Journal of Global Social Sciences **June 2024, Volume 5, Number 18, 7-17**
ISSN: Print 2735-9328, Online 2735-9336

REVIEW OF METHODS FOR MISSING DATA

Salam Mahmood HAMAD,
Department of business management, College of Humanities, University of Raparin
Salam.Mahmood@uor.edu.krd

Abstract

This paper reviews methods for handling missing data in a research study. Many researchers use as hoc methods such as complete case analysis, available case analysis (pairwise deletion), or single-value imputation, though these methods are easily implemented, they require assumptions about the data that rarely hold in practice. Model-based methods such as maximum likelihood using the EM algorithm and multiple imputations hold more promise for dealing with difficulties caused by missing data. While model-based methods require specialized computer programs and assumptions about the nature of the missing data, these methods are appropriate for a wider range of situations than the more commonly used ad hoc methods. The paper provides an illustration of the methods using data from an intervention study designed to increase students' ability to control their asthma symptoms.

Key words: *Missing Data, Statistics, Analytical Process.*

DOI: 10.58934/jgss.v5i18.261

1. INTRODUCTION

At some stage in their work, both researchers have faced the issue of missing quantitative data. Research informants may decline or forget to answer a question from the survey, les are lost or data is not correctly registered (Rashid, 2020; Jaf & Rashid, 2023). We cannot afford to start over or wait until we have established foolproof methods of collecting information, an unattainable target, given the cost of collecting data. When we do not have full information from all informants, we end up left with the decision on how to interpret data (Budur, 2020). We are not alone in this problem; the United States Census Bureau has been engaged in a

debate with the U.S. The U.S. Congress and On the handling of the undercount in the 2000 U.S. Supreme Court Census. Census. Provided that most scientists do not have the U.S. capital What are the choices open to the Census Bureau for review of data with missing information? The use of only those cases with full details is the most common approach and the simplest to apply (Rashid, 2019). In a statistical analysis, researchers either intentionally or by chance drop informants who do not have full data on the variables. Researchers may use a plausible value for the missing observations as an alternative to complete-case analysis, such as using the mean of the observed cases on that variable. More recently, statisticians have advocated approaches focused on data distribution models (such as maximum likelihood and multiple imputation). Much has been published in the statistical literature on missing data (Little, 1992; Little & Rubin, 1987; Schafer, 1997; Rashid & Sabir Jaf, 2023; Budur et al., 2018). Reviewing the data collection, data planning, data analysis, and outcome interpretation phases will illustrate the challenges that researchers need to consider when making a decision about how to deal with missing data in their work. The paper focuses on commonly used missing data methods: complete-cases, available-cases, single-value imputation, and more recent model-based methods, maximum likelihood for multivariate normal data, and multiple imputations.

2. LITERATURE REVIEW

The investigator has the ability to observe the potential reasons for missing data during data collection, evidence that will help direct the decision about what missing data approach is suitable for the analysis (Rashid, 2017; Mahmood & Sabir, 2023; and Rashid, 2023). From full-case analysis to model-based approaches, missing data techniques each bear assumptions about the existence of the process that causes the missing data. Several students in the asthma study have missing data on their symptom severity rating, as is expected for students aged 8 to 14 years. One possible explanation is that students simply forgot to visit the school clinic to ®ll out the form. If students randomly miss their symptom severity rating ± because they have forgotten or for any other reasons not relevant to their health, the results from the rest of the students should be indicative of the initial ratings of the care and control group. Rubin (1976) introduced the term ``missing completely at random'' (MCAR) to describe data where the complete cases are a random sample of the originally identified set of cases. Since the complete cases are representative of the sample originally described, the inferences based on the complete cases only refer to the wider sample and to the target population. Full case analysis for MCAR data produces findings that can be applied to the target population with one caveat

± the predictions would be less reliable than the researcher originally expected as a limited number of estimation cases are used.

Variables scale and distribution Another problem related to the process of data collection involves assumptions we make about the distribution of variables in the model (Budur et al., 2024). We select research procedures suitable to the size and distribution of the variables when choosing a statistical model for a sample. The researcher must make the assumption that the data is multivariate normal in the model-based methods I will address here, that the common distribution of all variables in the data set (including outcome measures) is a multivariate normal. The use of nominal (non-ordered categorical) variables seems to preclude this statement at the outset. As Schafer (1997) discusses, this assumption can be relaxed to the assumption that the data are multivariate normal conditional on the fully observed nominal variables. For example, if we gather information on gender and group assignment in a two-group experiment, we will assume that the variables in the data are multivariate normal within each cell defined by the crossing of gender and group (males and females in the treatment and control group). From this assumption, two consequences arise (Rashid, 2023). First, the use of the model-based approaches that I will explain here involves the full observation of the categorical variables in the model. As just discussed, collecting more than one measure of significant variables is one technique to help ensure fully observed categorical variables. Second, if high rates of missing observations are present in categorical variables in the data, then methods using the multivariate normal assumption should not be used. Schafer (1997) reports on simulation studies that provide proof of the robustness of the system to moderate deviations from normality when categorical variables have small amounts of missing values or are entirely observed.

