for their grouped or multilevel structure. For simplicity, however, the multilevel aspect of the data is ignored here and the employees are treated as independent agents. Consequently, the estimated relationships are somewhat overstated, and that the actual statistical significance of an effect is somewhat less than reported.
- The distribution of the data. Because this data set contains both continuous and binary variables, the MIX program might be the most appropriate one. However, it is also possible to do an acceptably good job using the simpler package, NORM, which treats the variables as if they are jointly normal.

Table 1 indicates the survey variables included in this demonstration of NORM, as well as the percentage of missing data for each variable. Overall, there could be as many as 60% missing if listwise deletion is applied. Figure 3 displays histograms indicating the distribution of responses for each of the included variables. All variables, except gender and frequency of direct usage, are normally distributed.

Prior to carrying out the imputation process, it is helpful to bear in mind the methods by which the

TABLE 1: VARIABLES AND RATES OF MISSING VALUES

Name	Description	Missing (%)	Remark
SEX	Gender	0.8	1 = Male; 2 = Female
AGE	Age	29.4	Age in years
ATTITUDE	Attitude toward change	19.2	Composite score
DIRECUSE	Frequency of direct usage	1.0	Unipolar Likert Scale
SATISFAC	User satisfaction	17.9	Composite score
PERFORMA	User performance	9.2	Composite score
Overall	Listwise deletion	60.0	384 complete cases

data will ultimately be analysed. In this case, because the main goal is to access possible factors affecting user performance, a linear regression model to predict PERFORMA is constructed.

$$PERFORMA_i = \beta_0 + \beta_1 SEX_i + \beta_2 AGE_i + \beta_3 ATTITUDE_i + \beta_4 DIRECUSE_i + \beta_5 SATISFAC_i + \epsilon_i$$

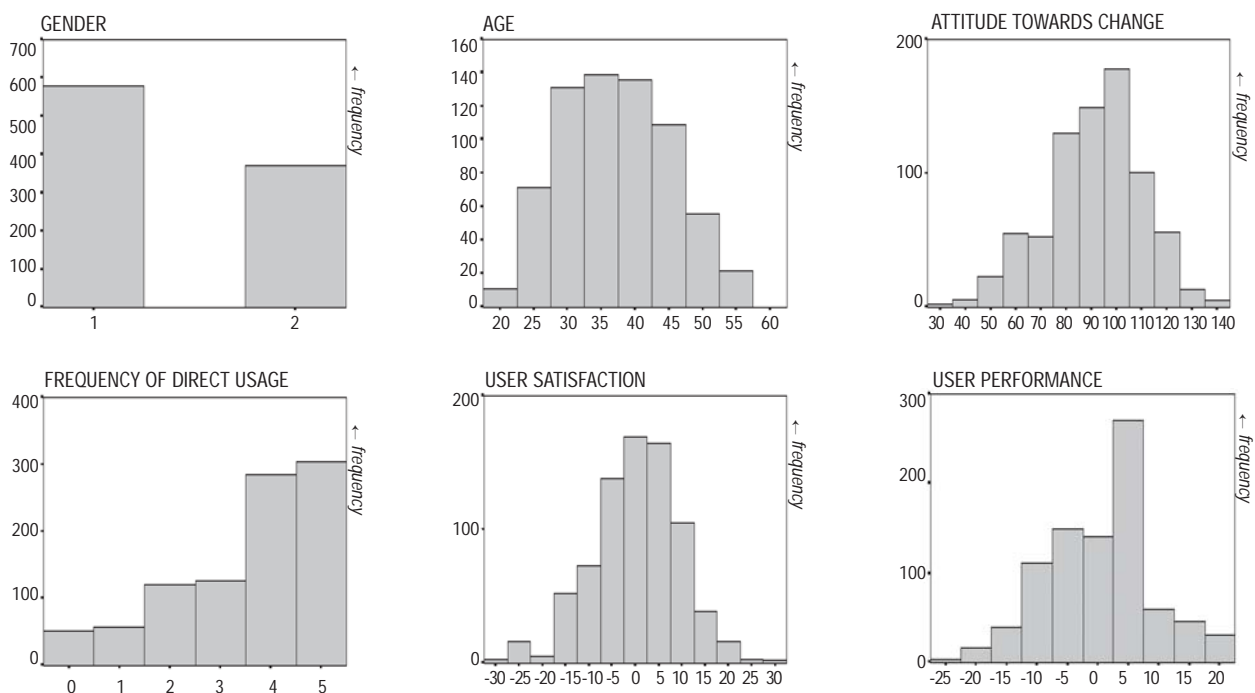
where b_0, b_1, b_2, b_3, b_4 and b_5 are constants and ϵ_i are residuals reflecting the variability of the estimates of the constants.

