Sains Malaysiana 46(2)(2017): 317–326
*http://dx.doi.org/10.17576/jsm-2017-4602-17*

# Missing Value Estimation Methods for Data in Linear Functional Relationship Model

(Kaedah Menganggar Data Lenyap menggunakan Model Linear Hubungan Fungsian)

ADILAH ABDUL GHAPOR, YONG ZULINA ZUBAIRI* & A.H.M. RAHMATULLAH IMON

ABSTRACT

*Missing value problem is common when analysing quantitative data. With the rapid growth of computing capabilities, advanced methods in particular those based on maximum likelihood estimation has been suggested to best handle the missing values problem. In this paper, two modern imputing approaches namely expectation-maximization (EM) and expectation-maximization with bootstrapping (EMB) are proposed in this paper for two kinds of linear functional relationship (LFRM) models, namely LFRM1 for full model and LFRM2 for linear functional relationship model when slope parameter is estimated using a nonparametric approach. The performance of EM and EMB are measured using mean absolute error, root-mean-square error and estimated bias. The results of the simulation study suggested that both EM and EMB methods are applicable to the LFRM with EMB algorithm outperforms the standard EM algorithm. Illustration using a practical example and a real data set is provided.*

*Keywords: Bootstrap; expectation-maximization; linear functional relationship model; missing value*

ABSTRAK

*Data lenyap sering terjadi dalam analisis data kuantitatif. Dengan berkembangnya keupayaan pengiraan, kaedah terkini iaitu kaedah kebolehjadian maksimum merupakan antara cara yang terbaik untuk menguruskan masalah data lenyap. Di dalam kertas ini, dua kaedah gantian moden diperkenalkan iaitu jangkaan pemaksimuman (EM) dan jangkaan pemaksimum bootstrap (EMB) untuk digunakan di dalam model linear hubungan fungsian (LFRM) iaitu LFRM1 bagi model penuh dan LFRM2 bagi model linear hubungan fungsian apabila parameter kecerunan dianggarkan menggunakan kaedah bukan berparameter. Prestasi EM dan EMB diukur berdasarkan purata ralat mutlak, punca purata kuasa dua ralat, dan anggaran terpincang. Melalui simulasi, kami dapati EM dan EMB kedua-duanya boleh digunakan oleh LFRM dan keputusan menunjukkan bahawa algoritma EMB adalah lebih baik daripada algoritma EM. Kajian ini disertakan dengan contoh data set yang sebenar.*

*Kata kunci: Bootsrap; data lenyap; jangkaan pemaksimum; model linear hubungan fungsian*

## INTRODUCTION

The presence of missing value is unavoidable in all fields of quantitative research, such as in the field of economics (Takahashi & Ito 2013), medical (Dziura et al. 2013), environmental (Razak et al. 2014; Zainuri et al. 2015), life sciences (George et al. 2015) and social sciences (Acock 2005; Schafer & Graham 2002). It has been established that ignoring missing values may result in biased estimates and invalid conclusions (Guan & Yusoff 2011). In short, inadequate approach of handling missing data in a statistical analysis will lead to erroneous estimates and incorrect inferences.

In general terms, techniques to deal with missing values can be categorised as traditional or modern approach. Some commonly used traditional ways are listwise deletion and pairwise deletion. As for imputation methods, mean imputation, hot-deck imputation, and stochastic imputation are among the commonly used (George et al. 2015). On the other hand, the modern approaches include those based on maximum likelihood and multiple imputations (Acock 2005). EM algorithm is an example of maximum likelihood and some examples of multiple imputations include Markov Chain Monte Carlo (MCMC), Fully Conditional Specification (FCS) and EMB algorithm (Baraldi & Enders 2010; Barzi & Woodward 2004; Gold & Bentler 2000; Little & Rubin 1987).

Studies on handling missing values are largely for univariate or regression model data. In this paper, we investigate the application of the EM and EMB methods in dealing with missing values for a type of model called the linear functional relationship model (LRFM). EM algorithm has become popular in handling missing data because of its simplicity and its wide applicability (Dempster et al. 1977). However, a major drawback of using EM is its slow convergence rate (Couvreur 1996). In order to improvise the existing EM method, we propose a bootstrap version of EM, known as the EMB. As EMB involves multiple imputation, we anticipate it will make the estimation less bias and will increase its efficiency when dealing with the missing data. A LRFM is employed to compare two

sets of data with both observable errors. The parameter estimates of LRFM can be obtained by maximum likelihood estimates, which we refer the full model LFRM with the acronym LFRM1 and when the slope parameter is estimated using nonparametric approach, which we refer it with the acronym LFRM2.

Thus, in this study, we aimed to investigate the performance of two modern approaches in dealing with missing data namely the EM and EMB for data sets that can be modelled by the LFRM1 and LFRM2. Simulation studies are done to investigate this study and illustration is provided using a practical example.