3. METHODOLOGY

We carefully gather as much information as possible during the data collection process, try to get complete data on all informants, and use more than one way to obtain important variables such as income. The next stage includes the analysis of the data, and knowing the number and pattern of missing observations is the crucial task for the researcher. To understand the complexity of the missing data problem, the researcher needs to have an understanding of which variables are missing observations. Typical univariate statistics also do not give the missing data a complete account; researchers also need to consider the amount of missing details about relationships in the data between variables. I will use data from the asthma study

(Velsor-Friedrich, in preparation) to illustrate issues that arise at this stage of a research study. When first processing the data, In order to verify the amount of missing data, we also look at univariate statistics like the mean, standard deviation, and frequencies. A collection of variables from a study evaluating the results of a program to improve the awareness of students about their asthma is listed in Table 1. I am interested in examining how a measure of a student's self-efficacy beliefs about controlling their asthma symptoms relates to a number of predictors. These predictors are Group, participation in a treatment or control group; Docvis, the number of doctor visits in a specified period post-treatment; Symsev, rating of the severity of asthma symptoms post-treatment; Reading, score on state-wide assessment of reading; Age in years; Gender; and Allergy, the number of allergies suffered by the student. Commonly-Used Missing Data Methods Complete-Case Analysis Many statistical packages use listwise deletion by default when a researcher is estimating a model, such as a linear regression. Cases in the proposed model that are missing variables are dropped from the study, leaving only full cases. A researcher using complete cases assumes that the observed complete cases are a random sample of the originally targeted sample, or in Rubin's (1976) terminology, that the missing data are MCAR. Where there are only a few incomplete findings in a data set, the assumption of MCAR data is more likely to apply; when only a few cases are missing, there is a better chance of full cases representing the population. When much data is lacking, a researcher faces a more difficult decision as in the case of the asthma study. Approximately 88% of the informants fail to report one or more variables, leaving only 19 out of 154 cases for analysis. As seen earlier, some preliminary evidence from Little's (1988) MCAR test indicates that the assumption of MCAR data may not apply.

Case analysis available, or pairwise deletion, uses all data available to estimate model parameters (Rashid, 2018). If a researcher looks at univariate descriptive statistics of a data set with incomplete observations, the means and variances of the variables found in the data set are analyzed using accessible case analysis. This method is shown in Table 1, where descriptive statistics are computed for various sets of cases. The potential problems increase when interest focuses on bivariate or multivariate relationships. A plain two information matrix with only one nonresponse subject variable is illustrated in Figure 1. All cases will be used to approximate the mean of X1 in pairwise deletion, but only the full cases would contribute to the X2 estimate and the correlation between X1 and X2. To estimate parameters of interest in the data, different sets of cases are used. Although Kim and Curry (1977) argue that by using available cases instead of full cases, estimates can be improved, others argue that (Anderson,

Basilevsky, & Hum, 1983; Haitovsky, 1968; Little, 1992; Little & Rubin, 1987) have pointed out problems with the procedure.

Many researchers are constrained by the statistical computing packages available in their data analysis choices, as well as knowledge of alternatives (Rashid, 2017; Budur et al., 2023). Total case analysis, available case analysis, and single-value imputation are the most commonly used techniques. Another group of methods, maximum probability and multiple imputations, are based on data models, in this case the multivariate normal distribution Analysis OF MISSING DATA 361. This section focuses on these ®ve methods for analyzing data with missing observations, examining the assumptions of each method, and describing how one analyzes the data with current software. While current software offers options such as full case analysis, case analysis available, and mean substitution, there are no equivalent data assumptions in these approaches, nor do they provide comparable results. Due to their applicability in a wide variety of contexts and data given a collection of explicitly defined assumptions, several statistical techniques appeal to researchers. Only complete case analysis and model-based methods can apply broadly across a number of data contexts with the fewest number of assumptions (Rashid, 2021). Although the debate so far has not differentiated between problems with missing data and missing predictor variables in an analysis, the existence of the variables with missing measurements and the purpose of the statistical analysis must be taken into account by the researcher. Missing results face different issues from missing predictors, as Little (1992) addresses. If the results are MAR, those cases with missing results and fully observed predictors do not add any data to a linear model that looks at the relationships between the outcome and the predictors. If results and predictors are MAR, some knowledge about the joint distribution of the missing predictors leads to cases missing both outcomes and predictors. When outcomes are not MAR, then more complex modeling procedures are needed. The methods described here will focus on MAR data only.

