Comparison between EM Algorithm and Multiple Imputation on Predicting Children’s Weight at School Entry

S S Avtar¹, G P Khuneswari¹, A A Abdullah¹, J H McColl², C Wright³ and GMS Team⁴

¹ Faculty of Science and Technology, University Tun Hussein Onn Malaysia, 84600, Muar, Johor, Malaysia
² School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, United Kingdom
³ School of Medicine, University of Glasgow, Glasgow G12 8QQ, United Kingdom
⁴ Newcastle University, Newcastle upon Tyne, Tyne and Wear NE1 7RU, United Kingdom

Abstract. EM Algorithm and Multiple Imputation are widely used methods in dealing with missing data. Although Multiple Imputation always be the favourite choice of researcher due to its accuracy and simple application, but the issue arises whether EM algorithm perform better with several times of imputation. Both methods will be tested using different number of imputations with the help of Amelia and Mice package in R software. The imputed data sets are compared using model averaging with Corrected Akaike Information Criteria ($\text{AIC}_c$) as model selection Criterion. External validation and mean squared error of prediction (MSE(P)) are used to determine the best imputation method. Gateshead Millennium Study (GMS) data on children weight will illustrate the comparison between EM Algorithm and Multiple imputation. The results show that Multiple imputation performs slightly better compared to EM Algorithm.

1. Introduction

Model-building process become more challenging when there exists missing data. 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: (1) missing completely at random (MCAR) (2) missing at random (MAR) (3) 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.

Dempster et al. proposed EM algorithm as iterative solution while assuming multivariate normal data. EM algorithm involves two step, E-step and M-step, where E-step is expectation and M-step is maximizing. However, EM leads to biased parameter estimates and underestimates the standard errors. For this reason, this research will focus on adjusting of EM algorithm to overcome the problem [1, 2]. EM algorithm will be combining with multiple imputation concept which several plausible data sets will be obtained before averaged the final output. Multiple imputation is a general approach to overcome missing data problem. It is a simple but powerful method for handling missing data. Multiple Imputation as originally conceived proceeds in two stages: (1) A data disseminator creates a small number of completed datasets by filling in the missing values with samples from an imputation model. (2) Analysts
compute their estimates in each completed dataset and combine them using simple rules to get pooled estimates and standard errors that incorporate the additional variability due to the missing data [3].

The accuracy of imputed data will be tested using model averaging. Model Averaging is an alternative of model selection, which the model will weighted across the model to obtained final best model. Model averaging tends to shrink the estimates on the weaker terms, yielding better predictions. The “best” models will hold higher weights [4]. In order to determine the “best” model, model selection criteria are needed. AICc will be used as model selection criterion. AICc is known to be less biased when there is small sample size [5]. Therefore, this research will focus on model-building approach for model averaging with comparison between EM algorithm and Multiple Imputation. The effects of interaction effects also will be explored.

2. Missing Data

2.1 Multiple Imputation

Multiple imputation (MI) is a method to encounter missing data problem by substituting each missing data by D ≥ 2 imputed values in order to create multiple complete dataset. After carry out MI analysis on completed dataset, then will combine the results to reflect the variability between-imputation and within-imputation. MI involves three stage in the process [6].

Stage 1: Generating
The missing values are replaced with D times to generate D complete data sets.

Stage 2: Analyzing
Once the MI have been generated, each imputed dataset is analyzed separately as though it was a complete dataset. Parameters are estimated from each imputed dataset.

Stage 3: Combining estimates
The results from D analyzed are combined into a single inference.

Rubin’s rules are as follow. The $\hat{\theta}_d$ is an estimate of a univariate or multivariate quantity of interest obtained from the $d^{th}$ imputed datasets and $V_d$ is the estimated variance of $\hat{\theta}_d$. The combined estimate $\bar{\theta}$ is the average of the individual estimates [3].

$$\bar{\theta} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d$$  \hspace{1cm} (1)

The total variance of $\bar{\theta}$ is formed from within-imputation variance $V = \frac{1}{D} \sum_{d=1}^{D} V_d$.

Multiple imputation method will be conducted with help of Mice package. Mice is a package in R software which is designed to help with imputing missing values with plausible data values. A specific distribution is obtained to drawn these plausible values for each datapoint [7].