## THE MODEL AND EXPECTATION MAXIMIZATION TECHNIQUES

### LINEAR FUNCTIONAL RELATIONSHIP MODEL FOR FULL MODEL

Linear Functional Relationship Model can be expressed by:

$$Y = \alpha + \beta X, \tag{1}$$

where both variables $X$ and $Y$ are linearly related but observed with error. Parameter $\alpha$ is the intercept value, and $\beta$ is the slope parameter. For any fixed $X_i$ we observe $x_i$ and $y_i$ from continuous linear variable subject to errors $\delta_i$ and $\varepsilon_i$ respectively, i.e.

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i, \tag{2}$$

where the error terms $\delta_i$ and $\varepsilon_i$ are assumed to be mutually independent and normally distributed random variables, i.e.

$$\delta_i \sim N(0, \sigma_\delta^2) \text{ and } \varepsilon_i \sim N(0, \sigma_\delta^2). \tag{3}$$

To avoid an unbounded problem in our equation, we assume an additional constraint $\sigma_\delta^2 = \lambda\sigma_\delta^2$, where $\lambda$ is known (Sprent 1969). Therefore, the log likelihood function can be given by:

$$\log L(\alpha,\beta, \sigma_\delta^2, X_1, \ldots, X_n; \lambda, x_1, \ldots, x_n, y_1, \ldots, y_n) =$$

$$-n \log(2\pi) - \frac{n}{2}\log\lambda - n\log\sigma_\delta^2 -$$

$$\frac{1}{2\sigma_\delta^2}\left\{\sum(x_i - X_i)^2 + \frac{1}{\lambda}\sum(y_i - \alpha - \beta X_i)^2\right\}. \tag{4}$$

There are $(n + 3)$ parameters to be estimated, namely $\alpha, \beta, \sigma_\delta^2$ and $X_1, \ldots, X_n$, the incidental parameters respectively. Differentiating with respect to parameters $\alpha, \beta, \sigma_\delta^2$ and $X_i$, we can obtain $\hat{\alpha},\hat{\beta},\hat{\sigma}_\delta^2$ and $\hat{X}_i$ given by:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + \left\{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2\right\}^{\frac{1}{2}}}{2S_{xy}},$$

$$\hat{\sigma}_\delta^2 = \frac{1}{(n-2)}\left\{\sum(x_i - \hat{X}_i)^2 + \frac{1}{\lambda}\sum(y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i)^2\right\},$$

and $\hat{X}_i = \dfrac{\lambda x_i + \hat{\beta}(y_i - \hat{\alpha})}{\lambda + \hat{\beta}^2}$,

where $\bar{y} = \frac{1}{n} \pm \sum y_i$, $\bar{x} = \frac{1}{n}\sum x_i$,

$S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$, and

$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$. (5)

Further details of the parameter estimation can be found in the literature (Al-Nasser 2005; Kendall & Stuart 1973).

### LINEAR FUNCTIONAL RELATIONSHIP MODEL WITH ASSUMED KNOWN SLOPE

From (4), by differentiating $\log L$ with respect to parameters $\alpha, \beta, \sigma_\delta^2$ and $X_i$ we can obtain $\hat{\alpha},\hat{\beta},\hat{\sigma}_\delta^2$ and $\hat{X}_i$ as given in (5). Alternatively, the parameter $\hat{\beta}$ can be obtained first using nonparametric estimation (Ghapor et al. 2015). The steps involved in finding the slope using the nonparametric method are as follows:

*Step 1:* Arrange the observations in ascending order, based on $x$ value, i.e., $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. The associated values of which may not be in ascending order are taken, i.e., $y_{[1]}, y_{[2]}, \ldots, y_{[n]}$. The new pairs will be $(x_{(i)}, y_{[i]})$.

*Step 2:* All the data are divided into $m$ - subsamples. These subsamples contains $r$ elements, such that $m \times r = n$. The samples are then arranged in the following form:

$$\begin{matrix} \left(x_{(1)}, y_{[2]}\right) & \left(x_{(2)}, y_{[1]}\right) & \cdots & \left(x_{(r)}, y_{[r]}\right) \\ \left(x_{(1)}, y_{[2]}\right) & \left(x_{(1)}, y_{[2]}\right) & \cdots & \left(x_{(1)}, y_{[2]}\right) \\ \vdots & \vdots & \vdots & \vdots \\ \left(x_{((m-1)+(r+1))}, y_{[(m-1)+(r+1)]}\right) & \cdots & \cdots & \left(x_{(mr)}, y_{[mr]}\right) \end{matrix},$$

where $m$ is the maximum divisor of $n$, such that $m \le r$; as an example, if $n = 50$, then $m = 5$ and $r = 10$ respectively.

