Innovations

Iterative Approaches Used in Obtaining Missing Observations in Nigeria Health Service Data

Efosa Michael Ogbeide

Department of Mathematics and Statistics,
Ambrose Alli University, Ekpoma, Edo state, Nigeria

Abstract

Missing data are common in predictive research in some reported data. This negatively affect the inference from such data set. Some researchers do use values from available observed data to make an estimation of the missing data in a non-ignorable non responses survey. Missing data are issues occur in most researches in education and health sectors records. It could occur in survey and non-survey researches. In this research, we apply and examined the performance of some missing imputations approaches. A secondary data from a national Health Service data was used based on the National Health and Nutrition Examination Survey (NHANES) from 2011 - 2012 website. This work is based on the use of iterative imputation approaches of the Least squared (LS) method, Expectation Maximization (EM) method and Multiple Imputations (MI) method. The efficiencies and performances of these approaches were evaluated via their absolute Raw Bias (RB), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and variance through the use of R Software. The data from the National Health and Nutrition Examination Survey (NHANES) data set has 24 observations, each having 4 features – age, body mass index, hypertension status and cholesterol level. Our results showed that Multiple Imputations had the lowest MSE of value of 0.284353, which happened to be the best performance approach for the missing data. This was closely followed by EM approach with MSE value of 0.35637 and LS with MSE value of 0.6453678.

Keywords: *missing data, imputation, iterative, estimation, error*

Introduction:

Missing data are issues that most researchers encounter on a routine basis. In survey research there can be many reasons for missing data such as respondents ignoring a few or all questions, questions being irrelevant to the respondent's situation, or inability of survey administrators to locate the respondent. This is when questionnaires are administered. Missing data can also occur in non-survey data, such as experimental data like agricultural experiment and administrative data including medical field (Grooves and Fowler, 2004; Acock, 2005). In non-survey samples, missing data can arise due to carelessness in observation recording, errors made during data entry or data loss due to misplacement. In this study, health service data is considered. Regardless of the reason why data is missing, once it is missing it becomes part of the dataset that is then used by researchers to perform analytical procedures. The quality of such analytical procedures directly depends on the quality of underlying data which in turn can be affected by the nature of missing data (Allison, 2001; Schafer and Graham, 2002).

Unfortunately there are many different methods of handling missing data which can have profoundly different effects on estimation. For this reason it is important to select the correct missing data handling method that is suitable to a researcher's particular circumstances, according to Ogbeide and Osemwenkhae (2014). These circumstances can be expressed as factors, such as sample size, proportion of missing data and method of analysis. Some which may fall under the control of the researcher in a given scenario and thus can be manipulated, while others are more difficult to control. For example, a researcher working with secondary data may likely not find it possible to increase the sample size to offset the effect of missing data but may have flexibility regarding the choice of analytical method. On the other hand, a researcher who is gathering her own data and who is relying on a specific method of analysis to answer her research questions may find it easy to increase her sample size in order to lower the proportion of missing cases. As these illustrations suggest, the scenario under which a researcher handles missing data can vary considerably depending on the researcher's circumstance. There were many investigations and comparisons of the performance of missing data handling methods in Peng, T.D., Harwell, M., Liou, S.M. and Ehman, L.H. (2006). Missing data occur in most studies in the behavioral sciences (Alosh, 2009, Allison, 2001; Pigott, 2001; Streiner, 2002 and Acock, 2005). Although adequate reporting and handling of missing data are essential for understanding results, this element of data analysis is often omitted from reports of research,
(Peng, *et al.*2006). With increased computing memory and processing speed, sophisticated analyses of missing data can now be accomplished by researchers without costly specialized software programs.
However, many researchers are unaware of the importance of reporting and managing missing data, and editors have generally not insisted that authors provide this essential information. Best practices relating to missing data in research call for two items of essential information that should be reported in every research study according to Baraldi and Enders (2010) as the extent and nature of missing data and the procedures used to manage the missing data, including the rationale for using the method selected.

Missing data problems nearly all standard statistical methods presume complete information for all variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, statistical inference, the precision of confidence interval is harmed, statistical power is weakened and the parameter estimates may be biased, according to Ogbeide and Osemwenkhae (2014). Appropriately dealing with missing data and measurement error can be challenging as they require a careful examination of the data to identify the type and pattern of missing data and also a clear understanding of how the different imputation methods work, sooner or later all researchers carrying out empirical research will have to decide how to treat missing data and measurement error. In a survey respondents may be unwilling to reveal some private information, a question may be inapplicable or the study participant simply may have forgotten to answer it. The reason why missing data can be a problem is that it reduces sample size and it is potential for bias, Ogbeide (2020).

