Are missing data lost data?

Blerina Metanj (Subashi)
Senior Research Analyst
IDRA Research & Consulting

Abstract

Data analysis is sometimes compromised by missing data. In recent years the development of statistical methods to address missing data has been an active area of research. In this context, multiple imputationsis a general-purpose method for analyzing datasets with missing data that is broadly applicable to a variety of missing data. In this paper firstly, issues on missing data will be presented with a brief theoretical introduction to the multiple imputations as an analytic strategy under this field. Besides,a discussion on multiple imputation techniques with a data example for illustration and clarity will be presented. In the end, some conclusions on a particular analysis will be discussed.

Keywords: multiple imputations, missing data, data analysis.

Introduction

Missing data can be a cause of problems and biases in many datasets. These can be a challenge to many researchers of different fields that have to deal with data. In the last years, a lot of methodologies handle these kinds of problems. Neglecting missing data can harm the results and bring to wrong conclusions (Little & Rubin, 1987; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997; Schafer & Graham, 2002).

Multipleimputationsin this context, are a methodology for handling missing data which can be used by researchers on many analytic levels. Many research studies have used multiple imputations (e.g., Graham et al., 1997; Wayman, 2002a) and good general reviews on multiple imputations have been published (Graham, Cumsille, &Elek-Fisk, 2003; Graham & Hofer, 2000; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). However, many researchers still do not use it due to the lack of information on the benefits that it can bring to their work.

This paper aims to be additional material to the existing literature giving a concrete example of using multiple imputations and showing its benefits. Many researchers can understand it and might use it in their work.

1. Background

Item missing data has long been recognized as a problem for data analysts. Early solutions to the problem of missing data were directed to specific distributions for the variables of interest and patterns of missing data. For example, Buck’s (1960) method introduced imputations of conditional mean values for each pattern of missing observations in a multivariate normal vector of variables.

Broad, formal recognition of imputation as a statistical technique for dealing with missing data may have been originated with the National Research Council (NRC) Panel on Incomplete Data. Many of the earliest papers on imputation concepts and theory appear in the 1985 three-volume publication produced by the panel (Madow and Olkin 1983). Throughout the 1980′ statisticians continued to conduct research and to publish on the imputation method (Kalton1983;Rubin 1980; Sande 1983). The general theory and methods were greatly extended by the introduction of the multiple imputation method (Rubin 1987). Despite these developments, the introduction of the imputation method to statistical practice at that time was a slow process and by no means universal.

Before the mid-1980, the accepted procedure among most data analysts was to explicitly denote values as missing (eg '." symbol) but to take no corrective steps in actual data analysis other than to analyze complete data case. During the 1980s, major federal survey programs in the US and Canada took the lead in the development and application of basic imputation methods such as regression imputation and the hot deck imputation method.

During the 1990s and continuing to the present, the demand for practical methods to address increasingly large and complex missing data problems in surveys and other statistical investigations led to an explosion of new theoretical work during the next two decades, much of it focused on methods of multiple imputations (van Buuren 2012).

2. Methods of treatment for missing values

Any analysis aims to make inferences for the eligible population under the study. Any missing data can threaten this goal. Therefore, we have to respond to the missing data problem in a way that reflects the population of inference.

There are many different methods of managing missing data. Some of the most population missing data methods involve ad-hoc or replacement of missing data. These methods edit missing data to produce a complete data set and because of ease of treatment are very attractive to users. However, researchers have been cautioned against using these methods because they have been shown to have serious drawbacks (e.g., Little & Schenker, 1995; Graham & Hofer, 2000; Graham et al. 1997; Schafer & Graham, 2002). For example, simply eliminating missing data ("listwise deletion" or "complete case analysis") will bias results if the remaining cases are not representative of the entire sample. In most statistical software this is the default method. Mean substitution is another method available in statistical packages. This method replaces missing data with the average of the valid data for the variable in question. Since the same value is being replaced for each missing case, this method artificially reduces the variance of the variable, and besides it diminishes the relationship with other variables. Graham et al (2003) refer to these traditional methods ad "unacceptable methods".

