*CJNR 2005, Vol. 37 N° 4, 156–171*

*Résumé*

# Les données manquantes : introduction aux concepts de base à l'intention du chercheur novice

**Maher M. El-Masri et Susan M. Fox-Wasylyshyn**

Les données manquantes posent un problème fréquent en recherche; s'il n'est pas traité correctement, il peut fausser les conclusions concernant une population. Il existe un ensemble de méthodes statistiques permettant d'interpréter les données manquantes, certaines simples, et d'autres complexes, sur le plan théorique et mathématique. Le présent article propose une vue d'ensemble du problème des données manquantes à l'intention des chercheurs débutants. Les auteurs expliquent les modèles de données manquantes, discutent des questions qu'elles soulèvent et présentent certaines méthodes de traitement courantes. Parmi les techniques abordées, on compte la suppression dans la liste (*listwise suppression*), la suppression par paires (*pairwise suppression*), la substitution moyenne par cas, par échantillon ou par groupe, l'imputation par régression et la maximisation de l'estimation.

Mots clés : données manquantes, modèles de données manquantes, suppression, imputation, substitution moyenne

CJNR 2005, Vol. 37 Nº 4, 156–171

*Best Practices in Research Methods*

# Missing Data: An Introductory Conceptual Overview for the Novice Researcher

**Maher M. El-Masri and Susan M. Fox-Wasylyshyn**

Missing data is a common issue in research that, if improperly handled, can lead to inaccurate conclusions about populations. A variety of statistical techniques are available to treat missing data. Some of these are simple while others are conceptually and mathematically complex. The purpose of this paper is to provide the novice researcher with an introductory conceptual overview of the issue of missing data. The authors discuss patterns of missing data, common missing-data handling techniques, and issues associated with missing data. Techniques discussed include listwise deletion, pairwise deletion, case mean substitution, sample mean substitution, group mean substitution, regression imputation, and estimation maximization.

Key words: missing data, patterns of missingness, deletion, imputation, case mean substitution, group mean substitution

## Introduction

Missing data is a common issue in research, and it can lead to inaccurate conclusions about populations if improperly handled. Missing data is a problem because analysis of incomplete or improperly imputed data sets threatens the external validity of the findings by yielding non-generalizable results. The problem of missing data is often attributed to either design issues or extraneous factors (Kline, 1998). Missing data attributed to design issues is often intentional, as when the investigator administers only a section of a long questionnaire due to time constraints, or when an inexpensive measure is used for the whole sample and a more expensive measure is used with a randomly selected smaller group. However, undesirable design-related "missingness" can also be attributed to preventable factors such as lengthy questionnaires, unclear instructions, and the use of high-level language. Missing data attributed to extraneous factors relate specifically to the respondent, and are often beyond the control of the investigator.

 

A variety of statistical techniques are available to treat missing data. Some of these techniques are simple while others are conceptually and mathematically complex. The purpose of this paper is to provide the novice researcher with a conceptual overview of the issue of missing data. The focus of this introductory paper will be *patterns of missing data* and simple *missing data handling techniques* such as sample mean substitution, group mean substitution, case mean substitution, pairwise deletion, listwise deletion, regression imputation, and estimation maximization. Techniques such as hot-deck imputation, maximum likelihood, and multiple imputation are relatively complex and are not readily available in traditional statistical software packages. Therefore, these techniques are beyond the scope of this paper. With the exception of mean substitution, the techniques described in this paper are appropriate for treating missing data measured at nominal, ordinal, and interval levels. Sample and group mean substitution can be applied to treat missingness only in variables that are measured at the interval level. However, case mean substitution can be used to impute ordinal missing data such as when item values are missing in a psychometric Likert-type scale.

