Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 16 Issue 5 Version 1.0 Year 2016
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Effect of Missingness Mechanism on Household Survey Estimates in Nigeria

By Faweya Olanrewaju & G. N. Amahia

Ekiti State University

Abstract- In this study, we employed three missingness mechanisms - MCAR, MAR and MNAR to investigate the effects of proportion of Missing data on descriptive and analytic statistics: Mean ($\bar{Y}$), Variance ($\sigma^2 y$), correlation coefficient ($\rho_{yx_1x_2}$), coefficient of variation (cv), skewness (sk) and Kurtosis (K) which are likely situation a researcher may encounter in the field when dealing with household surveys.

This study reveals that sometimes, missing data introduce systematic distortion in survey estimates and bias flows from missing data when the causes of the missing data are linked to the survey statistic measured.

Keywords: missing data, MCAR, MAR, MNAR, descriptive statistics.

GJSFR-F Classification : FOR Code : 70B15

EFFECTOFMISSINGNESSMECHANISMONHOUSEHOLDSURVEYESTIMATESINNIGERIA

Strictly as per the compliance and regulations of :

Notes

Effect of Missingness Mechanism on Household Survey Estimates in Nigeria

Faweya Olanrewaju [α] & G. N. Amahia [σ]

Abstract- In this study, we employed three missingness mechanisms - MCAR, MAR and MNAR to investigate the effects of proportion of Missing data on descriptive and analytic statistics: Mean ($\bar{\mathbf{y}}$), Variance ($\boldsymbol{\sigma^2 y}$), correlation coefficient ($\boldsymbol{\rho_{yx_1x_2}}$), coefficient of variation (cv), skewness (sk) and Kurtosis (K) which are likely situation a researcher may encounter in the field when dealing with household surveys.

This study reveals that sometimes, missing data introduce systematic distortion in survey estimates and bias flows from missing data when the causes of the missing data are linked to the survey statistic measured.

Keywords: missing data, MCAR, MAR, MNAR, descriptive statistics.

I. Introduction

One of the greatest threats compromising the accuracy of most surveys estimate during design and analysis is the problem of missing data. This may occur when some individuals provide no information because of non-contact of refusal to respond (unit non-response) or when other individuals are contacted and provide some information, but fail to answer some of the questions (item non-response).

Unfortunately, unit and item non-response are often neglected or not properly handled during analysis, and this leads to bias in the estimate. Thus, this study focused on the detection and minimization of bias associated with missing data. Though missing data is a common problem in most research studies, yet no commonly agreed upon solution exists.

Consequently, researchers have developed a wide variety of approaches for handling missing data, however, no single approaches is without pitfalls. Thus, researchers facing a missing data problem should thoroughly investigate the sources of the missing data as well as the options for handling missing data under different missingness mechanism with different amount of missing data. Otherwise, when researchers use missing data techniques without considering the mechanism of the missingness, they run the risk of obtaining biased estimates and misleading conclusions. In such cases, analysis and publication of the data may be of dubious value and jeopardize credibility of the organization preparing the report (Little & Smith, 1983).

II. Missing Data Mechanism

a) Missing Completely At Random

The distribution of the missing values R is assumed independent of both the target variable Y and auxiliary variable X. Thus,

Author α: Department of mathematical sciences, ekiti state university, iworoko, ekiti state. e-mail: lanrefaweya@gmail.com
Author σ: Department of statistics, university of ibadan.

$$P(R/Y, X) = P(R) \quad (1)$$

b) Missing At Random (MAR)

In general, MAR occur when there is no direct relation between the target variable Y and response behaviour R and the same time there is a relation between the auxiliary variable and the response behaviour R. This is expressed as:

$$P(R/Y, X) = P(R \, / \, Y^0, X) \quad (2)$$

c) Missing Not At Random

Missing data Mechanism where missing values are assumed to be related to the unobserved dependent variable vector Y_i^m, in addition to the remaining observed values is called Missing not at Random (MNAR). This is expressed as:

$$P(R/Y, X) = P(R/ \, Y^m \, Y^0 \; X) \quad (3)$$

III. Non-Response in Survey

Non-response is the failure of a sample survey (or a census) to collect data for all items in the survey questionnaire from all the population units designated for data collection. Non-response can be manifested either as item or as non-response.

a) Unit Non-Response

This refers to outright failure of a sampled subject to participation in a study.

b) Item Non-Response

Item non-response occurs in any kind of multivariate study (e.g. a survey) in which a subject responds to some, but not all survey items.

