A COMPARATIVE STUDY OF MULTIPLE IMPUTATION AND MAXIMUM LIKELIHOOD METHODS OF IMPUTING MISSING DATA IN A RANDOMISED COMPLETE BLOCK DESIGN

Supported by tetfund

Segun Peter Alade[1]* and Abraham Okolo[2]
[1]Department of Statistics, School of Science & Technology, Plateau State Polytechnic, Barkin-Ladi, Nigeria
[2]Department of Statistics & Operations Research, Modibbo Adama University of Technology, Yola, Nigeria
Corresponding author: segunpeteralade@yahoo.com

Received: December 04, 2019 **Accepted:** February 13, 2020

Abstract: In this study, the multiple imputation and maximum likelihood methods of imputing missing data in a randomised complete block design are compared. The aim is to seek for a more efficient technique for imputing missing data. The Multiple Imputation (MI) method involves imputing missing values repeatedly in order to account for variability due to imputations, while the Maximum Likelihood (ML) method (EM algorithm) first takes the estimate of variances, covariances and means from listwise deletion. These estimates are then used to solve for regression coefficients and the estimation of missing data. Data was collected from the Department of Animal Health and Production Technology, NVRI Vom. The data was that of an experiment on the effect of temperature and storage length on protein content of table eggs. The MI and ML methods were compared at levels 4, 5, 10 and 15 missing observations at m=20, 30 and 40 imputations using SPSS version 25 for the analysis. It was observed that the ML method performed better than the MI method at four (4) missing observations, except for m=40 imputation. Apart from that situation, the MI method performed better than the ML in other levels of missing observations. It was concluded that the ML is more efficient when the number of missing observations are few, although the MI can perform equally efficiently for the same situation when the number of imputation is excessively increased. The MI method performs excellently better than the ML method when the number of missing observations are more than 10% of the entire number of observations.

Keywords: Multiple imputation, maximum likelihood, missingness, listwise

Introduction

Collecting, analysing and drawing conclusions or inferences from data are central to research in many fields of life. Unfortunately, for quite a number of reasons, it is hardly possible to collect all the intended data. Incomplete data or missing values has become a common phenomenon in most fields of life and this has become a common obstacle for a balanced data analysis. Some of the reasons for missingness of data include; failure to answer a survey question, drop-out, planned missing values, intermittent measurements, latent variables, equipment malfunctions, to mention just a few. In experimental work, it sometimes happens that the results of one or more observations are lost, animals may become sick or die during the process of experimentation, field plots maybe ravaged by fire or by some pest, or labelling maybe laid wrongly. No matter the amount of care taken, missing observations can still find its way into a data set.

Missing data is capable of rendering the analysis unbalanced and in some instances rendering it unfeasible. Handling missing values is very critical to drawing conclusions for a parameter of interest. Many methods of handling missing values inappropriately fail to account for the uncertainty due to missing values. This failure can lead to biased estimates and loss of efficiency. Most missing value techniques fail to identify the nature of missingness. These techniques assume data are missing at random, which is not usually the case. Data could be missing at random (MAR), missing completely at random (MCAR) or missing not at random (MNAR). The nature of missingness has a role to play in estimating the missing observation. It is only appropriate to ignore the process that causes missing data when the sampling distribution inferences about the parameter indicates that the missing data are missing at random and the observed data are observed at random, but generally, inferences on missing data are conditioned on the observed pattern of missingness (Rubin, 1976).

The MI method accounts for variability of the incomplete data. The method involves replacing the missing data set by more than one plausible set of values and then approximately combining the complete data estimates using specific combining rules. The method involves generating completed datasets and then the computation of the multiple-imputed values. The estimates of missing data are obtained by simulation of random draws from the distribution of the missing variables given the observed variables.

A lot of studies have investigated conditions under which missing data techniques are able to reduce bias and increase the efficiency and the sensitivity of the statistical analysis. Until the late 1980s, deletion methods (i.e., listwise and pairwise deletion) dominated the analysis of missing data. However, with later developments in statistical software, ML estimation methods have become readily available for handling missing data in unbalanced designs (Shin *et al.*, 2017). There are various methods of estimating missing values, some of the most commonly used methods are; complete cases, available cases, single value imputation, and more recent model-based methods, maximum likelihood for multivariate normal data, and multiple imputation (Pigott, 2000).