Little and Rubin (2002) came up with the classification system that is in use today; Missing Completely At Random (MCAR*),* Missing At Random (MAR), and Missing Not At Random (MNAR). These mechanisms describe relationship between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, they are essentially three different explanations that dictate the performance of different missing data techniques. We give a conceptual description of each mechanism in this section and supplementary resources are available (Schafer and Graham, 2002; Baraldi and Enders, 2010).
Data are MCAR when the probability of missing data on a variable X is unrelated to other measured variables and to the values of X itself. In other words, missing data is completely unsystematic and the observed data can be thought of as a random subsample of the hypothetically complete data. As an example, consider a child in an educational study that moves to another district midway through the study. The missing values are

MCAR if the reason for the move is unrelated to other variables in the data set (e.g. socioeconomic status, disciplinary problem, or other study related variables). Other examples of MCAR occur when a participant misses a survey administration due to computer randomly misreads grids-in-sheets, or an administrative blunder causes several test of the entire sample but limits time-consuming behavioral observations to a random subset of participants. We describe a number of these so-called planned missing data designs at the end of the paper, because MCAR requires missingness to be unrelated to study variables, methodologist often argue that it is very strict assumption that is unlikely to be satisfied in practice, (Raghunathan, 2004).

The MAR mechanism requires a less stringent assumption about the reason for missing data. Data are MAR if missing data is related to other measured variables in the model of the incomplete variable (i.e. the hypothetical values that would have resulted had the data been completed). This terminology is often confusing and misleading because of the use of the word "random". In fact, MAR mechanism is not random at all and describes systematic missingness where the propensity for missing data is correlated with other study-related variables in an analysis. As an example of MAR mechanism, consider a study that is interested in assessing the relationship between substance use and self esteem measure (e.g. because students tend to be absent on the days that the researcher administered the self esteem questionnaires). This example qualifies as MAR if the propensity for missing data on the self esteem measure is completely determined by a student's substance use score (i.e. there is no residual relationship between the probability of missing data and self-esteem after controlling for substance use). As a second example, suppose that a school district administers a math aptitude exam, and student that score above a certain cut-off mark participate in an advanced math course.

Finally, data are MNAR if the probability of missing data is systematically related to the hypothetical values that are missing. In other words, the MNAR mechanism describe data that are missing based on the would-be values of missing scores. For example, consider a reading test where poor readers fail to respond to certain test items because they do not understand the accompanying vignette. Notice that the probability of a missing reading score is directly related that to reading ability. As another example, consider a self report alcohol assessment administered to high school students. MNAR data would result if heavy drinkers are more likely to skip questions out of fear of getting into trouble.

It is only possible to empirically test the MCAR mechanism. Methodologists have proposed a number of MCAR tests, although these procedures tend to have low power and do a poor job of detecting deviations from a purely random mechanism (Thoemmes and Enders 2007). In contrast, the MAR and MNAR mechanisms are impossible to verify because they depend on the unobserved data. That is, demonstrating a relationship between the probability of missing data and the would-be values of the incomplete variables that require knowledge of the missing values. Consequently, the MAR mechanism that underlies maximum likelihood and multiple imputations is ultimately an un-testable assumption. Some authors argue that the MAR assumption may be tenable in school based studies where student mobility is often the most common reason for attrition, although there is ultimately no way to verify this contention. Although MNAR mechanism is certainly possible in educational studies, this type of missing data is probably more common in clinical trials where attrition is a result of declining health or death.

Finally, it is important to point out that the data mechanisms are not characteristics of an entire data set, but they are assumptions that apply to specific analysis. Consequently, the same data set may produce analyses that are MCAR, MAR, or MNAR depending on the variables that are included in the analysis. Returning to the previous math aptitude example, suppose that the parameter of interest is the mean course grade. Recall that the missing course grades are related to math aptitude, such that students with low aptitude scores have missing data. Even though aptitude is not of substantive interest, the MAR, assumptions are only satisfied if this variable is part of the statistical analysis. Excluding the aptitude scores would likely bias the resulting mean estimate because the analysis is consistent with MNAR mechanism (in effect, omitting the "cause" of missing data induces a spurious correlation between the probability of missing data and course grades).

Although the magnitude of the bias depends on the correlation between the omitted aptitude variable and the course grades (bias increases as the correlation increases), the analysis is nevertheless consistent with MNAR mechanism. Later in the manuscript, we describe methods for incorporating so-called auxiliary variables that are related to missing data into a statistical analysis. Doing so can mitigate bias (i.e. by making the MAR mechanism more plausible) and can improve power (i.e. by recapturing some of the missing information).The purpose of this study was to investigate the impact of missing data in predictor variables used in iterative models with continuous outcome variables.
Missing data can be on the independent variables in which we classified observations into groups on the basis of self report, and then try to compare those groups on some independent variables. Also missing data can be on the dependent variable in which we have design that reduces to a one-way analysis of variance or test and the treatment of it is usually straight forward if only we can assume the data are at least MAR.