IV. Method of Analysis

We employed three missingness mechanisms - MCAR, MAR, and MNAR to investigate the effects of proportion of Missing data on descriptive and analytic statistics (Mean($\bar{Y}$)), Variance ($\sigma^2 y$), correlation coefficient ($\rho_{yx_1x_2}$), coefficient of variation (cv), skewness (sk) and Kurtosis (K) which are likely situation a researcher may encounter in the field when dealing with household surveys.

We denote by S the sample, Y = $(y_1, y_2, \ldots, y_n)^T$ where y_i denote the value of targeted random variable for unit i. Let X_i, I = 1, 2, …, k be some auxiliary variable which is available for all $i \in S$. Let R = $(r_1, \ldots, r_n)$ where $r_i = 0$ for unit that are observed and $r_i = 1$ for units that are missing. MCAR assumed distribution of missing values R to be independent of both targeted variable Y and auxiliary variable X_i, thus,

$$P(R/ \, Y, X_i \,) = P(R).$$

However, under MAR, there is no direct relationship between the targeted variable Y and the response behavior R and at the same time; there is no relationship between the auxiliary variable X_i and the response behaviour R. Thus, P(R/ Y, X_i) = P(R/ Y^0, X_i). In MNAR, missing values assumed to be related to unobserved dependent variable Y^m in addition to the remaining observed values Y^0 and this relationship cannot be explained by an auxiliary variable X_i. Thus, P(R/ Y^m, Y^0, X_i). A simple random sample of n = 100 households was selected from the record of survey data on "household income" from Akure North Local Government, Iju/ Ita- Ogbolu in Ondo

Notes

Global Journal of Science Frontier Research (F) Volume XVI Issue V Version I Year 2016

Notes

State to demonstrate the effect of missingness on descriptive an inferential statistics when different proportions of data are missing.

Three demographic variables; Y (Income N'000), Age (X_1) and year of schooling (X_2) were considered.
The variable Y was a combination of explanatory variables with added random components.

Then, differing amounts were deleted at random causing MCAR data, which had 0, 1, 5, 12, 23 and 44% missing data.

In MAR situation y become missing as follows: 0% for complete data set, 5% when $X_1 < 5$, 12% when $X_2 \geq 55$, 23% when $X_1 \leq 6$ and 44% when $X_1 \leq 6$ or $X_2 \geq 50$.

Sorting according to the actual y values in deleting the cases to give 6 different rate created MNAR data.

Table 1 : Table of Means, variance, correlation, skewness and kurtosis when different amounts of data are missing, under different assumption of missingness. The first row shows the mean, variances, correlation, skewness and kurtorsis of the household income data when no data are missing. That is the data are complete

	Missing Completely At Random (MCAR)					
Missing	$\bar{y}$	$\bar{\sigma}^2{}_y$	$\rho y_1 x_2$	CV	S_k	K
0	13.814	46.577	0.946	49.62	0.217	2.616
1	13.754	46.691	0.946	49.68	0.238	2.633
5	13.682	48.02	0.943	50.68	0.258	2.580
12	13.371	44.688	0.952	49.90	0.135	2.470
23	14.288	45.84	0.997	46.98	0.071	2.419
44	14.260	48.077	0.995	48.83	-0.35	2.437

Table : The table cont

	Missing At Random (MAR)					
Missing	$\bar{y}$	$\bar{\sigma}^2{}_y$	$\rho y_1 x_2$	CV	S_k	K
0	13.814	46.577	0.946	49.62	0.217	2.616
1	13.794	47.014	0.945	49.71	0.228	2.596
5	14.396	41.933	0.944	44.50	0.294	2.659
12	14.210	40.639	0.916	47.29	0.243	2.797
23	16.244	32.17	0.910	34.95	0.358	2.980
44	16.371	23.216	0.844	29.95	0.582	2.925