*Step 3:* Find all the possible paired slopes.

$$\left\{b_x(k)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}; i = 1,2,\ldots,j-1; j = 2,3,\ldots,r\right\}; k = 1,2,\ldots,m.$$

*Step 4:* Repeat Steps 1 to 3 by interchanging $y$ and $x$ to get another possible paired slopes of $b_y(k)_{ij}$.

$$\left\{b_y(k)_{ij} = \frac{y_{(j)} - y_{(i)}}{x_{[j]} - x_{[i]}}; i = 1,2,\ldots,j-1; j = 2,3,\ldots,r\right\}; k = 1,2,\ldots,m.$$

*Step 5:* Find the median of all these slopes.

$$\hat{\beta}_{new} = median\left\{b_x(k)_{ij}, b_y(k)_{ij}\right\}.$$

Hence, by assuming known slope as given by the $\hat{\beta}_{new}$, we may have a robust estimate of the parameters in linear functional relationship model and denoted by LFRM2.

## EXPECTATION-MAXIMIZATION ALGORITHM

EM algorithm is one example of imputation method using maximum likelihood, where it finds the maximum likelihood estimates through an iterative algorithm when there are missing values in the dataset (Little & Rubin 1987). In short, EM will 'fill in' the $Y_{mis}$, which are the missing data, based on an initial estimate of $\theta$ (whereby the estimate of $\theta$ is found by using only the data that are observed). Then, $\theta$ is re-estimated based on $Y_{obs}$, which are the data that we observe and the filled-in $Y_{mis}$ and this process is iterated until the estimates converge (Howell 2008). In order to elaborate, EM comprises of two steps namely the expectation or E-Step and the maximization or M-Step. In the E-step, we impute the missing values by replacing $Y_{mis}$ with the expected value of $E(Y_{mis}|Y_{obs}, \theta)$, by assuming $\theta = \theta^{(t)}$. Next, in the M-step, the expected value that we obtained from E-step will be maximised. These two steps will be done iteratively until it converges to a local maximum of the likelihood function (Schafer 1997). A detailed explanation on the convergence properties of EM algorithm can be found in some literature, as an example by Wu (1983).

EM algorithm has become popular because of its simplicity, the generality of its theory and because of its wide application (Dempster et al. 1977). Several examples of the applicability of EM include handling missing data in air pollutants studies (Schafer 1997), in linear regression model (Junger & de Leon 2015) and also in survival model (Wang & Miao 2009).

## EXPECTATION-MAXIMIZATION WITH BOOTSTRAPPING ALGORITHM

The emerging expectation-maximization with bootstrapping (EMB) algorithm is similar to the regular expectation-maximization (EM) algorithm. However, it involves multiple nonparametric bootstrap samples of the original incomplete data. The EMB algorithm performs multiple imputations that 'fills in' the missing values in the incomplete data set. Multiple imputations are less biased and its efficiency is higher than the listwise deletion (Honaker et al. 2013; Rancoita et al. 2015). Applying multiple imputations can be quite challenging as the nature of its algorithm can be quite complicated, but with the available of high performance computing, it can help perform the multiple imputations in a much advanced way (Honaker et al. 2013).

## PERFORMANCE INDICATORS

In order to measure the performance of our imputation using EM and EMB algorithm, we use several measurements namely the mean absolute error (MAE), root-mean-square error (RMSE) and estimated bias (EB). MAE is the average of the difference between the predicted and actual data points (Junnninen et al. 2004) and is given by

$$MAE = \frac{1}{N}\sum\left|P_i - O_i\right|, \tag{6}$$

where $N$ is the number of imputations, $P_i$ are the imputed values; and $O_i$ are the observed data values. Values of MAE can be from 0 to infinity in which a value of zero is an indicative of a perfect fit.

RMSE measures the differences between the predicted and actual data points and is given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[P_i - O_i\right]^2}. \tag{7}$$

with $N$ is the number of imputations $P_i$ and $O_i$ are the imputed and observed data points, respectively (Lindley 1947). A small value of RMSE suggests a good fit and large value otherwise.

Mean of estimated biased (EB) of a parameter on the other hand is defined by the mean of the absolute difference of the estimated value of the parameters obtained from the observed data and the estimated value of the parameters obtained from the data after imputing the missing values. A small EB is indicative of a reliable performance estimator (Lindley 1947).

## SIMULATION STUDY

A simulation study is conducted to investigate the performance of these two imputation methods namely the EM and EMB. For the first simulation study, we use the LFRM1 as in (1), where without any loss of generality, the parameters are set to $\alpha = 1$, $\beta = 1$, $\sigma_\delta^2 = 0.1$, $\lambda = 1$ with sample sizes, $n = 50$ and 100, respectively. For the simulation study, we assume the missing data are missing at random (MAR) and are inserted randomly at 5, 10, 20 and 30% levels, respectively (Howell 2008). We conducted this simulation for 5000 trials and the MAE, RMSE and EB of these two imputation methods, namely EM and EMB were analysed.

