THE ENHANCEMENT OF LINEAR REGRESSION ALGORITHM IN HANDLING MISSING DATA FOR MEDICAL DATA SET

Anirah Ahmad and Hasimah Hj. Mohamed
School of Computer Sciences, Universiti Universiti Sains Malaysia, Pulau Pinang, Malaysia
E-Mail: anirah@psp.edu.my

ABSTRACT

Missing data is a common problem faced by researchers in many studies. The occurrence of missing data can produce biased results at the end of the study and affect the accuracy of the findings. There are various techniques to overcome this problem and multiple imputation technique is the best solution. Multiple imputation can provide a valid variance estimation and easy to implement. This technique can produce unbiased result and known as a very flexible, sophisticated approach and powerful technique for handling missing data problems. One of the advantages of Multiple Imputation is it can use any statistical model to impute missing data. Hence the selection of the imputation model must be done properly to ensure the quality of imputation values. However the selection of imputed model is actually the critical step in Multiple Imputation. This research study a linear regression model (LR) as the selected imputation model, and proposed the new algorithm named Linear Regression with Half Values of Random Error (LReHalf). The proposed algorithm is used to improve the performance of linear regression in the application of Multiple Imputation. Furthermore this research makes comparison between LR and LReHalf. The performance of LReHalf is measured by the accuracy of imputed data produced during the experiments. Future research is highly suggested to increase the performance of LReHalf model. LReHalf was recommended to enhance the quality of MI in handling missing data problems, and hopefully this model will benefits all researchers from time to time.

Keywords: missing data, multiple imputation, linear regression with half values of random error.

INTRODUCTION

Information plays a very important role in our life (Magnani, M., 2004). It has been estimated that the amount of information in the world doubles every 20 months (Frawley, *et al.* 1992). Therefore information is considered as an important asset or a priority asset for organization, and it also can guide organization in decision making and business prediction for the future. It is noted that the number of such databases keeps growing rapidly because of the availability of powerful and affordable database systems (Chen, M.S. *et al*., 1996). Data mining is one of the vital tool in managing data in organization. Data mining is responsible in analysis of large data sets and the process involves computer-based methodology to discover knowledge from data. Kantardzic, M. (2002) noted that data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in exploratory analysis and need cooperative effort between humans and computers. Al Shalabi, L. *et al*. (2006) summarized that data mining is the process of analyzing data and generating new knowledge, hopefully understandable by humans, which was previously hidden and detected.

Data mining is a component of Knowledge Discovery Databases (KDD) which comprises of stages such as Data Selection, Data Preprocessing, Data Transformation and Data Mining. The data preprocessing stage contained a problem known as missing data. Missing data can produce low quality of data for KDD, and it will decrease the efficiency and relevancy of databases in KDD. As to handle this problem, the proposed method called multiple imputation (MI) technique is applied.

Over recent years, MI has gained popularity as a powerful statistical tool for handling missing data (Mackinnon, 2010). Sterne, J.A. *et al*. (2009) claimed MI has become increasingly popular for handling missing data in epidemiology. Spratt, M. *et al.* (2010) supported this statement and made conclusion that result from analyses based on MI are increasingly being reported in the epidemiologic and medical literature. According to Cummings P. (2013), MI will reduce bias and increase precision compared with complete-case analysis. Elizabeth, *et al*., (2009) also agreed that MI is a powerful and flexible technique for dealing with missing data.

According to Gebreab, S. Y. *et al*. (2015), MI can reduce potential bias estimates and avoid loss of statistical power due to missingness. MI also allow the researcher to account for uncertainty due to the missing data when making inferences (Murray J.S. and J.P. Reiter., 2014). Furthermore, MI can represent a good balance between quality of results, and it is now a well established technique to analyze data sets where some units have incomplete observations. Another advantages are with MI, it can provide the imputation model correctly, and the resulting estimates are consistent (Carpenter, J. R. *et al.*, (2006), and MI can be used with any kind of data and model with conventional software (Soley-Bori, M., 2013).