Table : The table cont

	Missing Not At Random (MNAR)					
Missing	$\bar{y}$	$\bar{\sigma}^2{}_y$	$\rho y_1 x_2$	CV	S_k	K
0	13.814	46.577	0.946	49.62	0.217	2.616
1	13.880	46.608	0.496	49.21	0.198	2.623
5	13.045	36.965	0.944	45.84	0.005	2.204
12	12.22	30.331	0.936	44.52	-0.096	2.268
23	11.103	24.457	0.931	44.93	-0.089	2.348
44	9.07	17.48	0.916	44.09	-0.072	2.490

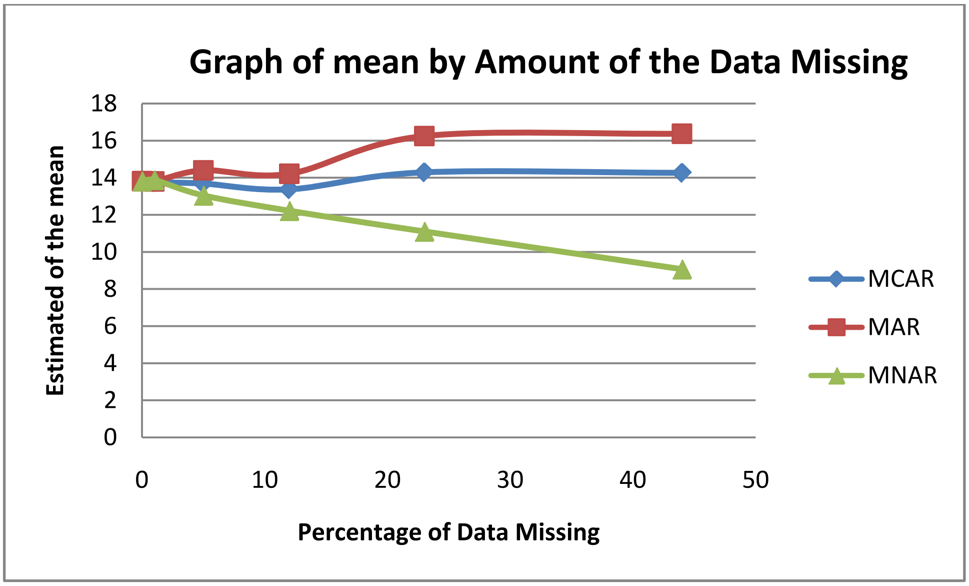

Figure 1 : Graph of Mean by Amount of the Data Missing

Comment: MCAR is approximately constant, while for MAR increases and MNAR decreases.

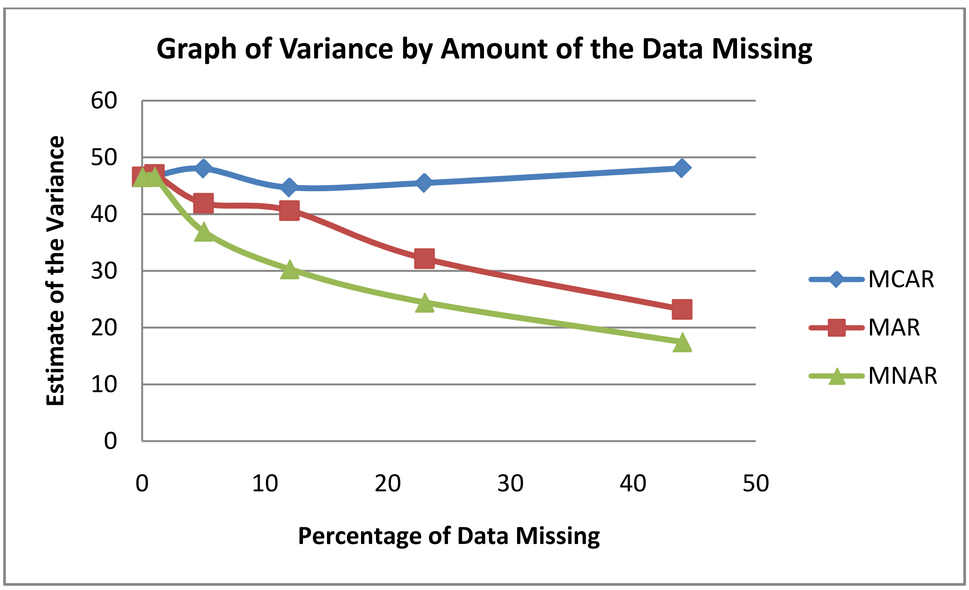

Figure 2 : Graph of Variance by Amount of the Data Missing

Comment: Under MCAR values for variances is approximately constant as proportion of missingness increases, while for MAR and MNAR decreases but MNAR is more drastical.