From Tables 1 and 2, we observe that both methods perform well, with small MAE and small RMSE at each $n$. EMB has significantly smaller MAE and RMSE values compared to EM for $n = 50$ and 100, respectively. For each level of percentage missing namely at 5, 10, 20 and 30%, respectively, the EMB consistently gives smaller MAE and RMSE values as compared to the EM. Looking at the RMSE for $n = 50$, at 5% missing values, the EMB values are different at only two decimal points from EM. It is worthwhile to note that the percentage change from EM

TABLE 1. MAE and RMSE for LFRM1 using two imputation methods for $n$ =50

| Percentage of missing (%) | Performance indicator / Methods | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 3.7530 | 25.13 | 5.4943 | 8.99 |
| | EMB | 2.8100 | | 5.0003 | |
| 10% | EM | 6.2616 | 19.82 | 5.1894 | 6.62 |
| | EMB | 5.0210 | | 4.8457 | |
| 20% | EM | 5.3612 | 17.85 | 4.9344 | 15.23 |
| | EMB | 4.4042 | | 4.1827 | |
| 30% | EM | 5.2312 | 13.21 | 5.4744 | 11.31 |
| | EMB | 4.5404 | | 4.8550 | |

TABLE 2. MAE and RMSE for LFRM1 using two imputation methods for $n$ =100

| Percentage of missing (%) | Performance indicator / Methods | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 4.6860 | 30.04 | 5.6699 | 32.43 |
| | EMB | 3.2781 | | 3.8314 | |
| 10% | EM | 5.2109 | 31.64 | 4.7434 | 16.36 |
| | EMB | 3.5623 | | 3.9672 | |
| 20% | EM | 5.6734 | 49.06 | 6.1135 | 26.60 |
| | EMB | 2.8900 | | 4.4872 | |
| 30% | EM | 4.4952 | 25.22 | 5.5477 | 17.99 |
| | EMB | 3.3617 | | 4.5497 | |

to EMB shows a significant difference of about 8.99% of improvement in the RMSE values. Another example, for the 20% missing value at $n$ = 50, the difference is significant with 15.23% improvement from EM to EMB. This proves that even though the difference of RSME is at two decimal places, it shows a huge improvement of EM to EMB. We note that as the sample size increase from $n$ = 50 to $n$ = 100 , the RMSE values of EMB decrease at all levels of percentage of missingness. This suggest that at a higher $n$, it leads to a smaller RMSE and bias.

From Tables 3 and 4, we observe that using EM and EMB, both methods give small value for the mean of the estimated bias for all the parameters $\alpha$, $\beta$, and $\sigma_\delta^2$. Imputation using EMB, however gives better precision with consistently even smaller bias values for all parameters as compared to the EM. Looking at the standard error of each parameter in the parenthesis, it shows that at each level of missingness, the EMB outperforms the EM by having smaller values of standard error. These observations clearly indicate the superiority of EMB in comparison to the EM.

The study is also replicated for the LFRM2, in which the slope parameter $\beta$ is estimated using a nonparametric method. From the results as presented in Tables 5 and 6, both methods of imputations are good, but EMB algorithm shows consistently smaller MAE and RMSE as compared to the EM algorithm for both $n$ = 50 and 100. We note that, as the percentage of missing data increases, EMB outperforms EM in terms of smaller MAE and RMSE values. Similar to LFRM1, the RMSE values of EM and EMB differs at only two decimal places but if we look at the percentage of improvement from EM to EMB, the change is significant. Again, the superiority of EMB applies for the LFRM2. Likewise, as $n$ increases from 50 to 100, both MAE and RMSE suggest a better precision for the LFRM2.

For the measure of estimated bias as given in Tables 7 and 8, both methods give small values for the mean of the estimated bias for all the parameters $\alpha$, $\beta$, and $\sigma_\delta^2$. Imputation using EMB, however gives better precision with smaller bias values for all parameters as compared to the EM. From the standard error of each parameter in the parenthesis, it shows that at each level of missing data, the EMB outperforms the EM by having smaller values of standard error. These observations clearly indicate the superiority of EMB in comparison to the EM.

In summary, the results of simulation studies suggested that imputing missing values using both EM and EMB are good, with EMB outperforms EM for models of the linear functional relationship type as they give smaller values of MAE, RMSE and smaller values of the standard error of the estimated bias in the parameters.

