American Journal of Theoretical and Applied Statistics
2015; 4(3): 192-200
Published online May 28, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150403.26
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Comparison of Methods of Handling Missing Data: A Case Study of KDHS 2010 Data

Shelmith Nyagathiri Kariuki, Anthony Waititu Gichuhi, Anthony Kibira Wanjoya

Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:
kariukishelmith@gmail.com (S. N. Kariuki), agwaititu@gmail.com (A. G. Waititu), awanjoya@gmail.com (A. K. Wanjoya)

To cite this article:
Shelmith Nyagathiri Kariuki, Anthony Gichuhi Waititu, Anthony Kibira Wanjoya. Comparison of Methods of Handling Missing Data: A Case Study of KDHS 2010 Data. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 3, 2015, pp. 192-200.
doi: 10.11648.j.ajtas.20150403.26

Abstract: Missing data poses a major threat to observational and experimental studies. Analysis of data having ignored missingness results to estimates that are inefficient and unbiased. Various researches have been done to determine the best methods of dealing with missing data. The analysis used in these researches involved simulating missing data from complete data. Missing data are then imputed using the various methods, and the best method is arrived at by looking at the biasness of the imputed estimates, from the complete data estimates and the magnitude of standard errors. This study aimed at establishing the best method of dealing with missing data, based on the goodness of fit tests. The study made use of data from KDHS 2010. The overall rate of missingness was about 80%. The missing data mechanism was tested and proved to be MAR. The missing data was then imputed using Expectation Maximization Algorithm and Multiple Imputation. Later, logistic models were fitted to both datasets. Afterwards, goodness of fit tests were carried out to determine which of the two methods was the better method for imputing data. These tests were the AIC, Root Mean Square Error of Approximation (RMSEA) and Cox and Snell's R-Squared. The predictive ability of the two models was also examined using confusion matrices and the area under receiver operation curve (AUROC). From these tests, multiple imputation was seen to be the better method of imputation since logistic regression model fitted the data better as compared to data imputed using expectation maximization. From the results of the study, the researchers recommend that the type of missingness present in data should be examined. If the amount of missing data is large, and the data is MAR, then data should be imputed using multiple imputation before any inference are made. The researchers suggested more research to be done to determine the maximum rate of missing data that should be imputed.

Keywords: Missingness, Missing at Random, Multiple Imputation, Expectation Maximization

1. Introduction

The goal of any analysis is to obtain unbiased estimates of population parameters, Graham (2012). But this is occasionally impossible due to non-response. Non-response refers to failure to obtain a measurement on one or more study variables for one or more subjects selected for a survey. As a result, the survey tends to have missing data. There are two forms of non-response. Item non-response and unit non-response. Unit non-response occurs when no information is collected for a sampled element. For example when a subject selected to participate in a study, fails to show up. Item non-response occurs when some but not all information is collected for a sampled element. For example a subject dropping out of the study for various reasons e.g. ill health. Non-response occurs frequently in observational and experimental studies. There are varied reasons for missing/ incomplete data in a survey. One of the reasons includes failure of a subject included in the study sample, to show up for the survey. Another reason would be equipment failure leading to incomplete data collection from all subjects. In addition, the questionnaire may contain ambiguous questions. The study may also contain different study designs. For example, the study may contain different questionnaires and thus some information on some variables in the questionnaires may not be collected for all subjects. Most frequently, subjects fail to provide some information, particularly due to matters of confidentiality. Finally, the researcher may also miss out on some information from particular variables during data entry, i.e. poor data entry. Most researchers, especially in observational surveys, mostly assume the issue of non-response in their studies and often proceed to analyze their data as it is. Others remove cases that involve missing data, from their studies, and analyze their data with only the cases that contain all information needed for the study. Analysis of data, having ignored missingness, leads to inefficient analysis and results that are biased.