Notes

Notes

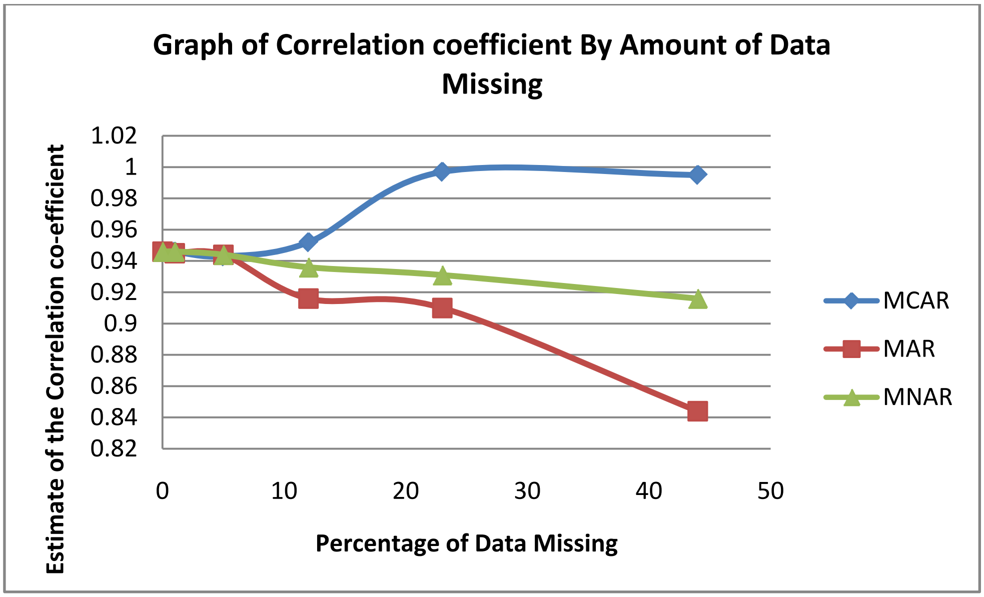

Figure 3 : Graph of Correlation Coefficient by Amount of the Data Missing

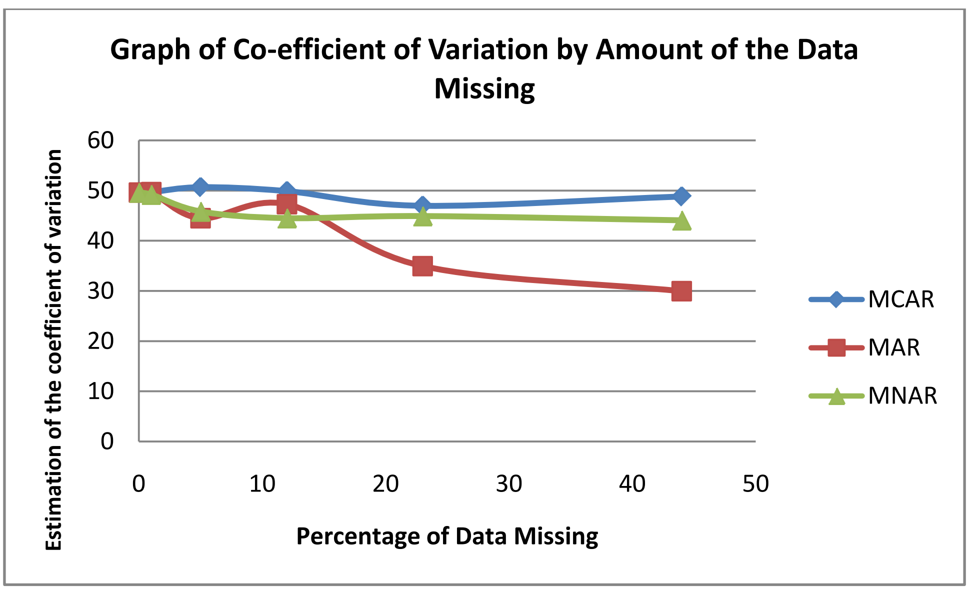

Figure 4 : Graph of Coefficient of Variation by Amount of the Data Missing

Comment: MCAR is approximately constant, while for MAR decreases.

