Missing data: a simple introduction

Marius Dioli

February 2019

1 Introduction

If one is to do anything of import in science, one has to manipulate and analyze data. However, due to a variety of factors, data is often flawed. One major flaw one can have, and the topic of this thesis, is missing data. The consequence of missing data is always a reduction in efficiency and may also lead to bias.

//A variety of ad hoc approaches are commonly used to deal with missing data, and many of the techniques described in this thesis are either//

//This thesis will approach the problem of missing data from the bayesian perspective and with the goal of imputing or removing data. Other methods motivated from the frequentist perspective such as likelihood based methods will be touched on breifly as alternatives. The strength of the bayesian approach is that it provides a principled way to account for uncertainty about the missingness mechanism and the true distribution of the full data. //

To quote MENG: "imputation is not (merely) a compu- tational tool but rather a mode of inference, which al- lows hierarchical and sequential input of assessment and information" [Multiple-Imputation Inferences with Uncongenial Sources of Input]

//Missing data is almost always considered in the context of fitting a model or making inference.//

.. We will measure its efficacy on synthetic data and real data before comparing it with an alternate method that is gaining in popularity, machine learning.

This paper is by no means comprehensive, but strives to give the reader a proper introduction to the problem of missing data and the technique of multiple imputation. // fyldig

As detailed in Xie and Meng [DISSECTING MULTIPLE IMPUTATION FROM], much of modern data analysis is done in a multi-phase paradigm, where the distinct and sequential phases data acquisition, data pre-processing, and data analysis, can have different or even contradictory assumptions.

2 Terminology and basic missingness theory

In the following section I will mostly adhere to the notation of Rubin in (citation), as it is both well established and quite general. However, I will adapt it slightly by using the observed/missing notation in (missing data) for readability, as it is easier for an unfamiliar audience and non-specialists to understand.

2.1 What is missingness?

When we say a variable is missing we mean that we have not observed it. The variable still has a value, but this value is hidden from us for some reason. What this reason might be will be described in detail later in this section. By treating missingness as a probabalistic phenomenon, statistically rigorous tools for dealing with it become available.

Let $Y = (y_{ij})$ be a nxk complete data matrix (i.e. with no missing values), with the ith row $y_i = (y_{i1}, ..., y_{ik})$ where y_{ij} is the value of variable Y_j for unit i. Let also the missingness indicator matrix $M = (m_{ij})$ be an nxk such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. Both Y and M are random variables, and we assume for simplicity that the rows (y_i, m_i) are independent and identically distributed over i. It is possible to have more than two possible values for each entry of the M, for example when we wish to differentiate between different kinds of non-response. However, we will not explore this in this paper. Now, given m_i, we can partition y_i into y_i^o and y_i^m, corresponding to the components that are observed and missing respectively (where the corresponding entry in m_i is 0 or 1 respectively).

2.2 The missingness mechanism

The missingness mechanism is characterized by the conditional distribution of m_i given y_i, i.e. $f_{M|Y}(m_i|y_i, \theta)$, where θ denotes unknown parameters. Stated differently the missingness mechanism is the probability of data being missing given the values of Y, missing and observed.

The missingness pattern in Y does not matter for the following definitions, and can be any of the following: *(Figure from page 9 Rubin)* However,

the missingness pattern will affect the efficacy of methods used to deal with missingness, and we shall explore this in subsequent sections.

We distinguish between three different missingness mechanisms by considering how M is related to Y: Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR). These can be roughly thought of as being in order of strictness from strictest (MCAR), to least strict (MNAR). Rubin(1976) [missing data book] introduced these missingness mechanisms, and they determine which techniques are appropriate to use on a given problem. **These mechanisms are rarely the case in practice, but it will be shown later that assuming them is reasonable in many circumstances. **

Missing Completely At Random (MCAR): This kind of missingness occurs when, $\forall i$ and any distinct values y_i, y_i^* in the sample space of Y, $f_{M|Y}(m_i|y_i, \theta) = f_{M|Y}(m_i|y_i^*, \phi)$. In other words, it means that the probability of missingness is independent of both the observed and the missing data, and we can view the observed data as a random sample of the complete data.

Missing At Random (MAR): In this case, $\forall i$ and any distinct values y_i^m, y_i^{m*} in the sample space of y_i^m, $f_{M|Y}(m_i|y_i^o, y_i^m, \theta) = f_{M|Y}(m_i|y_i^o, y_i^{m*}, \theta)$. This means that the probability of missingness, conditioned on the observed data, is independent of the missing data. Assuming MAR is less restrictive than MCAR, and may be considered a more plausible assumption about missing data in many contexts.

Missing Not At Random (MNAR): If the above equation does not hold for some i and some y_i^m, y_i^{m*}, then the missingness mechanism is called missing not at random (MNAR). This means that the probability of missingness is dependent on the missing data even after conditioning on the observed data.

