Discussion about Missing Data*

Understanding and Addressing Missing Data in Exploratory Data Analysis

Jiwon Choi

March 4, 2024

Table of contents

1 Introduction

Missing data is a common challenge in data collection and analysis, posing significant obstacles to obtaining accurate insights and making informed decisions. This issue, widely recognized in exploratory data analysis (EDA), requires a thorough understanding of its types and the implementation of strategic measures to lessen its effects. Gelman, Hill, and Vehtari (2020) categorize missing data into three main types: Missing Completely At Random (MCAR), Missing at Random (MAR), and Missing Not At Random (MNAR), each with distinct characteristics and implications for data analysis. This paper follows the contents of several chapters from "Telling Stories with Data"(Alexander 2023) and examines these categories, provides examples, and discusses various approaches for managing missing data, stressing the need for specific strategies to ensure the integrity and reliability of analytical results.

*Code and data are available at: https://github.com/jwonc4602/Discussion-about-Missing-Data

2 Discussion

2.1 Categories of Missing Data

1. **Missing Completely At Random (MCAR):** When data is MCAR, the absence of data points is unrelated to any observed or unobserved variables. An example of MCAR is the removal of population data for three randomly selected states from a dataset. This scenario poses the least concern for bias in summary statistics and inference since the missingness is completely random. However, confirming that data is indeed MCAR is challenging.
2. **Missing at Random (MAR):** Data is considered MAR when the missingness is related to other observed variables in the dataset but not to the missing data itself. For instance, if males are less likely to respond to a question about income, this missingness is related to the observed variable of gender but not directly to the income data. Simulating a MAR scenario involves removing income data for states with the highest populations, affecting the analysis based on income-related variables.
3. **Missing Not At Random (MNAR):** MNAR occurs when the missingness is related to the missing data itself or unobserved variables. An example is higher-income respondents being less likely to report their income, possibly due to privacy concerns or other factors not captured in the dataset. This type of missingness is the most challenging to address because it can introduce significant bias into the analysis.

2.2 Strategies for Handling Missing Data

Handling missing data requires a thoughtful approach that considers the nature of the missingness and the specific context of the analysis. The following strategies offer different ways to mitigate the impact of missing data:

- **Dropping Observations:** This straightforward approach involves removing data points with missing values. While simple, it can lead to the loss of valuable information, especially if the missingness is not MCAR.
- **Imputing Values:** Imputation involves replacing missing values with substitutes, such as the mean of the available data. This method can help retain data points but may introduce bias, particularly if the data is MNAR.
- **Multiple Imputation:** A more sophisticated technique, multiple imputation creates several imputed datasets and combines them to account for the uncertainty around the missing values. This approach is often more robust but requires careful implementation and interpretation.

3 Conclusion

Missing data is a complex issue that necessitates careful consideration and strategic handling in the context of EDA. Understanding the type of missingness—MCAR, MAR, or MNAR—is crucial for choosing an appropriate strategy to address it. Whether through data omission, imputation, or more advanced methods like multiple imputation, the chosen approach must be applied judiciously, with an awareness of its limitations and potential biases. Ultimately, the goal is to minimize the impact of missing data on the analysis, ensuring that the findings are as accurate and reliable as possible. Furthermore, acknowledging and documenting the handling of missing data is essential for transparency and reproducibility in research. As the field evolves, ongoing exploration of methods for managing missing data will remain a key aspect of ensuring the quality and integrity of data analysis.

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC. https://doi.org/10.1201/9781003229407.

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