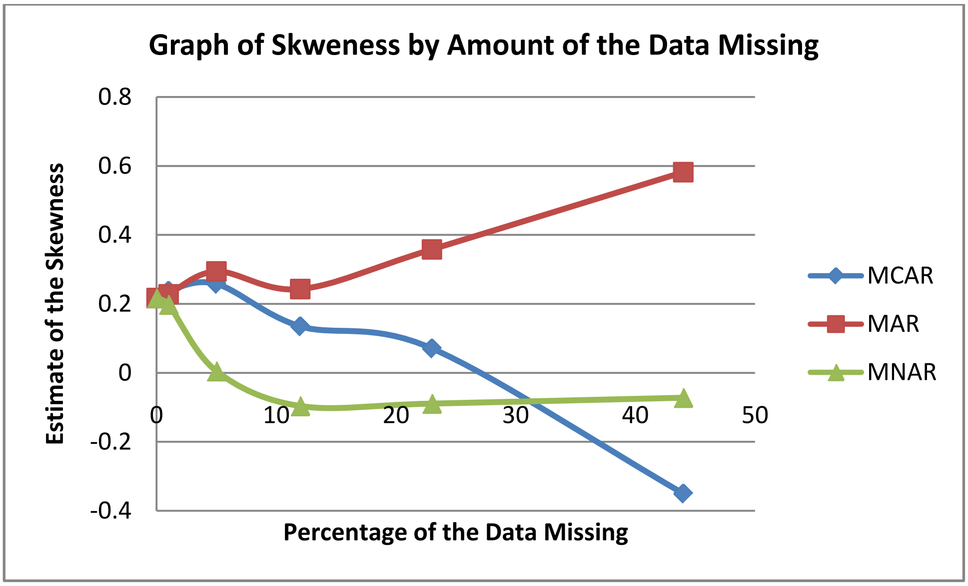

Figure 5 : Graph of Skewness by Amount of the Data Missing

Comment: MNAR formerly decreased but later constant as the proportion of missingness increases but MCAR increases while MAR decreases.

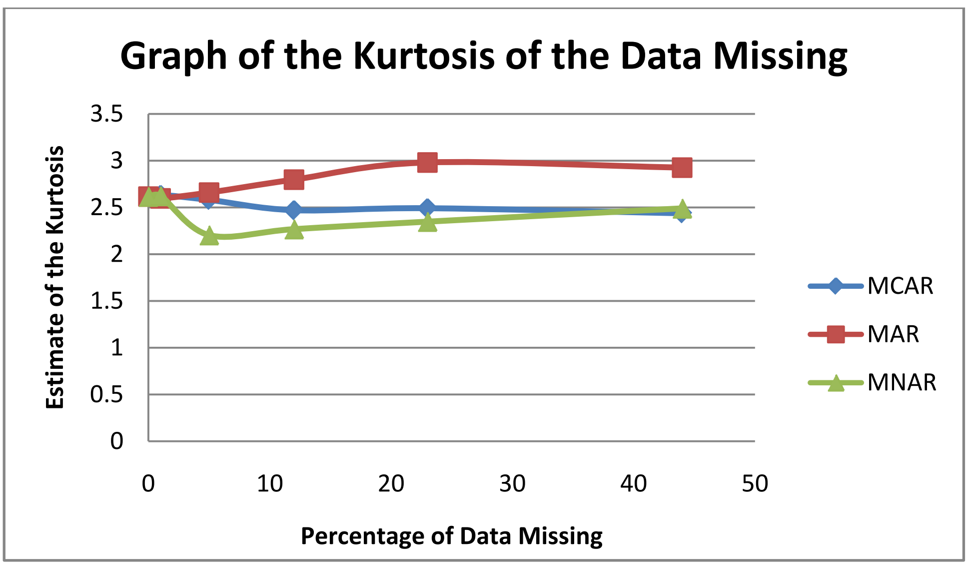

Figure 6 : Graph of Kurtosis by Amount of the Data Missing

Comment: MCAR and MNAR is approximately constant for kurtosis even as the proportional of missingness increases but MAR is a bit different.