Figure 4 shows the functionality of the window display for NORM. The variables in the data set are managed by the 'variables' grid. On this grid it is possible to:

- edit variable names;
- select variables for the model;
- select variables to be written to imputed data sets;
- apply transformations to variables;
- round off variables to a specified precision; and
- examine a variable's distribution.

When variables are not normally distributed, it often helps to apply transformations before imputing. For example, if a variable is right-skewed it may be sensible to impute the variable on

FIGURE 3: HISTOGRAMS OF THE SIX VARIABLES



a square root or log scale, and then transform it back to the original scale. When NORM creates an imputed version of the data set (*.imp file), it automatically rounds each variable to a precision which must be specified. Rounding, perhaps in conjunction with transformations, helps to impute values that resemble the observed data.

Even though use of the NORM package presupposes that all variables are normally distributed, the binary variable SEX can be imputed under normality assumptions and then rounded off to one or two, a procedure that tends to work quite well in practice (Schafer 1997a). Similarly a transformation of the variable DIRECUSE might be considered in order to make the normality assumption more plausible. However, the extent of the missing data for these variables is extremely low (1%) so that they would almost never be imputed and including them without transformation would produce very little distortion of their distributional shapes.

Once the data have been read, NORM can display a printed summary of the observed data. This summary includes the number and per cent missing of each variable, as well as the means and standard deviations of the observed data. It also provides a matrix of 0s and 1s, which indicate missingness patterns in the data. The window in NORM, which summarises the data, is shown in Figure 5.

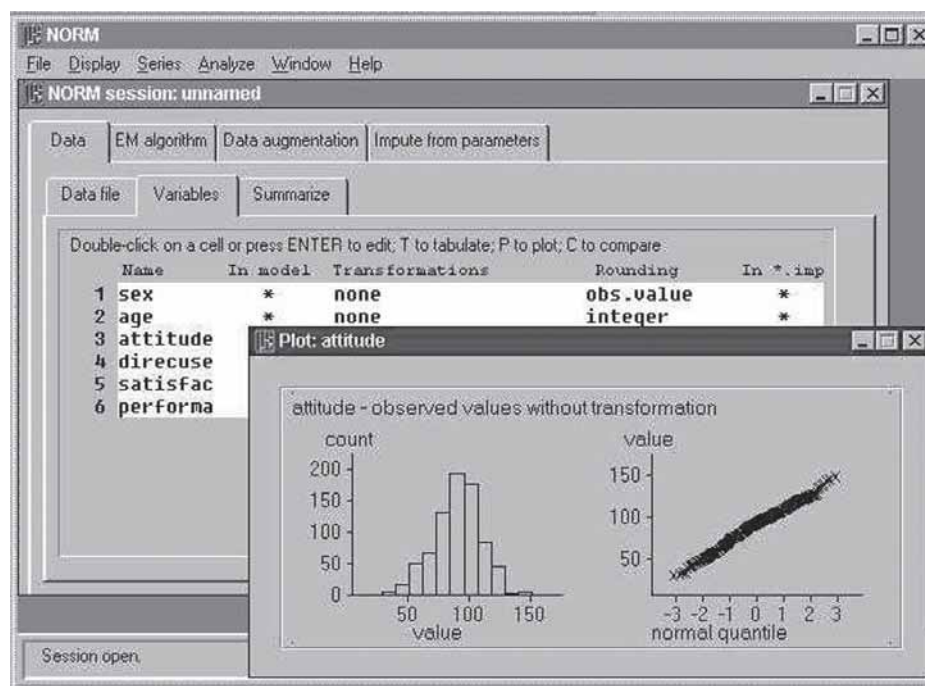
After examining the data summary, the Expectation-Maximisation procedure was run. This algorithm is a general technique for fitting models to incomplete data based on a process with two steps:

- 1 the Expectation (E) step in which missing sufficient statistics are replaced by their expected values given the observed data, using estimated values for the parameters; and
- 2 the Maximisation (M) step where the parameters are updated by their maximum-likelihood estimates, given the sufficient statistics obtained from the E-step.