This paper discussed about the usage of MI with the enhancement of Linear Regression model, and the paper is divided into five sections such as Introduction, Literature Review, Proposed Framework, Result and Analysis, and Conclusion.

LITERATURE REVIEW

Missing data

Missing data problems are commonly faced by a researcher in a wide range of field. Missing data can be described as no value exists in one or more of data observation.

The intend purpose of any analysis is to make valid inferences regarding a population of interest, missing data threatens this goal if it is missing in a manner which makes the sample different than the population from which it was drawn, that is, if the missing data creates a biased sample (Wayman, J.C., 2003). Newman D.A. (2014) concluded that the purpose of data analysis is to give inbiased estimates of population parameters, as well as to provide accurate (error-free) hypothesis testing. Therefore biased data is bad because the information about the population of interest is not correct and not accurate to any references in the future. Barnard, J. and Meng, X.L. (1999) concluded that three types of problems are usually associated with missing values such as loss of efficiency, complication in handling and analyzing the data, and bias resulting from differences between missing and complete data

There are various factors of missing data problems. Values may be missed in the course of data collection by a device faults or human errors (Jun Ma, *et al*., 2009). According to Allison, P.D. (2001), missing data which are caused by human errors are such as respondents are refusing to answer the questions and for example normally respondents ignore questions related about their incomes, respondents forget to answer some questions, respondents do not know the suitable answer.

Missing data can be categorized in several patterns and several mechanisms such as monotone pattern and arbitrary pattern. Berglund, P.A., (2010) noted that arbitrary missing data is used to describe a missing data pattern that has missingness interspersed among full data values while monotone missing data is a pattern in which the missing data exists at the end (reading from left to right) of the data record with no gaps between full and missing data. Table-1 shows the example of arbitrary pattern and Table-2 shows the example of monotone pattern.

Table-1. Arbitrary pattern
1 = value present; 0 = value missing

X	Y	Z
0	0	1
1	0	0
1	1	1
1	0	0

Table-2. Monotone pattern
1 = value present; 0 = value missing

X	Y	Z
1	1	1
1	1	0
1	0	0
1	0	0

The mechanisms of missing data can be divided into three types, first is missing completely at random (MCAR), then missing at random (MAR) and lastly isn't missing at random (NMAR). According to Elizabeth, *et al.*, (2009), MCAR occurs when the missingness is unrelated to the variables under study. In the other words, the missingness is purely random, and the individuals with missing data are a simple random sample of the full sample. MAR means that the probability of an observation being missing may depend on observed values but not on unobserved values. Finally, NMAR means that the probability of missingness depends on both observed and unobserved values.

The selection of missing data patterns and missing data mechanisms are very important when the missing data problem occurs. Magnani, M. (2004) claimed that the effectiveness of a missing data technique depends closely on the missing data mechanism. Horton, N. J. and Kleinman, K. P. (2007) stated, when missingness is non-monotone, models for the missingness of one variable may include covariates which are also missing values, and simpler methods can be utilized if the pattern is monotone, though a monotone pattern is uncommon in most realistic settings. The method of choice depends on the pattern of missingness in the data and the type of the imputed variable (Soley-Bori, M., 2013). Thus both missing data patterns and missing data mechanisms are important for the selection of the plausible statistical imputation model.

Missing data treatment

Various techniques have been proposed to deal with missing data (Geert, J.M.G. *et al*, 2006). Laird, R.J. and Rubin, D.B. (2002) stated that there are three categories which can be used as a technique to deal with missing data. The categories known as ignoring and discarding data, parameter estimation, and imputation techniques.

Ignoring and discarding data is a technique that used two methods which are known as complete case analysis and discarding instances and/or attributes. The complete case analysis will delete all missing data and only use an observed data. The second method is discarding instances and/or attributes. According to Mehala, B *et al*. (2009), this method consists of determining the extent of missing data on each instance and attribute, and deleting the instances and/or attributes with high levels of missing data. But before delete any attribute, evaluation of its relevance to the analysis should be conducted.

