What is Missing Data and What Should You Do About It

Rahul Gopeesingh

2024-03-05

Missing Data

Very frequently data is missing. This can be due to the method of data collection or certain observations that are difficult to measure. There are three types of missing data(Alexander 2023):

1. Missing Completely at Random(MCAR)
2. Missing at Random(MAR)
3. Missing Not at Random(MNAR)

MCAR

MCAR is by far the easiest to deal with as it suggests that missing data is unrelated to any other variable, whether observed or not. Despite this being the easiest of the three, it is usually the least common. Suppose we had a dataset of footballers and their stats throughout the season(goals, assists, headers, dribbles etc.) An example of MCAR may be missing data for the amount of time spent with the ball as this can be extremely difficult to measure.

MAR

MAR occurs when the data missing is dependent on some variable that is in the dataset. In the same example of football, suppose that there is a lot of missing entries for the number of assists, but it is way more likely that a player who played less games would be missing this entry than one who played more(maybe it is the case that whoever was collecting data paid less attention to the players that played less) This would be considered MAR as the missing data is related to an observed variable. It is important to really think about the way in which the data is missing

MNAR

MNAR occurs when the missing data is related to a variable that is not in the dataset. Considering the same example of footballers, maybe the observer only counted the goals of the strikers and forwards. The defenders' goals would be missing however the observer did not document the position of the players in the dataset. This is more difficult to deal with than MAR because little is known about the entries for which missing data occurs. (This will be understood more clearly once the methods are discussed.)

How to Deal With Missing Data

There are three methods to deal with missing data and choosing between them requires careful thought about the data that is missing. These include:

1. Ignoring all observations with missing data
2. Imputing the mean of observations without missing data
3. Using a technique called multiple imputation

Each of these have a place and none are perfect. In many cases the actual value will vary greatly from the simulated value and hence lead to inaccuracies. Regardless, one must deal with missing data somehow. A good practice would be to 'hide' some values(create missing data) and try all three methods to simulate missing data. This will allow us to compare simulated data with actual data. This should be done more than once, hiding a different set of data each time. Whichever of the three methods seem the most accurate should be used for the truly missing data.

References

[1]

Alexander, Rohan. 2023. *Telling Stories with Data*. https://tellingstorieswithdata.com/.

[1]Code and data supporting this analysis is available at: https://github.com/Rahul-Uoft/Missing-Data.git