1.2 Justification for the study.
When missing data are encountered in reported data, the statistical inference from such data set will be affected, particularly when making prediction from the data. So there is need for data cleaning or corrections. A standard method of correcting missing data when they occur and are non-ignorable (essentially when required for update), according to Little and Rubin (2002) is imputation. This is necessary to correct abnormality in the data. So, this study is set out to apply iterative approaches. The study is a research and application of imputation approaches to Nigeria health service data. This work will help further researchers especially first-timers to understand how missing data pose problems and how they can be handled. It will also encourage further work on missing data by future researchers.
Missing data is almost everywhere in medical research, in both observational studies and randomized trials. Until the advent of sufficiently powerful computers, much of the research in this area was focused on the problem of how to handle, in a practicable way, the lack of balance caused by incompleteness. As routine computation became less of a problem, attention moved to the much more subtle issue of the consequences of missing data on the validity of subsequent analyses. Little and Rubin (2002) came up with the classification system that is in use today; Missing Completely At Random (MCAR*),*Missing At Random (MAR), and Missing Not At Random (MNAR). These mechanisms describe relationship between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, they are essentially three different explanations that dictate the performance of different missing data techniques.

Missing data problems nearly all standard statistical methods presume complete information for all variables included in the analysis. A relatively few absent observations on some variables can reduce the sample size. As a result, statistical inference is affected, the precision of confidence interval is harmed, statistical power is weakened and the parameter estimates may be biased, according to Ogbeide (2018). Accordingly, dealing with missing data and measurement error can be very challenging as they require a careful examination of the data to identify the type and pattern of missingness and also a clear understanding of how the different imputation methods work, sooner or later all researchers carrying out empirical research will have to decide how to treat missing data and measurement error. The basic methods for handling missing data are the use of non iterative and the iterative approaches.

Unfortunately, there are many different methods of handling missing data that dwell more on the non iterative approaches mainly due to its simplicity. According to Little and Rubin (2002) , non iterative approaches suffer model specifications and non convergence assessments. This according to Little and Rubin (2002) has led to the need for computational approaches that has model specifications and convergence assessment efficiencies. For this reason, according to Ogbeide (2020), it is important to select the correct missing data handling method that is suitable to a researcher's particular circumstances. These circumstances

can be expressed as factors, such as sample size, proportion of missing data, method of analysis etc. Some which may fall under the control of the researcher in a given scenario and thus can be manipulated, while others are more difficult to control. However, many researchers are unaware of the importance of reporting and managing missing data, and editors have generally not insisted that authors provide this essential information. Best practices related to missing data in research call for two items of essential information that should be reported in every research study according to Baraldi and Enders (2010) as the extent and nature of missing data and the procedures used to manage the missing data, including the rationale for using the method selected. Accordingly, the purpose of this work is to clearly present some Iterative approaches necessary to successfully deal with missing data when they occur in a given situation. We shall also compare the different approaches and choose the best of the approaches in terms of computation and analysis of missing data using Nigeria health service data.

So, some outlined approaches for handling missing data according to Carpenter and Kenward, (2008) and Ogbeide **(**2020) are Least squared (LS) method, Expectation Maximization (EM) method and Multiple Imputations (MI) method. These approaches will be used in the study when missing data occur in the Nigeria health service data. It is therefore pertinent to use iterative imputation methods in solving missing data problems as it can be implemented on a number of simple and more complex data set for utmost reliability, stability, efficiency and consistency, Demirtas, Freels and Yucel (2008).The iterative imputation methods can generate health service data in Nigeria in terms of computation and analysis of missing data.
The overall aim of this research is to examine some iterative approaches to obtaining missing observations in Nigeria health service data. The specific objectives are to: apply some iterative imputation approaches to a Nigeria health service data. Compare the results from the different approaches from the data and provide general inference about the missing data handling performance approaches with application to health service data in Nigeria.

Literature Review

We review some of the work done by previous researchers in this field which are fundamental for the understanding of this current research work in this chapter. The problem of missing data is almost ubiquitous in medical research, both in observational studies and randomized trials. Until the advent of sufficiently powerful computers, much of the research in this area was focused on the problem of how to handle, in a practicable way, the lack of balance caused by incompleteness. As routine computation became less of a problem, attention moved to the much more subtle issue of the consequences of missing data on the validity of subsequent analyses. An example of such a development was the key idea of the Expectation Maximization (EM) approach according to Dempster, A.P.; Laird, N.M. and Rubin, D.B. (2007). Multiple Imputations is another approach (Acock,2005).

Missing data are problems nearly all standard statistical methods presume complete information for all variables included in the analysis. A relatively few absent observations on some variables can dramatically shrink the sample size. As a result, statistical inference is affected, the precision of confidence interval is harmed, statistical power is weakened and the parameter estimates may be biased according to Ogbeide (2018). Appropriately dealing with missing data and measurement error can be challenging as they require a careful examination of the data to identify the type and pattern of missing data and also a clear understanding of how the different imputation methods work, sooner or later all researchers carrying out empirical research will have to decide how to treat missing data and measurement error. In a survey respondents may be unwilling to reveal some private information, a question may be inapplicable or the study participant simply may have forgotten to answer it. The reason why missing data can be a problem is that they reduce efficiency from reduced sample size and it is a potential for bias, Little and Rubin (2002) and Ogbeide (2020).