The procedure is run until convergence is obtained. In this case the procedure converged in 28 iterations, which took less than 10 seconds on a 133 MHz Pentium computer. Upon convergence, NORM provided an iteration history, including the observed data likelihood, as well as the final estimates of the means, standard deviations and correlations.

Following convergence of the E-M procedure a Data Augmentation (DA) procedure was implemented. DA is an iterative process which, utilising the observed data, provides estimates of both the missing data and distributional parameters using a two-step iteration procedure:

FIGURE 4: DATA PLOT WINDOWS



- 1 the Imputation (I) step in which the missing data are imputed by drawing values from the conditional distribution, given the observed values and the parameters;
- 2 the Posterior (P) step in which new values for the parameters are imputed by drawing them from a Bayesian posterior distribution given the observed data and the most recent estimates for the missing data.

With the rapid convergence of the Expectation-Maximisation algorithm, it was estimated that 28 cycles of data augmentation (DA) would be sufficient for DA to converge. For an extra margin of safety, 250 cycles of DA were employed and an imputed data set was created at every 50th cycle, producing a total of $m = 5$ imputations. The entire data augmentation and imputation procedures took less than 30 seconds. Following imputation, a series of linear regression analyses were conducted to predict PERFORMA. The analyses were carried out in SPSS.

The final step in the analysis was to combine the coefficients and standard errors, a step carried automatically by NORM. The printed report generated by NORM, shown in Figure 6, resembles a table of coefficients produced by traditional regression software. For each coefficient, the report provides the overall estimate based on the five estimates and the associated standard errors, degrees of freedom, and p-values. The second table provides the lower and upper limits of a 95% confidence interval and the estimated per cent missing information for each coefficient.

The estimated coefficients are highly significant and the effects are in the expected direction. Performance appears to be positively associated with attitude toward change, frequency of direct usage, and the level of satisfaction, and males, as well as younger employees, seem to report a higher

impact of information technology on their performance.

In this application of NORM only 384 of 957 subjects provided complete data on all variables. If an analysis involving all these variables was performed using standard statistical packages, the built-in case deletion procedures would discard up to 60% of the subjects, see Table 1, resulting in a substantial loss of information. As a comparison, a linear regression using listwise deletion was undertaken using SPSS and the results are presented in Table 2.

Note that this analysis suggests that the effect of SEX is now insignificant. This example is in conformity with Schafer's (1997a) suggestion that when the incomplete cases comprise only a small fraction of all cases (5% or less) then listwise deletion may be a perfectly reasonable solution for the missing data problem. However, as noted by Little and Rubin (1987) in multivariate settings, where missing values occur on more than one variable, the loss in sample size can be considerable, particularly if the number of variables is large. If so, deleting cases may be unsatisfactory and problematic, causing large amounts of information to be discarded. In addition, case-deletion procedures may bias the results if the subjects who provide complete data are unrepresentative of the entire sample, as demonstrated in this case.

General comments

NORM version 2.03 runs under Windows 95, 98, and NT. The size of the file is less than 1 MB.