Most of the methods mentioned above, apart from the model based methods, do not take care of the pattern of missingness of an observation. The pattern of missingness of an observation has a lot to do with predicting the missing value (Rubin, 1976). The pattern of missingness and missingness mechanism has a lot to do with the validity of inferences drawn from incomplete data. Both deletion methods always require MCAR conditions. In addition, for listwise deletion, there are two MNAR special cases, regression imputation and stochastic regression imputation, which can yield unbiased estimates under MAR. For efficiency and consistency, the model needs to be correctly specified. LOCF is incapable of providing consistent estimates, even under MCAR. Listwise deletion produces standard errors that are correct for the subset of complete cases, but in general too large for the entire dataset. Calculation of standard errors under pairwise deletion is complicated. The standard errors after single imputation are too small since the standard calculations make no distinction between the observed data and the imputed data. Though correction factors for some situations have been developed according to Schaffer and Schenker (2000), but a more

convenient solution is multiple imputation (Buuren, 2012). It is in this light that this study seeks to explore the use of MI methods.

In this study, a comparison of the MI method with the maximum likelihood (ML) is undertaken. This is necessary because investigators have continued to argue for and against which method, between ML and MI is most appropriate in different situations. Some others have equally said the two methods are just as equally efficient (Graham, 2009). Allison (2012), reported that the two approaches that have good statistical assumptions are the maximum likelihood estimation and the multiple imputation technique. But he also argued that the maximum likelihood method is a better technique than the method of multiple imputation. This requires further investigation, especially with RCBD. There is therefore the need to compare the MI method to the ML method in order to determine which of them produces an efficient estimate when applied in an agricultural setting with RCBD layout.

The aim of this work is therefore to compare the multiple imputation and the maximum likelihood methods for use in a randomised complete block design (RCBD)with missing observations.

The objectives of this work is to determine a suitable imputation method for RCBD with replication, to access the precision of imputed value based on the adopted technique and to find the relative efficiency of the MI method in comparison to the ML methods.

Materials and Method

The multiple imputation method

The multiple imputation method which was introduced by Donald Rubin in 1987 is currently the prevailing method of estimating missing values (Garson, 2015). Buuren (2012) reported that it is almost the universally accepted method for estimating missing data nowadays and it also serve as a benchmark against which newer methods are being compared. In this technique, instead of filling the missing value with a single value, the missing value is replaced with a set of possible values. The idea behind this method is to fill in *m* plausible values for the missing data several times to account for model uncertainty. The reasoning behind the *m* replications of imputed values is to create m complete data sets each of which is to be analysed by standard complete data methods. To create these m plausible values, a regression model needs to be identified, this model would be used to create imputes based on other variables (predictors) in the data set. Here, m different but similar regression models would be identified to give different values that would be imputed to form m complete data sets that would each be analysed using standard procedures. This method retains the advantage of single imputation by using a complete data set and thereby allowing standard statistical procedure to be used for the analysis. Also by allowing more than one value on a missing variable to be estimated, MI corrects for sampling variability and thus improves upon the single imputation techniques that uses only a single value. The imputed *m* values on the variable of interest can therefore be aggregated to produce inferential results. In addition, random error in the imputation process yields approximately unbiased estimates of all parameters, which no deterministic method can perform. Also, repeated imputation allows for good estimates of the standard errors. Rubin (1987) argued that an important limitation of single imputation methods is that standard variance formulas applied to filled-in data systematically underestimated the variance of estimates, thus he proposed multiple imputation. In this method, the first step is to specify one encompassing multivariate model for the entire data set. There are four different types of multivariate complete data models, which are as follows: (i) normal model, which perform imputation under a multivariate normal distribution; (ii) log linear model, which has been traditionally used by social scientists to describe associations among variables in cross-classified data; (iii) general location model, which combines a log-linear model for the categorical variables with multivariate normal regression model for the continuous variables, and (iv) two level linear regression model, which is commonly applied to multi-level data. The chosen imputation model should be compatible with the subsequent analysis or to be precise, the model should be flexible enough to preserve the relationships among variables that will be the focal point of later analysis. Multiple imputations are similar to single imputation in that it imputes a set of likely values from a distribution for each missing variable (Yarandi, 2014). The normal model described as the best by Graham (2012), is the most implemented of all the models and it can handle a wide array of analytic problems. The normal model estimates without bias for, means, variances, covariances, and related quantities. But it does not give unbiased estimates for proportional estimations. Since the observed values are a random sample from the population, we can obtain consistent estimates of the mean and variance from the observed values alone, and use those estimates to impute the missing values with random draws from a normal distribution. We then analyze the mix of imputed and observed values as though it were complete.