Missing Data are problems that often occur in empirical studies too. Imagine that propositi-test a new medicine that should lower the blood pressure and have to visit the doctor several times for the duration of the study. It is possible that for some proposition measurements of the blood pressure or other covariates are missing. In which way one can deal with missing data depends strongly on the missing-process. It is possible that people just forget a visit or that their blood pressure is so high that they cannot keep the appointment. The different types of missing data will be described in the following part. In the second part of this section we will outline some existing methods to handle missing data.

Missing observation analysis was popularized by Rubin in the seventies. (Little and Rubin 2002) came up with the classification system that is in use today; Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). These mechanisms describe relationship between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, they are essentially three different explanations that dictate the performance of different missing data techniques. We give a conceptual description of each mechanism in this section and supplementary resources are available (Schafer and Graham, 2002; Baraldi and Enders, 2010).

Data are MCAR when the probability of missing data on a variable X is unrelated to other measured variables and to the values of X itself. In other words, missing data is completely unsystematic and the observed data can be thought of as a random subsample of the hypothetically complete data. As an example, consider a child in an educational study that moves to another district midway through the study. The missing values are MCAR if the reason for the move is unrelated to other variables in the data set (e.g. socioeconomic status, disciplinary problem, or other study related variables). Other examples of MCAR occur when a participant misses a survey administration due to computer randomly misreads grids-in-sheets, or an administrative blunder causes several test of the entire sample but limits time-consuming behavioral observations to a random subset of participants. We present some iterative imputation approaches in this work. The purpose of this work is to clearly present these iterative imputation approaches necessary to successfully deal with missing data when they occur in a given situation. We shall also compare the different approaches and choose the more efficient of the approaches in terms of computation and analysis of missing data using Nigerian health service data. So, some outlined approaches for handling missing data according to Carpenter and Kenward, (2008) and Ogbeide **(**2018) are Least squared method, Expectation Maximization (EM) method and Multiple Imputations method. These approaches will be used in the study when missing data occur in health services in Nigeria.

Research Methods

The research data that is used in this study is a secondary data on Nigeria Health Service data. Data from the National Health and Nutrition Examination Survey (NHANES) served as the basis for this study. The data was made available in 2011-2012. The NHANES is a program of studies designed to assess the health and nutritional status of adults and children in Nigeria. The survey is unique in that it combines interviews and physical examinations. The NHANES dataset has 25 observations, each having 4 features - age, body mass index, hypertension and cholesterol level. The age values are only 1, 2 and 3 which indicate the age bands 20-39, 40-59 and 60+ respectively. The short description of the data set is given below:

- **Age:** Age in years at screening of study participant.
- **BMI: Body mass index:** Reported for participants aged 2 years or older.
- **Cholesterol level:** Total Cholesterol of participants.
- **Hypertension status:** Blood Pressure Level of the participants.

The following imputation approaches would be considered -Least Squared (LS) method, Expectation Maximization (EM) and Multiple Imputations (MI) approaches. Imputation was completed on the outcome and all variables. Where applicable, the outcome and all variables were used as a predictor for the variable being imputed.

Performance evaluations of all the approaches would be done using Raw Bias (RB), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). All statistical procedures would be carried out in R-statistical software.

3.1.1 Least squared method.

In a randomized block least squared approach uses information row and column to decide missing values. This approach uses the treatment of data set design for finding missing values. Allen and Wishart (2008) provided estimator for the least squared estimates of one missing value if a randomized block design and of one missing value in a Latin square design. In a randomized block with T treatments and B blocks, the least square estimates of the missing values in the treatment $t$and block b is:

$$Y_{ij} = \frac{Ty^{(t)} + By^{(b)} - Y_{+}}{(T-1)(B-1)} \quad 1$$

where, $y_{+}^{(t)}$ and $y_{+}^{(b)}$ = sum of the observed values of Y for treatment t and block b, respectively and Y_{+}= sum of all observed values of Y. T = Treatment size. B = Block size.

N.B: The statistical model of the missing observation (incomplete data) is

$$Y_{ij} = \mu + T_i + B_j + \varepsilon_{ij} \text{ and } \varepsilon_i \overset{iid}{\sim} N(0,\delta^2) \quad 2$$

Where iid = independently and identically distributed.

Here, equation (1) and equation (2) are the estimators for the missing data from the given available data set.

3.1.2 Expectation Maximization (EM) method.

The Expectation Maximization (EM) approach formalizes a relatively old ad hoc idea for handling missing data: (1) Replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct,

(4) re-estimate parameters, and so forth, iterating until convergence. Such methods are EM algorithms for models where the complete data loglikelihood is computed. Given X_{ij} observed data with $X_{ij} = X_{obs} + X_{mis}$ the loglikelihood is given by $\ell(\theta|X_{obs,}\, X_{mis})$.

where, X_{obs} is the observed values, X_{mis} the missing values and θ unknown parameter.

