Missing Data*

What is missing data and what can we do about it?

Talia Fabregas

March 5, 2024

Datasets can have missing variables. These variables can be missing completely at random missing at randon; or missing not at random. Missing data can never be made up and there is no one-size-fits-all way to handle missing data. Understanding what data is missing, why it is missing, and how to handle it is an important part of exploratory data analysis.

1 Introduction

Sometimes, a data set can have missing variables. In fact, missing data is basically inevitable regardless of how thourough the data acquisition process is Alexander (2023). In surveys, an example of missing data is non-responses; non-response is actually an important thing to explore, especially in non-probability samples, because people often abstain from responding to a particular survey question for a reason, and there are often systemic differences between the people who responded and the people who abstained Alexander (2023).

This paper will explore the three different categories of missing data: missing completely at random; missing at random; and missing not at random, as well as examples of each. Furthermore, this paper will discuss how, and when, missing data can be appropriately handled by dropping observations with missing data, imputing the mean, using multiple imputations, or considering additional explanatory variables.

It is important to note that there is no "best" way to deal with missing data. This depends on our data set, category of missing data, and why certain data is missing. However, we can never make up missing data or compensate for it. Doing so would be highly unethical and would undermine the accuracy and credibility of our research.

*Code and data are available at: https://github.com/taliafabs/HandlingMissingData.git

2 Exploratory Data Analysis

Exploratory data analysis (EDA) is a dynamic, customized process. There is no one-size-fits-all approach to EDA. However, when conducting EDA our main goal is to understand the distribution and properties of individual variables, relationships about variables, and what is not there Alexander (2023). Thus, determining what data is missing, why it might be missing, and how we might be able to handle it appropriately is an important part of exploratory data analysis.

3 Types of Missing Data

There are three main categories of missing data: Missing Completely at Random; Missing at Random; and Missing not at Random Alexander (2023).

3.1 Missing Completely at Random

When data is missing completely at random (MCAR), the reason why certain values, variables, or observations are missing or incomplete is entirely independent of any variable present within or outside of the data set Alexander (2023). In a perfect world, missing values will be MCAR because this situation is the easiest one to deal with and. In this case, there is minimal concern about the effects that the removal observations with missing values might have on the summary statistics. In the context of a study aiming to learn about the effect of age, gender, income, and religion on the preferred presidential candidate of American adults, data might be missing completely at random if no age group, gender, income bracket, or religion is more or less likely to have answered the question about who their preferred presidential candidate is.

3.2 Missing at Random

When data is missing at random (MAR), the reason why certain observations are missing is related to other variables within the data set Alexander (2023). For example, if we are interested in learning about the effect of age, gender, income, and religion on one's preferred presidential candidate, and for some reason, females were less likely to respond to the survey question about who their preferred presidential candidate is, this data might be missing at random. If the increased likelihood of abstaining from answering the question is solely related to gender, and not another factor such as partisan affiliation, home state, or home county, then it might be just fine to continue as is as long as we take gender into account in the regression model or post-straify based on gender Alexander (2023). However, the likelihood of this independence holding is pretty low, and in all likelihood, there is a relationship between whether or not an individual respondend to the survey and political support (**mnar?**).

3.3 Missing not at Random

When data is missing not at random (MNAR), certain observations are missing from the data set for reasons relating to other variables present in the data set in a way that is related to unobserved variables or the missing variable itself Alexander (2023). If female respondents' decreased likelihood of answering the question about their preferred presidential candidate is in fact related to another variable outside the data set,such as partisan affiliation, home state, or home county, then the data would be missing not at random. Another possible example of MNAR in this same context is non-religious, young females and males also being less likely to answer who their preferred presidential candidate is. This could be due to the fact that left-leaning members of generation z are dissatisfied with President Joe Biden, or due to the fact that members of generation z just tend to be less interested in politics (interest in politics or perception of how important voting is could be an unobserved variable).

4 How to Handle Missing Data

There are multiple options for handling missing data. There is no "best" option or clear-cut way of doing this, so it is important to always examine how different ways of handling missing data will impact our results and to document exactly what was done and why it was done Alexander (2023). It is important to understand which category the missing data falls under before determining how to handle it.

4.1 Drop observations with missing data

The base R mean() function automatically excludes observations with missing values in its calculations R Core Team (2023). If data is missing completely at random, then omitting the observations that are missing a certain value should not have any significant effect on the summary statistics Alexander (2023). A way in which we can determine whether or not this is an option could be by randomly removing observations a few different times, and then comparing summary statistics. However, as previously noted, data is often not missing completely at random, so this approach could be risky and might not always be appropriate.

4.2 Impute the Mean of Observations Without Missing Data

Another option is to impute the mean of observations without missing data.To do this, we create a second data set that does not have any of the observations with missing data, compute the mean of the respective column, and then impute the mean values into the original data set where the missing values are Alexander (2023).

4.3 Multiple Imputation

Multiple imputation involves creating various potential data sets, doing inference, then combining them by averaging Hill (2007). However, multiple imputation is not a very versatile approach. It is often not the way to go because it can create biased results and is done under the assumption that the missing values would have no effect on the summary statistics. Before doing imputation, it is important to simulate the removal of observations and implement various options in order to understand what the trade offs might be Alexander (2023).

4.4 Additional Explanatory Variables

When data is MNAR, dropping incomplete observations and data could be considered, but they will often amplify biases and threaten the accuracy of the data analysis Alexander (2023). This is because when the reason why certain observations have certain missing values because of a variable outside of the data set, the best way to understand the missing values is to take the missing external variable into account. The addition of additional explanatory variables, such as party identification, ideological self-placement, and past vote might help us gain a better understanding of MNAR preferred presidential candidate response values Rothschild (2016).

5 Discussion

6 Missing values can be an important part of the story that a data set tells

Handling missing data appropriately is a crucial step towards producing an accurate study, model, or research paper. In the same hypothetical survey example previously used, if gen z non-religious respondents were more likely to have a missing preferred presidential candidate value, then this could be an important part of the story that the data set tells. The absence of a value in the preferred candidate column can be due to a respondent's political preferences, partisan affiliation (or lack thereof), or issue evaluations. This would mean that the data is missing not at random, and simply removing observations with missing data would almost certainly produce biased, and therefore inaccurate results. Therefore, in this case, we would need to find a way to incorporate the missing data into our data analysis, model, research paper, and most importantly, the story that we are using the data set to tell. If we tell a story with data that has observations that are missing not at random, but fail to choose an appropriate way to address this and explain why the data is missing, then the story will be very incomplete.

7 Conclusion

There's no such thing as a uniform, step-by-step approach for handling missing data. Each of the approaches outlined in this paper has its own pros and cons. Missing data is inevitable, and it is something that usually needs to be incorporated into the data analysis, rather than ignored. Understanding what data is missing, why it is missing, and if the reason why it is missing relates to the missing variable itself, another variable within the data set, or a variable outside the data set is a crucial first step towards finding an appropriate way to address it.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. University of Toronto Department of Statistical Sciences. https://www.tellingstorieswithdata.com.

Hill, Andrew Gelman; Jennifer. 2007. *Data Analysis Using Regression and Multilevel/Hierarchial Models*.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rothschild, Andrew Gelman; Sharad Goel; Douglas Rivers; David. 2016. *The Mythical Swing Voter*. Quarterly Journal of Political Science. https://www.nowpublishers.com/article/Details/QJPS-15031.