## The Issue

The issue of missing data is not a trivial one. The majority of statistical analyses can be conducted only on complete data sets (Allison, 2000; Rubin, 1987) — that is, cases with missing data on even one variable will be dropped from computer analyses. This leads to reduced sample size, compromises statistical power, and could affect the accuracy of parameter estimates (Patrician, 2002). Listwise deletion of missing data affects statistical power in two ways. First, in multivariate analysis, deleting a relatively large number of cases that have missing data on a given variable may mask the true relationship between this variable and the remaining variables, which could render the whole analysis invalid (Patrician). Second, deletion of a large number of cases with missing values on one or more variables may lead to a significant reduction in sample size, thus compromising statistical power (Patrician; Roth & Switzer, 1995). When data are missing in a systematic pattern, it is assumed that there are differences between respondents and non-respondents with regard to the variables on which data are missing. This is because systematic missing data are often the result of respondents' choosing to withhold certain types of information. The inability to account for systematically missed data in deletion procedures leads to misrepresentation of the true characteristics of the sample. Therefore, limiting the analysis to cases with complete information may lead to non-response bias, and may produce inaccurate parameter estimates (Barnard & Meng, 1999; Tabachnick &

Fidell, 2001), which ultimately limit the generalizability of the findings (Cohen & Cohen, 1983; Huisman, 1998).

### *Extent of Missing Data*

When faced with missing data, the researcher should first determine the extent and pattern of missingness (Kline, 1998; Tabachnick & Fidell, 2001). Several authors recommend deleting variables, rather than cases, when the amount of missing data on the variables is large. Tabachnick and Fidell suggest that if missing values are limited to a few variables and those variables are not critical to the analysis or are highly correlated with other complete variables, it is best to delete these variables from the analysis, as they may not carry any clinically significant data. Although several authors recommend deleting a variable with a large amount of missing data, there is no consensus among them with regard to what constitutes a large amount of missingness. Cohen and Cohen (1983) suggest that up to 10% missing data on a variable is not large and that the variable should therefore be retained for analysis. Raymond and Roberts (1987) recommend a more liberal estimate, suggesting that a variable should be deleted when 40% or more of the data are missing. Tabachnick and Fidell and Kline suggest that the pattern of missing data is more important than the extent of missingness. Tabachnick and Fidell classify patterns of missing data as either random or systematic (also known as non-random or non-ignorable), and suggest that systematic missing data pose a greater threat to the generalizability of findings than randomly missing data.

### *Patterns of Missing Data*

Treatment of missing data is dependent on the pattern of missingness, which essentially determines the potential generalizability of research findings. It is therefore important that the investigator determine the pattern of missingness prior to deciding which missing data technique will be employed. The patterns of missing data can be classified into three categories: missing completely at random (MCAR), missing at random (MAR), and systematic (Heitjan, 1997; Kline, 1998; Patrician, 2002). By identifying the pattern of missingness, investigators can better determine the probability that missing data are dependent on the values of available data (i.e., observed values).

When the probability of missing data on one variable is independent of the values of that variable and of the values of the other variables in the data set, the data are assumed to be MCAR (Heitjan, 1997; Patrician, 2002). Suppose, for example, that a group of obese women are enrolled in a study to examine the impact of weight reduction on self-esteem. During the first session, participants are classified as having borderline,

moderate, or severe obesity. The pattern of missingness is assumed to be MCAR if follow-up data on weight reduction are missing only because some participants could not attend a given session for reasons such as illness or inability to secure transportation. This is because the missing value is not related to the participants' weight loss or to any other variables in the data set. However, if the probability of non-response is independent of the participant's weight loss but is related to the values of one or more of the other variables in the data set, the data are considered to be MAR (Kline, 1998; Little & Rubin, 1987). In other words, when data are MAR, missingness is not attributed to the value of the variable on which data are missing but is related to values of other variable(s) in the data set (University of Texas Statistical Services, 2000). For example, the pattern of missingness would be MAR if participants who were diagnosed as severely obese in the first session decide not to attend a follow-up session because they are embarrassed at being the largest members of the group, regardless of whether they have lost weight. In this case, the missingness is not related to the weight loss itself but is related to the initial classification of severe obesity. Both MCAR and MAR assume that missing data are not related to participants' true scores on the variable with missing data. However, the definition of MCAR carries a stronger assumption that missing data are truly random (Kline; Patrician).

