International Journal of Mathematical Analysis and
Optimization: Theory and Applications
Vol. 2020, No. 1, pp. 733 - 740

# Multiple Imputation: An Iterative Regression Imputation

Traore, Bintou[1*], Adeleke, A. Ismaila[2]

1*, Department of Mathematics, University of Lagos, Lagos, Nigeria.
2, Department of Actuarial Science and Insurance, University of Lagos, Lagos, Nigeria.
*Corresponding author: bintou.traore64@yahoo.fr

**Article Info**
Received: 22 April 2020 Revised: 10 July 2020
Accepted: 10 July 2020 Available online: 19 July 2020

**Abstract**

Multiple imputation (MI) is a commonly applied method of statistically handling missing data. It involves imputing missing values repeatedlyto account for the variability due to imputations. There are different techniques of MI that have proven to be effective and available in many statistical software packages. However, the main problem that arises when statistically handling missing data, namely, bias, still remains. Indeed, as multiple imputation techniques are simulation-based methods, estimates of a sample of fully complete data may substantially vary in every application using the same original data and the same implementation method. Therefore, the uncertainty is often under- or overestimated, exhibiting poor predictive capability. A new approach of MI based on regression method is presented. The proposed approach consists of constructing a possible lower and upper bound around the sum of square of residuals (SSE) that would have been obtained in a complete case (that is, if there were no missing data). Then, iteratively implement regression imputation (RI) to replace the missing values and compute a new SSE with fully completed data. If the new SSE does not fall within the constructed bounds, the RI method is repeated until the SSE estimated falls into those bounds.The SSEs of the prediction are used to assess the performance of the proposed approach compared to expectation-maximization (EM) imputation and multiple imputation by chained equations (MICE). 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). However, when the proportion of missingness is severe and the data are missing not at random (MNAR), the proposed method performs better than MICE and EM algorithms.

**Keywords And Phrases:** Missing data, multiple imputation, sum of squares of residues, algorithms.
**MSC2010: 47H09; 47H10**

## 1 Introduction

Multiple imputation (MI) is a highly praised simulation-based method to provide consistent and asymptotically efficient estimates for the statistical analysis of missing data. This method, first proposed by Rubin (1986) to impute missing data while solving some of issues, relies on the efficiency of classical missing data handling methods such as case deletion and single imputation. Indeed, case deletion and single imputation are known to be sensitive to missing data mechanisms (MCAR, MAR, and MNAR) and to underestimate the standard error,leading to an overestimation of test statistics (Schafer and Graham, 2002; Rubin, 1996). Multiple imputation addresses these issues and provides more consistent estimates by increasing the number of imputations to reduce bias in the standard error introduced by the additional uncertainty due to imputations (Allison, 2002; Rubin, 1996; Schafer and Graham, 2002); Little and Rubin, 2002). In addition, unlike other methods, MI tends to be less sensitive to the different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1987). Various methods of multiple imputation have been developed to handle missing data in different circumstances. These methods include the expectation-maximization (EM) algorithm, multiple imputation by chained equations (MICE) based on a MonteCarlo Markov chain (MCMC) algorithm, the imputation-posterior (IP) method and the multiple imputation bypredictive mean matching (PMM) technique (Dempster et al., 1977; Rubin, (1986, 1987); Oudshoornet al., 1999); King et al., 2001; White et al., 2011; Azur et al. 2011; Morris et al.,2014; and Kleinke, 2018). However, as multiple imputation techniques are inherently simulation-based methods, estimates of a sample of multiplyimputed data may substantially vary in every application using the same original data and the same implementation method (Nakai and Weiming, 2011; Hippel, 2018). Therefore, uncertainty is often under- or overestimated, exhibiting poor predictive capability. The determination of the full additional uncertainty is not straightforward. In addition, the discrepancy between the true and the estimated parameters becomes considerably large as the fraction of missing data increases. A possible reduction in this bias requires much more imputation, which requires more resources to generate, store and analyze the multiplyimputed data.

The present work proposes a new MI approach that addresses these issues by avoiding or at least reducing bias and improving precision. In contrast to existing MI techniques, the proposed approach consists of constructing a possible lower and upper bound around the sum of square of residuals (SSE) that would have been obtained in a complete case (that is, if there were no missing data). Then,iteratively implement regression imputation (RI) to replace the missing values and compute a new SSE with fully completed data. If the new SSE does not fall within the constructed bounds, the RI method is repeated until the SSE estimated falls into those bounds. For a multiple imputation process, this procedure is repeated for a predefined number of times. The rest of this paper is organized as follows: Section 2 provides a brief description of MICE and EM algorithms and presents the proposed method with a detailed discussion of the framework. An illustrative example using real data and the conclusion are given in Sections 3 and 4, respectively.