Analysis with MI is a three-step process, firstly, one imputes the data, and generating m imputed data sets. With each data set, a different imputed value replaces each missing value. Early writers suggested that very few imputed data sets were required. However, more recent work has suggested that more imputations (e.g., *m*= 20 to 40 or more) are required to achieve the statistical power of equivalent with ML procedures (Graham, 2012). Though, Rubin and Schenker (1986) demonstrated that even in extreme cases where the proportion of missing information make up about one-third of the data set, no more than 5 replicates of the model provides efficient estimates.

Secondly, one analyses the m data sets with usual, complete data, procedures (statistical software package), saving the parameter estimates and standard errors from analysis of each data set.

Thirdly, one would combine the results to get MI inference. Following what are commonly known as Rubin's rules (Rubin, 1987), the two most important quantities for MI inference are the point estimate of the parameters of interest and the MI-based standard errors. Some of the important quantities to combine these results are described in Chowdhury (2014), as indicated in equations (2.1) to (2.6).

The point estimate for each parameter is simply the arithmetic average of that parameter estimate (e.g., a regression coefficient) over the *m* imputed data sets.

$$\hat{\theta}_{MI} = \sum_{k=1}^{m} \frac{\hat{\theta}_k}{m} \qquad (2.1)$$

The within imputation variance, $\hat{\sigma}_w^2$, is the average variance within the imputed data sets,

$$\hat{\sigma}_w^2 = \sum_{k=1}^{m} \frac{\hat{\sigma}_k^2}{m} \qquad (2.2)$$

The between imputation variance, $\hat{\sigma}_b^2$, is the sample variance of the parameter estimate over the m imputed data set

$$\hat{\sigma}_b^2 = \sum_{k=1}^{m} \frac{\left(\hat{\theta}_k - \hat{\theta}_{MI}\right)^2}{m-1} \qquad (2.3)$$

The total variance, $\hat{\sigma}_{MI}^2$, is the sum of the within and the between variances;

$$\hat{\sigma}_{MI}^2 = \hat{\sigma}_w^2 + \left(1 + \frac{1}{m}\right)\hat{\sigma}_b^2 \qquad (2.4)$$

We need to note that $\hat{\sigma}_b^2$ is the variance that is due to missing data. If there were no missing data, then the variance of the parameter over the m imputed data sets would be zero and $\hat{\sigma}_b^2$component of variance would be zero. The MI inference standard error is simply the root of$\hat{\sigma}_{MI}^2$.

The fraction of missing information (FMI) in Rubin's rules is given by;

$$FMI = \frac{\frac{r+2}{(v_m+3)}}{r+1} \qquad (2.5)$$

FMI represents the amount of information that is missing from a parameter estimate because of the missing data, in other words it is the same as the amount of missing data. Though this value in any analysis estimated FMI usually differ from the hypothetical value, the formula given above just gives an estimate.

The relative efficiency, *RE* of using the finite m imputation estimator, rather than using an infinite number for the fully efficient imputation, in units of variance, is approximately a function of m and λ;

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1} \qquad (2.6)$$

Multiple imputation has been shown to be robust to departures from normality assumptions and provides adequate results in the presence of low sample size or high rates of missing data. Undoubtedly, certain requirements must be met for multiple imputations to meet these desirable properties. First, the data must be missing at random (MAR) meaning that the missing data are dependent on the observed variables not the missing observations. Secondly, the selection of model used to generate the imputed values must be well suited with the subsequent analysis so that it can conserve the associations among variables that will be the focus of later analysis. Thirdly, the model used for the analysis must agree with the model used in the imputation (Wayman, 2003).