If there is a probability that missing data on a variable are dependent on the value of the missing variable itself, then the pattern of missingness is said to be systematic. In this case, the missing data are dependent not on other variables in the data set but on the missing value itself (Heitjan, 1997). Using the aforementioned example, if participants who had no weight loss decided not to attend the second session because they did not see a benefit in participating, then missingness on follow-up measures of weight reduction is related to the missing value itself (weight loss) and is said to be systematic. This is because the missing value is explainable *only* by the variable on which the data are missing (weight loss) (University of Texas Statistical Services, 2000). Unlike in the case of random patterns of missing data, systematically missing data are not due to chance, but are intentional. Eliminating cases with missing data may result in non-generalizable findings when missingness is systematic (Kline, 1998).

### *Determining the Pattern of Missing Data*

Knowledge concerning the pattern of missing data helps the investigator to determine the most appropriate approach to dealing with missingness. This is especially important when the pattern is systematic, because, if not treated appropriately, such a pattern tends to yield biased parameter estimates and invalid results. However, the process of determining the

pattern of missing data can be a difficult one. Although there are several techniques available to help the researcher examine the data in order to determine the nature of the missingness (Cohen & Cohen, 1983; Orme & Reis, 1991), frequently the data provide little information that can be used in identifying patterns (Heitjan, 1997; Rubin, 1987). Although these techniques allow investigators to rule out MCAR, they cannot confirm that this is the cause of missingness. Assume, for example, that a researcher examines the data and finds no systematic explanation for missingness in a data set. In this case, there may be a tendency to infer that the missing data are MCAR. However, it is quite possible that the values of missing items are related only to the values themselves, or to other variables that are not included in the data set. These two possibilities cannot be tested because they are unknown and/or inaccessible to the researcher (Huisman, 1998). In addition, direct testing for the MAR assumption is not possible because investigators have no access to the missing values (Allison, 2000). Investigators are thus advised to make every effort to prevent the occurrence of missing data and to familiarize themselves with their study population such that they can anticipate the types of respondents who will omit certain data and develop strategies to facilitate prediction of the missing values (Kline, 1998; Patrician, 2002). For instance, respondents' postal codes may serve as a proxy to help the investigator estimate the social class or income of respondents who did not provide data on those variables.

Although it is tempting to assume that missing data are attributable to random factors and that they will have no impact on the generalizability of the findings, the researcher should test this assumption, especially if the amount of missing data is large. One way to examine data for evidence of systematic missingness is to create a missing data dummy variable that will be treated as the dependent variable in a predictive logistic regression model that includes the remaining variables in the data set as independent variables, to determine which of the variables predicts the presence of missing data on that variable (Acock, 1997; Hair, Anderson, Tathman, & Black, 1998; Huisman, 1998; Little & Rubin, 1987; University of Texas Statistical Services, 2000). Thus, other variables in the data set can be used to explain missing data and to provide information that can be used to mitigate the bias caused by missing data and to identify the pattern of missingness. Suppose, for example, that a study on job satisfaction collected data on level of education, type of profession, age, and income, but some respondents failed to report their income. The investigator may decide to create a missing data variable (coded as reported income = 1, no report of income = 0), which can be entered as the dependent variable in a logistic predictive regression model to examine whether this variable can be predicted by level of

education, profession, and/or age. If the presence of missing data on this variable is predicted by other variables, then data cannot be MCAR. When this approach to determining the pattern of missing data is used, the missing data variable (i.e., report vs. no report of income) could be included in the main analysis of job satisfaction, because it may provide important information concerning the attributes of the respondents in relation to their income and its relation to job satisfaction.