## 2 Methodology

### 2.1 Brief Description of MICE and EM Algorithms

The MICE procedure fits a regression model for each variable having missing data and uses fully observed variables as covariates. In cases where all variables have missing values, the procedure initially fills in all missing variables at random and then regresses each missing variable on the other fully observed variables. Missing values are imputed using posterior predictive distribution (see Azur et al., 2011); Raghunathan et al., 2001; Van Buuren, 2007).

The EM algorithm (Dempsteret al., 1977) is a general method for obtaining maximum likelihood estimates;it involves two steps: the E-step and the M-step. The first essentially calculates the expected values of the complete-data sufficient statistics given the observed data, $X_{obs,i}$ and current estimates $\Theta^t = (\mu, \sigma^2)$. The second step computes new parameter estimates, $\Theta^{t+1} = (\mu^{t+1}, \sum^{t+1})$ where $\mu_j^{t+1} = \sum_{i=1}^{n} \frac{x_{ij}^t}{n}$ and $\sigma_{jk}^{t+1} = \frac{1}{n}\sum_{i=1}^{n}\{(x_{ij}^t - \mu_j^{t+1})(x_{ik}^t - \mu_k^{t+1}) + \gamma_{jki}^t\}$ .

The algorithm iteratively proceeds between the E-step and the M-step until the discrepancy between $\Theta^t$ and $\Theta^{t+1}$ converges to a specified criterion. At the final E-step, imputed values are supplied for missing values.

## 2.2 Proposed Method

Basically, imputation-based techniques involve replacing missing values using available observations. The purpose of imputation is to provide consistent test statistics, which means providing a sampling variance that is as close as possible to the sampling variance without missing data. The aim of our method is to improve accuracy by constructing a limit around the true SSE even though it is unknown.Indeed, the main idea is to restrict the imputation to values for which the SSE is as close as possible to the true SSE. The method requires at least one fully observed variable and can be applied to any missing data pattern. For the first step of iteration, a regression model is fitted for each variable having missing values, and the estimation is restricted to individuals with observed values. Then, missing values are replaced by the predicted values increased with residuals drawn from a normal distribution. For the remaining iterations, new values are imputed with respect to the observed values and current imputed values for the missing data. In each iteration, missing values are replaced under the condition that the SSE obtained from the completed data set is within the constructed interval. The proposed iterative method can be summarized as follows:

Step 1: Define the number of missing and non-missing variables.

Step 2: Fit a regression model with available observations:
$Y_j^{obs} = \beta_0 + \sum_{k=1}^{q} \beta_k X^{obs} + \epsilon_j$
where $Y_j^{obs}$ are the available part of the missing variables $Y_j (j = 1, 2, ..., p); X_k^{obs}$ are the available part of the fully observed variables; p and q are the number of missing and nonmissing variables, respectively; $\beta_0$ and $\beta_k$ are the coefficients of the regression; and $\epsilon_j$ is the residual $(\epsilon_j \sim N(0, \sigma^2))$. Compute the corresponding sum of squares of residuals, $SSE_{obs}^j$:
$SSE_{obs}^j = \sum_{i=1}^{n_{obs}^j} \widehat{\epsilon}_{ij}^2 = \sum_{i=1}^{n_{obs}^j} Y_{ij}^{obs} - \widehat{Y}_{ij}^{obs}, \quad i = 1, 2, ..., n_{obs}^j$

Step 3: Use the estimated regression coefficients $\widehat{\beta} = \widehat{\beta_0}, \widehat{\beta}_{j-1}$ to replace the missing values, $Y_j^{mis}$
$Y_j^{mis} = \widehat{\beta}_0 + \sum_{j=1}^{q} \widehat{\beta}_{j-1} X_k^{mis} + \widehat{\epsilon}_j$ where $\widehat{\epsilon}_j$ is drawn from $N(0, \widehat{\sigma}_j, \upsilon)$ with $\widehat{\sigma}_j$ being the sum of squares of residuals for observed data, and $\upsilon$ is generated from a chi-square distribution with *df* degrees of freedom.

