بسم الله الرحمن الرحيم

Sudan University of Science and Technology

College of Graduate Study

A Thesis Submitted in fulfillment of Requirement for the Ph.D. Degree in Statistics

Title:

Multiple Imputation, Regression Imputation and Expectation Maximization Methods for Handling Missing Data

طرق الإدخال المتعدد، إدخال الإنحدار وتعظيم التوقع لتقدير البيانات المفقودة

Presented by:

Montasir Abbas Ahmed Mohammed

Supervisor:

Dr. Adil M. Y. Waniss
Associate Professor of Statistics

Co-Supervisor:

Khalid R. K. Genawi
Associate Professor of Statistics

July 2019

الآية الكريمة

﴿ اللَّهُ نُورُ السَّمَاوَاتِ وَالْأَرْضِ مَثَلُ نُورِهِ كَمِشْكَاةٍ فِيهَا مِصْبَاحٌ الْمِصْبَاحُ فِي زُجَاجَةٍ الزُّجَاجَةُ كَأَنَّهَا كَوْكَبٌ دُرِّيٌّ يُوقَدُ مِنْ شَجَرَةٍ مُبَارَكَةٍ زَيْتُونَةٍ لَا شَرْقِيَّةٍ وَلَا غَرْبِيَّةٍ يَكَادُ زَيْتُهَا يُضِيءُ وَلَوْ لَمْ تَمْسَسْهُ نَارٌ نُورٌ عَلَى نُورٍ يَهْدِي اللَّهُ لِنُورِهِ مَنْ يَشَاءُ وَيَضْرِبُ اللَّهُ الْأَمْثَالَ لِلنَّاسِ وَاللَّهُ بِكُلِّ شَيْءٍ عَلِيمٌ ﴾

سورة النور–آية 35

DEDICATION

”Dedicated to my beloved parents & family” for their endless love, support and encouragement. And to the Spirit of my brother, Dr. Magdi Mahmoud, Allah’s Mercy be upon him.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to all those who helped me to make my thesis work a success. First and foremost I would like to thank ALLAH for he has enabled and guided me to finally finish this research.

I express my sincere and whole hearted thanks, to my supervisor, esteemed mentor Dr. Adil M. Y. Waniss, for his regular advice, magnitude of dynamic and unceasing guidance, suggestion and encouragement throughout the course of present research.

I express my heart full thanks to Dr. Khalid R. K. Genawi, Dr. Mohammedelameen Qurashi and Dr. Hesham Reyad for valuable guidance, technical and moral support given to me throughout the entire course of dissertation work

There are no words to express my gratitude and thanks to My Beloved parents, family members and friends for always standing by me. Their love has been the major spiritual support in my life.

As a final word, I would like to thank each and every individual who have been a source of support and encouragement and helped me to achieve my goal and complete my dissertation work successfully.

ABSTRACT

Missing data are widespread, and pose problems for many statistical procedures. We all should be using methods that treat missing data properly, rather than deleting data or using single imputation. Importantly, researcher should pay attention by using most appropriate analysis of his data, in order to arrive to conclusions that have more accurate parameters. To achieve this objective, an appropriate method of handling treating missing data must be chosen before starting the analysis.

This research aims to a comparative study to the Multiple Imputation (MI) method of estimation against two other methods; the Regression Imputation of estimation and the Expectation-maximization (EM) algorithm of estimation, for estimating missing data.

The study is based in application on data randomly generated, some of them were missed by different percentages (5%, 10%, 15%, 20% and 30%). It also uses SPSS Program as statistical package to help in estimating and analyzing the data.

Data randomized was tested using little's test on which this data was divided into missing completely at random and missing not completely at random. The study proved that based on descriptive statistics, there is considerable differences between means and variances of estimated missing values, and to test the statistical significance of differences, the study used ANOVA test, and the consequently results proved that there is no significant difference between means.

The study also found that (98%) of the correlations were not significant based on the correlation matrix. finally, the study compared the estimated missing values after calculating the mean absolute error (MAE), based on the results, the study concluded that the Expectation-maximization (EM) method of estimation is better than the other two methods in producing more efficient estimates.

This study recommends to give attention should be paid to the missing data in the design and performance of the studies and in the analysis of the resulting data. And the application of the sophisticated statistical analysis techniques should only be performed after the maximal efforts have been employed to reduce missing data as they can cause bias and lead to invalid conclusions. It also recommends using the Expectation-maximization (EM) method of estimation because its estimates are the most efficient.

المستخلص

البيانات المفقودة منتشرة، وتطرح العديد من المشاكل في العمليات الإحصائية. يجب علينا استخدام طرق لمعالجة البيانات المفقودة على نحو سليم، بدلاً من حذفها او استخدام الادخال الفردي. الأهم من ذلك، ينبغي على الباحث الانتباه عند استخدام التحليل الأكثر ملاءمة لبياناته ، من أجل التوصل إلى بيانات متوافقة لديها معلمات أكثر دقة. و لتحقيق هذا الهدف، يجب اختيار طريقة مناسبة لمعالجة البيانات المفقودة قبل بدء التحليل.

يهدف هذا البحث إلى تقديم دراسة للمقارنة بين طريقة الادخال المتعدد للتقدير مع طريقتين أُخرتين وهما طريقة ادخال الإنحدار للتقدير و طريقة تعظيم التوقع للتقدير، و ذلك لتقدير البيانات المفقودة.

إعتمدت الدراسة في الجانب التطبيقي على بيانات تم توليدها عشوائياً، و فقد جزء منها بنسب مختلفة (5%، 10%،15%،20% و 30%). كما تم استخدم الحزمة SPSS كأداة تحليلية لتقدير البيانات المفقودة وتحليلها.

تم اختبار عشوائية البيانات المفقودة باستخدام اختبار ليتل (little’s) بناءً عليه قسمت هذه البيانات إلى بيانات مفقودة بصورة عشوائية كاملة، و بيانات ليست مفقودة بصورة عشوائية كاملة. وقد اثبتت الدراسة و بناءً على الاحصاءات الوصفية أنه توجد فروق ظاهرية بين متوسطات وتباينات القيم المقدرة باستخدام طرق التقدير الثلاثة، ولمعرفة هذه الفروق إستخدمت الدراسة اختبار تحليل التناين (ANOVA) واثبتت النتائج أنه لا توجد فروق بين هذه المتوسطات. كما توصلت الدراسة إلى ان (98%) من الارتباطات كانت غير معنوية إعتماداً على مصفوفة الارتباطات. كما قامت الدراسة بمقارنة مقدرات طرق التقدير الثلاثة السابقة بعد حساب متوسط الخطأ المطلق (MAE)، واستناداً إلى النتائج، خلُصت الدراسة إلى أن طريقة تعظيم التوقع للتقدير تعتبر أكثر كفاءة من الطريقتين الأخرتين فى تقديم مقدرات أكثر كفاءة.

توصي هذه الدراسة بإيلاء اهتمام أكبر للبيانات المفقودة ذلك عند تصميم و سير عمل الدراسات وفي تحليل البيانات الناتجة. كما يجب تطبيق تقنيات التحليل الإحصائي المتطورة بعد بذل أقصى جهد لتقليل البيانات المفقودة لأنها يمكن أن تسبب التحيز وتؤدي إلى إستنتاجات غير صحيحة. كما توصي بإستخدام طريقة تعظيم التوقع لتقدير البيانات المفقودة، لان مقدراتها هي الاكثر كفاءة.

TABLE OF CONTENTS

Chapter III: Methods for Estimating Missing Data

Chapter IV: Comparative Study of the Methods of Estimating Missing Data

Chapter I

Introduction

1.1 Introduction:

Survey considers as one of the basic methodologies in the descriptive research where interested in the study of the social, economic and other conditions in a particular community, with a view to gather the facts and to draw the necessary conclusions to solve the problems of this society. The survey methods are differ according to the fields followed by the researcher to achieve the work. Surveying has many characteristics and features such as shortening the time, effort and cost. The aim of survey is to get groups of classified data and its interpretation and then generalizing it, with the aim of rationalizing the practical implementation in future.

This research is concerned with methods of estimating missing data in surveys. missing data may be ignored or neglected, which may lead to estimates that have less efficient, and may limit the use of some statistical methods that require no missing data, in this case may occur some bias in results and weak the power of statistical tests and measurements used. Attention has been taken for missing data, their processing and the mechanism of dealing with them increasing in development progresses in statistical programs. To achieve these objectives, an appropriate method must be chosen to deal with data before starting the analysis.

1.2 Problem of the study

It has been demonstrated repeatedly that missing data have large effects on the results of a survey. Moreover, increasing the sample size without targeting nonresponse does nothing to reduce bias in missing data; a larger sample size merely provides more observations from the class of persons that would respond to the survey. Increasing the sample size may actually worsen the nonresponse bias, as the larger sample size may divert resources that could have been used to reduce or remedy the nonresponse, or it may result in less care in the data collection [1]. Most small surveys ignore any nonresponse or missing data that remains after callbacks and follow-ups, and report results based on complete records only. The main problem caused by nonresponse or missing data is potential bias. Some factors affecting non-response; interviewers, data-collection method, questionnaire design, incentives financial or otherwise

1.3 Importance of the study

The importance of this study are as follows:

(1) Describes the pattern of missing data. Where are the missing values located? How extensive are they? Do pairs of variables tend to have values missing in multiple cases? Are data values extreme? Are values missing randomly?

(2) Estimates means, variances, covariances, and correlations for different missing data methods: the regression imputation method or the EM Algorithm method.

(3) Fills in (imputes) missing values with estimated values using MI, regression or EM methods; however, multiple imputation is generally considered to provide more accurate results.

(4) Reduce bias by estimating missing data by an exact statistical scientific methods to include all the categories of the original community of study.

(5) Shortening the time, effort and cost throughout collecting new data.

1.4 Objectives of the study

The aims of this study is to identifying the efficiency of (MI) method, regression method and (EM) methods of estimation of missing data, by comparing their estimators with each other. The main objectives are as follows:

(1) Evaluate and understand missing data.
(2) Explain common missing data methods in surveys, and know their advantages and disadvantages.
(3) Apply the theories related to how to estimate missing data according to the characteristics of the community covered by survey.
(4) Comparison between different methods of estimating missing data.

1.5 Data of the study

The study is based on the applied side of generating random data with mean (1000) and variances (1.04, 26.88 and 83.74). It also uses SPSS Program as analytical tool to estimate and analyze data.

1.6 Hypotheses of the study

This study assumes the following hypotheses:

(1) Is the missing completely at random?

(2) Main hypothesis, not there is a statistically significant difference between generating random data and the estimated parameters by the three methods.

(3) The correlation between generating random data and the estimated parameters by the three methods is significant.

(4) (EM) the more efficient method for estimating missing data.

(5) Regression and (MI) the less efficient methods for estimating missing data.

1.7 Methodology of the study

In this study, a descriptive and analytical approaches are used to determine the efficiency of (MI), regression and (EM) methods; for estimating missing data. Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions. This research reviews the problems and types of missing data, along with the techniques for handling missing data. The mechanisms by which missing data occurs are illustrated, and the methods for handling the missing data are discussed. The research concludes with recommendations for the handling of missing data. It also uses SPSS Program as analytical tool to estimate and analyze data.

1.8 Researches and Previous Studies

This section concerned to illustrate the most researches and studies using the estimation methods of missing data.

(1) In 2016 Mohammad Taghi Sattari, Ali Rezazadeh Joudi and A. Kusiak introduced a study entitled (Assessment of different methods for estimation of missing data in precipitation studies). The study considered various techniques for filling in missing precipitation data. To assess the suitability of the different methods for filling in missing data.

They used the arithmetic averaging method, the multiple linear regression method, the non-linear iterative partial least squares algorithm and the multiple imputation method

All results in this study proved that the multiple linear regression method; provided a successful estimation of the missing precipitation data, in addition; the multiple imputation method produced the most accurate results for precipitation data.

(2) In 2007 Hyun Kang conducted a study entitled (The prevention and handling of the missing data). This manuscript reviewed the problems and types of missing data, along with the techniques for handling missing data. And the methods for handling the missing data are discussed. The paper concluded with recommendations for the handling of missing data. From their most important, more attention should be paid to the missing data in the design and performance of the studies and in the analysis of the resulting data, in addition; Application of the sophisticated statistical analysis techniques should only be performed after the maximal efforts have been employed to reduce missing data in the design and prevention techniques.

(3) In 2009 SPSS conducted a study entitled (Missing data: the hidden problem). This white paper presented a case study demonstrating how missing data can affect your analysis and the decisions you make based on your results.

In this case study, missing data did in fact affect the analysis and results. By thoroughly analyzing the missing data and imputing the missing data, a more valid conclusion was reached. SPSS Missing Value Analysis provides the tools needed to diagnose missing data and take action.

