UNIVERSITÀ DI GENOVA
SCUOLA DI SCIENZE SOCIALI
DIPARTIMENTO DI ECONOMIA

Tesi di laurea magistrale in

Statistical Models

Techniques of multiple imputation to deal with missing data in large datasets

Relatore: Fabio Rapallo

Candidato: Gianluca Olcese

Anno accademico 2022-2023

Index

Abstract

In recent decades' literature, missing data occurrence represented a major issue in various fields of research. Considering that a partial loss of information affecting the dataset under analysis may represent a relevant constraint when conducting an inferential procedure, such topic has been in the recent years – and is still nowadays – a relevant and common topic of discussion in the literature.

The first part of this work has the objective the provide a brief overview of the main types of missing data and of the mechanisms which may be responsible of such missing values. Given that the aim of an inference methodology is to obtain unbiased results, it is particularly relevant to analyze statistical methods which allow to handle missing data and to obtain such desired unbiased figures. The second section of this paper provides therefore an overview of the main single imputation techniques, which are able to fill in the missing values by the use of a number of different statistical techniques. A more complex and accurate imputation methodology is then represented by the multiple imputation procedure, presented in chapter III, which may also be applied to different types of data as explained in the last section of this paper.

Chapter I

A general overview about the issue of missing data

In various types of statistical and inferential analysis, a phenomenon that can frequently occur is the issue of missing data. It can appear in various forms and with different characteristics, based on which it is necessary to equip oneself in the best way to obtain unbiased results.

1.1 Missing data: reasons and definitions

In recent decades, the issue of missing data has acquired an increasingly important role within the analyzes conducted by researchers in various fields. The lack of some relevant information within a dataset can in fact easily lead, if not managed in the best way, to obtaining biased results.

There are multiple possible reasons that can lead to missing data. In this framework, the use of the term "possible" is not casual but takes on a precise meaning: when the data is only partially observed, it is very difficult to have precise knowledge of the mechanisms that can lead to missing data. It is therefore possible to formulate a series of hypotheses regarding the mechanism of causality underlying the phenomenon, provided they are consistent with the data analysed.

Moreover, in each dataset there are, as respondents, some *units* which are asked to provide information about a series of *items*. In a classic individual questionnaire, the unit consists of the individual while the item is represented by the answers to the questions asked.

In this sense, it is important to distinguish between two types of missing data. The first one, called *unit nonresponse*, consists of a situation in which all the responses of a specific unit are missing: substantially, in the previous example, there will be no information available regarding a specific individual.

The second typology is represented by the *item nonresponse*, in which one or more units provide only a part of the requested information (e.g., an individual who does not respond to one or more questions in a questionnaire).

In addition to correctly managing the presence of missing data within a dataset, it is particularly important to understand the origin of the phenomenon in question.

Missing data can then be classified according to the assumptions underlying the “missingness” mechanism, that is, the assumed mechanisms that are believed as causing the data to be missing (Pampaka et al., 2016).

Missing data mechanisms are described as falling into one of the three categories briefly described below (Allison, 2000), which sometimes are called “distribution of missingness” (Schafer and Graham, 2002):

- Missing Completely At Random (MCAR): independence is assumed between the missingness and observed and missing responses. That is, every case is characterized by the same missingness probability.
- Missing At Random (MAR): the missingness is assumed to be conditional independent of the missing responses, given the observed responses. Therefore, the probability of observing missing data regarding a particular variable of interest may depend on other observed variables, but not on such variable itself.
- Missing Not At Random (MNAR): missingness depends on both observed and unobserved (missing) data.

Missing data are called *Missing Completely At Random* (MCAR) if, given a certain value, the probability of it being missing is assumed to be unrelated to the observed and unobserved data on that unit (Carpenter and Kenward, 2013). When MCAR data occur, there is no relation between the chance of the data being missing and the values: the observed data are therefore representative of the population of interest but, of course, the fact that some information has been lost has to be taken into consideration.

MCAR data, in real world experiments, may arise in a lot of potential situations. In medical research it may consist in a tube containing a blood sample of a study subject broken by accident (such that it is not possible to measure the blood parameters of interest) or in an accidental loss of a questionnaire of a study subject (Donders et al., 2006), while in educational research we may observe a situation in which – conducting some sort of school study – some pupils might be missing from a sample because they might have been away for school for random and unpredictable reasons (Pampaka et al., 2016).

Data are said to be *Missing At Random* (MAR) if *given, or conditional on, the observed data* the probability distribution of the missing data is independent on the unobserved data (Carpenter and Kenward, 2013).

In MAR data framework, missing data depend on known values and they are, consequently, fully described by the variables actually observed in the dataset. Missingness does not depend on the variable of interest: it could instead depend on the other variables which are observed. Therefore, accounting for values "causing" MAR data will result in obtaining unbiased results (Wayman, 2003).

In real world studies, Missing At Random data are likely to occur in a variety of different fields. In medical research, MAR data could take the form of older patients who might be more likely to miss "insurance" than younger ones: in this framework, "insurance" will be MAR if the study of interest has collected the age for all the subjects of the study (He, 2010). In educational research we may observe MAR data in a situation in which, in a school survey, a part of pupils may be missing because they are representing their school in some sort of competition.

If in a dataset we observe missing data which cannot be classified as MCAR nor MAR, then these data are called *Missing Not At Random* (MNAR). In this framework, the probability of an observation being missing depends on the underlying value, and this dependence remains even given the observed data (Carpenter and Kenward, 2013).

In this situation, the reason leading to missing data is not completely at random and is related to patient characteristics which are not observed.

When a MNAR data mechanism occurs, it leads to a relevant loss of valuable information and there is no universal method of handling in a proper way the missing data (Donders et al., 2006).

This typology of missing data, in real world studies, reveals to be observable in many different fields. In the framework of a socio-economic study, MNAR data may occur in a situation in which asking for a subject for his or her income level missing data may be more likely to occur when the income is relatively high (Donders et al., 2006). In such case, the probability of nonresponse characterizing the income variable depends on values which might be missing.

In educational research, we may observe MNAR data in a school study when pupils might not respond to sensitive questions about their special educational needs, supposed to be related to the outcome variable of interest (Pampaka et al., 2016).

To conduct correct statistical inferences leading to unbiased results, it is crucial to understand of which class the missing data mechanism falls into.

Under MNAR data mechanism, from the observed data something is not available to the researchers conducting the study of interest. Therefore, it is possible to state that MAR data can never be proved or falsified using data alone (He, 2010).

In many situations, however, it is actually possible to test if missing data belong to MCAR class. If for some variable there exist meaningful differences between the subjects with and without missing data, it is possible to state that the missing data of interest are not driven by a MCAR mechanism.

Under MAR assumptions (including MCAR as special case) it is possible to ignore missingness models and focus on the missing-data models, while in MNAR framework generally it is necessary to specify missingness models to obtain the correct inferences.

When dealing with missing data, it is necessary to adopt adequate techniques in order to handle in a proper way the data issue. Moreover, in last decades literature have been provided some basic recommendations as to what should be done about missing data (Pampaka et al., 2016):

- Always report details of missing data.
- Adjust results for what is known about the missing data, if possible.
- Report the sensitivity of the reported results to the distribution of missing observations.

1.2 Some examples and main implications for inference

In real-world analyses, it may be necessary to deal with the issue of missing data in a plurality of situations belonging to various fields of research. In such contexts, it is therefore necessary to understand the possible mechanism driving to missing data and, consequently, to handle their presence in a proper way to obtain correct and unbiased results.

In this perspective, in the last decades the literature has provided examples of real-world studies in which such issues are – or may be – present in different forms and with different classifications (MCAR, MAR, MNAR[1]).

The aim of this section is to present a brief overview on some relevant missing data real-world reported – or potential – situations that the literature has provided in the last twenty years, for each of the main three categories of missing data cited above[2].

One of the main fields in which missing data reveals to be a common issue consists of the medical research. In this perspective, a first relevant work is the one by Donders et al. (2006, "Review: A gentle introduction to imputation of missing values). In the first section of the paper the authors explain the split of missing data in the three main categories, for which one of them they provide examples of real-world medical research situations in which data may miss due to specific reasons.

According to the authors, typical examples in which MCAR missing data – characterized by the fact that subjects who have missing data are a random subset of the complete sample of subjects – are the accidental breaking of a tube containing a blood sample of a study subject (making it impossible to measure the blood parameters of interest) or an accidental loss of a questionnaire of a study subject. In these two situations it is possible to state that the reason for missingness is completely random and that the probability that an observation is missing is not related to any other patient characteristics. Therefore, the set of subjects with no missing data will undoubtedly be a random sample from the

[1] Missing Completely At Random, Missing At Random, Missing Not At Random

[2] MCAR, MAR, MNAR

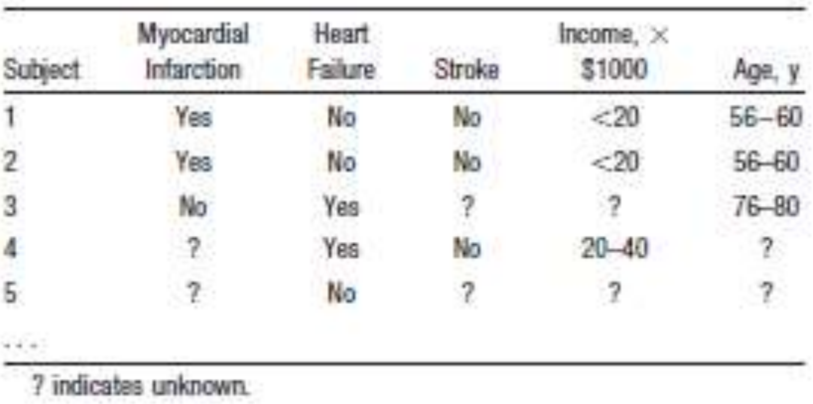
source population. We may instead observe the occurrence of MNAR missing data – in the framework of which the probability that an observation is missing depends on information that is not observed – when, asking a subject about her or his income level, missing data may be more likely to occur when the underlying income level is high. In such framework the reason for missingness is therefore related to patient characteristics that cannot be observed. MAR missing data, which reason for missingness is based on other observed characteristics, need to be handle with care: missing data can indeed be considered random conditional on these other patient characteristics that determined their missingness and that are available at the time of analysis (Rubin, 1976). In this sense, Donders et al. provide a practical example: if, in a medical research framework, the aim is to evaluate the predictive value of a diagnostic test of interest and the results of the tests are known for all the diseased subjects but unknown for a random sample of non-diseased subjects, then such missing data fall into the classification of MAR because, conditional on observed patient characteristics, missing data are random (provided that missingness does not only depend on the outcome variable).

He (2010, "Missing Data Analysis Using Multiple Imputation – Getting to the Heart of the Matter"), to explain the difference between the three main categories of missing data, considers the study of Huskamp et al. (2009), who investigated the patterns of hospice discussion with providers by patients with late-stage cancer. In this study, the authors use data collected from a multisite cohort study of care for patients with lung or colorectal cancer by the Can-CORS Consortium[3]. In such dataset, as typically happens in any large health or social dataset, a substantial amount of missing data may occur, characterized by no systematic pattern. In the example provided by the authors, the fractions of missing observations range from 0.04% to 19.48% for the variables, including both the predictors and the outcome. The relevance of this phenomenon is confirmed by the fact that removing from the dataset the patients with missing data would result in a loss of around 30% of the sample, inevitably leading to a massive issue about the validity of the obtained results.

[3] Cancer Care Outcome Research and Surveillance Consortium

Some lines of such dataset are shown in Table 1.1, in which the missing data are the elements that we do not observe, marked by question marks.

Missing Data Matrix.



Subject	Myocardial Infarction	Heart Failure	Stroke	Income, × $1000	Age, y
1	Yes	No	No	<20	56–60
2	Yes	No	No	<20	56–60
3	No	Yes	?	?	76–80
4	?	Yes	No	20–40	?
5	?	No	?	?	?
...					

? indicates unknown.

Table 1.1. Source: He, 2010, "Missing Data Analysis Using Multiple Imputation – Getting to the Heart of the Matter"

Moreover, the author provides, for each of the main three missing data classifications[4], examples of reasons that can lead to missingness in the dataset of analysis. MCAR missing data, in the context of analysis, may be difficult to observe because most missingness is not completely random: older patients, for example, are more likely than younger ones to have nonresponse on either income or insurance questions. We may instead observe MAR missing data, which rely on the more general assumption that the probability a variable is missing depends only on the observed characteristics, in the case in which older patients might be more likely to miss "insurance" with respect to younger patients. In this framework, the variable "insurance" is said to be MAR if the study has collected information on age for all patients in the sample. MNAR missing data may instead arise when people with higher income are less likely to reveal them; therefore, the probability of nonresponse for the variable "income" depends on values that are or can be missing.

[4] Missing Completely At Random, Missing At Random, Missing Not At Random

One of the most important references in terms of missing data consists in the book “Multiple Imputation and its Applications”, published by James R. Carpenter and Michael G. Kenward (Department of Medical Statistics – London School of Hygiene and Tropical Medicine, UK) in 2013.

In the first section of the book the authors provide an overview of the differentiation between the different categories of missing data with some examples of application.

The first one consists of the so called “Mandarin tableau”: in Figure 1.1 it is shown part of the frontage of a senior mandarin’s house in the New Territories, Hong Kong.

Detail from a senior mandarin’s house front in New Territories, Hong Kong. Photograph by H. Goldstein.

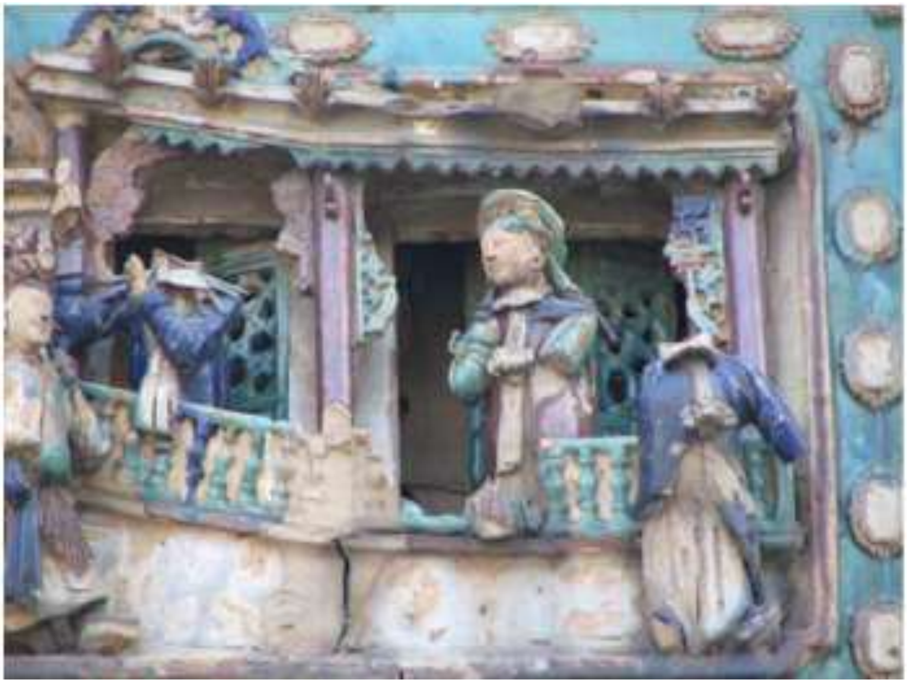


Figure 1.1. Source: Carpenter and Kenward, 2013, “Multiple Imputation and its Application”

Assuming that interest is about the figurines’ characteristics – such as their number, height, facial characteristics, and dress – unit nonresponse will correspond to missing figurines, while item nonresponse will arise in the case of damages figurines. In this example, if the aim is to summarize facial characteristics of the figurines and missing heads are supposed to behave as MCAR missing data, from the observed heads a valid estimate is obtained, even if imprecise compared to an estimate obtained observing all the heads. On the other hand, a MAR classification would imply to assume that the distribution of head characteristics given body characteristics does not depend on whether the head is present. Therefore, under this assumption, it would be possible to estimate the distribution of the characteristics of the figurines with missing heads from the ones with similar body characteristics. Completely different scenario is the one in which we assume to have MNAR missing data: in this case, it would be possible that the figurines with missing heads were wearing some sort of head dress which identified them as a member of some class or group which was the cause for the heads to be smashed. Under this mechanism, it is not possible to state anything about typical characteristics of head dress without making assumptions (which, of course, cannot be verified) about the characteristics of the missing head dresses. Moreover, this type of assumption implies a different distribution of head dress given body dress for the figurines with and without heads.

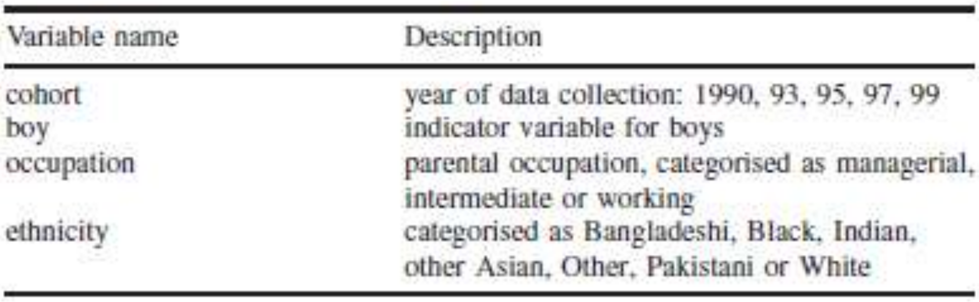
Carpenter and Kenward provide then another fundamental example of real missing data scenario. The framework is the one of YCS[5] of England and Wales, an ongoing UK government funded representative survey of pupils at school-leaving age (School year 11, age 16-17)[6]. The authors consider a harmonized dataset deposited by Croxford et al. (2007) that comprises YCS cohorts from 1984 to 2002 and consider data from pupils attending comprehensive schools from five YCS cohorts and who reached the end of Year 11 in 1990, 1993, 1995, 1997 and 1999.

In Table 1.2 it is possible to observe the covariates from the YCS considered by the authors; the variables “cohort” and “boy” do not present any missing data.

[5] Young Cohort Study

[6] UK Data Archive, 2007

YCS variables for exploring the relationship between Year 11 attainment and social stratification.

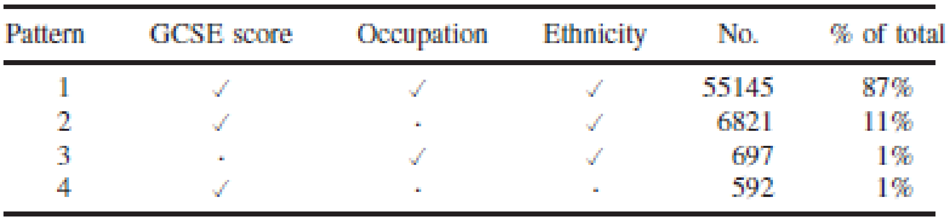


Variable name	Description
cohort	year of data collection: 1990, 93, 95, 97, 99
boy	indicator variable for boys
occupation	parental occupation, categorised as managerial, intermediate or working
ethnicity	categorised as Bangladeshi, Black, Indian, other Asian, Other, Pakistani or White

Table 1.2. Source: Carpenter and Kenward, 2013, "Multiple Imputation and Its Application"

Moreover, the pattern of missingness for GCSE[7] score and the remaining two variables are shown in Table 1.3. It is important to point out that, in this example, it is not possible to re-order the variables to obtain a monotone pattern.

