Evaluation of Expectation Maximization and Full Information Maximum Likelihood as a Handling techniques for Missing Data

Alaa A. Abd Elmegaly

Lecturer of Statistics, Higher Institute of Advanced Administrative Sceinces & Computers, Ministry of Higher Education and Scientific Research, Egypt.

Abstract: *Data analysis and its importance cannot be denied in our life. Most of the time the researchers face the problem of missing data in their analysis. This has a high effect on the reliability of the study results. This study aims to evaluate each Expectation-Maximization (**EM**) and Full Information Maximum Likelihood (**FIML**) as two procedures for handling missing data. Effect size has been used as a criterion to determine which method is better. Simulation has been done to compare these methods at different sample sizes and different mechanisms for missing data by the **R** program. The results indicated that the two procedures were good, but **FIML** was more reliable than **EM** in the case of small sample sizes and large missing percentages, so **FIML** has been considered as the better method in small sample sizes and large missing percentages. It contains values of effect size more reliable than the original data.*

Keywords: *Effect Size; EM; FIML; Missing Data.*

1. Introduction

Missing data is one of the most important problems that researchers face during their analysis of their research data. It affects the quality of the data and the quality of statistical analysis. This missing data appears in different patterns and different mechanisms. Statisticians have tried to develop and find better methods to estimate the missing values that are lost in various scientific research data, each of these methods has its conditions and assumptions according to the pattern and mechanism of missing data (**Mcknight et al, 2007**).

Despite the great development witnessed by the problem of missing data in terms of developing methods for its estimation and treatment, the traditional methods that depend on the method of deletion are still used due to the ease of their procedures. And the new methods need a lot of time to be applied instead of the old ones (**Enders**, 2010).

This was warned by the American Society of Psychological and Educational Sciences and called for the use of developed statistical methods that take into account the conditions under which data are lost (**Peng et al, 2006**). The presence of the missing data when using multivariate statistical analysis methods is more influential and dangerous to the results of the analysis. Particularly for multiple regression models, where lost data affect the efficiency and predictability of the regression model, resulting in the inaccuracy and credibility of the researcher's results. Much research refers to the problem of missing data and methods of handling it. The demonstration of loss patterns and mechanisms also has huge importance from researchers. More researchers focused on comparing different treatment methods and providing proposed methods of estimating missing values. The researchers compare the proposed modern methods and the traditional ones to determine the effectiveness of the proposed method.

2. Problem of the Study

The presence of the missing data problem in multiple regression models with different types of loss and at different rates may result in wrong conclusions and biased estimates as a result of the deletion of part of the sample data. Although early pioneers have contributed to the development and finding of methods to handle missing data, and each method has its assumptions in use, its pattern, and mechanism. Researchers find it difficult to determine the better method to deal with missing data, all methods of dealing with it require some kind of additional assumptions and at the same time, there is general agreement on their strengths and weaknesses. Hence the need for such a study, which compares two methods to handle the missing data at different sample sizes, different missing percentages, different mechanisms of missing data. The main question of the study can therefore be formulated as follows:

Which the best method to handle missing data, Expectation-Maximization (EM) algorithm or full information (FIML) in case of taking effect size ES to compare them?

3. Objective of the Study

The study aimed to evaluate the methods of the Expectation-Maximization algorithm (**EM**) and Full Information Maximum Likelihood (**FIML**)in case of effect size has been used as a criterion to compare between the two methods in estimating the missing values in the multiple regression model, through the use of simulation.

4. Importance of the Study

The study derives its theoretical importance by talking about the subject of missing data, methods of handling them, between different patterns of missing, and its mechanisms as well as their contribution together with many studies in this area to evaluate the methods of algorithms that maximize expectation and full information of the maximum likelihood to handle missing data. The second field is the applied importance of providing researchers with the information needed to handle missing data to contribute to more accurate estimates of missing data that can be depended upon to replace missing data.

5. Theoretical Framework

5.1. Meaning of missing data

Data was left unanswered due to failure to obtain data from the community or the United Nations Development Organization (**Abdulrahman**&**Attia**, 2013). (**Little**&**Rubin**, 2002) said that the missing data is considered to be incomplete observation data, and it hinders statistical analysis.