Parameter estimation commonly implements the Expectation Maximization algorithms. Expectation Maximization algorithms will first define a model for the complete data, then assumptions regarding the missing data mechanism is made, and the parameters of the model are estimated by maximum likelihood or maximum a posteriori (MAP) procedures (Lakshminarayan, K. *et al*., 1999). Maximum likelihood is the model to estimate the parameters in a statistical model.

Maximum likelihood procedures that use variants of Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data (Lakshminarayan, K. *et al*., 1999). It also can produce efficient parameter estimation because this technique uses all observed data. However according to Lakshminarayan, K. *et al.* (1999), disadvantages of this technique are model based approaches involving parameter estimation are more suitable for users who are familiar with the missing data mechanism and have the necessary expertise and tools for analyzing the incomplete data.

Imputation techniques are process of filling in a value in missing data. Single imputation and Multiple Imputation are two examples of imputation techniques. Single imputation replaces missing data by imputing only once, and after that proceed to complete case analysis, to produce completed a new data set. Thus the predicted value cannot reflect that uncertainty about the value (Ding, Y. and Ross, A., 2012). Furthermore the single imputation concept generally underestimates the standard errors of estimates because choosing a single imputation pretends that we know the unobserved value with certainty, when actually it is unknown but estimated by the imputation method (He, Y., 2010).

However it is different with MI that estimate missing data multiple times and produce multiple sets of complete data. According to He, Y. (2010), MI retains the advantages of Maximum Likelihood estimates while also allowing the uncertainty caused by imputation, which is ignored in single imputation. Therefore MI is better compared with single imputation because it can overcome the problems of single imputation and at the same time produce an unbiased result.

Multiple imputation technique

Multiple Imputation is a missing data treatment that was introduced by Donald B. Rubin in 1987. Rubin D.B. (2003) noted that in the 1970's and 1980's, he was deeply involved with problems of non-response in public use sample surveys, and that was the initial impetus for his proposal of MI. Rubin's first article with the title Inference and Missing Data just discussed about the problems of missing data and the solutions, after that more discussion about this problem occurs until Rubin created the new solution named multiple imputation. Stuart, E.A. (2009) noted that multiple imputation was conceived by Rubin in 1987 and described further by Little and Rubin in 2002, and Schafer in 1997. Horton, N.J. and Lipsitz, S.P. (2001) described the process of MI as shown in Table-3:

Table-3. The process of MI.

Steps	Process
Imputation	Generate a set of *m*>1 plausible values for $Z^{mis} = (Y^{mis}; X^{mis})$.
Analysis	Analyze the *m* data sets using complete-case models.
Combination	Combine the results from the *m* analyses

The imputation step is a process to fill in the missing values in data sets. The variable *m* refers to the numbers of imputation process. He, Y. (2010) noted that it involves creating more than 1 set of replacements for the missing values based on plausible models for data. Commonly, researchers choose between 3 to 10 data sets (Wayman, J.C., 2003). Imputation values are drawn from a distribution which produces different values for each missing data.

In other words, with MI any missing values for any variable will be predicted using any existing values from other variables. The imputes or the predicted values then replace the missing values and thus produce "imputed data sets". This process will repeat in multiple times and produce multiple imputed data sets. Then the standard statistical analysis is applied to each imputed data set, to produce multiple analysis results. And lastly combine all the analysis results to produce one overall analysis.

MI is now a well established technique to analyze data sets where some units have incomplete observations. There are several standard statistical models which can be used in MI are such as regression, propensity scores and Markov Chain Monte Carlo (MCMC).

Linear regression

The selected standard statistical model for this research is a linear regression model (LR). When using MI, the first step is to impute missing data with new values using the standard statistic model. LR will estimate values for missing data in data sets. Multiple LR will be used in this research because more than one independent variables are used for imputing missing data. The equation of multiple LR is:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + e_i \quad (1)$$

In this regression model, Y_i is a missing value or also known as a dependent variable. The dependent variable is the variable that need to be predicted. Then β_1 is the slope and β_0 is the intercept in the equation. The variable X is an independent variable which is also known as predictor variables. The variable X represents observed or present data in the data set. It is also called as a covariates.