2. Review of Previous Studies

There exists no literature regarding an acceptable percentage of missing data for valid statistical inferences. According to (Schafer, 1999), a missing rate of 5% or less would be acceptable. Bernett (2001) claimed that a missing data of 10% or more would lead to biased results. However, Tabachnick and Fidel (2012) claimed that missing data mechanism and missing data patterns have a greater impact on research results as compared to the proportion of missing data. Prior to 1980, various ad-hoc methods of dealing with missing data existed. This included list wise deletion, pairwise deletion and mean substitution. These methods were easy to use, but often produced biased results Peng and Zhu (2007). After some few years, Little and Rubin (1987) introduced other two methods, Expectation Maximization algorithm (EM) and Full Information Maximum Likelihood (FIML). These methods were seen to be more superior to the previous adhoc methods in that they produced better estimates with smaller and acceptable standard errors. Finally, in the late 80's, more superior methods such as Multiple Imputation, were developed. These methods were proved to be flexible and produced smaller standard errors as compared to earlier methods. Ibrahim (1990) proposed Expectation-Maximization (EM) method of weights to obtain maximum likelihood estimates of regression coefficients for the logistic regression model with missing categorical covariates. Later, Little (1992) looked at different methods for handling missing values in covariates in regression analysis and concluded that the preferred methods were model-based estimation methods. Graham (2002) considered both likelihood based methods and Parametric Methods (MI methods). They further suggested that more research should be done on comparison of MI and weighting techniques such as Ibrahim's EM method. Didelez (2002) investigated Maximum Likelihood estimates of logistic regression coefficients with missing values in covariates when the distribution of these covariates was misspecified. From her work, it was concluded that the parametric approach could cause major biasness of the results if the assumed distribution was different from the true distribution. Raghunathan (2004) compared the magnitude of bias by using three methods of dealing with data missingness in logistic Regression. Complete Case (Listwise deletion), a weighting technique and MI. From the results, the bias was greatest in the Complete Case method. The method of MI produced estimates whose sampling distributions were close to the true population. She also noted that the method of weights was not as efficient as MI. She also noted that both methods were valid only under the MAR assumptions. Raghunathan also discussed the ML method and its non-suitability for practical purposes due to technical difficulties. Peng and Zhu (2007) later carried out an analysis to compare the MI technique and EM technique. These two approaches were compared for dealing with missing data in categorical explanatory variables in logistic regression. The results were then compared to those obtained when Complete Case Method was used. From this study, it was noted that MI was more efficient, as compared to EM. The biasness of the results was worst in the CC method. Generally, the study concluded that MI method was better than EM method. He (2010) carried out a study to determine the association between patients' cardiovascular disease variables and hospice discussion. The outcome variable was patients' hospice discussion and the predictor variables included myocardial infarction, heart failure, stroke, and diabetes. Multiple Imputation analysis was carried out using the MICE approach and logistic regression model was built. Complete Case Analysis was also carried out and the logistic regression model built. The results of the study indicated that regression estimates from the CC and MI are somewhat different and the latter produces smaller standard errors than the former for all regressors, illustrating the superior efficiency in the Multiple Imputation Method. This research study aimed at testing the missing data mechanism present in KDHS 2010 data and comparing the methods of dealing with missing data, based on the predictive ability and goodness of fit tests.

3. Methodology

3.1. Missing Data Mechanism

According to Little and Rubin (1987), missing data mechanism may be classified as one of the following three types. The first one is missing Completely At Random (MCAR) which is a mechanism where the probability of a value missing is not dependent on observed or the unobserved values but on some unknown parameter. The second one is missing At Random (MAR) assumption which states that the probability of a value missing does not depend on the missing value itself but on the observed values and an unknown parameter. The third one is missing not at Random, (MNAR). In this case, missingness is no longer at random. MNAR assumption states that the probability of missing values depends on the unobserved values themselves. For example, people having high income are less likely to reveal them. The probability of missing data in the income variable is dependent on the variable itself.

3.2. Patterns of Missing Data

There are three patterns of missing data: Univariate, monotone and arbitrary. Let $X_1, X_2, \ldots, X_j, \ldots, X_p$ be variables in a study. Also let X_{ij} be the entry of the ith case in the jth variable. A dataset is said to have a univariate missing data pattern if there's data missing for one or more of the j variables, for a particular case i. A monotone pattern exists when you can order the variables such that, if a variable has a missing value, all preceding variables also have missing values. This mostly happens when subjects drop out of a study. A dataset is said to have an arbitrary missing data pattern if missing data occur in a random manner for any case and for any variable.

3.3. Methods of Dealing with Missing Data

3.3.1. Complete Case Analysis

According to Graham (2012), the method involves dropping cases that contain missing information for some variables and analyzing the data with only cases that contain full information for every variable

