ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AN INTEGRAL STUDY ON MISSING VALUE DATA IMPUTATION

S.Swetha*
Assistant Professor, Department of CSE, DIET Andhra Pradesh, India.

ABSTRACT

Imputation is the process of replacing missing data with substituted values. Missing data can create problems for analyzing data, imputation is seen as a way to avoid pitfalls involved with list wise deletion of cases that have missing values. That is to say, when one or more values are missing for a case affects the representativeness of the results [1]. Once all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data. Many theories have been proposed for missing data computation but the majority of them introduce large amounts of bias. A very few techniques to deal with missing data include: hot deck and cold deck imputation; list wise and pair wise deletion; mean imputation; regression imputation; last observation carried forward; stochastic imputation; and multiple imputation.

This paper reviews missing value data imputation methods for analyzing missing data, including basic concepts and applications of imputation techniques.

KEYWORDS: Bias, Random Process, Sample Space, Statistics, Imputed data set, Mean, Variance, Standard error.

INTRODUCTION

Educational researchers are facing many problems and biases which can be caused by missing data. What is missed data? Missed data, or missed values occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. To deal with missing data variables, proposed data imputation.

Actually what is Data Imputation? Data Imputation is the substitution of estimated values for missing or inconsistent data items (fields). The substituted values are intended to create a data record that does not fail edits. [2]

Why the data are missing? Missing of data can be occurred informally in some combination of three ways: Random processes, processes which are measured, and processes which are not measured. First, data can be ―Missing completely at Random‖, or MCAR. When data are MCAR,

Missing cases are no different than non-missing cases, in terms of the analysis being performed. These cases can be considered as randomly missing from the data and the only real price in failing to account for missing data is loss of power. Second, data can be missing ―Missing at Random‖, or

MAR. In this case, missing data depends on known values and thus is described fully by variables observed in the data set. Accounting for the values which ―cause‖ the missing data will produce unbiased results in an analysis. Third, data can be missing in an unmeasured fashion i.e. ―non-ignorable‖ (also called ―Missing Not at Random‖ (MNAR) and ―Not Missing at Random‖ (NMAR)). Since the missing data depends on events or items which the researcher has not measured, this is a damaging situation. [3] ―Reference [4]‖ referred missing data mechanisms as ―accessible‖ and ―inaccessible.‖ An accessible mechanism is one it gives where the cause of miss can be. An inaccessible mechanism is one where the missing data mechanism cannot be measured. The missing data mechanism is actually made up of both accessible and inaccessible factors. So, the mechanism is rarely inaccessible. Although a researcher may not be confident that the data present a purely accessible mechanism, covering as much of the mechanism possible will usually produce sound results ([5], [6], [7]). To overcome these missing data problems advanced methodologies have been implementing to handle these problems and biases. Some of techniques to deal with missing data are: hot deck and cold deck imputation; list wise and pair wise deletion; mean imputation; regression

imputation; last observation carried forward; stochastic imputation; and multiple imputation.

Data Imputation techniques have lot of scope in various fields. It performs statistical analysis of missing data in various areas like: Clinical data sets (lung cancer, prostate cancer, oncology), Bio-informatics (protein sequence, DNA, RNA analysis), Wireless sensor networks, Neural Networks.

This paper will first present a brief discussion of some data imputation issues. Following this will be a description of the workings of the imputation process, with a data example interspersed throughout the description to provide illustration and clarity. Finally, the paper will conclude with a brief discussion of issues surrounding this particular analysis.

IMPUTATION TECHNIQUES

There are many theories have been proposed by scientists to account for missing data but the majority of them introduce large amounts of bias. Imputation Techniques are mainly classified into 3 categories.

A. *Case Deletion*

List-wise Deletion:

The most common and simple way of dealing with missing data is list-wise deletion, which is used by deleting missing value. If the data are missing completely then list-wise deletion does not add any bias, but it will decrease the power of the analysis. This is because deleting all cases with missing values decreases the effective sample size, this directly affect interrelationships present in the variables. For example, if 1000 cases are collected but 80 have missing values, the effective sample size for a complete-case analysis is on 920. If the cases are not missing completely at random, then list-wise deletion will introduce bias because the sub-sample of cases represented by the missing data are not representative of the original sample.

Pair-wise Deletion

Pair-wise deletion (or "available case analysis") involves deletion of specific variable cell values along with the outcome variable when a particular variable is required in an analysis and has a missing value, but the case will exist in all other situations. Pair-wise deletion introduces impossible mathematical situations such as correlations that are over 100% [8].