2.3 Illustrative example

The following example illustrates the three different types of missingness.

Suppose you are doing a study in which you want to investigate people's sleep patterns. You send N people a survey asking them for the average amount of sleep they got each night in 2018. You send a follow-up survey for 2019. Table 1 shows simulated data for N = 15 people drawn from a bivariate normal distribution with $\mu_x = \mu_y = 7.5$, $\sigma_x = \sigma_y = 2$, and a correlation $\rho = 0.4$. The first two columns show the complete data for X and Y, while the other three columns show the values for 2019 after imposing missingness by three methods.

In the MCAR scenario, we see a random spread of missing data, i.e. a random subset of those who answered the 2018 survey also answered the 2019 one. Things get more interesting in the next two columns. Suppose you are

	2018	2019 (Complete)	2019 (MCAR)	2019 (MAR)	2019 (MNAR)
1	8.931665	7.417407	8.931665	0.000000	7.417407
2	9.371777	7.271854	9.371777	0.000000	7.271854
3	8.419175	7.894072	0.000000	0.000000	0.000000
4	5.280704	5.987574	0.000000	5.987574	5.987574
5	11.124287	7.430975	11.124287	0.000000	7.430975
6	8.444062	10.664983	0.000000	0.000000	0.000000
7	6.412436	8.766714	0.000000	8.766714	0.000000
8	5.562886	3.465600	0.000000	3.465600	3.465600
9	5.563854	7.068254	5.563854	7.068254	7.068254
10	8.170268	6.869599	8.170268	0.000000	6.869599
11	6.596475	5.423074	0.000000	5.423074	5.423074
12	6.613481	6.713958	0.000000	6.713958	6.713958
13	8.594411	9.179908	0.000000	0.000000	0.000000
14	7.724661	6.144757	7.724661	0.000000	6.144757
15	11.241391	7.978949	0.000000	0.000000	0.000000

Figure 1: Different missingness mechanisms

only interested in those who've slept less than 7.5 hours on average, so you only send a follow-up survey to those who reported less than 7.5 hours of sleep for 2018 (2018¡7.5). This is MAR, since the missingness of the 2019 entries depends on the values for 2018. For the final column, suppose you sent out your 2019 survey to all the individuals who answered the 2018 one, but include a line stating "Only answer this survey if you slept less than 7.5 hours on average in 2019" (2019 ¡ 7.5). This is MNAR, since the missingness of the 2019 entries depends on the values for those entries. Remember, the full dataset exists, but it has been censored due to some mechanism, in this case through your own design.

2.4 The validity of missingness mechanisms

In controlled environemnts, like the one from the our example, we can guarantee that data are MAR since we control the censoring of the data (in this case by placing a cutoff between 2018 and 2019). However, it is rare that we have this much control over our data and a compelling reason to censor it so. In most real world scenarios we only have the data at hand to inform us about the missingnsess mechanism. Unfortunately, we cannot infer the missingness mechanism from data ... We can check the validity of MCAR against MAR, but only under the unverifiable assumption that the data is not MNAR. Neither can we use the data at hand to support or refute on specific MNAR mechanism over another. The literature stresses the importance of sensitivity analyses in this case, checking the sensitivity of our inference under a variety of plausible assumptions. Finally MAR cannot be checked empirically against MNAR (Cite from page 8, 9, and 10 in missing data). We therefore have to assume the missingness mechanism, most often MAR, and there are strong arguments for this. First and foremost is that assuming MAR yields good empirical results, which will be shown in later sections.

3 Imputation

3.1 What is imputation?

Imputation refers to substituting or "filling in" missing values with plausible estimates. The main strength of imputation is that it does not change the nature of the subsequent analysis. It makes it much easier for the data processor and analyst to work together since the processor's techniques won't interfere with the analysts'. Another benefit of imputation is that it augments the dataset, meaning that the dataset being analysed is larger than the original dataset, potentially yielding higher performance.

3.2 Congeniality: when the imputer's and analysts models differ

As mentioned in the introduction, modern data analysis' multiple phases mean that the assumptions underlying each phase and the relationship between the phases is an important object of study. Meng's [meng 1994] concept of congeniality provides a theoretical framework from which to evaluate these differences.

Congeniality is an important concept in imputation and has a big impact on the statistical validity of MI. In short, uncongeniality arises when the analysis procedure differs substantially from the imputation model. However, as Meng (cite) states, "If the imputer's model is reasonably accu- rate, then following the multiple-imputation recipe prevents the analyst from producing inferences with serious nonresponse biases" Congeniality [Meng]

"When the imputation model class and the (embedded) analyst's model class differ, the behavior of Rubin's rules becomes very complicated, capable of producing inconsistent variance estimators, a matter that has received recurrent criticisms" [Meng 2017]

3.3 Techniques of interest

To properly evaluate the performance of MI, we should compare it to other similar and commonly used techniques, i.e. techniques that manipulate the dataset by either imputing or removing values. When choosing to apply a method it is important to keep two questions in mind. First, does the method under consideration give us consistent estimators for our model? Second, does the method give us appropriate measures of precision? How the following methods answer these questions is how we will decide the superiority of one method over another.