Generally, the loglikelihood $\ell(\theta\,|X)$ itself needs to be estimated at each iteration of the algorithm after estimation of the expectation of the complete available data.

The E step finds the conditional expectation of the ‘‘missing data’’ given the observed data and current estimated parameters, and then substitutes these expectations for the ‘‘missing data.’’ Specifically, let $\theta^{(r)}$ be the current estimate of the parameter θ unknown from X_{ij}. The E step of Expectation Maximization finds the expected complete-data loglikelihood of θ. According to Little and Rubin (2002), the expectation is given by:

$$E\left(\ell(\theta|X_{obs,}\, X_{mis})\right) = E(\, X_{obs} + (n-r)\mu^{(r)}) \quad 3$$

The value estimated from this step is substituted into the original data set with missing observation using equation (3).

The M step is particularly simple to describe Maximum Likelihood (ML) of the "E" step imputation. It performs ML estimation of θ just as if there were no missing data, that is, as if they had been filled in. Thus the M step of EM uses the identical computational method as ML estimation from $\ell(\theta|X)$.

The M step of Expectation Maximizations determines $\theta^{(r+1)}$ by maximizing this expected complete-data log likelihood of the data set.

$$Q\left(\theta^{(r+1)}\middle|\theta^{r}\right) \geq Q(\theta|\theta^{r}), \text{ for all } \theta. \qquad 4$$

where, $Q\left(\theta^{(r+1)}\middle|\theta^{r}\right)$ is the partial derivative of the loglikelihood of X_{ij} containing the unknown θ in the current estimated sequence r. The equation (4) is computed repeatedly until convergence is achieved.

3.1.2 Multiple Imputations (MI).

Multiple Imputation (MI) is an important and influential approach in the analysis of missing data under the MAR and multivariate normality assumptions. Although proposed in the context of non response in sample survey, the technique is quite general and can be readily used in other settings as well. The merits of MI and recent software developments have been discussed elsewhere (Harel and Zhou, 2007; Carpenter and Kenward, 2013). Generally, MI works as follows: every missing datum is replaced by two or more imputed values in order to reflect uncertainty about which value to impute and then each completed data set is analyzed separately by standard statistical tools just as if the imputed data were real. By imputing missing data m times, it introduces statistical uncertainty into the model and uses the uncertainty to emulate the sample variability of a complete data set.

Estimates of parameters are combined by using Rubin's rule (the three characteristic systems) to make the final inferences about the data. The validity of MI depends on the imputed values. If an inappropriate model for the imputation process is used, it will result in biased and inconsistent conclusions (Collins, Schafer and Kam (2001). A full understanding of the methodology to obtain imputed values is important to obtain unbiased estimates with correct confidence intervals. Failure to consider all relevant aspects of creating imputation models can impact the validity of inferences. For example, if our scientific interest focuses on the correlation between two variables, then both variables must be present in the imputation process even if only one of them has missing values. If we mistakenly remove the other variable from the imputation model, then inferences derived from MI will be biased.

In general, the imputation model should not impose unnecessary restrictions on the variables that will be the subject of the later analysis. An imputation model should also preserve the relations among the variables in the post-imputation analysis. As a general guideline, imputation methods should be general enough to encompass the intended analyses. To produce high quality imputed values for a particular incomplete variable, the imputation model should include variables that are (i) potentially related to the incomplete variable and (ii) potentially related to the missingness of the incomplete variable. It should be noted that a particular variable in this case may have different roles as it is part of the complete data model or the non response model or both.

3.1.4 MI General Procedure

MI was first proposed by Rubin (2004) as a method for handling missing data in survey. It seems to be one of the most promising methods for general-purpose handling of missing data in multivariate analysis; this is because MI has several desirable features:

1. the fact that MI introduces appropriate random error into the imputation process makes it imputations are sufficient to obtain excellent results (Enders, 2010).possible to get approximately unbiased estimates for all parameters. No deterministic imputation method can do this in general settings (Allison, 2001).
2. MI can be used with any kind of data and any kind of analysis without specialized software. Repeated imputation allows one to get good estimates of the standard errors (Allison, 2001).
3. MI can be highly efficient even for small values of m. MI imputations are sufficient to obtain excellent results using (Enders, 2010) as follows;.

Imputing data: The most challenging step, which imputes the missing values m> 1 times. Imputed values are drawn from the distribution of incomplete data set, and each drawn value will be different for the missing item. The results of this step are m complete data sets. There are a variety of imputation models that can be used. The choice of imputation models depends on the assumptions regarding the missingness mechanisms and patterns, as well as the data distribution. Next, **Imputation analysis:** The repeated analysis step on the imputed data. This step analyses each of the m complete data sets by using a standard complete data method, which should be compatible with the imputation model used by the imputation step (Alzola and Harrell 2006). The results of this step are m analysis steps for the final results to emerge. This step integrates the m analysis results into a final result for the inference. It consists of computing the mean over the m repeated analysis, the corresponding variance and confidence interval or p value. Simple rules exist for combining the m analysis results.