According to Guan, S. *et al*. (2009), a regression model is fitted with the observed values for the variable Y_i and its covariates $X_1, X_2, ..., X_k$ and based on the fitted regression model is simulated from the posterior predictive distribution of the parameters (regression parameter estimates and associated covariance matrix) and is used to impute the missing values for each variable.

A symbol e_i, i=1,2,...,n in the above equation are known as a random error. When applying the regression model in MI, the researcher need to include root mean squared error (RMSE) in the regression equation.

The conventional regression imputation normally regresses variable *X* on variable *Y* to get complete data set. This is done to get the result which can be used to impute on missing data in variable *Y*. Normally the result is overestimated and the standard deviation of variable *X* is underestimated. The scenario also means that the variance is too low. This is because the predicted value is a perfect linear function of *Y*. According to Allison P.D. (2008), the solution for this problem is by adding random components to a deterministic prediction equation.

Random components are two components such as random error and RMSE. By doing this step, the variance can be increased and the result will be closed to the real value. Thus when apply MI, the conventional regression imputation with adding a random component need to be done multiple times. Therefore the result would better with good estimation result and also accurate result compare with the conventional regression imputation.

The application of linear regression in medical domains

Missing data problems are very common and mostly occurs in medical domains. Mackinnon, A. (2010) concluded that missing data are ubiquitous problem in nearly all fields of medical research. The observation from Mackinnon, A. (2010) shown a drastic increase in articles on applying MI to data analyses published in four leading medical journals such as the British Medical Journal (BMJ), The journal of the American Medical Association (JAMA), Lancet (weekly peer-reviewed general medical journal) and the New England Journal of Medicine (NEJM).

Comparison of the standard statistic models

Table-4 shows a comparison of the three standard statistical models which are normally used to impute missing data in MI. Each model has advantages and disadvantages according to the issue discusses in the Table-4 such as pattern, fitting model, procedure of process to estimate missing values and form of output.

The LR was selected based on the features of LR which shown in Table-4, and also as it is good in production values and simple to use.

Table-4. Comparison of three standard statistic models.

Models & Issues	Regression	Propensity Scores	MCMC
Pattern	Parametric and monotone	Nonparametric and monotone	Nonparametric and arbitrary
Fitting model	Regression model	Conditional probability	Markov chains
Process	Simulate a posterior prediction of the parameters to create a new logistic regression model	Observation are stratified into a number of strata based on propensity scores	Simulate the result of the complete data posterior distribution
Output	Parameter estimates	Means of logistic regression model	Assumption of multivariate normality

PROPOSED FRAMEWORK

The proposed framework for this research has three main processes. The three main processes consist of input, proposed algorithm and output.

Input

Four data sets from UCI Machine Learning Repository (UCI) are chosen as the sample data in this research. The main reasons of the selection are UCI offered an open source data source, and all data sets are real-world data. Furthermore UCI data sets also easy to be accessed and implemented for any research. Data sets from UCI are widely used as test data set for benchmarking in the area of data mining and artificial intelligence. Table-5 shows the general characteristics for each data set.

Table-5. General characteristics for data sets.

No	Name	Instances	Attributes
1	Ecoli	336	8
2	Haberman’s Survival	306	3
3	Vertebral Column	310	6
4	Liver Disorders	345	7

Proposed algorithm

The second step of the proposed framework is proposed algorithm. The processes involved in this step are implementation of the proposed algorithm and produce imputed data sets.

LR can be implemented using simple linear regression or multiple linear regression. Based on the characteristics of the data sets for this research, a multiple linear regression is more suitable to be used. This is because most of the data sets contain more than one attributes that has a relationship with missing data, thus the other attributes can be used as a predictor in this research.