(4) In 2016 Maria Pampaka, Graeme Hutcheson & Julian Williams introduced a study entitled (Handling missing data: analysis of a challenging data set using multiple imputation). This paper depended on compare methods, that is, step-wise regression (basically ignoring the missing data) and MI models, with the model from the actual enhanced sample.

This study demonstrated that even with this very difficult data set, MI still proved to be useful. But the most important conclusion from this paper is that missing data can have adverse effects on analyses and imputation methods should be considered when this is an issue.

(5) In 2018 Alvira Swalin introduced a study entitled (How to Handle Missing Data). She compared different methods of estimating Missing Data as (Time-series Analysis, ML, Regression, K Nearest Neighbors (KNN), etc.) The main results of the research are: among all the methods, multiple imputation and KNN are widely used, and multiple imputation being simpler is generally preferred.

(6) In 2017 Y Susianto, K A Notodiputro, A Kurnia and H Wijayanto introduced a study entitled (A Comparative Study of Imputation Methods for Estimation of Missing Values of Per Capita Expenditure in Central Java).

the paper discussed three imputation methods namely the Yates method, expectation-maximization (EM) algorithm, and Markov Chain Monte Carlo (MCMC) method. These methods were used to estimate the missing values of per-capita expenditure data at sub-districts level in Central Java.

they evaluated the performance of these imputation methods is evaluated by comparing the mean square error (MSE) and mean absolute error (MAE) of the resulting estimates using linear mixed models. It is showed that MSE and MAE produced by the Yates method are lower than the MSE and MAE resulted from both the EM algorithm and the MCMC method. Therefore, the Yates method is recommended to impute the missing values of per capita expenditure at sub-district level.

(7) In 2015 Aureliano Crameri, Agnes von Wyl, Margit Koemeda, Peter Schulthess and Volker Tschuschke introduced a study entitled (Sensitivity analysis in multiple imputation in effectiveness studies of psychotherapy). They presented a sensitivity analysis technique based on posterior predictive checking. And they demonstrated the possibilities this technique can offer with the

example of irregular longitudinal data collected with the outcome by questionnaire in a sample of 260 persons.

they presented the importance of sensitivity analysis in (1) quantify the degree of bias introduced by missing not at random data (MNAR) in a worst reasonable case scenario, (2) compare the performance of different analysis methods for dealing with missing data, or (3) detect the influence of possible violations to the model assumptions.

Finally, this study demonstrated that repeated measurements analyzed with MI are useful to improve the accuracy of outcome estimates in quality assurance assessments and non-randomized effectiveness studies in the field of outpatient psychotherapy.

Chapter II

Basic Concepts of Missing Data

2.1 Introduction:

This chapter reviews the definition of missing data, effects of Ignoring missing data, the three different classes of missing data (Mechanisms for missing data) , explain how different missing data mechanisms can be detected at least for some of the classes using Little's MCAR Test and Techniques for Handling the Missing Data. The chapter concludes with final message about missing data.

2.2Definition of missing data

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [2]. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research [3] and [4].

However, until recently, most researchers have drawn conclusions based on the assumption of a complete data set.

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions [5].

2.3 How to dealing with missing data

In an ideal world, your data set would always be perfect without any missing data. But perfect data sets are rare in ecology and evolution, or in any other field. Missing data haunts every type of ecological or evolutionary data: observational, experimental, comparative, or meta-analytic. But this issue is rarely addressed in research articles. Why? Researchers often play down the presence of missing data in their studies, because it may be perceived as a weakness of their work [6]; this tendency has been confirmed in medical trials [7]; [8] and [9].

The most common way of handling missing data is called list-wise deletion: researchers delete cases (or rows/lists) containing missing values and run a model, using the data set without missing values (known as complete case analysis). While common, few researchers explicitly state that they are using this approach.

What is wrong with deletion? The problems are twofold: (1) loss of information (i.e., reduction in statistical power) and (2) potential bias in parameter estimates under most circumstances (bias here means systematic deviation from population or true parameter values; [10].

The good news is that we now have solutions that combat missing data problems. They come in two forms: Estimation methods (e.g., MI, EM and Reg.), and data augmentation (DA; in the statistical literature, the term data augmentation is used in different ways, but we follow the usage of McKnight et al. [11]. The bad news is that very few researchers use such statistical tools [12] MI and DA have been available to us since the late 1980s, with some key publications in 1987 [13]; [14]; [15] and [16].

It is high time for us to finally start using missing data procedures in our analyses. This is especially so given the recent growth in the number of R libraries that can handle missing data appropriately using MI and DA [6]; [17].

2.4 Effects of ignoring missing data

It has been demonstrated repeatedly that missing data can have large effects on the results of a survey, Moreover, increasing the sample size without targeting missing data does nothing to reduce missing data bias; a larger sample size merely provides more observations from the class of persons that would respond to the survey. Increasing the sample size may actually worsen the missing data bias, as the larger sample size may divert resources that could have been used to reduce or remedy the missing data, or it may result in less care in the data collection. Most small surveys ignore any nonresponse that remains after callbacks and follow-ups, and report results based on complete records only. The main problem caused by missing data is potential bias. Some Factors Affecting Non-Response. Interviewers, Data-collection method, Questionnaire design, Incentives, financial or otherwise.

Results reported from an analysis of only complete records should be taken as representative of the population of persons who would respond to the survey, which is rarely the same as the target population. If you insist on estimating population means and totals using only the complete records and making no adjustment for nonrespondents, at the very least you should report the rate of nonresponse.

The main problem caused by nonresponse is potential bias. Think of the population as being divided into two somewhat artificial strata of respondents and nonrespondents. The population respondents are the units that would respond if they were chosen to be in the sample; the number of population respondents, N_R, is unknown. Similarly, the N_M (M for missing) population nonrespondents are the units that would not respond. We then have the following population quantities:

Stratum	Size	Total	Mean	Variance
Respondents	N_R	t_R	$\bar{y}_{RU}$	S_R^2
Nonrespondents	N_M	t_M	$\bar{y}_{MU}$	S_M^2
Entire population	N	t	$\bar{y}_U$	S^2

The population as a whole has variance $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$, mean $\bar{y}_U$, and total t. A probability sample from the population will likely contain some respondents and some nonrespondents. But, of course, on the first call we do not observe y_i for any of the units in the nonrespondent stratum. If the population mean in the nonrespondent stratum differs from that in the respondent stratum, estimating the population mean using only the respondents will produce bias.

Let $\bar{y}_R$ be an approximately unbiased estimator of the mean in the respondent stratum, using only the respondents. As

$$\bar{y}_U = \frac{N_R}{N}\bar{y}_{RU} + \frac{N_M}{N}\bar{y}_{MU}$$

The bias is approximately

$$E[\bar{y}_R] - \bar{y}_U \approx \frac{N_M}{N} (\bar{y}_{RU} - \bar{y}_{MU}).$$

The bias is small if either (1) the mean for the nonrespondents is close to the mean for the respondents, or (2) N_M/N is small—there is little nonresponse. But we can never be assured of (1), as we generally have no data for the nonrespondents. Minimizing the nonresponse rate is the only sure way to control nonresponse bias [1].

2.5 Mechanisms for missing data

Most surveys have some residual nonresponse even after careful design and follow-up of nonrespondents. All methods for fixing up nonresponse are necessarily model-based. If we are to make any inferences about the nonrespondents, we must assume that they are related to respondents in some way.

Dividing population members into two fixed strata of would-be respondents and would-be nonrespondents is fine for thinking about potential nonresponse bias and for two-phase methods. To adjust for nonresponse that remains after all other measures have been taken, we need a more elaborate setup. Define the random variable

$$R_i = \begin{cases} 1 \text{ if unit } i \text{ responds} \\ 0 \text{ if unit } i \text{ dose not responds} \end{cases}$$

After sampling, the realizations of the response indicator variable are known for the units selected in the sample. A value for y_i is recorded if r_i, the realization of R_i, is 1. The probability that a unit selected for the sample will respond,

$$\Phi_i = P(R_i = 1)$$

Is of course unknown but assumed positive. [18] call suppose that y_i is a response of interest, and that x_i is a vector of information known about unit *i* in the sample. Information used in the survey design is included in x_i. We consider three types of missing data, using [19] terminology of nonresponse classification.

2.5.1 Missing at random (MAR)

If Φ_i depends on x_i, but not on y_i, the data are missing at random (MAR); the nonresponse depends only on observed variables. We can successfully model the nonresponse, since we know the values of *xi* for all sample units [1].

As we tend to consider randomness as not producing bias, we may think that (MAR) does not present a problem. However, (MAR) does not mean that the missing data can be ignored. If a dropout variable is (MAR), we may expect that the probability of a dropout of the variable in each case is conditionally independent of the variable, which is obtained currently and expected to be obtained in the future, given the history of the obtained variable prior to that case [5].

The practical problem with the MAR mechanism is that there is no way to confirm that the probability of missing data on y is solely a function of other measured variables

2.5.2 Missing Completely at Random (MCAR)

If Φ_i does not depend on x_i y_i, or the survey design, the missing data are missing completely at random (MCAR). Such a situation occurs if, for example, someone at the laboratory drops a test tube containing the blood sample of one of the survey participants—there is no reason to think that the dropping of the test tube had anything to do with the white blood cell count.

If the response probabilities Φ_i are all equal and the events $\{R_i = 1\}$ are conditionally independent of each other and of the sample selection process given n_R, then the data are MCAR. If an SRS of size n is taken, then under this mechanism the respondents will be a simple random subsample of variable size n_R. The sample mean of the respondents, $\bar{y}_R$ is approximately unbiased for the population mean. The MCAR mechanism is implicitly adopted when nonresponse is ignored [1].

The statistical advantage of data that are MCAR is that the analysis remains unbiased. Power may be lost in the design, but the estimated parameters are not biased by the absence of the data [5].

2.5.3 Missing Not at Random (MNAR)

Finally, data are missing not at random (MNAR) when the probability of missing data on a variable y_i is related to the values of y_i itself, even after controlling for other variables. Or; if the characters of the data do not meet those of (MCAR) or (MAR), then they fall into the category of missing not at random (MNAR).

The cases of (MNAR) data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data. The model may then be incorporated into a more complex one for estimating the missing values [5].

The probabilities of responding, Φ_i, are useful for thinking about the type of nonresponse. Unfortunately, they are unknown, so we do not know for sure which type of nonresponse is present. We can sometimes distinguish between (MCAR) and (MAR) by fitting a model attempting to predict the observed probabilities of response for subgroups from known covariates. If the coefficients in a logistic regression model predicting nonresponse are significantly different from 0, the missing data are likely not (MCAR). Distinguishing between (MAR) and (MNAR) is more difficult. In practice, we expect most nonresponse in surveys to be of the (MNAR) type. It is unreasonable to expect that we can construct a perfect model that will completely explain the nonresponse mechanism. But we can try to reduce the bias due to nonresponse.

2.6 Diagnosing the Mechanism

2.6.1 MAR vs. MNAR

The only true way to distinguish between MNAR and MAR is to measure some of that missing data. It's a common practice among professional surveyors to, for example, follow-up on a paper survey with phone calls to a group of the non-respondents and ask a few key survey items. This allows you to compare respondents to non-respondents.

If their responses on those key items differ by very much, that's good evidence that the data are MNAR.

However in most missing data situations, we don't have the luxury of getting a hold of the missing data. So while we can't test it directly, we can examine patterns in the data get an idea of what's the most likely mechanism.

The first thing in diagnosing randomness of the missing data is to use your substantive scientific knowledge of the data and your field. The more sensitive the issue, the less likely people are to tell you. They're not going to tell you as much about their cocaine usage as they are about their phone usage.

Likewise, many fields have common research situations in which non-ignorable data is common. Educate yourself in your field's literature.

2.6.2 MCAR vs. MAR

There is a very useful test for MCAR, Little's test. But like all tests of assumptions, it's not definitive. So run it, but use it as only one piece of information.

A second technique is to create dummy variables for whether a variable is missing.

1 = missing

0 = observed

You can then run t-tests and chi-square tests between this variable and other variables in the data set to see if the missingness on this variable is related to the values of other variables.

For example, if women really are less likely to tell you their weight than men, a chi-square test will tell you that the percentage of missing data on the weight variable is higher for women than men.

The SPSS Missing Data module has a very nice procedure for doing this automatically–you don't have to create all those dummy variables. I don't know of other software packages having this built in, but it's not hard to program. [20]

2.7 Rerunning the Analysis for Little's MCAR Test

The results of Little's MCAR test appear in footnotes to each EM, Reg. estimate tables. The null hypothesis for Little's MCAR test is that the data are missing completely at random (MCAR). If the test has a not significance level of P<0.05 the data can be considered as not missing completely at random NMCAR.

