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 ($\rho_{yx_1,x_2}$), 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
Effect of Missingness Mechanism on Household Survey Estimates in Nigeria

Faweya Olanrewaju & G. N. Amahia

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 ($r_{y_1,y_2}$), 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,
\[ P(R/Y, X) = P(R) \] (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:

\[ P(R/Y, X) = P(R/ Y^0, X) \] (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:

\[ P(R/Y, X) = P(R/ Y^m Y^0 X) \] (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 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 (\( \sigma^2Y \)), correlation coefficient (\( \rho_{yx_1,x_2} \)), 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, \( Y = (y_1, y_2, ..., y_n)^T \) 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 \( R = (r_1, ..., r_n) \) 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
State to demonstrate the effect of missingness on descriptive and 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 kurtosis of the household income data when no data are missing. That is the data are complete.

<table>
<thead>
<tr>
<th>Missing</th>
<th>$\bar{y}$</th>
<th>$\sigma^2_y$</th>
<th>$\rho Y_1X_2$</th>
<th>CV</th>
<th>$S_k$</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13.814</td>
<td>46.577</td>
<td>0.946</td>
<td>49.62</td>
<td>0.217</td>
<td>2.616</td>
</tr>
<tr>
<td>1</td>
<td>13.754</td>
<td>46.691</td>
<td>0.946</td>
<td>49.68</td>
<td>0.238</td>
<td>2.633</td>
</tr>
<tr>
<td>5</td>
<td>13.682</td>
<td>48.02</td>
<td>0.943</td>
<td>50.68</td>
<td>0.258</td>
<td>2.580</td>
</tr>
<tr>
<td>12</td>
<td>13.371</td>
<td>44.688</td>
<td>0.952</td>
<td>49.90</td>
<td>0.135</td>
<td>2.470</td>
</tr>
<tr>
<td>23</td>
<td>14.288</td>
<td>45.84</td>
<td>0.997</td>
<td>46.98</td>
<td>0.071</td>
<td>2.419</td>
</tr>
<tr>
<td>44</td>
<td>14.260</td>
<td>48.077</td>
<td>0.995</td>
<td>48.83</td>
<td>-0.35</td>
<td>2.437</td>
</tr>
</tbody>
</table>

Table: The table cont

<table>
<thead>
<tr>
<th>Missing</th>
<th>$\bar{y}$</th>
<th>$\sigma^2_y$</th>
<th>$\rho Y_1X_2$</th>
<th>CV</th>
<th>$S_k$</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13.814</td>
<td>46.577</td>
<td>0.946</td>
<td>49.62</td>
<td>0.217</td>
<td>2.616</td>
</tr>
<tr>
<td>1</td>
<td>13.794</td>
<td>47.014</td>
<td>0.945</td>
<td>49.71</td>
<td>0.228</td>
<td>2.596</td>
</tr>
<tr>
<td>5</td>
<td>14.396</td>
<td>41.933</td>
<td>0.944</td>
<td>44.50</td>
<td>0.294</td>
<td>2.659</td>
</tr>
<tr>
<td>12</td>
<td>14.210</td>
<td>40.639</td>
<td>0.916</td>
<td>47.29</td>
<td>0.243</td>
<td>2.797</td>
</tr>
<tr>
<td>23</td>
<td>16.244</td>
<td>32.17</td>
<td>0.910</td>
<td>34.95</td>
<td>0.358</td>
<td>2.980</td>
</tr>
<tr>
<td>44</td>
<td>16.371</td>
<td>23.216</td>
<td>0.844</td>
<td>29.95</td>
<td>0.582</td>
<td>2.925</td>
</tr>
</tbody>
</table>

Table: The table cont

<table>
<thead>
<tr>
<th>Missing</th>
<th>$\bar{y}$</th>
<th>$\sigma^2_y$</th>
<th>$\rho Y_1X_2$</th>
<th>CV</th>
<th>$S_k$</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13.814</td>
<td>46.577</td>
<td>0.946</td>
<td>49.62</td>
<td>0.217</td>
<td>2.616</td>
</tr>
<tr>
<td>1</td>
<td>13.880</td>
<td>46.608</td>
<td>0.496</td>
<td>49.21</td>
<td>0.198</td>
<td>2.623</td>
</tr>
<tr>
<td>5</td>
<td>13.045</td>
<td>36.965</td>
<td>0.944</td>
<td>45.84</td>
<td>0.005</td>
<td>2.204</td>
</tr>
<tr>
<td>12</td>
<td>12.22</td>
<td>30.331</td>
<td>0.936</td>
<td>44.52</td>
<td>-0.096</td>
<td>2.268</td>
</tr>
<tr>
<td>23</td>
<td>11.103</td>
<td>24.457</td>
<td>0.931</td>
<td>44.93</td>
<td>-0.089</td>
<td>2.348</td>
</tr>
<tr>
<td>44</td>
<td>9.07</td>
<td>17.48</td>
<td>0.916</td>
<td>44.09</td>
<td>-0.072</td>
<td>2.490</td>
</tr>
</tbody>
</table>
**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 drastic.
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.
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.