2.2 EM Algorithm

The Expectation Maximisation (EM) algorithm is a general algorithm for maximum likelihood (ML) estimation for incomplete data [3]. It involves two steps: (1) the expectation step (E-step), (2) the maximisation step (M-step). The distribution of the complete data X in any incomplete data set can be factorised as

$$f(Y, X; \theta) = f(Y, X_{obs}; \theta) f(X_{mis}|Y, X_{obs}; \theta)$$  \hspace{1cm} (2)

Considering each term in equation above as a function of $\theta$, it follows that

$$ln(Y, X; \theta) = ln(Y, X_{obs}; \theta) + ln(X_{mis}|Y, X_{obs}; \theta) + c$$  \hspace{1cm} (3)

$ln(Y, X; \theta)$ is the log-likelihood of the complete data, $ln(Y, X_{obs}; \theta)$ is the log-likelihood of the observed data and c is an arbitrary constant. E-step takes the average of complete data log-likelihood with respect
to the distribution \( f(X_{mis}|X_{obs}; \theta^{(r)}) \), where \( \theta^{(r)} \) is the current parameter estimate of \( \theta \). This log-
likelihood yields
\[
\int \ln(Y, X; \theta) f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis} \\
= \int \ln(Y, X_{obs}; \theta) f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis} \\
+ \int \ln(X_{mis}|Y, X_{obs}; \theta) f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis}
\] (4)

Equation (4) can be written in the form of a \( Q \)-function and \( H \)-function as follows
\[
Q(\theta|\theta^{(r)}) = \int \ln(Y, X_{obs}; \theta) f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis} + H(\theta|\theta^{(r)}) \\
= \ln(Y, X_{obs}; \theta) \int f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis} + H(\theta|\theta^{(r)}) \\
= \ln(Y, X_{obs}; \theta) + H(\theta|\theta^{(r)})
\] (5)

where the \( H \)-function is
\[
H(\theta|\theta^{(r)}) = \int \ln(X_{mis}|Y, X_{obs}; \theta) f(X_{mis}|Y, X_{obs}; \theta^{(r)}) dX_{mis}
\] (6)

The E-step is based on the evaluation of the \( Q \)-function in Equation (5). The M-step involves
maximizing \( Q(\theta|\theta^{(r)}) \) with respect to \( \theta \) to obtain \( \theta^{(r+1)} \). The iteration between the E-step and M-step
will continue until convergence [3, 7].

Amelia package in R software will be used for EM algorithm analysis. This package allows
multiple imputes of missing data using EM algorithm which the number of imputations can be change
according to the user preference [8].

3. Model-Building

3.1 Multiple Regression
Multiple linear regression is an extended version of simple linear regression model that involve two or
more explanatory variables in a prediction equation. A more complex model that contain more
explanatory variables typically is more useful in providing sufficiently precise of the response variable.
The general regression model with two predictor variables \((X_1, X_2)\) is [9]
\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i
\] (7)

The model with more than two predictor variables \((p - 1)\) predictor variables, \(X_1, X_2, \ldots, X_{p-1}\) is
\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i(p-1)} + \epsilon_i
\] (8)

where \( \beta_0, \beta_1, \ldots, \beta_{p-1} \) are parameters, \(X_{i1}, X_{i2}, \ldots, X_{i(p-1)} \) are known constants and \( \epsilon_i \) are independent
\( N(0, \sigma^2) \) for \( i = 1, 2, \ldots, n. \)

3.2 Model Averaging
Model selection will include additional uncertainty into the model building process. The properties of
parameter estimates obtained from the selected model method do not represent the stochastic nature of
the model selection process. Model averaging had been proposed as an alternative method to model
selection to overcome the under-estimation of standards errors in model selection. A model average
estimator weighs across all possible models rather than picking a single best model. Model averaging
will give less weight on the estimates of the weaker variables and will yield better predictions. The
'better' models will receive higher weights compare to others model. Suppose that there are \( M \) candidate
models. In one approach, the weight \( w_m \) for model is [4, 10]
\[ w_M = \frac{\exp \left( \frac{l_M}{2} \right)}{\sum_{m=1}^{M} \exp \left( \frac{l_M}{2} \right)} \]  