Data are MCAR when the pattern of missing values does not depend on the data values. If the test has a significance level of P>0.05 the data can be considered as missing completely at random MCAR. [21]

Note: It is important to note that you’re not able to test whether your missing data is MAR or MNAR. The above mentioned procedures will only give you an indication for MCAR data or MAR/MNAR data. Pay attention to the possibility of MNAR, because all analyses have serious problems when you’re missing data is MNAR. [22]

2.8 Final question: What if our data is missing but not at random?

We must specify a model for the probability of missing data, which can be pretty challenging as it requires a good understanding of the data generating process. The Sample Selection Bias Model, by James Heckman, is a widely used method that you can apply in SAS using PROC QLIM [23].

2.9 Missing data Analysis

Missing data analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

The Missing Value Analysis procedure performs three primary functions:

• Describes the pattern of missing data. Where are the missing values located? How extensive are they? Do pairs of variables tend to have values missing in multiple cases? Are data values extreme? Are values missing randomly?

• Estimates means, standard deviations, covariances, and correlations for different missing value methods: expectation maximization method, multiple imputation method and regression imputation method

• Fills in (imputes) missing values with estimated values using expectation maximization or regression imputation; however, multiple imputation is generally considered to provide more accurate results

2.10 Techniques for Handling the Missing Data

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully [24]. The following are suggested to minimize the amount of missing data in the surveys: [5]

First, the study design should limit the collection of data to those who are participating in the study. This can be achieved by minimizing the number of follow-up visits, collecting only the essential information at each visit, and developing the userfriendly case-report forms.

Second, before the beginning of the research, a detailed documentation of the study should be developed in the form of the manual of operations, which includes the methods to screen the participants, protocol to train the investigators and participants, methods to communicate between the investigators or between the investigators and participants, implementation of the treatment, and procedure to collect, enter, and edit data.

Third, before the start of the participant enrollment, a training should be conducted to instruct all personnel related to the study on all aspects of the study, such as the participant enrollment, collection and entry of data, and implementation of the treatment or intervention

Fourth, if a small pilot study is performed before the start of the main trial, it may help to identify the unexpected problems which are likely to occur during the study, thus reducing the amount of missing data.

Fifth, the study management team should set a priori targets for the unacceptable level of missing data. With these targets in mind, the data collection at each site should be monitored and reported in as close to real-time as possible during the course of the study.

Sixth, study investigators should identify and aggressively, though not coercively, engage the participants who are at the greatest risk of being lost during follow-up.

Finally, if a participant decides to withdraw from the follow-up, the reasons for the withdrawal should be recorded for the subsequent analysis in the interpretation of the results.

2.11 Final messages

Missing data are pervasive, and pose problems for many statistical procedures. We hope we have convinced you that we all should be using methods that treat missing data properly (i.e., MI, EM or Reg.), rather than deleting data or using single imputation. Importantly, it is not difficult to implement these missing data. We also hope that you will now think about the missingness mechanisms when planning studies (i.e., collecting auxiliary variables). Especially, we think that researchers can probably benefit a lot from learning the planned missing design [25]; [26]; [27]; [28] and [29], although such a concept is nearly unheard of in our field.

We also presented you with some current difficulties associated with missing data. There are no easy solutions for missing values in multilevel data, especially when missing values occur in multiple levels and when clustering occurs at more than two levels. Nor is the implementation of MNAR models straightforward. But missing data theory is an active area of research, so who knows what the future will bring to us? [30] comments that "Until more robust MNAR analysis models become available (and that may never happen), increasing the sophistication level of MAR analysis may be the best that we can do." [31]

Chapter III

Methods for Estimating Missing Data

3.1 Introduction

Missing data arise in almost all serious statistical analyses. In this chapter we discuss a variety of methods to handle missing data, including some relatively simple approaches that can often yield reasonable results. We used three methods, and we applied on generated data). And we would like to simply clean the dataset so it could be analyzed as if there were no missingness.

3.2 Methods for handling missing data

It is not uncommon to have a considerable amount of missing data in a study. One technique of handling the missing data is to use the data analysis methods which are robust to the problems caused by the missing data. An analysis method is considered robust to the missing data when there is confidence that mild to moderate violations of the assumptions will produce little to no bias or distortion in the conclusions drawn on the population. However, it is not always possible to use such techniques. Therefore, a number of alternative ways of handling the missing data has been developed.

3.3 Conventional methods

3.3.1 Listwise deletion (or complete case analysis):

If a case has missing data for any of the variables, then simply exclude that case from the analysis. It is usually the default in statistical packages. [32].

Advantages: It can be used with any kind of statistical analysis and no special computational methods are required.

Limitations: It can exclude a large fraction of the original sample. For example, suppose a data set with 1,000 people and 20 variables. Each of the variables has missing data on 5% of the cases, then, you could expect to have complete data for only about 360 individuals, discarding the other 640. It works well when the data are missing completely at random (MCAR), which rarely happens in reality [33].

3.3.2 Pairwise deletion

Pairwise deletion eliminates information only when the particular data-point needed to test a particular assumption is missing. If there is missing data elsewhere in the data set, the existing values are used in the statistical testing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data. This approach presents the following problems: 1) the parameters of the model will stand on different sets of data with different statistics, such as the sample size and standard errors; and 2) it can produce an intercorrelation matrix that is not positive definite, which is likely to prevent further analysis [34].

Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

3.4 Advanced Methods

3.4.1 Imputation methods:

Substitute each missing value for a reasonable guess, and then carry out the analysis as if there were not missing values.

There are two main imputation techniques:

- *Marginal mean imputation:* Compute the mean of X using the non-missing values and use it to impute missing values of X.

  Limitations: It leads to biased estimates of variances and covariances and, generally, it should be avoided.

- *Conditional mean imputation:* Suppose we are estimating a regression model with multiple independent variables. One of them, X, has missing values. We select those cases with complete information and regress X on all the other independent variables. Then, we use the estimated equation to predict X for those cases it is missing.

  If the data are MCAR, least-squares coefficients are consistent (i.e. unbiased as the sample size increases) but they are not fully efficient (remember, efficiency is a measure of the optimality of an estimator. Essentially, a more efficient estimator, experiment or test needs fewer samples than a less efficient one to achieve a given performance). Estimating the model using weighted least squares or generalized least squares leads to better results [32], [35], [36].

Limitations of imputation techniques in general: They lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error [35].

3.4.1.1 Regression Imputation Method of estimation:

A much more promising method is to use standard regression analysis to provide estimates of the missing data conditional on complete variables in the analysis. For example, for the simple case of univariate missingness in a single continuous variable Y, we fit a regression model to explain Y by the remaining p variables represented by the vector X using the complete cases (subscripted by i):

$$Y_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i \qquad (1)$$

Predicted values for the expected values of the missing cases of Y (subscripted by j) can be obtained from

$$\hat{Y}_j = \hat{\alpha} + \sum_{k=1}^{p} \hat{\beta}_k X_{jk} \qquad (2)$$

It should be emphasized that the equations above could be generalized to include models for non-continuous data such as binomial or count data.

Missing data are usually multivariate and it is possible to extend the procedure of regression based imputation from the univariate case to deal with multivariate missingness. For each missing value in the data set a model can be fitted for that variable employing the complete cases of all the other variables [37]. Where the number of variables with missing values is large, the number of models to be fitted will also be large, however, efficient computational methods (such as Little & Rubin's sweep operator) can be employed [38]. Alternatively, an iterative regression approach can be adopted [39] whereby missing values in a given variable are predicted from a regression of that variable on the complete cases of all other variables in the dataset. This process is repeated for all variables with missing values using complete cases of the other variables *including previously imputed values* until a completed rectangular data set has been generated. The imputation of missing values for each variable is then re-estimated in turn using the complete set of data and the process continues until the imputed values stop changing

Advantages: The imputation retains a great deal of data over the listwise or pairwise deletion and avoids significantly altering the standard deviation or the shape of the distribution. However, as in a mean substitution, while a regression imputation substitutes a value that is predicted from other variables, no novel information is added, while the sample size has been increased and the standard error is reduced.[40]

3.4.1.2 Multiple Imputation method of estimation:

It is important to recognize that when employing any imputation method we are estimating a missing value that is not observed. It is straightforward to see that in the case of unconditional mean imputation, the variance of the completed variable will be too low, since the imputed means do not contribute to the variance. However, the same is true with the other forms of imputation – if the expected value of the missing data point is imputed, although this is the 'best' prediction of the missing value (in the sense of mean squared error), there will be no allowance for the uncertainty associated with the imputation process. For example, if imputations are based on a regression equation, as in Equation (2) for the simple univariate missingness example, then there will be no variation between predicted values for observations with the same values for all of the other non-missing variables. Such 'deterministic' imputation approaches [39] will therefore underestimate the variance of any estimators in subsequent statistical analysis of the imputed data set. Therefore, imputed values of missing data should include a random component to reflect the fact that imputed values are estimated (using so-called 'stochastic' imputation methods [39]) rather than treating the imputed values as if they are known with certainty.

For the regression example, two components to the uncertainty in the imputation process can be distinguished. The first component is the mean squared error from the regression which represents the between observation variability not explained by the regression model. Two approaches to including this error term are either: to select a value at random from a normal distribution with variance equal to the mean squared error from the regression; or to compute the residuals from the regression and to add one of these residuals at random to each of the imputed values from the regression. Of these two approaches, the second non-parametric bootstrap approach is probably preferred since it is straightforward to do and does not rely on the parametric assumption of normally distributed errors. The second component of uncertainty comes from the fact that the coefficients of the regression model are themselves estimated rather than known. The variance of the prediction error for each covariate pattern can be obtained from the variance–covariance matrix and, assuming multivariate normality, this component of uncertainty can also be incorporated into the stochastic imputation procedure.

Clearly, once missing values are imputed with a random component, then a complete data set will no longer be unique and the results of any analysis of will be dependent on the particular imputed values. The principle of multiple imputation uses this fact directly in order to allow estimation of variance in statistics of interest in an analysis that include representation of uncertainty in the true values of the missing information.

With multiple imputation, an incomplete data set will have the missing values imputed several (M) times, where the values to fill in are drawn from the predictive distribution of the missing data, given the observed data. Each imputed data set is then separately analyzed with the desired methods for complete data. The variability in the statistic of interest across the alternative data sets then gives an explicit assessment of the increase in variance due to missing data. Thus this variance of each final parameter estimate is composed of two parts: the estimated variance within each imputed data set and the variance across the data sets.

Suppose that the statistic of interest in the analysis is given by y.

The steps in the multiple imputation procedure are then:

1. Generate M sets of imputed values for the missing data points, thus creating M completed data sets.

2. For each completed data set, carry out the standard complete data analysis, obtaining estimate $\hat{\theta}_i$ of interest and its estimated variance $\hat{a}r\left(\hat{\theta}_i\right) for\ i = 1 \dots M$.

3. Combine the results from the different data sets. The multiple imputation estimate of θ is

$$\hat{\theta} = \frac{1}{M}\sum_{I=1}^{M}\hat{\theta}_i$$

(i.e. the mean across the imputed data sets) and multiple imputation estimate of variance is

$$v\hat{a}r(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M} v\hat{a}r(\hat{\theta}_i) + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{i=1}^{M}(\hat{\theta}_i - \hat{\theta})$$

The first term on the right hand side of this equation relates to the variance within the imputed data sets, whereas the term on the far right captures the uncertainty due to the variability in the imputed values, i.e. between the imputed data sets. The term 1+1/M is a bias correction factor.

The approximate reference distribution for interval estimates and significance tests is a t distribution with degrees of freedom $= (M - 1)(1 + r^{-1})^2$; [40] where r is the estimated ratio of the between-imputation component of variance (numerator) to the within-imputation component of variance (denominator).

Rubin [42] shows that the relative efficiency of an estimate based on M complete data sets to one based on an infinite number of them is approximately $(1 + \gamma/M)^{-1}$ where γ is the rate of missing data. With 50% missing data, an estimate based on M ¼ 5 complete data sets has a standard deviation that is only about 5% wider than one based on infinite M. Unless rates of missing data are very high, there is little advantage to using more than five complete data sets [43].

Advantages: It has the same optimal properties as ML, and it removes some of its limitations. Multiple imputation can be used with any kind of data and model with conventional software. When the data is MAR, multiple imputation can lead to consistent, asymptotically efficient, and asymptotically normal estimates. *Limitations:* It is a bit challenging to successfully use it. It produces different estimates (hopefully, only slightly different) every time you use it, which can lead to situations where different researchers get different numbers from the same data using the same method. [33], [36]

3.4.2 Last observation carried forward

In the field of anesthesiology research, many studies are performed with the longitudinal or time-series approach, in which the subjects are repeatedly measured over a series of time-points. One of the most widely used imputation methods in such a case is the last observation carried forward (LOCF). This method replaces every missing value with the last observed value from the same subject. Whenever a value is missing, it is replaced with the last observed value [44].

This method is advantageous as it is easy to understand and communicate between the statisticians and clinicians or between a sponsor and the researcher.