FIGURE 5: DATA SUMMARY

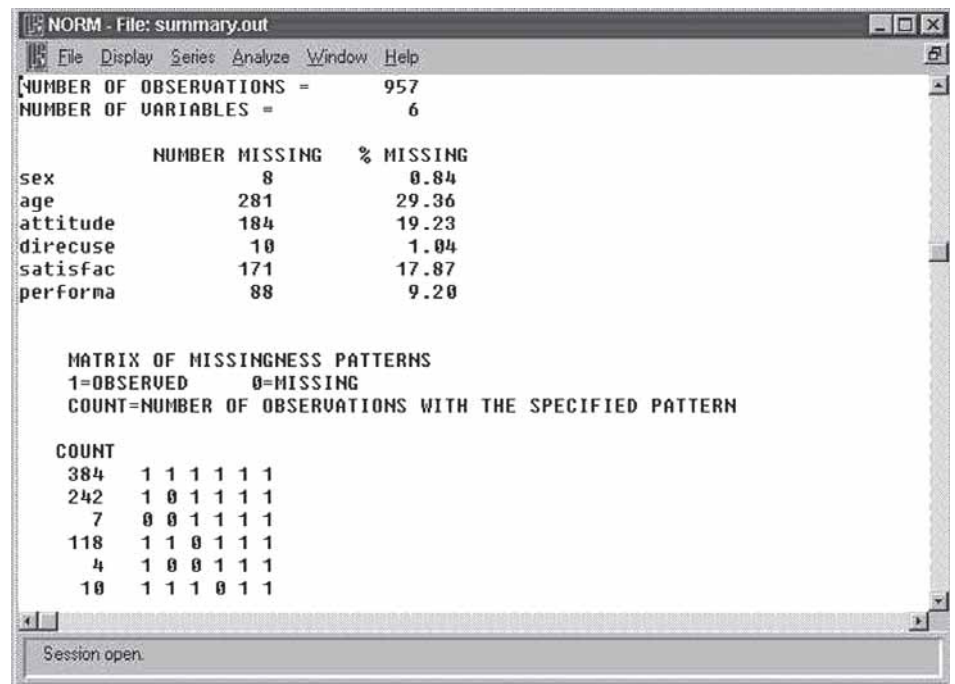
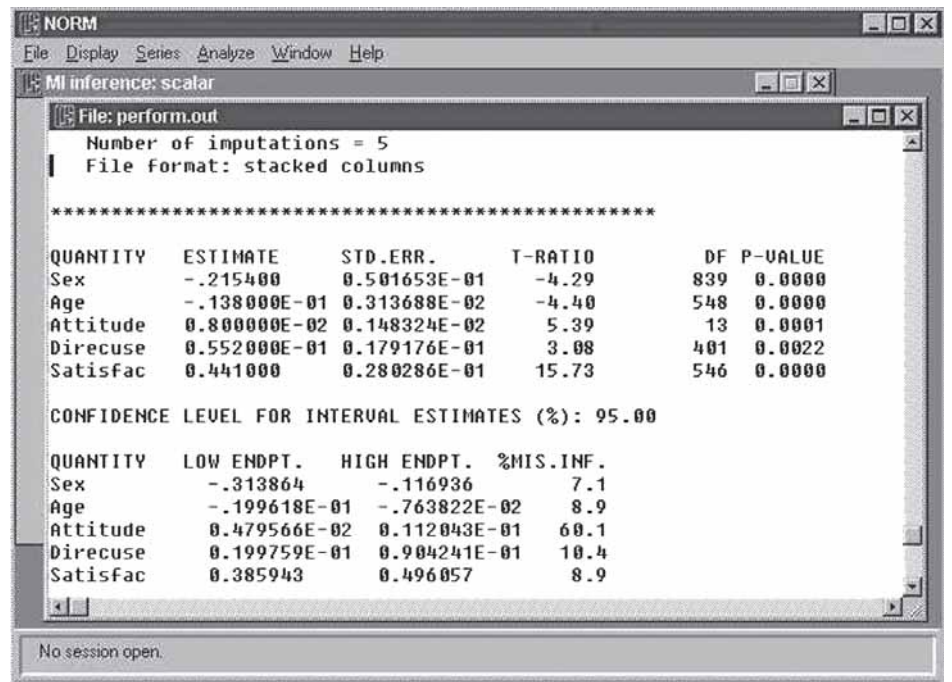


FIGURE 6: MULTIPLE INFERENCE RESULT



Once the NORM program has been downloaded and saved in the desired directory, it can be installed easily. Huge computing power is not required; a personal computer with a Pentium 100 MHz processor, 16 MB of RAM and 5-10 MB of free HDD space can produce multiple imputed data sets in a matter of minutes.

The NORM program is relatively easy to use. Once a NORM session is opened, a new window with multiple sheets will appear. The sequence of the sheets is in accordance with the sequence of procedures that have to be done with the sequence of the sheets indicating the order of required actions. The variable grid in the data sheet has very useful functions including tabulate (to view the frequency table), plot (to show the histogram and normal quantile), and compare (to compare the original data against the imputed data). These functions provide an opportunity to check data

TABLE 2: LINEAR REGRESSION RESULTS WITH LISTWISE DELETION ^(S)

	Unstandardised	Std error coefficients	Standardised coefficients	t	Sig.
(Constant)	-.548	.306		-1.787	.075
SEX	3.311E-02	.077	.018	.428	.669
AGE	-1.174E-02	.004	-.124	-2.690	.007
ATTITUDE	8.566E-03	.002	.202	4.657	.000
DIRECUSE	9.908E-02	.023	.199	4.242	.000
SATISFAC	.327	.039	.369	8.302	.000

(S) Dependent variable: *PERFORM*

distributions before deciding any transformation.