3.3.2. Multiple Imputation

Multiple Imputation involves making repeated random draws from the predictive distribution of the missing data conditional on the observed data, Bouhlila and Sellaouti [2013]. This method involves creating more than one set of replacements for the missing values. As a result, multiple completed data sets are obtained. Each completed data set is analyzed separately. Point estimates and standard errors for each of the variables, from each analysis are then obtained. These sets of point estimates and standard errors are combined to obtain a single point estimate, standard error, associated confidence interval and p-value. This step involves calculating the average of the estimates across multiple imputations and variances of estimates both within and between imputations. According to John W. Graham and Gllreath (2007), the main idea of multiple imputation is that plausible values may be used in place of the missing values in a way that allows parameter estimates to be unbiased thus making them more important and secondly the uncertainty of parameter estimation in the missing data case to be estimated in a reasonable way. According to Rubin (1987), multiple imputation involves three steps. The first step is the Imputation Step. Here missing values are imputed using Multiple Imputation by Chained Equations (MICE) model. It is also known as Sequential Regression Multiple Imputation, He (2010). In MICE, multivariate data have different conditional models for each incomplete variable. The imputation model is specified separately for each variable, using other variables as predictors. For example linear regression is used for continuous variables and logistic regression is used for binary variables, He (2010). This step involves drawing random samples of missing data based on information obtained from the observed data. The second step is the statistical analysis step. In this step the m sets of data are analyzed separately using statistical procedures MICE. Point estimates and estimated standard errors are extracted from the analysis. The final step involves combining results from the previous steps. This involves combining the m point estimates and the estimated standard errors to arrive at a single point estimate, its estimated standard error, and the associated confidence interval or significance test. From Rubin (1997), the following rules for multiple imputations are defined. These are known as the Rubin rules. According to Dong and Peng (2013), let S^i denote the estimate of parameter S, from the i^{th} dataset.

The pooled estimate of the parameter estimate S is calculated as the average of the m estimates of the same parameter.

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m} \widehat{S_\iota}$$

The within imputation variance is the average sampling variance derived by treating the imputed values as though they were real. Let the estimated variance of $\widehat{\boldsymbol{S_\iota}}be\widehat{\boldsymbol{U_\iota}}$. It is given as:

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} \widehat{U_\iota}$$

The between imputation variance is the variability across the imputed values. It is given as:

$$B = \frac{1}{m-1}\sum_{i=1}^{m} \left(\widehat{S_\iota} - \bar{S}\right)^2$$

The variance of the pooled estimate, is the weighted sum of two variances, i.e. the within imputation and the between imputation variances.

$$var(\bar{S}) = \bar{W} + \left(1 + \frac{1}{m}\right) \times B$$

The overall standard error is given as the square root of the variance of the pooled estimate. The number of imputations needed to produce the most accurate results will be determined by the efficiency of the estimate based on the imputations. The relative efficiency of using the finite m imputation estimator rather than using an infinite number for the fully efficient imputation, in units of variance, is approximately a function of m and fraction of missing information. The relative efficiency is given as:

$$r.e = \left(1 + \frac{\gamma}{m}\right)^{-1}$$

Where γ is the fraction of missing information given by the equation:

$$\gamma = \frac{riv + (2/(df_m + 3))}{riv + 1}$$

riv is the relative increase in variance due to missing data? It is the adjusted between-imputation variance standardized by the within-imputation variance. It is given as

$$riv = \frac{(1 + \frac{1}{m}) \times B}{\bar{W}}$$

3.3.3. Expectation Maximization Algorithm

According to Chang and Kim (2007) the EM algorithm estimates the parameters directly by maximizing the complete data log likelihood function. It does this by iterating between the E and M steps. The first step is the E-Step (Expectation Step) .At the E (expectation) step, expectation of the log-likelihood function of the parameters, given the observed data is calculated.

Given that$Y = (Y_{observed}, Y_{missing})$, the distribution of Y conditional on an unknown parameter $\emptyset$ is:

$$Pr\ (Y/\emptyset) = Pr\ (Y_{observed}, Y_{missing}/\emptyset)$$
$$= Pr\ (Y_{observed}/\emptyset) \times Pr\ (Y_{missing}/Y_{observed}, \emptyset)$$

In terms of the likelihood function, the posterior probability of each estimate or rather the complete-data likelihood is given as:

$$L\ (\emptyset/Y) = L(\emptyset/Y_{observed}, Y_{missing})$$
$$= L(\emptyset/Y_{observed}) \times Pr\ (Y_{missing}/Y_{observed}, \emptyset)$$

Taking logs on both sides,

$$l(\emptyset/Y) = l(\emptyset/Y_{observed}, Y_{missing})$$
$$= l(\emptyset/Y_{observed}) \times logPr\ (Y_{missing}/Y_{observed}, \emptyset)$$