Although simple, this method strongly assumes that the value of the outcome remains unchanged by the missing data, which seems unlikely in many settings (especially in the anesthetic trials). It produces a biased estimate of the treatment effect and underestimates the variability of the estimated result. Accordingly, the National Academy of Sciences has recommended against the uncritical use of the simple imputation, including LOCF and the baseline observation carried forward, stating that:

Single imputation methods like last observation carried forward and baseline observation carried forward should not be used as the primary approach to the treatment of missing data unless the assumptions that underlie them are scientifically justified [45].

3.4.3 Maximum Likelihood

We can use this method to get the variance-covariance matrix for the variables in the model based on all the available data points, and then use the obtained variance- covariance matrix to estimate our regression model. [46]

Compared to MI, MI requires many more decisions than ML (whether to use Markov Chain Monte Carlo (MCMC) method or the Fully Conditional Specification (FCS), how many data sets to produce, how many iterations between data sets, what prior distribution to use-the default is Jeffreys-, etc.). On the other hand, ML is simpler as you only need to specify your model of interest and indicate that you want to use ML. [47]

There are two main ML methods:

3.4.3.1 Direct Maximum Likelihood: It implies the direct maximization of the multivariate normal likelihood function for the assumed linear model. *Advantage:* It gives efficient estimates with correct standard errors. *Limitations:* It requires specialized software (it may be challenging and time consuming).

3.4.3.2 The Expectation-maximization (EM) algorithm of estimation:

This algorithm is a parametric method to impute missing values based on the maximum likelihood estimation. This algorithm is very popular in statistical literatures and has been discussed intensively by many researchers, such as: [48], [49], [50], and [51]

This algorithm uses an iterative procedure to finding the maximum likelihood estimators of parameter vector through two step described in Dempsteret al. [49] and [50] as follows:

a). The Expectation step (E-step)

The E step is the stage of determining the conditional expected value of the full data of log likelihood function $l(\theta|Y)$ given observed data. Suppose for any incomplete data, the distribution of the complete data Y can be factored as

$$f(Y|\theta) = f(Y_{mis}, Y_{obs}|\theta)$$

$$= f(Y_{obs}|\theta)\, f(Y_{mis}|Y_{obs}, \theta) \quad (1)$$

Where $f(Y_{obs}|\theta)$ the distribution of the data is observed Y_{obs} and $f(Y_{mis}, Y_{obs}|\theta)$ is the distribution of missing data given data observed. Based on the equation (1), we obtained log likelihood function

$$l(\theta|Y) = l(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs}, \theta) \quad (2)$$

Where $l(\theta|Y)$ is log likelihood function of complete data, $l(\theta|Y_{obs})$ is log likelihood function of observed data, and $f(Y_{mis}|Y_{obs}, \theta)$ is the predictive distribution of missing data given θ

Objectively, to estimate θ is done by maximizing the log likelihood function (2). Because Y_{mis} not known, the right side of equation (2) can not be calculated. As a solution, $l(\theta|Y)$ is calculated based on the average value $\log f(Y_{mis}|Y_{obs}, \theta)$ using predictive distribution $f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right)$, where $\theta^{(t)}$ is temporary estimation of unknown parameters. In this context, an initial estimation $\theta^{(0)}$ be calculated using the complete case analysis. With this approach, the mean value of equation (4) can be expressed

$$Q\left(\theta\middle|\theta^{(t)}\right) = l(\theta|Y_{obs}) + \int \log f(Y_{mis}|Y_{obs}, \theta)\, f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right) \partial Y_{mis}$$

$$= \int [l(\theta|Y_{obs}) + \int \log f(Y_{mis}|Y_{obs}, \theta)]\; f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right) \partial Y_{mis}$$

$$= \int l(\theta|Y)\, f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right) \partial Y_{mis} \quad (3)$$

The equation (3) basically a conditional expected value of log likelihood function for complete data $l(\theta|Y)$ given observed data and initial estimate of unknown parameter.

b). the maximization step (M-step)

The M step is to obtained the iteratively estimation $\theta^{(t+1)}$ with maximizes $Q(\theta|\theta^{(t)})$ as follow

$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$ (4)

Both E and M steps are iterated until convergent.

Advantage: We can use SAS, since this is the default algorithm it employs for dealing with missing data with Maximum Likelihood.

Limitations: Only can be used for linear and log-linear models (there is neither theory nor software developed beyond them). [35], [36], [52] and [53]

3.5 Other advanced methods

3.5.1 Bayesian simulation methods

There are two main methods:

Firstly: Schafer algorithms:

It uses Bayesian iterative simulation methods to impute data sets assuming MAR. Precisely, it splits the multivariate missing problem into a series of univariate problems based on the assumed distribution of the multivariate missing variables (e.g. multivariate normal for continuous variables, multinomial log linear for categorical variables). In other words, it uses an iterative algorithm that draws samples from a sequence of univariate regressions.

Secondly: Van Buuren algorithm:

It is a semi-parametric approach. The parametric part implies that each variable has a separate imputation model with a set of predictors that explain the missingness. The non-parametric part implies the specification of an appropriate form (e.g. linear), which depends on the kind of variables [32] and [54]

3.5.2 Hot deck imputation methods

It is used by the US Census Bureau. This method completes a missing observation by selecting at random, with replacement, a value from those individuals who have matching observed values for other variables. In other words, a missing value is imputed based on an observed value that is closer in terms of distance. SAS macro developed by Lawrence Altmayer, of the U.S. Census Bureau. Can be found in Ahmed Kazi et al; 2009. [32]

3.6 Sensitivity analysis

Sensitivity analysis is defined as the study which defines how the uncertainty in the output of a model can be allocated to the different sources of uncertainty in its inputs.

When analyzing the missing data, additional assumptions on the reasons for the missing data are made, and these assumptions are often applicable to the primary analysis. However, the assumptions cannot be definitively validated for the correctness. Therefore, the National Research Council has proposed that the sensitivity analysis be conducted to evaluate the robustness of the results to the deviations from the MAR assumption [55]

3.7 Final messages:

- Make every effort to avoid missing data, or failing that, to understand how much and why data is missing.
- Understand missing data mechanisms (MCAR, MAR, MNAR) and their implications.
- Avoid default methods (listwise deletion, pairwise deletion).
- Avoid default fixups (mean imputation, etc.) where possible.
- Use multiple imputation to take proper account of missings.
- Do a sensitivity analysis.

Chapter IV

Comparative Study of the Methods of Estimating Missing Data

4.1: Introduction

This chapter includes the applied aspect to what explained in the theoretical chapter and we will describe the data, test the little’s MCAR of the missing data, one way ANOVA test, calculate covariances matrix and correlations, lastly Std. Error Mean and MAE to comparative between estimation methods, this chapter shows that

4.2: Description of study's data

We applied this study on generated data, it has the same mean with three different variances such:

- Normally distributed data and missing completely at random (MCAR)- missing value ((5%), (10%), (15%), (20%)) and (30%) respectively. With variance (1.04).

- Normally distributed data and missing completely at random (MCAR)- missing value ((5%), (10%), (15%), (20%)) and (30%) respectively. With variance (26.88).

- Normally distributed data and missing completely at random (MCAR)- missing value ((5%), (10%), (15%), (20%)) and (30%) respectively. With variance (83.74).

- Normally distributed data and missing not completely at random - missing value ((10%), (20%) and (30%)) respectively. With variance (1.04).

- Normally distributed data and missing not completely at random - missing value ((10%), (20%) and (30%)) respectively. With variance (26.88).

- Normally distributed data and missing not completely at random - missing value ((10%), (20%) and (30%)) respectively. With variance (83.74).

4.3 Results obtained by MCAR: missing value 5% with variance 1.04

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.1): little's test results

little’s test	
chi-square	0.023
Sig.	0.881

Source: The researcher from applied study, SPSS Package, 2018

From table (4.1), it shows the sig. value of little's test (0.881) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.2): Statistics results

	MI method	***Regression method***	***EM method***
Mean	999.79	999.85	999.82
Variance	1.07	0.99	1.05

Source: The researcher from applied study, SPSS Package, 2018

From table (4.2), the results revealed that the regression method has a mean (999.85) greater than the means of EM method and MI method (999.82) and (999.79) respectively. With a variance of regression method (0.99) lower than the variances of EM method and MI method (1.05) and (1.07) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.3): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	0.091	3	0.030	0.044	0.987
Within Groups	10.978	16	0.686		
Total	11.068	19			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.3), it shows the sig. value of the F-test (0.987) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.4): covariances results

	MI Method	*Regression method*	*EM method*
covariance	0.044	0.012	- 0.466

Source: The researcher from applied study, SPSS Package, 2018

From table (4.4), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.5): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.085	0.410	- 0.427
Sig.	0.892	0.493	0.473

Source: The researcher from applied study, SPSS Package, 2018

From table (4.5), it has been shown that according to the sig. values of the Chi square test (0.892), (0.493) and (0.473) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.6): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.227	0.013	0.476
MAE	1.742	1.671	1.825

Source: The researcher from applied study, SPSS Package, 2018

From table (4.6), the results revealed that MAE of the regression method was lower than MAE of MI method and EM method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***the regression method*** is more efficient than the other two methods.

4.4 Results obtained by MCAR: missing value 5% with variance 26.88

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.7): little's test results

little’s test	
chi-square	1.057
Sig.	0.304

Source: The researcher from applied study, SPSS Package, 2018

From table (4.7), it shows the sig. value of little's test (0.304) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.8): Statistics results

	MI method	***Regression method***	***EM method***
mean	1000.15	1000.00	1000.02
variance	24.78	25.11	24.29

Source: The researcher from applied study, SPSS Package, 2018

From table (4.8), the results revealed that MI method has a mean (1000.15) greater than the means of EM method and the regression method (1000.02) and (1000) respectively. With a variance of EM method (24.29) lower than the variances of MI method and the regression method (24.78) and (25.11) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.9): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	110.176	3	36.725	2.126	0.137
Within Groups	276.379	16	17.274		
Total	386.555	19			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.9), it shows the sig. value of the F-test (0.137) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.10): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 6.593	- 3.564	0.088

Source: The researcher from applied study, SPSS Package, 2018

From table (4.10), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.11): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.438	- 0.120	0.334
Sig.	0.461	0.847	0.583

Source: The researcher from applied study, SPSS Package, 2018

From table (4.11), it has been shown that according to the sig. values of the Chi square test (0.461), (0.847) and (0.583) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.12): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.015	1.997	0.018
MAE	2.005	2.332	1.673

Source: The researcher from applied study, SPSS Package, 2018

From table (4.12), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.5 Results obtained by MCAR: missing value 5% with variance 83.74

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.13): little's test results

little’s test	
chi-square	0.122
Sig.	0.727

Source: The researcher from applied study, SPSS Package, 2018

From table (4.13), it shows the sig. value of little's test (0.727) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.14): Statistics results

	MI method	***Regression method***	***EM method***
mean	1002.34	1002.42	1002.11
variance	80.73	85.40	79.15

Source: The researcher from applied study, SPSS Package, 2018

From table (4.14), the results revealed that the regression method has a mean (1002.42) greater than the means of MI method and EM method (1002.34) and (1002.11) respectively. With a variance of EM method (79.15) lower than the variances of MI method and the regression method (80.73) and (85.40) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.15): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	106.059	3	35.353	0.629	0.607
Within Groups	899.108	16	56.194		
Total	1005.167	19			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.15), it shows the sig. value of the F-test (0.607) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.16): covariances results

	MI Method	*Regression method*	*EM method*
covariance	21.927	28.905	0.796

Source: The researcher from applied study, SPSS Package, 2018

From table (4.16), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.17): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.596	0.276	0.059
Sig.	0.288	0.654	0.926

Source: The researcher from applied study, SPSS Package, 2018

From table (4.17), it has been shown that according to the sig. values of the Chi square test (0.288), (0.654) and (0.926) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.18): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.621	4.625	0.599
MAE	2.207	3.208	1.866

Source: The researcher from applied study, SPSS Package, 2018

From table (4.18), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.6 Results obtained by MCAR: missing value 10% with variance 1.04

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.19): little's test results

little’s test	
chi-square	0.236
Sig.	0.627

Source: The researcher from applied study, SPSS Package, 2018

From table (4.19), it shows the sig. value of little's test (0.627) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.20): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.88	999.85	999.88
variance	1.00	1.09	0.94

Source: The researcher from applied study, SPSS Package, 2018

From table (4.20), the results revealed that MI and EM methods have the same mean (999.88); and it greater than a mean of the regression method (999.85). With a variance of EM method (0.94) lower than the variances of MI method and the regression method (1.00) and (1.09) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.21): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	2.111	3	0.704	0.850	0.476
Within Groups	29.820	36	0.828		
Total	31.931	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.21), it shows the sig. value of the F-test (0.476) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.22): covariances results