Let's take an example:

TABLE – I SAMPLE DATA WITH MISSING VALUE

Unit	Variables	
	Variable1	variable2
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.47

8	2.4	4.94
9	1.7	5.73
10	3.6	-Missing-

In case deletion, all units with incomplete data will be deleted from the analysis (e.g. unit 10). Handling missing data by eliminating cases with missing data (―List-wise deletion‖ or ―Pair-wise deletion) will bias results if the remaining cases are not representative of the entire sample. This method is the default in most statistical software [9].

B. Single Imputation

[1] ***Hot-deck Imputation:*** Another simple common method of imputation was hot-deck imputation where a missing value was imputed from a randomly selected similar record. The term "hot deck" originated from the storage of data on punch cards, and in this model donors and recipients are from the same data set. The stack of cards was "hot" because it was currently being processed.

[2] ***Last Observation Carried Forward****:* One form of hot-deck imputation is called "last observation carried forward". Which involves sorting of a dataset according to number of variables,
thus creating an ordered dataset. The technique then finds the first missing value and uses the cell value which is immediately prior to the missing data they are used to impute the missing value. The process is repeated for the next cell with a missing value until all missing values have been imputed [8].

This method is specific to longitudinal data problems. Each individual, missing values are replaced by the last observed value of that variable. For example:

TABLE II IMPUTED DATA

	Observation Time					
Unit	**1**	**2**	**3**	**4**	**5**	**6**
1	3.8	3.1	2.0	?->2.0	?->2.0	?->2.0
2	4.1	3.5	3.8	2.4	2.8	3.0
3	2.7	2.4	2.9	3.5	?-> 3.5	?-> 3.5

Here the three missing values for unit 1, at times 4, 5 and 6 are replaced by the value at time 3, namely 2.0. Likewise the two missing values for unit 3, at times 5 and 6, are replaced by the value at time 4, which is 3.5. Using LOCF, once the data set has been completed in this way it is analysed as if it were fully observed. For full longitudinal data analyses this is clearly disastrous [10].

*[5]****Cold-deck Imputation***: Cold-deck imputation by contrast, selects donors from another dataset. Since computer power has advanced rapidly and punched cards are no longer used, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques.

[6] ***Mean Substitution Imputation:*** Another imputation technique involves in replacement of any missing value with

the mean of that variable for all other cases i.e. missing data replaces with the average of valid data for the variable in question. Which has the benefit that those variables sample mean value does not change. However, mean imputation may affect the correlations among variables because it changes the way two variables co-vary. Because the same value is being substituted for each missing case, this method artificially reduces the variance of the variable in question, in addition to diminishing relationships with other variables. Thus, mean imputation has some attractive properties for unvaried analysis but becomes problematic for multivariate analysis [8].

For table I, the Simple mean imputation can be calculated as; we replace the unit 10 variable 2 value with the arithmetic average of the observed data for that variable.

Mean = (5.67+ 4.81+ 4.93 + 6.21 + 6.83 + 5.61 + 5.47 + 4.94 +

5.73+ 5.58)/10= 5.58

TABLE IIII MPUTED DATA BY MEAN SUBSTITUTION

Unit	Variables	
	Variable1	**variable2**
1	3.4	5.67
2	3.9	4.81
3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.47
8	2.4	4.94
9	1.7	5.73
10	3.6	5.58

TABLE IV IMPUTED DATA BY REGRESSION TECHNIQUE

Unit	Variables	
	Variable1	**variable2**
1	3.4	5.67
2	3.9	4.81

3	2.6	4.93
4	1.9	6.21
5	2.2	6.83
6	3.3	5.61
7	1.7	5.47
8	2.4	4.94
9	1.7	5.73
10	3.6	5.24

This approach is clearly inappropriate for categorical variables. It does not lead to proper estimates of measures of association or regression coefficients. Rather, associations become weak. In addition, variances will be wrongly estimated if the imputed values are treated as real. Thus inferences will be wrong too [11].

5) ***Regression Imputation:*** Regression imputation has the opposite problem of mean imputation. Regression model is estimated to predict the missing values and the missing data is imputed in relation to this. Whether a value on a specific variable is missing or not, is initially found out by using available information from complete and incomplete cases. Then using proper fitted values from the regression model impute the missing values. The problem is that the imputed data do not have an error term included in their estimation, thus the estimates fit perfectly along the regression line without any residual variance. This causes relationships to be over identified and suggest greater precision in the imputed values than is warranted. The regression model predicts the most likely value of missing data but does not supply uncertainty about that value [8].

