# The strategies of managing missing data*

Ben Li

March 4, 2024

## Introduction

Missing data is a common problem in research, data analysis, and any data-related profession. It is a phenomenon in which information is not stored efficiently during data collection and processing due to various reasons, such as respondents skipping specific questions, data entry errors, etc. This phenomenon leads not only to incomplete datasets but also to inaccurate research results. Therefore, identifying and dealing with missing data becomes a critical step in ensuring the validity of a study. The problem of missing data spans multiple fields and affects any profession related to data, such as social sciences, medical research, economics, and so on. As big data and machine learning continue to advance, so do the methods of processing data, aiming to minimize the impact of missing data on research.

## Types of Missing Data

based on Gelman, Hill, and Vehtari (Gelman, Hill, and Vehtari 2020) we consider three main categories of missing data:

1. Missing Completely At Random;
2. Missing at Random; and
3. Missing Not At Random.

Missing at complete random (MCAR): The probability of missing data is the same in all cases, and in such cases, the missing data are independent of observed and unobserved data. In this case, the cause of the missing data is independent of any data values in the dataset, and therefore, the missing data do not bias the data analysis. For example, if a respondent accidentally

*Code and data supporting this analysis is available at: https://github.com/UofT-DailinLi/toronto-crime-report.git skipped a question unrelated to either the individual's characteristics or the content of the question itself, it would be MCAR. But this is usually rare.

Missing at Random (MAR): Missing data are not correlated with the missing data, although they are correlated with at least one of the observed variables. This means that the missing data becomes uncorrelated with the actual value once the observed variable with which it is correlated is considered. For example, the new generation of young people prefers to avoid answering questions about income. Still, this response needs to be more related to their age, not directly to the question about income. Then, this missingness can be considered as MAR.

The last case is Missing Not at Random (MNAR), in which the missing data is correlated with the unobserved data itself, which can complicate data processing. This is because, often, we are still determining the reasons and mechanisms behind it. For example, if a person is sick, they are likely to hide that they are sick when doing the questionnaire. In such a case, the missing value is correlated with the missing mechanism, which is considered minor.

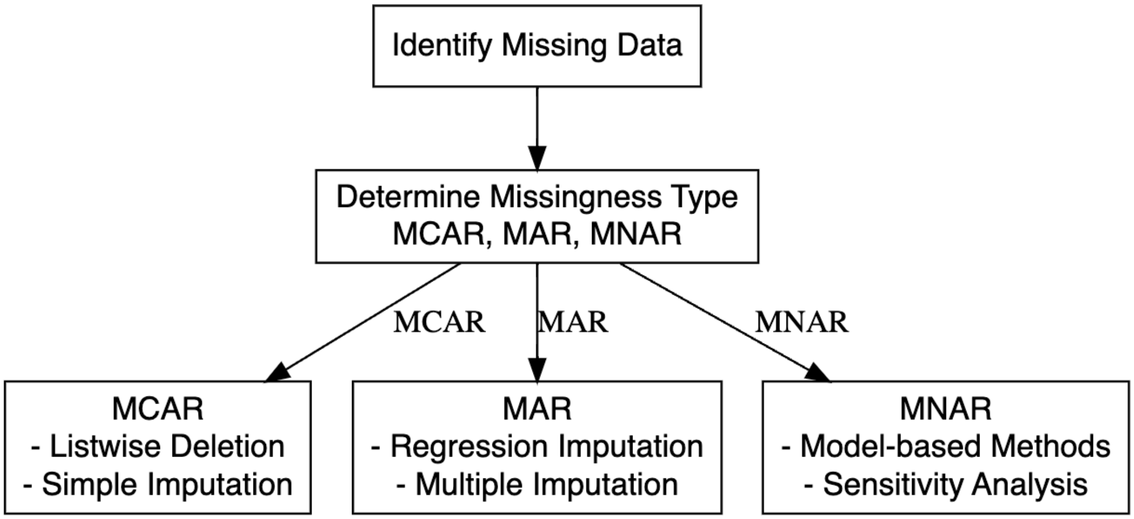

Figure 1: Flowchart for Handling Missing Data

## Dealing with Missing Data

According to this flowchart in figure 1, we can find three different solutions for missing data; we have the deletion strategy regarding Missing Completely at Random (MCAR). Direct deletion is feasible if the number of missing data is insignificant because the missing data are not correlated with any observed or unobserved data. The second method we can use is simple imputation, such as using the mean and median of the remaining data to fill out the missing value.

Regarding Missing at Random (MAR), we can use regression imputation to predict the missing value based on other data in the dataset using linear and logistic regression. The second method can be multiple imputations, which generate multiple datasets by estimating the missing values multiple times, such as using regression modelling and Bayesian modelling based on the data's nature and the missingness mechanism, etc. The results are then combined and analyzed and discussed. The results are then combined and analyzed. This approach is more flexible and considers uncertainties in the estimation process.

We can use the model-based method for Missing Not at Random (MNAR). We create a model to predict if this data will be missing based on other variables and then use this model to GUIDE this estimation. It can also be analyzed through sensitivity testing to determine the results if different assumptions were made about the missing data.

## Acknowledgment

Peer reviewed by Hechen Zhang

## Reference

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press. https://avehtari.github.io/ROS-Examples/.