A second method of examining the pattern of missingness involves computation of *t*-tests to compare respondents and non-respondents on an item or measure. In this method, the sample is split into into two groups, those who responded to the variable in question and those who did not. Differences in the means of the observed values of the other measures in the data set are then tested (Acock, 1997; Huisman, 1998). Significant differences between respondents and non-respondents with respect to other observed variables indicate that the data cannot be MCAR (Huisman). In this approach, however, the sample size must be considered, because statistical significance is very sensitive to sample size. The absence of statistical difference with a small sample size might not necessarily mean that missingness was random.

In a third approach to examining the pattern of missingness, the missing data are incorporated as an independent dummy variable into a multiple regression model (Orme & Reis, 1991). The missing data variable is entered hierarchically into the regression equation such that the complete observations are entered in step one and the missing data variable is entered in step two. This approach partials out the effect of variables with complete observations from the relationship between the missing data variable and the outcome variable. In general, the degree of association between the missing data variable and the dependent variable indicates the degree to which data are missing on a non-random basis in relation to the dependent variable (Orme & Reis). However, Orme and Reis caution that a zero correlation between a missing data variable and the dependent variable indicates only that the missing data are unrelated to the dependent variable; it does not indicate that the obtained values on the predictor variables are a random subset of the sampled values. This approach thus provides a way to rule out other possibilities, but it cannot confirm the assumption that data are MCAR. Its main advantage is that it allows the investigator to examine the pattern of missing data while at the same time treating missing data. In addition, it allows for inclusion of the entire sample in the data analysis, and thus preserves statistical power, which may be compromised if a large number of respondents with missing data were to be eliminated. Finally, it reduces bias in the parameter estimates (Orme & Reis) that may result from deletion if missingness is systematic. Incorporating missing data as a dummy code in the

analysis, as described above, is not recommended when it results in a severely uneven split (90-10) between the two levels of the variable (reported vs. not reported). This is because the variance of the missing data variable will be quite small, which will constrain its correlation with other variables. In addition, when the same respondents have missing data on more than one variable, this approach may yield high correlations (multicollinearity) with the other missing data variables. Multicollinearity among the dummy variables for missing data may lead to data redundancy, which can subsequently impede any meaningful interpretation of the possible causes of missing data.

## Techniques for Handling Missing Data

The techniques for handling missing data can be classified into deletion techniques and imputation techniques (Kline, 1998; Little & Rubin, 1987). With deletion techniques, cases with missing data are excluded from statistical calculations. With imputation techniques, in contrast, an estimate of each missing datum is calculated and the missing data points are replaced, or *imputed*, by their estimates. In the imputation techniques discussed in this paper, each missing datum is replaced with a single estimate. A more complex imputation procedure that is beyond the scope of this paper is *multiple imputation,* which involves the creation of multiple estimates of each data point. The choice of missing data handling technique can affect the amount of dispersion around true scores, and therefore affect the degree of bias in the final results (Roth & Switzer, 1995). Thus, the choice should be based on the amount and pattern of missing data.

### *Deletion Techniques*

***Listwise deletion*** eliminates a case when any of its variables or items has a missing data point, regardless of whether that particular data point is being used in the analysis (Kline, 1998; Patrician, 2002; Tabachnick & Fidell, 2001). In other words, it restricts the analysis to those cases with complete data. To illustrate, assume an investigator wishes to conduct an analysis using the variables self-care, self-care agency, health, and well-being. Listwise deletion would result in elimination of an entire case if it has missing data on any of these variables, regardless of whether the variable was used in the analysis. This strategy is the default function on many statistical programs, such as SPSS and SAS. The primary advantage of listwise deletion is that it allows for all analyses to be conducted on the same number of cases (Kline), and not on an overlap of different samples, as is the case with pairwise deletion. However, deletion of all cases with missing data may result in the loss of a large number of cases. Hence, one

of the main criticisms of listwise deletion is that the reduction in sample size can substantially diminish statistical power (Kline; Little & Rubin, 1987; Raymond & Roberts, 1987; Roth, 1994; Tabachnick & Fidell).