In order to maintain the number of cases originally defined for the analysis, some researchers may provide a plausible missing value, such as the mean for the cases in which the variable is observed (Rashid, 2018). With the statistical process, the researcher proceeds as though the knowledge is fully observed. Although this method allows all cases to be included in a standard analysis technique, substituting missing values with a single value modifies the distribution of that variable by reducing the variance that is likely to be present. Little (1992) points out that the variance of these same variables is underestimated, whereas mean imputation results in

overall means that are equal to the complete case values. Two sources derive from this underestimation. First, ®lling in the missing values with the same mean value does not account for the variation that would likely be present if the variables were observed. The true values are probably different from the average. Second, the smaller standard errors do not sufficiently reject the ambiguity that occurs in the data due to the increased sample size. In certain cases, lack important variables, a researcher may not have the same amount of information present as he or she would have with fully observed data. When calculating multivariate parameters such as regression coef ® cents, bias in the calculation of variances and standard errors are compounded. In no circumstances does imputation suggest impartial outcomes.

Single-Value Imputation Some researchers all have a logical missing value, such as the mean for the cases that observe the attribute, in order to maintain the number of cases originally defined for the analysis (Rashid, 2021). With the statistical process, the researcher proceeds as though the knowledge is fully observed. Although this method allows all cases to be included in a standard analysis technique, substituting missing values with a single value modifies the distribution of that variable by reducing the variance that is likely to be present. Little (1992) points out that the variance of these same variables is underestimated, whereas mean imputation results in overall means that are equal to the complete case values. Two sources derive from this underestimation. First, ®lling in the missing values with the same mean value does not account for the variation that would likely be present if the variables were observed. The true values are probably different from the average. Second, the smaller standard errors do not sufficiently reject the ambiguity that occurs in the data due to the increased sample size. In certain cases lack important variables, a researcher may not have the same amount of information present as he or she would have with fully observed data. In calculating multivariate parameters such as regression coefficients, bias in the calculation of variances and standard errors is exacerbated. In no circumstances does imputation suggest impartial outcomes (Rashid, 2021).

The small number of cases with complete data in this intervention study poses a problem. Given that 154 students participated in the analysis, it does not seem appropriate to use only 19 cases to describe the whole; we do not think that the 19 cases are a random sample of the entire data set, although we do not have clear proof of this assumption. Instead, by considering the MAR results, we can make a much less demanding assumption (Rashid, 2020). We can use maximum likelihood methods or multiple imputations with MAR data. Assuming that the multivariate

data is normal, we obtain our regression model estimates. As predicted, there are few variations in the magnitude of the maximum likelihood and multiple imputation estimates, as multiple methods of imputation overlap with those of maximum likelihood. However, the distinction lies in the estimation of the standard errors and in the overall ease of calculation. Programs such as NORM (Schafer, 1999) provide both maximum likelihood and multiple imputation results, and more programs are likely to be available in the near future (Barnard, 2000). Between the model-based and full case estimates, the understanding of the models varies. Although adjusting for reading capacity, gender, age, intensity of symptoms, and total number of allergies experienced, the treatment group did not score significantly higher on the self-efficacy measurement. However, children who report more serious symptoms appear to doubt their capacity to manage their asthma. However, the analysis does suffer from a common problem ± a lot of missing data on significant steps. One measure of the risk that a child can experience an acute episode, especially if asthma is not under control, is the number of allergies experienced by a child. Children have attended schools in the inner city in this population, and are likely to frequent areas where their allergies and symptoms of asthma are intensified. An alternative measure of the risk of asthma episodes from a student's allergy may provide more information to analyze the model. In addition, the inclusion of Age and Reading can lead to problems of collinearity in the model itself. However, given the association between these two variables, the use of both in the model to establish multiple imputations is justified.

4. CONCLUSION

In the statistical literature, a large amount of research has appeared to persuade social scientists to use approaches other than the available case and median imputation to manage missing data. Full case analysis approaches may provide unbiased estimates when few cases are lacking values. The number of full cases is a small fraction of the total in other conditions, as in the asthma intervention report. The cost and expense in the analysis warrant our use of techniques that use as much knowledge as possible.

REFERENCES

Budur, T., Demirer, H., & Rashid, C. A. (2023). The effects of knowledge sharing on innovative behaviours of academicians; mediating effect of innovative organization culture and quality of work life. *Journal of Applied Research in Higher Education*.

Budur, T., Abdullah, H., Rashid, C. A., & Demirer, H. (2024). The Connection Between Knowledge Management Processes and Sustainability at Higher Education Institutions. *Journal of the Knowledge Economy*, 1-34.

