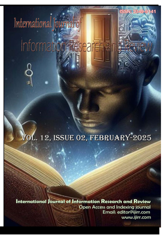
IJIRR

International Journal of Information Research and Review
Vol. 12, Issue, 02, pp.7920-7924, February, 2025



RESEARCH ARTICLE

MULTIPLE IMPUTATION, REGRESSION IMPUTATION AND EXPECTATION MAXIMIZATION METHODS FOR HANDLING MISSING DATA

Monastir Abbas Ahmed Mohammed, Ashraf Hassan Idris Brama, Mohammed Abdalwahab Mohammed Salim and Alaa Alfadel Ahmed Abuzaid

Department of Management Information System, Collage of Business & Economics, Qassim University, Buraydah, Kingdom of Saudi Arabia, Saudi Arabia

ARTICLE INFO

Article History:
Received 18th November, 2024
Received in revised form
24th December, 2024
Accepted 25th January, 2025
Published online 27th February, 2025

Key words:
Multiple Imputation (MI), Regression Imputation, Expectation-maximization (EM) algorithm, missing data, mean absolute error (MAE).

ABSTRACT

The objective of this study is to analyze the negative effects of anthropogenic activities on the wood resources of the Faïra classified forest in the rural commune of Pélengana. To do this, we proceeded by the mixed method, in order to collect information. The results lay bare anthropogenic practices in the decline of woody forest species, agriculture and its practices, in particular the abusive cutting of wood for the unbridled conquest of space and the use of phytosanitary and chemical products. Solutions are proposed for sustainable resource management to safeguard biodiversity. To this end, it is urgent to intervene in order to protect these precious resources on the environment.

INTRODUCTION

Missing data is a normal issue that researcher have to counter. Missing data may happen due to human error or machine error. There are three type is missing data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). Hence, missing data have to be solved before continue with model-building process. There are several methods available for missing data. However, in this research will discuss on EM algorithm and multiple imputation. Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data (11). Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research.

Methods for handling missing data: It is not uncommon to have a considerable amount of missing data in a study. One technique of handling the missing data is to use the data analysis methods which are robust to the problems caused by the missing data. An analysis method is considered robust to the missing data when there is confidence that mild to moderate violations of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population. However, it is not always possible to use such techniques. Therefore, a number of alternative ways of handling the missing data have been developed.

Advanced methods

1.1 Multiple Imputation: The imputed values are draws from a distribution, so they inherently contain some variation. Thus, multiple imputation (MI) solves the limitations of single imputation by introducing an additional form of error based on variation in the parameter estimates across the imputation, which is called “between imputation errors”. It replaces each missing item with two or more acceptable values, representing a distribution of possibilities (1). MI is a simulation-based procedure. Its purpose is not to re-create the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference (2). It involves 3 steps:

Running an imputation model defined by the chosen variables to create imputed data sets. In other words, the missing values are filled in m times to generate m complete data sets. m=20 is considered good enough. Correct model choices require considering:

- Firstly, we should identify which are the variables with missing values.
- Secondly, we should compute the proportion of missing values for each variable.
- Thirdly, we should assess whether different missing value patterns exist in the data (SAS helps us doing this), and try to understand the nature of the missing values. Some key questions are:
- Are there a lot of missing values for certain variables? (E.g. Sensitive question, data entry errors?)
- Are there groups of subjects with very little information available? (E.g. Do they have something in common?)
- Which is the pattern of missingness? Monotone or arbitrary?

The m complete data sets are analyzed by using standard procedures
The parameter estimates from each imputed data set are combined to get a final set of parameter estimates.
Advantages: It has the same optimal properties as ML, and it removes some of its limitations. Multiple imputation can be used with any kind of data and model with conventional software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates. Limitations: It is a bit challenging to successfully use it. It produces different estimates (hopefully, only slightly different) every time you use it, which can lead to situations where different researchers get different numbers from the same data using the same method (3) and (1).

