What is missing data? Methods of Handling Missing Data*

Zhijun Zhong

March 4, 2024

Table of contents

1 Introduction

Missing data in all areas of research is a common challenge that destroyed the honesty and integrity of statistical analysis. As mentioned by Derrick A in his comprehensive examination of the issue (Bennett 2001), missing data will distort the result of the research making them biased and inaccurate. Note that Bennett is from Department of Medicine which often take missing data on patients which will have a lot of experience on this. He divided the missing data into 3 difference category including, Missing at random (MAR), Missing completely at

*Paper is Availiable at https://github.com/JerrZzzz/Handling-Missing-data.git

random(MCAR) and not missing at random (NMAR). This classification is crucial in using different ways to handle different missing data for each category.

2 Classcification for Missing data

We call the missing data which does not bias the results MCAR(Bennett 2001). It means that there is no connection between observed values and the unobserved values. In this situation, it is like all the missing data is a set of random number which will provide no bias to the dataset. However, when the missing data is relate to the observed data but not the missing data itself, then we call it MAR(Bennett 2001). It means that even though there is relationship, but we can provide an explanation though modeling and so on. Lastly, when the probability of missing data is relate to the unobserved data which implies a bias, then we call it NMAR(Bennett 2001). Bennett highlights that we have to create models for these NMARs so that we can accurately estimate parameters in our interest otherwise we will probably made wrong conclusions.

3 Ignore Missing data

A popular way of handling missing data is ignoring them. But, it does not mean that we can just delete them and pretend they never exist. As Bennett mention in his paper (Bennett 2001) that there is 2 primary approaches: complete case analysis and available case analysis.

For complete case analysis, it is defined and used by may companies which only participants who had a complete file will be included in our research. Bennett explained to us that it can be biased when the missing data is classified as MCAR(Bennett 2001). Since that it is completely random, so it will rarely happen in our world. Thus, using this method, our analysis of a potential non-representative sub-dataset will lead to a biased results.

For an Available case analysis commonly known as pairwise delection means we can use all the available observations which unlike the complete case analysis, will not delete any observation because of any missing values. Bennett also warned us the limitation of this method where it can lead to potential bias and inconsistency. Since that the number of sample size are different for each column, so inconsistency can be explained when we focus on specific columns.

4 Other methods

There are different methods mentioned by Bennett (Bennett 2001) including imputation methods, Likelihood-based methods and Indicator methods.

4.1 Imputation methods

Single imputation methods is a way to replace the missing values(Bennett 2001). For example, Last value carried forward is a method involves putting the last value to fill the missing value. It can be seriously biased if the data is MAR and NMAR. We can also use the mean to fill the missing value. It seemed simple without deleting any observation but it will lead to an underestimation of the variance. Regression model can also be used to fill the missing value. By creating a model and predicting the missing value is a good way. However, it is complex and we have to assume that the missing data is classified as MAR. For a hot deck imputation, we can replace the missing value as some similar case within the dataset. It is simple and can be less biased than mean substitution. However, it will be complex when we trying the match the cases. Lastly, cold deck imputation different from hot deck imputation where it uses external information to match different observations.

There is also some other imputations that are less common but can be useful in some specific situations. For a multiple imputation where are create multiple imputed datasets by replacing each missing values which a set of possible values. Then, do an analyze on those datasets which we can put the results together to produce estimate of the uncertainty of the missing data. This method works well on the cross section of the data where we need to assume the data is MAR. We can also use Markov-Chain Imputation which works primarily to the longitudinal of the data. In this method, it combines the progression of a condition over time to the imputation process. If we can understand the transition between different times, we can understand the likelihood of the value when they transtition from time to time.

4.2 Likelihood-based methods

Likelihood-based methods includes 2 approaches: expectation-maximization approach and raw maximum likelihood. For expectation-maximization approach, we will alternate between "estimate the missing value"(Bennett 2001) and "maximizing the likelihood functions for parameter estimation"(Bennett 2001). The estimating of missing value involves using our non-missing data to estimate the distribution of our missing values while maximizing likelihood function for parameter estimation updates as there were no missing data. It is most effective when our missing data is MAR. However, if we have extensive missing data, it can be slow to converge. On the other hand, Raw maximum likelihood (Bennett 2001) uses observed data to estimate missing values through the maximum likelihood. In this approach, unlike the drawback of expectation-maximization approach, it can handle large amount of missing data.

4.3 Indicator method

For indicator method, Bennett uses 2 approaches including indicator method and Pattern-Mixture Models (Bennett 2001). For indicator methods, every missing value will create a indicator where it indicated the value is missing(1) or not missing(2). So the indicator value

and the original value are all combined in the dataset which we can include the missing value in our model. Bennett highlights that this is popular but it can be biased when the missing value is a predictor in the model. For Pattern-Mixture Models, they stratify the analyze by the pattern of the missing data, seeing the dataset as a mixture of sub-populations. In this way, we are able to handle complex situations when we analyze patterned separately then combine the results.

5 Conclusion

Missing values common in our data analysis. Using the correct way of treating and dealing with them is important and challenging. Eliminate all the bias from missing value is difficult to achieve. We need to determine the classification of the missing value so that we can choose the right way to handle our missing value.

Reference

Bennett, Derrick A. 2001. "How Can i Deal with Missing Data in My Study?" *Australian and New Zealand Journal of Public Health* 25 (5): 464–69. https://doi.org/https://doi.org/10.1111/j.1467-842X.2001.tb00294.x.