Another problem associated with listwise deletion is bias. Listwise deletion assumes that data are MCAR. If data are MCAR, deleting cases with missing data does not pose a problem with bias, because the remaining cases with complete data are essentially a random subsample of the original sample (Tabachnick & Fidell, 2001) and will result in unbiased population values (Little & Rubin, 1987). However, when data are not MCAR, listwise deletion may inflate or deflate parameter estimates and lead to biased results. This is because respondents with missing data are likely to be different in some way from respondents who provide complete information. Therefore, respondents contributing to statistical analyses may be unrepresentative of the target population (Little & Rubin; Patrician, 2002; Schafer & Olsen, 1998).

When data are MCAR, listwise deletion often yields unbiased parameter estimates but may result in larger standard errors due to the decrease in sample size (Patrician, 2002). Thus, listwise deletion should be used only when the amount of missing data is small (Roth & Switzer, 1999) and is assumed to be randomly scattered (Tabachnick & Fidell, 2001). Hertel (1976) recommends that listwise deletion not be used if it leads to loss of more than 15% of cases. However, Roth (1994) considers listwise deletion to be appropriate only if less than 5% of the data are missing and if the data are MCAR.

***Pairwise deletion,*** also known as *available case analysis,* eliminates a case only when that case has missing data on the variables that are under analysis. However, that case will be included in other analyses that do not involve the variables on which data are missing (Roth, 1994). Using the example discussed earlier, if a case was missing a score on the variable *self-care,* it would be excluded from analyses involving *self-care*. However, the case could still contribute data towards analyses that involve other variables in the data set such as *self-care agency, health,* and *well-being* — if the case had no missing data for any of these variables. Thus, although pairwise deletion results in loss of data, it preserves sample size and statistical power (Roth, 1994; Tabachnick & Fidell, 2001).

Pairwise deletion is based on the assumption that estimates of linear models are functions of the first and second moments (i.e., mean and standard deviation) of any pair of variables. According to this assumption, either of these moments can be estimated using all cases with complete data on each variable or pair of variables (Allison, 2003). Thus, pairwise deletion involves the creation of a correlation matrix in which each correlation is calculated using only those cases that contain complete data points for both variables being correlated. Multiple regression analysis

could then be computed on the resulting correlation matrix of nonmissing data (Orme & Reis, 1991; Patrician, 2002). Given that cases with missing data contribute to the calculation of some correlations but not to others, pairwise deletion produces a correlation matrix with correlations that are based on slightly different subjects and/or different numbers of subjects (Cohen & Cohen, 1983; Patrician). Thus, pairwise deletion may result in a series of analyses that represent different overlapping samples that may be representative of slightly different populations. This problem often complicates interpretation of correlations and somewhat impedes generalization to a specific population (Raymond & Roberts, 1987). In addition, it is difficult to determine the appropriate sample size on which to base reporting of statistical tests. An additional problem with pairwise deletion is that it can result in mutually inconsistent correlations that would be impossible to obtain with a complete data set (Cohen & Cohen). With complete data, the correlation between any two variables is constrained by their correlation with a third variable. This constraint may not hold true when pairwise deletion is used. Further, use of mutually inconsistent bivariate correlations can yield multiple regression coefficients that are less than zero or greater than one, both of which are theoretically impossible (Cohen & Cohen).

Allison (2000) suggests that when data are MCAR, pairwise deletion yields unbiased parameter estimates of sample means, variances, and correlation coefficients, because available pairs of scores are a random subset of the pairs of scores for the entire sample. Roth (1994) indicates that if data are MCAR, pairwise deletion is an appropriate technique if the proportion of missing data does not exceed 20%.

### *Imputation Strategies*

Imputation entails the calculation of an estimate of each missing datum based on the values of other variables or the making of a reasonable guess to complete the data set. Data analysis is then carried out on a complete data set that includes both actual and imputed data (Little & Rubin, 1987). In general, imputation strategies are superior to deletion strategies, because they retain sample size and therefore maintain statistical power. In addition, some imputation strategies do not require that data be MCAR, an assumption that is often difficult to confirm (Raymond, 1986).