Pattern of missing values in the YCS data.



Pattern	GCSE score	Occupation	Ethnicity	No.	% of total
1	✓	✓	✓	55145	87%
2	✓	·	✓	6821	11%
3	·	✓	✓	697	1%
4	✓	·	·	592	1%

Table 1.3. Source: Carpenter and Kenward, 2013, "Multiple Imputation and Its Application"

In this study, if missing data are assumed to be of MCAR typology, it would be possible to obtain valid inference results from the 55145 complete records (Table 1.3). However, without having the data for the 8110 individuals characterized missing data, the partial loss information would lead to less precise results with respect to the case of no missing data.

Another relevant real case of missing data provided by Carpenter and

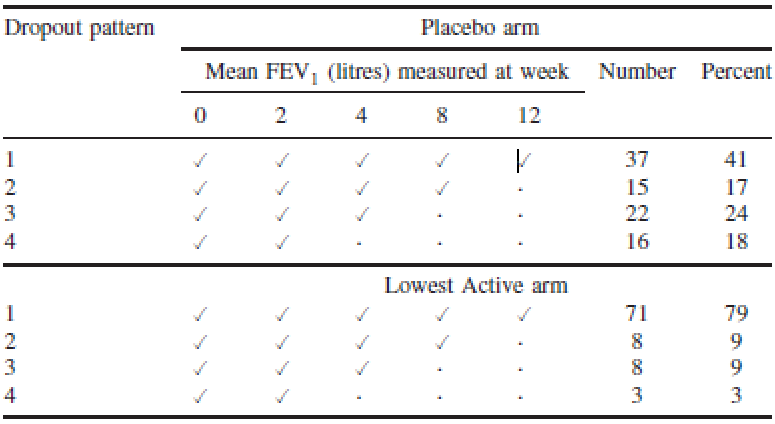
[7] General Certificate of Secondary Education

Kenward is the “Randomized controlled trial of patients with chronic asthma”. In this framework, the authors consider data from a 5-arm asthma clinical trial to assess the safety and efficacy of budesonide, a second-generation glucocorticosteroid, on 473 patients with chronic asthma who were enrolled in the 12-week randomized, double-blind, multi-centre parallel-group trial, which compared the effect of a daily dose of 200, 400, 800 or 1600 mcg of budesonide with placebo. The principal outcomes of clinical interest include patients’ peak expiratory flow rate[8] and their FEV$_1$[9]. The trial found a statistically significant at a 95% confidence level dose-response effect for the mean change from baseline over the study for both morning and evening peak expiratory flow and FEV$_{1.}$

The aim of the study was to compare FEV$_1$ across treatment arms at 12 weeks; however, excluding 3 patients with intermittent participation in the study, only 37 out of 90 patients in the placebo arm, and 71 out of 90 patients in the lowest active dose arm, at twelve weeks had remained in the trial.

The withdrawal pattern for the placebo and lowest active dose arms is shown in Table 1.4. It is possible to observe that the missingness pattern is monotone in both treatment arms.

Asthma study: withdrawal pattern by treatment arm.



Dropout pattern	Placebo arm						
	Mean FEV$_1$ (litres) measured at week					Number	Percent
	0	2	4	8	12		
1	✓	✓	✓	✓	✓	37	41
2	✓	✓	✓	✓	.	15	17
3	✓	✓	✓	.	.	22	24
4	✓	✓	.	.	.	16	18
	Lowest Active arm						
1	✓	✓	✓	✓	✓	71	79
2	✓	✓	✓	✓	.	8	9
3	✓	✓	✓	.	.	8	9
4	✓	✓	.	.	.	3	3

Table 1.4. Source: Carpenter and Kenward, 2013, “Multiple Imputation and Its Application”

[8] The maximum speed of expiration in litres/minute

[9] “Forced Expiratory Volume”: the volume of air, in litres, the patient with fully inflates lungs can breathe out in one second

In this study, if missing data are assumed to be MCAR, it is possible to get a valid estimate of the overall mean in each group at 12 weeks by averaging the 37 available observations in the placebo group and the 71 in the active group, obtaining respectively 2.05 litres (s.e.[10] 0.09) and 2.23 litres (s.e. 0.10) leading to a treatment effect of 2.23 – 2.05 = 0.18 litres.

However, is a MNAR mechanism is assumed to drive the missing data, it is possible to assume a pattern mixture model and the treatment effect varies as we move away from the MAR mechanism assumption (Figure 1.2). Moreover, since the placebo group is characterized by many more missing patients, the treatment effect estimate reveals to be much more sensitive to departures for MAR in such group.

Contour plot of the difference in average FEV$_1$ (litres) between active and placebo groups, as we move away from MAR. Under MAR, the difference is 0.18 litres.

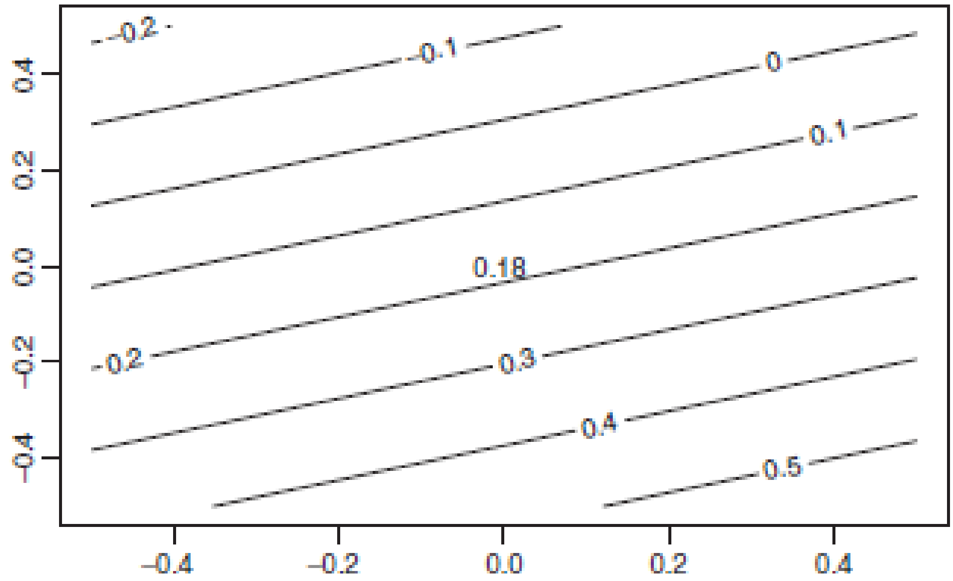


Figure 1.2. Source: Carpenter and Kenward, 2013, "Multiple Imputation and its Application"

Regarding the same topic, another relevant publication consists in the paper by Katherine J. Lee and Julie A. Simpson (2014, "Introduction to multiple

[10] Standard error

imputation for dealing with missing data"). Indeed, the aim of such work is to estimate whether current asthma status is associated with FEV_1, after adjusting for some covariates as age, gender, socio-economic status, smoking status, height, and waist circumference by the means of a multivariable linear regression (Kasza and Wolfe, 2014).

Lee and Simpson use a random data sample from the fifth decade of follow-up from the TAHS[11], a population-based longitudinal cohort study of 8683 children born in 1961 and attending school in Tasmania in 1968.

Considering such dataset, waist circumference data were not available for approximately one quarter of the subjects, leading to a material loss of relevant information. Moreover, the analysis is restricted to 316 TAHS participants with complete data on all the covariates except from waist circumference. A brief overview of the results obtained by Lee and Simpson will be provided later in this work, when various techniques to deal with missing data in large dataset will be analyzed.

Maria Pampaka, Graeme Hutcheson and Julian Williams (2016, "Handling missing data: analysis of a challenging data set using multiple imputation) provide some real-world scenarios in which different forms of missing data may arise in educational research.

The authors explain how, in the context of a school survey, different missingness mechanisms can lead to different missing data categories: if, for example, a researcher gets permission to administer a questionnaire about bullying to the students during class time, on the administration day there are various scenarios which could verify:

- some students may not have been present at random without any predictable reason;
- some pupils may have been absent because that day they might be representing their school in some sort of competition, being them the most engaged and keenest;
- some students may choose not to respond to some particular questions, maybe because they are the ones being bullied or because

[11] Tasmanian Longitudinal Health Study

they may have special needs.

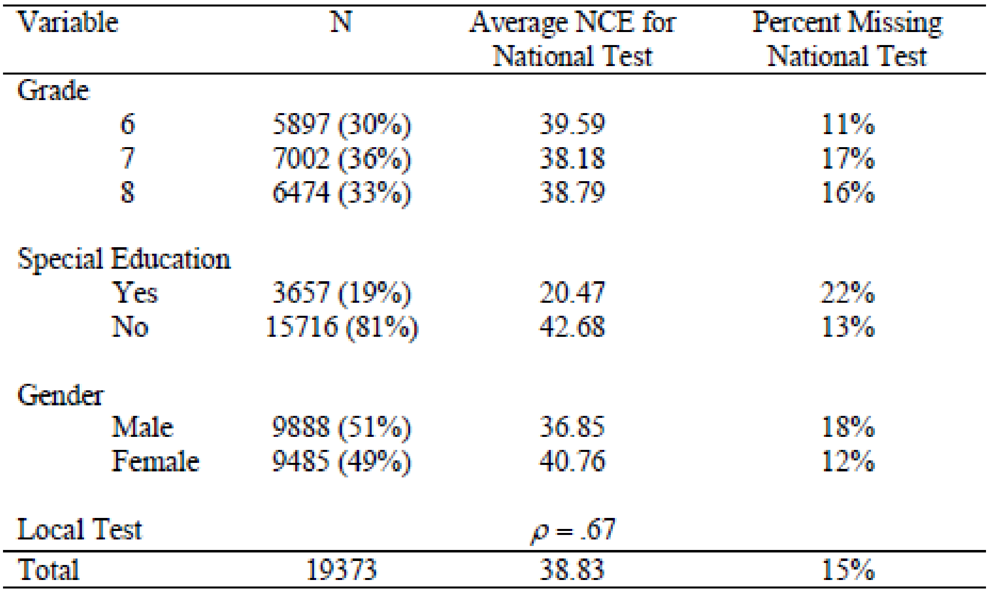
Each one of the three situations above would therefore lead to a different category of missing data. MCAR missing data, for which all the cases are characterized by the same probability of being missing, may arise in the situation in which the students are missing from school for random and unpredictable reasons: the missingness is said to be independent of the observed and missing responses. In the second framework, in which some pupils may be absent because of representing their school in some sort of competition, missing data would be classified as MAR: missingness does not depend on the variable of interest but it could actually depend on other observed variables. In the last situation, in which a number of students choose not to respond to specific sensitive questions about their special educational needs (which are assumed to be also related to the outcome variable of interest), MNAR missing data would instead be observed because missingness would depend on both observed and unobserved information.

Jeffrey C. Wayman, in the paper "Multiple Imputation For Missing Data: What Is It And How Can I Use It'", presented at the 2003 Annual Meeting of the American Educational Research Association (Chicago, Illinois), provides another relevant example of missing data issue in the framework of educational research.

Wayman uses a dataset coming from a large United States school district; variable of interest are: a participant's grade, gender, participation in special education, NCE[12] on a nationally-administered reading test, and row score on a locally-administered reading test. The author explains how local test scores ranged between 232 and 430, however approximately 95% of the data points fell between 303 and 383 and therefore the sample was restricted to such observations in order to provide the clearest possible explanation. Moreover, it was decided to include participants with grades 6, 7, or 8, and with no missing responses for gender, special education status, and local reading test score. With these adjustments, the sample resulted in 19373 subjects, of which 2896 (15%) had missing information about the national test score (Table 1.5).

[12] Norman curve equivalent

Description of the sample.

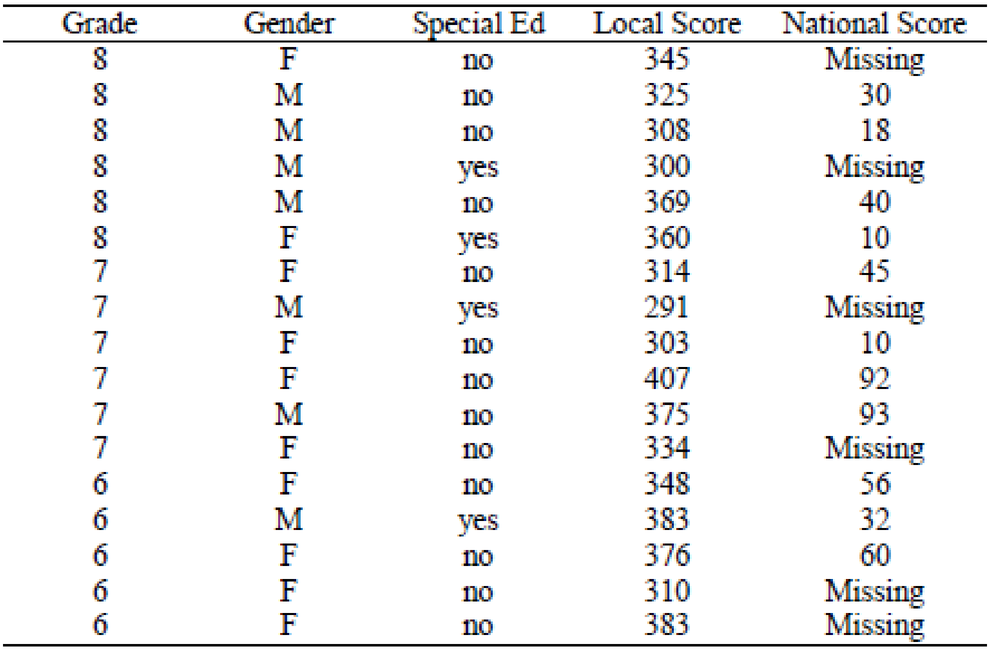


Variable	N	Average NCE for National Test	Percent Missing National Test
Grade			
6	5897 (30%)	39.59	11%
7	7002 (36%)	38.18	17%
8	6474 (33%)	38.79	16%
Special Education			
Yes	3657 (19%)	20.47	22%
No	15716 (81%)	42.68	13%
Gender			
Male	9888 (51%)	36.85	18%
Female	9485 (49%)	40.76	12%
Local Test		$\rho = .67$	
Total	19373	38.83	15%

Table 1.5. Source: Wayman, 2003, "Multiple Imputation For Missing Data: What Is It And How Can I Use It?"

Missing data bias reveals to be evident because special education students, males and pupils who had bad results on the local test typically reveal to do worse on the national test. Moreover, these groups of subjects are the ones who are more likely to present missing data. With this purpose, Table 1.6 reports a subset of participants from the dataset under analysis.

Selected data from the full dataset.



Grade	Gender	Special Ed	Local Score	National Score
8	F	no	345	Missing
8	M	no	325	30
8	M	no	308	18
8	M	yes	300	Missing
8	M	no	369	40
8	F	yes	360	10
7	F	no	314	45
7	M	yes	291	Missing
7	F	no	303	10
7	F	no	407	92
7	M	no	375	93
7	F	no	334	Missing
6	F	no	348	56
6	M	yes	383	32
6	F	no	376	60
6	F	no	310	Missing
6	F	no	383	Missing

Table 1.6. Source: Wayman, 2003, “Multiple Imputation For Missing Data: What Is It And How Can I Use It?”

Having observed, through a general overview of recent relevant literature, how frequently missing data can appear in various types of analyzes belonging to a variety of research fields, it is also important to understand the impacts of the lack of information on the main statistical inference techniques.

In a framework in which some sort of missing data arises, it is necessary to have some specific assumptions under which computational methods lead to valid inference. Therefore, in this context, it is easy to observe misleading inference processes.

The missingness mechanism leading to missing data issue is usually unlikely to be definitively identified from the observed data, even if the latter may indicate possible plausible missing data mechanisms. Therefore, it is needed to take into account some sort of assumption about the missingness mechanism underlying the data in order to be able to draw statistical inference.

Even if some assumptions can be made about the reason underlying missing information in the dataset, it is important to point out that the precise mechanism causing the missing data can rarely be definitively established: with the aim to verify the robustness of an inference to a range of different possible missingness mechanisms, it can be useful to implement a process dubbed as *sensitivity analysis* (Carpenter and Kenward, 2013).

An inference methodology in which there is some kind of missing data issue can be affected mainly by two problems: loss of efficiency and bias. The former is an inevitable missing data consequence and the extent of information loss is not directly linked to the proportion of incomplete records but instead it is said to be intrinsically linked to the question of interest. When dealing with a dataset characterized by missing data, most statistical software in an automatic way restricts the analysis to complete records; however, this loss of information leads to consequences which are not always easy to predict in advance. Moreover, many times the information deriving from partial complete records is fundamental for the study itself: it is therefore necessary to implement some techniques to handle in a proper way missing data and minimize the loss of information.

The subset made of complete records may also not be representative of the whole population under analysis. In this case, a restriction of the sample to complete records may lead to a biased inferential procedure, where the extent of such bias depends on the statistical behavior of the missing data affecting the dataset under study.

If a specific assumption about the reason leading to missing data is made, it is possible to implement a valid analysis that does not require to include the model for the missingness mechanism underlying. In this specific situation, such mechanism is dubbed as *ignorable*.

In this context, it is relevant to explore the implications of missing data, in terms of loss of information and bias, in the response and/or in the covariates under different mechanisms driving missing data.

The first case is the one in which we observe a partially observed response. In this sense, Carpenter and Kenward (2013) take into consideration the following regression model:

$$Y_i \;=\; \beta_0 + \;\beta_1 x_i + \;e_i, \quad e_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

In such model, Y (the outcome variable of interest) is only partially observed, and R_i indicates whether the outcome is observed or not for an individual, while the x_i are supposed to be known without any error. Then, the contribution to the likelihood for $\beta = (\beta_0, \beta_1)$ from unit i, conditional on x_i, can be defined as the following:

$$L_i = \Pr(R_i, Y_i|x_i) = \Pr(R_i|Y_i, x_i) \,|\, \Pr\,(Y_i|x_i)$$

Assuming that the parameters of $\Pr(Y_i|x_i)$, β, are different from the parameters of $\Pr(R_i|Y_i, x_i)$, it is possible to state that the contribution to the above likelihood for an individual characterized by missing response can be obtained by integrating over all possible values of the missing outcome variable Y_i, given x_i:

$$\int \Pr\,(Y_i|x_i)\, dY_i = 1$$

All individuals with missing Y, conditional on x, contribute 1 (the total probability, since it is an integration over all possible values of Y_i given β, x_i) to likelihood for β. Therefore, it is possible to state that there is no effect on the maximum likelihood estimate of β. This is possible because the parameter space of the conditional distribution of Y given X is separate from the one of the marginal distribution of X. Then, as a direct consequence, the conditional distribution of $\Pr(Y|X)$ has not any information on the marginal distribution of X and does not place any restriction on it.

It is then relevant to consider the opposite framework, that is the one in which, considering a regression of *Y* on *X*, the former is fully observed while the latter is only partially observed.

Letting $R_i = 1$ if X_i is observed and $R_i = 0$ otherwise, it is useful to consider the regression of *Y* on *X* estimated using only the complete records (that is, $R_i = 1$):

$$\begin{aligned}\Pr(Y_i|X_i, R_i = 1) = \frac{\Pr(Y_i, X_i, R_i = 1)}{\Pr\,(X_i, R_i = 1)} = \frac{\Pr(R_i = 1|Y_i, X_i)\Pr\,(Y_i, X_i)}{\Pr(R_i = 1|X_i)\Pr\,(X_i)}\\ = \left\{\frac{\Pr\,(R_i = 1|Y_i, X_i}{\Pr\,(R_i = 1|X_i)}\right\}\Pr\,(Y_i|X_i)\end{aligned}$$

When the mechanism underlying the missingness for the covariates involves the response variable *Y*, just restricting the sample to the complete records leads to obtain biased point estimator and, therefore, invalid inference methodology. This situation holds whether the missingness mechanisms only depends on *Y* (MAR, with MCAR as a special case) or when it also includes *X* (MNAR).