TABLE 3. Mean of estimated bias and (standard error) of the parameters for LFRM1 using two imputation methods for $n$ =50

| Percentage of missing (%) | Methods \ Parameters | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 3.621E-02 | 6.600E-03 | 5.101E-04 |
| | | (3.001E-02) | (5.321E-03) | (4.231E-04) |
| | EMB | 3.024E-02 | 6.071E-03 | 4.403E-04 |
| | | (1.503E-02) | (2.952E-03) | (3.395E-04) |
| 10% | EM | 2.986E-02 | 5.643E-03 | 8.028E-04 |
| | | (2.228E-02) | (4.225E-03) | (7.220E-04) |
| | EMB | 2.865E-02 | 5.510E-03 | 6.829E-04 |
| | | (2.208E-02) | (4.217E-03) | (5.599E-04) |
| 20% | EM | 3.086E-02 | 5.613E-03 | 1.147E-03 |
| | | (2.291E-02) | (4.193E-03) | (1.010E-03) |
| | EMB | 2.939E-02 | 5.496E-03 | 9.250E-04 |
| | | (2.176E-02) | (4.076E-03) | (7.372E-04) |
| 30% | EM | 2.144E-02 | 3.915E-03 | 7.425E-04 |
| | | (1.619E-02) | (2.942E-03) | (6.027E-04) |
| | EMB | 2.079E-02 | 3.895E-03 | 5.904E-04 |
| | | (1.552E-02) | (2.909E-03) | (4.477E-04) |

TABLE 4. Mean of estimated bias and (standard error) of the parameters for LFRM1 using two imputation methods for $n$ =100

| Percentage of missing (%) | Methods \ Parameters | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 2.012E-02 | 3.909E-03 | 3.837E-04 |
| | | (1.538E-02) | (2.983E-03) | (3.437E-04) |
| | EMB | 2.000E-02 | 3.907E-03 | 3.157E-04 |
| | | (1.485E-02) | (2.899E-03) | (2.456E-04) |
| 10% | EM | 2.064E-02 | 3.944E-03 | 5.340E-04 |
| | | (1.545E-02) | (2.960E-03) | (4.489E-04) |
| | EMB | 1.989E-02 | 3.827E-03 | 4.301E-04 |
| | | (1.514E-02) | (2.894E-03) | (3.375E-04) |
| 20% | EM | 2.132E-02 | 3.892E-03 | 7.952E-04 |
| | | (1.616E-02) | (2.974E-03) | (6.492E-04) |
| | EMB | 2.053E-02 | 3.847E-03 | 6.271E-04 |
| | | (1.567E-02) | (2.915E-03) | (4.905E-04) |
| 30% | EM | 2.244E-02 | 3.923E-03 | 9.732E-04 |
| | | (1.675E-02) | (2.956E-03) | (7.761E-04) |
| | EMB | 2.093E-02 | 3.835E-03 | 7.636E-04 |
| | | (1.612E-02) | (2.921E-03) | (5.913E-04) |

TABLE 5. MAE and RMSE for the LFRM2 by using two imputation methods for $n$ =50

| Percentage of missing (%) | Method \ Performance indicator | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 7.3911 | 27.67 | 9.0257 | 39.02 |
| | EMB | 5.3460 | | 5.5041 | |
| 10% | EM | 5.6922 | 48.60 | 7.3679 | 15.72 |
| | EMB | 2.9257 | | 6.2096 | |
| 20% | EM | 5.3877 | 4.52 | 5.7500 | 7.44 |
| | EMB | 5.1443 | | 5.3224 | |
| 30% | EM | 3.4405 | 6.86 | 4.8878 | 10.43 |
| | EMB | 3.2045 | | 4.3782 | |

TABLE 6. MAE and RMSE for the LFRM2 by using two imputation methods for $n$ =100

| Percentage of missing (%) | Method \ Performance indicator | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 6.0935 | 36.73 | 5.4978 | 2.60 |
| | EMB | 3.8556 | | 5.3549 | |
| 10% | EM | 5.1017 | 53.49 | 5.2083 | -0.67 |
| | EMB | 2.3729 | | 5.2433 | |
| 20% | EM | 3.7023 | 6.07 | 5.4531 | 25.83 |
| | EMB | 3.4775 | | 4.0445 | |
| 30% | EM | 3.9048 | 18.18 | 4.3791 | 4.73 |
| | EMB | 3.1950 | | 4.1721 | |