where \( l_M = 2 \log L(M) - c_{n,p} \) is model selection criterion for \( M \), while \( c_{n,p} \) is the penalty term for model \( M \). The estimate of a parameter \( \beta_p \) is

\[ \hat{\beta}_p = \sum_{m=1}^{M} w_M \hat{\beta}_{(p,M)} \]  

where \( \hat{\beta}_{(p,M)} \) is the estimate of \( \hat{\beta}_p \) under model \( M \) for \( m = 1,2, \ldots, M \). The modified weights will be used based on model selection criteria AIC, AICc and BIC. A modification was carried out for calculating the weights in order to avoid numerical error. The weights \( w_M \) were calculated as

\[ w_M = \frac{\exp \left( \frac{l_M - \bar{i}}{2} \right)}{\sum_{m=1}^{M} \exp \left( \frac{l_M - \bar{i}}{2} \right)} \]  

where \( \bar{i} = \frac{1}{M} \sum_{m=1}^{M} l_M \) with \( l_M \) is log-likelihood function of model \( M \) for \( m = 1, 2, \ldots, M \). Model selection criterion used for \( M \) is Akaike’s information criterion (AIC).

AIC is a popular and had been widely used as model selection criteria. AIC is calculated using the amount of fitted parameters, maximum likelihood estimate, and including intercept of model, \( (p) \). AIC for model \( M \) is [11, 12]

\[ AIC = -2 \ln L(M) + 2p \]  

For small sample sizes (approximated as being when \( n/p \) is less than 40 and \( p \) is the number of fitted parameters in the most complex model), a corrected version of Akaike’s information criterion, \( (AIC_c) \) is recommended. The general form of \( AIC_c \) is [13]

\[ AIC_c = AIC + \frac{2p (p + 1)}{n - p - 1} \]  

4. Cross-validation

The data will be divided into two parts before analysis is conduct. 10% of the data will be used as training set which contains of complete observation. While the other 90% will be used as test set to conduct missing data imputation and regression. The evaluating of the model performance is by calculating its mean squared error of prediction (MSE(P)). In general, MSE(P) can be describe as [14, 15]

\[ MSE (P) = \frac{1}{t} \sum_{i=1}^{t} (\hat{Y}_t - Y_t)^2 \]  

where \( \hat{Y}_t \) is estimated \( Y \) of test values and \( Y_t \) is the actual test values used for prediction. MSE(P) is usually used to assess the performance of regressions.

5. Result and Discussion

Study shows there are relationship between rapid weight gain and later overweight, leading to the suggestion that prevention and treatment of childhood obesity should begin as early as first year of life. The Gateshead Millennium Study (GMS) is a study of feeding and growth in infancy. The original objectives of the study are to explore the relationship between child development and feeding in the year of life, but it was later extended to follow up the children throughout childhood. The study was conducted at Gateshead area of northeast England. There are 1029 babies that born between 1st June 1999 till 31 May 2000 that involve in this study. The sample size represents 83% of all births in the
region on that year. The children were studied prospectively using parent report questionnaire shortly after birth at 6 weeks and at 4, 8, 12 months. The cohort has since been re-traced at school entry, parent report questionnaires completed at 5-8 years, and a range of anthropometric and body composition measures collected at age 7-8 years [10, 16].

Table 1 show the variable that will be using to carry of this analysis. Two variables (gestational age and Weight at 8 Month) have been remove from this analysis due to high value of VIF. Table 2 shows the descriptive statistics for male and female children. The mean weight of the children at school entry is 19.80kg while the median is 19.20kg. The first quartile is 17.80kg while third quartile is 21.20kg. It can be concluded that the smallest weight is 13.00kg while the heaviest child weighted 56.00kg. 25% of the overall weight falls below 17.80kg and 25% of the overall weight is falls above 21.20kg. There are 305 cases with missing value in weight of children at school entry with 29.64%, the highest percent of missing variable. It follows by weight at 6 weeks (17.20%), weight at 12 months (16.62%) and lastly weight at 4 months (14.97%). However, birthweight have a complete information. Since there are outlier, 3 observations will be removed. The new sample size will be 1026.