1.2 Regression Imputation Method of estimation: A much more promising method is to use standard regression analysis to provide estimates of the missing data conditional on complete variables in the analysis. For example, for the simple case of univariate missingness in a single continuous variable Y, we fit a regression model to explain Y by the remaining p variables represented by the vector X using the complete cases (subscripted by i):

$$Y_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

Predicted values for the expected values of the missing cases of Y (subscripted by j) can be obtained from

$$\hat{Y}_j = \hat{\alpha} + \sum_{k=1}^{p} \hat{\beta}_k X_{jk} \qquad (2)$$

It should be emphasized that the equations above could be generalized to include models for non-continuous data such as binomial or count data. Missing data are usually multivariate and it is possible to extend the procedure of regression-based imputation from the univariate case to deal with multivariate missingness. For each missing value in the data set a model can be fitted for that variable employing the complete cases of all the other variables(4). Where the number of variables with missing values is large, the number of models to be fitted will also be large, however, efficient computational methods (such as Little & Rubin's sweep operator) can be employed(5).Alternatively, an iterative regression approach can be adopted(6)whereby missing values in a given variable are predicted from a regression of that variable on the complete cases of all other variables in the dataset. This process is repeated for all variables with missing values using complete cases of the other variables *including previously imputed values* until a completed rectangular data set has been generated. The imputation of missing values for each variable is then re-estimated in turn using the complete set of data and the process continues until the imputed values stop changing.

Advantages: The imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced (7)

1.3The Expectation-maximization (EM) algorithm of estimation: This algorithm is a parametric method to impute missing values based on the maximum likelihood estimation. This algorithm is very popular in the statistical literature and has been discussed intensively by many researchers, such as:

This algorithm uses an iterative procedure to find the maximum likelihood estimators of the parameter vector through two-step described in Dempsteret al. (8) and (9).as follows:

The Expectation step (E-step)

The E step is the stage of determining the conditional expected value of the full data of log likelihood function l(θ|Y) given observed data. Suppose for any incomplete data, the distribution of the complete data Y can be factored as

$$f(Y|\theta) = f(Y_{mis}, Y_{obs}|\theta)$$

$$= f(Y_{obs}|\theta)\, f(Y_{mis}|Y_{obs}, \theta) \qquad (1)$$

Where $f(Y_{obs}|\theta)$ the distribution of the data is observed Y_{obs} and $f(Y_{mis}, Y_{obs}|\theta)$ is the distribution of missing data given data observed. Based on the equation (1), we obtained log likelihood function

$$l(\theta|Y) = l(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs}, \theta) \qquad (2)$$

Where $l(\theta|Y)$ is the log-likelihood function of complete data, $l(\theta|Y_{obs})$ is the log-likelihood function of observed data, and $f(Y_{mis}|Y_{obs}, \theta)$ is the predictive distribution of missing data given θ

Objectively, to estimate θ is done by maximizing the log likelihood function (2). Because Y_{mis} not known, the right side of equation (2) cannot be calculated. As a solution, $l(\theta|Y)$ is calculated based on the average value $\log f(Y_{mis}|Y_{obs}, \theta)$ using predictive distribution $f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right)$, where $\theta^{(t)}$ is a temporary estimation of unknown parameters. In this context, an initial estimation $\theta^{(0)}$ be calculated using the complete case analysis. With this approach, the mean value of equation (2) can be expressed

$$Q\left(\theta\middle|\theta^{(t)}\right) = l(\theta|Y_{obs}) + \int \log f(Y_{mis}|Y_{obs}, \theta)\, f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right)\, \partial Y_{mis}$$

$$= \int (l(\theta|Y_{obs}) + \int \log f(Y_{mis}|Y_{obs}, \theta))\; f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right)\, \partial Y_{mis}$$

$$= \int l(\theta|Y)\, f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right)\, \partial Y_{mis} \qquad (3)$$

The equation (3) basically a conditional expected value of the log-likelihood function for complete data $l(\theta|Y)$ given observed data and initial estimate of an unknown parameter.

The maximization step (M-step)

The M step is to obtain the iteratively estimation $\theta^{(t+1)}$ with maximizes $Q\left(\theta\middle|\theta^{(t)}\right)$ as follow

$$Q\left(\theta^{(t+1)}\middle|\theta^{(t)}\right) \geq Q\left(\theta\middle|\theta^{(t)}\right) \qquad (4)$$

Both E and M steps are iterated until convergent.

Advantage: We can use SAS, since this is the default algorithm it employs for dealing with missing data with Maximum Likelihood.