Where $l(\emptyset/Y)$ is the complete data log likelihood and $l(\emptyset/Y_{observed})$ is the observed-data log likelihood. $Pr\ (Y_{missing}/Y_{observed}, \emptyset)$is the predictive distribution of the missing data, given$\emptyset$.We cannot predict the likelihood of the complete-data log likelihood $l(\emptyset/Y)$, since the distribution of the missing data is unknown. But we can compute the expectation of the likelihood of the complete data set given an initial guess of the parameter$\emptyset$ $\boldsymbol{ie\emptyset^0}$. This first guess can be determined by first analyzing the data by complete case. This gives a rough idea of the estimate of the parameter. The expectation of the complete data log-likelihood function is calculated as follows:

$$Z(\emptyset/\emptyset^0) = E\{l(\emptyset/Y)/Y_{observed}, \emptyset^0\}$$

Since Y is a missing random variable under an assumed distribution $Pr\ (Y_{missing}/Y_{observed}, \emptyset)$, then the expectation of the complete data log-likelihood function can be written as:

$$Z\left(\emptyset\frac{}{\emptyset^0}\right) = \int l\left(\emptyset\frac{}{Y}\right) \times \Pr\left(\frac{Y_{missing}}{Y_{observed}}, \emptyset^0\right) dY_{missing}$$

The second step is the M-Step (Maximization Step). At this step, the next guess of $\emptyset$is obtained by maximizing the expectation of the complete data log likelihood from the previous E-step.

$$\emptyset^t = argmax_{\emptyset}(Z(\emptyset/\emptyset^0))$$

This guess of$\emptyset$is thus used in the next E-step. The EM algorithm then alternates between E-Step and M-step. The algorithm is terminated when the successive estimates of$\emptyset$ are almost identical.

4. Results and Discussion

4.1. Description of the Data

The study involved secondary data analysis of the 2010 Kenyan Demographic and Health Survey (KDHS) dataset for children. The data involved 950 cases randomly chosen from the total cases interviewed. The dependent variable in the data was created following the age of death of the child. If the child had lived for twelve months or less before death, the case was treated as infant mortality. Table 1 shows the explanatory variables that were used in the study.

Table 1. *Table of Explanatory Variables.*

Variable	Definition
BORD	Birth Order Number
DELIVERY	Place where the child was born
A.VISITS	Number of antenatal visits
SIZE	Size of child at birth
BF	Months of Breastfeeding
AGE_1stBIRTH	Age of the Mother at her first birth
WORKING	Whether the respondent is working or not
RESIDENCE	Type of Place of Residence
BWEIGHT	Birth Weight in grams
EDUC.LEVEL	Highest Education Level
SMOKE	Whether the respondent smokes or not

Table 2. shows the distribution of the missing values under each variable.

Table 2. *Table of distribution of missing values per variable.*

Variable	Frequency	Percentage
BORD	0	0
DELIVERY	772	81.3%
A.VISITS	685	72.1%
SIZE	685	72.1%
BF	691	72.7%
AGE_1stBIRTH	0	0.0%
WORKING	0	0.0%
RESIDENCE	0	0.0%
BWEIGHT	685	72.1%
EDUC.LEVEL	0	0.0%
SMOKE	0	0.0%

From the table, the variables containing missing data are size of the child at birth, months of breastfeeding, place of delivery, birth weight and number of antenatal visits. Those that do not contain missing data are birth order, age at first birth, whether the respondent is working or not, residence, education level, and whether the respondents smokes or not.

4.2. Assessing the Missing Data Pattern

Table 3. *Table of Missing Data Pattern.*

No. of. Cases	Residence	Educ. Level	Age_Ist Birth	Smoke	Birth Order	Working	Inf. Mortality	Size	Delivery	Birth weight	Months of BF	A. Visits	Missing Var
178	1	1	1	1	1	1	1	1	1	1	1	1	0
81	1	1	1	1	1	1	1	1	1	1	1	0	1
6	1	1	1	1	1	1	1	1	1	1	0	0	2
685	1	1	1	1	1	1	1	0	0	0	0	0	5

From Table 3, the data contained four missing data patterns. As seen in the first column, 178 cases had no missing values in all the variables, 81 cases had missing values in the number of antenatal visits variable, 6 cases had missing values in both the months of breastfeeding and number of antenatal visits and 685 cases had missing values in five variables, namely, size of child at birth, place of delivery, birth weight, months of breastfeeding and number of antenatal visits. From the last column, 178 cases did not have any missing values, 81 cases had missing values in only one column i.e. number-of-antenatal-visits, 6cases had missing values in two columns i.e. months of breastfeeding and number of antenatal visits variable and 685 cases had missing values in 5 columns i.e. size of child at birth, place of delivery, birth weight, months of breastfeeding and number of antenatal visits.