	MI Method	*Regression method*	*EM method*
covariance	0.003	- 0.457	0.002

Source: The researcher from applied study, SPSS Package, 2018

From table (4.22), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.23): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.003	- 0.361	0.039
Sig.	0.993	0.305	0.914

Source: The researcher from applied study, SPSS Package, 2018

From table (4.23), it has been shown that according to the sig. values of the Chi square test (0.993), (0.305) and (0.914) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.24): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.261	0.406	0.019
MAE	3.420	3.469	3.340

Source: The researcher from applied study, SPSS Package, 2018

From table (4.24), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.7 Results obtained by MCAR: missing value 10% with variance 26.88

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.25): little's test results

little’s test	
chi-square	0.015
Sig.	0.904

Source: The researcher from applied study, SPSS Package, 2018

From table (4.25), it shows the sig. value of little's test (0.904) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.26): Statistics results

	MI method	*Regression method*	*EM method*
mean	999.82	999.87	999.81
variance	26.21	28.77	24.56

Source: The researcher from applied study, SPSS Package, 2018

From table (4.26), the results revealed that the regression method has a mean (999.87) greater than the means of MI method and EM method (999.82) and (999.81) respectively. With a variance of EM method (24.56) lower than the variances of and MI method and the regression method (26.21) and (28.77) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.27): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	3.984	3	1.328	0.059	0.981
Within Groups	813.748	36	22.604		
Total	817.733	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.27), it shows the sig. value of the F-test (0.981) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.28): covariances results

	MI Method	*Regression method*	*EM method*
covariance	1.188	0.972	- 0.470

Source: The researcher from applied study, SPSS Package, 2018

From table (4.28), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.29): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.063	0.041	- 0.498
Sig.	0.863	0.911	0.143

Source: The researcher from applied study, SPSS Package, 2018

From table (4.29), it has been shown that according to the sig. values of the Chi square test (0.863), (0911) and (0.143) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.30): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.369	1.728	0.068
MAE	3.790	3.909	3.356

Source: The researcher from applied study, SPSS Package, 2018

From table (4.30), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.8 Results obtained by MCAR: missing value 10% with variance 83.74

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.31): little's test results

little’s test	
chi-square	2.471
Sig.	0.116

Source: The researcher from applied study, SPSS Package, 2018

From table (4.31), it shows the sig. value of little's test (0.116) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.32): Statistics results

	MI method	*Regression method*	*EM method*
mean	1002.05	1001.81	1001.82
variance	92.45	83.94	77.67

Source: The researcher from applied study, SPSS Package, 2018

From table (4.32), the results revealed that MI method has a mean (1002.05) greater than the means of EM method and the regression method (1001.82) and (1002.81) respectively. With a variance of EM method (77.67) lower than the variances of the regression method and MI method (83.94) and (92.45) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.33): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	140.181	3	46.727	0.702	0.557
Within Groups	2397.460	36	66.596		
Total	2537.641	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.33), it shows the sig. value of the F-test (0.557) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.34): covariances results

	MI Method	*Regression method*	*EM method*
covariance	15.518	10.469	- 0.377

Source: The researcher from applied study, SPSS Package, 2018

From table (4.34), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.35): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.192	0.194	- 0.146
Sig.	0.595	0.592	0.687

Source: The researcher from applied study, SPSS Package, 2018

From table (4.35), it has been shown that according to the sig. values of the Chi square test (0.595), (0.592) and (0.687) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.36): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	3.934	2.633	0.126
MAE	4.645	4.211	3.375

Source: The researcher from applied study, SPSS Package, 2018

From table (4.36), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.9 Results obtained by MCAR: missing value 15% with variance 1.04

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.37): little's test results

little's test	
chi-square	2.768
Sig.	0.096

Source: The researcher from applied study, SPSS Package, 2018

From table (4.37), it shows the sig. value of little's test (0.096) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.38): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.86	999.83	999.85
variance	0.96	1.00	0.86

Source: The researcher from applied study, SPSS Package, 2018

From table (4.38), the results revealed that MI method has a mean (999.86) greater than the means of EM method and the regression method (999.85) and (999.83) respectively. With a variance of EM method (0.86) lower than the variances of MI method and the regression method (0.96) and (1.00) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.39): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	0.305	3	0.102	0.145	0.933
Within Groups	39.331	56	0.702		
Total	39.636	59			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.39), it shows the sig. value of the F-test (0.933) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.40): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.488	0.485	0.004

Source: The researcher from applied study, SPSS Package, 2018

From table (4.40), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.41): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.550	0.452	0.238
Sig.	0.034	0.091	0.393

Source: The researcher from applied study, SPSS Package, 2018

From table (4.41), it has been shown that according to the sig. values of the Chi square test of generated values and estimated missing values of MI method (0.034) is less than significant level (0.05) that means, the correlation is significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.42): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.208	0.251	0.004
MAE	5.069	5.084	5.001

Source: The researcher from applied study, SPSS Package, 2018

From table (4.42), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.10 Results obtained by MCAR: missing value 15% with variance 26.88

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.43): little's test results

little’s test	
chi-square	0.058
Sig.	0.809

Source: The researcher from applied study, SPSS Package, 2018

From table (4.43), it shows the sig. value of little's test (0.809) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.44): Statistics results

	MI method	*Regression method*	*EM method*
mean	1000.36	1000.33	1000.28
variance	21.15	23.13	19.03

Source: The researcher from applied study, SPSS Package, 2018

From table (4.44), the results revealed that EM method has a mean (1000.28) greater than the means of the regression method and MI method (1000.33) and (1000.36) respectively. With a variance of EM method (19.03) lower than the variances of MI method and the regression method (21.15) and (23.13) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.45): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	126.984	3	42.328	1.866	0.146
Within Groups	1270.587	56	22.689		
Total	1397.571	59			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.45), it shows the sig. value of the F-test (0.146) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.46): covariances results

	MI Method	*Regression method*	*EM method*
covariance	7.903	- 7.930	- 0.016

Source: The researcher from applied study, SPSS Package, 2018

From table (4.46), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.47): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.300	- 0.215	- 0.297
Sig.	0.277	0.442	0.282

Source: The researcher from applied study, SPSS Package, 2018

From table (4.47), it has been shown that according to the sig. values of the Chi square test (0.277), (0.442) and (0.282) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.48): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.989	1.388	0.002
MAE	5.330	5.463	5.001

Source: The researcher from applied study, SPSS Package, 2018

From table (4.48), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.11 Results obtained by MCAR: missing value 15% with variance 83.74

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.49): little's test results

little's test	
chi-square	0.547
Sig.	0.460

Source: The researcher from applied study, SPSS Package, 2018

From table (4.49), it shows the sig. value of little's test (0.460) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.50): Statistics results

	MI method	*Regression method*	*EM method*
mean	1002.62	1002.04	1002.46
variance	72.13	79.65	65.73

Source: The researcher from applied study, SPSS Package, 2018

From table (4.50), the results revealed that MI method has a mean (1002.62) greater than the means of EM method and the regression method (1002.46) and (1002.04) respectively. With a variance of EM method (65.73) lower than the variances of MI method and the regression method (72.13) and (79.65) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.51): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	125.535	3	41.845	0.616	0.608
Within Groups	3806.583	56	67.975		
Total	3932.117	59			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.51), it shows the sig. value of the F-test (0.608) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.52): covariances results

	MI Method	*Regression method*	*EM method*
covariance	17.611	-15.543	-2.225

Source: The researcher from applied study, SPSS Package, 2018

From table (4.52), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.53): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.225	- 0.143	- 0.111
Sig.	0.421	0.612	0.693

Source: The researcher from applied study, SPSS Package, 2018

From table (4.53), it has been shown that according to the sig. values of the Chi square test (0.421), (0.612) and (0.693) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.54): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.789	2.485	0.457
MAE	5.596	5.828	5.152

Source: The researcher from applied study, SPSS Package, 2018

From table (4.54), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.12 Results obtained by MCAR: missing value 20% with variance 1.04

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.55): little's test results

little's test	
chi-square	0.689
Sig.	0.407

Source: The researcher from applied study, SPSS Package, 2018

From table (4.55), it shows the sig. value of little's test (0.407) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.56): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.84	999.76	999.82
variance	0.96	1.16	0.85

Source: The researcher from applied study, SPSS Package, 2018

From table (4.56), the results revealed that MI method has a mean (999.84) greater than the means of EM method and the regression method (999.82) and (999.76) respectively. With a variance of EM method (0.85) lower than the variances of MI method and the regression method (0.96) and (1.16) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.57): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	2.436	3	0.812	1.069	0.368
Within Groups	57.760	76	0.760		
Total	60.196	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.57), it shows the sig. value of the F-test (0.368) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.58): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.061	0.434	- 0.018

Source: The researcher from applied study, SPSS Package, 2018

From table (4.58), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.59): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.083	0.362	- 0.197
Sig.	0.728	0.117	0.406

Source: The researcher from applied study, SPSS Package, 2018

From table (4.59), it has been shown that according to the sig. values of the Chi square test (0.728), (0.117) and (0.406) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.60): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.168	0.275	0.021
MAE	6.723	6.758	6.674

Source: The researcher from applied study, SPSS Package, 2018

From table (4.60), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.13 Results obtained by MCAR: missing value 20% with variance 26.88

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.61): little's test results

little’s test	
chi-square	0.006
Sig.	0.937

Source: The researcher from applied study, SPSS Package, 2018

From table (4.61), it shows the sig. value of little's test (0.937) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.62): Statistics results

	MI method	*Regression method*	*EM method*
mean	999.71	999.14	999.44
variance	30.68	28.69	21.57

Source: The researcher from applied study, SPSS Package, 2018

From table (4.62), the results revealed that EM method has a mean (999.71) greater than the means of MI method and the regression method (999.44) and (999.14) respectively. With a variance of EM method (21.57) lower than the variances of the regression method and MI method (28.69) and (30.68) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.63): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	141.096	3	47.032	1.782	0.158
Within Groups	2005.596	76	26.389		
Total	2146.691	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.63), it shows the sig. value of the F-test (0.158) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.64): covariances results

	MI Method	*Regression method*	*EM method*
covariance	9.116	3.628	- 0.043

Source: The researcher from applied study, SPSS Package, 2018

From table (4.64), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.65): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.272	0.123	- 0.266
Sig.	0.246	0.604	0.257

Source: The researcher from applied study, SPSS Package, 2018

From table (4.65), it has been shown that according to the sig. values of the Chi square test (0.246), (0.604) and (0.257) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.66): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.514	1.326	0.007
MAE	7.171	7.109	6.669

Source: The researcher from applied study, SPSS Package, 2018

From table (4.66), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.14 Results obtained by MCAR: missing value 20% with variance 83.74

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.67): little's test results

little’s test	
chi-square	0.012
Sig.	0.915

Source: The researcher from applied study, SPSS Package, 2018

From table (4.67), it shows the sig. value of little's test (0.915) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.68): Statistics results

	MI method	*Regression method*	*EM method*
mean	1002.66	1002.58	1002.52
variance	76.61	92.95	68.03

Source: The researcher from applied study, SPSS Package, 2018

From table (4.68), the results revealed that MI method has a mean (1002.66) greater than the means of the regression method and MI method (1002.58) and (1002.52) respectively. With a variance of EM method (68.03) lower than the variances of MI method and the regression method (76.61) and (92.95) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.69): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	43.903	3	14.634	0.226	0.878
Within Groups	4914.883	76	64.670		
Total	4958.786	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.69), it shows the sig. value of the F-test (0.878) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.70): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 17.472	- 9.453	1.824

Source: The researcher from applied study, SPSS Package, 2018

From table (4.70), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.71): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.287	- 0.092	0.199
Sig.	0.219	0.701	0.401

Source: The researcher from applied study, SPSS Package, 2018

From table (4.71), it has been shown that according to the sig. values of the Chi square test (0.219), (0701) and (0.401) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.72): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.505	2.558	0.227
MAE	7.168	7.519	6.742

Source: The researcher from applied study, SPSS Package, 2018

From table (4.72), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.15 Results obtained by MCAR: missing value 30% with variance 1.04

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.73): little's test results

little's test	
chi-square	0.863
Sig.	0.353

Source: The researcher from applied study, SPSS Package, 2018

From table (4.73), it shows the sig. value of little's test (0.353) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.74): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.86	999.69	999.79
variance	1.06	1.12	0.75

Source: The researcher from applied study, SPSS Package, 2018

From table (4.74), the results revealed that MI method has a mean (999.86) greater than the means of EM method and the regression method (999.79) and (999.69) respectively. With a variance of EM method (0.75) lower than the variances of MI method and the regression method (1.06) and (1.12) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.75): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	3.366	3	1.122	1.500	0.218
Within Groups	86.764	116	0.748		
Total	90.129	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.75), it shows the sig. value of the F-test (0.218) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.76): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.257	0.006	- 0.002

