Types of Missing Data and Strategies for Handling Missing Data in Research*

Wentao Sun

March 5, 2024

Missing data are prevalent in quantitative research. It is a situation where there is no value for an observed variable in the collected dataset. This common problem in research can affect the quality of the analysis and lead to estimation bias making the whole study to get a wrong conclusion. Missing data are not problematic, per se—how we approach and treat missing data, on the other hand, can be highly problematic(Little et al. 2014). Finding the cause of the missing data and taking appropriate measures to deal with it is crucial for the integrity of the study results. It is not surprising that researchers often use missing data processing techniques. There are huge problems to be faced in the data collection phase of research in the first place. The main purpose of this paper is to discuss missing data and methods of remediation after data loss to safeguard the proper conduct of research. Understanding the types of missing data is essential for choosing the appropriate treatment. There are three main types of missing data: missing completely at random (MCAR), missing at random (MAR) and non-random missing (NMAR). Once the type of missing data is identified, the method of processing the missing data depends on the number and nature of the missingness. By using R(R Core Team 2020) as shown below.

Determining the type of missing data usually requires an in-depth understanding of the data and its collection process. No really satisfactory solution exists for missing data, which is why it is important to try to maximise data collection(Altman and Bland 2007). In practical research, data are considered to be missing completely at random when the missing data are not part of the data being observed. This means that the missing data are not related to the data that need to be obtained in the research project. For example, in a research survey on sleeping habits, some of the participants in the questionnaire forgot to fill in their age in the questionnaire they answered, and this was random. Their sleeping habits were not related to the age group collected and the different age groups did not have a great impact on the results of the study. This situation of missing age data we can consider as MCAR. The second type of missing data is if the missing data are associated with an observed variable but not with an unobserved variable. This case is considered to be one in which the data are missing

*Available at: https://github.com/TonySun1107/miniessay8.git. Thanks to Yunzhao Li for providing support.

at random. This case of missingness is much more likely to occur and unavoidable. Suppose that in a scientific study, younger people are more likely than older people not to answer a question about income. These questions relate to the personal privacy of the participants. So often the data collected for such questions are missing or inaccurate. When faced with this kind of problem in a study, one may choose to refer to other survey variables to estimate the approximate interval. As an example, if the age data is complete, then the missing income data can be analyzed based on the age data. By analyzing the exact data that has been obtained to estimate the missing data, the study can be carried out and this study still maintains its credibility. Item nonresponse seems to be an important problem in marketing research and occurs widely across a range of different types of market research data, including scanner panels(Kamakura and Wedel 2000). In this case, the missing income data is considered as MAR. The third category of missing data is when the absence of data depends on the value of the missing data itself, in which case we consider the data to be non-randomly missing. This case is more complex to deal with. The missing data are directly correlated with the unobserved data, which means that reasonable estimates cannot be made from the available data. The implications of this situation for scientific research are enormous, and alternative ways of refining the data obtained or secondary data collection must be thought of. For example, in a study on the prevalence of underage smoking, those individuals who were underage may have been more reluctant to provide basic information and answers to questions. At this point a different survey would have to be considered. Or changing the questionnaire questions and answering them sideways to get the desired answers. Making the participants more willing to answer would also make the data closer to the real situation. In this case is NMAR.

When dealing with data in scientific studies, researchers often encounter the problem of missing data. Historically, researchers have relied on a variety of ad hoc techniques to deal with missing data(Baraldi and Enders 2010). Missing data not only reduces the effective sample size for analysis, but also may lead to huge bias in research results. Therefore, it is crucial to choose a reasonable method to deal with missing data. Three common processing methods will be discussed below.

The first method of dealing with missing data is listwise deletion. It is a straightforward and simple method of dealing with missing data that directly removes observations that contain any missing values. The main advantage of this method is that it maintains the credibility and accuracy of the study because the results obtained from the analysis are based on all the data that were fully obtained and no data were estimated. However, it also has the obvious disadvantage that it may lead to the loss of a large amount of data, which may make the whole study incomplete. For example, in a survey study, the researcher may be interested in several variables (e.g., age, income, sleep status, etc.). If some of the participants do not answer any of the questions about income, then all of the data for that participant will not be analyzed and processed. The direct deletion of data seriously compromises the integrity of the study. Again this causes bias in the data. The second method of dealing with missing data is pairwise deletion. It allows for the deletion of missing data for only the variables involved in the specific analysis of the study. This method deals with the missing data based on the correlation between the different variables, using the data already available for estimation or

deletion. The biggest advantage of this method is that it utilizes more data and improves data utilization. Similarly it has the disadvantage that it leads to inconsistency between the results every time. There is no guarantee that the results afterward will match the reality. The third method of dealing with missing data is mean interpolation. Mean value interpolation is a common method of interpolating missing data. Missing data is filled in by replacing the missing data with the mean value of the corresponding variable. This method is straightforward and simple to operate and can be applied to most of the datasets studied. However, an unavoidable problem with mean interpolation is that it may underestimate the variance of the variables, introducing greater bias into the study data affecting the inaccuracy of the results of the study analysis. This method is applicable in market research where the researcher needs to collect consumer ratings for multiple products. If some consumers only rated some of the products, the researcher may use the mean to estimate the ratings of the unrated products in place of the missing data in order to include that consumer's data in the analysis. Although this method is quick and easy to handle missing data, it may lead to errors in the mean ratings of products. It is concluded that pairwise deletion and listwise deletion are among the least effective methods in terms of approximating the results that would have been obtained had the data been complete(Raymond 1986).

Identifying the type of missing data is a crucial first step when dealing with missing data. Different types of missing data have different impacts on the analysis results. Therefore different methods are required to deal with missing data. Missing data can be broadly categorized into three types: missing completely at random (MCAR), missing at random (MAR), and non-random missing (NMAR). Understanding these distinctions can help in choosing the most appropriate processing strategy to minimize the potential bias of missing data on study results. The use of appropriate processing methods not only mitigates the bias that may be introduced by missing data, but also improves the reliability of the study's conclusions. When reporting the results of a study, the study should clearly state the processing methods and possible effects of missing data. This maintains the credibility of the study.

References

Altman, Douglas G, and J Martin Bland. 2007. "Missing Data." *Bmj* 334 (7590): 424–24.

Baraldi, Amanda N, and Craig K Enders. 2010. "An Introduction to Modern Missing Data Analyses." *Journal of School Psychology* 48 (1): 5–37.

Kamakura, Wagner A, and Michel Wedel. 2000. "Factor Analysis and Missing Data." *Journal of Marketing Research* 37 (4): 490–98.

Little, Todd D, Terrence D Jorgensen, Kyle M Lang, and E Whitney G Moore. 2014. "On the Joys of Missing Data." *Journal of Pediatric Psychology* 39 (2): 151–62.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raymond, Mark R. 1986. "Missing Data in Evaluation Research." *Evaluation & the Health Professions* 9 (4): 395–420.