Output

Lastly the process of comparison will be conducted in order to proof that the proposed algorithm is better. The comparison will use two methods and the methods area imputed data sets produce from the linear regression model, and imputed data sets produce from proposed algorithm

The comparison and analysis activities are based on the accuracy of the values in imputed data sets. Accuracy is about how accurate the data produced from proposed algorithm compare to the linear regression model. According to Richman, M.B., Trafalis, T.B. and Adrianto, I. (2008), the accuracy can be measured by using a mean absolute error formula (MAE):

$$\text{MAE}=1/N * T_j - P_j \tag{2}$$

The MAE measure a performance by differentiating variance between the original data set and imputed data set. The variable N in the MAE equation is the total of iteration, variable T refers to the values of the original data set, variable P is the values of imputed data sets, and variable j refers to the values of the iteration.

Linear regression with half values of random error

The pseudo code for the proposed algorithm in this research is shown in Figure-1. The proposed algorithm is known as LReHalf which means Linear Regression with Half values of random error.

```
Procedure of LReHalf
Input  : Dataset in XLS file, all data
Output :  5 sets of imputed  dataset without missing data
Begin
  Replace missing value with constant numeric values (0 or 0.00)
  For i =1 to N
    For j = 1 to M
      If D (i,j) is not a Number (missing value), then
         Substitute zero to D (i,j)
     End
  End
Read file
    Estimate correlation coefficient
    Perform multiple linear regression to get intercept and slope
    Estimate random errors for covariance parameter
    Estimate RMSE parameter
    For m = 1 to 5
      Perform LReHalf using intercept, slope, half of random
      error and RMSE
      Replace missing data in a new data set
     End
End Read File
End
```

Figure-1. Pseudocode of LReHalf.

According to Patil, D.V. and Bichkar, R.S. (2010), the usage of random error in LR with the larger number of predictors can increase the chance for the imputed estimation molding noise in the data compare to the actual variants to missing data. Therefore the main function of the proposed algorithm is to reduce the value of random error.

Random error which is normally represented by symbol e_i in LR equation, need to be reduced to decrease the values of needless noise. As discussed in the previous chapters, noise is an outlier which is incorrect attribute values and can give bad effects to the estimation for missing data. This problem also reduces the quality of cleaning data when the data set with noise problem is provided to Knowledge Discovery Databases (KDD). Hence the proposed algorithm implements this by dividing the random error with values 2, in other words only half of the actual random error value will be used in this algorithm. The random error cannot be omitted from the LR equation because as discussed before, the random provides the estimated correlation between observed and unobserved values which can produce prediction values that closed to the real values. Therefore the proposed algorithm suggested reducing only half of the random error in the new equation. The new equation for proposed algorithm is:

$$Y_i = \beta_{0\,+}\,\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + s_{x.y.}(e_i\ /2) \tag{3}$$

The imputed values for each data set will be compared to the imputed data sets using (LR) to measure the accuracy of the imputed data sets. The measurement of accuracy will be implemented to proof the proposed algorithm is better and faster than (LR), when handling missing data using multiple imputation technique. Details for the data accuracy and the application of data accuracy will be described in the following section. *j* refers to the values of the iteration.

RESULT AND ANALYSIS

Two experiments were conducted in this research. The first experiment is implemented using two algorithms which are Linear Regression algorithm (LR) and proposed algorithm (LReHalf). These two algorithms are applied to all data sets to obtain the result which shows a performance of each algorithm. After that, followed by the second experiment that measures the accuracy of these two algorithms. The results of the first experiment for all four data sets are shows in Figure-2, Figure-3, Figure-4 and Figure-5.

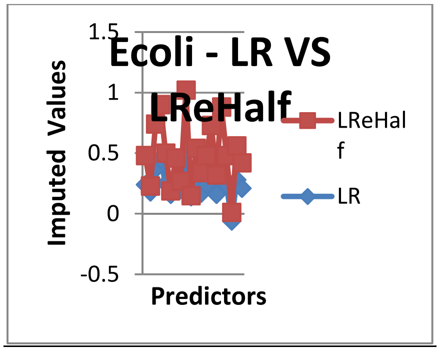

Figure-2. Differences in ecoli.

