What to do about missing values*

Ahnaf Alam

February 21, 2024

Table of contents

1 General idea

Missing data points are a regular feature in almost all datasets. The defining task for researchers is figure out the source of the missing data as these can lead to bias and error (Newman 2014). Missing data can arise for many reasons, however, we can categorize missing values into three distinct categories:

*For further information, please visit https://github.com/AhnafAlam1/missing-values

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

The subsequent paragraph will focus in discussion what these categories of missing values, their features, reasons and what can researchers do to mitigate the issue of missing data.

2 Missing completely at random

This situation arises when data points are missing independent of any other variables (Alexander 2023). In layman's terms, this points to a situation when data points are missing at random and it is not related to any other factors or variables in the data set. This coincidental situation implies the data set in hand, is still representative of the broader population. Amongst the categories of missing data, this form of missing data is the only one that is completely random. However, this form of missing data is also extremely rare as this type of data requires reserchers to make "untenable assumptions" (Graham 2009).

3 Missing at random

MAR differentiates from MCAR in that missing values are related to one of the of the other variables in the data set. Take for example, a situation where a researcher is interested in differing Indian caste system's opinion on country's current federal government. In this situation, we will have MCAR, if we have missing values for opinion variable for lower castes compared to higher castes. MAR is systematic missing data mechanism (Newman 2014), unlike MCAR.

4 Missing not at random

MCAR occurs when missing data is related to unobserved or latent variable or that missing data is related to the missing variable itself. Suppose you are collecting data on patients weights in a medical clinic. You ask patients to self-report their weight to researcher every two weeks. Now, imagine individuals with higher weight level may feel insecure to share their weight details and therefore, they may not report their weight on a regular basis. In this situation, higher weights is related to missing data for weight. It is often impossible to determine if data is MNAR or not as this requires researcher to compare observed values to missing values and researcher does not have access to missing values (Newman 2014).

5 What to do about missing data

There are five common ways to deal with missing values in data sets (Newman 2014). They are:

1. Listwise deletion
2. Pairwise deletion
3. Single imputation
4. Maximum likelihood
5. Multiple imputation

5.1 Listwise deletion

The basic idea behind listwise deletion is to delete all cases where data is missing and proceed with the analysis. The obvious issue with this type of mechanism is we risk reducing sample size significantly, especially if there's a lot of missing data points. This is bad as lowering count of observations means the power for statistical tests significance, and this could lead to biased estimate, particularly for MAR and MNAR.

5.2 Pairwise deletion

Pairwise deletion omits data based on variables used in the analysis. In other words, analysis may be completed based on subset of data depending on where the missing value is (Moran 2024). For example, if person X has missing data for income, however, has complete data for race, height and weight, then X will be included in all analysis that involves race, height and weight but excluded from analysis that includes income. This, again, could lead to biased estimate, espcially for MAR and MNAR.

5.3 Single imputation

Single imputation involves filling in the missing value using some form of estimate, usually being mean, median or mode. However, doing this leads to underestimation of variance and this methods ignores relationship with other variables (Zhang 2016).

5.4 Maximum likelihood

This method involves directly estimating parameters of interest from incomplete data and proceeding with analysis based on this (Newman 2014). This form of estimation relies on probability when estimating missing data. It considers likelihood of scenarios and estimates based on what is more probable given the existing data. One of the positives for this mechanism is that it is unbiased under MCAR and MAR (Newman 2014).

5.5 Multiple imputation

This mechanism works by drawing “repeated simulated datasets from a posterior distribution defined by the missing data conditional observed data” (Pepinsky 2018). In simpler terms, this method generates several potential solution for missing data, conditional on what you already know from the data. This is a very handy model as improves as you add more variables to the model and further, it is unbiased under MCAR and MAR (Newman 2014).

6 Conclusion

Missing data can arise with for various and there are various methods to deal. Some are more preferred my researchers as analysis leads to unbiased estimates. However, for most cases, it’s the type of missing value that dictates what sort of mechanism is needed to deal with this issue. Table 1 provides a summary for all the different mechanisms used to deal with missing data, along with their benefits and drawbacks, based on (Newman 2014).

Table 1: Table display major benefits and solutions for each of the missing data mechanism. It is based on Newman (2014).

Potential solutions	Benefits and drawbacks
Likewise Deletion	Biased under MAR and MNAR
Pairwise Deletion	Biased under MAR and MNAR
Single Imputation	Biased under MCAR
Maximum Likelihood	Unbiased under MAR and MCAR. The model improves as you add more variable
Multiple Imputation	Unbiased under MCAR and MAR. Also improves as you add more variables to the model. However, the model gives slightly different estimate each time

Citations

Alexander, Rohan. 2023. "Exploratory Data Analysis." *Telling Stories with Data - 11 Exploratory Data Analysis*. https://tellingstorieswithdata.com/11-eda.html#missing-data.

Graham, J. W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60: 549–76.

Moran, Melissa. 2024. "Handling Missing Data: Listwise Versus Pairwise Deletion." *Statistics Solutions*. https://www.statisticssolutions.com/handling-missing-data-listwise-versus-pairwise-deletion/#:~:text=Pairwise%20deletion%20omits%20cases%20based,on%20where%20values%20are%20missing.

Newman, Daniel A. 2014. "Missing Data: Five Practical Guidelines." *Organizational Research Methods* 17 (4): 372–411.

Pepinsky, Thomas B. 2018. "A Note on Listwise Deletion Versus Multiple Imputation." *Political Analysis* 26 (4): 480–88.

Zhang, Z. 2016. "Missing Data Imputation: Focusing on Single Imputation." *Annals of Translational Medicine* 4 (1): 9. https://doi.org/10.3978/j.issn.2305-5839.2015.12.38.