Results and Discussions

This section deals with the application of the approaches of Least Squared, Expectation Maximization and Multiple Imputations for the data from health service data in Nigeria based on National Health and Nutrition Examination Survey (NHANES). The approaches error estimates will also be presented in this section.

Table 4.1: Original data of missing values analysis

		Original Data		
	Age	**Bmi**	**Hyp**	**Chl**
1	1	NA	NA	NA
2	2	22.7	1	187
3	1	NA	1	187
4	3	NA	NA	NA
5	1	20.4	1	113
6	3	NA	NA	184
7	1	22.5	1	118
8	1	30.1	1	187
9	2	22	1	238
10	2	NA	NA	NA
11	1	NA	NA	NA

12	2	NA	NA	NA
13	3	21.7	1	206
14	2	28.7	2	204
15	1	29.6	1	NA
16	1	NA	NA	NA
17	3	27.2	2	284
18	2	26.3	2	199
19	1	35.3	1	218
20	3	25.5	2	NA
21	1	NA	NA	NA
22	1	33.2	1	229
23	1	27.5	1	131
24	3	24.9	1	NA

Source: National Health and Nutrition Examination Survey (NHANES), 2011-2012.

Ages 1, 2, and 3 represent the age interval of 20-39, 40-59 and 60+ respectively. The serial Number 1, 2 --- 24 represent the number of months from January 2011 as 1, February 2011 as 2 until 24 respectively. Bmi implies body mass index, Hyp implies hypertension status (blood Pressure level of the participants) and Chl represents cholesterol level.

Table 4.2: Missing values analysis using least squared method.

		Original			**Least Square Imputation**		
	Age	**Bmi**	**Hyp**	**Chl**	**Bmi**	**hyp**	**Chl**
1	1	NA	NA	NA	**28.981**	**1.073**	**175.666**
2	2	22.7	1	187	22.700	1.000	187.000
3	1	NA	1	187	**29.177**	1.000	187.000
4	3	NA	NA	NA	**21.147**	**1.390**	**214.968**
5	1	20.4	1	113	20.400	1.000	113.000
6	3	NA	NA	184	**20.442**	**1.375**	184.000
7	1	22.5	1	118	22.500	1.000	118.000
8	1	30.1	1	187	30.100	1.000	187.000
9	2	22	1	238	22.000	1.000	238.000
10	2	NA	NA	NA	**26.934**	**1.400**	**208.769**
11	1	NA	NA	NA	**29.177**	**1.083**	**177.082**
12	2	NA	NA	NA	**26.019**	**1.292**	**202.162**
13	3	21.7	1	206	21.700	1.000	206.000
14	2	28.7	2	204	28.700	2.000	204.000
15	1	29.6	1	NA	29.600	1.000	**180.612**
16	1	NA	NA	NA	**31.203**	**1.188**	**191.713**
17	3	27.2	2	284	27.200	2.000	284.000

18	2	26.3	2	199	26.300	2.000	199.000
19	1	35.3	1	218	35.300	1.000	218.000
20	3	25.5	2	NA	25.500	2.000	**244.662**
21	1	NA	NA	NA	**29.177**	**1.083**	**177.082**
22	1	33.2	1	229	33.200	1.000	229.000
23	1	27.5	1	131	27.500	1.000	131.000
24	3	24.9	1	NA	24.900	1.000	**244.733**

Imputation Table 4.2 shows the variable characteristics of the original database compared to the least squared method of dealing with missing data

4.1.2 Expectation Maximization.

Expectation–Maximization (EM) algorithm is an iterative method to find maximum likelihood or Maximum A posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved variables. This technique iteratively goes through the data while still preserving the covariance structure of the data. Expectation Maximization imputation was implemented through the R package mvdalab by using the function impute EM.

Table 4.3: Missing values analysis using Expectation Maximization approach.

		Original			**Expectation Maximum**		
	Age	Bmi	Hyp	Chl	Bmi	Hyp	Chl
1	1	NA	NA	NA	**29.478**	**0.964**	**173.659**
2	2	22.7	1	187	22.700	1.000	187.000
3	1	NA	1	187	30.867	1.000	187.000
4	3	NA	NA	NA	**21.718**	**1.633**	**223.740**
5	1	20.4	1	113	20.400	1.000	113.000
6	3	NA	NA	184	**17.529**	**1.413**	184.000
7	1	22.5	1	118	22.500	1.000	118.000
8	1	30.1	1	187	30.100	1.000	187.000
9	2	22	1	238	22.000	1.000	238.000
10	2	NA	NA	NA	**25.598**	**1.298**	**198.699**
11	1	NA	NA	NA	**29.478**	**0.964**	**173.659**
12	2	NA	NA	NA	**25.598**	**1.298**	**198.699**
13	3	21.7	1	206	21.700	1.000	206.000
14	2	28.7	2	204	28.700	2.000	204.000
15	1	29.6	1	NA	29.600	1.000	**177.218**
16	1	NA	NA	NA	**29.478**	**0.964**	**173.659**
17	3	27.2	2	284	27.200	2.000	284.000
18	2	26.3	2	199	26.300	2.000	199.000
19	1	35.3	1	218	35.300	1.000	218.000