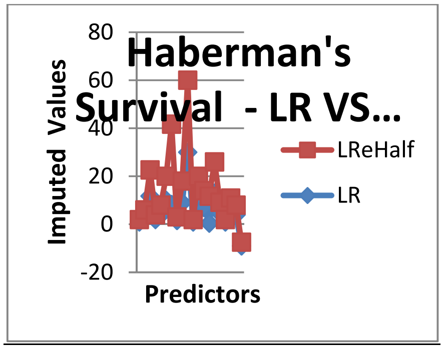

Figure-3. Differences in haberman's survival.

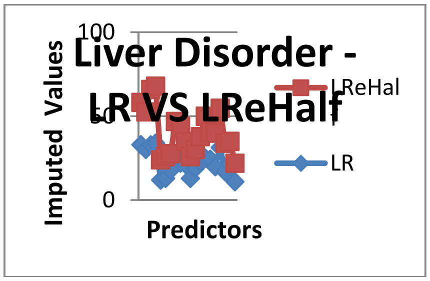

Figure-4. Differences in liver disorder.

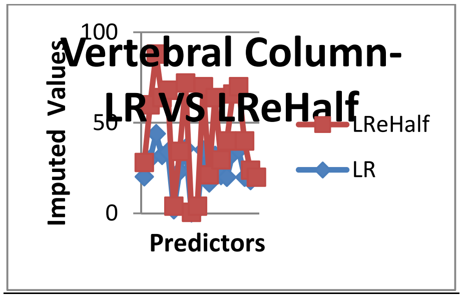

Figure-5. Differences in vertebral column.

The LReHalf produced the scattered output of estimating values and have major differences from LR. This is based on the result from the values of estimation from both algorithms, but it still follows the pattern as the estimation of LR. Except for data set Liver Disorder which shows no duplicate estimation values between LR and LReHalf. Therefore the three main findings from this experiment are shows in Table-6.

Table-6. The main findings of the first experiment.

No.	Result
1	The LReHalf produces estimation values with major differences from LR. The Liver Disorder data set shows the major differences compared to other data sets, which some of the estimation has higher values of estimation. This happens because of the nature of data in each data set.
2	The result of both algorithms has negative estimation values. This indicates that some of the estimation is not close to the real values. In this case, LReHalf performs better than LR as the positive estimated values are higher compared to LR.
3	Overall, the result from LR is better because the estimation values are consistent compared to LReHalf, even though some of the estimations are in a negative value.

The second experiment is the analysis of the data accuracy of all the data sets. Table-7 show the accuracy from each data sets. According to Allison, P.D. (2008), the smaller the variation between the imputed estimation is better as it shows that the value is close to the real value. Therefore the accuracy of the estimation values produced from LReHalf is not good because the variations in the data sets follow the rules highlighted by Allison, P.D. (2008).

Table-7. The accuracy result from all data sets.

Ecoli		**Haberman**		**Liver**		**Verteral**	
0.06	**0.04**	**-5.46**	**-4.49**	**9.93**	**15.9**	**11.05**	**13.5**
0.05	0.18	-5	-5	18.32	22.6	3.6	9.61
0.07	-0.02	-16.74	-5	5.03	9.76	10	6.85
0.04	0.11	-0.3	-0.17	0.77	1.24	15.35	6.75
0.03	0.03	-5.6	-2.11	5.74	4.15	1.49	1.81
0.05	0.25	-2.2	-4.18	2.92	8.72	16.75	6.2
0	0.18	-3.81	-5.1	23.03	21.3	11.7	13.6
0.9	0.13	-0.76	-2	6.31	21.83	15.1	-1.6

The second experiment compares variances of imputed values in data sets using LReHalf. The main findings from this experiment are shows in the Table-8.

Table-8. The main findings of the second experiment.