There are two other competing methods which we should touch on breifly before continuing. Weighting procedures and likelihood-based methods are ... alternatives to multiple imputation. However, there is reason to prefer MI as the default options over these for a few reasons. First that, given congeniality, MI tends to outperform likelihood-based methods for small sample sizes, while delivering similar performance for large sample sizes [Schafer, 2016].

3.4 Ad hoc methods

Carpenter and Kenward (Missing data handbook) ...

Case deletion, also called complete case analysis (CCA): A very common practice, CCA involves restricting analysis to entries with com-

plete data. For those familiar with programming, a typical example would be to remove all rows from a dataframe with NaN entries. The first and most obvious problem with this approach is that it drastically reduces the amount of useable data. Additionally, this technique introduces large biases in the data if missingness is MAR or MNAR (CITE REVIEW PAPER FROM TURID). Only under MCAR is the data not biased when conducting CCA, but one is still faced with reduced accuracy due to missing data.

Mean imputation:

Hot Deck imputation:

3.5 Statistically principled methods

Single imputation

Maximum likelihood and mixed linear models... This is the current alternative to the main method which we will discuss in the next chapter, multiple imputation. MI is often preferred over ML and MLM for two reasons: 1. It often performs similarly to multiple imputation with only a small advantage on certain kinds of data. 2. It is less easy to use for non-statisticians A

4 Multiple imputation

Highly robust technique that yields gains even with relatively small portions of missing data.

Only become feasible in the last 20 years due to computing power required.

Bayesian approach.

Between imputation variance

4.1 Specifying the imputation model

4.2 Rubin's rules and Meng's adaptations

Under congeniality, Rubin's rules for calculating standard errors, parameter values, and variance suffice.

If models are uncongenial, we use Meng's rules.

4.3 Potential pitfalls and criticism

When there is missing data in several variables of different types ,for example continuous and binary, and the data is structured, specifying the imputation model can be quite difficult. This is because it has to be a joint model...

Multiple imputation is motivated from the Bayesian perspective, yet survey inferences, its primary application area thus far, are traditionally dominated by frequentist analyses. Among all the criticisms of multiple imputation, the recent work of Fay (1991, 1992) is the most intense, as it is directed at the validity of multiple-imputation in- ference in practice. Fay provided several examples (see also Kott, 1992) to show that the variance esti- mator obtained from the repeated-imputation combining rules disagrees asymptotically with the sampling variance of the repeated-imputation estimator, even when the imputation model is correctly speci- fied. [MENG]

Taking all this into account, it is important to see MI in context of the options available to someone working with data. As we have established earlier, simply ignoring missing data can lead to serious biases, while alternative methods yield similar if not inferior performance in certain cases. As long as a potential user is aware of the potential pitfalls, MI is the superior method for dealing with missing data.

4.4 When to use multiple imputation?

Computational resources and time to create the full joint imputation model. (Handbook 236)

A common measure used to decide when to use multiple imputation is percentage of missing data.

However, Madley-Dowd et. al. (Proportion of missing data paper) propose a different measure.

MI can be used to perform sensitivity analysis by varying the imputation model.

5 Experiments

In this section we will use Madley-Dowd et. al.'s missingness measure to...

5.1 MICE

"The increased flexibility in modeling these conditional distributions may outweigh the lack of clear theoretical justification of the method." Rubin

5.2 Synthetic example

"The MCAR missingness mechanism removed the first p observations such that pn gives the required proportion of missing data. MAR missingness was simulated under a logistic regression model using

The value of a was manipulated for the different simu- lation settings to provide the required proportion of missing data on average across data sets." from (P. Madley-Dowd et a)

Also check out their motivation for the generating of the dataa

5.2.1 Inference

Test statistics Different tests

Congenial vs uncongenial models, see the annotated paper for how they did things with regard to imputation vs analysis.

5.2.2 Prediction

5.2.3 Classification

5.3 Performance across different data types

Continuous Categorical (binary, multicategory) Time to event Timeseries

5.4 Sensitivity analysis

5.5 Real data

Description of the data

HUNT Norwegian data

Andreas data webSite balanced number of classes for classification

5.6 TODO

Difference between baysian and frequentist paradigm. Citation example: [Grund et al., 2016, P.10].

References

[Grund et al., 2016] Grund, S., Lüdtke, O., and Robitzsch, A. (2016). Pooling ANOVA results from multiply imputed datasets: A simulation study. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 12(3):75–88.

[Schafer, 2016] Schafer, J. L. (2016). Multiple imputation: a primer:. *Statistical Methods in Medical Research*.