20	3	25.5	2	NA	25.500	2.000	**272.573**
21	1	NA	NA	NA	**29.478**	**0.964**	**173.659**
22	1	33.2	1	229	33.200	1.000	229.000
23	1	27.5	1	131	27.500	1.000	131.000
24	3	24.9	1	NA	24.900	1.000	**190.152**

Imputation Table 4.3 shows the variable characteristics of the original database compared to expectation maximization of dealing with missing data

4.1.3 Multiple Imputations

The mice package in R was used to impute MAR values only. As the name suggests mice uses multivariate imputations to estimate the missing values. The package uses PMM (Predictive Mean Matching) method to impute continuous variables. If any variable contains missing values, the package regresses over the other variables and predicts the missing values.

Table 4.4: Missing values analysis using Multiple Imputations approach.

		Original			**Multiple Imputation (MI)**		
	Age	Bmi	Hyp	Chl	Bmi	Hyp	Chl
1	1	NA	NA	NA	35.3	1	218
2	2	22.7	1	187	22.7	1	187
3	1	NA	1	187	33.2	1	187
4	3	NA	NA	NA	22.7	2	229
5	1	20.4	1	113	20.4	1	113
6	3	NA	NA	184	25.5	2	184
7	1	22.5	1	118	22.5	1	118
8	1	30.1	1	187	30.1	1	187
9	2	22	1	238	22	1	238
10	2	NA	NA	NA	30.1	1	184
11	1	NA	NA	NA	22	1	118
12	2	NA	NA	NA	27.5	1	204
13	3	21.7	1	206	21.7	1	206
14	2	28.7	2	204	28.7	2	204
15	1	29.6	1	NA	29.6	1	238
16	1	NA	NA	NA	30.1	1	187
17	3	27.2	2	284	27.2	2	284
18	2	26.3	2	199	26.3	2	199
19	1	35.3	1	218	35.3	1	218
20	3	25.5	2	NA	25.5	2	218
21	1	NA	NA	NA	35.3	1	229
22	1	33.2	1	229	33.2	1	229
23	1	27.5	1	131	27.5	1	131
24	3	24.9	1	NA	24.9	1	186

Imputation Table 4.4 shows the variable characteristics of the original database compared to multiple imputations of dealing with missing data.

4.1.4 Evaluation Criteria

The goal of multiple imputations is to obtain statistically valid inferences from incomplete data. The quality of the imputation method should thus be evaluated with respect to this goal. There are several measures that may inform us about the statistical validity of a particular procedure. These are:

1. **Raw Bias (RB):** The raw bias of the estimate $\bar{Q}$ is defined as the difference between the expected value of the estimate and truth: $RB = E\left(\bar{Q}\right) - Q.$ RB should be close to zero. Bias can also be expressed as percent bias: $PB = 100 \times \left| \left(E\left(\bar{Q}\right) - Q \right) / Q \right|$. For acceptable performance we use an upper limit for PB of 5%. (Demirtas, Freels, and Yucel, 2008)
2. **Mean Squared Error (MSE)*:*** The mean squared error indicates how close the predicted values are to the actual values; hence a lower MSE value signifies that the model performance is good. One of the key properties of MSE is that the unit will be the same as the target. $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \left(E\left(\bar{Q}\right) - Q\right)^2$
3. **Root Mean Squared Error (RMSE)*:*** The RMSE is the square root of MSE. $RMSE = \sqrt{\left(E\left(\bar{Q}\right) - Q\right)^2}$. RMSE is a compromise between bias and variance, and evaluates $\bar{Q}$ on both accuracy and precision.

Table 4.5: Error estimates from the various approaches.

	LS	EM	MI
RB	0.758954	0.42667	0.453813
MSE	0.6453678	0.35637	0.284353
RMSE	0.8033478	0.59696	0.533246
Variance	0.6453678	0.35637	0.284353

Keys: LS–Least Squared, EM–Expectation Maximization, MI–Multiple Imputation.

This iterative imputation approaches values for the Least Square (Yates method) Regression, Expectation Maximization (EM) and Multiple Imputations (MI) in obtaining missing data in health services in Nigeria. Imputation was completed on the available data. The performances of the approaches were evaluated using Raw Bias (RB), Mean Square Error (MSE) and Root Mean Square Error (RMSE). All statistical procedures were carried out in R-statistical software.