In the context of the linear regression, there is one last possible case, that is the one in which missing data issue affects both the response variable and the covariates. Supposing (1) to have the three variables *X, Y, Z* and that (2) *Y* and *X* are MAR given *Z*, in a linear regression of *Y* on *X, Z,* units with *X, Y* missing will contribute to the likelihood of Pr(*Y*| *β*; *X, Z*) in the behavior described by the following integration:

$$\int \Pr(Y|\beta; X.Z)\,dY = 1$$

Therefore, it possible to state the complete records analysis described above in this scenario will be unbiased. Additional variables predictive of Y and/or X, therefore, may be useful to recover more information about the missing values and, consequently, the estimate of β.

Having presented and defined the different classifications of missing data, provided examples of real studies in which this issue has played an important role, and analyzed the consequences of having at disposal incomplete information in the context of statistical inference methodologies, the next chapters of this work aim to present different categories of techniques to handle missing data issue in the best possible way.

Chapter II

Methods of single imputation to handle missing data

In the recent decades' literature, it is possible to find several methodologies to treat missing data and the consequences deriving from such recurrent issue.

Most of the above-mentioned methods have been developed to handle missing data issue in sample surveys; moreover, they have some drawbacks when applied to the Data Mining context.

When dealing with the replacement of such missing data, it reveals to be fundamental to pay particular attention to three key factors (Patel, 2012):

- estimated values should not be affected by bias;
- the relation between attributes should be maintained;
- the overall cost needs to be minimized.

A very important role is inevitably played by the choice of the right technique, which depends on the problem domain, the data's domain and the goal characterizing the data mining process (Somasundaram and Nedunchezhian, 2011).

In the following section of this work will be provided a general overview of some relevant methods to handle missing data issue in a proper way, with the aim to obtain unbiased inference results.

2.1 Ignoring and discarding data

The methodology of ignoring and discarding data is usually implemented when, assuming that the data mining aim is the classification, the class label reveals to be missing or many attributes (not just one) are missing from the dataset row of interest. However, it is important to point out that, if the share of such rows is sufficiently high, the performance obtained by implementing the methodology will be poor.

To discard data with missing values, it is possible to use two main techniques.

The first one, known as *complete case analysis*, available in all statistical programs and used as default methodology in many of them, consists of discarding all instances with missing data. The second method, known as *discarding instances and/or attributes*, consists of determining the extent of missing data for each instance and attribute and, as second stage, deleting the instances and/or attributes characterized by an high level of missing data. However, before deleting any attribute, it is fundamental to control for its relevance to the analysis of interest: a relevant attribute should be kept even if characterized by a high share of missing values.

A fundamental characteristic of the above-mentioned methods lies in the fact that they should both be applied only in the case in which missing data derive from a missingness mechanism allowing them to be classified as MCAR: missing data belonging to the other two main categories (MAR, MNAR) are characterized by non-random elements which may lead to have some bias in the obtained results.

Somasundaram and Nedunchezhian (2011, "Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values") provide a brief example of discarding data application.

The authors suppose to consider a database of students enrollment data (age, SAT score, state of residence, etc.) and a variable which classifies their success in college between "Low", "Medium" and "High".

If the aim of the work is to build a statistical model able to predict the students' success in college, data rows containing missing data for the outcome variable of interest (the success in college) are not useful to predict the success in college and, therefore, these rows could be ignored and removed from the dataset before starting the analysis (under the assumption that the underlying missingness mechanism leads to MCAR missing data).

2.2 Mean and median substitution

Mean substitution is a single imputation technique which consists of replacing missing values on a covariate with the mean value of the observed data points. Therefore, this method involves the implementation of the replacement of the missing data for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute.

In this framework, the imputed missing values are said to be contingent upon one and only one variable – the between subjects mean for that variable based on the available data (Malarvizhi et al., 2012).

Mean substitution may be implemented using two slightly different methodologies: *question mean* and *individual mean*.

The question mean imputation method consists of imputing the overall mean of the specific question from the entire cohort (Shrive et al., 2006): if a subject is characterized by a missing value for question 17, the imputed value is the mean value computed from the completed question 17 for the entire cohort analyzed.

Individual mean, instead, is a methodology which may also be used as a simple form of imputation in such scenarios: the imputed value is obtained by computing the mean of a given participant's complete responses to other questions. Therefore, if a subject is characterized by two missing responses, the missing values will be filled using the calculated average of the remaining (completed) 18 questions.

Regarding the framework of mean substitution, a further slight distinctions reveals to be relevant. Indeed, it is possible to use the overall attribute mean or the attribute mean for all samples belonging to the same class.

Computing the imputed value based on the overall attribute mean, missing values for such attribute are replaced with its mean value of the whole database (Somasundaram and Nedunchezhian, 2011): considering for example a database of United States family incomes and assuming the average income of a US family to be X, it is possible to use X as value to replace the missing occurrences of the income.

Using the attribute mean for all samples belonging to the same class, the mean value to impute in order to fill the missing data is computed not with the mean of a certain attribute for all the rows of the given dataset. The calculation may indeed be limited to the relevant class to make the value more relevant for the data row of interest. In this framework, Somasundaram and Nedunchezhian (2011) provide the example of cars pricing database: among other things, it classified cars to “Luxury” and “Low budget” and missing values is dealt in the cost field. In such a situation, replacing the missing data about the cost of a luxury car with the average cost of all luxury cars would probably be way more accurate with respect to the value computed by the factor in the low budget cars.

Median substitution of covariates and outcome variables, just like mean substitution, is still frequently used when missing data issue occurs in statistical inference procedures. It is possible to state that the median substitution imputation methodology reveals to be slightly improved: this improvement is obtained by first stratifying the data into subgroups and then using the average of the subgroup of interest. As a direct consequence, median imputation results in the median of the entire dataset being the same as it would be with case deletion, but the variability between subjects’ responses reveals to be decreased, causing a bias in the variances and covariances toward zero (Malarvizhi and Thanamani, 2012, “K-Nearest Neighbor in Missing Data Imputation”).

Even if the above-mentioned imputation methods are still very used, there are some issues that may arise when they are implemented in the framework of a statistical analysis.

While mean substitution results in overall means equal to complete case values, the variances of these same covariates reveal indeed to be underestimated (Little, 1992) and such underestimation derives from two sources.

First, by filling the missing data points with the same mean value one does not account for the variation that would likely arise in the case in which the variables of interest would instead be observed. Indeed, if the true values would be observed they would probably vary from the imputed mean.

The other driver of the variances underestimation is the fact that the smaller standard errors obtained due to the increased sample size do not reflect adequately the uncertainty characterizing the dataset under analysis in the research.

Bias in the estimation of standard errors and variances are compounded when estimating multivariate parameters such as regression coefficients. Therefore, there is not any circumstance in which a mean substitution imputation may lead to obtain unbiased results (Pigott, 2001).

2.3 Regression

Regression imputation method is based on the assumption of linear relationship between the different variables. That is, it is assumed that the value of one variable changes in a linear way with the other ones. In the framework of this technique, the missing values are replaced using a linear regression function instead of imputing all missing data with some statistics of particular interest as the mean or the median.

When implementing a regression-based imputation, the regression of each variable *j* is used to fill missing values.

In detail, for each variable *j* present the dataset under analysis, regression imputation technique involves the following steps:

(a) Remove the records characterized by missing values for the variable *j*.

(b) Fit the regression of the reduced *j* (without missing values) on other variables.

(c) Use the regression coefficients to fill the missing values in variable *j*.

There are two main different regression techniques: the *predictive regression* and the *random regression*.

In the predictive regression (deterministic regression or conditional mean), the linear regression is used for numeric variables while, when dealing with categorical missing data, a logistic regression is implemented.

The linear regression, by its nature, is characterized by a linear function based on probability; the logistic regression, instead, works on logistic function based on probability but is characterized by only two possibilities for probability.

Moreover, the predictive regression may be characterized by the presence of an auxiliary variable to find the missing values which relates missing values Y_i to such auxiliary variable X_i and the predicted values used for the missing data in Y.

Therefore, the aim of these methods is to create a predictive model able to estimate imputed values which will substitute the missing ones. The attribute affected by missing data, in this methodology, is used as class-attribute, while the remaining ones are used as input in the predictive model.

It is important to point out, in favor of this methodology, that often it may occur that the different attributes in a model reveal to be correlated among themselves. These correlations could therefore be used to implement a resilient predictive model for classification or regression and some of the relationship may be maintained if captured by the constructed regression model.

However, one has also to take into account that the model estimated values usually tend to be more well-behaved with respect to the true values which, unfortunately, are not observed in the reality. Being the missing values predicted from a set of attributes, it is likely to happen that the predicted values are more consistent with the set of attributes used than they would be with the unknown true attributes.

Another relevant drawback regarding this imputation method consists of the requirement for correlation among the attributes of the model: in a situation in which there are no relevant relationships among one or more attributes of the dataset and the attribute affected by missing values, inevitably the regression model implemented will not lead to obtain a precise estimation of the missing values.

The other main regression imputation method is the so-called random regression. This methodology has the aim to find missing values for any variables based on the conditional distribution.

The random regression, therefore, leads to the imputation of the missing value of interest based on the conditional distribution of *Y* given *X*. In concrete *applications*, this type of approach reveals to be more effective in situations in which numeric data are present (Patel, 2012).

2.4 Hot deck and cold deck

Hot deck and cold deck imputation methods are generally used when the components of the data under analysis are skewed (or twisted), that is they present a long tail of data point (usually on the right, “right-skewed data”).

These methods involve the replacement of missing values of one or more variables for a non-respondent (called the recipient) with observed values taken from a respondent (the donor), who has to be similar to the non-respondent with respect to characteristics observed by both cases.

The term “hot deck” derives from the use of computer punch cards for data storage. It refers to the deck of cards for donors available for a non-respondent. The deck was “hot” since it was currently being processed, opposed to the “cold deck” referring to the use of pre-processed data as the donors.

When the donor is selected following a random procedure from a set of potential donors (the donor pool), the method is called *random hot deck method*.

However, sometimes a single donor is identified and the missing values are imputed from that one case, usually the “nearest neighbour” based on some metric. When this happens, the imputation methodology is called *deterministic hot deck method*, since there is no any randomness present in the selection of the donor.

It is important to point out that, in this framework, the term “deterministic” only describes the procedure under which a donor is selected; while in the general imputation framework the same term may be used to describe methods that impute the mean or other non-random relevant values.

Hot deck is implemented, typically, through two stages. In the first stage there is a partition of the data into clusters, while in the second stage each instance characterized by missing data is associated with one cluster. Then, complete cases in a cluster are used to fill in the missing values: this can be implemented simply by calculating the mean or mode of the attribute of interest within a certain cluster.

The hot deck imputation method does not rely on model fitting for the variables characterized by missing values and that therefore needs to be imputed; therefore, this methodology is potentially less sensitive to model misspecification with respect to imputation methods based on parametric models, such as regression imputation (Andridge and Little, 2010).

However, one needs to keep in mind that hot deck imputation method is characterized by implicit assumptions through the choice of metric to match donors to recipients, and the variables included in such metric.

Another relevant feature of this imputation technique consists of the fact that, since values come from responses actually observed in the so-called donor pool, only plausible values may ne imputed.

Moreover, since information in the incomplete cases is being retained, hot deck implementation may represent a gain in efficiency compared to complete-case analysis.

In the end, it is also possible to observe a reduction in non-response bias, to the extent that there is some sort of association between the covariates defining the imputation classes and both the propensity to respond to the questions and the variable which needs to be imputed.

The hot deck imputation procedure is commonly used by United States government statistics agencies and survey organizations with the aim to provide rectangular dataset for users. For example, the NCES[1] uses, even within a survey, different forms of hot deck and alternative imputation methods: out of twenty recent surveys, eleven used a form of adjustment cell hot deck while the remaining nine used a form of cold deck imputation, deterministic imputation, or a Bayesian method for MI[2]. Within the eleven surveys characterized by the hot deck imputation procedure, many of them used both random within sequential imputation and class imputation (NCES, 2002).

Moreover, the hot deck method has been also applied in epidemiologic and

[1] National Center for Education Statistics

[2] Multiple Imputation

medical settings, even if parametric imputation methods still reveal to be more common.

In literature it is possible to find applications of the hot deck and some comparisons with the other imputation methods in the works by Barzi & Woodward (2004) and Perez et al. (2002) regarding cross-sectional studies, while for longitudinal studies it is relevant to cite the papers by Twisk & de Vente (2002) and Tang et al. (2005). However, it is important to point out that the lack of software in commonly used statistical packages may deter applications of the hot deck methodology in these settings.

Cold deck imputation slightly differs from hot deck because it involves imputing missing values of a record using anything other than reported values for the same variable in the current dataset. Therefore, cold deck imputation requires the availability of at least one additional dataset from which the donor will be selected.

An application of cold deck imputation could consist of a framework in which one is using a company’s revenue for March from the previous year’s dataset to fill the missing revenue for March in the current year’s dataset to calculate the turnover of the current year (Figure 2.1).

Multiply imputed datasets (m = 3).

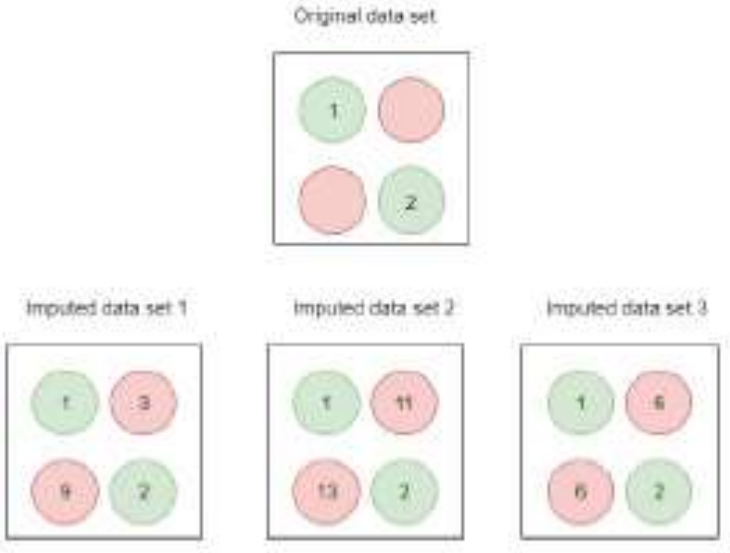


Figure 2.1. Source: Narayan, 2017, "A nearest neighbor based cold-deck imputation for X-ray tube wear estimation"

2.5 K-Nearest Neighbor

A relevant imputation method for missing values in large dataset is the so-called K-Nearest Neighbor Algorithm for Classification.

Malarvizhi and Thanamani (2012, "K-Nearest Neighbor in Missing Data Imputation") provide an exhaustive explanation regarding this imputation technique.

The authors explain that, given a certain dataset, each sample has *n* attributes combined to form an n-dimensional vector such as:

$$x = (x_1, x_2, .. , x_n)$$

Then, these *n* attributes are considered to be the independent variables within the analysis.

Moreover, each sample is also characterized by another attribute, which is denoted by *Y* and is called the dependent variable; with its value depending on the other *n* attributes *x*.

Assuming that *Y* behaves as a categoric variable, it is possible to define a scalar function *f* such as:

$$Y = f(x)$$

This function, therefore, assigns a class to every above-mentioned vector.

Supposing a set of *T* vectors given together with their corresponding classes:

$$x_i, y_i \qquad i = 1, 2, \dots , T$$

Then, the set T is referred to as the so-called *training set*.

The idea underlying the K-Nearest Neighbor methodology is to identify, in such training set, *k* samples whose independent variables *x* are similar to *u*, then to use the *k* samples previously identified to classify such new sample into a certain class, *v*.

Assuming the function *f* to be characterized by a smooth behavior, it is reasonable to look for samples in the training set which are, in terms of the independent variables *x,* near it. Then, it is sufficient to compute *v*, for the samples of interest, from the values of *Y*.

The so-called distance or dissimilarity measure can then be computed, between two different samples, by measuring distance using the Euclidean distance between points:

$$J = \sum_{j-1}^{k} \sum_{i-1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

In the simplest case, that is the one in which $k = 1$ and the sample in the training set is the closest (the so-called *nearest neighbor*) to u, v is equal to Y which is the class of the nearest neighboring sample (Malarvizhi and Thanamani, 2012).

When higher values of k occur, the instance is characterized by a major smoothing which reduces the over-fitting risk due to noise in the training data. However, in real applications usually it happens that k reveals to be in units or tens rather than in the order of hundreds or thousands.

In practice, for handling missing values, the K-Nearest Neighbor methodology outperforms internally two well known Machine Learning Algorithms: C4.5 and CN2. These two algorithms induce propositional concepts: decision trees and rules, respectively. C4.5 algorithm seems to have a good internal algorithm for the treatment of missing data, while CN2 seems to use a simpler method to deal with missing values.

C4.5 algorithm uses a probabilistic approach to handle missing data: missing values can arise in any attribute in training and test files, except the class attribute.

CN2 algorithm, instead, uses a rather simple imputation method to treat missing values: each one of these if filled in with its attribute's most common known value.

Just like any other imputation technique, K-Nearest Neighbor method is characterized by both benefits and drawbacks.

A first advantage of this approach relies in the fact that it is able to predict both the discrete attributes (which reveal to be the most frequent value among the K-Nearest Neighbors) and the continues attributes (the mean among the K-Nearest Neighbors).

Moreover, this technique is characterized by the fact that there is not any need to create a predictive model for each attribute of the study affected by missing values. Indeed, the K-Nearest Neighbor methodology does not implement any explicit model (like, for example, a set of rules or a decision tree), since the dataset is used as a "lazy" model (Batista and Monard, 2002).

The K-Nearest Neighbor imputation method may also be adapted easily to work with any attribute as class, simply modifying the attributes considered in the distance metric. It is therefore considered an approach which can easily handle situations in which multiple missing values are present.

Even if the K-Nearest Neighbor approach looks to be characterized by many advantages, it is important to point out also a relevant drawback: whenever the running algorithm looks for the most similar instances, it searches through all the dataset under analysis. This may represent a very relevant issue, since often a researcher is analyzing very large datasets.

In literature, it is possible to find several wors with the aim to deal with and solve this limitation affecting K-NN.

One proposed method consists of the creation of the creation of a reduced training set for the K-Nearest Neighbor composed only by prototypical examples (Wilson and Martinez, 2000).

Batista and Monard (2002) use, instead, an access method called M-trees, which can organize and search datasets based on a generic metric space. The authors explain how the M-trees methodology is able to reduce drastically the number of distance computations in similarity queries.

2.6 Non-negative matrix factorization

Non-negative matrix factorization approach consists of a matrix decomposition applied to decompose a non-negative matrix into two low-rank non-negative matrices (Li and Ngom, 2013), which has been successfully applied in the mining of biological data.