Table 1: Description of variable for GMS

| Variables | Descriptions         | Unit               |
|-----------|----------------------|--------------------|
| Y         | Weight at school entry | Kilograms (kg)     |
| X₁        | Birthweight          | Kilograms (kg)     |
| X₂        | Weight at 6 weeks    | Kilograms (kg)     |
| X₃        | Weight at 4 months   | Kilograms (kg)     |
| X₄        | Weight at 12 months  | Kilograms (kg)     |
| X₅        | Gender               | 1=Male, 2=Female   |

Table 2: Descriptive Statistics of GMS

|               | Y     | X₁     | X₂     | X₃     | X₄     |
|---------------|-------|--------|--------|--------|--------|
| Minimum       | 13.00 | 0.840  | 2.760  | 4.090  | 6.240  |
| 1st Quartile  | 17.80 | 3.000  | 4.287  | 6.000  | 9.300  |
| Median        | 19.20 | 3.350  | 4.720  | 6.540  | 10.160 |
| Mean          | 19.80 | 3.325  | 4.727  | 6.605  | 10.170 |
| 3rd Quartile  | 21.20 | 3.700  | 5.120  | 7.125  | 10.940 |
| Maximum       | 56.00 | 5.370  | 6.800  | 9.750  | 15.700 |
| Number of Missing Data | 305  | 177    | 177    | 154    | 171    |
| Percentage of Missing Data | 29.640% | 17.200% | 14.970% | 16.620% |

Two different R package were used to carry out the imputation of missing value which are MICE and AMELIA respectively. MICE package was used to conduct imputation of multiple imputation and Amelia was used on EM Algorithm analysis. The missing value was generated five times, $D=5$ and 10 times, $D=10$ for MICE and AMELIA. The analysis for model averaging will be conducted using AIC. Table 3 and Table 4 show the result for Multiple Imputation and EM Algorithm.

Table 5 shows the value of MSE(P) for all 4 models. Based on Table 5, five imputation is better compared to 10 imputation due to the lower MSE(P). However, Multiple Imputation still works slightly better compare to EM Algorithm for five and 10 imputation. In conclusion, model of Multiple Imputation with $D=5$ is the best.
Table 3: Coefficient value for model based on EM Imputation

|      | D = 5      | D = 10     |
|------|------------|------------|
| Constant | 11.6676    | 9.0560     |
| X1    | -1.2461    | -0.8138    |
| X2    | 0.7073     | 1.0322     |
| X3    | 0.2930     | -0.0264    |
| X4    | 0.2248     | 0.5183     |
| X5    | 0.6422     | 0.6697     |
| X12   | -0.1805    | -0.0912    |
| X13   | 0.0588     | 0.1267     |
| X14   | 0.2586     | 0.2045     |
| X23   | -0.0823    | -0.0255    |
| X24   | -0.0040    | -0.0762    |
| X34   | -0.0152    | 0.0230     |
| X123  | 0.0547     | 0.0059     |
| X124  | 0.0206     | 0.0293     |
| X234  | -0.0523    | -0.0400    |
| X1234 | 0.0537     | 0.0496     |

Table 4: Coefficient value for model based on Multiple Imputation

|      | D = 5      | D = 10     |
|------|------------|------------|
| Constant | 3.9527     | 6.2457     |
| X1    | -1.0299    | -0.4802    |
| X2    | 1.6461     | 1.6753     |
| X3    | 1.3324     | 0.7879     |
| X4    | 0.8491     | 0.5756     |
| X5    | 0.6761     | 0.5977     |
| X12   | -0.1195    | -0.4033    |
| X13   | -0.0181    | -0.0708    |
| X14   | 0.2608     | 0.2308     |
| X23   | -0.1312    | -0.1507    |
| X24   | -0.1289    | -0.0875    |
| X34   | -0.1732    | -0.0697    |
| X123  | 0.0274     | 0.0638     |
| X124  | 0.0238     | 0.0364     |
| X234  | 0.0044     | 0.0045     |
| X1234 | 0.0664     | 0.0586     |

According to the best model, starting weight at school entry is 3.9527 kg. However, it will decrease 1.0299 kg for every 1 kg of birthweight due to negative relationship between birthweight and weight at school entry. Variables weight at 6 weeks, 4 months, 12 months will increase the weight at school entry by 1.6461, 1.3324, and 0.8491 respectively. Female children tend to have higher weight compare to male children by 0.6761 kg. Interaction of birthweight weight with weight at 8 months; weight at 6 weeks, 4 months with 12 months; and birthweight, weight at 6 weeks, 4 months with 12 months give positive effects of increasing weight at school entry with 0.2608, 0.0044 and 0.0644 respectively. The higher number of interactions contributes less effects to the model.