Here, we use the completers to calculate the regression of the incomplete variable on the other complete variables. Then, we substitute the predicted mean for each unit with a missing value. In this way we use information from the joint distribution of the variables to make the imputation.

Example: Use Table I for Regression mean imputation.

To perform regression imputation, we first use regression on variable 2 on variable 1.

By applying Regression Mean imputation xpected value of v2 = 6.56 – 0.366 V1.

For unit 10, this gives: 6.56 – 0.366 x 3.6 = 5.24. This value is placed in unit 10 variable 2.

Results of regression mean imputation. Regression mean imputation can generate unbiased estimates of means, associations ad regression coefficients in a much wider range of settings than simple mean imputation. One important problem remains. The variability of the imputations is too small, so the estimated precision of regression coefficients will be wrong and inferences will be misleading [12].

6) ***Stochastic Regression:*** Stochastic regression was a fairly successful attempt to correct the lack of an error term in regression imputation by adding the average regression variance to the regression imputations to introduce error. Stochastic regression is an excellent technique, and shows much less bias than the above mentioned techniques, but it still missed one thing - if data are imputed then intuitively one would think that more noise should be introduced to

the problem than simple residual variance [8].

C. Multiple Imputation

1) ***Multiple Imputation:*** Multiple imputation is one of the best method for dealing with missing data because it is easy to use and produce balancing results. It produces unbiased results even in the case of low sample size or very high rates of missing data.

Missing values for any variable in multiple imputation are predicted by using existing values from other variables. These Predicted values are called as ―imputes‖ and they are substituted for the missing values, resulting in a full data set called an ―imputed data set.‖ This process is repeated multiple times for producing multiple imputed data sets. Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined to produce one overall analysis.

It is important to note that imputed values produced from an imputation model are not intended to be ―guesses‖ and this modeling is intended to create an imputed data set which maintains the overall variability in the population while preserving relationships with other variables. Thus, in performing multiple imputation, a researcher is interested in preserving important characteristics of the data set as a whole. Creating imputes is merely a mechanism to deliver an analysis which makes use of all possible information.

2) The Multiple Imputation Process*:* Multiple imputation is mainly implemented by using three steps: creating imputed data sets which are plausible representations of the data; statistical analysis on each of these imputed data sets; combining the results of these analyses (―average‖ them) to produce one set of results. To proceed into multiple imputation process we have to fist identify imputed data sets, that's the crucial steps of the multiple imputation process. In order to create imputed values, we need to identify some model (we'll call it a regression line) which will allow us to create imputes based on other variables in the data set (predictor variables). Do this multiple times for multiple imputed data sets. Proper predictor variables i.e. variables correlated with the missing variable, the reason for missingness, or both; should be chosenfor regression [13].

For example, if the missing variable of interest is a high school achievement test score, variables such as the student's previous test scores could be included since they are likely correlated with achievement test scores. Using the same example suppose students are more likely to be missing the achievement test if they are in the upper grades. Grade in school could then be included in the imputation model as a reason for missingness. Choosing variables to include in the imputation model is important [14]. In this example, the variable of interest is a nationally administered reading test score, herein referred to as the ―national test‖ and given in normal curve equivalents (NCEs) Almost 15% of the data on this test is missing. Four variables were chosen for the imputation model: score on a locally administered reading test (―local test‖) grade, gender, and special education status To illustrate the data, Table - V lists a subset of participant data from this data set.

TABLE V SELECTED DATA FOR MULTIPLE IMPUTATION

Grade	*Gender*	*Special Ed*	*Local Score*	*National Score*
8	F	No	345	-Missing-
8	M	No	325	30
8	M	No	308	18
8	M	Yes	300	-Missing-
8	M	No	369	40
8	F	Yes	360	10
7	F	No	314	45
7	M	Yes	291	-Missing-
7	F	No	303	10
7	F	No	407	92
7	M	No	375	93

7	F	No	334	-Missing-
6	F	No	348	56
6	M	Yes	383	32
6	F	No	376	60
6	F	No	310	-Missing-
6	F	No	383	-Missing-

The regression lines to impute national test scores from grade, gender, special education, and local test scores are:

(1) National Score = -135.78 + .31(Grade) + 1.14(Male) + - 10.68(Special Ed) + .50(Local Score) + error

(2) National Score = -133.40 + .34(Grade) + .81(Male) + - 9.96(Special Ed) + .50(Local Score) + error

(3) National Score = -131.51 + .11(Grade) + 1.43(Male) + - 10.19(Special Ed) + .49(Local Score) + error

(Note: Male=1, Female=0; Special Ed=1, not Special Ed=0)

If the analyst had to do the imputation by hand, (s) he would begin by identifying cases which were missing the national test. For each of these participants, the analyst would observe that participant's values for grade, gender, special education, and local test, fill these values into the first regression line, add a random error5, and compute predicted values. These predicted values would be substituted for the missing values to create the first imputed data set. The same procedure would be followed using the second regression line to create the second imputed data set, and also for the third.