5.2. Pattern of missing data

Knowing the pattern and mechanism of missing data helps the researcher to choose the appropriate statistical treatment to estimate the parameters of the model, especially since there are statistical treatments that are suitable for special patterns of missing data and the steps are clear and easy to apply. While there are statistical treatments suitable for the general pattern and are more complex than the treatments of special patterns (**Little**, 1992). The pattern of missing data refers to the review of the observed and missing values within the data set (**Enders**, 2010). That is, the missing data pattern simply describes the location of those data gaps and does not explain why the data was lost. The missing data takes the form of the following patterns:

- **The first pattern**(the univariate pattern): in this pattern, missing data loses data in one independent variable, and the rest of the variables are complete (**Little**&**Rubin**, 2002).
- **The second pattern**(Multivariate missing data pattern): In this type of missing data, data is lost in more than one independent variable, so that the number of cases of loss is equal in all independent variables that contain missing data (**Little**&**Rubin**, 2002).
- **The third pattern** is the Monotone pattern: In this pattern of missing data, the loss of the data is in an orderly shape for some independent variables. So that the variable with the largest number of missing cases is the first of the variables. Then comes the variable that contains the second order of the largest number of the cases of loss. And so on for the rest of the independent variables (**Schafer**, 1997).
- **The Fourth pattern**(General pattern): In this type of missing data, the loss of data does not take a specific form, but the lost data are random and scattered. Among the other missing data patterns, the general pattern is the dominant one (**Graham**, 2012).
- **Fifth pattern**(File matching pattern): This pattern of missing data occurs in two independent variables only, so that there is no common missing data between the two variables (**Molenberghs** et al., 2015).

5.3. Mechanism of missing data

The missing data mechanism refers to the relationship between the measured variables and the probability of data loss (**Enders**, 2010). Although the mechanisms of missing data do not provide us with a causal explanation for the loss of data, they do provide us with a general mathematical relationship between the data and the process of losing it. Understanding these mechanisms and determining their nature helps in choosing the appropriate method in which the missing data will be treated. Therefore, the mechanisms of data loss are:

- **Missing Completely At Random Mechanism MCAR**: is said to be lost at random if the missing data does not depend on the values of the independent variable containing these missing values or any other independent variable in the database.
- **Missing At Random Mechanism MAR**: It is said that the data is lost randomly if the missing data does not depend on the values of the independent variable containing these missing values, but depends on the other independent variable(s).
- **Not Missing At Random Mechanism NMAR**: It is said that data is lost non-randomly if the missing data in a particular independent variable depends on the values of the same variable.

5.4. Methods of handling missing data

- **Leastwise Deletion (LD):** This method is based on deleting the case that has missing data in at least one variable, and it uses only cases with complete data on all variables. So it is also called Complete Case Analysis (**Dudin**&**Muhammad**, 2013). This method is valid in the case of few missing data and when the sample size is large, allowing these cases to be deleted. It is also used in the case of completely random missing data, where the lack of randomness leads to bias results (**Rizkallah**, 2002). One of the problems with this method is that it may produce biased standard estimates, meaning that a bias may occur in the results of this method because the case, in effect, did not represent the entire population (**Nakai**&**Ke**, 2011).

- **Pairwise Deletion (PD):** In this method, the item that contains missing data is deleted from the variable used in the analysis only, but in the rest of the variables, this item is not deleted. Somewhat on the collected data and did not delete it at all, and the disadvantage of this method is that the samples vary from one analysis to another as a result of deleting some cases once and not deleting them again (**Hoyt**&**Kramer**, 2016).

- **Mean Imputation**: In this method, the missing value in a variable is replaced with the average of the available data in that variable. And here the researcher calculates the mean of the variable that contains missing data, using all the data available to him, and then puts the resulting average value in the cells null to replace each missing value (**Al-Qahtani**&**Saad**, 2015).

- **Last Observation Carried Forward**: This method is limited to handle longitudinal data, and this method is used to replace each missing value with the last observed value in the same topic. Whenever values are missing, they are replaced with the last observed value. This method is relatively rarely used in the behavioral and social sciences, and despite the frequent use of this method in medical studies and clinical trials, many experimental studies have demonstrated the weakness of this method in dealing with longitudinal missing data (**Molenberghs et al**, 2015).

- **Hot Deck Imputation**: Hot Deck estimation for missing data is a set of techniques that are used to impute missing data with similar degrees of similar observations for the missing observations. the most common application of Hot Deck estimation is to fill in the missing values of the observations with random values from Other observations are similar to observations with incomplete data in the same dimensions and items (**Little**&**Rubin**, 2002).

- **Cold Deck Imputation**: This method depends on replacing the missing value with a fixed value chosen from a source other than the current database, where the missing value is replaced with another value from another source, such as taking the value from a previous version of the same survey. The researcher should ensure that the value derived from previous research is more correct and accurate than any value derived internally. Unfortunately, the values that can be used are not always available using the Cold Deck method (**Al-Hamami**&**Hussein**, 2007).