Budur, T. (2020). The role of online teaching tools on the perception of the students during the lockdown of Covid-19. *International Journal of Social Sciences & Educational Studies*, *7*(3), 178-190.

Budur, T., Abdullah Rashid, C., & Poturak, M. (2018). Students perceptions on university selection, decision making process: A case study in Kurdistan Region of Iraq. *International Journal of Social Sciences & Educational Studies*, *5*(1), 133-144.

Fay, R.E. (1991). A Design-Based Perspective on Missing Data Variance. Paper presented at the 1991 Annual Research Conference, U.S. Bureau of the Census.

Fay, R.E. (1992). When are Inferences from Multiple Imputation Valid? Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1993). Valid Inferences from Imputed Survey Data. Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1994). Analyzing Imputed Survey Data Sets With Model-Assisted Estimators. Paper presented at the Section on Survey Research Methods, American Statistical Association.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. Journal of the American Statistical Association, 91, 490±498.

Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, B30, 67±82.

Heitjan, D.F., & Basu, S. (1996). Distinguishing `Missing at Random' and `Missing Completely at Random'. American Statistician, 50, 207±213.

Jaf, R. A. S., & Rashid, C. A. (2023). The Role of Accounting Measurement and Disclosure of Social Capital in Improving Financial Performance. *Academic Journal of Nawroz University*, *12*(4), 469-477.

Jones, M.P. (1996). Indicator and strati®cation methods for missing explanatory variables in multiple linear regression. Journal of the American Statistical Association, 91, 222±230.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods & Research, 6, 215±240.

Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83, 1198±1202.

Little, R.J.A. (1992). Regression with missing X's: A review. Journal of the American Statistical Association, 87, 1227±1237.

Little, R.J.A., & Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons.

Mahmood, S., & Sabir, R. A. (2023). The Impact of time driven activity based costing on Competitive Advantage in the Kurdistan Region of Iraq Economic Unit. *Central European Management Journal*, *31*(2), 674-690.

Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Sciences, 9, 538±573.

Meng, X., & Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. Journal of the American Statistical Association, 86, 899±909.

Rashid, C. A. (2017). An Evaluation of Mobile Instant Messaging Applications' Preferences: Case of Kurdistan Region of Iraq. *International Journal of Social Sciences & Educational Studies*, *3*(4), 20.

Rashid, C. A. (2021). The importance of statistical analysis in accounting research. *Journal of Global Social Sciences*, *2*(7), 71-84.

Rashid, C. A. (2021). The Effect of International Financial Reporting Standards 7 on Financial Crisis. *Journal of Global Economics and Business*, *2*(5), 87-100.

Rashid, C. A. (2018). Efficiency of financial ratios analysis for evaluating companies' liquidity. *International Journal of Social Sciences & Educational Studies*, *4*(4), 110.

Rashid, C. A. (2019). Pricing policy and its impact on the profitability. *International Journal of Finance & Banking Studies*, *8*(3), 101-108.

Rashid, C. A. (2017). The Importance of Audit Procedure in Collecting Audit Evidence/Case of Kurdistan Region/Iraq. *International Journal of Social Sciences & Educational Studies*, *4*(2), 15.

RASHID, C. A. (2020). Balanced Score Card and Benchmarking as an Accounting Tool to Evaluate Morrison's Performance. *Journal of Global Economics and Business*, *1*(3), 59-72.

Rashid, C. A. (2023). Social capital accounting and financial performance improvement: the role of financial information reliability as a mediator. *Journal of Islamic Accounting and Business Research*.

Rashid, C. A., & Sabir Jaf, R. A. (2023). The Role of Accounting Measurement and Disclosure of Social Capital in Improving Quality of Accounting Information. *Iranian Journal of Management Studies*.

Rashid, C. A. (2023). A requirement for fraud investigation professionals. *Journal of Global Economics and Business*, *4*(12), 75-89.

Rashid, C. A., & Jaf, R. A. S. (2023). The Usefulness of The Capital Asset Pricing Model in Predicting Total Shareholder Return. *Zanco Journal of Human Sciences*, *27*(1), 408-416.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581±592.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: JohnWiley & Sons.

Rubin, D.B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91, 473±489.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman & Hall.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http:// www.stat.psu.edu/~jls/misoftwa.html. SPSS. (1999). SPSS for windows (Version Rel. 9.0). Chicago: SPSS, Inc.

Tanner, M.A. (1993). Tools for Statistical Inference. New York: Springer-Verlag. Velsor-Friedrich, B. (in preparation). Results of an asthma intervention program in eight innercity schools.

Torlak, N. G., Budur, T., & Khan, N. U. S. (2024). Links connecting organizational socialization, affective commitment and innovative work behavior. The Learning Organization, 31(2), 227-249.