TABLE 7. Mean of estimated bias and (standard error) of the parameters for LFRM2 using two imputation methods for $n$ =50

| Percentage of missing (%) | Methods \ Parameters | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 3.069E-02 | 5.944E-03 | 4.899E-04 |
| | | (2.329E-02) | (4.511E-03) | (4.812E-04) |
| | EMB | 3.032E-02 | 5.915E-03 | 4.320E-04 |
| | | (2.280E-02) | (4.451E-03) | (3.857E-04) |
| 10% | EM | 3.125E-02 | 5.930E-03 | 7.795E-04 |
| | | (2.407E-02) | (4.588E-03) | (7.035E-04) |
| | EMB | 3.027E-02 | 5.828E-03 | 6.445E-04 |
| | | (2.294E-02) | (4.432E-03) | (5.249E-04) |
| 20% | EM | 3.258E-02 | 6.007E-03 | 1.153E-03 |
| | | (2.436E-02) | (4.523E-03) | (9.966E-04) |
| | EMB | 3.169E-02 | 5.976E-03 | 9.235E-04 |
| | | (2.404E-02) | (4.489E-03) | (7.311E-04) |
| 30% | EM | 3.390E-02 | 5.968E-03 | 1.449E-03 |
| | | (2.543E-02) | (4.511E-03) | (1.225E-03) |
| | EMB | 3.292E-02 | 6.081E-03 | 1.190E-03 |
| | | (2.451E-02) | (4.498E-03) | (9.659E-04) |

TABLE 8. Mean of estimated bias and (standard error) of the parameters for LFRM2 using two imputation methods for $n$ =100

| Percentage of missing (%) | Parameters / Methods | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 5.895E-02 | 1.165E-02 | 9.344E-04 |
| | | (4.489E-02) | (8.898E-03) | (1.128E-03) |
| | EMB | 5.889E-02 | 1.163E-02 | 8.523E-04 |
| | | (4.353E-02) | (8.614E-03) | (1.055E-03) |
| 10% | EM | 2.283E-02 | 4.389E-03 | 5.323E-04 |
| | | (1.721E-02) | (3.285E-03) | (4.367E-04) |
| | EMB | 2.240E-02 | 4.366E-03 | 4.339E-04 |
| | | (1.690E-02) | (3.265E-03) | (3.334E-04) |
| 20% | EM | 2.365E-02 | 4.391E-03 | 8.042E-04 |
| | | (1.771E-02) | (3.303E-03) | (6.577E-04) |
| | EMB | 2.258E-02 | 4.299E-03 | 6.466E-04 |
| | | (1.712E-02) | (3.244E-03) | (4.948E-04) |
| 30% | EM | 2.432E-02 | 4.377E-03 | 1.064E-03 |
| | | (1.880E-02) | (3.344E-03) | (8.295E-04) |
| | EMB | 2.337E-02 | 4.358E-03 | 8.284E-04 |
| | | (1.738E-02) | (3.255E-03) | (6.307E-04) |

The EM algorithm has largely been used in solving maximum-likelihood parameter estimation problems (Bilmes 1998; Bock & Murray 1981; Dempster et al. 1977). It has also become popular in handling missing data because of its simplicity, in spite of its slow convergence rate (Couvreur 1996). Nevertheless, EM has wide application in addressing missing data in medical (Dziura et al. 2013) and environmental data (Razak et al. 2014; Zainuri et al. 2015).

In this paper, we improved the EM algorithm by integrating bootstrap in the EM procedure. Simulation studies indicate the superiority of EMB in both LFRM1 and LFRM2 models. The re-sampling method of EMB made the estimator improved by creating a multiply-imputed values for each missing data. As a result, the average value of the imputed dataset contributes towards making the estimates more accurate with smaller standard errors.

## APPLICATION TO REAL DATA

In order to illustrate with a practical example, we consider a data set which consists of 96 observations which are free from any outliers (Goran et al. 1996). The study was to examine the accuracy of some widely used body-composition techniques for children, using the dual-energy X-ray absorptiometry (DXA). The sample comprises of children ages from four to ten years. They assessed children's body fat by using two variables, namely the skinfold thickness (ST) and bioelectrical resistance (BR). We assume that the measurement error can take place in either variable of this experiment and the relationship between these two variables can be expressed in a LFRM as given in (1).

In the interest of measuring the performance of EM algorithm and EMB algorithm, we randomly make the dependent variable missing at 5, 10, 20 and 30%, respectively. Both LFRM1 and LFRM2 models are applied in this experiment.

Table 9 shows the values of MAE and RMSE for LFRM1, using both imputation methods of EM and EMB. We observe that there is a consistency in the results whereby the EMB algorithm has smaller MAE and RMSE values as compared to using the EM algorithm. Similar conclusion can be made for the results in Table 10, in which the values of bias using EMB are smaller in comparison to the EM.

As mentioned earlier, we also consider the LFRM model where the slope parameter is estimated using a nonparametric method, namely LFRM2. Table 11 indicates the MAE and RMSE values of the slope for LFRM2 while Table 12 illustrates the EB of the parameters of the slope for LFRM2. From both Tables 11 and 12, we note that, EMB algorithm proves to be better with smaller values of EB, MAE and RMSE.