***Case mean substitution*** entails the replacing of a missing data point with the mean for that case on the items that have complete data for that case (Raymond, 1986). It is applicable for missing data on psychometric measures in which all items are indicators of a higher-level abstract concept, because psychometric measures are deliberately constructed such that each item is correlated with the remaining items in the

measure. Hence, case mean substitution is based on the assumption that, for any given case, the score for one item is closely related to the scores of the remaining items. The main advantage of case mean substitution is that it acknowledges differences across respondents by using data provided by the individual to estimate missing data for that individual, rather than using data provided by other respondents. For example, assume there are missing data on three items of a 20-item psychometric instrument that measures depression via Likert-type items; in this scenario, the mean of the 17 remaining items would be calculated and assigned to each of the missing values for that case.

Roth, Switzer, and Switzer (1999) examined the impact of several imputation techniques (listwise deletion, case mean substitution, item mean substitution, and regression) on correlation and regression coefficients using data sets that had missing data on 20% of the items, in both random and systematic patterns. They conclude that case mean substitution is the most robust approach to handling missing data in psychometric measures. Further, Downey and King (1998) found that, when data were MCAR, case mean substitution reproduced a fairly robust alpha if up to 30% of the items were missing, but found about 5% inflation in the alpha when 70% of the items were missing. In addition, they report that correlations between true and estimated scores were greater than 0.95 when (a) the number of missing items did not exceed 60%, or (b) the number of respondents with missing data did not exceed 15%. These findings suggest that case mean substitution is a robust imputation technique for psychometric data as long as the extent of missingness does not exceed 30%.

***Sample mean substitution*** is one of the most commonly used imputation techniques (Acock, 1997; Raymond & Roberts, 1987). It entails the substitution of the missing value on a variable with the sample mean of available data for that variable (Acock; Kline, 1998; Tabachnick & Fidell, 2001). For example, a missing score for a case on the variable *self-care* would be assigned the sample mean value of *self-care* that was obtained from all other cases that provided scores on this variable. This approach assumes that the best guess of a score for a normally distributed variable is the mean (Acock). It also assumes that missing and available data are normally distributed because they are assumed to be random subsets of the total sample. The mean for available data is therefore assumed to represent an unbiased estimate of the mean for the total sample (Hertel, 1976). If the variable with missing data is not normally distributed (i.e., skewed), median substitution may be more accurate than mean substitution (Acock).

Although sample mean substitution is easy to compute and although it preserves data, it tends to decrease variance-covariance between the variable with missing data and the other variables (Acock, 1997; Raymond & Roberts, 1987; Roth, 1994; Tabachnick & Fidell, 2001). Because of its insensitivity to the response pattern of an individual subject, sample mean substitution also ignores response bias (Kline, 1998; Patrician, 2002). If data are not MCAR, attenuation of variance could reduce estimates of standardized coefficients (such as $R^2$, $\beta$ in regression analysis), increase standard errors, and reduce true estimates of *t* values (Acock). However, since *t* values are also dependent on sample size, sample mean substitution may artificially inflate the *t* value, because it retains cases with missing data by replacing their missing data with invariant values that do not accurately represent the true scores of missing values (Acock). The main advantage of sample mean substitution is that it is a conservative approach in which the mean for the distribution as a whole does not change (Tabachnick & Fidell). Nonetheless, this does not outweigh the aforementioned disadvantages, because the ascribed mean value is more likely closer to the available values of other respondents than to the real missing value. The use of sample mean substitution should therefore be restricted to situations in which data are assumed to be MCAR and the extent of missingness is very small (Roth, 1994).