The maximum likelihood method

The principle of maximum likelihood is a fairly simple technique but it is not quite easy to carry out because of computational complexity. There are a number of ways to obtain maximum likelihood estimators, and one of the most common is called the Expectation-Maximization algorithm usually known as the EM algorithm. There are various versions of the EM algorithm (Graham, 2012). The version that would be discussed here is the EM algorithm for covariance matrices. The basic idea is simple enough, but the calculation is a bit complicated.

In order to solve the EM algorithm manually, we would first take estimates of the variances, covariances and means, perhaps from listwise deletion. We would then use those estimates to solve for the regression coefficients, and then estimate missing data based on those regression coefficients. For example, we would use whatever data we have to estimate $Y = bX + a$, and then use X to estimate Y wherever it is missing. This is the estimation step or "E-step" of the algorithm. Having filled in missing data with these estimates, we would then use the complete data (including estimated values) to recalculate the regression coefficients. This is the maximization step or "M step". Having new regression coefficients, we would re-estimate the missing data, calculate new coefficients, etc. We would continue this process until the estimates no longer change noticeably, at this point, it is said that EM has converged. Then we can say we have maximum likelihood estimates of the parameters, and we can use them to make the maximum likelihood estimates of the regression coefficients.

The solution from the EM algorithm is better than we can do with coding for missing data, but it will still underestimate the standard errors of the coefficients. There are alternative maximum likelihood estimators that will be better than the ones obtained by the EM algorithm, but they assume that we have an underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data (Azadeh, 2012). The EM algorithm assumes that data are multivariate normal and that missingness is MAR. The EM estimates of the mean vector and covariance matrix can then be used in multivariate analyses to obtain estimates of the model parameters and standard errors, to test hypotheses, and to score values for observations using the model selected. The problem with using the EM covariance estimates as input for subsequent analyses is that standard errors might still be biased because the covariance matrix is treated as though it came from complete data (Truxillo, 2004).

Allison (2012) reported that the ML is a great technique for handling missing data and that the most important advantage it has over MI is that there is no conflict between the imputation model and the analysis model. Also, both MI and ML are asymptotically efficient because they have minimum efficient but can attain full efficiency with infinite number of data sets, which is difficult to achieve. One further advantage ML has over the MI method is that it is not cumbersome to compute and it gives the same estimate each time it is computed but the MI gives different estimates for every time it is computed. Graham (2012) observed that MI is not a good method of imputing missing observations when it comes to analyses that lend themselves well to normal-method. Shin, Davison and Long (2017), reported that the ML does quite well with smaller sample sizes than the MI, this is similar to the opinion that MI does quite well when the percentage of missing data is more than 20 percent (Graham, 2012).

Method of data analysis

The data collected was used to exemplify the comparisons between the MI method and the ML of estimating missing observations in a two-way factorial design. The data were secondary collected from the department of Animal Health and Production Technology, National Veterinary Institute Vom. The responses obtained were from an experiment carried out on the effect of storage length and temperature on crude protein content of table eggs. The data simply known as protein content has four missing observations from a total of forty-eight (48) observations.

The data with k missing observation was analysed using a two-way analysis of variance technique with the model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

Where μ is the overall mean, α_j is the deviation of group j from the first factor to the overall mean, β_k is the deviation of group k from the second factor to the overall mean, $\alpha\beta_{jk}$ is the interaction between Factors 1 and 2, and ε_{ijk} is the error term that is assumed to be normally distributed with mean zero and variance σ^2. The estimates of interest are the mean squared error, (MSE) and the standard error (SE) of the mean of observations. Apart from the original data which had 4 missing observations, other data sets with 5, 10, and 15 randomly deleted observations were also analysed. For the MI method, the analysis was carried out at m=20, 30 and 40 imputations. The regression model with fully conditioned specification model was used to generate all imputed values. The ML analysis was also carried out using the EM algorithm method. The analysis of data was done with the aid of SPSS version 25.