- **Imputation Using Regression Analysis**: This method depends on the use of regression for the purpose of estimating and predicting the value of a variable in terms of another variable. This method has many disadvantages. It may lead to an adjustment in the data more than necessary, as well as assume the existence of a relationship Correlation with the variable that contains missing data. If the correlation is weak, the missing data cannot be predicted. Despite its shortcomings, it is suitable in cases where the missing data is spread moderately with a sufficient correlation between the variables (**Abu Alam**&**Mahmoud,** 2007).

- **Multiple Imputation (MI)** method: The problem of imputation using the mean and using regression is that: it underestimates the variance in the data. The solution to this problem is to use several imputations to estimate each of the variables that contain missing values, and the different imputations are distinguished from each other only by adding a small amount of random error. Then end up with many datasets with different imputations for the missing values can be done. Each of these data sets is analyzed as if they were full-data normal sets. As a final step, the different model estimates are combined taking into account the regression error. So the multiple imputation techniques go through the following three steps:

- Create multiple complete datasets using multiple imputation.
- Carrying out the required analysis on each set of data sets.
- Combining different analyzes (**Blanch**&**Niles**, 2017) and (**Lee**&**Shi**, 2021).

- **The Expectation-Maximization Algorithm (EM Algorithm)**: it is a method for finding the maximum likelihood estimation of the model parameters. It is an iterative algorithm that produces the maximum probability of estimates for the missing data. This method is characterized by its simplicity and comprehensiveness of the cases that can be processed. This method depends on imputing for missing data through estimates obtained from a repeated estimation process to take advantage of all available information from complete and incomplete cases. Where the missing data has been predicted based on initial estimates of the model parameter values, and then these predictions are used to modify the parameter values, and the prediction process is repeated using the modified model parameters until the parameters approach the maximum likelihood estimates (**Davey&Savla**, 2010)

- Full Information Maximum Likelihood (**FIML**):**FIML** method is one of the most famous methods for estimating parameters in structural modeling. it is a basic method for estimating mean and variances based on incomplete data and assuming that they are lost completely or randomly. In this method, the missing data is estimated using only the entire case data without depending on the data of other cases (**Penget al**, 2006). This method also allows takinginformation of all available data without resorting to any of the missing imputation methods (**Blanch**, 2017) and (**Lee&Shi**, 2021).

6. Methodology of the Search

Simulations have been used to generate random numbers that follow a standard normal distribution to make a regression model consists of three independent variables using **R** program version 4.0.2. Provided that all conditions for multiple regression were achieved in the model. Then, percentage of loss for these data was done at 10%, 20%, 30%, respectively. Sample sizes of 10, 20, 30, 50 and 100 have been used to evaluate **EM** and **FIML** methods. Finally, the effect size criterion was used to compare the two methods.

7. Steps to Conduct the Research:

- Generating random numbers that follow a standard normal distribution with sample sizes of 10, 20, 30, 50 and 100
- Making data loss with different percentages and mechanisms
- Use of two imputation methods to make up for the missing values (**EM** and **FIML**)
- Using the effect size criterion to compare the previous Imputation methods
- Repeat the previous steps 1000 times

8. Results

The results of the main study question, which was, "**Which methods are better to treat missing data, the Expectation-Maximization algorithm (EM) or the full information of the greatest possibility (FIML) in the presence of an effect size criterion for comparison between them**"?

The following table shows the results obtained from the simulation of the two methods (**EM**) and (**FIML**) in case of the effect size was the criterion for comparison between them as follows:

Table (1): effect size for EM and FIML at different sample sizes and different Percentage of missing data for MAR, NMAR, and MCAR mechanisms

N	Percentage of missing data	Method of estimation	MAR	NMAR	MCAR
10	0.1	ORIGINAL DATA	2.3757	2.3886	2.3175
		EM	3.3797	2.5959	3.0354
		FIML	1.9515	1.5966	2.0532
	0.2	ORIGINAL DATA	2.3263	2.3201	2.3829
		EM	3.3839e+14	3.5876	4.4476
		FIML	1.4850	1.1426	1.7838
	0.3	ORIGINAL DATA	-	2.3566	2.3442
		EM	-	8.8244	5.4637
		FIML	-	0.9554	1.5833
20	0.1	ORIGINAL DATA	1.9908	1.9528	1.9831
		EM	2.4105	1.8682	2.3194
		FIML	1.5351	1.1541	1.6733
	0.2	ORIGINAL DATA	1.9711	1.9728	1.9826
		EM	3.5750	1.8497	2.8763
		FIML	1.0405	0.7920	1.4640
	0.3	ORIGINAL DATA	-	1.9892	1.9717
		EM	-	2.1667	3.5607
		FIML	-	0.6017	1.2523