The *standard-NMF*[3]decomposes a non-negative matrix *X* into two non-negative factors *A* and *Y*, that is:

[3] Non-negative matrix factorization

$$X_+ = A_+ Y_+ + E$$

Referring to the above equation, the term E represents the error (or residual), while M_+ indicates that the matrix M is, as assumed above, non-negative. Then, the optimization for such matrix in the Euclidean space is formulated as follows:

$$\min_{A,Y} \frac{1}{2} ||X - AY||_F^2 \, , subject\ to\ A, Y \ \geq 0$$

From a statistical point of view, the above formulation is obtained, under the relevant assumption of a Gaussian error, from the so-called log-likelihood function.

Moreover, assuming that the multivariate data points are arranged in the columns of *X*, then *A* becomes the so-called *basis matrix* while *Y* is called the *coefficient matrix*; therefore, each column of *A* is a *basis vector*. In this context, each data point is therefore a non-negative linear combination of the basis vectors.

Since the above-mentioned optimization problem is a non-convex optimization problem, the main prescribed optimization technique to solve it is the so-called *block-coordinate descent* algorithm (Li and Ngom, 2013), which, even if relatively easy to implement, is not guaranteed to converge to a stationary point.

A drawback of the standard-NMF method is that it only works for non-negative data, obviously leading to limits in its applications.

To this purpose, Ding et al. (2010) extended the framework to the so-called *semi-NMF*, which peculiarity is to remove the non-negativity constraint on the data *X* and the basis matrix *A*. Semi-NMF can therefore be expressed as follows:

$$\min_{A,Y} \frac{1}{2} \|X - AY\|_F^2 , subject\ to\ Y\ \geq 0$$

Since semi-NMF can be applied to matrix of mixed signs, it allows to extend the framework of NMF to various fields.

Brunet et al. (2004) and Kim & Park (2007) implemented a NMF methodology as a clustering method in order to discover the metagenes[4] and interesting molecular patterns.

Carmona-Saez et al. (2006) proposed an implementation of *non-smooth NMF* (NS-NMF) to study the biclustering of gene expression data; while Wang et al. (2006) provided a *least-squares NMF* (LS-NMF) to take into account the uncertainty of information characterizing the gene expression data.

In the end, Li and Ngom (2012) proposed *kernel-NMF* for reducing dimensions of gene expression data.

However, most authors provide their own NMF implementation with their publications so that the scientific community may use such implementations to re-perform such data mining tasks. However, it is important to point out some issues which prevent researchers and practitioners in the fields of data mining, biological, health medical and bioinformatics areas from using such implementations in a complete way.

The first relevant issue consists of the fact that the above-mentioned NMF techniques usually reveal to be implemented in different programming languages, such as R, MATLAB, C++ and Java, with only one optimization algorithm usually provided in the publication. Therefore, for a researcher who wants to choose an appropriate and suitable mining method for the data under analysis it is difficult to choose the appropriate control parameters and criteria.

Another relevant drawback derives from the way in which the authors usually provide their implementations: in most of the cases, scientific papers only provide NMF optimization algorithms at a basic level and not data mining

[4] E.g., groups of similarly behaving genes

implementation at higher level, making it harder for a researcher to properly investigate and understand the data under analysis.

In the end, the currently existing in literature NMF implementations are application-specific: it does not exist a systematic NMF implementation or package with the aim to perform recurring data mining tasks on biological data.

Even if there exists some NMF toolboxes, as of now there is not any of them which is able to solve the above-mentioned three issues altogether.

Chapter III

Techniques of multiple imputation

Multiple imputation consists of more sophisticate techniques to handle the presence of missing data in large datasets in the framework of statistical inferential analyses. In the next section, an overview of such statistical methodologies will be provided together with some real examples of application.

3.1 The multiple imputation procedure

Multiple imputation methodology consists, substantially, of a two-stage process.

In the first stage, the missing values that the researcher is dealing with are imputed by sampling from an imputation model.

Such model should, therefore, include all variables characterizing the analysis model (outcome, exposure, confounders), as well as additional – at least partially – observed covariates which are not originally included in the model under analysis but that are assumed to be associated with the variables affected by missing data. These additional covariates are called, in this framework, *auxiliary variables*.

Then, the above-mentioned imputation process is repeated multiple times: in this way, a number of completed datasets are implemented (Figure 3.1) with the aim to capture the uncertainty characterizing the missing values.

Illustration of the method of multiple imputation.

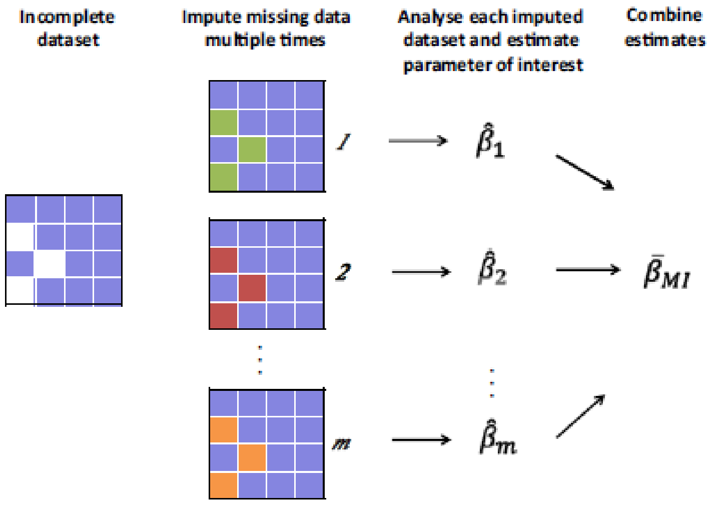


Figure 3.1. Source: Lee and Simpson, 2014, "Introduction to multiple imputation for dealing with missing data"

In such illustration, each box represents a given data point, characterized by columns representing variables and rows representing individuals. In the end, intuitively, blank spaces represent missing values.

β_i represents the estimate of interest characterizing a given completed dataset *I*, while β_{MI} is the estimate obtained from multiple imputation.

The second stage is characterized by the fact that the analysis of interest is performed on each one of the completed datasets, constituted by the observed and the imputed values.

Then, the final MI[1] estimate is defined as the average of the estimates derived from each completed dataset (Kasza and Wolfe, 2014).

Moreover, it is possible to state that the standard error characterizing the multiple imputation estimates incorporates both the uncertainty in the estimate within the completed datasets and the uncertainty across the completed datasets due to the missingness (Lee and Simpson, 2014).

When applying a multiple imputation method, it is necessary to consider that the results will be affected by two typologies of uncertainty:

[1] Multiple Imputation

- Uncertainty *within* imputation, represented by the confidence intervals;
- Uncertainty *between* imputation, consisting of a horizontal shift between results.

A standard procedure may consist of simply considering the mean of the results deriving from separate analyses: such method is called *pooled point estimate*. However, applying such technique, it is likely that the averages confidence intervals underestimate the total variation (made up by uncertainty within imputation and uncertainty between imputation).

In such framework, Rubin (1987) proposed the so-called *Rubin's rules* to pool results from analyses of multiply imputed data. In order to present such rules, it is necessary to define the following quantities:

- *m* represents the number of imputed datasets;
- Q_l is the quantity of interest from *l*-th imputation;
- U_l represents the variance of Q_l.

In such framework, the *pooled parameter estimate* is calculated as follows:

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l$$

Then, the variance characterizing the above pooled parameter estimate is obtained, from the within and between imputation variance, according to the following equation:

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})^T\,(\hat{Q}_l - \bar{Q})$$

The total variance is, indeed, calculated as follows:

$$T = \bar{U} + B + B/m$$

As direct consequence of the presence of missing data, the variance will show an increase represented by the following parameter r_m:

$$r_m = \frac{(B + B/m)}{U}$$

Moreover, it is possible to obtain confidence intervals characterizing pooled estimates by using the *pooled standard error* $\sqrt{T}$ and a *reference t distribution* with the following degrees of freedom:

$$v = (m - 1)(1 + r_m^{-1})^2$$

Then, the *100% confidence interval* is defined by the following expression:

$$\bar{Q} \pm t_v\,(\propto/2)\sqrt{T}$$

Where t_v represents the $\propto/2$ quantile of the *t* distribution characterized by *v* degrees of freedom.

In such framework, the associated *p*-value is constituted by the following probability:

$$\Pr\left\{F_{1,v} > (Q_0 - \bar{Q})^2/T\right\}$$

Where $F_{1,v}$ is a random variable characterized by an *F* distribution with 1 and *v* degrees of freedom, while Q_0 represents the null hypothesis value (which is typically zero).

Consequently, the multiple imputation methodology allows to produce a valid 95% confidence interval and *p*-value for the multiple imputation estimate of the regression parameter under analysis.

In a situation – the simplest case – in which missing data affect only a single continuous variable, the imputation model of interest simply consists of a linear regression model for the variable characterized by missing data, which is regressed on the other covariates used for imputation (therefore, the other variables present in the analysis plus the above-mentioned auxiliary variables).

Instead, when missing data do not affect just a single variable, there are two main approaches which can be used to impute the missing values of interest.

The first one consists of imputing the missing values by using a series of conditional regression models. In such framework, it is needed to set up a regression model for each variable affected by missing data, cycling through the regression models sequentially to impute the missing values for each variable, conditional on the imputed values for the other covariates characterized by the presence of missing data (Lee and Simpson, 2014).

The second method consists, instead, of imputing all the covariates affected by missing values simultaneously, by using a joint normal distribution.

It is important to point out that both the above-mentioned multiple imputation approaches are currently available in standard computerized statistical packages, such as Stata and SAS.

Just like any of the above-mentioned single imputation methodologies, the multiple imputation procedure is characterized by both advantages and backwards.

The first relevant benefit that multiple imputation procedure offers is reducing bias (Lee and Simpson, 2014). In some scenarios in which there may be differences among participants with and without missing data: for example, in medical research in the framework of a respiratory study, we may observe a situation in which the patients affected by asthma and/or allergies may be the ones to be more motivated to attend follow-up visits. In this case, conducting a simple *complete case analysis* may lead to obtain biased results, since it would be biased on a sub-sample which is not representative of the whole population of the study.

If the missingness mechanism underlying the occurrence of missing data is known and the missingness depends on the observed data and not on the unobserved ones, that is if MAR (with MCAR as special case) missing data occur, the multiple imputation procedure has the possibility to fill the missing values by using the observed data.

Therefore, filling in the missing data enable to include all subjects in the analysis and allows to correct the bias that characterizes the complete case analysis.

However, it is relevant to point out that such bias correction is possible only in the case in which there are appropriate auxiliary variables to include in the model used to impute the missing values. If this does not happen, analysis and imputation models reveal to be analogous.

Whenever the missingness mechanism cannot assumed to be random, and therefore missing data behave as MNAR, multiple imputation still allows to reduce the bias deriving from a complete case analysis under the assumption to have auxiliary variables which are strong predictors of missingness. Indeed, some bias is very likely to remain since the estimation of the imputed values is based only on the observed data.

The other main benefit characterizing the multiple imputation procedure consists of an improving in precision.

In a situation in which missingness occurs completely at random (MCAR missing data) a complete case analysis leads to obtain unbiased results because it includes a random sample of the original study subjects and therefore a random sample from the population (assuming the randomness of the original sample with respect to the whole population). Even if these results reveal to be unbiased, a relevant aspect of the complete case analysis is that it is conducted throwing away some information about the study participants affected by missing values in one or more covariates of interest. Then, a complete case analysis may lead to obtain wide confidence intervals around parameters estimates, i.e., it might be inefficient.

In this perspective, it is possible to state that the multiple imputation procedure can obtain narrower confidence intervals and consequently to obtain an improvement in efficiency. This is possible simply because the multiple imputation methodology, for its nature, allows to include all the participants in the analysis. Then it is interesting to understand under which circumstances the implementation of a multiple imputation procedure can lead to a major (or minor) increase in efficiency.

Multiple imputation may lead to a largest potential gain of efficiency over a complete case analysis in a framework in which the variables of interest, the exposure of interest and the outcome are all fully observed, but there are missing values in relevant confounders. In such case, excluding incomplete cases leads to a loss of information about the exposure-outcome relationship in cases in which the covariate is missing. Such information can, then, be recovered through the implementation of a multiple imputation mechanism.

Instead, when a dataset is characterized by missing exposure or outcome values, multiple imputation reveals to be less likely to gain information about the exposure-outcome association, unless to be in the case in which there are some auxiliary variables which present a high correlation with the covariate affected by missing data (Lee & Carlin 2012, Marshall et al. 2010).

In the end, the above-mentioned multiple imputation possible efficiency gain derives from the inclusion, in the imputation model, of the auxiliary variables. Therefore, the stronger reveals to be the association characterizing the incomplete variables and the auxiliary ones, the more the imputed values will be accurate and the more the multiple imputation methodology implemented will lead to an improvement in efficiency.

However, it is important to point out that in real analyses it is needed to exist a reasonably strong correlation between the incomplete variables and the auxiliary ones to observe, after the application of a multiple imputation procedure, a relevant gain in efficiency with respect to a complete case analysis (Graham, 2012). Moreover, in most cases it is hard to observe variables characterized by such a strong correlation (Karahalios et al., 2010).

In contrast to the above-mentioned advantages characterizing the multiple imputation procedure, it is relevant to point out the main drawback which may affect this type of methodology.

Indeed, multiple imputation may introduce bias over a complete case analysis if not carried out appropriately (Lee and Carlin, 2012). Specifically, when setting up the imputation model of the first stage of the multiple imputation process, there are some decisions which, if not taken in a proper way, may affect the validity of the inference results.

A first feature to take into consideration is the share of missing data affecting the dataset under analysis. If a researcher has to deal with a lot of missing data, any bias deriving from the decisions of the setting up framework regarding the imputation model will inevitably be inflated since a large amount of data will be imputed based on a potentially mis-specified model (Rubin, 1996).

The second relevant step regards the choice of the variables to include in the imputation model. Indeed, it is relevant to include all the covariates of the analysis model in the imputation model, plus any interaction term and the ones which may describe a nonlinear association, such as quadratics or logarithms transformations (Graham, 2012). If a researcher leaves one or more of these variables out of the imputation model, the inference results may be biased. Moreover, it is relevant to include auxiliary variables in the perspective of the recovery of information lost.

The third relevant feature is about the inclusion of non-normally distributed continues variables into the imputation model. In this perspective, both the above-mentioned methodologies of multiple imputation are characterized by the assumption of normality for continuous variables. In this perspective, including the original scale values of any non-normally distributed covariate in such imputation model may lead to obtain imputed values quite different from the observed ones. This may, consequently, lead to flow on effects to the inference obtained. However, it has been suggested that transforming data prior to imputation to improve normality may, on the other hand, lead to biased results in some cases (von Hippel, 2013).

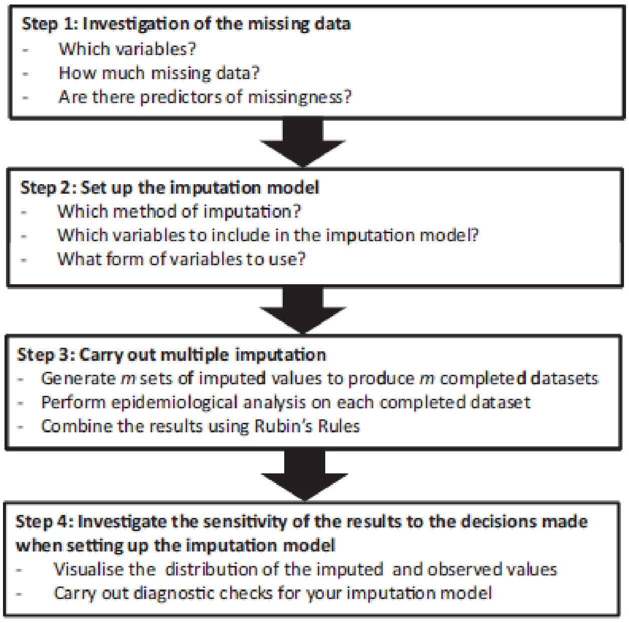
The fourth situation to take into consideration consists of the imputation of categorical variables when using a joint normal distribution. Indeed, since both main multiple imputation methodologies actually assume normality for all the variables present in the imputation model, it remains unclear, in this framework, which may be the best way to impute missing values affecting a categorical covariate (Galati et al. 2012, Lee et al. 2012).

Then, it is important to choose how to impute and analyze covariates characterized by a restricted range of values. In this perspective, various approaches have been suggested of the imputation and consequent analysis of restricted range variables (von Hippel 2013, Enders 2010, Royston et al. 2009).

It is crucial to take into consideration that all five the above-mentioned potential issues should be considered prior to imputation and with respect to the dataset under analysis in the research. Indeed, the flexibility characterizing the multiple imputation process suggests that it would be desirable to have some expertise in the methodology prior to using it and making the above reported relevant decisions (Lee and Simpson, 2014).

In the end, a relevant step consists of exploring the sensitivity of the obtained results to the decisions made in all the imputation process. Intuitively, the best scenario is the one in which all tested imputation models lead to the same general conclusion (Figure 3.2).

Process for carrying out multiple imputation.



Step 1: Investigation of the missing data
- Which variables?
- How much missing data?
- Are there predictors of missingness?

Step 2: Set up the imputation model
- Which method of imputation?
- Which variables to include in the imputation model?
- What form of variables to use?

Step 3: Carry out multiple imputation
- Generate *m* sets of imputed values to produce *m* completed datasets
- Perform epidemiological analysis on each completed dataset
- Combine the results using Rubin's Rules

Step 4: Investigate the sensitivity of the results to the decisions made when setting up the imputation model
- Visualise the distribution of the imputed and observed values
- Carry out diagnostic checks for your imputation model

Figure 3.2. Source: Lee and Simpson, 2014, "Introduction to multiple imputation for dealing with missing data"

3.2 Algorithms for multiple imputation

Multiple imputation procedure can be implemented with the use of algorithm within several different package presents within various statistical software.

In this perspective, Pampaka et al. (2016, “Handling missing data: analysis of a challenging data set using multiple imputation” provide a general overview of the main algorithms presented in literature to apply the multiple imputation methodology.

First, it is relevant to specify that, while some authors (Schafer and Graham, 2002) distinguish between multiple imputation and maximum-likelihood estimation[2] approaches for dealing with missing data, Pampaka et al. assume the two methodologies to be interconnected, since maximum-likelihood usually is used for the estimation of the imputation model.

Indeed, the essential element characterizing any approach is assumed to be the distribution of the observed data as a function of the population distribution (complete dataset) with respect to the missing values.

Maximum-likelihood approach estimation is based on maximizing the (log of the) so-called *likelihood function*. In most situations, such maximization is computed in an iterative way by using the so-called *expectation-maximization* (EM) *algorithm*, a very established statistical technique.

In general, maximum-likelihood methods summarize a likelihood function averaged over a predictive distribution for the missing values (Schafer 1999, Schafer and Graham 2002, Ibrahim et al. 2005).

In the last decades, Bayesian[3] multiple imputation methods are becoming more popular. These methodologies have the peculiarity to be performed using a Bayesian predictive distribution to generate the imputations (Nielsen, 2003) and

[2] Maximum-Likelihood estimation is a statistical method for estimating population parameters (i.e. mean and variance) from sample data that selects as estimates those parameter values maximizing the probability of obtaining the observed data

[3] Bayesian statistical methods assign probabilities or distributions to events or parameters (e.g. a population mean) based on experience or best guesses (more formally defined as prior distributions) and then apply Bayes’ theorem to revise the probabilities and distributions after considering the data, thus resulting in what is formally defined as posterior distribution

then specifying prior values for all the parameters of interest (Ibrahim et al., 2005). Moreover, Schafer and Graham (2002) state that Bayesian methodologies bring together multiple imputation methods and maximum-likelihood methods.