There are more statistically principled methods of handling missing data which have been shown to perform better (e.g., Little & Rubin, 1987; Grahamet al., 1997; Schafer & Graham, 2002).

These methods do no deal only with the replacement of the missing value, but they use the available information to preserve relationships in the entire data set. Maximum likelihood estimation is such a method. This method finds the most likely value for the parameter based on the data set collected. This is a sound method for treating missing data, but often difficult for less advanced analysts. The Expectation-Maximization (EM) algorithm is another method for applying in missing data. This algorithm starts with an initial guess for the parameters.

Multiple imputation techniquesare more commonly used because of ease of use and the availability of software.

3. Mechanisms responsible for missing data

The missing data mechanism describes the process that generates the missing values. It is crucial to understand why there are missing data in a data set. Graham et al. (2003) have categorized the causes in three ways:

- Random processes
- Processes that are measured
- Processes that are not measured

Modern missing data methods generally work well with the first two causes.

More formally missing data mechanism can be grouped (Little and Rubin (1987):

1. MCAR – Missing Completely at Random. In this case, there are no differences between missing and not missing data, in terms of the analysis being performed. A variable is missing completely at random, neither other variables in the dataset nor the unobserved value of the variable itself predict whether a value will be missing. Missing data are MCAR if the events that lead to missing-ness are independent of the observed variables and the unobserved parameters of interest. Missing data are dispersed randomly throughout data. In practice is a very unlikely situation. They are equally unlikely as MNAR.
2. MAR – Missing at Random. Missing data depends on known values and thus is described fully by variables observed in the data set. A variable is said to be missing at random if other variables (but not the variable itself) in the dataset can be used to predict missing-ness on a given variable. Data are MAR if the probability of missing depends **only** on some (or all) of the observed data. Missing data are related to other data. For example, men may be more likely to decline to answer some questions in a survey. Gender predicts missing-ness on other variables.
3. MNAR – Missing not at Random or NMAR (Not Missing at Random) – The missing data depends on events or items which the researcher has not measured. If the value of the unobserved variable itself predicts missing-ness. This is a damaging situation. Often is called non-ignorable missing-ness. In the reality, it is impossible to know definitively if data truly are MNAR, so they are treated as MAR or MCAR. For example, individuals with very high incomes are more likely to decline to answers about incomes compared to other individuals with moderate incomes.

Dealing with missing data is important, as the mechanisms choose can dramatically alter the results. Different types of missing data require different treatment.

Graham & Donaldson (1993) classify missing data mechanism as:

a. Accessible – where the cause of missing-ness can be accounted for. These situations encompass MCAR and MAR circumstances.
b. Inaccessible – missing data mechanisms cannot be measured. They include

nonignorable mechanisms and MAR mechanisms where the cause of missingness is known, but not measured.

As Graham and Hofer (2000) state, the missing data mechanism is rarely completely inaccessible. Often, the mechanism is made up of both accessible and inaccessible factors. Thus, the researcher should cover as much of the mechanism causing missing data as possible, to produce sound results (Graham et al., 1997; Little, 1995; Rubin, 1996). A sensitivity analysis conducted by Graham et al. (1997) showed that the effects of an inaccessible mechanism are often surprisingly minimal in the implementation of multiple imputations.

4. Multiple imputation

This technique, use the existing values of other variables to predict missing values for any variable. The predicted values, called "imputes", are substituted for the missing values, resulting in a full data set called an "imputed data set."

This process is done multiple times and multiple data sets are produced (hence the term "multiple imputations"). Then on each imputed data set statistical analysis is carried out. These analyses at the end are combined to produce an overall analysis.

Multiple imputations not only take care of the missing data by restoring the natural variability in the missing data but at the same time, incorporates the uncertainty caused by estimating missing data. Maintaining the original variability of the missing data is done by creating imputed values thatare based on variables correlated with the missing data and causes of missing-ness. Uncertainty isaccounted for by creating different versions of the missing data and observing the variabilitybetween imputed data sets.