***Group mean substitution*** ascribes the group mean value to missing data points within that group, based on the assumption of within-group homogeneity. Therefore, this technique is applicable only to analyses involving grouped data, such as *t* test comparisons, ANOVA, and logistic regression analysis. Group mean substitution is believed to yield more accurate estimates of missing data than sample mean substitution because it minimizes the risk of attenuation of between-group variance that occurs when an overall sample mean is used to replace missing values (Acock, 1997; Tabachnick & Fidell, 2001). In other words, this approach assumes that scores for different groups (levels) of a given variable are heterogeneous, and that scores for subjects within a group are *homogenous* (Acock). Suppose, for instance, that stroke patients and healthy individuals are compared on *self-care abiliites*, and that some data are missing on this, a continuous variable. Using group mean substitution, missing data on self-care abilities could be estimated by dividing the sample into two groups based on their health state (healthy versus stroke). Cases from the healthy group who are missing data on self-care abilities would then be assigned the healthy group mean value of the self-care abilities variable and vice versa. A significant disadvantage of group mean substitution is that the assumption of within-group homogeneity may be violated if the within-group variance is relatively large. In this case, group mean substi-

tution may yield parameter estimates that are not different from those produced by sample mean substitution (Tabachnick & Fidell).

***Regression-based imputation*** uses knowledge of other variables to predict the values of missing data on a given variable. This technique entails the creation of a dummy code for missing data and treating it as a dependent variable. The values of the missing data are then estimated using the logistic regression equation that results from regressing other variables with complete observations on the missing data dummy code. This approach is based on the principle that if the missing data variables can be predicted by the other variables in the data set, then the resulting regression equation could be used to predict missing values for incomplete cases (Hair et al., 1998; Patrician, 2002; Tabachnick & Fidell, 2001). If more than one variable in the data set has missing data, a prediction equation will be needed for each missing data variable, which can be a very tedious and complicated process.

The main advantage of regression-based imputation is that it strives to methodologically estimate the missing data and thus is a relatively objective technique (Tabachnick & Fidell, 2001). Use of regression-based imputation yields reasonable estimates of means, particularly when normality assumptions are plausible. However, the covariance matrix that results from a data set with imputed values tends to underestimate the true variances and covariances, because regression techniques project the value of missing data onto the regression line, thus decreasing deviation about the line. The extent of underestimation resulting from regression imputation is, however, less than that which results from mean substitution techniques (Little & Rubin, 1987). Empirical studies indicate that regression methods are more accurate than the previously described approaches to dealing with missing data (Raymond & Roberts, 1987). Raymond and Roberts suggest that regression methods are most useful when data are 10% to 40% incomplete and the variables are at least moderately correlated. When correlations between variables are low, regression will not perform much better than mean substitution or pairwise deletion. Roth (1994) suggests that regression methods are appropriate when 6% to 20% of data are MCAR, up to 15% of data are MAR, or up to 10% of data are missing in a systematic pattern.

Despite its strength as an empirical imputation technique, regression-based imputation has several disadvantages. It can lead to over-prediction of the missing data if the explained variance ($R^2$) in the missing data variable was inflated due to multicollinearity (Acock, 1997; Cohen & Cohen, 1983). Also, the scores may fit together better than they should because the predicted missing value is likely to be more consistent with the variables that predicted it than with the actual value of the missing

score (Tabachnick & Fidell, 2001). A third disadvantage stems from the fact that the variables used to predict the missing variable(s) may not be good predictors and may therefore lead to inaccurate estimation of the missing value(s). One way to minimize inflation or underestimation of estimates is to use only the best predictor or set of predictors in the regression model (Acock). In addition, researchers using regression methods to estimate missing values are cautioned not to include the dependent variable of the study in the prediction equation that will be used to estimate missing data, because this may artificially inflate the $R^2$ (Raymond & Roberts, 1987).

***Expectation maximization (EM) algorithm*** uses an iterative procedure in order to produce the best parameter estimates. It begins with an estimation of missing data based on assumed values for the parameters. The actual data and missing estimates are then used to update the parameter estimates, which are, in turn, used to re-estimate missing data. The process continues until there is convergence in the parameter estimates (Roth, 1994; Schafer & Olsen, 1998), which indicates that more iterations will not produce any significant change in parameter estimates (University of Texas Statistical Services, 2000). EM is considered superior to the aforementioned techniques because it produces unbiased parameter estimates when data are MCAR and less biased parameters when data are MAR or systematic (Acock, 1997). Despite its complex mathematical and conceptual foundations (Roth, 1994), EM can be easily carried out using several software packages such as SPSS under the missing data analysis option.