To illustrate, consider the first student shown in Table V, an eighth-grade female not in special education who scored 345 on the local test, but is missing the national test. In order to produce an imputed value for this student, we would substitute these values into the first regression equation. We also need to randomly draw an error value to use in this equation.

Thus, the imputed value for this student is 42.91, figured thusly: 42.91 = -135.78 + .31(8) + 1.14(0) + -10.68(0) + .50(345) + 3.71. We ignore the students who are not missing the national test and proceed to the next student missing the national test, an 8th grade male in special education who scored 300 on the local test. Again, we need to randomly draw an error value for the equation; that value is 2.86.

The imputed value for this student is 10.02, again figured thusly:

10.02 = -135.78 + .31(8) + 1.14(1) + -10.68(1) + .50(300) + 2.86.

This procedure would continue for each of the 2894 other students in the data set who are missing the national test score. Once this procedure is finished, each missing score has an imputed value substituted for it, resulting in a fully-complete imputed data set. Note that all of these imputations were created using the first imputation equation, so this procedure has produced imputed data set #1. Since we have chosen to work with three imputed data sets, we must create two more. To begin creating the second imputed data set, we will follow the same procedure, but this time using Equation 2. Once again, we start with the first student shown in Table V, the eighth-grade female not in special education who scored 345 on the local test. We randomly draw an error for this Imputation, resulting in an imputed value of 42.27:

42.27 = -133.40 + .34(8) + 0.81(0) + -9.96(0) + .50(345) + 0.45.

As before, we would continue to use this equation to impute values for each student missing the national test score, resulting in imputed data set #2. Like imputed data set #1, the second data set is a plausible, but different representation of the population. Imputed data set #3 would then be created using Equation 3, in the same fashion as the first two data sets.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

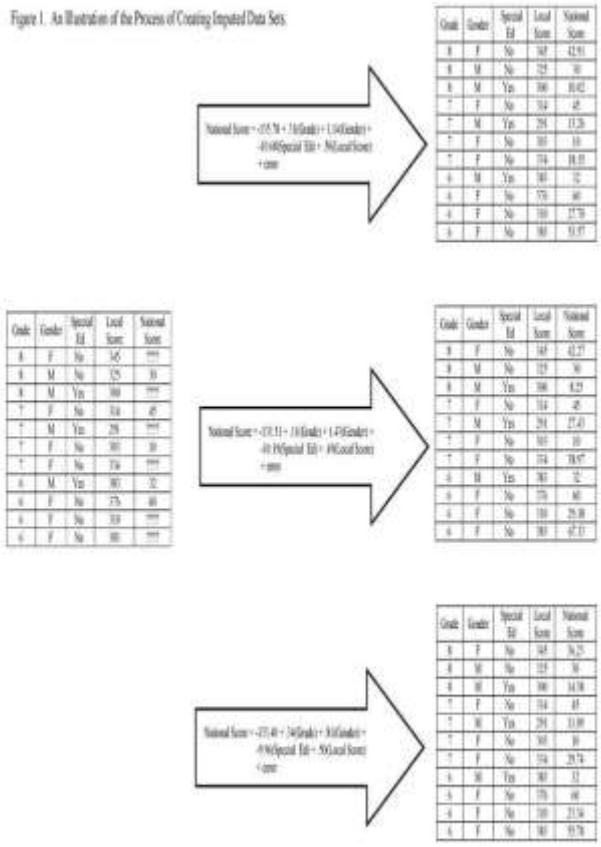

Fig. 1 An Illustration of process of creating imputed data sets.

3) ***Analyzing Imputed Data Sets:*** Once the imputed data sets have been created, the analysis of choice is conducted separately for each data set. This analysis can be any analysis you would perform if there were no missing data (e.g., means, regression, ANOVA), in fact, analysis proceeds just like there were no missing data, and except the analysis is performed on each imputed data set. Point estimates and variances of these point estimates will be collected from the analyses for combination in the next step. To illustrate, consider a very simple analysis which computes the overall mean of the national test. For each imputed data set, we will compute a separate mean and variance; these sets of estimates are then saved so they can be combined to produce an overall estimate of the mean and an overall estimate of the standard error.

