Mini Essay 8

Cristina Burca

2024-03-05

Missing data is a poses a common challenge encountered by most statisticians across various research domains. It is first important to deal with missing data correctly, so that results are not skewed or biased from missing information. Firstly, one must establish what type of data is missing to know how to correctly replace or account for the incomplete data. The three main categories of missing data are 1. Missing Completely at Random, 2. Missing at Random, and 3. Not missing at Random.

When data is 'Missing Completely at Random' (MCAR), this data is independent of any other observed data. The missing values are not related to any other values, and are randomly distributed throughout the data. The fact that these values are missing does not relate to or affect any of the other recorded values. An example would be if students in a classroom had independent math and science test scores recorded, and some students had a missing score, which could be due to the fact that they missed the test, or human error caused a mistake in the recording of the scores. The fact that some of these test scores does not affect the rest of the data.

Data that is 'Missing at Random' (MAR) is different because these missing values are related to other variables in the dataset. These values that are missing can be explained or predicted by the other variables. An example would be the recording of annual income- men might be more likely to disclose their income to a survey compared to women, and thus the missing income values for women may be accounted for from the fact that women may not be as comfortable disclosing their income. The missing income data depends on the gender, but within the gender variable, the data that is missing is completely up to chance.

Data that is 'Missing not at Random' (MNAR) is characterized by its relationship to the unobserved 'missing' variables. This type of missing data cannot be predicted from the observed data. If we recorded annual incomes of men and women, those with a lower income might be less inclined to disclose their income. Then, the missing values depend on the value of the income itself, but other factors like gender do not matter, and it is not of importance who did or did not disclose that information.

Each type of missing data is dealt with separately, and that is why it is important to classify the type of missing data it is. For data that is MCAR, it does not depend or relate to other

observed or unobserved values, and thus the row of data can be removed and unconsidered without skewing or altering the result. This is because the parameter estimates are unbiased, and thus the missing data can come from anywhere on the distribution. The only disadvantage for this method is that the sample size decreases in size as rows of data are removed from the dataset, giving you a smaller dataset to work with.

Furthermore, these missing values can be replaced so that instead of these observations not being considered in the dataset at all, we can still use them, so we reduce the bias factor and keep the same sample size. We do this by dropping observations with missing data, and using the function mean() to calculate the mean of a variable, excluding all observations with missing values in the calculation. We can then create a new dataset, transferring any data we already have, and replace the missing values with the new mean value we obtained.

Moreover, multiple imputation is a more flexible method that can be used for any type of missing data. Simulating the removal of observations and implementing various options can help us understand the trade-offs we face when comparing different simulated imputations. When we having MCAR or MAR, multiple imputation is a considered method to replace missing values. This is done by creating several versions of the original dataset, with the only difference being the imputed values that were calculated. The data should be first fit to an appropriate model, missing points in the dataset should be estimated through simulation. The data should be fitted and simulated multiple times, to get different variations of the data, and to have variables to compare when considering different models for each option. Then, data analysis tests should be run on each variation of the model. Then, the values of the parameter estimates should be averaged, and any other values calculated like standard deviation or variance should be obtained from each model and compared.

This approach allows for variation in the results, and allows for comparison of the different variations of the simulated results to visualize the trade-off of values when considering certain parameters. But overall, nothing can make up for missing data, as simulated values will never be equal to real life observations. It is evident that the consequences of missing data extend far beyond statistical inconvenience, but can cause doubt over the validity and reliability of research findings.