Such combination has been implemented in the last decades statistical packages: for example, Pampaka et al. (2016) in their work use a package called *Amelia II*.

In general, King et al. (2001), summarize multiple imputation algorithms as follows:

> computing the observed data likelihood [. . .] and taking random draws from it, is computationally infeasible with classical methods. Even maximizing the function takes inordinately long with standard optimization routines. In response to such difficulties, the Imputation-Posterior (IP) and Expectation-Maximization (EM) algorithms were devised and subsequently applied to this problem. From the perspective of statisticians, IP is now the gold standard of algorithms for multivariate normal multiple imputations, in large part because it can be adapted to numerous specialized models. Unfortunately, from the perspective of users, it is slow and hard to use. Because IP is based on Markov Chain Monte Carlo (MCMC) methods, considerable expertise is needed to judge convergence, and there is no agreement among experts about this except for special cases. (54)

Considering the above-reported issues, it has been developed the so called *EMB algorithm*, combining the typical EM[4] algorithm with a bootstrap approach to take draws from the posterior distribution. Such new algorithm, therefore, expands the range of computationally feasible data types and sized to which it is possible to apply the multiple imputation methodology.

Currently, in most used statistical software there are various packages to apply the multiple imputation method.

In R, various procedures are available to perform imputation of missing values:

- Amelia II (Honaker et al., 2011)
- arrayImpute (Lee et al., 2009)
- cat (for categorical-variables datasets affected by missing values; Schafer, 1997)

[4] Expectation-maximization

- EMV (for the Estimation of Missing Values for a Data Matrix; Gottardo, 2004)
- impute (Hastie et al., 2014)
- mi (Su et al., 2011)
- mice (Van Buuren and Groothuis-Oudshoorn, 2011)
- Hmisc (Harrell, 2008)

Other tools for performing multiple imputations of missing data are available within other statistical packages such as ICE in STATA, the SAS PROC MI, Missing Data Library, and NORM for S-Plus and SOLAS. Moreover, multiple imputation may also be applied using MLwiN or SPSS (Pampaka et al., 2016).

Horton and Kleinman (2007, "Much Ado about Nothing: A Comparison of missing Data Methods and Software to Fit Incomplete Data Regression Models.") applied imputation with Amelia II, Hmisc, mice and other statistical packages finding similar parameters estimates for all different analyses. Moreover, they obtained a relevant reduction regarding the standard error estimates with respect to the complete case analysis estimators.

Hutcheson and Pampaka (2012) also published a practical tutorial about the imputation of missing data using Amelia II.

Chhabra et al. (2017, "A Comparison of Multiple Imputation Methods for Data with Missing Values") implemented in the statistical software R a comparison between six multiple imputation methods included in the imputation package *mice*:

- Predictive Mean Matching
- Multiple Random Forest Regression Imputation
- Multiple Bayesian Regression Imputation
- Multiple Linear Regression using Non-Bayesian Imputation
- Multiple Classification and Regression Tree (CART)
- Multiple Linear Regression with Bootstrap imputation

The results obtained by Chhabra et al. will be analyzed in the next section, in the framework of a brief overview of some relevant applications of the multiple imputation procedure in literature.

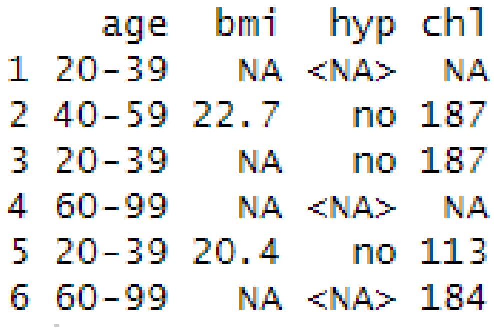
A simple application of the multiple imputation methodology has been implemented using the above-mentioned package *mice* available in the statistical software *R*.

Such package allows to implement a method to handle missing data by creating multiple imputation (replacement values) for multivariate missing data. The methodology applied by the package is based on Fully Conditional Specification technique, characterized by the fact that each incomplete covariate is imputed by a separate model.

Moreover, the *mice* algorithm is able to impute mixes of continuous, binary, unordered categorical and ordered categorical data. Indeed, such package can be used to impute continuous two-level data maintaining consistency between imputations by means of the so-called passive imputation.

For this example of application has been used the “nhanes2” dataset, that is a small dataset available in *R* designed for missing data examples. It is made of four variables, which are quantitative and binary (Figure 3.3).

The nhanes2 dataset.

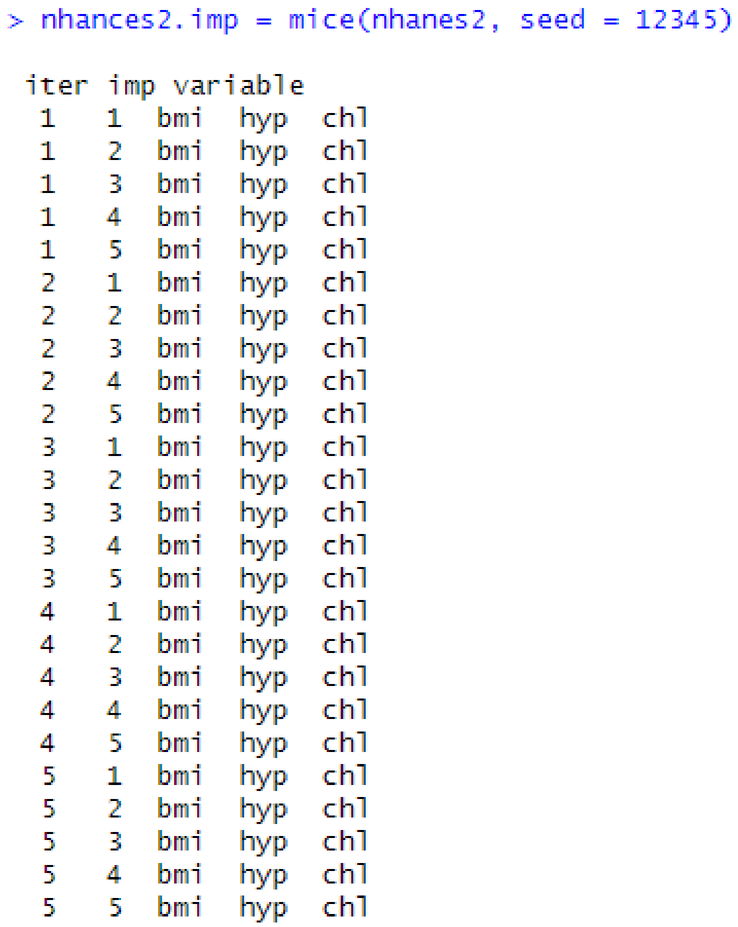


```
    age  bmi  hyp chl
1 20-39   NA <NA>  NA
2 40-59 22.7   no 187
3 20-39   NA   no 187
4 60-99   NA <NA>  NA
5 20-39 20.4   no 113
6 60-99   NA <NA> 184
```

Figure 3.3.

Then, the *mice* function has been used to create the default five imputed datasets (Figure 3.4):

The mice function.



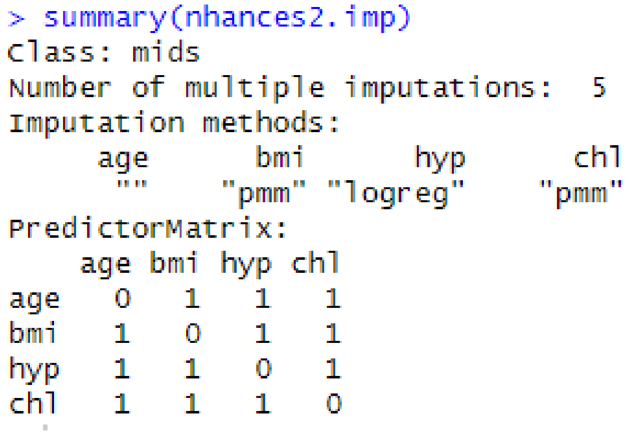
```
> nhances2.imp = mice(nhanes2, seed = 12345)

 iter imp variable
  1   1  bmi  hyp  chl
  1   2  bmi  hyp  chl
  1   3  bmi  hyp  chl
  1   4  bmi  hyp  chl
  1   5  bmi  hyp  chl
  2   1  bmi  hyp  chl
  2   2  bmi  hyp  chl
  2   3  bmi  hyp  chl
  2   4  bmi  hyp  chl
  2   5  bmi  hyp  chl
  3   1  bmi  hyp  chl
  3   2  bmi  hyp  chl
  3   3  bmi  hyp  chl
  3   4  bmi  hyp  chl
  3   5  bmi  hyp  chl
  4   1  bmi  hyp  chl
  4   2  bmi  hyp  chl
  4   3  bmi  hyp  chl
  4   4  bmi  hyp  chl
  4   5  bmi  hyp  chl
  5   1  bmi  hyp  chl
  5   2  bmi  hyp  chl
  5   3  bmi  hyp  chl
  5   4  bmi  hyp  chl
  5   5  bmi  hyp  chl
```

Figure 3.4.

Using the *summary* function, it is possible to have an overview of the tasks completed by the imputation method (Figure 3.5):

Summary of the imputation procedure.



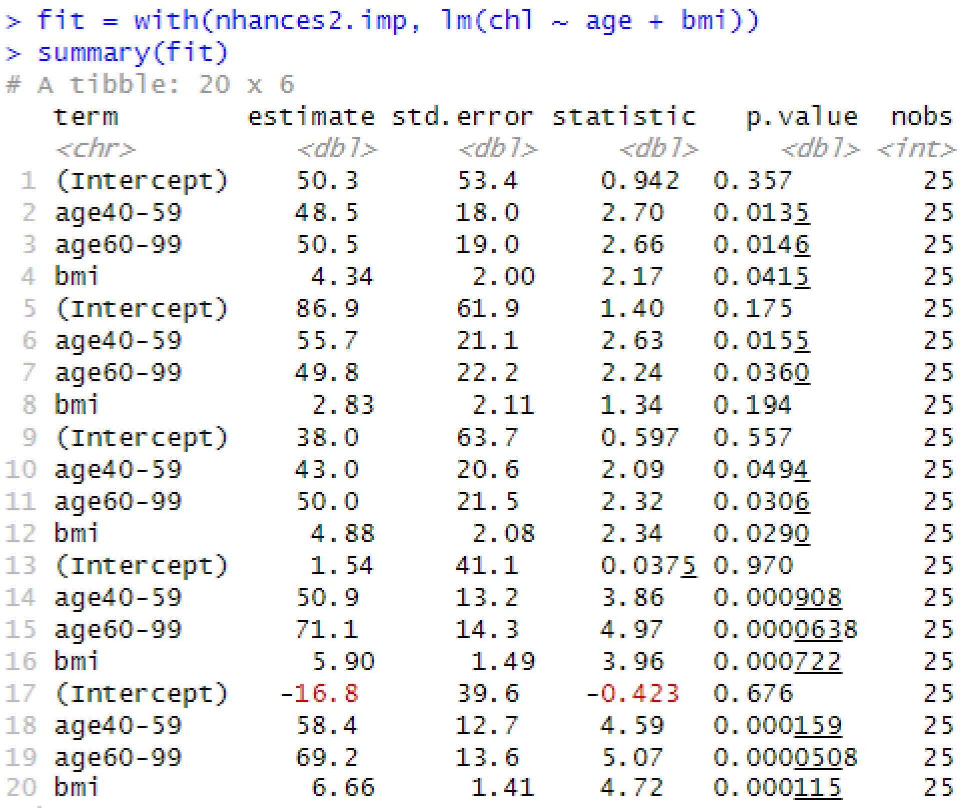
```
> summary(nhances2.imp)
Class: mids
Number of multiple imputations:  5
Imputation methods:
     age       bmi       hyp       chl
      ""     "pmm"  "logreg"     "pmm"
PredictorMatrix:
    age bmi hyp chl
age   0   1   1   1
bmi   1   0   1   1
hyp   1   1   0   1
chl   1   1   1   0
```

Figure 3.5.

The above summary shows the number of imputed datasets (in this case, five), the imputation method used by the algorithm (for categorical variables a log regression since the data is not continuous) and, in the end, the predictor matrix which reports the variables used in predicting missing values for a specific variable (e.g., for age the model used all the other three included variables).

Then, it is possible to run a standard regression using the above-reported imputed datasets. Specifically, it has been run a regression model which regresses *chl* on the covariates *age* and *bmi* (Figure 3.6).

Standard regression using the imputed datasets.

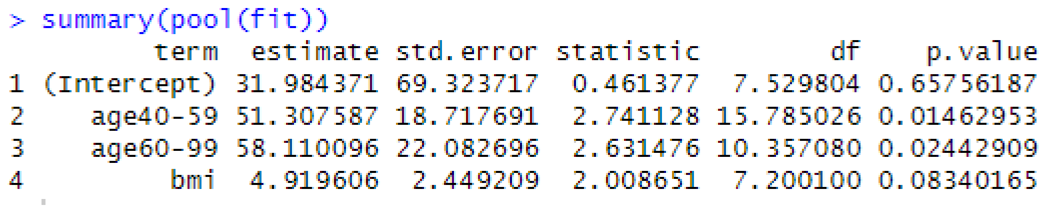


```
> fit = with(nhances2.imp, lm(chl ~ age + bmi))
> summary(fit)
# A tibble: 20 x 6
   term        estimate std.error statistic   p.value  nobs
   <chr>          <dbl>     <dbl>     <dbl>     <dbl> <int>
 1 (Intercept)    50.3      53.4     0.942  0.357        25
 2 age40-59       48.5      18.0     2.70   0.0135       25
 3 age60-99       50.5      19.0     2.66   0.0146       25
 4 bmi             4.34      2.00    2.17   0.0415       25
 5 (Intercept)    86.9      61.9     1.40   0.175        25
 6 age40-59       55.7      21.1     2.63   0.0155       25
 7 age60-99       49.8      22.2     2.24   0.0360       25
 8 bmi             2.83      2.11    1.34   0.194        25
 9 (Intercept)    38.0      63.7     0.597  0.557        25
10 age40-59       43.0      20.6     2.09   0.0494       25
11 age60-99       50.0      21.5     2.32   0.0306       25
12 bmi             4.88      2.08    2.34   0.0290       25
13 (Intercept)     1.54     41.1     0.0375 0.970        25
14 age40-59       50.9      13.2     3.86   0.000908     25
15 age60-99       71.1      14.3     4.97   0.0000638    25
16 bmi             5.90      1.49    3.96   0.000722     25
17 (Intercept)   -16.8      39.6    -0.423  0.676        25
18 age40-59       58.4      12.7     4.59   0.000159     25
19 age60-99       69.2      13.6     5.07   0.0000508    25
20 bmi             6.66      1.41    4.72   0.000115     25
```

Figure 3.6.

In the end, it is possible to use the *pool* function to pool the regression results together over the five imputed datasets to obtain just a unique final result (Figure 3.7).

Regression output after the application of the pool function.



```
> summary(pool(fit))
         term  estimate std.error statistic        df    p.value
1 (Intercept) 31.984371 69.323717  0.461377  7.529804 0.65756187
2    age40-59 51.307587 18.717691  2.741128 15.785026 0.01462953
3    age60-99 58.110096 22.082696  2.631476 10.357080 0.02442909
4         bmi  4.919606  2.449209  2.008651  7.200100 0.08340165
```

Figure 3.7.

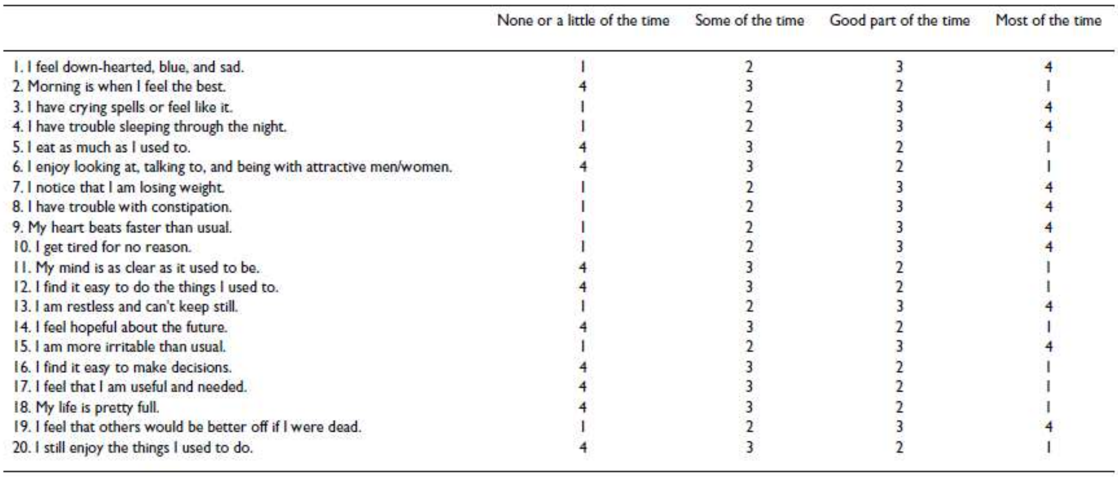
3.3 Examples of application in literature

In recent decades' literature many authors have conducted research, belonging to many different fields, applying various multiple imputation methodologies to deal with missing information in the datasets under analysis.

Shrive et al. (2006, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods") compare six different imputation techniques for dealing with missing data in the Zung Self-reported Depression scale (SDS).

The Self-reported depression scale questionnaire consists of a 20 question scale (Table 3.1).

The Zung Self-rating Depression scale (SDS).



	None or a little of the time	Some of the time	Good part of the time	Most of the time
1. I feel down-hearted, blue, and sad.	1	2	3	4
2. Morning is when I feel the best.	4	3	2	1
3. I have crying spells or feel like it.	1	2	3	4
4. I have trouble sleeping through the night.	1	2	3	4
5. I eat as much as I used to.	4	3	2	1
6. I enjoy looking at, talking to, and being with attractive men/women.	4	3	2	1
7. I notice that I am losing weight.	1	2	3	4
8. I have trouble with constipation.	1	2	3	4
9. My heart beats faster than usual.	1	2	3	4
10. I get tired for no reason.	1	2	3	4
11. My mind is as clear as it used to be.	4	3	2	1
12. I find it easy to do the things I used to.	4	3	2	1
13. I am restless and can't keep still.	1	2	3	4
14. I feel hopeful about the future.	4	3	2	1
15. I am more irritable than usual.	1	2	3	4
16. I find it easy to make decisions.	4	3	2	1
17. I feel that I am useful and needed.	4	3	2	1
18. My life is pretty full.	4	3	2	1
19. I feel that others would be better off if I were dead.	1	2	3	4
20. I still enjoy the things I used to do.	4	3	2	1

Table 3.1. Source: Shrive et al., 2006, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods"

As shown in the above table, each question has a score between 1 and 4; then, the sum of the responses is calculated. Moreover, such sum of scores across the 20 questions is converted to a 100-point scale by dividing the sum by 0.8.

The participants of the study were 1931 surgical patients: among them, 1580 patients completed all 20 questions of the SDS[5] questionnaire, while the remaining 351 did not fully complete the instrument. In detail, the quantity of missing values affecting the 351 subjects occasionally involved only one missing response in the entire instrument. Indeed, most participants were characterized by four or less missing items.