The relative efficiency of the results obtained from the multiple imputation and the maximum likelihood methods of

analysis is computed by comparing the error variance using the ratio;

$$Relative\ Efficiency\ (A{:}B) = \frac{\sigma_A^2}{\sigma_B^2}$$

The relative efficiency can be expressed in terms of percentage by multiplying by 100. If Relative Efficiency (*A:B*) is greater than 100, it implies that *B* is more efficient or precise than A. If Relative Efficiency (*A:B*) is less than or equal to 100, it implies that B is less efficient to *A*.

Results and Discussions

Fraction of missing information

The FMI represented in Table 1, shows the amount of information missing from the parameter estimate. The lowest value being at m=40 when there are 4 missing observations and the highest at m=20 for 15 missing observations.

Table 1 shows the amount of missing information, the lower the value the better. The amount is higher at m=20 for 15 missing observations, this indicates that when the number of missing observation is becoming higher the number of imputations should also go higher so as to be able to recover a considerable amount of the missing information (Yuan, 2000). The lowest value was at m=40 when there are 4 missing observations, this seems to buttress the point made by Allison (2012) that MI can attain full efficiency when the number of imputation becomes infinitely high.

Table 1: Fraction of missing information over levels of imputations

Number of missing observations	*Number of Imputations*		
	m=20	***m=30***	***m=40***
4	0.240	0.196	0.160
5	0.162	0.244	0.242
10	0.202	0.254	0.260
15	0.564	0.322	0.442

Table 2: Efficiency of imputations

Number of missing observations	*Number of Imputations*		
	m=20	***m=30***	***m=40***
4	0.988	0.994	0.996
5	0.992	0.992	0.994
10	0.990	0.994	0.994
15	0.973	0.989	0.989

Efficiency of imputations

Table 2 shows the efficiencies of imputation at various levels of imputation and number of missing observations. Most of the values stood at about 99%, with the smallest value being 0.973 at m=20 for 15 missing observations, followed by m=20 at 4 missing observation. This is indicative that the MI performs well when the number of imputation is high and the number of missing observations is not so low. The efficiency seems dropping generally at 15 missing observations, which shows that the MI begins to lose its efficiency when the number of missing data is getting higher and higher; Garson (2015) suggested its use for missing data with percentages between 20 and 50. In other words, the MI should not be used as a substitute for data collection. Even though the MI begins to lose its efficiency as the number of missing observation increases, it still has a better performance than the ML. The error becomes larger and larger with the ML as the number of missing observations increases.

Tables 3 and 4 shows the relative efficiencies of the standard errors and mean squared errors respectively, for the MI to ML methods. It can be observed that the ML performed better than the MI at four (4) missing observations, except for when m=40. This is in agreement to the point that ML performs better than MI when the number of missing observations are very few just as reported in Shin *et al*. (2017). Apart from the already mentioned situations the MI performed better than the ML in other levels of missing observation.

Table 3 shows the standard errors obtained from both the MI and ML methods of missing data imputations. The relative efficiencies of the MI:ML were also computed; they are in parenthesis in each cell. A value less than 100% indicates that the MI is more efficient than the ML, if the relative efficiency is greater than 100% however, this indicates that the ML is more efficient.

Table 3: Relative efficiency of MI:ML for standard errors

Number of imputations	Number of missing observations			
	4 (%)	**5 (%)**	**10 (%)**	**15 (%)**
m=20	0.041 (102.5)	0.042 (84.0)	0.049 (98.0)	0.058 (89.2)
m=30	0.041 (102.5)	0.041 (82.0)	0.050 (100.0%)	0.063 (96.9)
m=40	0.038 (95.0)	0.043 (86.0)	0.049 (98.0)	0.058 (89.2)
ML	**0.040**	**0.050**	**0.050**	**0.065**

Table 4: Relative efficiency of MI:ML for mean squared errors

Number of imputations	Number of missing observations			
	4 (%)	**5 (%)**	**10 (%)**	**15 (%)**
m=20	0.061 (102.5)	0.063 (53.4)	0.085 (69.7)	0.117 (57.1)
m=30	0.073 (102.5)	0.058 (49.2)	0.083 (68.0)	0.118 (57.6)
m=40	0.103 (177.6)	0.060 (50.8)	0.078 (63.9)	0.118 (57.6)
ML	**0.058**	**0.118**	**0.122**	**0.205**

Table 4 shows the mean squared errors obtained from both the MI and ML methods of missing data imputations. The relative efficiencies of the MI:ML were also computed; they are in parenthesis in each cell. A values less than 100% indicates that the MI is more efficient than the ML, if the relative efficiency is greater than 100% however, this indicates that the ML is more efficient.

