Definition of missing data and approches to do about it*

Yunshu Zhang

March 5, 2024

Table of contents

*Code and data are available at: https://github.com/Yunshu921/Article_missing_data.git

1 Introduction

When we prepare to do the statistic analysis, usually we need to clean our data at first. One of most important process is to eliminate the bias of the data. In particular, missing data may cause biased estimates. In this article, I want to talk about the the definition of missing data and how to deal with it under different conditions.

2 Categories of missing data

2.1 MCAR

When data is MCAR, the data is missing regardless of the observed value or unobserved value. It is typically safe to remove MCAR data because analyzing solely the cases that have no missing values can lead to valid conclusions ((2015a)).

2.2 MAR

In cases where missing data is MAR, the likelihood that a value is absent is typically determined by its observed values. For example, we want to gather information on the property tax band and income. Higher earners are typically less inclined to reveal them. As a result, a simple average of respondents' incomes will have a right skew. However, now assume we firstly know all people's property tax band and secondly given property tax band people give no response to the income question randomly. Then the missing income data is considered to be missing at random. This means the reason for the absence of income data depends on the property tax band. Once the property tax band is given, the missingness of income data is not dependent on the income level itself. At this time, a simple summary metric, such as the mean income determined from the given data, was biased. However, we can get an accurate result from a simple model that anticipates income based on property tax band, accounting for the feature that causes the data to be MAR ((2015b)).

2.3 MNAR

The MNAR category applies when the probability that data is missing dependes on the unobserved or missing values themselves. Because the missing information is unknown, knowing the appropriate model for the missingness process is quite hard ((2015c)).

2.4 Summery

if missing data are MCAR or MAR, the missingness mechanism is ignorable. But for data are MNAR, the missingness mechanism is non-ignorable.

3 Methods

In this section, we explore will single imputation methods and likelihood methods.

3.1 Single Imputation Methods

By replacing missing values with estimates, single imputation completes the dataset and prepares it for study. Based on Little, Roderick and Rubin, Donald (Little and Rubin (2020)), we need to firstly get the distribution of observed data. Then use the distribution to do single or multiple imputation. To generate this distribution, there are two general methods which are explicit modeling and implicit modeling. In this article, we only focus on explicit modeling.

Explicit modeling methods contain the following:

3.1.1 Mean imputation

In this way, we suppose the distribution of mean is constant for observed and unobserved values. Thus we can use the mean of all observed values to fill into missing data.

3.1.2 Substitution

A method for dealing with missing data are in a large fieldwork stage of research, replaces missing data with alternative units not yet selected into the sample (Little and Rubin (2020)).

3.1.3 Regression imputation:

Firstly, we need to make a regression model based on observed values. Then we can replace missing data with their predicted values from the relationships between the missing data and other observed values in the data.

3.2 Likelihood Estimation

3.2.1 The likelihood function

For discrete data, the likelihood is the joint probability of observed data considered as a function of the unknown parameter . ### Maximum likelihood estimation In the discrete data, given sample data x1,…,xn the maximum likelihood estimate for is the value $\hat{\theta}$ that maximizes the joint probability of the observed data, i.e. that maximizes the value of the likelihood function L( ).

Now, based on Little, Roderick and Rubin, Donald (Little and Rubin (2020)), we can define the likelihood of ignoring the missingness mechanism and then we should think about the circumstances in which this simpler likelihood can be used to draw conclusions about . Under different conditions, we have specific formulas and prerequisite to estimate the missing data.

4 Acknowledgements

I would like to express my deepest thinks to Dingning Li, whose invaluable guidance and insightful feedback were instrumental in the completion of this paper.

References

2015a. *Web.archive.org*. https://web.archive.org/web/20150910180725/http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=75:missing-completely-at-random-mcar&catid=40:missingness-mechanisms&Itemid=96.

———. 2015b. *Web.archive.org*. https://web.archive.org/web/20150910180057/http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=76:missing-at-random-mar&catid=40:missingness-mechanisms&Itemid=96.

———. 2015c. *Web.archive.org*. https://web.archive.org/web/20151006215548/http://missingdata.lshtm.ac.uk/index.php?option=com_content&view=article&id=77%3Amissing-not-at-random-mnar&catid=40%3Amissingness-mechanisms&Itemid=96.

Little, Roderick J A, and Donald B Rubin. 2020. *Statistical Analysis with Missing Data*. Hoboken, Nj: John Wiley & Sons, Inc.