Considering the 1580 subjects who completed all the items of the questionnaire, the authors simulated missing values in these complete cases by assigning each observation a number between 0 and 1 selected in a random way from a uniform distribution (0, 1)[6]. Then, the assigned value was used to assign missing values to selected observations.

Initially, the authors simulated three MCAR scenarios in which the probability of missingness is assumed not to be linked to any other patient characteristic. In such framework, observations assigned a value lower than 0.10 were deleted, consequently simulating a study characterized by the missingness of 10% of the originally collected data. Then, for the subsequent MCAR simulations the threshold value was increased first to 0.20 and then to 0.30. In the end, subjects without deleted values were removed from the analysis, since they had no missing values to impute.

Moreover, the authors considered an unbalanced MCAR scenario in which the probability of missing question 6 was 20%, while the same probability for the other questions was of 10%, Such simulation is referred to as the “Q6” simulation.

Next, a MAR simulation was implemented, characterized by the fact that the probability of missingness was linked to known patient characteristics. Specifically, the probability of a missing value was linked to the subject’s gender of the patient: females over 65 were assigned a non-response probability of 20%, while for all the other patients such probability was assumed to be 10%.

In the end, the authors considered a MNAR framework in which the probability of missingness is assumed to depend on unknown patient characteristics. In such context, all questions except for question 6 were assigned

[5] Self-rating Depression scale

[6] Each number between 0 and 1 is characterized by an equal probability of beign assigned

a missingness probability of 10%. Moreover, with a response to question 6 equal to 1 or 2, the probability of missingness for question 6 was assumed to be 5%, while with the above-mentioned responses being 3 or 4 the missingness probability for question 6 such probability increased to 20%.

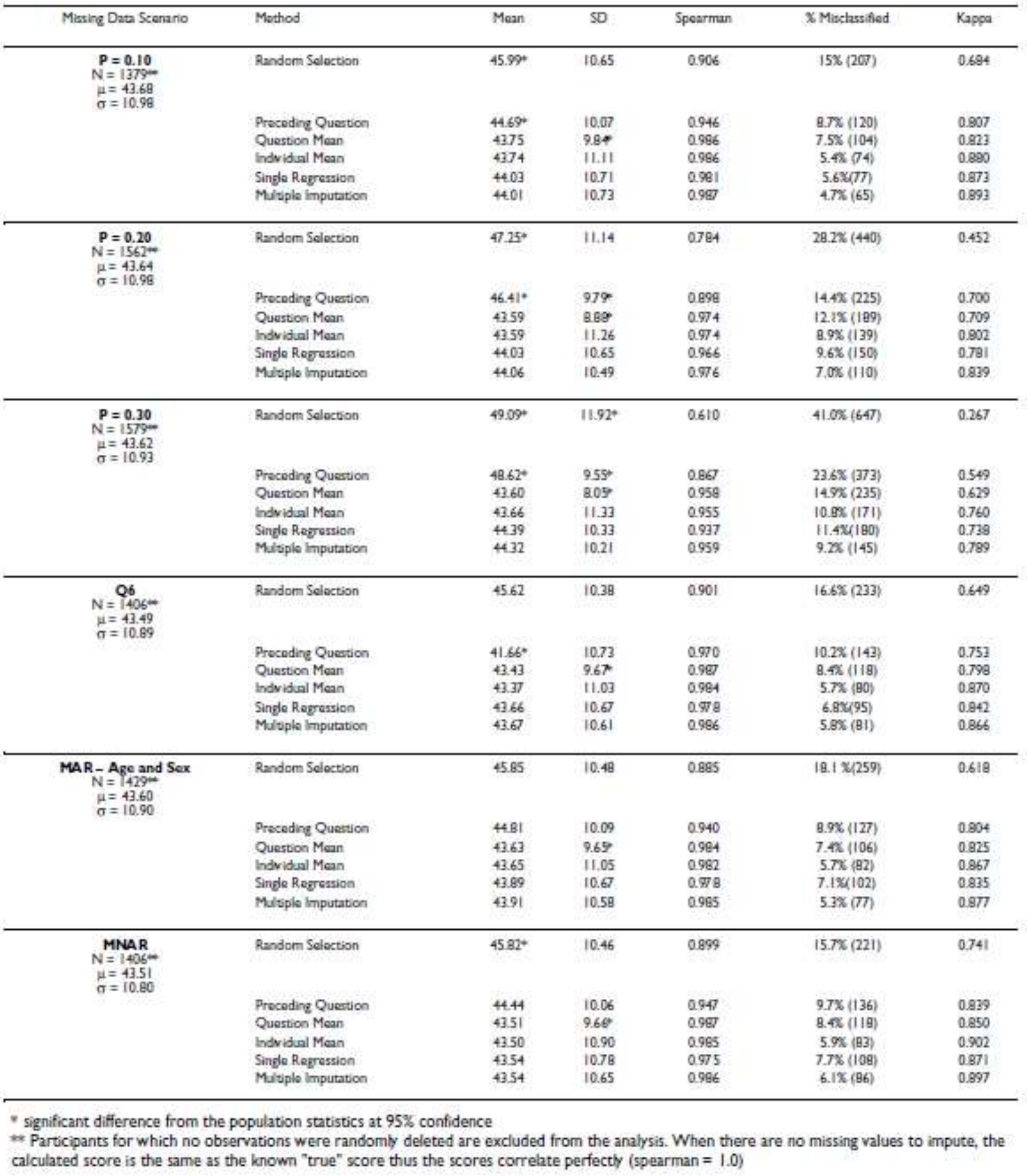
As tool for the analysis, Shrive et al. compared six different imputation methodologies:

- Random Selection
- Proceding Response
- Question Mean
- Individual Mean
- Single Regression
- Multiple Imputation

As for multiple imputation, an experimental version of multiple imputation available in SAS 8.1 was applied. Moreover, the missing data are filled five times generating five unique and completed datasets, with each of them analyzed separately to calculate a mean and a standard deviation. The following step consists of combining the results from the different analyses to produce, for each missing value of interest, a mean and a standard deviation. In this context, the predictors used in the multiple imputation procedure to predict missing values were the responses to completed questions.

A comparison between the six above-mentioned imputation method applied allows to state that the multiple imputation procedure reveals to be, without any doubt, the most accurate imputation methodology for this analysis (Table 3.2, Figure 3.8).

Diagnostic measures for imputation methods.



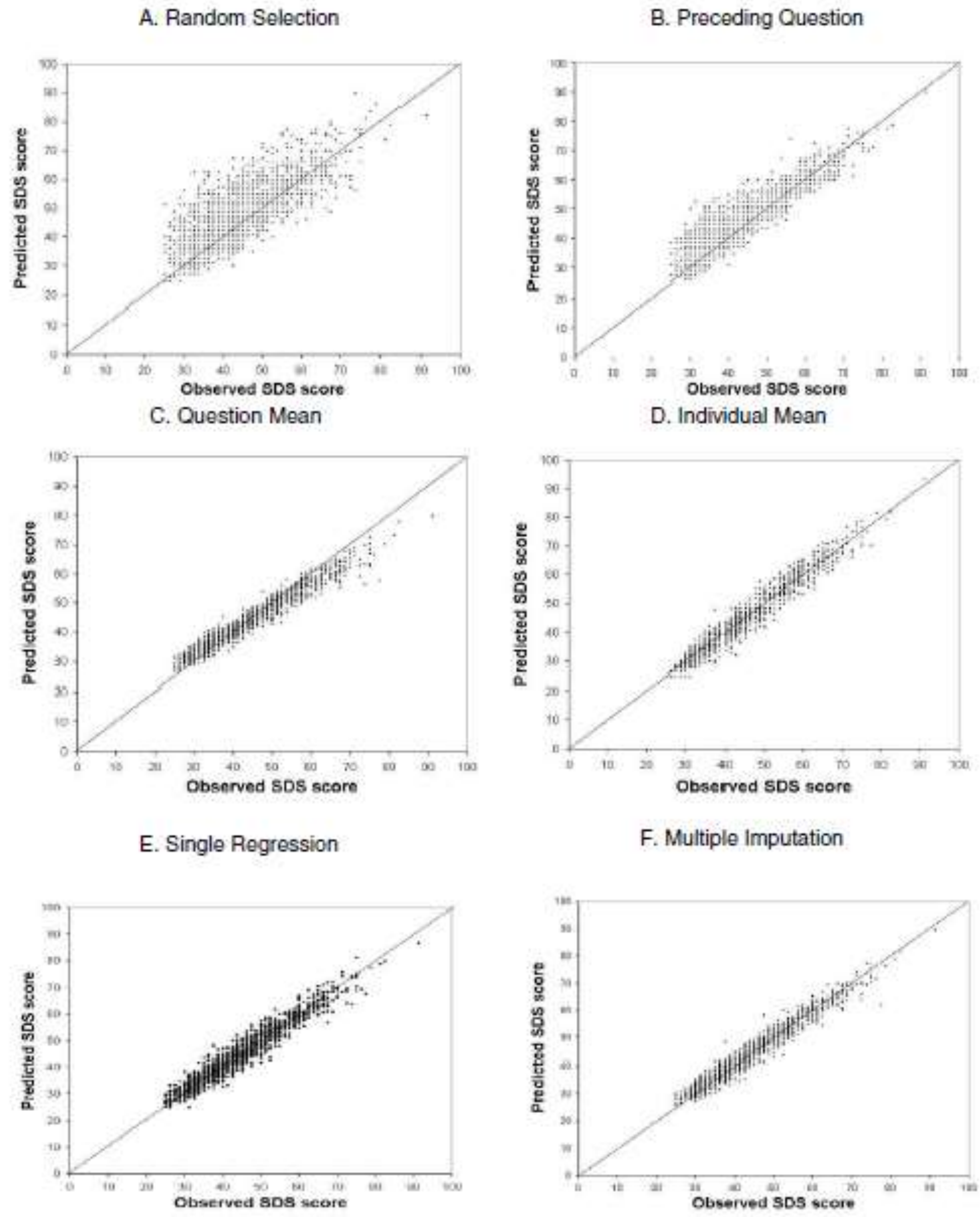
Missing Data Scenario	Method	Mean	SD	Spearman	% Misclassified	Kappa
P = 0.10 N = 1379** μ = 43.68 σ = 10.98	Random Selection	45.99*	10.65	0.906	15% (207)	0.684
	Preceding Question	44.69*	10.07	0.946	8.7% (120)	0.807
	Question Mean	43.75	9.84*	0.986	7.5% (104)	0.823
	Individual Mean	43.74	11.11	0.986	5.4% (74)	0.880
	Single Regression	44.03	10.71	0.981	5.6%(77)	0.873
	Multiple Imputation	44.01	10.73	0.987	4.7% (65)	0.893
P = 0.20 N = 1562** μ = 43.64 σ = 10.98	Random Selection	47.25*	11.14	0.784	28.2% (440)	0.452
	Preceding Question	46.41*	9.79*	0.898	14.4% (225)	0.700
	Question Mean	43.59	8.88*	0.974	12.1% (189)	0.709
	Individual Mean	43.59	11.26	0.974	8.9% (139)	0.802
	Single Regression	44.03	10.65	0.966	9.6% (150)	0.781
	Multiple Imputation	44.06	10.49	0.976	7.0% (110)	0.839
P = 0.30 N = 1579** μ = 43.62 σ = 10.93	Random Selection	49.09*	11.92*	0.610	41.0% (647)	0.267
	Preceding Question	48.62*	9.55*	0.867	23.6% (373)	0.549
	Question Mean	43.60	8.05*	0.958	14.9% (235)	0.629
	Individual Mean	43.66	11.33	0.955	10.8% (171)	0.760
	Single Regression	44.39	10.33	0.937	11.4%(180)	0.738
	Multiple Imputation	44.32	10.21	0.959	9.2% (145)	0.789
Q6 N = 1406** μ = 43.49 σ = 10.89	Random Selection	45.62	10.38	0.901	16.6% (233)	0.649
	Preceding Question	41.66*	10.73	0.970	10.2% (143)	0.753
	Question Mean	43.43	9.67*	0.987	8.4% (118)	0.798
	Individual Mean	43.37	11.03	0.984	5.7% (80)	0.870
	Single Regression	43.66	10.67	0.978	6.8%(95)	0.842
	Multiple Imputation	43.67	10.61	0.986	5.8% (81)	0.866
MAR – Age and Sex N = 1429** μ = 43.60 σ = 10.90	Random Selection	45.85	10.48	0.885	18.1 %(259)	0.618
	Preceding Question	44.81	10.09	0.940	8.9% (127)	0.804
	Question Mean	43.63	9.65*	0.984	7.4% (106)	0.825
	Individual Mean	43.65	11.05	0.982	5.7% (82)	0.867
	Single Regression	43.89	10.67	0.978	7.1%(102)	0.835
	Multiple Imputation	43.91	10.58	0.985	5.3% (77)	0.877
MNAR N = 1406** μ = 43.51 σ = 10.80	Random Selection	45.82*	10.46	0.899	15.7% (221)	0.741
	Preceding Question	44.44	10.06	0.947	9.7% (136)	0.839
	Question Mean	43.51	9.66*	0.987	8.4% (118)	0.850
	Individual Mean	43.50	10.90	0.985	5.9% (83)	0.902
	Single Regression	43.54	10.78	0.975	7.7% (108)	0.871
	Multiple Imputation	43.54	10.65	0.986	6.1% (86)	0.897

* significant difference from the population statistics at 95% confidence

** Participants for which no observations were randomly deleted are excluded from the analysis. When there are no missing values to impute, the calculated score is the same as the known "true" score thus the scores correlate perfectly (spearman = 1.0)

Table 3.2. Source: Shrive et al., 2006, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods"

Predicted versus observed scores for each imputation technique with a probability of missing of 20%.



A. Random Selection

B. Preceding Question

C. Question Mean

D. Individual Mean

E. Single Regression

F. Multiple Imputation

Figure 3.8. Source: Shrive et al., 2006, "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods"

Pampaka et al. (2016, “Handling missing data: analysis of a challenging data set using multiple imputation”) applied multiple imputation in the field of educational research.

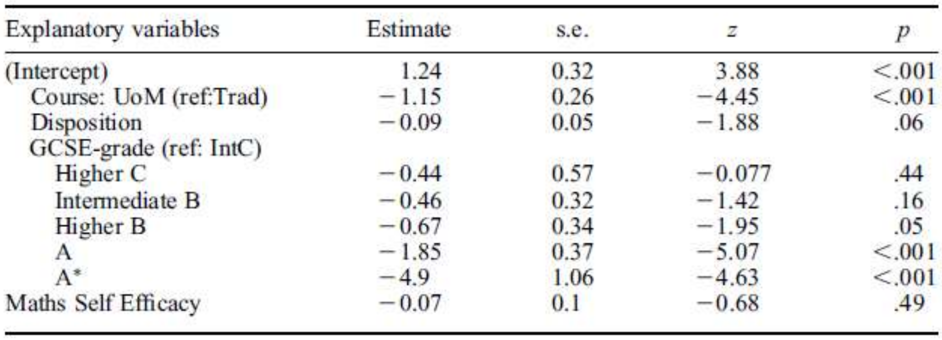
Specifically, Pampaka et al. conducted an analysis, in the context of UK schooling system. In particular, the authors were interested in modelling whether students dropped out of the mathematics courses they were enrolled on. The authors start from an existing original study (Hutcheson et al., 2011), in which such drop-out variable was found to be related to the typology of course they were on, their previous GCSE[7] score in mathematics, their disposition to study mathematics at high level and their self-efficacy rating.

Indeed, the analysis carried out by Pampaka et al. is restricted to the data in the model used in the above-mentioned original paper.

First, the outcome variable of interest (the dropout) is modeled using the initial data affected by missing values (which amount to 495 out of 1374), then these results are compared to a model in which missing data are imputed (consequently, n = 1374).

The results obtained carrying out the analysis, by the means of a logistic regression, using only the 495 completed data points are shown in Table 3.3:

A logistic regression model of “dropout” using the 495 cases available at the end of the initial study.



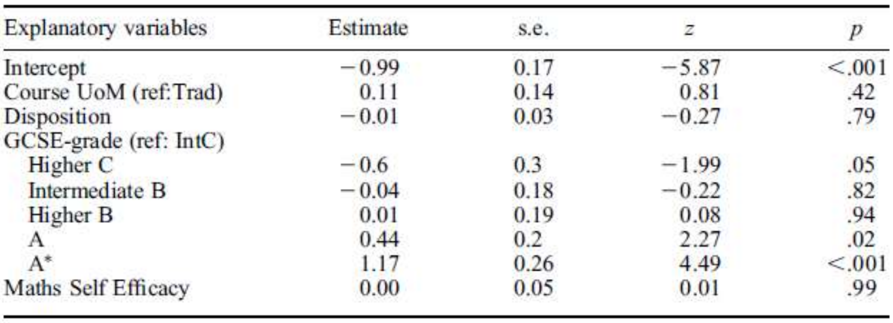
Explanatory variables	Estimate	s.e.	*z*	*p*
(Intercept)	1.24	0.32	3.88	<.001
Course: UoM (ref:Trad)	−1.15	0.26	−4.45	<.001
Disposition	−0.09	0.05	−1.88	.06
GCSE-grade (ref: IntC)				
Higher C	−0.44	0.57	−0.077	.44
Intermediate B	−0.46	0.32	−1.42	.16
Higher B	−0.67	0.34	−1.95	.05
A	−1.85	0.37	−5.07	<.001
A*	−4.9	1.06	−4.63	<.001
Maths Self Efficacy	−0.07	0.1	−0.68	.49

Table 3.3. Source: Pampaka et al., 2016, “Handling missing data: analysis of a challenging data set using multiple imputation”

[7] GCSE qualifications are usually taken at the end of compulsory education in a range of subjects. Students typically take about 8-10 of these in a range of subjects that must include English and mathematics.

However, since the missing data are unlikely to behave as MCAR, the above model is likely to provide a biased picture of the outcome variable of interest. In this perspective, the authors use a logistic regression to show the non-random nature of the missing data which occur in this framework (Table 3.4).

A logistic regression model of missingness on dropout variable.



Explanatory variables	Estimate	s.e.	z	p
Intercept	−0.99	0.17	−5.87	<.001
Course UoM (ref:Trad)	0.11	0.14	0.81	.42
Disposition	−0.01	0.03	−0.27	.79
GCSE-grade (ref: IntC)				
Higher C	−0.6	0.3	−1.99	.05
Intermediate B	−0.04	0.18	−0.22	.82
Higher B	0.01	0.19	0.08	.94
A	0.44	0.2	2.27	.02
A*	1.17	0.26	4.49	<.001
Maths Self Efficacy	0.00	0.05	0.01	.99

Table 3.4. Source: Pampaka et al., 2016, "Handling missing data: analysis of a challenging data set using multiple imputation"

Considering such missingness mechanism, it is possible to state that the model shown in Table 3.3 is likely to overestimate the effect of the high-achieving pupils.

To address the potential bias characterizing the 495 subjects' sample, Pampaka et al. imputed the 879 missing values using, in the statistical software *R*, the above-mentioned *Amelia II package*, which assumes that the complete data are multivariate normal, and that the missing data follow a MAR missingness mechanism.

The above Table 3.4, together with Figure 3.9, show that the occurrence of missingness depends on GCSE grades, which is an observed variable. Moreover, Amelia II is assumed to be an appropriate package for this analysis because the missing values are binary.

Probability of providing information.