No.	**Result**
1	Variances produced by LReHalf are not consistent because some of the data sets show small variances, bigger variances and also negative variances. It is a good result when variation is smaller between imputed estimation and the true values.
2	The scenario happens as the nature of data in each data set such as low and high values, and the characteristics of the proposed algorithm is not good for estimation of missing values.

CONCLUSIONS

The aim of this research is to qualify linear regression model as a selected imputation method in MI. This aim is achieved by proposing an enhancement for linear regression model which named as LReHalf, in order to produce high quality data cleaning outputs. The outputs are known as multiple imputed datasets and the quality of the outputs are examined by the accuracy of the outputs.

The two main objectives of this research are handling missing data problems using the selected imputation model in MI, and enhancing the selected statistical model to yield better data cleaning result. As stated previously, the selected statistical model is a linear regression model (LR). This model is able to predict a missing data based on the observed data.

The datasets which are used in this research have arbitrary missing data patterns and it involves two or more predictors in linear regression equation. Therefore the multiple linear regressions have been used to produce the outputs.

According to the results, the outputs from the two experiments show that LReHalf is not efficient to be used in handling missing data. During the implementation of Experiment 1, first the imputed result shows that LReHalf is better than LR. However after imputed multiple times and accuracy performance, the result shows that LR is better than LReHalf.

Therefore the suggestion to reduce the value of random error which contains the needless noise is not relevant when using LR in handling missing data. The effects of reducing the values of random error by dividing with value two, has produced the highest estimation for missing data. Hence this scenario has given major differences when comparing the output from LReHalf and the output from LR.

There are several factors that make the outputs from LReHalf are not relevant to this research. The factors such as the characteristics of LReHalf which only provides half of the random error, the nature of the data in the data sets and the assumption of the mechanism of missing data (Missing at Random). The selection of data set also important as different format of data can reduce the efficiency of the outputs, such as standardization of the decimal places, and value zeroes (0) is also exist in the data set instead of missing data (also need to replace with values zero).

Future research is highly suggested to enhance this research finding such as to reduce the value of random error which in the same time can lower the estimation for missing data. LReHalf was recommended to improve the quality of MI in handling missing data problems, and hopefully this model will benefits all researchers from time to time.

REFERENCES

Allison, P.D. 2001. Missing Data. Thousand Oaks, CA: Sage.

Allison , P.D. 2008. Missing Data 1. Instructor Notes. University of Pennyselvia.

Al Shalabi, L. *et al.* 2006. A Framework to Deal with Mssing Data in Data Sets. Journal of Computer Sciences. 2(9). Pp. 740-745.

Barnard, J. and Meng, X.L. 1999. Applications of Multiple Imputation in Medical Studies: From AIDS to NHANES. Stat. Methods Med. Res. 8(1), 17-36.

Berglund, P.A. 2010. An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2. SAS Global Forum 2010-Statistics and Data Analysis.

Carpenter, J. R. *et al.* 2006. A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. Journal of the Royal Statistical Society. Series A (Statistics in Society). Vol. 169, No. 3 (2006), pp. 571-584.

Chen, M.-S., Han, J. and Yu, P. 1996. Data Mining: An Overview From A Database Perspective. Knowledge and Data Engineering, IEEE Transactions on 8(6): 866-883.

Cummings. P. 2013. Missing Data and Multiple Imputation. JAMA pediatricsVolume 167, Number 7, Pages 656-661.

Ding, Y. and Ross, A. 2012. A Comparison of Imputation Methods for Handling Missing Scores in Biometric Fusion. Pattern Recognition 45. 919 – 933.

Elizabeth A. S. *et al*. 2009. Multiple Imputation with Large Data Sets: A Case Study of the Children's Mental Health Initiative. Am J Epidemiol. 169(9): 1133–1139.

Frawley, W. J., Piatetsky-Shapiro, G. and Matheus, C. J. 1992. Knowledge discovery in databases: an overview.