The results of this analysis are as follows:
Imputed Data Set #1: Mean = 37.8105, Variance = .0187 Imputed Data Set #2: Mean = 37.8488, Variance = .0185 Imputed Data Set #3: Mean = 37.8166, Variance = .0185

ANALYSIS OF THE AVAILABLE METHODOLOGIES

In our example, data were missing from a national reading assessment. The assessment and the missing data were thought to be correlated with gender, grade, special education status, score from a local reading test, so these variables

were used to help account for missing information from the national test. A basic analysis – computing the overall mean – was undertaken to describe the process of multiple imputation. Given the goals of this paper, the analysis and imputation were understandably rudimentary, but the example still provides good illustration of some important concepts. As stated earlier, multiple imputation (MI) almost always produces estimates which are more representative of the population than do the more popular methods of handling missing data, list-wise deletion (LD) and mean substitution (MS).

Means and standard errors were also computed using these methods in order to illustrate earlier points regarding multiple imputation:

List-wise Deletion (LD):

Mean: 38.83, Standard error: 0.146

Mean Substitution (MS):

Mean: 38.83, Standard error: 0.124

Multiple Imputation (MI):

Mean: 37.83, Standard error: 0.138

One criticism of LD and MS is that both produce biased point estimates because both assume that the missing set of participants is similar to the set with valid values. Since males, students in higher grades, students participating in special education, and students with lower local test scores had lower average national test scores, and since these students were more likely to be missing the national test, that assumption is suspect. Given this bias description, we would assume that the sample of students with valid values would be an artificially higher-scoring sample, thus biasing upward any estimate of mean national test score. In fact, the overall mean calculated using MI is a full NCE lower than that calculated using LD or MS6. This is expected, since MI attempts to account for the sample bias described here. Since MS handles missing data by substituting the same value for each missing data point, standard error estimates from the MS method are necessarily biased downward. This is illustrated in our results. Standard error estimates obtained using MS are 0.022 (15%) and 0.014 (10%) less than the LD and MI estimates, respectively [14].

CONCLUSION

Missing of data leads to adverse effects in data analysis and produce biased results while dealing with missing data. In this paper, we have attempted to provide a basic and clear description of the ideas behind, process of imputation and imputation techniques. The description offered here is mostly conceptual, aimed at providing a clear understanding of the basic ideas of imputation, a good base from which to learn more about the use of imputation, and an understanding for reading research which uses imputation.

REFERENCES

[1] Gelman, Andrew, and Jennifer Hill, *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press, 2006. Ch.25.
[2] Glossary of Terms Used in Statistical Data Editing , Located on *K-Base, the knowledge base on statistical data editing*, UN/ECE Data Editing Group.
[3] Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.
[4] Graham, J. W., & Donaldson, S. I. (1993). *Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data.* Journal of Applied Psychology,78, 119-128.
[5] Graham, J. W., Hofer, S.M., Donaldson, S.I., MacKinnon, D.P., & Schafer, J.L. (1997). *Analysis with missing data in prevention research. In K. Bryant, M. Windle, & S. West (Eds.), The science of prevention: methodological advances from alcohol and substance abuse research.* (325-366).
[6] Little, R. J. A., Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, D. B. Sobel (Eds.) *Handbook*

of Statistical Modeling for the Social and Behavioral Sciences. New York, NY: Plenum.
[7] Rubin, D. B. (1996). *Multiple imputation after 18+ years. Journal of the American Statistical Association*, 91 (434), 473-489.
[8] Enders, C.K. (2010). *Applied missing data analysis.* New York: Guilford Press.
[9] [Online].Available: *http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=67%3Acompleters-analysis&catid=39%3Asimple-ad-hoc-methods-for-coping-with-missing-data&Itemid=96*
[10] [Online].Available: *http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=71%3Alast-observation-carried-forward-locf&catid=39%3Asimple-ad-hoc-methods-for-coping-with-missing-data&Itemid=96*
[11] [Online].Available: *http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=68%3Asimple-mean-imputation&catid=39%3Asimple-ad-hoc-methods-for-coping-with-missing-data&Itemid=96*
[12] [Online].Available: *http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=69%3Aregression-mean-imputation&catid=39%3Asimple-ad-hoc-methods-for-coping-with-missing-data&Itemid=96*
[13] Jeffrey C. Wayman (2003), Multiple Imputation For Missing Data: What Is It And How Can I Use It? Annual Meeting of the American Educational Research Association, Chicago, IL.
[14] Collins, L. M., Schafer, J. L., & Kam, C. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6 (4), 330-351.