V. Discussion of Results

Among all the parameters considered, the one where there was no major significance difference under the three mechanisms is Kurtosis, which is the degree of peakedness of the curve of the distribution of the variable under consideration. Thus from the study, it implies that as the sample size 'n' increases, the curve tends to normality irrespective of the nature of irrespectively of the nature missingness. In addition, this study revealed that sometimes, missing data introduce systematic distortion in survey estimates and bias flows from missing data when the causes of the missing data are linked to the survey statistics measured.

Notes

References Références Referencias

Notes

1. Amahia, G.N (2010) Factors, prevention and correction methods for non-response in sample surveys. Central Bank of Nigeria, Journal of Applied Statistics 1(1), pp 79-89, (Nigeria).
2. Arbuckle, J.L. (1996). Full information likelihood estimation in the presence of incomplete data. In G.A. Marcoulides and R.E. Schumaker (Eds.), Advance structural equation modeling. Mahwah, Nj: Lawrence Erlbaum.
3. Beale, E.M.L., & Little, R.J.A (1975). Missing values in multivariate analysis.Journal of the Royal Statistical Society, Series B, 37, 129- 146.
4. Cochran, W.(1968). "The Effective of Adjustment by Subclassification in Removing Bias in Observational studies." Biometrics, 24, pp. 295-313.
5. Cohen, J and Cohen, P, (1983). Missing data. In J Cohen and P. Cohen, Applied multiple regression:Correction analysis for the behavioral sciences (pp. 275-300). Hillsdale, NJ Erlbaum.
6. Cool, A.L. (2000). A review of methods for dealing with missing data. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, TX.(ERIC Document ReproductioN Service No.ED 438 311).
7. Diggle P.J Testing for random dropouts in repeated measurement data.Biometrics,45 (1989) 1255-1258.
8. Diggle P.J., Heagerty P., Liang, K.Y. et al. Analysis of longitudinal Data. 2nd ed. Oxford University Press Inc.,(2002).
9. Dixon W. J BMDP Statistical software.Los Angeles. University of California Press,(1988).
10. Ender C.K., and Bandalos, D.L (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models.Structural Equation Modeling, 8(3),430-457.
11. Groves, R Couper M. (1998); Non-response Interview Survey, New York Wiley.
12. Groves, R Singer. and Corning, A (2000); Leverage- Salience Theory of Survey Participation Description and an Illustration. Public Opinion Quarterly,64,299-308.
13. Groves, R. Cialdini R and Couper M (1992): "Understanding the Decision to participate in a Survey" The public Opinion Quaterly,54(4), 475-495.
14. Howell, D.C (2007) The analysis of missing data.Inoutwaite, W and Turner, S .Handbook of social science Methodology London: Sage. Return.
15. Kalton, G. and Flores-Cervantes, I (2003), "Weighting Methods" Journal of official statistics, 19,pp.81-97.
16. Kim, J.O., & Curry, J. (1977). The instrument of missing data in multivariate analysis Sociological Methods & Research, 6(2), 215-240.
17. Little R.J.A. A test of missing completely at random for multivariate data with missing values. Journal of American statistical Association, 83 (1988),1198-2002.
18. Little, J. & An, H., (2004). Robust likelihood-based analysis of multivariate data with missing values.Statisticasinica Vol. 14, pp 949-968.
19. Madlow, W.G., Nisselson, H., Olkin, I. (Eds.), 1983. Incomplete data in sample surveys.Report and Case Studies, vol.1. Academic Press, New York.
20. Quinten, A., Raaijmakers, W., 1999. Effectiveness of different missing data treatments in survey with Likert-type data: introducing the relative mean substitution approach. Educational and psychological Measurement 59 (5), 725-748.
21. Schafer J.L., Graham J. W. Missing data: Over view of the state of the art. Psychological Methods, 7 (2002),147:177.

22. Schafer, J (2002). Dealing with missing data.Research letter information mathematics science. Vol. 3, pp 153-160.
23. Tshering, S., Okazaki T and Endo S., March 2013: A Method to Identify Missing Data Mechanism in Incomplete Dataset. IJCSNS International journal of Computer Science and Network Security,Vol.13 No14 Page (14-21).
24. Utazi C.E., Onyeagu S.I. & Osuji G.A (2010). On the efficiency of some techniques for estimating covariance and correlation matrices from incomplete data. Journal of the Nigerian Statistical Association Vol. 22, 44-63.

Notes