Table 4.6: Missing values variables analysis

Variable	**Missing (%)**
Age	0%
BMI	36%
Hyp	32%
Chl	40%

The study sample consisted of 3 variables with 24 cases. BMI had 9 (30.4%) of missing data, while hyp had 32% of missing data. Chl had the highest proportion of missing data accounting for 40%. Overall, least squared imputation resulted in the lowest model performances. A single imputation using the expectation maximization algorithm provides unbiased parameter estimates and improves statistical power of analyses (Enders, 2001; Schaffer and Graham, 2002). Multiple Imputations only had the lowest MSE of value 0.284353, which happens to be the best performance approach. This was closely followed by EM algorithm with MSE value of 0.35637 and the Least squared has 0.6453678 from table 4.5 above.

The RB for LS is 0.758954, the EM has 0.42667 and MI is 0.453813. The RMSE for LS is 0.8033478, EM has 0.59696 while MI has 0.533246.

These error estimates indicate the improvement using Multiple Imputations for the data at hand.

Conclusion

Statistical estimations are affected by missing data. A good imputation approach is one way of reducing consequences. This studies uses the iterative approaches due to statistical advantages. The results of this study revealed MI has the lowest MSE value than the MSE computed from LS and EM algorithms. These results were consistent with absolute bias (AB) of the MI method which was also lower than the variance computed for other methods. Hence, based on the approaches performance, we concluded that a careful judgment be made between MI and EM algorithms when imputing the NHANES dataset. Other researchers are advised to use these approaches to explore for missing data as used to get more efficient missing observations in health services and in other fields.

References

1. *Allen, F.G. and Wishart, J. (2008). A method of estimating the yield of a missing plot field experiments, J. Agric. Sci. 20: 399-406.*
2. *Allison, P.D. (2001). Missing data-Quantitative applications in the social sciences. Thousand Oaks. CA: Sage.Vol.136.*
3. *Alosh, M. (2009). The impact of missing data in a generalized integer value auto regressing model for count data. Journal of Biopharmaceutical statistics 19: 1039-1054.*
4. *Baraldi, A. N. and Enders, C.K. (2010). An introduction to modern missing data analysis. Journal of school psychology 48: (1), 5-37*
5. *Carpenter, J.R. and Kenward, M.G. (2008). Missing data in clinical trials – a practical guide. National Health Service Coordinating Centre for Research Methodology: Birmingham, UK.*
6. *Collins, L.M.; Schafer, J.L. and Kam, C.M.(2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Meth 6(4):330–351.*
7. *Demirtas, H.; Freels, S. and Yucel, R. (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes.A simulation assessment. Journal of Statistical Computation and Simulation 78: 69-84.*
8. *Dempster, A.P. Laird N.M. and Rubin, D.B. (2007): Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1-38.*
9. *Donders ART; van der Heijden, G.J; Stijnen, T and Moons, K.G. (2006). Review: a gentle introduction to imputation of missing values. J. of Clin. Epidemiol. 59(10):1087–91.*
10. *Grooves, R. and Fowler, F. (2004). Survey Methodology, New York: Wiley and sons Inc. Enders, C. K. (2010). Applied missing data analysis. New York: Guilford Press.*
11. *Harel, O. and Zhou, X. H. (2007).Multiple imputation review of theory of implementation and software. Statistics in Medicine (in press).*

12. *Ogbeide, E.M. (2018): A new imputation based on Expectation Maximization of the data set. Science and Technology journal. 3(1);139-142.*
13. *Ogbeide, E.M. (2020): Imputation statistics value for missing data in Nigeria crime rate reported. Journal on Demography and Social Statistics. 7(1); 9-16.*
14. *Ogbeide, E.M. and Osemwenkhae, J.E. (2014). A new Imputation Method based on Expectation Maximization of the dataset. Proceeding of the Statistics Research Group (SRG), National conference, university of Benin, Nigeria.*
15. *Peng, C.Y. J.; Harwell, M.; Liou, S.M. and Ehman, L.H. (2006). Advances in missing data methods and implications for educational research in Sawilowsky ed. Greewhich, CT: Information Age. Real Data Analysis, (pp.31-78).*
16. *Pigot, T.D. (2001). A Review of Methods for Missing data. Educational Research and Evaluation 7:353-383.*
17. *Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. Annual review of public health 25: 88-117.*
18. *Rubin, D.B. (2004) 'Basic ideas of multiple imputation on non-response', Survey Methodology, Vol. 12, No. 1, pp.37–47.*
19. *Schafer, J. L. and Graham, J.W. (2002). "Missing Data: our view of state of the state of the Art". psychological methods 2: 147-149.*
20. *Streiner, D.L. (2002). The case of the missing data: methods of dealing with dropouts and other research vagaries. Canadian Journal of Psychiatry 47: 68-74.*
21. *Song, Q., Shepherd, M., Cartwright, M., (2005). A short note on safest default missingness mechanism assumptions. Empirical Software Engineering: An International Journal 10 (2), 235–243.*
22. *Thoemmes, F. and Enders, C.K. (2007). A structural equation model for testing whether data are missing completely at random. Paper presented at the annual meeting of the American Educational research Association, Chicago.*