It is important to note that imputed values produced from an imputation model are not intended to be "guesses" as to what a particular missing value might be; rather, this modeling isintended to create an imputed data set which maintains the overall variability in the populationwhile preserving relationships with other variables.

Thus, in performing multiple imputations, aresearcher is interested in preserving important characteristics of the data set as a whole (e.g.,means, variances, regression parameters). Creating imputes is merely a mechanism to deliver ananalysis that makes use of all possible information.

Multiple imputationsis an attractive choice because it represents a good balance between quality of results and ease of use. The performance of multiple imputations in a variety of missing data situations has been well-studied and it has been shown to perform favorably (Graham et al., 1997; Graham & Schafer, 1999; Schafer & Graham, 2002).

Multiple imputations have been shown to produce unbiased parameter estimates which reflect the uncertainty associated with estimating missing data. Further, multiple imputationsare robust to departures from normality assumptions and provide adequate results in the presence of low sample size or high rates of missing data. Multiple imputations also represent a tractable solution to missing data problems. This procedure is computationally simpler than other statistically principled methods such as maximum likelihood estimation, and as will be shown later, is a method that is intuitive and easy to understand. Although the statistical principles behind multiple imputations are not trivial, user-friendly software exists which employs these procedures such that the researcher can concentrate on learning and implementing the process of multiple imputations rather than the underlying statistics.

Finally, one of the great requests of multiple imputations is that the required user interaction isfamiliar to many researchers – multiple imputations produce full, complete data sets on which to

perform analyses, and these analyses can be performed by nearly any method or software package the analyst chooses.Schafer and Olsen (1998) point out that like any statistical technique, multiple imputationsdepend on some assumptions, and responsible use of multiple imputations involves a basic understanding of these assumptions and their implications.

The concept of multiple imputations was formulated by Rubin (1987) in large part to address the need for a robust method that could be applied to large data sets with many variable types.

Multiple imputationshave several desirable features:

- MI is *model-based* – It ensures statistical transparency and integrity of the imputation process.
- MI is *stochastic*- It imputes missing values based on draws of the model parameters and error terms from the predictive distribution of the missing data.
- MIS is *multivariate*- It preserves not only the observed distributional properties of every single variable but also the associations among the many variables that may be included in the imputation model, under the assumption that the data are MAR.
- MI *employs multiple independent repetitions* of the imputation procedure that permit the estimation of the uncertainty (the variance) in parameter estimates that is attributable to imputing missing values.
- MI is *robust*
- MI is *usable* in real statistical applications.

5.1 Creating Imputed Data Sets

Multiple imputations (MI) appear to be one of the most attractive methods for the general-purpose handling of missing data in multivariate analysis.

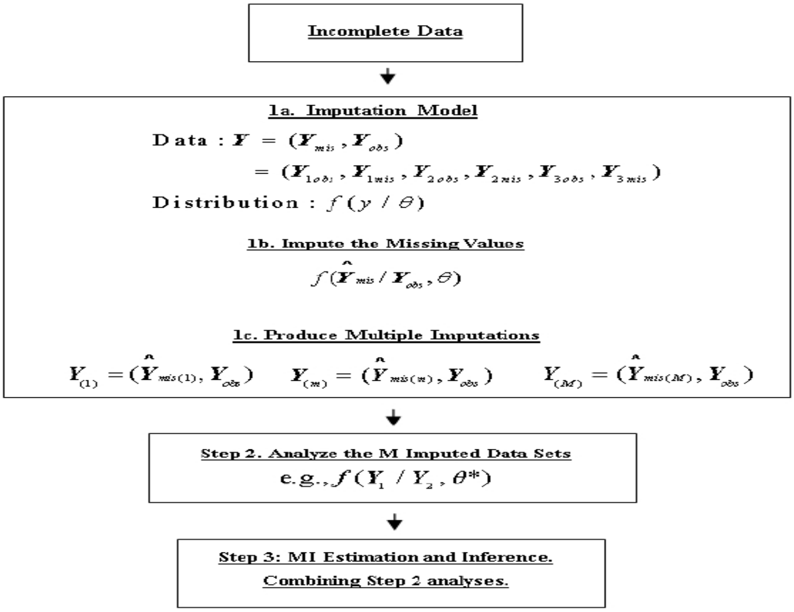
The basic idea, first proposed by Rubin (1977) and elaborated in his (1987) book, is quite simple:

1. Impute missing values using an appropriate model that incorporates random variation. The first step in multiple imputations is to create values (“imputes”) to be substituted for the missing data. To create imputed values, we need to identify some model (regression line) that will allow us to create imputes based on other variables in the data set (predictor variables).
2. Do this M times (usually 3-5 times), producing M complete data sets.
3. Perform the desired analysis on each data set using standard complete-data methods.
4. Average the values of the parameter estimates across the M samples to produce a single point estimate.

5. Calculate the standard errors by (a) averaging the squared standard errors of the M estimates (b) calculating the variance of the M parameter estimates across samples, and (c) combining the two quantities using a simple formula (given below)

This procedure can be organized into three sequential steps:

Figure 1. Multiple imputation steps


Incomplete Data
1a. Imputation Model
Data : $Y = (Y_{mis}, Y_{obs})$
$= (Y_{1obs}, Y_{1mis}, Y_{2obs}, Y_{2mis}, Y_{3obs}, Y_{3mis})$
Distribution : $f(y / \theta)$
1b. Impute the Missing Values
$f(\hat{Y}_{mis} / Y_{obs}, \theta)$
1c. Produce Multiple Imputations
$Y_{(1)} = (\hat{Y}_{mis(1)}, Y_{obs})$
$Y_{(m)} = (\hat{Y}_{mis(m)}, Y_{obs})$
$Y_{(M)} = (\hat{Y}_{mis(M)}, Y_{obs})$
Step 2. Analyze the M Imputed Data Sets
e.g., $f(Y_1 / Y_2, \theta^*)$
Step 3: MI Estimation and Inference.
Combining Step 2 analyses.

Source: Author

The choice of variables to include in the imputation model should not be limited to only variables that have an item missing data or variables that are expected to be used in the subsequent analysis. As a general rule of thumb, the set of variables included in the imputation model for an MI analysis should be much larger and broader in scope than the set of variables required for the analytic model. Based on recommendations from Schafer (1999) and van Buuren (2012) some practical guideline for choosing which variables to include in the imputation model are the following:

6. Include all key analysis variables:(dependent and independent)
7. Include other variables that are correlated or associated with the analysis variables
8. Include variables that predict item missing data on the analytic variables

Failure to include one or more analysis variables (1) in the imputation model can result in bias in the subsequent MI estimation and inference. Including additional variables (2) that are good predictors of the analytic variables improves the precision and accuracy of the imputation of item missing data. Under the assumption that item missing data is MAR, incorporation variables (3) that are correlated with the variables that have missing data and predict the propensity for response will reduce bias associated with the item missing data mechanism. When in doubt, including more variables in the imputation model is better.

Discussion of exactly how this set of regression lines is identified is beyond the scope of this paper; for this paper, we will assume this set is easily produced. The number of imputed data sets to create is up to the analyst. Commonly, researchers choose between 3 and 10 data sets. Choosing variables to include in the imputation model is important and the reader is directed to further reference (Collins et al., 2002; Wayman, 2002b). Theoretically, the statistical efficiency of multiple imputation methods is maximized when the number of repetitions is infinite, M= ∞. The same theory tells us that if we make the practical choice of using only a modest, finite number of repetitions (e.g., M=5, 10, 20) that loss of efficiency compared to the theoretical max. is relatively small. A measure of relative efficiency is:

$$RE = \left(1+\frac{\lambda}{M}\right)^{-1}$$

◎ – is the fraction of missing information;
M – is the number of MI repletion.

If the rates of missing data are modes (<20%), MI analyses based on as few as M=5 or M=10 will achieve >96% of efficiency. If the missing information is high (30-50%), analysts are advised to take M=20 or M=30 to maintain a relative efficiency of 95% or greater.