It can be inferred that from this practical application, both methods of imputations namely EM and EMB demonstrate good results based on the EB, MAE and RMSE values. It is shown that imputing missing values using EMB gives a better approach than the EM in handling missing values for data that can be modelled by the linear functional relationship formulation. In this practical example, it

TABLE 9. MAE and RMSE for LFRM1 for real data using two imputation methods

| Percentage of missing (%) | Performance indicator / Method | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 5.2256 | 10.97 | 4.6518 | 28.20 |
| | EMB | 4.6521 | | 3.3400 | |
| 10% | EM | 5.5593 | 27.81 | 5.3013 | 6.14 |
| | EMB | 4.0135 | | 4.9756 | |
| 20% | EM | 4.9928 | 26.13 | 4.9781 | 5.25 |
| | EMB | 3.6883 | | 4.7166 | |
| 30% | EM | 5.1355 | 20.27 | 5.5337 | 5.95 |
| | EMB | 4.0946 | | 5.2044 | |

TABLE 10. Estimated bias of parameters using LFRM1 for real data

| Percentage of missing (%) | Parameters / Methods | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 0.4926 | 0.0997 | 0.0782 |
| | EMB | 0.3975 | 0.0997 | 0.0573 |
| 10% | EM | 0.6098 | 0.0997 | 0.1821 |
| | EMB | 0.4895 | 0.0997 | 0.1036 |
| 20% | EM | 0.5243 | 0.0997 | 0.1625 |
| | EMB | 0.4366 | 0.0997 | 0.0772 |
| 30% | EM | 0.6315 | 0.0997 | 0.1017 |
| | EMB | 0.6236 | 0.0997 | 0.0524 |

TABLE 11. MAE and RMSE for LFRM2 for real data using two imputation methods

| Percentage of missing (%) | Performance indicator / Method | MAE | Percentage change of MAE (%) | RMSE | Percentage change of RMSE (%) |
|---|---|---|---|---|---|
| 5% | EM | 3.6671 | 39.49 | 5.5610 | 40.53 |
| | EMB | 2.2190 | | 3.3070 | |
| 10% | EM | 2.7472 | 12.79 | 3.7241 | 4.69 |
| | EMB | 2.3959 | | 3.5494 | |
| 20% | EM | 2.6680 | 36.36 | 5.2740 | 15.77 |
| | EMB | 1.6978 | | 4.4424 | |
| 30% | EM | 3.2698 | 18.21 | 3.8403 | 25.61 |
| | EMB | 2.6744 | | 2.8568 | |

TABLE 12. Estimated bias of parameters for LFRM2 for real data

| Percentage of missing (%) | Parameters / Methods | $\alpha$ | $\beta$ | $\sigma_\delta^2$ |
|---|---|---|---|---|
| 5% | EM | 0.0236 | 0.0080 | 0.0963 |
| | EMB | 0.0067 | 0.0080 | 0.0128 |
| 10% | EM | 0.0508 | 0.0080 | 0.1865 |
| | EMB | 0.0406 | 0.0080 | 0.0196 |
| 20% | EM | 0.1538 | 0.0080 | 0.1775 |
| | EMB | 0.0373 | 0.0080 | 0.1194 |
| 30% | EM | 0.1950 | 0.0080 | 0.2707 |
| | EMB | 0.1543 | 0.0080 | 0.1381 |

is proven that EMB has improved the precision in the algorithm and this is reflected by its superior performance.

## CONCLUSION

In this paper, we have investigated two modern approaches of handling missing values namely EM and EMB for datasets that can be modelled by the linear functional relationship model. The results from simulation study suggested both methods of imputation can be applied for two forms of the linear functional relationship model. Even in the presence of high percentage of missing values (to as high as 30%), both methods adequately handle the problem. These can be seen with small bias measure of parameter and small MAE and RMSE. When comparing the two imputation methods, EMB is superior to EM. Again, this is evidenced by the MAE and RMSE values. EMB has several advantages where it can be easily applied to LFRM, the computing time is much faster as compared to the EM and the bootstrapping method gives better precision to the parameter estimates.

We have illustrated using a real data set that compares the relationship between two variables measurements. The results obtained showed that if in the case when the real data set has missing values for a percentage to as high as 30%, both methods of imputation are suitable for handling missing values with EMB being superior than EM.

Nowadays, modern techniques are increasingly more popular with recent computing capabilities that perform at very high speed and have very good precision in the algorithm. Both EM and EMB provide practical approaches in handling missing data sets that is of the linear functional relationship model with better performance for EMB.

## ACKNOWLEDGEMENTS

The authors would like to express our appreciation and thanks to the editors and reviewers for their valuable comments and feedback on this paper. This work was supported by the Postgraduate Research Grant, University of Malaya (PG058-2015A).

## REFERENCES

Acock, A.C. 2005. Working with missing values. *Journal of Marriage and Family* 67: 1012-1028.