4.3. Assessing the Missing Data Mechanism

4.3.1. Test for MCAR Mechanism

Test for MCAR was done using Little MCAR Chi-square test. The chisquare statistic obtained was 132.2474 with 28 degrees of freedom and a p-value of 1.776357×10^{-15}. Since this p-value is less than 0.05, the null hypothesis was rejected thus the researcher concluded that the data was not Missing Completely at Random. The researcher therefore went ahead to test the MAR assumption.

4.3.2. Test for MAR Mechanism

The MAR assumption was tested by creating dummy variables for each variable that had missing values in it. The researcher coded 1where the value in the original variable was missing and 0 where the value was observed. The researcher then ran chi-square tests between each dummy variable and other variables in the dataset to see if the missingness in this variable was related to the values of the other observed variable. If the p-value between a dummy variable and an observed variable was less than 0.05, the null hypothesis was rejected concluding that missingness was dependent on the observed variable. For all the five variables containing missing values, i.e. size of child at birth, months of breastfeeding, place of delivery, birth weight, and number of antenatal visits, missingness was dependent on residence, educational level, age at first birth and infant mortality. Since the p-values of the four variables were less than 0.05, the null hypothesis was rejected and thus the research concluded that the missingness was dependent on the stated variables. So in general, the study concluded that there was dependence between completely observed variables and missingness and thus missing observations in the data were Missing at Random.

4.4. Methods of Dealing with Missing Data

4.4.1. Complete Case Analysis

The three methods that were intended to be used in this study include the complete case, the multiple imputation method and expectation maximization method. But the missing data mechanism present in this data is MAR, thus the use of complete case would result into biased estimates. The researchers thus decided to use multiple imputation and expectation maximization to deal with the missing data.

4.4.2. Multiple Imputations

(i). The Imputation Procedure

The missing data was imputed using Multiple Imputation. One of the assumptions of using Multiple Imputation is that the missing data needs to be MAR, as had already been tested and proved. The method used to impute the missing values was dependent on the type of variable itself. For size and delivery, since they are categorical variables with more than two levels, polytomous regression was used for imputation. For numeric variables such as months of breastfeeding, birth weight and number of antenatal visits, predictive mean matching was used. The algorithm imputed each incomplete column in the data from left to right. The variables that predicted missingness under each variable containing missing values were used to impute the missing values in that variable. To determine the number of imputations, the study first made use of 100 imputations. This was motivated by White et.al's claim that the number of imputations should be at least equal to the percentage of missing cases in the data. Since the percentage of missing cases was 82%, the study made use of 100 imputations. However, there is no literature on the optimal number of imputations to use for a given rate of missing information. The researcher thus decided to impute the data using various numbers of imputations in order to examine the behavior of the logistic model. The research made use of 100, 200,500, 750 and 1000 imputations. The statistics used to judge on the optimal number of imputations include the the within imputation variance, the fraction of missing information as a result of the missing data and the relative efficiency of using the specified number of imputations given certain fraction of missing data. The within imputation variances generally decreased as the number of imputations approached 1000. The fraction of missing information due to non-response generally decreased, albeit by a small margin, as the number of imputations approached 1000. The relative efficiencies for using the m imputations were all very high, 99%.This is because the fraction of missing information were relatively small as compared to the number of imputations. From the results, the research concluded that using a large number of imputations yielded smaller fractions of missing information and very relatively efficiencies. Since there was a general decrease in the statistics as the number of imputations approached 1000, the researchers thus decided to use 1000imputations for the study.

(ii). Diagnostic Checking of the Imputations

From the results achieved from using 1000 imputations, summary statistics of the observed and imputed values were examined. This was in order to achieve the extent to which the imputed values matched or differed from the imputed values. Kobi Abayomi and Levy (2008) in their paper of 2008, titled ‘Diagnostics of Multivariate Imputations’ suggested the use of Kolmogorov Smirnov test to compare the distribution of the observed data and the imputed data, for each variable, and raise a flag when statistically significant differences were found. Moreover, they indicated that a difference in distribution does not necessarily signal a problem with imputation, unless the difference is very big. Let F_O be the distribution of the observed dataset and let F_I be the distribution of the imputed data set. The hypothesis to be tested here is:

$$H_O:F_O=F_I$$

vs

$$H_A:F_O\neq F_I$$

Table 4 shows the p-values obtained from the Kolmogorov Smirnov Test of the data before and after imputation through Multiple Imputation.