The current version of NORM can only read and produce ASCII format data. Therefore, in order to be able to use NORM, an ASCII data file should be produced beforehand. This is also true for the output and it is necessary to convert the imputed data sets into SPSS, for example, before any subsequent analyses can be undertaken. By comparison SOLAS 3.0 allows data to be imported directly from a wide variety of packages including SAS (Unix/Windows), SPSS and S-PLUS (Unix/Windows). In SOLAS 3.0, once the data are imported, the missing data pattern can be displayed and a decision upon the most appropriate technique made. Once imputation is complete the imputed data sets can be analysed within SOLAS or exported to a variety of other packages.

Conclusions

Evaluation practice needs the methodological sophistication that new statistical analysis techniques can offer. Standard programs for data analysis such as SPSS, SAS and LISREL were never intended to handle data sets with a high percentage of incomplete cases. The analysis of data sets with missing values is one area of statistics where major advances have recently been made. This can increase the strength of the findings of evaluation studies or surveys with missing data. Among these new techniques, multiple imputation is especially powerful because of its generality.

Multiple imputation seems to be an attractive way of dealing with missing values. Listwise deletion maybe a perfectly reasonable solution for the missing data problem when the incomplete cases comprise only a small fraction of all cases. However, in multivariate settings the loss in sample size can be considerable, particularly if the number of variables is large. If so, deleting cases may be unsatisfactory, causing large amounts of information to be discarded. In addition, case-deletion procedures may bias the results. MI methods accommodate the notion of uncertainty and allow valid estimates of the variance to be calculated using standard complete data procedures. MI has now become a major applied approach to the general problem of obtaining valid statistical inferences when faced with missing data, and may become the dominant approach in social science data analysis in the future. One reason for this growth of MI is its suitability for modern computing environments.

In addition, there are freeware software packages available to be downloaded. The software is quite user-friendly if the user understands the multiple imputation process. Evaluation practitioners and survey researchers could find these techniques useful for data sets where there are large number of missing values.

Acknowledgments

I wish to thank both Associate Professor Colin A Sharp and Dr Herbert Stock for their valuable time and suggestions, and Professor John P Keeves for his help in understanding multiple imputation methods.

Notes

- 1 The Editors acknowledge the technical assistance of Dr Herbert Stock, of the Flinders Institute of Public Policy and Management, in the finalisation of this paper.

References

- Arbuckle, JL 1995, *AMOS user's guide*, SmallWaters Corporation, Chicago.
- Darmawan, IGN 2001, 'Adoption and implementation of information technology in Bali's local government: a comparison between single level path analyses using PLSPATH 3.01 and AMOS 4 and multilevel path analyses using MPLUS 2.01', *International Education Journal*, vol. 2, no. 4, pp. 100-125.
- Kline, RB 1998, *Principles and practice of structural equation modeling*, The Guilford Press, New York.
- Little, RJA & Rubin, DB 1987, *Statistical analysis with missing data*, John Wiley, New York.
- Little, RJA & Rubin, DB 1990, 'The analysis of social science data with missing values', in J Fox & JS Long (eds), *Modern methods of data analysis*, Sage Publications, Newbury Park California.
- Rubin, DB 1987, *Multiple imputation for nonresponse in surveys*, J. Wiley & Sons, New York.
- Schafer, JL 1997a, *Analysis of incomplete multivariate data*, Chapman & Hall, New York.
- Schafer, JL 1997b, *Introduction to multiple imputations for missing data problems*, viewed 6 May 2002, <www.stat.psu.edu/~jls/asa97/slide7.html>.
- Schafer, JL 1999, 'NORM: multiple imputation of incomplete multivariate data under a normal model, version 2.03', software for Windows 95/98/NT, available from <www.stat.psu.edu/~jls/misoftwa.html>.
- Schafer, JL & Olsen, MK 1998, 'Multiple imputation for multivariate missing-data problems: a data analyst's perspective', *Multivariate Behavioral Research*, vol. 33, pp. 545-571.
- Statistical Solutions 2001, *What is multiple imputation?*, viewed 6 May 2002, <www.statsol.ie/solas/solas.htm>.