Geert, J.M.G. *et al*. 2006. Imputation of Missing Values is Superior to Complete Case Analysis and The Missing-Indicator Model in Multivariable Diagnostic Research: A Clinical Example, Journal of Clinical Epidemiology. 59. Pp 1102-1109.

Gebreab, S. Y. *et al*. 2015. The Impact of Lifecourse Socioeconomic Position on Cardiovascular Disease Events in African Americans: The Jackson Heart Stud. Journal of the American Heart Association.

Guan, S. 2009. Stein-Type Improved Estimation of Standard Error under Asymmetric LINEX Loss Function. Statistics, 43, pp. 121-129.

He, Y. 2010. Missing Data Analysis Using Multiple Imputation: Getting to the Heart of The Matter. Journal of The American Heart Association, 3, 98-105.

Horton, N.J. and Lisitz, S.R. 2001. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. The American Statistician. Vol. 55, No. 3. pp 244-254.

Horton, N. J. and Kleinman, K. P. 2007. Much Ado about Nothing: A Comparison Of Missing Data Methods And Software To Fit Incomplete Data Regression Models. The American Statistician. Volume 61, Issue 1, pp79-90.

Jun, Ma, *et al*. 2009. Impute Missing Assessments by Opinion Clustering in Multi-Criteria Group Decision Making Problems. IFSA-EUSFLAT. pp 555-560.

Kantardzic, M. 2002. Data Mining: Concepts, Models, Models, and Algorithms. Journal of Computing and Information Science in Engineering - JCISE, vol. 5, no. 4

Lakshminarayan, K. *et al*. 1999. Imputation of Missing Data in Industrial Databases, Applied Intelligence 11, pp 259-275.

Laird, R.J. and D. B. Rubin. 2002. Statistical Analysis with Missing Data. Second Edition. John Wiley and Sons, New York.

Little, R.J.A. Rubin D.B. 2002. Statistical Analysis with Missing Data, 2nd Edition. Wiley, New York.

Mehala, B. *et al*. 2009. Selecting Scalable Algorithms To Deal With Missing Values. International Journal of Recent Trends in Engineering. Vol. 1. No. 2.

Mackinnon, A. 2010. The Use And Reporting of Multiple Imputation in Medical Research-A Review. Journal of Internal Medicine, 268 :586 – 593.

Magnani, M. 2004. Techniques For Dealing With Missing Data In Knowledge Discovery Tasks. 40127 Bologna – ITALY.

Murray J.S. and J.P. Reiter. 2014. Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models with Local Dependence. arXiv preprint arXiv:1410.0438, 2014. 2, 2014.

Newman, D.A. 2014. Missing Data: Five Practical Guidelines. Organizational Research Methods 2014. Vol 17(4). pp 372-411.

Patil, D.V. and Bichkar, R.S. 2010. Multiple Imputation of Missing Data With Genetic Algorithm Based Techniques, IJCA Special Issue on Evolutionary Computation for Optimization Techniques. pp. 74-78.

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. J. Wiley and Sons, New York.

Rubin, D.B. 2003. Discussion on Multiple Imputation, International Statistical Review, Vol. 71, No. 3, pp. 619-625.

Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data, London. Chapman and Hall.

Schafer, J. L. and Graham, J.W. 2002. Missing Data: Our View Of The State Of The Art, American Psychological Association, Vol.7, No. 2, pp. 147 -177.

Spratt, M. *et al*. 2010. Strategies for Multiple Imputations in Longitudinal Studies. American Journal of Epidemiology. Vol. 172, No. 4, 478-487.

Sterne, J.A. *et al*. 2009. Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls. BMJ.

Soley-Bori, M. 2013. Dealing With Missing Data: Key Assumptions And Methods For Applied Analysis. Journal of the American Statistical Association 89, 425, 278-288.

Stuart E. A. *et al*. 2009. Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative, American Journal of Epidemiology, pp 1-7.

Wayman, J. C. 2003. Multiple Imputation For Missing Data: What Is It And How Can I Use It? Annual Meeting of the American Educational Research Association, pp. 1-16.