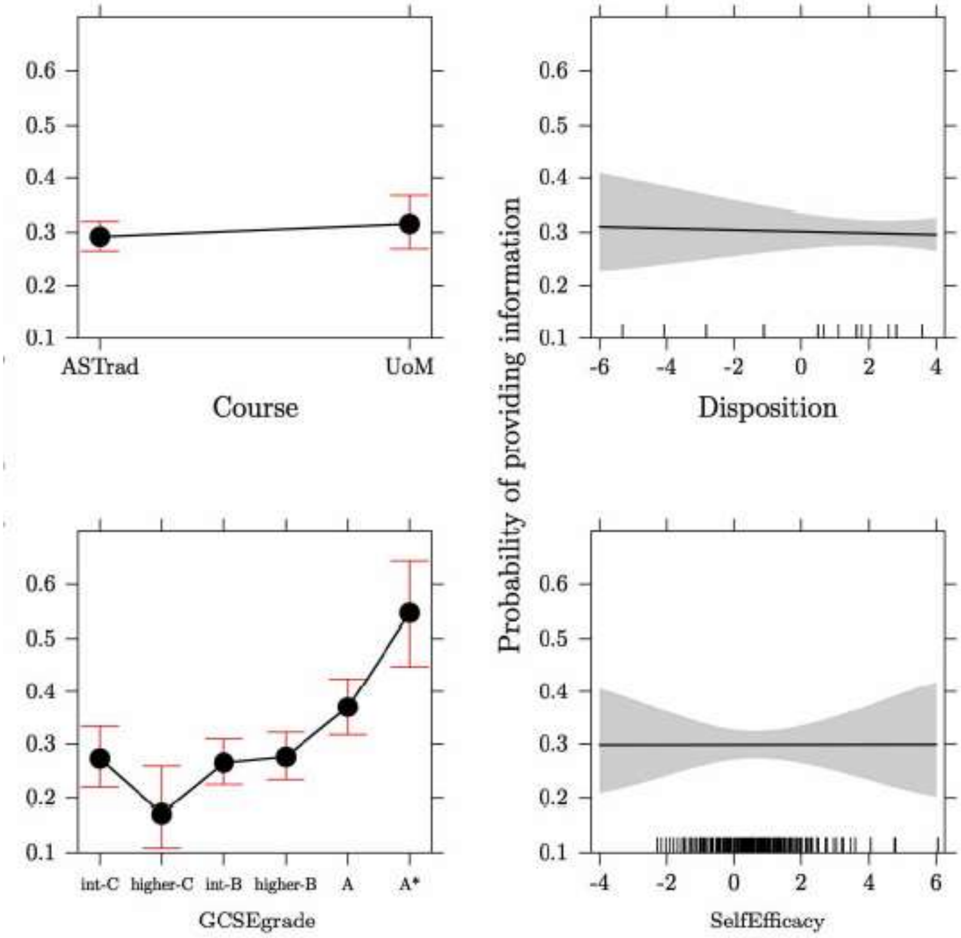


Figure 3.9. Source: Pampaka et al., 2016, "Handling missing data: analysis of a challenging data set using multiple imputation"

The imputation model included a number of variables available in the full dataset (Course, Disposition, GCSE-Grade, Maths Self Efficacy) plus some additional covariates such as information about EMA[8], ethnicity, gender, Language, LPN[9], uniFAM[10] and HEFCE[11].

Amelia II imputed 100 separate datasets and, in order to get parameter estimates for the overall imputed model, such imputation models were combined

[8] i.e. whether the student was holding Educational Maintenance Allowance

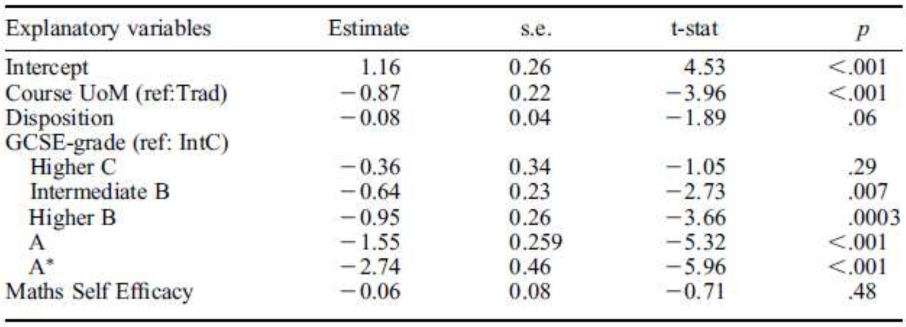
[9] i.e. whether the student was from Low Participation Neighborhood

[10] Whether the student was not first generation at HE

[11] An ordered categorical variable denoting socio-economic status

obtaining the combined estimates and standard errors using the Zelig library (Owen et al., 2013) available in the *R* package. The overall statistics for the imputed models computed using such library are shown in Table 3.5.

A logistic regression model of "dropout" using imputed data (n = 1374).



Explanatory variables	Estimate	s.e.	t-stat	*p*
Intercept	1.16	0.26	4.53	<.001
Course UoM (ref:Trad)	−0.87	0.22	−3.96	<.001
Disposition	−0.08	0.04	−1.89	.06
GCSE-grade (ref: IntC)				
Higher C	−0.36	0.34	−1.05	.29
Intermediate B	−0.64	0.23	−2.73	.007
Higher B	−0.95	0.26	−3.66	.0003
A	−1.55	0.259	−5.32	<.001
A*	−2.74	0.46	−5.96	<.001
Maths Self Efficacy	−0.06	0.08	−0.71	.48

Table 3.5. Source: Pampaka et al., 2016, "Handling missing data: analysis of a challenging data set using multiple imputation"

Even if the conclusions for the model based on the imputed data are similar to the ones for the model affected by missing data (n = 495), it is important to notice the relevant difference found in the standard error estimates for the GCSE grades. In this sense, the model characterized by the use of the multiple imputation methodology allows for a better differentiation of the covariate "GCSE-grade", leading to significant differences between more categories with respect to the initial model (the Higher B and Intermediate B groups are now significantly different to the reference category).

As introduced in the previous section of this work, Chhabra et al. (2017, "A Comparison of Multiple Imputation Methods for Data with Missing Values") apply six different multiple imputation techniques all available in the statistical package *mice* within the software *R*. Then, it is useful to provide an high level overview of some of the six multiple imputation methodologies (Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, Multiple Linear Regression using Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART), Multiple

Linear Regression with Bootstrap Imputation).

The Predictive Mean Matching technique consists of an attractive technique available for missing value substitution under the occurrence of quantitative variables. Such methodology uses the linear regression and the nearest-neighbor together to estimate the values of interest.

Multiple Random Forest Regression Imputation is characterized by the fact that a forest of classification or regression trees is constructed using bootstrap – or subsamples of the original data and the majority vote or overall average of trees generate the prediction rule for the target variable (Chhabra et al., 2017).

Multiple Classification and Regression Tree (CART) consists of an algorithm for both classification and regression, which makes use of decision trees that are binary to classify new data.

In the end, Multiple Linear Regression with Bootstrap Imputation uses any test or metric which relies on random sampling with replacement.

In their work, Chhabra et al. use iris dataset from UC Irvine Machine Learning Repository, composed by three classes, each one of them having 50 cases. Moreover, the dataset is characterized by four continuous features (sepal width, sepal length, petal width, petal length) introduced artificially with a percentage of missing values around 20% (Figure 3.10).

Missing value mechanism in sample data.

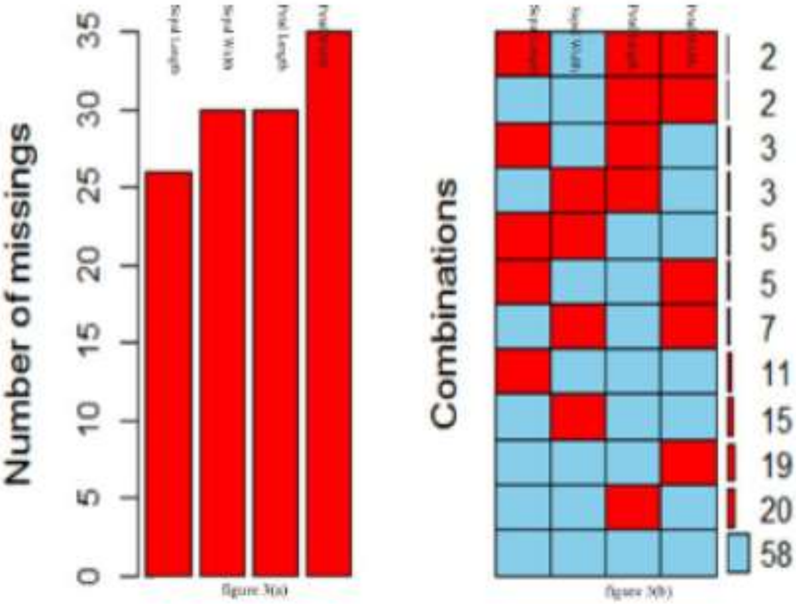
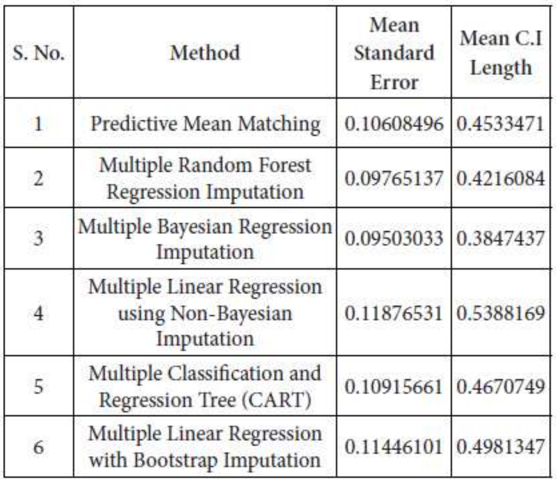


Figure 3.10. Source: Chhabra et al., 2017, "A Comparison of Multiple Imputation Methods for Data with Missing Values"

Applying the six above-mentioned multiple imputation methods available in the package *mice*, the authors observed the performance shown in Table 3.6:

Comparison of different Multiple Imputation methods.



S. No.	Method	Mean Standard Error	Mean C.I Length
1	Predictive Mean Matching	0.10608496	0.4533471
2	Multiple Random Forest Regression Imputation	0.09765137	0.4216084
3	Multiple Bayesian Regression Imputation	0.09503033	0.3847437
4	Multiple Linear Regression using Non-Bayesian Imputation	0.11876531	0.5388169
5	Multiple Classification and Regression Tree (CART)	0.10915661	0.4670749
6	Multiple Linear Regression with Bootstrap Imputation	0.11446101	0.4981347

Table 3.6. Source: Chhabra et al., 2017, "A Comparison of Multiple Imputation Methods for Data with Missing Values"

Observing the above table, it is possible to state that standard error and mean confidence interval length is the least in the case of Multiple Imputation combined with Bayesian Regression. Moreover, the results obtained by applying the Multiple Random Forest Regression Imputation reveal to be quite similar to the ones obtained with the Multiple Bayesian Regression Imputation.

In the end, the authors explain that a possible explanation driving the gain of efficiency applying the Multiple Imputation combined with Bayesian Regression is represented by the fact that such technique can make better use of the available data by accommodating nonlinearities among the predictors (Chhabra et al., 2017).

Chapter IV

Multiple imputation for different types of cross-sectional data

A relevant feature of the above-described multiple imputation techniques consists of the fact that such methodologies may be applied to a variety of different fields and situations.

Among these, multiple imputation methods can be applied in context characterized by different types of cross-sectional data.

In this perspective, in the next section will be provided an overview of the application of multiple imputation in situations in which multiple types of cross-sectional data.

Specifically, the following cross-sectional data frameworks will be analyzed:

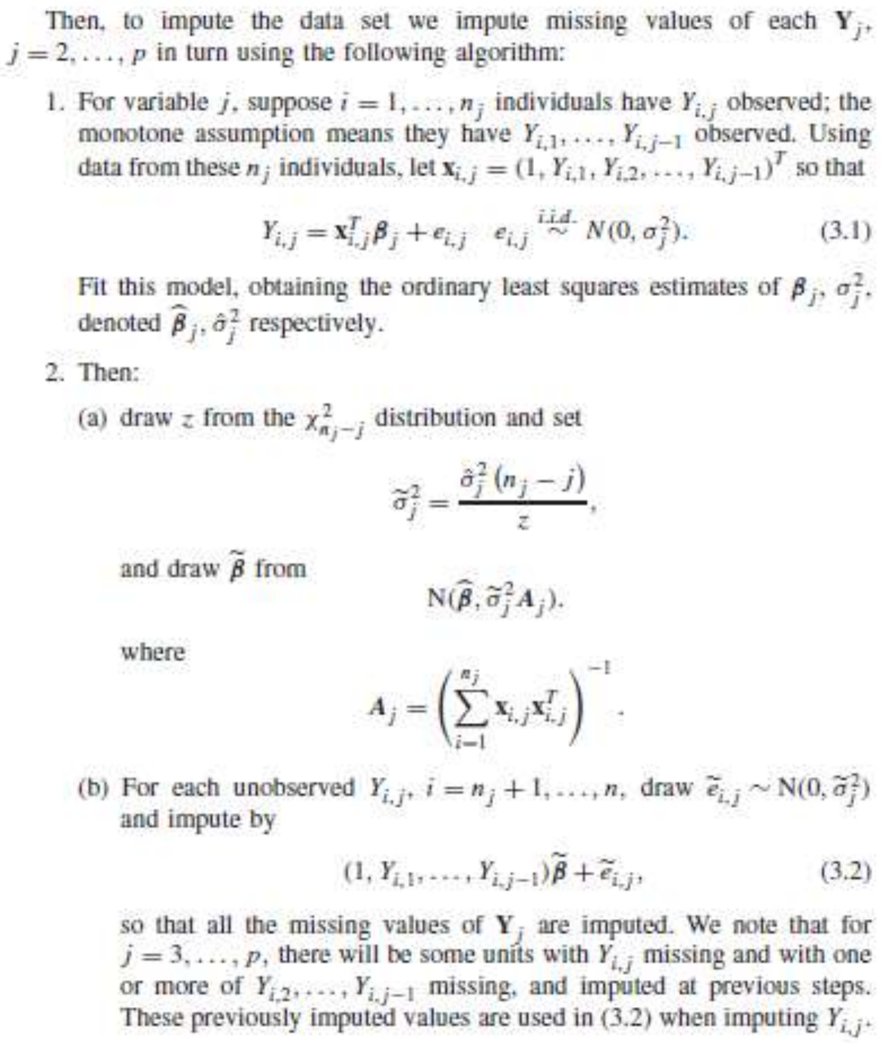
- Quantitative data
- Binary and ordinal data
- Unordered categorical data

4.1 Quantitative data

Multiple imputation procedure can be applied, first, to cross-sectional missing data which joint distribution can be considered to be multivariate normal (in this context, so-called quantitative data).

Assuming to observe missing data characterized by a monotone missing pattern and by a MAR missingness mechanism (with MCAR as a special case), it is sufficient to use a regression-based imputation algorithm to fill in the missing values, as explained in detail by Carpenter and Kenward (2013) (Figure 4.1):

Regression imputation with a monotone missingness pattern.



Then, to impute the data set we impute missing values of each $\mathbf{Y}_j$, $j = 2, \ldots, p$ in turn using the following algorithm:

1. For variable j, suppose $i = 1, \ldots, n_j$ individuals have $Y_{i,j}$ observed; the monotone assumption means they have $Y_{i,1}, \ldots, Y_{i,j-1}$ observed. Using data from these n_j individuals, let $\mathbf{x}_{i,j} = (1, Y_{i,1}, Y_{i,2}, \ldots, Y_{i,j-1})^T$ so that

$$Y_{i,j} = \mathbf{x}_{i,j}^T \boldsymbol{\beta}_j + e_{i,j} \quad e_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_j^2). \tag{3.1}$$

Fit this model, obtaining the ordinary least squares estimates of $\boldsymbol{\beta}_j$, σ_j^2, denoted $\widehat{\boldsymbol{\beta}}_j$, $\hat{\sigma}_j^2$ respectively.

2. Then:

(a) draw z from the $\chi^2_{n_j - j}$ distribution and set

$$\widetilde{\sigma}_j^2 = \frac{\hat{\sigma}_j^2 (n_j - j)}{z},$$

and draw $\widetilde{\boldsymbol{\beta}}$ from

$$\mathrm{N}(\widehat{\boldsymbol{\beta}}, \widetilde{\sigma}_j^2 \mathbf{A}_j).$$

where

$$\mathbf{A}_j = \left(\sum_{i=1}^{n_j} \mathbf{x}_{i,j} \mathbf{x}_{i,j}^T \right)^{-1}.$$

(b) For each unobserved $Y_{i,j}$, $i = n_j + 1, \ldots, n$, draw $\widetilde{e}_{i,j} \sim \mathrm{N}(0, \widetilde{\sigma}_j^2)$ and impute by

$$(1, Y_{i,1}, \ldots, Y_{i,j-1}) \widetilde{\boldsymbol{\beta}} + \widetilde{e}_{i,j}, \tag{3.2}$$

so that all the missing values of $\mathbf{Y}_j$ are imputed. We note that for $j = 3, \ldots, p$, there will be some units with $Y_{i,j}$ missing and with one or more of $Y_{i,2}, \ldots, Y_{i,j-1}$ missing, and imputed at previous steps. These previously imputed values are used in (3.2) when imputing $Y_{i,j}$.

Figure 4.1. Source: Carpenter and Kenward, 2013, "Multiple Imputation and Its Application"

In such framework, performing the above steps 1-2 for $j = 2, \ldots, p$ it is possible to obtain the first imputed dataset. Then, it is necessary to repeat the whole sequence to generate the successive imputed datasets.

The second approach to deal with quantitative missing data consists of the so-called *joint modelling*.

In such framework, no assumption is made about the missingness pattern underlying the missing data. However, the missingness mechanism is assumed to behave as MAR.

Then, the imputation model for the data is defined as the following multivariate normal model:

$$Y \sim \mathrm{N}(\beta, \Omega)$$

Where Ω consists of the unstructured covariance matrix. In the end, for the imputation it is used the so-called *Gibbs sampler*.

Another fundamental approach to handle quantitative missing data is the *full conditional specification* methodology.

This method is originated by relaxing the assumption that all covariate values in the sequential regressions are actually observed. By relaxing such hypothesis, in literature was proposed an approach called *imputation using chained equations* (ICE), known nowadays as *full conditional specification* (FCS).

Such methodologies have been proposed, among others, by van Buuren et al. (1999), Raghunathan et al. (2001) and van Buuren (2007).

In such approach, the first thing that is necessary is to re-order the variables so that the missingness pattern is as close as possible to be a monotone pattern. Then, it is necessary to fill in the missing values for each variable, typically by simply drawing, with replacement, from the observed values characterizing each variable.

In practice, the algorithm works as follows:

1. It implements a regression of *the observed part* of the variable of interest an all the remaining ones, with the missing values set at their current imputed values;
2. It imputes the missing values by using the regression imputation algorithm.

Running through for several times the two above-reported steps, it is called a *cycle*. After finishing the first cycle, all the initial starting values have already been replaced by the imputed values.

Then, it is necessary to run a number of cycles for the algorithm to *converge*. When this happens, the current values at that time originate the first imputed dataset. Such procedure is repeated for a desired number of times, with the imputed values stochastically independent from the first imputation. At the end, the desired number of imputed datasets is created from the algorithm.

Indeed, such process is called "full conditional specification" because each covariate is imputed from its full conditional distribution on all the other variables.