Step 4: Construct an approximate limit around the true sum of squares of residuals. This is done as follows:
i. Compute: $SSE_{ref}^j = \frac{SSE_{obs}^j}{c_j}$ with $c_j = \frac{n_{obs}^j}{n}$ the sample size and $n_{obs}^j$ the number of non-missing values in the corresponding variable $Y_j$.
ii. Compute $\delta_j = c_j(1 - c_j) + 0.05$. Then, generate a sequence $r_{ij}$ from $c_j - \delta_j$, with 0.01 as the increment of the sequence.
iii. Calculate the quantities $SSE_{r_l}^j = \frac{SSE_{obs}^j}{r_l}$ with $r_l \in R_j$ (s is the length of the sequence).
Set B as the set of $SSE_{r_l}^j <$ the integer rounding of $SSE_{ref}^j + \frac{1}{2} SSE_{ref}^j$.
Set $SSE_{low}^j$ as the mean of the set of B less than the integer rounding of $SSE_{ref}^j$.
Set $SSE_{up}^j$ as the mean of the set of B greater than the integer rounding of round $SSE_{ref}^j$.

Step 5: Fit a regression model for the fully complete data. If the corresponding
$SSE^j$ does not fall into the interval $[SSE_{low}^j - SSE_{up}^j]$ repeat steps 2-4 until this condition is met.

Step 6: For each missing value, draw new $\widehat{\epsilon}_{lj}$ from $N(0, \widehat{\sigma}_{lj}.v)$, with $\widehat{\sigma}_{lj}$ being the sum of squares of residuals ($SSE^{ij}$) for the current fully complete data, and add to the initial predicted value, $Y_{ij}^{mis}$.

Step 7: Repeat steps1-5 for each missing variable for a fixed number of times.

# 3 Results

The dataset used in this study is the estimate of government effectiveness collected by the Word Bank for 213 countries in the world over 17 years. Originally, the dataset contained missing values, but we took the complete observations available (n=182), almost ignoring the possible dependencies of the missing values in the data. As variables, we used the estimate of government effectiveness collected over 1996, 2003, 2007, and 2010, with the first two years being predictors and the remaining two years being missing variables. Missing values were generated under the three main missing data mechanisms (MCAR, MAR and MNAR) using R software with the "ampute" function included in the MICE package. For purposes of demonstration, each missing value is imputed five times for each missing variable using EM, MICE and the proposed method, and the results are presented in Table 1 and 2.

Table 1 and2 show the sum of squares of residuals arising from the use of the three different techniques under the condition that the data are MCAR, MAR and MNAR. Column 2 provides the number of missing values in each variable (Y1 and Y2), while columns 3 and 4 give the constructed bounds (lower and upper) around the true SSE in column 5. The three remaining columns show the SSE arising from the three imputation techniques: MICE, EM and PM.

**Table 1: Comparison of three imputation techniques under MCAR based on SSE for the first variable Y1 .**

| $Y1$ | $N.mis$ | $Lower$ | $Upper$ | $True$ | $MICE$ | $EM$ | $PM$ |
|---|---|---|---|---|---|---|---|
| **MCAR** | **5** | **8.6735** | **10.1258** | **9.3151** | **9.7395** | **9.7722** | **9.36** |
| | **10** | **8.5599** | **10.5904** | **9.3151** | **10.3643** | **10.2819** | **9.4595** |
| | **18** | **8.4089** | **11.3886** | **9.3151** | **10.5842** | **10.2833** | **9.9149** |
| | **36** | **7.8078** | **13.2014** | **9.3151** | **11.89947** | **12.2057** | **9.1846** |
| | **55** | **7.8226** | **11.2932** | **9.3151** | **11.6032** | **12.4745** | **8.9501** |
| | **73** | **6.6857** | **10.2436** | **9.3151** | **9.7966** | **10.8701** | **8.0189** |
| | **91** | **7.2597** | **11.658** | **9.3151** | **14.9306** | **12.716** | **8.4517** |
| | **109** | **5.2714** | **8.8164** | **9.3151** | **10.398** | **9.8927** | **7.9908** |
| | **127** | **6.6419** | **11.7138** | **9.3151** | **13.5721** | **17.3383** | **10.1268** |
| **MAR** | **5** | **8.7966** | **10.2696** | **9.3151** | **9.5222** | **9.6094** | **9.2506** |
| | **10** | **8.5541** | **10.5833** | **9.3151** | **9.8496** | **9.8428** | **9.3514** |
| | **18** | **7.5688** | **10.2508** | **9.3151** | **9.8847** | **9.9969** | **8.9988** |
| | **36** | **7.6738** | **12.9748** | **9.3151** | **11.1334** | **12.0294** | **9.4578** |
| | **55** | **7.7929** | **11.2607** | **9.3151** | **11.8589** | **11.5096** | **9.1952** |
| | **73** | **9.0928** | **13.7472** | **9.3151** | **14.3103** | **14.8817** | **10.8612** |
| | **91** | **6.3904** | **10.262** | **9.3151** | **11.2638** | **10.7526** | **8.5284** |
| | **109** | **8.4276** | **14.0008** | **9.3151** | **16.6353** | **17.8809** | **11.3495** |
| | **127** | **5.8627** | **10.3734** | **9.3151** | **17.714** | **16.8891** | **8.7946** |
| **MNAR** | **5** | **8.7843** | **10.2552** | **9.3151** | **9.738** | **9.9855** | **9.3294** |
| | **10** | **7.5983** | **9.4007** | **9.3151** | **8.813** | **8.9204** | **8.5477** |
| | **18** | **7.7749** | **10.5299** | **9.3151** | **9.9857** | **9.9386** | **9.0218** |
| | **36** | **7.8203** | **13.2225** | **9.3151** | **11.4748** | **10.9633** | **9.335** |
| | **55** | **6.9131** | **10.2094** | **9.3151** | **10.6906** | **9.1451** | **8.5186** |
| | **73** | **6.6697** | **10.2062** | **9.3151** | **11.5968** | **12.0234** | **8.3331** |
| | **91** | **8.569** | **13.5828** | **9.3151** | **19.7374** | **22.4621** | **10.0657** |
| | **109** | **8.4264** | **13.9988** | **9.3151** | **21.8685** | **18.6521** | **11.4969** |
| | **127** | **6.5209** | **11.2585** | **9.3151** | **17.0477** | **15.7509** | **10.3996** |