Table 4. *Table of Kolmogorov Smirnov Test of MI.*

Kolmogorov Smirnov Test	
Variable	P-Value
A.VISITS	0.9477
DELIVERY	0.4887
SIZE	1.0000
BF	0.5272
BWEIGHT	0.2276

From the table, the p-values were all greater than 0.05, so there was no sufficient evidence to reject the null hypothesis. The study thus concluded that the observed data and the imputed data had the same distribution and thus the imputations were viable.

4.4.3. Expectation Maximization Algorithm

(i). The Imputation Procedure

The incomplete data was then imputed using Expectation Maximization criterion. This involved imputing the missing data only once, by getting the maximum likelihood estimates of the available data and using these estimates to impute the missing data.

(ii). Diagnostic Checking of the Imputations

The imputation was accessed to check whether the imputed values matched the observed data. Table 5shows the p-values obtained from the Kolmogorov Smirnov Test of the data before and after imputation through Expectation Maximization Algorithm.

Table 5. *Table of Kolmogorov Smirnov Test of EM.*

Kolmogorov Smirnov Test	
Variable	P-Value
A.VISITS	0.0000
DELIVERY	0.0000
SIZE	0.0000
BF	0.0000
BWEIGHT	0.0000

From the table, the p-values were all less than 0.05, so there was insufficient evidence to reject the null hypothesis. The study thus concluded that the distribution of the observed data and the imputed data did not match and thus the imputation was not viable. This was due to the fact that the percentage of missing information was larger than that of the available information. So using a small percentage of information, to impute a large percentage of information, leads to the distortion of the distribution of the data

4.5. Fitting the Logistic Model

Since the occurrence of infant mortality was a binary outcome, logistic models were fitted to both datasets; i.e the dataset imputed using Multiple Imputation and the one imputed using Expectation Maximization Algorithm. The models were fitted in order to establish how significant and important the explanatory factors were in explaining infant mortality. Table 6 shows the estimates, standard errors, width of confidence intervals, odds ratio and P-Values obtained for each of the explanatory variables

Table 6. *Logistic Model Estimates of Expectation Maximization Algorithm and Multiple Imputation.*

Variable	Category	Estimate	SE	Width	OR	P-Value
Constant		-3.4965 (-4.231)	1.3014 (1.3869)	5.1024 (5.7689)	0.0303 (0.0145)	0.0072 (0.0023)
Birth Order		0.1046 (0.1721)	0.0715 (0.0615)	0.2802 (0.2416)	1.1102 (1.1878)	0.1435 (0.0051)
Number of Antenatal Visits		0.0040 (0.1894)	0.1752 (0.1577)	0.6873 (0.6185)	1.0039 (1.2085)	0.9820 (0.2298)
Delivery	Public	1.6907 (0.3950)	0.6800 (0.3167)	2.6669 (1.2433)	5.4236 (1.4844)	0.013 (0.2123)
	Private	1.7494 (-0.3165)	0.8071 (0.8068)	3.165 (3.2332)	5.7513 (0.7287)	0.0303 (0.6949)
	Others	-0.0958 (-13.38)	242.4931 (851.985)	950.56 (-)	0.9086 (0.0000)	0.9997 (0.9875)
Size	>Average	-2.0008 (0.1448)	0.7870 (1.2349)	3.0861 (5.29)	0.8180 (1.1559)	0.7986 (0.9066)
	Average	0.3816 (1.3072)	0.6920 (1.1075)	2.7135 (4.761)	1.4647 (3.6957)	0.5813 (0.2379)
	<Average	0.1665 (0.6983)	0.7827 (1.2416)	3.0691 (5.3079)	1.1812 (2.0104)	0.8315 (0.5738)

Variable	Category	Estimate	SE	Width	OR	P-Value
	Very Small	1.0369 (1.8006)	0.8580 (1.3338)	3.3647 (5.6515)	2.8205 (6.0533)	0.2269 (0.1770)
Birth Weight		-0.0003 (0.0000)	0.0002 (0.0001)	0.0009 (0.0005)	0.9997 (1.0000)	0.2368 (0.7389)
Months of Breastfeeding		-0.2006 (-0.1377)	0.0405 (0.0274)	0.1588 (0.1077)	0.8183 (0.8714)	0.0000 90.878
Age at 1st Birth		0.0073 (0.0002)	0.0429 (0.0333)	0.168 (0.1309)	1.0073 (1.0003)	0.8654 (0.9938)
Working	Yes	0.1744 (-0.0828)	0.3426 (0.2414)	1.3431 (0.9458)	1.1905 (0.9205)	0.6108 (0.7315)
Residence	Rural	2.1798 (1.7785)	0.5247 (0.3957)	2.0571 (1.5627)	8.8450 (5.9212)	0.0003 (0.0000)
Education Level	Primary	0.1675 (-0.0601)	0.4272 (0.2654)	1.6749 (1.0436)	1.1823 (0.9416)	0.6951 (0.8206)
	Secondary	-1.0927 (-1.1591)	0.6328 (0.4099)	2.4808 (1.6172)	0.3353 (0.3138)	0.0842 (0.0047)
	Higher	0.0471 (0.3775)	1.0630 (0.7633)	4.1672 (3.0709)	1.0482 (1.4587)	0.9647 (0.6209)
Smoke	Yes	-12.195 (-12.14)	708.137 (547.89)	2775.8444 (-)	0.0000 (0.0000)	0.9863 (0.9823)

