Imputation of missing data
A brief introduction

Abdelwahid Benslimane
wahid.benslimane@gmail.com

I) Introduction

During the collection process of the data which will be used to develop decision models, it is not uncommon for observations to be incomplete, meaning that some values of one or more variables are missing. This can occur, for example, when certain individuals do not respond to specific survey questions, or when certain responses are not entered into the computer system. It may also be related to data coding issues (such as the presence of non-interpretable characters) that make some values unusable. In the case of measurements from sensors, one of the sensors may no longer function or may operate intermittently.

A simple and frequently employed solution is to ignore these incomplete observations, but this approach can lead to a decrease in model performance, or even to a strong modelling bias that renders the model unusable.

A better solution is to impute the missing data, after characterizing the phenomenon that leads to the absence of certain values, and treat the estimated values as measured values.

The present document provides a simple definition of various scenarios in which data may be missing. Following that, a few methods for data imputation will be briefly described.

II) Missing data scenarios

Missing data can be characterized as follows:

- **Data Missing Completely At Random (MCAR)**

In this case, the probability of the value of a variable being missing is identical for all observations (and all variables with missing values). This case is rare. Ignoring observations with missing data is equivalent to using a random sample of complete observations and does not introduce specific biases into the modelling but may decrease the quality of the resulting

model if the number of complete observations is not sufficiently high. In particular, in the presence of imbalanced classes, this reduction in the number of exploitable observations for modeling can have a significant impact, as the rarer classes may be represented by a very insufficient number of complete observations.

- **Data Missing Not At Random (MNAR)**

In this case, the probability of the absence of a variable's value depends on variables that have not been observed. We are still in the MNAR case if the lack of response to a question depends on the value that the response would have if it were present. During a survey, some questions may indeed be avoided because the answers could be difficult for the respondents to admit. A common example is the case where individuals with a high income refuse to disclose it. Ignoring observations with missing data introduces a bias into the modelling because it is equivalent to eliminating observations in a non-random manner. However, since the absence of a variable's value does not depend on observed variables, imputing missing data based on the values taken by observed variables is more challenging to justify.

- **Data Missing At Random (MAR)**

In this case, the probability of a variable's value being absent is totally independent of the value the variable would have had if it were not missing (hence the data are missing "at random"). However, this probability depends on the values taken by other variables that have been observed. For example, we are in this situation (MAR) when a participant is more likely to not respond to a survey question if they have given a certain response (properly recorded, not missing) to a previous question. In the MAR case, ignoring observations with missing data introduces bias into the modeling. Indeed, if the value of a variable is missing when other observed variables take certain (combinations of) values, disregarding incomplete observations is akin to ignoring most occurrences of these combinations of values for the observed variables. For instance, if survey participants do not answer question 2 when they have provided answer A to question 1, ignoring observations with missing data also means ignoring the responses A to question 1.

It is not possible to determine solely from the available observations in a modelling problem whether data are missing at random or not (MAR or MNAR). To address this question, it is recommended to explore other studies on the same modelling problem characterized by the same missing data, in order to gain insights into the mechanism by which data are missing. Including a maximum number of variables in the data collection that have the potential to explain missing data can be considered to make the MAR case more likely than MNAR case.

When data are missing in a non-random manner (MNAR), by definition, a model for the missing data is unavailable and it might be possible to enhance data collection (considering new variables that explain missing data) to move from the MNAR case to the MAR case.

III) A few data imputation methods

1. Imputation using a single value: default value, mean, median

The simplest method involves using a single value as an estimate for missing values. The value employed can be:

- A fixed value, independent of the data (and thus the same for all variables with missing values), for instance, the value 0. Of course, arbitrary choices should be avoided.
- A value representative of the distribution of the concerned variable, such as the mean or median for a quantitative variable, calculated from the available (non-missing) values.
- For a quantitative variable that can only take a very limited number of different values or for a nominal variable (with categories), it is possible to choose the most frequent value among those present.

It is important to note that using a single value to impute missing values for a variable can significantly reduce the estimated variance for that variable. This should be taken into consideration in interpreting the results in the subsequent modelling process.

2. Imputation from the group center

When natural groups are present in the data, it is possible to use the centers of these groups to impute missing values for certain observations. The procedure is straightforward. An automatic classification algorithm, such as K-means, is applied to complete observations (without missing data). Then, for each observation with missing data, the nearest group center is determined, and each non-informed value is assigned the value of the corresponding variable for the nearest group center.

To learn more about K-means, you can read this document: https://github.com/abenslimaneakawahid/clustering/blob/main/K-means.pdf.

3. Imputation based on the k-nearest neighbors (kNN)

The use of imputation by group centers can yield interesting results but is contingent upon the hypothesis of the presence of groups in the data. For an observation with missing data, considering that the nearest complete observations are more representative than others constitutes a less restrictive hypothesis. Of course, this proximity must be determined based on distance calculations limited to only the complete variables for the observation with missing data.

The procedure then involves, for each observation with missing data, finding its kNN, taking into account, in distance calculations, only the values of the complete variables for that observation. Subsequently, the non-informed value is assigned the mean of the values taken by the same variable for these k neighbours.

4. Imputation based on local regression (LOESS - LOcal regrESSion)

LOESS is a method for imputing missing data. In this approach, a low-degree polynomial is fitted around the missing data using the least squares method applied to complete observations in the vicinity of the incomplete observation.

For example, if only the value y_{ij} is missing for individual i, the kNN of i are selected from the set of complete observations. If these observations are denoted as $(1), \dots, (k)$, the least squares linear regression problem is then estimated as follows:

$$\hat{\beta} = arg \min_{\beta \in R^{p-1}} \sum_{i'=1}^{k} \left(\beta^T y_{(i')}^{-j} - y_{i'j} \right)^2$$

where $y_{(i')}^{-j}$ is the vector of observations for the $p-1$ variables other than Y_j for the individual (i') . The value y_{ij} is then imputed by $\hat{\beta}^T y_i^{-j}$

This approach can be easily generalized in the case where multiple variables need to be imputed for the same observation. Additionally, instead of selecting the k-nearest neighbours, decreasing weights can be assigned to the individuals used to estimate β as they distance themselves from individual i (source: Alyssa Imbert, Nathalie Vialaneix, Journal de la Societe Française de Statistique, 2018, 159 (2), pp.1-55. ffhal-02618033).

5. Imputation through singular value decomposition (SVD)

This method is interesting in cases where singular value decomposition (SVD) with rank reduction provides a good approximation of the (complete) observations. In other words, when observations described by d quantitative variables are close to a linear subspace of dimension k much smaller than d. In this scenario, an estimate for the missing value of a variable value for an observation is obtained at the intersection between the subspace, which is a good approximation of the complete observations, and the subspace obtained by fixing the values of the known variables for the incomplete observation.

The value of the reduced rank k has a significant impact on the estimation result. Specific methods for choosing an optimal k can be applied but will not be explained in this document. However, if you wish to learn more about the SVD method, you can refer to the following document: https://github.com/abenslimaneakawahid/iterative-methods/blob/main/SVD.pdf.

6. Additional methods

There are numerous other methods for imputing missing data, such as the NIPALS algorithm (Nonlinear Iterative Partial Least Squares), Amelia II (a multiple imputation program developed in 2011 by James Honak, Matthew Blackwell, and Gary King), missForest (a completion method based on random forests proposed in 2011 by Daniel J. Stekhoven and Peter Bühlmann), Bayesian inference, and more.