**Table 2: Comparison of three imputation techniques under MCAR based on SSE for the first variable Y2.**

| $Y2$ | $Lower$ | $Upper$ | $True$ | $MICE$ | $EM$ | $PM$ |
|---|---|---|---|---|---|---|
| **MCAR** | **12.3243** | **14.388** | **13.3086** | **13.3737** | **13.5326** | **13.2585** |
| | **12.3014** | **15.2194** | **13.3086** | **14.0902** | **13.8183** | **14.1601** |
| | **11.7613** | **15.9287** | **13.3086** | **15.5506** | **15.1741** | **13.4568** |
| | **10.6571** | **18.019** | **13.3086** | **15.3601** | **16.5956** | **12.9549** |
| | **11.4209** | **16.6117** | **13.3086** | **18.3583** | **17.1985** | **13.7546** |
| | **9.9318** | **15.1793** | **13.3086** | **14.3128** | **14.6456** | **12.0647** |
| | **11.8683** | **18.4889** | **13.3086** | **18.6878** | **19.1183** | **14.5525** |
| | **9.2833** | **15.1176** | **13.3086** | **14.4278** | **18.4212** | **13.7646** |
| | **10.7228** | **18.4555** | **13.3086** | **21.9708** | **31.6509** | **16.8492** |
| **MAR** | **12.5535** | **14.6555** | **13.3086** | **13.709** | **13.5618** | **13.3505** |
| | **12.2666** | **15.1764** | **13.3086** | **14.2115** | **13.5344** | **14.0913** |
| | **11.1989** | **15.1671** | **13.3086** | **13.8068** | **13.6584** | **12.889** |
| | **11.3107** | **19.1241** | **13.3086** | **16.4912** | **16.9257** | **13.7574** |
| | **11.8632** | **17.1424** | **13.3086** | **16.9726** | **16.5702** | **13.4449** |
| | **13.2819** | **20.033** | **13.3086** | **23.0195** | **21.8117** | **14.7588** |
| | **10.1789** | **16.1347** | **13.3086** | **14.4793** | **15.6275** | **14.9975** |
| | **12.2084** | **19.8811** | **13.3086** | **21.1317** | **20.7962** | **15.9713** |
| | **11.4998** | **19.7928** | **13.3086** | **26.9311** | **21.291** | **15.0145** |
| **MNAR** | **12.3138** | **14.3757** | **13.3086** | **13.5481** | **13.6318** | **13.2058** |
| | **11.3266** | **14.0134** | **13.3086** | **12.6284** | **12.9213** | **13.301** |
| | **11.8877** | **16.1001** | **13.3086** | **14.6709** | **15.403** | **13.4561** |
| | **11.7398** | **19.8495** | **13.3086** | **17.7909** | **18.1661** | **14.0918** |
| | **10.5839** | **15.5549** | **13.3086** | **17.1281** | **17.0028** | **12.9565** |
| | **9.9369** | **15.187** | **13.3086** | **15.0493** | **16.2449** | **12.9912** |
| | **11.0973** | **17.5623** | **13.3086** | **22.5529** | **22.9957** | **14.0125** |
| | **11.5177** | **19.1345** | **13.3086** | **28.9392** | **25.3928** | **14.5682** |
| | **11.0017** | **18.8785** | **13.3086** | **21.473** | **22.0808** | **13.8547** |