+ The values in brackets are for Expectation Maximization Algorithm model while those without brackets are for Multiple Imputation.

4.6. Evaluating the Logistic Model

4.6.1. Assessing the Importance of the Variables

From the MI model, for a one unit increase in the birth order, the odds of infant mortality increase by 11%.A one unit increase in the number of antenatal visits results in a 1% decrease in the odds infant mortality. A child whose size is greater than average is about 2% more likely to die as an infant, as compared to a child of large size. One that is of average size is 46%more likely to die as an infant, as compared to a child of large size. A child whose size is less than average is 18% more likely to die as an infant, as compared to a child of large size. A child born of a very small size is 2 times more likely to die as an infant, as compared to a child born of large size. Children born by working ladies have 19% higher chances of dying as infants as compared to children born by women who are not working while those bore by women living in rural areas are 9 times more likely to dye as infants, as compared to those living in urban areas. A one unit increase in the months of breastfeeding reduces the chances of infant mortality by 80% .Other factors kept constant, the odds of infant mortality, for a child born by a woman who delivers in a private place are 5 times greater than the odds of those born by women who delivers at home. The odds of infant mortality of a child born by a woman, who delivers in a public place, are 6 times higher than a child delivered at home. Children bore by women who have attained a secondary education are 66% less likely to die as infants as compared to those bore by women with no education, while those bore by women who's highest level of education is primary are 18%more likely to die as infants as compared to those bore by women with no education.

From the EM model, for a one unit increase in the birth order, the odd of infant mortality increase by 19%. A one unit increase in the number of antenatal visits results in a 21% increase in infant mortality. A child whose size is greater than average is 16% more likely to die as an infant, as compared to a child of large size. One that is of average size is 4 times more likely to die as an infant, as compared to a child of large size. A child whose size is less than average is 2 times more likely to die as an infant, as compared to a child of large size. A child born of a very small size is 6 times more likely to die as an infant, as compared to a child born of large size. Children born by working ladies have 8%lower chances of dying as infants as compared to children born by women who are not working while those bore by women living in rural areas are 6 times more likely of dying as infants, as compared to those living in urban areas. A one unit increase in the months of breastfeeding reduces the chances of infant mortality by 12%. Other factors kept constant, the odds of infant mortality, for a child born by a woman who delivers in a private place are 4 times greater than the odds of those born by women who deliver at home. The odds of infant mortality of a child born by a woman, who delivers in a public place, are 48 % higher than a child delivered at home. Children bore by women who have attained a secondary education are 70% less likely to die as infants as compared to those bore by women with no education, while those bore by women who's highest level of education is primary are 6% less likely to die as infants as compared to those bore by women with no education.

4.6.2. Assessing the Significance of the Variables

The Wald test was used to determine the significance of each parameter in the model. From the MI model, the significant variables were delivery in a public and private place, months of breastfeeding and residence. From the EM model, the significant variables were birth order, months of breastfeeding, residence and educational level (secondary).

Table 7. *Likelihood Ratio Test.*

Likelihood Ratio Test		
Imputation Method	Test Statistic	P-Value
Expectation Maximization	85.62	0.0001
Multiple Imputation	298.92	0.0000

The likelihood Ratio test was also used to determine the joint significance of the models. Table 7 shows the Likelihood Ratio test statistics and their p-values.

From Table 7, Multiple Imputation has a smaller test statistic value and p-value, of the two models. This means that the predictors of the Multiply Imputed dataset model were more significant in explaining infant mortality better than predictors of the model of the dataset imputed using Expectation Maximization method.

4.6.3. Assessing the Goodness of Fit of the Models

Three tests were used to test the goodness of fit of the two models. These tests include the Akaike's Information Criterion (AIC), the Root Mean Square Error of Approximation (RMSEA) and the Cox and Snell's R-Squared. Table 8 shows the values obtained from the three tests, for the EM and MI models.