Source: The researcher from applied study, SPSS Package, 2018

From table (4.76), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.77): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.257	0.006	- 0.022
Sig.	0.170	0.975	0.907

Source: The researcher from applied study, SPSS Package, 2018

From table (4.77), it has been shown that according to the sig. values of the Chi square test (0.170), (0.975) and (0.907) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.78): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.186	0.200	0.014
MAE	10.062	10.067	10.005

Source: The researcher from applied study, SPSS Package, 2018

From table (4.78), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.16 Results obtained by MCAR: missing value 30% with variance 26.88

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.79): little's test results

little’s test	
chi-square	1.595
Sig.	0.207

Source: The researcher from applied study, SPSS Package, 2018

From table (4.79), it shows the sig. value of little's test (0.207) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.80): Statistics results

	MI method	*Regression method*	*EM method*
mean	1000.19	1000.36	1000.22
variance	27.21	27.95	20.63

Source: The researcher from applied study, SPSS Package, 2018

From table (4.80), the results revealed that the regression method has a mean (1000.36) greater than the means of EM method and MI method (1000.22) and (1000.19) respectively. With a variance of EM method (20.63) lower than the variances of MI method and the regression method (27.21) and (27.95) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.81): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	52.493	3	17.498	1.032	0.381
Within Groups	1967.509	116	16.961		
Total	2020.001	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.81), it shows the sig. value of the F-test (0.381) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.82): covariances results

	MI Method	*Regression method*	*EM method*
covariance	-1.724	- 0.727	0.148

Source: The researcher from applied study, SPSS Package, 2018

From table (4.82), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.83): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.081-	- 0.032	0.111
Sig.	0.672	0.865	0.559

Source: The researcher from applied study, SPSS Package, 2018

From table (4.83), it has been shown that according to the sig. values of the Chi square test (0.672), (0.865) and (0.559) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.84): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.867	0.910	0.054
MAE	10.289	10.303	10.018

Source: The researcher from applied study, SPSS Package, 2018

From table (4.84), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

4.17 Results obtained by MCAR: missing value 30% with variance 83.74

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.85): little's test results

little's test	
chi-square	1.109
Sig.	0.292

Source: The researcher from applied study, SPSS Package, 2018

From table (4.85), it shows the sig. value of little's test (0.292) is greater than significant level (0.05) that mean the missing completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.86): Statistics results

	MI method	***Regression method***	***EM method***
mean	1001.93	1001.79	1002.03
variance	76.47	93.83	58.20

Source: The researcher from applied study, SPSS Package, 2018

From table (4.86), the results revealed that EM method has a mean (1002.03) greater than the means of EM method and the regression method (1001.93) and (1001.79) respectively. With a variance of EM method (58.20) lower than the variances of EM method and regression method (76.47) and (93.83) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.87): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	40.683	3	13.561	0.199	0.897
Within Groups	7916.175	116	68.243		
Total	7956.858	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.87), it shows the sig. value of the F-test (0.897) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.88): covariances results

	MI Method	*Regression method*	*EM method*
covariance	20.463	22.886	2.032

Source: The researcher from applied study, SPSS Package, 2018

From table (4.88), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.89): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.276	0.222	0.264
Sig.	0.140	0.239	0.158

Source: The researcher from applied study, SPSS Package, 2018

From table (4.89), it has been shown that according to the sig. values of the Chi square test (0.140), (0.239) and (0.158) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.90): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.450	2.017	0.150
MAE	10.483	10.672	10.050

Source: The researcher from applied study, SPSS Package, 2018

From table (4.90), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that ***EM method*** is more efficient than the other two methods.

Table (4.91): Shows summary of final results of Little's test in case the missing completely at random (MCAR).

No.	***Missing***	***Variances***	***Chi-square***	***Sig***	***Results***
1	5%	1.04	0.023	0.881	*Accept H_0*
2		26.88	1.057	0.304	
3		83.74	0.122	0.727	
4	10%	1.04	0.236	0.627	*Accept H_0*
5		26.88	0.015	0.904	
6		83.74	2.471	0.116	
7	15%	1.04	2.768	0.096	*Accept H_0*
8		26.88	0.058	0.809	
9		83.74	0.547	0.460	
10	20%	1.04	0.689	0.407	*Accept H_0*
11		26.88	0.006	0.937	
12		83.74	0.012	0.915	
13	30%	1.04	0.863	0.353	*Accept H_0*
14		26.88	1.595	0.207	
15		83.74	1.109	0.292	

Source: The researcher from applied study, SPSS Package, 2018

From above table, it shows the sig. values of little's test, all values are greater than significant level (0.05) that mean the missing completely at random (MCAR).

Table (4.92): Shows summary of final results of ANOVA in case the missing completely at random (MCAR).

No.	***Missing***	***Variances***	***F***	***Sig***	***Results***
1	5%	1.04	0.044	0.987	*Accept H_0*
2		26.88	2.126	0.137	
3		83.74	0.629	0.607	
4	10%	1.04	0.850	0.476	*Accept H_0*
5		26.88	0.059	0.981	
6		83.74	0.702	0.557	
7	15%	1.04	0.145	0.933	*Accept H_0*
8		26.88	1.866	0.146	
9		83.74	0.616	0.608	
10	20%	1.04	1.069	0.368	*Accept H_0*
11		26.88	1.782	0.158	
12		83.74	0.226	0.878	
13	30%	1.04	1.500	0.218	*Accept H_0*
14		26.88	1.032	0.381	
15		83.74	0.199	0.897	

Source: The researcher from applied study, SPSS Package, 2018

From above table, it shows the sig. value of the F-test, all values are greater than significant (0.05) that mean there is no statistical difference between means of estimated missing values.

Table (4.93): Shows summary of final results of variances and covariances matrix in case the missing completely at random (MCAR).

No.	***Missing***	***Variances***			***Covariances between***		
		MI method	***Reg. method***	***EM method***	***Generated data & MI***	***Generated data & Reg.***	***Generated data & EM***
1	5%	0.26	0.01	1.13	0.044	0.012	-0.466
2		5.16	19.95	0.02	-6.593	-3.564	0.088
3		13.14	106.95	1.80	21.927	28.905	0.796
4	10%	0.68	1.65	0.04	0.003	-0.457	0.002
5		18.74	29.88	0.07	1.188	0.972	-0.470
6		154.77	69.35	0.16	15.518	10.469	-0.337
7	15%	0.65	0.95	0.01	-0.488	0.485	0.004
8		14.68	28.90	0.01	7.903	-7.930	-0.016
9		48.00	92.65	3.13	17.611	-15.543	-2.225
10	20%	0.56	1.51	0.09	-0.061	0.434	-0.018
11		45.83	3519	0.01	9.116	3.628	-0.043
12		45.32	130.83	1.04	-17.472	-9.453	1.824
13	30%	1.04	1.20	0.06	-0.257	0.006	-0.002
14		22.56	24.87	0.09	-1.724	-0.727	0.148
15		63.05	122.04	0.68	20.463	22.886	2.032

Source: The researcher from applied study, SPSS Package, 2018

From the above table, it has been shows the variances of estimators of the three methods and covariances values between generated values and estimated missing values.

Table (4.94): Shows summary of final results of correlations in case the missing completely at random (MCAR).

No.	***Missing***	***correlations between***					
		Generated data & MI method		***Generated data & Reg. method***		***Generated data & EM method***	
		r	***Sig.***	***r***	***Sig.***	***r***	***Sig.***
1	5%	0.08	0.892	0.410	0.493	-0.427	0.473
2		-0.438	0.461	-0.120	0.847	0.334	0.583
3		0.596	0.288	0.276	0.654	0.059	0.926
4	10%	0.003	0.993	-0.361	0.305	0.039	0.914
5		0.063	0.863	0.041	0.911	-0.498	0.143
6		0.192	0.595	0.194	0.592	-0.146	0.687
7	15%	-0.550	0.034	0.452	0.091	0.238	0.393
8		0.300	0.277	-0.215	0.442	-0.297	0.282
9		0.225	0.421	-0.143	0.612	-0.111	0.693
10	20%	-0.083	0.728	0.362	0.117	-0.197	0.406
11		0.272	0.246	0.123	0.604	-0.266	0.257
12		-0.287	0.219	-0.092	0.701	0.199	0.401
13	30%	-0.257	0.170	0.003	0.975	-0.022	0.907
14		-0.081	0.672	-0.032	0.865	0.111	0.559
15		0.276	0.140	0.222	0.239	0.264	0.158

Source: The researcher from applied study, SPSS Package, 2018

From the above table, it has been shows Pearson (r)s (r) and the sig. values of the Chi square test between generated values and estimated missing values..

Table (4.95): Shows summary of final results of Std. Error of Mean and MAE in case the missing completely at random (MCAR).

No.	***Missing***	***MI method***		***Regression method***		***EM method***	
		S.E. mean	***MAE***	***S.E. mean***	***MAE***	***S.E. mean***	***MAE***
1	5%	0.227	1.742	0.013	1.671	0.476	1.825
2		1.015	2.005	1.997	2.332	0.018	1.673
3		1.621	2.207	4.625	3.208	0.599	1.866
4	10%	0.261	3.420	0.406	3.469	0.019	3.340
5		1.369	3.790	1.728	3.909	0.068	3.356
6		3.934	4.645	2.633	4.211	0.126	3.375
7	15%	0.208	5.069	0.251	5.084	0.004	5.001
8		0.989	5.330	1.388	5.463	0.002	5.001
9		1.789	5.596	2.485	5.828	0.457	5.152
10	20%	0.168	6.723	0.275	6.758	0.021	6.674
11		1.514	7.171	1.323	7.109	0.007	6.669
12		1.505	7.168	2.558	7.519	0.227	6.742
13	30%	0.186	10.062	0.200	10.067	0.014	10.005
14		0.867	10.289	0.910	10.303	0.054	10.018
15		1.450	10.483	2.017	10.672	0.150	10.050

4.18 Results obtained by Not - MCAR: missing value (10%) with variance (1.04)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.96): little's test results

little’s test	
chi-square	4.832
Sig.	0.028

Source: The researcher from applied study, SPSS Package, 2018

From table (4.96), it shows the sig. value of little's test (0.028) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.97): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.84	999.82	999.87
variance	1.07	1.12	0.96

Source: The researcher from applied study, SPSS Package, 2018

From table (4.97), the results revealed that EM method has a mean (999.87) greater than the means of MI method and the regression method (999.84) and (999.82) respectively. With a variance of EM method (0.96) lower than the variances of MI method and the regression method (1.07) and (1.12) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.98): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	1.325	3	0.442	0.483	0.696
Within Groups	32.913	36	0.914		
Total	34.237	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.98), it shows the sig. value of the F-test (0.696) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.99): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.647	- 0.072	- 0.021

Source: The researcher from applied study, SPSS Package, 2018

From table (4.99), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.100): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.633	- 0.062	- 0.248
Sig.	0.050	0.866	0.490

Source: The researcher from applied study, SPSS Package, 2018

From table (4.100), it has been shown that according to the sig. values of the Chi square test (0.050), (0.866) and (0.490) are greater than or equal significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.101): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.347	0.397	0.028
MAE	3.449	3.466	3.343

Source: The researcher from applied study, SPSS Package, 2018

From table (4.101), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.19 Results obtained by Not - MCAR: missing value (10%) with variance (26.88)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.102): little's test results

little’s test	
chi-square	5.267
Sig.	0.022

Source: The researcher from applied study, SPSS Package, 2018

From table (4.102), it shows the sig. value of little's test (0.022) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.103): Statistics results

	MI method	*Regression method*	*EM method*
mean	1000.15	1000.31	1000.27
variance	25.85	25.95	23.24

Source: The researcher from applied study, SPSS Package, 2018

From table (4.103), the results revealed that the regression method has a mean (1000.31) greater than the means of EM method and MI method (1000.27) and (1000.15) respectively. With a variance of regression method (23.24) lower than the variances of MI method and EM method (25.85) and (25.95) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.104): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	146.321	3	48.774	2.569	0.069
Within Groups	683.539	36	18.987		
Total	829.861	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.104), it shows the sig. value of the F-test (0.069) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.105): covariances results

	MI Method	*Regression method*	*EM method*
covariance	11.423	0.972	- 0.470

Source: The researcher from applied study, SPSS Package, 2018

From table (4.105), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.106): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.504	0.041	- 0.498
Sig.	0.138	0.911	0.143

Source: The researcher from applied study, SPSS Package, 2018

From table (4.106), it has been shown that according to the sig. values of the Chi square test (0.138), (0.911) and (0.143) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.107): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.642	1.728	0.068
MAE	3.881	3.909	3.356

Source: The researcher from applied study, SPSS Package, 2018

From table (4.107), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.20 Results obtained by Not - MCAR: missing value (10%) with variance (83.74)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.108): little's test results

little’s test	
chi-square	5.970
Sig.	0.015

Source: The researcher from applied study, SPSS Package, 2018

From table (4.108), it shows the sig. value of little's test (0.015) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.109): Statistics results

	MI method	*Regression method*	*EM method*
mean	1002.39	1001.62	1001.51
variance	82.34	70.08	80.57

Source: The researcher from applied study, SPSS Package, 2018

From table (4.109), the results revealed that MI method has a mean (1002.39) greater than the means of the regression method and EM method (10001.62) and (1001.51) respectively. With a variance of the regression method (70.08) lower than the variances of EM method and MI method (80.57) and (82.34) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.110): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	597.608	3	199.203	2.841	0.051
Within Groups	2524.088	36	70.114		
Total	3121.695	39			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.110), it shows the sig. value of the F-test (0.051) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.111): covariances results

	MI Method	*Regression method*	*EM method*
covariance	-10.706	- 0.498	18.196

Source: The researcher from applied study, SPSS Package, 2018

From table (4.111), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.112): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.135	- 0.049	0.169
Sig.	0.710	0.893	0.640

Source: The researcher from applied study, SPSS Package, 2018

From table (4.112), it has been shown that according to the sig. values of the Chi square test (0.710), (0.893) and (0.640) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.113): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	2.523	0.324	3.422
MAE	4.174	3.441	4.474

Source: The researcher from applied study, SPSS Package, 2018

From table (4.113), the results revealed that MAE of the regression method was lower than MAE of MI method and EM method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that the regression method is more efficient than the other two methods.