The following table shows the relative efficiencies with different values of m and λ. For cases with little missing information, only a small number of imputations are necessary for the MI analysis.

Table 1. Relative efficiencies with different values of m and λ

m	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Source: Author

5.2 Example for Multiple Imputation

A researcher is hoping to model the quantity of a good based on the price and the income. Let suppose the following data.

To build an adequate model he chooses to use the price and income. Although not detailed here, choosing variables to include in the imputation model is important and the reader is directed to further reference (Collins et al., 2002; Wayman, 2002b).

Table 2. Example of using Multiple Imputation

Quantity	Price	Income
19	12.66	16.0
24	7.61	29.3

27	11.08	13.5
30	7.84	21.3
24	8.65	29.1
.	6.32	27.5
39	7.04	16.2
41	8.45	27.8
20	9.89	14.0
39	4.47	23.4
45	9.38	39.1
.	4.91	30.3
46	6.02	29.3
49	7.12	29.7
.	2.77	17.7
52	5.18	20.2
35	10.25	17.7
57	5.68	30.6
58	3.59	39.2
61	1.69	32.7
66	3.45	37.0

Source: Author

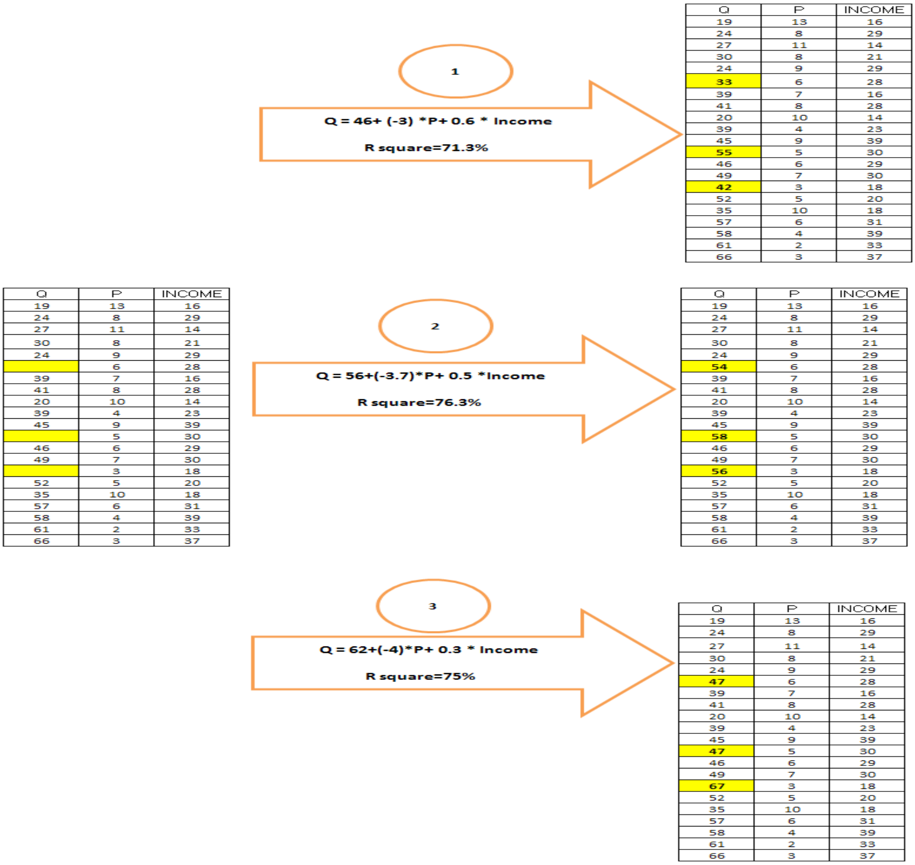
The figure below explains the process of creating imputed values.[1]

Once the imputed data sets have been created, the analysis of choice is conducted separately foreach data set. The multiple regression model with all two predictors produced different R^2 for each linear regression. Based on the figure above the analyst can choose the model with higher R square statistics.