## Summary

This paper provides an overview of commonly recommended approaches to handling missing data. Despite the interesting features of each of these techniques, the most effective way of handling missing data is to prevent its occurrence. However, when missing data becomes a problem, it is essential for the researcher to determine the pattern of missingness and choose the proper approach to handling missing data. Almost all of the missing data techniques discussed in this paper have advantages and disadvantages. Some techniques, such as deletion procedures and mean substitution, are technically simple but empirically weak. Others are technically challenging but tend to yield more robust estimates. Because the validity of research results may be dependent on the investigator's approach to handling missing data, we recommend that nurse researchers inform their readers about how the problem of missing data was addressed. This practice serves to highlight the rigour and validity of nursing research.

## References


Acock, A. (1997). Working with missing data. *Family Science Review, 1*(10), 76–102.

Allison, P. (2000). *Missing data*. Thousand Oaks, CA: Sage.

Allison, P. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545–557.

Barnard, J., & Meng, X. L. (1999). Applications of multiple imputation in medical studies: From AIDS to NHANES. *Statistical Methods for Medical Research, 8*(1), 17–36.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences,* 2nd Ed. Hillsdale, NJ: Lawrence Erlbaum.

Downey, R. G., & King, C.V. (1998). Missing data in Likert ratings: A comparison of replacement methods. *Journal of General Psychology, 125,* 175–191.

Hair, J., Anderson, R., Tathman, R., & Black, W. (Eds.). (1998). *Multivariate data analysis with readings,* 4th Ed. Upper Saddle River, NJ: Prentice-Hall.

Heitjan, D. F. (1997). Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health, 87*(4), 548–550.

Hertel, B. (1976). Minimizing error variance introduced by missing data routines in survey analysis. *Sociological Methods and Research, 4,* 459–474.

Huisman, M. (1998). Missing data in behavioral science research: Investigation of a collection of data sets. *Kuantitieve Methoden, 57,* 63–93.

Kline, R. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Orme, R. G., & Reis, J. (1991). Multiple regression with missing data. *Journal of Social Service Research, 15,* 61–91.

Patrician, P. A. (2002). Multiple imputation for missing data. *Research in Nursing and Health, 25*(1), 76–84.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation and the Health Professions, 9,* 395–420.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement, 47,* 13–26.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47,* 537–561.

Roth, P. L., & Switzer, F. S. (1995). A Monte Carlo analysis of missing data techniques in a HRM setting. *Journal of Management, 21,* 1003–1023.

Roth, P. L., & Switzer, F. S. (1999). Missing data: Instrument-level heffalumps and item-level woozles. Retrieved April 26, 2004, from http://www.aom.pace.edu/rmd/1999_RMD_Forum_Missing_Data.htm.

Roth, P. L., Switzer, F. S., & Switzer, D. M. (1999). Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. *Organizational Research Methods, 2*(3), 211–212.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley.

Schafer, F. L., & Olsen, M. K. (1998). Multiple imputation or multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics,* 4th Ed. Boston: Allyn & Bacon.

University of Texas Statistical Services. (2000). Handling missing or incomplete data. Retrieved June 4, 2004, from http://www.utexas.edu/cc/faqs/stat/general/gen25.html

## Authors' Note

Comments or queries may be directed to Maher M. El-Masri, Faculty of Nursing, University of Windsor, Health Education Centre, 401 Sunset, Room 328, Windsor, Ontario N9B 3P4 Canada. Telephone: 519-253-3000, ext. 2400. Fax: 519-973-7084. E-mail: melmasri@uwindsor.ca

---

*Maher M. El-Masri, PhD, RN, is Associate Professor, Faculty of Nursing, University of Windsor, Ontario, Canada. Susan M. Fox-Wasylyshyn, PhD, RN, is Assistant professor, Faculty of Nursing, University of Windsor.*