Limitations: Only can be used for linear and log-linear models (there is neither theory nor software developed beyond them). (10)

Application: In this section, we will introduce the applied side

RESULTS AND DISCUSSION

Descriptive statistics of the three methods: We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table 1. Statistics results

Missing percentages%	*Statistics*	*MI method*	*Regression method*	*EM method*
5%	***Mean***	1000.15	1000.00	1000.02
	Variance	24.78	25.11	24.29
15%	***Mean***	1000.36	1000.33	1000.28
	Variance	21.15	23.13	19.03
30%	***Mean***	1000.19	1000.36	1000.22
	Variance	27.21	27.95	20.63

IBM SPSS 25

Results obtained by descriptive statistics; the results revealed that there are no ostensible differences between means and variances.

ANOVA for the estimated means: To test the second hypothesis; is there a statistically significant difference between means or not?We calculated the sig. value of the F-test.

Table 2. Shows summary of the final results of ANOVA

Missing percentages %	*F*	*Sig*	*Results*
5%	2.126	0.137	*Accept H_0*
15%	1.866	0.146	
30%	1.032	0.381	

IBM SPSS 25

From above table, it shows the sig. value of the F-test, all values are greater than significant (0.05) that mean there is no statistical difference between means of estimated missing values.

iii.Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the chi-square test.

Table 3. Correlations results

Missing percentages %	*correlations between*					
	Generated data & MI method		*Generated data & Reg. method*		*Generated data & EM method*	
	r	*Sig.*	*r*	*Sig.*	*r*	*Sig.*
5%	-0.438	0.461	-0.120	0.847	0.334	0.583
15%	0.300	0.277	-0.215	0.442	-0.297	0.282
30%	-0.081	0.672	-0.032	0.865	0.111	0.559

IBM SPSS 25

From above table, it shows the sig. values of the Chi-square test are greater than the significant level (0.05) that mean the correlations are not significant.

iv.Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table 4. Std. Error of Mean and MAE results

Missing percentages %	*MI method*		*Regression method*		*EM method*	
	S.E. of mean	*MAE*	*S.E. of mean*	*MAE*	*S.E. of mean*	*MAE*
5%	1.015	2.005	1.997	2.332	0.018	1.673
15%	0.989	5.330	1.388	5.463	0.002	5.001
30%	0.867	10.289	0.910	10.303	0.054	10.018

Source: The researcher from the applied study, SPSS Package, 2019

From the above table, the results revealed that MAE of EM method was lower than MAE of the MI method and the regression method. These results were consistent with the values of S.E. mean. Hence, based on those results, we concluded that **the EM method** is more efficient than the other two methods.

Note: SPSS * SPSS Missing Value Analysis1 is an additional module for SPSS (version 18) that provides graphical tools to investigate missing data, and imputes missing data using the MI, EM and regression imputation. (www.spss.com).

RESULTS

- Results obtained by descriptive statistics; the results revealed that there are no ostensible differences between means and variances.
- the results revealed that MAE of EM method was lower than MAE of the MI method and the regression method these results were consistent with the values of S.E. mean
- we concluded that the EM method is more efficient than the other two methods
- The results indicate that the three methods work reasonably well in many situations, particularly when the amount of missingness is low and when data are missing at random (MAR) and missing completely at random (MCAR).

RECOMMENDATION

- Further research can be conduct involving cross-validation
- Internal cross-validation strategies such as k-fold cross-validation, bootstrapping or subsampling are suggested to carry out after imputation of data
- Another problem with EM is that it leads to biased parameter estimates and underestimates the standard errors. For this reason, statisticians do not recommend EM as a final solution

REFERENCES

1. Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.
2. Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
3. Nakai M and Weiming Ke., 2011. Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. Vol. 5, no.1, 1-13.

4. Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J Roy Stat Soc Series B 1960; 22(2): 302–306.

5. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons, New York, 1987.

6. Brick JM, Kalton G. Handling missing data in survey research. Stat Meth Med Res 1996; 5: 215–238.

7. Hyun Kang. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May; 64(5): 402–406

8. Dempster A P, Laird N M, and Rubin D B 1977 Journal of the Royal Statistical Society Series B 39 (1) 1-38

9. Little R J A and Rubin D B 2002 Statistical Analysis with Missing Data Second Edition (Hoboken, New Jersey: John Wiley & Son Inc.)

10. Enders, C.K., Bandalos, D.L., 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modelling: A Multidisciplinary Journal 8, 430–457.

11. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–576. doi: 10.1146/annurev.psych.58.110405.085530

****