In summary, the result is indicative that the ML method has a lower standard error and MSE for four (4) missing observations except when the number of imputation was m=40 for standard error. This shows that ML is a more efficient technique to use than the MI when it comes to imputation of missing values with fewer missing observations say, less than 10%. But precision could be achieved with the MI method by increasing the number of imputations to m=40 or more. However, it has been observed that the MI method is preferable to the ML method as the number of missing observations becomes large, because it produces a lower standard error of the mean and also a lower MSE.

Conclusions

From the foregoing, both the ML and MI method of imputing missing observations in a RCBD can be adopted, but the ML is more efficient if the number of missing observations are few. The MI can only perform equally efficiently for few missing observations only if the number of imputation is excessively increased, but this is tedious. The MI method performs excellently better than the ML method when the number of missing observations are much more, although it should not be used as a substitute to complete data collection

because it loses its power as the number of missing observations increases say beyond thirty percent of the total expected observations.

Based on the result obtained in this work, it is recommended that the Maximum Likelihood method be adopted for imputation of missing observations in a two-way randomised complete block design with more than one observation per cell when missing items in the data are few, but when the missing observation are many, say 10% and above, the multiple imputation method should be adopted.

Conflict of Interest

Authors declare that there is no conflict of interest related to this study.

References

Allison PD 2012. Handling Missing Data by Maximum Likelihood. *Statistical Horizons*, Haverford, PA, USA.

Azadeh A & Saberi Z 2012. *Estimating The Missing Values in Analysis of Variance Tables by a Flexible Adaptive Artificial Neural Network and Fuzzy Regression Models*. Paper presented at the 6th International Day of Statistics and Economics, Prague.

Buuren S 2012. Flexible Imputation of Missing Data. Boca Raton: CRC Press.

Chowdhury A 2014. Multiple Imputation: Better than Single Imputation in Pain Studies? Cytel Statistical Software & Services India, SPO4.

Garson GD 2015. *Missing Values Analysis and Imputation*, (2015 edition), Asheboro, USA: Statistical Associates Publishing.

Graham JW 2009. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60: 549-576.

Graham JW 2012. *Missing Data: Analysis and Design*. Statistics for Social and Behavioural Sciences. Springer Science and Business Media, New York.

Piggott TD 2000. A review of methods for missing data. *Edu. Res. and Evaluation*, 7(4): 353-383.

Rubin DB 1976. Inference and missing data. *Biometrika*, 63(3): 581-592.

Rubin DB 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons Inc:

Schaffer JL & Schenker N 2000. Inference with imputed conditional means. *J. Am. Stat. Assoc.*, 95(449): 144-154.

Schenker N & Rubin DB 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Stat. Assoc.*, 81: 366-374.

Shin T, Davison ML & Long JD 2017. Maximum likelihood versus multiple imputation for missing data in small Longitudinal samples with non-normality. *Psychological Methods*, 22(3): 426-449.

Truxillo C 2016. A Comparison of Missing Data Handling Methods. Institute Inc, Cary, NC.

Van Ginkel JR & Kroonenberg PM 2014. Analysis of variance of multiply imputed data. *Multivariate Behavioural Research*, 49: 78-91.

Wayman, J. C. (2003), Multiple imputation for missing data: What is it and how to use it? A paper presented at the 2003. Annual Meeting of the American Educational Research Association, Chicago, IL.

Yarandi HN 2012. Handling Multiple Data with Multiple Imputation Using PROC MI in SAS. University of Florida, Gainesville, Florida.

Yuan CY 2000. Multiple Imputation for Missing Data: Concepts and New Development, SAS Users Group International Proceedings.