4.21 Results obtained by Not - MCAR: missing value (20%) with variance (1.04)

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.114): little's test results

little's test	
chi-square	5.192
Sig.	0.023

Source: The researcher from applied study, SPSS Package, 2018

From table (4.114), it shows the sig. value of little's test (0.023) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.115): Statistics results

	MI method	***Regression method***	***EM method***
mean	999.89	999.85	999.86
variance	0.95	0.97	0.78

Source: The researcher from applied study, SPSS Package, 2018

From table (4.115), the results revealed that MI method has a mean (999.89) greater than the means of EM method and the regression method (999.86) and (999.85) respectively. With a variance of EM method (0.78) lower than the variances of MI method and the regression method (0.95) and (0.97) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.116): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	0.557	3	0.186	0.223	0.880
Within Groups	63.201	76	0.832		
Total	63.758	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.116), it shows the sig. value of the F-test (0.880) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.117): covariances results

	MI Method	*Regression method*	*EM method*
covariance	0.115	- 0.120-	- 0.043

Source: The researcher from applied study, SPSS Package, 2018

From table (4.117), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.118): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.104	- 0.102	- 0.181
Sig.	0.664	0.669	0.446

Source: The researcher from applied study, SPSS Package, 2018

From table (4.118), it has been shown that according to the sig. values of the Chi square test (0.664), (0.669) and (0.446) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.119): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.213	0.227	0.045
MAE	6.738	6.742	6.682

Source: The researcher from applied study, SPSS Package, 2018

From table (4.119), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.22 Results obtained by Not - MCAR: missing value (20%) with variance (26.88)

i. Little's test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little's test.

Table (4.120): little's test results

little's test	
chi-square	4.199
Sig.	0.040

Source: The researcher from applied study, SPSS Package, 2018

From table (4.120), it shows the sig. value of little's test (0.040) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.121): Statistics results

	MI method	*Regression method*	*EM method*
mean	999.99	999.53	999.83
variance	30.21	30.97	24.43

Source: The researcher from applied study, SPSS Package, 2018

From table (4.121), the results revealed that MI method has a mean (999.99) greater than the means of EM method and the regression method (999.83) and (999.53) respectively. With a variance of EM method (24.43) lower than the variances of the regression method and MI method (30.97) and (30.21) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.122): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	56.641	3	18.880	0.968	0.413
Within Groups	1482.869	76	19.511		
Total	1539.510	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.122), it shows the sig. value of the F-test (0.413) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.123): covariances results

	MI Method	*Regression method*	*EM method*
covariance	-7.356	- 0.768	- 0.346

Source: The researcher from applied study, SPSS Package, 2018

From table (4.123), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.124): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.365	- 0.036	- 0.106
Sig.	0.114	0.881	0.657

Source: The researcher from applied study, SPSS Package, 2018

From table (4.124), it has been shown that according to the sig. values of the Chi square test (0.114), (0.881) and (0.657) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.125): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.222	1.300	0.198
MAE	7.074	7.100	6.733

Source: The researcher from applied study, SPSS Package, 2018

From table (4.125), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.23 Results obtained by Not - MCAR: missing value (20%) with variance (83.74)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.126): little's test results

little’s test	
chi-square	5.984
Sig.	0.014

Source: The researcher from applied study, SPSS Package, 2018

From table (4.126), it shows the sig. value of little's test (0.014) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.127): Statistics results

	MI method	***Regression method***	***EM method***
mean	1004.44	1002.98	1002.82
variance	96.54	93.33	72.32

Source: The researcher from applied study, SPSS Package, 2018

From table (4.127), the results revealed that MI method has a mean (1004.44) greater than the means of the regression method and EM method (1002.98) and (1002.82) respectively. With a variance of EM method (72.32) lower than the variances of the regression method and MI method (93.33) and (96.54) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.128): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	1291.208	3	430.403	5.880	0.001
Within Groups	5563.384	76	73.202		
Total	6854.592	79			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.128), it shows the sig. value of the F-test (0.001) is less than significant level (0.05) that mean there is statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.129): covariances results

	MI Method	*Regression method*	*EM method*
covariance	1.702	1.608	-5.194

Source: The researcher from applied study, SPSS Package, 2018

From table (4.129), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.130): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	0.022	0.019	- 0.205
Sig.	0.926	0.936	0.386

Source: The researcher from applied study, SPSS Package, 2018

From table (4.130), it has been shown that according to the sig. values of the Chi square test (0.926), (0.936) and (0.386) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.131): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	2.259	2.471	0.748
MAE	7.420	7.490	6.916

Source: The researcher from applied study, SPSS Package, 2018

From table (4.131), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.24 Results obtained by Not - MCAR: missing value (30%) with variance (1.04)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.132): little's test results

little’s test	
chi-square	3.916
Sig.	0.048

Source: The researcher from applied study, SPSS Package, 2018

From table (4.132), it shows the sig. value of little's test (0.048) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.133): Statistics results

	MI method	*Regression method*	*EM method*
mean	999.70	999.80	999.84
variance	1.21	1.15	0.74

Source: The researcher from applied study, SPSS Package, 2018

From table (4.133), the results revealed that EM method has a mean (999.84) greater than the means of the regression method and MI method (999.80) and (999.70) respectively. With a variance of EM method (0.74) lower than the variances of the regression method and MI method (1.15) and (1.21) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.134): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	4.242	3	1.414	1.472	0.226
Within Groups	111.426	116	0.961		
Total	115.668	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.134), it shows the sig. value of the F-test (0.226) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.135): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.127	0.164	0.002

Source: The researcher from applied study, SPSS Package, 2018

From table (4.135), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.136): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.106	0.139	0.090
Sig.	0.579	0.464	0.635

Source: The researcher from applied study, SPSS Package, 2018

From table (4.136), it has been shown that according to the sig. values of the Chi square test (0.579), (0.464) and (0.635) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.137): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.220	0.215	0.005
MAE	10.073	10.072	10.002

Source: The researcher from applied study, SPSS Package, 2018

From table (4.137), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.25 Results obtained by Not - MCAR: missing value (30%) with variance (26.88)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.138): little's test results

little’s test	
chi-square	7.740
Sig.	0.005

Source: The researcher from applied study, SPSS Package, 2018

From table (4.138), it shows the sig. value of little's test (0.005) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.139): Statistics results

	MI method	*Regression method*	*EM method*
mean	998.99	999.51	999.22
variance	29.72	33.38	20.78

Source: The researcher from applied study, SPSS Package, 2018

From table (4.139), the results revealed that the regression method has a mean (999.51) greater than the means of EM method and MI method (999.22) and (998.99) respectively. With a variance of EM method (20.78) lower than the variances of MI method and the regression method (29.72) and (33.38) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.140): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	129.734	3	43.245	1.801	0.151
Within Groups	2785.110	116	24.010		
Total	2914.843	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.140), it shows the sig. value of the F-test (0.151) is greater than significant level (0.05) that mean there is no statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.141): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 3.015	5.975	- 1.475

Source: The researcher from applied study, SPSS Package, 2018

From table (4.141), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.142): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.121	0.199	- 0.325
Sig.	0.525	0.291	0.080

Source: The researcher from applied study, SPSS Package, 2018

From table (4.142), it has been shown that according to the sig. values of the Chi square test (0.525), (0.291) and (0.080) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.143): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	1.009	1.212	0.183
MAE	10.336	10.404	10.061

Source: The researcher from applied study, SPSS Package, 2018

From table (4.143), the results revealed that MAE of EM method was lower than MAE of MI method and the regression method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

4.26 Results obtained by Not - MCAR: missing value (30%) with variance (83.74)

i. Little’s test for randomness

To test the first hypothesis; is the missing completely at random or not? We calculated the sig. value of the little’s test.

Table (4.144): little's test results

little’s test	
chi-square	46.785
Sig.	0.00

Source: The researcher from applied study, SPSS Package, 2018

From table (4.144), it shows the sig. value of little's test (0.00) is less than significant level (0.05) that mean the missing is not completely at random.

ii. Descriptive statistics of the three methods

We calculate means and variances depending on the completed values of variables, to know is there ostensibly differences.

Table (4.145): Statistics results

	MI method	*Regression method*	*EM method*
mean	1002.23	1000.76	1001.34
variance	86.55	74.29	56.13

Source: The researcher from applied study, SPSS Package, 2018

From table (4.145), the results revealed that MI method has a mean (1002.23) greater than the means of EM method and the regression method (1001.34) and (1000.76) respectively. With a variance of EM method (56.13) lower than the variances of the regression method and MI method (74.29) and (86.55) respectively.

iii. ANOVA for the estimated means

To test the second hypothesis; is there a statistically significant difference between means or not? We calculated the sig. value of the F-test.

Table (4.146): ANOVA results

	Sum of Squares	*df*	*Mean Square*	*F*	*Sig.*
Between Groups	534.994	3	178.331	2.957	0.035
Within Groups	6995.097	116	60.303		
Total	7530.091	119			

Source: The researcher from applied study, SPSS Package, 2018

From table (4.146), it shows the sig. value of the F-test (0.035) is less than significant level (0.05) that mean there is statistical difference between means of the generated values and estimated missing values.

iv. Covariance matrix.

We calculated covariances between generated values and estimated missing values.

Table (4.147): covariances results

	MI Method	*Regression method*	*EM method*
covariance	- 0.064	- 1.186	- 21.090

Source: The researcher from applied study, SPSS Package, 2018

From table (4.147), it shows the covariances values between generated values and estimated missing values.

v. Correlation matrix

To test the third hypothesis; is the correlation significant or not? We calculated the sig. values of the Chi square test.

Table (4.148): Correlations results

	MI method	*Regression method*	*EM method*
Pearson (r)	- 0.027	- 0.016	- 0.234
Sig.	0.886	0.931	0.213

Source: The researcher from applied study, SPSS Package, 2018

From table (4.148), it has been shown that according to the sig. values of the Chi square test (0.886), (0.931) and (0.213) are greater than significant level (0.05) that means, the correlation is not significant.

vi. Std. Error of Mean and MAE

We calculated std. error of mean and MAE depend on generated values and estimated missing values.

Table (4.149): Std. Error of Mean and MAE results

	MI method	*Regression method*	*EM method*
S.E. mean	0.046	1.424	1.782
MAE	10.015	10.475	10.594

Source: The researcher from applied study, SPSS Package, 2018

From table (4.149), the results revealed that MAE of MI method was lower than MAE of the regression method and EM method. These results were consistent with values of S.E. mean. Hence, based on those results, we concluded that EM method is more efficient than the other two methods.

Table (4.150): Shows summary of final results of Little's test in case the missing is not - completely at random (Not - MCAR)

No.	*Missing*	*Variances*	*Chi-square*	*Sig*	*Results*
1	10%	1.04	4.832	0.028	*Reject H_0*
2		26.88	5.267	0.022	
3		83.74	5.970	0.015	
4	20%	1.04	5.192	0.023	*Reject H_0*
5		26.88	4.199	0.040	
6		83.74	5.984	0.014	
7	30%	1.04	3.916	0.048	*Reject H_0*
8		26.88	7.740	0.005	
9		83.74	46.785	0.000	

Source: The researcher from applied study, SPSS Package, 2018

From above table, it shows the sig. values of little's test, all values are is less than significant level (0.05) that mean the missing is not completely at random (Not - MCAR).