30	**0.1**	**ORIGINAL DATA**	1.9031	1.8711	1.8695
		EM	2.2739	1.7242	2.1383
		FIML	1.3867	1.0442	1.5810
	0.2	**ORIGINAL DATA**	1.8844	1.8896	1.8792
		EM	3.1396	1.5565	2.5167
		FIML	0.8991	0.6905	1.3614
	0.3	**ORIGINAL DATA**	-	1.8573	1.8940
		EM	-	1.4564	3.3273
		FIML	-	0.4941	1.1833
50	**0.1**	**ORIGINAL DATA**	1.8189	1.8154	1.8326
		EM	2.1685	1.6146	2.0733
		FIML	1.2463	0.9633	1.5339
	0.2	**ORIGINAL DATA**	1.8074	1.8122	1.8242
		EM	2.9060	1.3854	2.4290
		FIML	0.7925	0.6218	1.2944
	0.3	**ORIGINAL DATA**	1.8297	1.8128	1.7944
		EM	12.6823	1.1958	2.8156
		FIML	0.4352	0.4284	1.1098
100	**0.1**	**ORIGINAL DATA**	1.7678	1.7698	1.7741
		EM	2.1086	1.5397	1.9964
		FIML	1.1448	0.9153	1.4767
	0.2	**ORIGINAL DATA**	1.7756	1.7826	1.7653
		EM	2.8068	1.3047	2.3026
		FIML	0.7216	0.5854	1.2587
	0.3	**ORIGINAL DATA**	1.7834	1.7663	1.7791
		EM	6.0557	1.0676	2.7992
		FIML	0.3593	0.3912	1.0923

The previous table indicates that:

- Ina sample size of 10 observations and a 10% percentage of missing data, **EM** was better than **FIML** at the **NMAR** mechanism, while **FIML** was better than **EM** at the **MAR** and **MCAR** mechanisms.
- In a sample size of 10 observations and 20%, 30% percentages of missing data, the **FIML** method was better than **EM** of all missing data mechanisms.
- Ina sample size of 20 observations and a 10% percentage of missing data, **FIML** was better than **EM** at the **MCAR** mechanism, while **EM** was better than FIML with the **MAR** and **NMAR** mechanisms methods.
- In a sample size of 20 observations and the percentage of missing data are 20% and 30%, the **EM** method was better than **FIML** at the **NMAR** mechanism, while the **FIML** method was better at the **MAR** and **MCAR** mechanisms.
- In a sample size of 30 observations and a percentage of 10% loss data, the **FIML** method was better than **EM** at the **NMAR** mechanism, while the **EM** method was better than **FIML** at the **MAR** and **MCAR** mechanisms.
- In a sample size of 30 observations and a percentage of 20% for missing data, the **FIML** method was better than **EM** of all loss mechanisms
- In a sample size of 30 observations and a percentage of 30% loss data, the **EM** method was better than **FIML** at the **NMAR** mechanism, while the **FIML** method was better at **MAR** and **MCAR** mechanisms.
- In a sample size of 50 observations and a percentage of 10% missing data, the **EM** method was better than **FIML** of all mechanisms
- In a sample size of 50 observations and a percentage of 20% missing data, the **FIML** method was better than **EM** at the **MCAR** mechanism, while **EM** method was better than **FIML** at **MAR**, and **NMAR** mechanisms.
- In a sample size of 50 observations and a 30% percentage of loss, the **EM** method was better than **FIML** at **NMAR** mechanism, while the **FIML** method was better at **MAR** and **MCAR** mechanisms.
- In a sample size of 100 observations and a percentage of missing data 10%. The **EM** method was better than **FIML** with all the mechanisms
- In a sample size of 100 observations and a percentage of 20% for missing data, the **FIML** method was better than **EM** at **MCAR**

mechanism, while the **EM** method was better than **FIML** at **MAR**, and **NMAR** mechanisms.

- In a sample size of 100 observations and a percentage of 30% missing data, the **EM** method was better than **FIML** at **NMAR** mechanism, while the **FIML** method was better at **MAR** and **MCAR** mechanisms.

Summary

For the **MAR** mechanism of missing data at small sample sizes (n = 10), **FIML** was better than the method as its effect size values were closer to effect size values than **EM** relative to the effect size of original data. For sizes, more than 20 observations, **EM** and **FIML** methods were found to have approximately the same importance as the effect size. But with sample size rising to 30 and more, the results indicated that **EM** is better than in the case of the small percentage of missing data, while **FIML** is better than in the case of the large percentage of missing data.