This analysis stops at the third step as mentioned above. In future papers, the analysis will extend to combine these analyses to produce one overall set of estimates. Combining the estimates from the imputed datasets is done using rules established by Rubin (1987). These rules allow the analyst to produce one overall set of estimates like that produced in a non-imputation analysis.

[1] The software used in this paper is SPSS.

Figure 2. Process of Creating Imputed Data Sets



Q	P	INCOME
19	13	16
24	8	29
27	11	14
30	8	21
24	9	29
	6	28
39	7	16
41	8	28
20	10	14
39	4	23
45	9	39
	5	30
46	6	29
49	7	30
	3	18
52	5	20
35	10	18
57	6	31
58	4	39
61	2	33
66	3	37

Q = 46+ (-3) *P+ 0.6 * Income

R square=71.3%

Q	P	INCOME
19	13	16
24	8	29
27	11	14
30	8	21
24	9	29
33	6	28
39	7	16
41	8	28
20	10	14
39	4	23
45	9	39
55	5	30
46	6	29
49	7	30
42	3	18
52	5	20
35	10	18
57	6	31
58	4	39
61	2	33
66	3	37

Q = 56+(-3.7)*P+ 0.5 *Income

R square=76.3%

Q	P	INCOME
19	13	16
24	8	29
27	11	14
30	8	21
24	9	29
54	6	28
39	7	16
41	8	28
20	10	14
39	4	23
45	9	39
58	5	30
46	6	29
49	7	30
56	3	18
52	5	20
35	10	18
57	6	31
58	4	39
61	2	33
66	3	37

Q = 62+(-4)*P+ 0.3 * Income

R square=75%

Q	P	INCOME
19	13	16
24	8	29
27	11	14
30	8	21
24	9	29
47	6	28
39	7	16
41	8	28
20	10	14
39	4	23
45	9	39
47	5	30
46	6	29
49	7	30
67	3	18
52	5	20
35	10	18
57	6	31
58	4	39
61	2	33
66	3	37

Source: Author

Conclusions

Not long ago, missing data was viewed as something to discard. Researchers knew the bias problems this presented, but there were no methods available to account for missing data bias. Today, however, researchers are not bound by such constraints. Methods such as multiple imputations are available and usable for most researchers, so there is no need to publish studies that suffer from sample bias.

The description in this paper attempts is mostly conceptual, aimed at providing a clear understanding of the basic ideas of multiple imputations, a good base from which to learn more about the use of multiple imputations, and an understanding for reading research that uses multiple imputations.Scepticism about any methodology which is unfamiliar and is presented as an improvement upon traditional methodologies is understandable and often necessary. However, it is important to reiterate that the superiority of multiple imputations to traditional methods is based on mathematical fact, not belief or opinion. Other missing data methods will likely be developed which are superior to multiple imputations, but until such methods are available, multiple imputations provide a good solution to missing data problems.

References

Jeffrey C.Wayman, *"Multiple Imputation For Missing Data: What Is It And How Can I Use It?"*, Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL,pp . 2 -16, 2003.

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

Rubin, D. B. (1996). Multiple imputations after 18+ years. *Journal of the American Statistical Association*. , 473–489.

Software for Multiple Imputation Donald B. Rubin.

Goldstein, H., J. Carpenter, et al. (2009). "*Multilevel Models with multivariate mixed response types*." Statistical Modeling.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Little, R.J.A. and Rubin, D.B. (1987), *Statistical Analysis with Missing Data First Edition*, New York: John Wiley & Sons, Inc.

Multiple Imputation for Missing Data Paul D. Allison, Sociology Department, University of Pennsylvania,4-6.

Yuan, Y.C. *Multiple Imputation for Missing Data: Concepts and New Development*. P267-275.

Nicholas J. HORTON and Stuart R. LIPSITZ *Multiple Imputation in Practice: Comparison of Software Packages for Regression ModelsWith Missing Variables.*

Little RJA, Rubin DB. *Statistical Analysis with Missing Data. Second Edition* ed. New York: John Wiley & Sons, 1987.