## 4 Discussion and conclusion

In this work, we proposed an iterative method based on regression for the imputation of missing values. The proposed method is effective only if : (i) the chosen regression model describe adequately the data under study, and (ii) the increment of the sequence used to construct the lower and upper bounds is very small (0.01); otherwise, it will be very likely to obtain bounds that do not include the true SSE. We used data sets from real life to evaluate the performance of the proposed method compared to other imputation methods, such as EM and MICE algorithms. Some elements are removed from these data matrices following the three main missing data mechanisms (MCAR, MAR and MNAR), and the number of removed data varies from 5 to 127. The removed data are replaced five times for each variable, and the mean of the SSEs obtained from the individual analysis of the multiply imputed data is used for the comparison.

Through the results, we find that PM can perform either like or better than EM and MICE in estimating missing values. With respect to the sum of squares of errors (SSE), it is confirmed that the three methods work reasonably well in many situations, with slight deviation from the true SSE. However, this deviation becomes substantially large as the degree of missingness increases and under the MNAR mechanism. Nonetheless, even in such a situation, PM seems to be better than EM and MICE. Indeed, we observe that the lower and upper bounds of SSE estimated are close to the true SSE under the three missing data mechanisms and that PM always provide SSEs within these bounds. However, MICE and EM tend to provide SSEs that are considerably different than the true SSE when the number of missing data increases and under MNAR.

## References

[1] **Allison P. D (2001). Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences. 07-136. Thousand Oaks, CA: Sage.**

[2] **Allison, P. D. (2002). Missing data. Thousand Oaks, CA: Sage**

[3] **Carpenter J. R and Kenward MG (2013). Multiple Imputation and its Application. John Wiley and Sons Ltd, West Sussex.**

[4] **Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J, (2011). Multiple imputation by chained equations: what is it and how does it work?** *International Journal of Methods in Psychiatric Research,* **20 (1):40-9.**

[5] **Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion).** *Journal of the Royal Statistical Association,* **39(1), 1-38.**

[6] **Hippel, P. T. V (2018). How many imputations do you need? A two-stage calculation using a quadratic rule.** *Sociological Methods and Research*, **1-20.**

[7] **King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation.** *American Political Science Review*, **95(1), 49–69.**

[8] **Kleinke, K. (2018). Multiple imputation by predictive mean matching when sample size is small. Methodology:** *European Journal of Research Methods for the Behavioral and Social Sciences,* **14(1), 3-15.**

[9] **Little, R. J. A. and Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.**

[10] **McKnight, E. P, McKnight K. M, Sidani S. and Figueredo A. J. (2007). Missing data: A gentle introduction.** *The Guilford Press A Division of Guilford Publications, Inc*. **72 Spring Street, New York, NY 10012.**

[11] **Morris, T. P, White, I. R and Royston, P., (2014). Tuning multiple imputation by predictive mean matching and local residual draws.** *BMC Medical Research Methodology,* **14(1), 75.**

[12] **Nakai, M. and Weiming, K. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis.** *Int. Journal of Math. Analysis*, **5(1), 1 - 13.**

[13] **O'Kelly, Michael and Ratitch. B (2014). Clinical Trials with Missing Data. United Kingdom: John Wiley and Sons, Ltd**

[14] **Oudshoorn, C. G. M., van Buuren, S. and van Rijckevorsel, J. L. A. (1999). Flexible multiple imputation by chained equations of the AVO-95 survey. TNO PreventieenGezondheid, TNO/PG 99.045.**

[15] **Raghunathan. T (2015). Missing data in practice. Chapman and Hall/CRC Press, Boca Raton, London, New York.**

[16] **Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: JohnWiley and Sons.**

[17] **Rubin, D.B. (1996). Multiple imputation after 18- years.***Journal of the American Statistical Association*, **91(434), 473-489.**

[18] **Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art.** *Psychological Methods*, **7(2), 147–177.**

[19] **Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective.** *Multivariate Behavioral Research,* **33(4), 545–71.**

[20] **Templ M., Kowarik A. and Filzmoser P. (2011). Iterative stepwise regression imputation using standard and robust methods.** *Computational Statistics and Data Analysis*, **55 (10), 2793–2806.**

[21] **Van Buuren S (2012). Flexible Imputation of Missing Data. Chapman and Hall/CRC Press, Boca.**

[22] **White, I. R., Royston, P. and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice.** *Statistics in Medicine*, **30(4), 377–399.**