Table 8. *Goodness of Fit Tests.*

	Goodness of Fit Tests		
Imputation Method	AIC	RMSEA	COX
Expectation Maximization	684.61	2.8406	0.0862
Multiple Imputation	505.49	0.2709	0.2689

From the table, based on the RMSEA and Cox and Snell's R-Squared value, Multiple Imputation was seen to be the better model. Multiple Imputation had a smaller AIC value as compared to Expectation Maximization. Based on Anderson and Burhnam 2002, when the difference in AIC values is greater than 7, then there is strong evidence to support the conclusion of differences between the models. Hence generally, Multiple Imputation was a better method of Imputation as compared to Expectation Maximization.

4.6.4. Assessing the Predictive Power of the Models

The predictive abilities of the two models were later assessed using Area under Receiver Operating Curve (AUROC) and Accuracy of the models, as obtained from confusion matrices.

Table 9 shows the values obtained under the two measures.

Table 9. *Measures of Predictive Ability of the Models.*

Measures of Predictive Ability		
Imputation Method	AUROC	ACCURACY
Expectation Maximization	0.733	0.8695
Multiple Imputation	0.916	0.8916

From the table, the MI model was more accurate in predicting infant mortality as compared to EM model. This is because the MI model had a higher accuracy value (0.8916), as compared to the EM model (0.8695).

5. Conclusion and Recommendation

The objectives aimed by the study were met. The type of missing data present in the data was shown to be missing at Random. The missing data was imputed using Expectation Maximization Algorithm and Multiple Imputation. The study made use of 1000 imputations. Imputation using Multiple Imputation was seen to be more viable as compared to Expectation Maximization, since the distribution of the data after imputation was the same as that before imputation. Previous studies have shown that when data is MAR, complete case produces estimates that are biased, and hence it was not considered as a method of dealing with missing data. Results from fitting logistic regression model showed that Multiple Imputation model was a better fit as compared to Expectation Maximization. This was as a result of examining the distribution of the fitted values, before and after imputation, the predictive ability of the model and the goodness of fit tests of the models. The researchers recommend that the type of missingness present in the data should be examined. If the amount of missing data is large, and the data is MAR, then data should be imputed using multiple imputation before any inference are made.

Acknowledgements

The researchers would want to acknowledge www.measuredhs.com for providing the data used in this study.

Nomenclature

AUROC: Area under Receiver Operating Characteristic Curve
CC: Complete Case Analysis
EM: Expectation Maximization Algorithm
FIML: Full Information Maximization Likelihood
KDHS: Kenya Demographic and Health Survey
MAR: Missing at Random
MCAR: Missing Completely at Random
MNAR: Missing Not at Random
MI: Multiple Imputation
MICE: Multiply Imputed Chained Equations

References

[1] Alan Agresti. An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc.,Hoboken, New Jersey, 2007

[2] Shu-Ching Chang and Hyung Jin Kim. Em algorithm. December 9, 2007.

[3] Dong and Peng. Principled missing data methods for researchers. Springler Plus, 2013.

[4] Joseph L.Shafer and John W. Graham. Missing data: Our view of the state of the art. *Psychological Methods,* 2002*, 7*, 147-177

[5] Yulei He. Missing data analysis using multiple imputation: Getting to the heart of the matter. National Institute of Health Public Access, January 1 2010.

[6] Nicholas J. Horton. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. National Institute of Health Public Access, February 2007.

[7] Tamara Brian Wilfried Laubach Jochen Hardt, Max Herke. Multiple imputation of missing data: A simulation study on a binary response. Open Journal of Statistics, 3:370_378, 2013..

[8] Ting Hsiang Lin. A comparison of multiple imputation with em algorithm and mcmc method for quality of life missing data. Springer Science + Business Media B.V., September 2008.

[9] Joseph L.Shaferand John W. Graham. Missing data: Our view of the state of the art. Psychological Methods, 7(2):147-177, January 2002.

[10] Show-Mann Liou Chao-Ying Joanne Peng, Michael Harwell and Lee H. Ehman. Advances in missing data method and implications for educational research. page 6, June 2003.

[11] J.W Graham. Missing Data Analysis and Design. Springer, 2012.

[12] Gabriele B. Durrant. Imputation mmethod for handling item-nonresponse in the social sciences. June 2005.

[13] Andrew Gelman Kobi Abayomi and Marc Levy. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society, 57:273291, November 2008.