Table 5: MSE(P) values

| Model               | Selection Criteria | MSE(P) |
|---------------------|--------------------|--------|
| EM Algorithm        | D = 5              | 13.1159|
|                     | D = 10             | 13.1389|
| Multiple Imputation | D = 5              | 13.1071|
|                     | D = 10             | 13.1220|

6. Conclusion

Multiple Imputation excel the EM algorithm with 0.0088 for D=5 and 0.0169 for D=10. This may due to major drawbacks of EM algorithm. Little and Rubin (2002) stated that there are two major drawbacks of EM algorithm. First, it will converge very slowly in cases with large fractions of missing data. Second, the M-step will be difficult in some cases and then the theoretical simplicity of EM will not convert to simplicity in practice. 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 [17]. Besides that, 5 imputation works better than 10 imputation. Schafer and Olsen suggest that in many applications, just 3–5 imputations are sufficient to obtain robust results [18, 19].

Further research can be conduct involving cross-validation. This research only focuses on external cross-validation, where the data have been divided into training and test data before performing imputing missing data. Internal cross-validation strategies such as k-fold cross-validation, bootstrapping or subsampling are suggested to carry out after imputation of data [20].
7. References

[1] Dempster A P, Laird N M and Rubin D B 1997 *Journal of the Royal Statistics Society B* vol 39 pp 1-38.

[2] Schafer J L 1997 *Analysis of incomplete multivariate data* (London: Chapman and Hall)

[3] Little R J A and Rubin D B 2002 *Statistical analysis with missing data* (New Jersey: John Wiley and Sons)

[4] Buckland T S, Burnham K P and Augustin N H 1997 *Biometrics* vol 53(2) pp 603-618.

[5] Khuneswari G, Avtar S S and Abdullah A A 2018 *International Journal of Engineering & Technology* vol 7(30) pp 205-208

[6] White I R, Royston P and Wood A M 2011 *Statistics in Medicine* vol 30(4) pp 377-399

[7] Stef V B and Karin G O 2011 *Journal of Statistical Software* vol 45(3) pp 1-67

[8] James H, Gary K and Matthew B 2011 *Journal of Statistics Software* vol 45(7) pp 1-47.

[9] Kutner M H, Nachtsheim C J, and Neter J 2008 *Applied Linear Regression Models* (New York: McGraw-Hill/Irwin)

[10] Khuneswari G 2015 *Model selection and model averaging in the presence of missing data* PhD (University of Glasgow)

[11] Claeskens G, and Hjort N L 2008 *Model Selection and Model Averaging* (Cambridge: Cambridge University Press)

[12] Burnham K P and Anderson D R 2004 *Sociological Methods & Research*, vol 33(2) pp 261-304

[13] Symonds M R E and Moussalli A 2011 *Behavioral Ecology and Sociobiology* vol 65(1) pp 13–21

[14] Mevik B H and Cederkvist H R 2004 *Journal of Chemometrics* vol 18(9) pp 422–429

[15] Wallach D and Goffinet B 1989 *Ecological Modelling* vol 44 pp 299–306

[16] Wright M W, Cox K M, Sherriff A, Franco-Villoria M, Pearce M S, Adamson A J, and core team GMS 2011 T *Public Health Nutrition* vol 15(4) pp 656-662

[17] Pigott T D 2001 *Educational Research and Evaluation* vol 7(4) pp 353-383

[18] Schafer J L and Olsen M K 1998 *Multivariate Behavioral Research* vol 33 pp 545–571

[19] Osborne J W and Waters E 2002 *Practical Assessment, Research & Evaluation* vol 8(2) pp 1-14

[20] Wahl S, Boulesteix A L, Zierer A, Thorand B & Van de Wiel M A 2016 *BMC Med Res Methodology* vol 16(1) pp 144