Table (4.151): Shows summary of final results of ANOVA in case the missing is not - completely at random (Not - MCAR)

No.	***Missing***	***Variances***	***F***	***Sig***	***Results***
7	15%	1.04	0.483	0.696	Accept H0
8		26.88	2.569	0.069	Accept H0
9		83.74	2.841	0.051	Accept H0
10	20%	1.04	0.223	0.880	Accept H0
11		26.88	0.968	0.413	Accept H0
12		83.74	5.880	0.001	Reject H0
13	30%	1.04	1.472	0.226	Accept H0
14		26.88	1.801	0.151	Accept H0
15		83.74	2.957	0.035	Reject H0

Source: The researcher from applied study, SPSS Package, 2018

From above table, it shows the sig. values of the F-test, values are greater than significant (0.05) that mean there is no statistical difference between means of the three estimation methods. Except two values ( ) and ( ) are less than significant (0.05) that mean there is statistical difference between means of the three estimation methods.

Table (4.152): Shows summary of final results of variances and covariances matrix in case the missing is not - completely at random (Not - MCAR)

No.	***Missing***	***Variances***			***Covariances between***		
		MI method	***Reg. method***	***EM method***	***Generated data & MI***	***Generated data & Reg.***	***Generated data & EM***
1	5%	1.20	1.58	0.01	-0.647	-0.072	-0.021
2		26.97	29.88	0.05	11.423	0.972	-0.470
3		63.66	1.05	117.11	-10.706	-0.498	18.196
4	10%	0.91	1.03	0.04	0.115	-0.120	-0.043
5		29.89	33.78	0.79	-7.356	-0.768	-0.346
6		102.11	122.11	11.18	1.702	1.608	-5.194
7	15%	1.45	1.39	0.00	-0.127	0.164	0.002
8		30.55	44.07	1.01	-3.015	5.975	-1.475
9		0.06	60.86	95.24	-0.064	-1.186	-21.090

Source: The researcher from applied study, SPSS Package, 2018

From the above table, it has been shows the variances of estimators of the three methods and covariances values between parameters of generated data and estimators of the three methods

Table (4.153): Shows summary of final results of correlations in case the missing is not - completely at random (Not - MCAR)

No.	Missing	correlations between					
		Generated data & MI method		Generated data & Reg. method		Generated data & EM method	
		r	Sig.	r	Sig.	r	Sig.
1	5%	-0.633	0.050	-0.062	0.866	-0.248	0.490
2		0.540	0.138	0.041	0.911	-0.498	0.143
3		0.135	0.710	-0.049	0.893	0.169	0.640
4	10%	0.104	0.664	-0.102	0.669	-0.181	0.446
5		-0.365	0.114	-0.036	0.881	-0.106	0.657
6		0.022	0.926	0.019	0.936	-0.205	0.386
7	15%	-0.106	0.579	0.136	0.464	0.090	0.635
8		-0.121	0.525	0.199	0.291	-0.325	0.080
9		0.027	0.886	-0.016	0.931	-0.234	0.213

Source: The researcher from applied study, SPSS Package, 2018

From the above table, it has been shows Pearson correlations (r) and the sig. values of the Chi square test between parameters of generated data and estimators of the three methods.

Table (4.154): Shows summary of final results of Std. Error of Mean and MAE in case the missing is not - completely at random (Not - MCAR)

No.	***Missing***	***MI method***		***Regression method***		***EM method***	
		S.E. mean	***MAE***	***S.E. mean***	***MAE***	***S.E. mean***	***MAE***
1	10%	0.347	3.449	0.397	3.466	0.028	3.343
2		1.642	3.881	1.728	3.909	0.068	3.356
3		2.523	4.174	0.324	3.441	3.422	4.474
4	20%	0.213	6.738	0.227	6.742	0.045	6.682
5		1.222	7.074	1.300	7.100	0.198	6.733
6		2.259	7.420	2.490	7.490	0.748	6.916
7	30%	0.220	10.073	0.215	10.072	0.005	10.002
8		1.009	10.336	1.212	10.404	0.183	10.061
9		0.046	10.015	1.424	10.475	1.782	10.594

Chapter VI

Results and Recommendations

5.1 Introduction:

In this chapter the results and recommendations of the study are presented with reference to the aim of the study, which was a comparative study of the Multiple Imputation (MI) method of estimation against two other methods; the Regression Imputation of estimation and the Expectation-maximization (EM) algorithm of estimation, for estimating missing data.

5.2 Results:

Firstly: Results obtained by the first case; the missing is completely at random (MCAR).

[1] Results obtained by Little's MCAR test; all values of the sig. values of little's test are greater than significant level (0.05) that mean the missing is completely at random.

[2] Results obtained by descriptive statistics; the results revealed that there is no ostensibly differences between means and variances.

[3] Results obtained by ANOVA; the results revealed, all values of the sig. values of the F-test are greater than significant level (0.05) that mean there is no significant difference between means of the three estimation methods.

[4] Results obtained by the correlation; the results revealed, (98%) of values of the sig. values of the Chi-square test are greater than the significant level (0.05) that mean the correlations are not significant.

Note: if the sig. value of the Chi-square test is greater than the significance level (0.05), there is inconclusive evidence regarding the significance of the association between the variables.

[5] Results obtained by MAE; This study compared the three previous methods, and based on the results, we concluded there is no statistical difference between means of estimators of the three estimation methods; the results are shown as follows:

i. (98%) of the results shown that the Expectation-maximization (EM) method of estimation is more efficient than the other two methods.

ii. (2%) of the results shown that the Regression Imputation method of estimation is better than the other two methods in producing more efficient estimates.

Secondly: Results related by the second case; the missing is not - completely at random (Not-MCAR).

[6] Results obtained by Little's MCAR test; all values of the sig. values of little's test are less than significant level (0.05) that mean the missing is not - completely at random (Not-MCAR).

[7] Results obtained by descriptive statistics; the results revealed that there is no ostensibly differences between means and variances.

[8] Results obtained by ANOVA; shown that the (98%) of the sig. values of the F-test are greater than significant level (0.05) that mean there is no statistical difference between means of the estimation methods. And the remaining the sig. values of the F-test shown there is a significant difference between means of the three estimation methods.

[9] Results obtained by the correlation; the results revealed, (96%) of values of the sig. values of the Chi-square test are greater than the significant level (0.05) that mean the correlations are not significant.

Note: if the sig. value of the Chi-square test is greater than the significance level (0.05), there is inconclusive evidence regarding the significance of the association between the variables.

[10] Results obtained by MAE; This study compared the three previous methods, and based on the results, we concluded there is no statistical difference between means of estimators of the three estimation methods; the results are shown as follows:

i. (90%) of the results shown that the Expectation-maximization (EM) method of estimation is more efficient than the other two methods.

ii. (4%) of the results shown that the Multiple Imputation (MI) method of estimation is more efficient than the other two methods.

5.3 Recommendations:

This study is recommended the following:

[1] More attention should be paid to the missing data in the design and performance of the studies.

[2] Don't ignoring missing data, because it can reduce the statistical power of a study and can produce biased estimates.

[3] Application the Multiple Imputation (MI) method of estimation, the Regression Imputation of estimation and the Expectation-maximization (EM) algorithm of estimation in moderate and large sample sizes to guarantee to produce more efficient estimates.

[4] Application the above methods in the different probability distributions which are a very important area in statistics.

[5] Using the Expectation-maximization (EM) algorithm of estimation because it is better than the other two methods in producing more efficient estimates.

References

[1] Sharon L. Lohr 2010 Sampling: Design and Analysis, Second Edition, Arizona State University, 330-339.

[2] Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009; 60:549-576.

[3] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367:1355–1360

[4] O'Neill RT, Temple R. The prevention and treatment of missing data in clinical trials: an FDA perspective on the importance of dealing with it. Clin Pharmacol Ther. 2012;91:550–554.

[5] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May 24. doi: 10.4097/kjae.2013.64.5.402

[6] van Buuren, S. (2012). Flexible imputation of missing data. Boca Raton, FL: CRC Press

[7] Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clinical Trials, 1, 368-376.

[8] Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74, 525-556

[9] Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. Psychological Reports, 99, 675-680.

[10] Nakagawa, S., & Hauber, M. E. (2011). Great challenges with few subjects: Statistical strategies for neuroscientists. Neuroscience and Biobehavioral Reviews, 35, 462-473.

[11] McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). Missing data: a gentle introduction. New York, NY: The Guilford Press.

[12] Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: The dangers of ignoring missing data. Trends in Ecology & Evolution, 23, 592-596.

[13] Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. C. Clogg (Ed.), Sociological methodology, 1987 (pp. 71–103). San Francisco: Jossey-Bass.

[14] Tanner, M. A., & Wing, H. W. (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82, 528-540.

[15] Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley

[16] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, NY: J. Wiley & Sons.

[17] Nakagawa, S., & Freckleton, R. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. Behavioral Ecology and Sociobiology, 6, 103-116.

[18] Rosenbaum, P. R., and Rubin, D. B. (1983).The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55.

[19] Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, N.J.: Wiley.

[20] Data Analysis with SPSS: A First Course in Applied Statistics (4th Edition) by Stephen Sweet and Karen Grace-Martin

[21] SPSS - IBM Knowledge Center (https://www.ibm.com/support/knowledgecenter/en/SSLVMB_sub/spss/tutorials/mva_describe_rerun_mcartest.html)

[22] Handling missing data. Martijn Heymans. V2.0: 12 May 2015 (http://www.emgo.nl/kc/handling-missing-data/)

[23] Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998. Characterizing Selection Bias Using Experimental Data. Econometrica 66, 1017–1098.

[24] DeSarbo S, Green PE, Carroll JD. An alternating least-squares procedure for estimating missing preference data in product-concept testing. Decision Sciences . 1986;17:163–185.

[25] Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of School Psychology, 48, 5-37.

[26] Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. Psychological Methods, 11, 323-343.

[27] Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. Annual Review of Psychology, 60, 549-576

[28] Graham, J. W. (2012). Missing data : analysis and design. New York: Springer

[29] Rhemtulla, M., & Little, T. D. (2012). Planned Missing Data Designs for Research in Cognitive Development. Journal of Cognition and Development, 13, 425-438.

[30] Enders, C. K. (2010). Applied missing data analysis. New York: Guilford Press.

[31] Shinichi Nakagawa - Missing data: mechanisms, methods, and messages

[32] Briggs, A., Clark, T., Wolstenholme, J., Clarke, P., 2003. Missing.... presumed at random: cost-analysis of incomplete data. Health Economics 12, 377–392

[33] Nakai M and Weiming Ke., 2011. Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. Vol. 5, no.1, 1-13.

[34] Kim JO, Curry J. The treatment of missing data in multivariate analysis. Sociol Methods Res. 1977;6:215–241

[35] Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annu Rev Psychol 60, 549–576.

[36] Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.

[37] Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J Roy Stat Soc Series B 1960; 22(2): 302–306.

[38] Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley & Sons, New York, 1987.

[39] Brick JM, Kalton G. Handling missing data in survey research. Stat Meth Med Res 1996; 5: 215–238.

[40] Hyun Kang. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May; 64(5): 402–406

[41] Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable non-response. J Am Stat Assoc 1986; 81: 366–374.

[42] Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987 .

[43] Schafer JL. Multiple imputation: a primer. Stat.

[44] Hamer RM, Simpson PM. Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials. Am J Psychiatry. 2009;166:639–641.

[45] Panel on Missing Data in Clinical Trials. The prevention and treatment of missing data in clinical trials. 2nd ed. Washington DC: National Academies Press; 2010. pp. 107–114.

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

[47] SAS Institute, 2005. Multiple Imputation for Missing Data: Concepts and New Approaches.

[48] Little R J A and Rubin D B 1987 Statistical Analysis with Missing Data (New York: John Wiley & Son Inc.)

[49] Dempster A P, Laird N M, and Rubin D B 1977 Journal of the Royal Statistical Society Series B 39 (1) 1-38

[50] Little R J A and Rubin D B 2002 Statistical Analysis with Missing Data Second Edition (Hoboken, New Jersey: John Wiley & Son Inc.)

[51] Watanabe M and Yamaguchi K 2004. The EM Algorithm and Related Statistical Models (New York: Marcel Dekker, Inc.)

[52] Enders, C.K., Bandalos, D.L., 2001. The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Structural Equation Modeling: A Multidisciplinary Journal 8, 430–457.

[53] Allison P., 2003. Handling Missing Data by Maximum Likelihood. SAS Global Forum 2012. Development (Version 9.0)

[54] Kong A., Liu KJ and Hung Wong W., 1994. Sequential Imputations and Bayesian Missing Data Problems. Journal of the American Statistical Association 89, 425, 278-28.

[55] Panel on Missing Data in Clinical Trials. The prevention and treatment of missing data in clinical trials. 2nd ed. Washington DC: National Academies Press; 2010. pp. 107–114.