Al-Nasser, A.D. 2005. A new nonparametric method for estimating the slope of simple linear measure error model in the presence of outliers. *Pak. J. Statist.* 21(3): 265-274.

Baraldi, A.N. & Enders, C.K. 2010. An introduction to modern missing data analyses. *Journal of School Psychology* 48: 5-37.

Barzi, F. & Woodward, M. 2004. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology* 160(1): 34-45.

Bilmes, J.A. 1998. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*. pp. 2-7.

Bock, R.D. & Murray, A. 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* 46(4): 443-459.

Couvreur, C. 1997. The EM algorithm: A guided tour computer intensive methods in control and signal processing. New York: Springer. pp. 209-222.

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B (Methodological)* 39(1): 1-38.

Dziura, J.D., Post, L.A., Zhao, Q., Fu, Z. & Peduzzi, P. 2013. Strategies for dealing with missing data in clinical trials: From design to analysis. *The Yale Journal of Biology and Medicine* 86(3): 343-358.

George, N.I., Bowyer, J.F., Crabtree, N.M. & Chang, C.W. 2015. An iterative leave-one-out approach to outlier detection in RNA-Seq data. *PLoS ONE* 10(6): e0125224. doi:10.1371/journal. pone.0125224G.

Ghapor, A.A., Zubairi, Y.Z., Mamun, A.S.M.A. & Imon, A.H.M.R. 2015. A robust nonparametric slope estimation in linear functional relationship model. *Pak. J. Statist.* 31(3): 339-350.

Gold, M.S. & Bentler, P.M. 2000. Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modelling*: *A Multidisciplinary Journal* 7(3): 319-355.

Goran, M.I., Driscoll, P., Johnson, R., Nagy, T.R. & Hunter, G.R. 1996. Cross-calibration of body-composition techniques against dual-energy X-Ray absorptiometry in young children. *American Journal of Clinical Nutrition* 63: 299-305.

Guan, N.C. & Yusoff, N.S.B. 2011. Missing values in data analaysis: Ignore or Impute? *Education in Medicine Journal* 3(1): 6-11.

Honaker, J., King, G. & Blackwell, M. 2013. *Amelia II: A Program for missing data*. http://gking.harvard.edu/amelia.

Howell, D.C. 2008. The analysis of missing data. In *Handbook of Social Science Methodology*, edited by Outhwaite, W. & Turner, S. London: Sage.

Junger, W.L. & de Leon, A.P. 2015. Imputation of missing data in time series for air pollutants. *Atmospheric Environment* 102: 96-104.

Junninen, H., Niska, H., Tuppurrainen, K., Ruuskanen, J. & Kolehmainen, M. 2004. Methods for imputation of missing values in air quality data sets. *Atoms Environ*. 38: 2895-2907.

Kendall, M.G. & Stuart, A. 1973. *The Advance Theory of Statistics*. Vol. 2, London: Griffin.

Lindley, D.V. 1947. Regression lines and the linear functional relationship. *J. R. Statist. Soc., Suppl*., 9: 218-244.

Little, R.J.A. & Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. New York: Wiley.

Morita, T. & Kimura, M. 2014. A fundamental study on missing value treatment for software quality prediction. *Advanced Science and Technology Letters* 67: 70-73.

Rancoita, P.M.V., Zaffalon, M., Zucca, E., Bertoni, F. & Campos, C.P. 2015. Bayesian network data imputation with application to survival tree analysis. *Computational Statistics and Data Analysis* 93: 373-387.

Razak, N.A., Zubairi, Y.Z. & Yunus, R.M. 2014. Imputing missing values in modelling the $PM_{10}$ concentrations. *Sains Malaysiana* 43(10): 1599-1607.

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

Schafer, J.L. & Graham, J.W. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7: 147-177.
Sprent, P. 1969. *Models in Regression and Related Topics*. London: Methuen.
Takahashi, M. & Ito, T. 2013. Multiple imputation of missing values in economic surveys: Comparison of competing algorithms. *Proceedings 59th ISI World Statistics Congress*, Hong Kong, August 25-30th.
Wang, J. & Miao, Y. 2009. Note on the EM Algorithm in linear regression model. *International Mathematical Forum* 38: 1883-1889.
Wu, C.F.J. 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics* 11(1): 95-103.
Zainuri, N., Jemain, A. & Muda, N. 2015. A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana* 44(3): 449-456.

Adilah Abdul Ghapor
Institute of Graduate Studies
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Yong Zulina Zubairi*
Centre for Foundation Studies in Science
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

A.H.M. Rahmatullah Imon
Department of Mathematical Sciences
Ball State University
47306 Indiana
United States of America

*Corresponding author; email: yzulina@um.edu.my

Received: 1 December 2015
Accepted: 9 June 2016