Regarding the **NMAR** mechanism of the missing data, the results showed that the **EM** method was better than the **FIML** method in the case of large samples (greater than 30), where the effect size values of the **EM** method were closer to the effect size values of the complete data with all different loss ratios than the **FIML** method. In the case of small sample sizes, there is no clear way about the best method to select it. The results showed that the **FIML** method is better in the case of a small sample size (n = 10) with large data loss rates. The results also showed that the **FIML** method is better in the case of a large sample size (n = 30) with low data loss rates.

For the **MCAR** mechanism of the missing data, the results showed that the **FIML** method is better than in the case of small sample sizes and all percentages of missing data. In the case of large sample sizes, the **EM** method was better than in the case of low percentages of loss, and the **FIML** method was the best in the case of large loss rates. The next table indicates the best method in each case as follows:

Table (2): the best method in each case

n	Percentages of missing	MAR	NMAR	MCAR
10	0.1	FIML	EM	FIML
	0.2	FIML	FIML	FIML
	0.3	FIML	FIML	FIML
20	0.1	EM	EM	FIML
	0.2	FIML	EM	FIML
	0.3	FIML	EM	FIML
30	0.1	EM	FIML	EM
	0.2	FIML	FIML	FIML
	0.3	FIML	EM	FIML
50	0.1	EM	EM	EM
	0.2	EM	EM	FIML
	0.3	FIML	EM	FIML
100	0.1	EM	EM	EM
	0.2	EM	EM	FIML
	0.3	FIML	EM	FIML

References

[1] Abdulrahman, Attia, T. (2013) "A guide to designing and implementing research in the social sciences: An applied approach to building research skills" Riyadh, Institute of Public Administration.

[2] Abu Alam, Mahmoud, R. (2007) "Research methods in psychological and educational sciences. 6th edition" Cairo: Universities Publishing House.

[3] Al-Hamami, Alaa Hussein, A. (2007) "Data mining.1st edition" Amman: Ithraa for Publishing and Distribution.

[4] Al-Qahtani, Saad. (2015) "Applied Statistics - Basic Concepts and the Most Used Statistical Analysis Tools in Social and Human Studies and Research Using SPSS" Riyadh, Institute of Public Administration.

[5] Blanch, Niles J. (2017) "Introduction to Structural Equation Modeling using IBM SPSS statistics and AMOS" Riyadh, Institute of Public Administration.

[6] Davey, A. & Savla, J. (2010) *"Statistical powers analysis with missing data astructural equation modeling approach"* New York , Routledge .

[7] Dudin, Muhammad, H. (2013) "Advanced statistical analysis of data using SPSS. 2nd edition" Amman, Dar Al Masirah for Publishing and Distribution.

[8] Enders, C. K. (2010) "*Applied missing data analysis*" New York, The Guilford Press.

[9] Graham, J. W. (2012) "*Missing data analysis and design*" New York, springer.

[10] Hoyt, D. & Kramer, D. (2016) "Introduction to the SPSS Statistical Software Package in Psychology" Amman, Dar Al-Fikr for Publishing.

[11] Lee, T. & Shi, D. (2021) "A Comparison of full information maximum likelihood and multiple imputation in structrual equation modeling with missing data" Psychological Methods, Advances online publication.

[12] Little, R . J. A. (1992) "Regression with missing X's" *Journal of the American Statistical Association*, 87, 1227 – 1237.

[13] Little, R. J. A. & Rubin, D. B. (2002) "*Statistical analysis with missing data*" NewYork, Wiley.

[14] Mcknight, P. E., Mcknight, K. M., Sidani, S. & Figuredo, A . J. (2007) *"Missingdata a gentle introduction*" New York, The Guilford Press.

[15] Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2015) "*Handbook of missing data methodology*" New York, CRC Press.

[16] Nakai, M. & Ke, W. (2011) "Review of methods for handling missing data in longitudinal data analysis" *Journal of Math. Analysis*, 5, 1-13.

[17] Peng, C., Harwell, M., Liou, S. &Ehman, L. (2006) "*Advances in missing data methods and implications for educational research*" Real Data Analysis, PP. 31-78.

[18] Rizkallah, A. (2002) "Researchers' guide to statistical analysis of choice and interpretation" Cairo, House of Books.

[19] Schafer, J. L. (1997) "*Analysis of incomplete multivariate data"* London, Chapman & Hall.