The three above-mentioned imputation methodologies may naturally be applied by researchers in statistical packages.

The sequential regression imputation can be programmed in any one of the main statistical software. Moreover, such technique is available in *SAS PROC MI* (V9 onwards). It is relevant to point out that, since this method assumes that the missingness pattern is monotone, the software at the beginning checks for such assumptions and does not run if it is violated.

As for the joint modelling approach, the earliest commonly used software implementing is considered to be the so-called Schafer's *NORM* package (Schafer, 1997). Such package, which has been ported to *R* and *S-plus*, is also considered to be the inspiration for multivariate normal imputation available in *SAS PROC MI* and similar algorithms in *Stata*. In the end, the joint multivariate normal model may be applied also in Windows *REALCOM-impute* (Carpenter et al., 2011).

Full conditional specification methodology is instead implemented by a *SAS* macro, *IVEware*, as well as in *R* with the two packages *mice* and *mi*.

4.2 Binary and ordinal data

The second typology of cross-sectional data for which it is possible to apply the multiple imputation procedure consists of binary and ordinal data.

Considering this type of data, based on the different assumptions made on the characteristics of the missing values, it is possible to implement a number of approaches.

Assuming that the outcome variable of interest Y_i is fully observed, that the missing data affecting the dataset behave as MAR (with, as usual, MCAR as special case) with missing values are characterized by a monotone missingness pattern, and that the dataset of interest is made by a mix of binary and continuous variables, the procedure is the following.

First, it is necessary to put the variables in order to make a monotone missingness pattern, with first the fully observed covariates. Then, it is necessary to impute each partially observed variable in turn, conditional on previous covariates.

The whole procedure needs to be repeated with the aim to generate successive imputed datasets.

However, when missing data mechanism is characterized by a nonmonotone missing pattern, Carpenter and Kenward (2013) propose an approach consisting of treating binary, binomial and ordinal variables as continuous for the imputation purpose, and then in the imputed data to round their imputed values to the nearest valid discrete value before continuing to fit the model.

Assuming those variables not to be affected by missing values, handle them as if they were continuous in the framework of a multivariate normal imputation yields that the distribution of the other covariates reveals to be conditioned on a linear function of them.

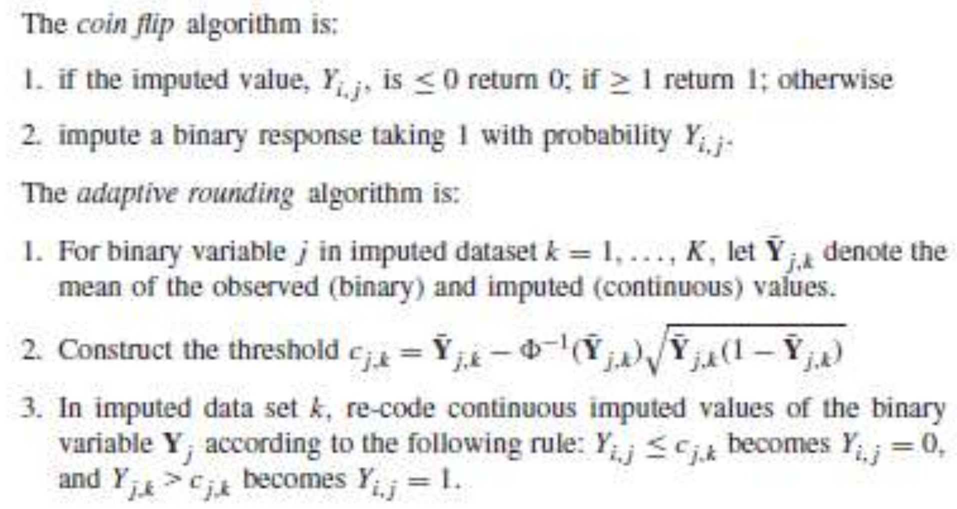
Instead, in the case in which fully observed binary variables are formally modeled, in most applications the results are likely to be almost indistinguishable (Carpenter and Kenward, 2013).

Specifically, Bernaards et al. (2007), considering the framework of binary data, propose and compare three methodologies to implement the above-described procedure:

- *Simple rounding*, consisting of simply round to the nearest of o or 1;
- *Coin flip*;
- *Adaptive rounding*.

Carpenter and Kenward (2013) provide a detailed explanation for the functioning of the *coin flip* algorithm and the *adaptive rounding* algorithm (Figure 4.2):

Coin flip algorithm and adaptive rounding algorithm.



The *coin flip* algorithm is:

1. if the imputed value, $Y_{i,j}$, is ≤ 0 return 0; if ≥ 1 return 1; otherwise
2. impute a binary response taking 1 with probability $Y_{i,j}$.

The *adaptive rounding* algorithm is:

1. For binary variable j in imputed dataset $k = 1, \ldots, K$, let $\bar{\mathbf{Y}}_{j,k}$ denote the mean of the observed (binary) and imputed (continuous) values.
2. Construct the threshold $c_{j,k} = \bar{\mathbf{Y}}_{j,k} - \Phi^{-1}(\bar{\mathbf{Y}}_{j,k})\sqrt{\bar{\mathbf{Y}}_{j,k}(1 - \bar{\mathbf{Y}}_{j,k})}$
3. In imputed data set k, re-code continuous imputed values of the binary variable $\mathbf{Y}_j$ according to the following rule: $Y_{i,j} \leq c_{j,k}$ becomes $Y_{i,j} = 0$, and $Y_{j,k} > c_{j,k}$ becomes $Y_{i,j} = 1$.

Figure 4.2. Source: Carpenter and Kenward, 2013, "Multiple Imputation and Its Application"

Therefore, it is possible to state that adaptive rounding is quite similar to simple rounding: the main difference consists of the application of the above-reported threshold. Indeed, for values closer to 0 or 1, the imputed binary variables will be characterized by a higher variability.

Horton et al. (2003) anticipate bias in parameter estimates in the case in which simple rounding is applied.

In this perspective, Bernaards et al. (2007) compare all three proposals in simulation studies and find that coin flipping performs worst, with adaptive rounding having a slight edge over simple rounding. Moreover, the adaptive rounding methodology reveals to perform satisfactorily in applications when the underlying probability is between 0.1 and 0.9.

A third relevant approach to handle binary and ordinal missing data consists of the so-called *general location model*. Such methodology to a joint imputation model for continuous and categorical data, described by Schafer (1997), makes use of the general location model provided by Olkin and Tate (1961).

Such model consists of first separating the data into continuous and categorical variables. Then, for each cell of the contingency table which has been defined by the categorical variable, it is necessary to fit a separate multivariate normal model to the continuous variables.

Another commonly used method to deal with binary and ordinal missing data consists of the *full conditional specification* approach.

First, it is necessary to first order the variables of interest to obtain a missingness pattern as close as possible to be monotone. Then, it is needed to fill in the missing values of each variable. This is typically done by drawing with replacement from the observed values of the covariate of interest. In the following step, the algorithm works as follows:

3. By the means of a logistic regression for the binary variables and of a linear regression for the continuous ones, it implements a regression of *the observed part* of the variable of interest on all remaining ones, which missing values are set at their current imputed values;
4. It uses the appropriate regression imputation methodology (depending on the nature of the data) to impute the missing values of interest.

Referring to such two steps, it is necessary to cycle through them until the algorithm looks lime to have converged to the stationary distribution. In the end, the current imputed values are kept to make the first imputed dataset. Such procedure is repeated several times, drawing each subsequent completed dataset.

Referring to real applications, it is necessary to mention the potential for explicit or implicit over-fitting of models with a number of correlated binary variables.

In this perspective, it is relevant to understand how to apply the above-mentioned methodologies in statistical packages.

In the situation of sequential imputation of missing data characterized by a monotone missingness pattern and having a mix of binary and continuous variables, it is possible to use the linear and logistic model fitting software available in most statistical packages.

In such framework, it is possible to avoid overfitting by simply checking that the results characterizing each regression model are sensible. It is important to point out that this check has to be done before starting the process of imputation.

In the case of the joint multivariate normal approach, some statistical packages include automatic rounding; however, if a researcher wants to use the adaptive rounding, she/he will have to write an own post-imputation data step. An advantage of using the joint multivariate normal approach consists of the fact that it reveals to be more robust to perfect prediction errors. By the way, some issues may arise in the case in which the variables reveal to be highly corelated. It is possible to successfully address this problem by using a ridge parameter.

Considering the full conditional specification algorithm, it is possible to use various statistical packages. Moreover, in *Stata* it is possible to implement automatic detection and adjusting for perfect prediction, even if it is relevant to point out that detection or perfect prediction is not guaranteed.

4.3 Unordered categorical data

Regarding the application of the multiple imputation procedure to unordered categorical data, in last decades the literature has provided many different possible approaches.

Assuming, as in the two previous frameworks, to deal with missing data characterized by a monotone missingness pattern, it is possible to apply the above-described sequential imputation with a unique but relevant difference in the methodology: the logistic regression (which is used in the case of binary and ordinal data) is replaced with a multinomial logistic regression.

After having implemented such above-mentioned crucial substitution, the procedure for this approach reveals to be exactly the same as described in the previous section.

Just like in the framework of binary and ordinal data, a second relevant multiple imputation methodology which can be used consists of the joint multivariate normal model. The main feature of applying multivariate normal imputation to categorical data is that, supposing to have a categorical variable characterized by M levels, it is necessary to generate $M - 1$ dummy variables indexing the categories of interest.

Even if this approach has not been widely explored in literature, Carpenter and Kenward (2013) suggest that it is likely to perform in an acceptable way if applied in several practical settings.

Another relevant methodology to deal with unordered categorical missing data is, as for binary and ordinal data, the general location model (Olkin and Tate, 1961).

However, with reference to this approach it is important to point out that the general location model is characterized by a saturated log-linear model the categorical variables and that, usually, both categorical and multivariate normal models need to be quite simplified before it is fitted.

In the end, it is also possible to apply to such typology of missing data the above-described full conditional specification approach, but with some points of attention.

First, when dealing with a M-level categorical variables which are included as predictors in the regression models constituting the full conditional specification, such covariates need to be included in the methodology as M-level categorical variables (that is, using $M - 1$ dummy indicators).

Moreover, in the case in which the missingness pattern is actually monotone, an appropriately specified full conditional specification leads to have imputed data characterized by the same distribution of a hypothetical sequential regression imputation, once the former has converged.

Regarding the application of the multiple imputation procedure to unordered categorical data, various statistical packages are available for researchers.

Schafer’s *CAT* package uses a joint log-linear model, which is extended by the so-called *MIX* package to a mix of categorical and continuous data by using the general location model (Schafer, 1999). Such packages have been ported to the software *R*.

In the end, the full conditional specification algorithm is available both in *Stata* and *R*.

Conclusions

As shown in chapter I, missing data (of any of the three main typologies presented) actually represent a major issue often affecting inferential procedures in many research fields.

Taking this into account, in the last decades various techniques have been developed to handle missing data and, consequently, to obtain unbiased inference results.

As for single imputation techniques, it has been shown that there exists a variety of different methodologies, based on different assumptions and statistical methods. Such approaches, even if still characterized by a number of not negligible drawbacks, under specific assumptions and situations are able to lead a researcher to obtain unbiased results.

Multiple imputation procedure, on the other hand, can be considered as an improvement with respect to the single imputation techniques. Indeed, such methodology not only is able to lead to unbiased results but, in some circumstances, it allows to get a reduction in bias and/or a gain in efficiency. Moreover, since multiple imputation can be applied to different types of cross-sectional data, it substantially represents a step forward regarding the possible fields of application.

Even if it consists of a more accurate and precise methodology, it is relevant to point out that multiple imputation has not to be considered a miracle cure and that the door is still widely open for the implementation of new approaches and methodologies, also considering the continuous evolution of statistical software and packages that a researcher has available when conducting a study.

Bibliography

Aitchison, J. and Bennett, J. A. (1970) Polychotomous quantal response by maximum indicant. Biometrika, 57, 253–262.

Allison, Paul D. "Multiple imputation for missing data: A cautionary tale." Sociological methods & research 28.3 (2000): 301-309.

Andridge, Rebecca R., and Roderick JA Little. "A review of hot deck imputation for survey non-response." International statistical review 78.1 (2010): 40-64.

Barzi, Federica, and Mark Woodward. "Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies." American journal of epidemiology 160.1 (2004): 34-45.

Batista, Gustavo EAPA, and Maria Carolina Monard. "A study of K-nearest neighbour as an imputation method." His 87.251-260 (2002): 48.

Bernaards, C. A., Belin, T. R. and Schafer, J. L. (2007) Robustness of a multivariate ormal approximation for imputation of incomplete binary data. Statistics in Medicine, 26, 1368–1382.

Brunet, Jean-Philippe, et al. "Metagenes and molecular pattern discovery using matrix factorization." Proceedings of the national academy of sciences 101.12 (2004): 4164-4169.

Carmona-Saez, Pedro, et al. "Biclustering of gene expression data by non-smooth non-negative matrix factorization." BMC bioinformatics 7 (2006): 1-18.

Carpenter, James, and Michael Kenward. Multiple imputation and its application. John Wiley & Sons, 2012.

Carpenter, J. R., Goldstein, H. and Kenward, M. G. (2011a) REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. Journal of Statistical Software, 45(4), 1–14.

Chhabra, Geeta, Vasudha Vashisht, and Jayanthi Ranjan. "A comparison of multiple imputation methods for data with missing values." Indian Journal of Science and Technology 10.19 (2017): 1-7.

Ciaccia, Paolo, Marco Patella, and Pavel Zezula. "M-tree: An efficient access method for similarity search in metric spaces." Vldb. Vol. 97. 1997.

Croxford, L., Ianelli, C. and Shapira, M. (2007) Documentation of the Youth Cohort Time-Series Datasets, UK Data Archive Study Number 5765, Economic and Social Data Service.

Ding, Chris HQ, Tao Li, and Michael I. Jordan. "Convex and semi-nonnegative matrix factorizations." IEEE transactions on pattern analysis and machine intelligence 32.1 (2008): 45-55.

Donders, A. Rogier T., et al. "A gentle introduction to imputation of missing values." Journal of clinical epidemiology 59.10 (2006): 1087-1091.

Enders, Craig K. Applied missing data analysis. Guilford Publications, 2022.

Finch, W. Holmes. "Imputation methods for missing categorical questionnaire data: A comparison of approaches." Journal of Data Science 8.3 (2010): 361-378.

Galati, J. C., et al. "Rounding non-binary categorical variables following multivariate normal imputation: evaluation of simple methods and implications for practice." Journal of Statistical Computation and Simulation 84.4 (2014): 798-811.

Gottardo, R. "EMV: Estimation of missing values for a data matrix." R package version 1.1 (2006).

Graham, John W. Missing data: Analysis and design. Springer Science & Business Media, 2012.

Harrell, Frank E., and Charles Dupont. "Hmisc: harrell miscellaneous." R package version 3.2 (2008): 437.

Hastie, Trevor, et al. "Imputing missing data for gene expression arrays." (1999).

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. "Statistical learning with sparsity." Monographs on statistics and applied probability 143 (2015): 143.

He, Yulei. "Missing data analysis using multiple imputation: getting to the heart of the matter." Circulation: Cardiovascular Quality and Outcomes 3.1 (2010): 98-105.

Honaker, James, Gary King, and Matthew Blackwell. "Amelia II: A program for missing data." Journal of statistical software 45 (2011): 1-47.

Horton, Nicholas J., and Ken P. Kleinman. "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." The American Statistician 61.1 (2007): 79-90.

Horton, N. J., Lipsitz, S. R. and Parzen, M. (2003) A potential for bias when rounding in multiple imputation. The American Statistician, 57, 229–232.

Huisman, Mark. "Imputation of missing item responses: Some simple techniques." Quality and Quantity 34 (2000): 331-351.

Huskamp, Haiden A., et al. "Discussions with physicians about hospice among patients with metastatic lung cancer." Archives of Internal Medicine 169.10 (2009): 954-962.

Hutcheson, Graeme. "Missing data: Data replacement and imputation." Journal of Modelling in Management 7.2 (2012).

Hutcheson, Graeme D., Maria Pampaka, and Julian Williams. "Enrolment, achievement and retention on 'traditional'and 'Use of Mathematics' pre-university courses." Research in Mathematics Education 13.2 (2011): 147-168.

Ibrahim, Joseph G., et al. "Missing-data methods for generalized linear models: A comparative review." Journal of the American Statistical Association 100.469 (2005): 332-346.

Karahalios, Amalia, et al. "The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study." Emerging themes in epidemiology 10.1 (2013): 1-11.

Kasza, Jessica, and Rory Wolfe. "Interpretation of commonly used statistical regression models." Respirology 19.1 (2014): 14-21.

Kim, Hyunsoo, and Haesun Park. "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis." Bioinformatics 23.12 (2007): 1495-1502.

King, Gary, et al. "Analyzing incomplete political science data: An alternative algorithm for multiple imputation." American political science review 95.1 (2001): 49-69.

Laaksonen, Seppo Sakari. "Multiple imputation for a continuous variable." International Journal of Mathematical and Statistical Sciences (2016).

Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." Nature 401.6755 (1999): 788-791.

Lee, E.-K., D. Yoon, and T. Park. 2009. "ArrayImpute: Missing Imputation for Microarray Data."

Lee, Katherine J., and John B. Carlin. "Recovery of information from multiple imputation: a simulation study." Emerging themes in epidemiology 9 (2012): 1-10.

Lee, Katherine J., and Julie A. Simpson. "Introduction to multiple imputation for dealing with missing data." Respirology 19.2 (2014): 162-167.

Lee, Katherine J., et al. "Comparison of methods for imputing ordinal data using multivariate normal imputation: a case study of non-linear effects in a large cohort study." Statistics in medicine 31.30 (2012): 4164-4174.

Li, Yifeng, and Alioune Ngom. "A new kernel non-negative matrix factorization and its application in microarray data analysis." 2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE, 2012.

Li, Yifeng, and Alioune Ngom. "The non-negative matrix factorization toolbox for biological data mining." Source code for biology and medicine 8.1 (2013): 1-15.

Lichman M. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science; 2013. PMid: 24373753.

Lin, Xihui, and Paul C. Boutros. "Optimization and expansion of non-negative matrix factorization." BMC bioinformatics 21.1 (2020): 1-10.

Little, Roderick JA. "Regression with missing X's: a review." Journal of the American statistical association 87.420 (1992): 1227-1237.

Malarvizhi, R., and Antony Selvadoss Thanamani. "K-nearest neighbor in missing data imputation." Int. J. Eng. Res. Dev 5.1 (2012): 5-7.

Marshall, Andrea, et al. "Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study." BMC medical research methodology 10.1 (2010): 1-16.

Narayan, Nitin. "A nearest neighbor based cold-deck imputation for X-ray tube wear estimation."

National Center for Education Statistics. Tech rep. U.S. Department of Education; 2002. NCES statistical standards.

Nielsen, Søren Feodor. "Proper and improper multiple imputation." International Statistical Review 71.3 (2003): 593-607.

Olkin, I. and Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. Annals of Mathematical Statistics, 32, 448–465.

Owen, Art B., and Patrick O. Perry. "Bi-cross-validation of the SVD and the nonnegative matrix factorization." (2009): 564-594.

Owen, Matt, et al. "Zelig v4. 0-10 Core Model Reference Manual." Letzter Abruf am 8 (2013): 2015.

Pampaka, Maria, Graeme Hutcheson, and Julian Williams. "Handling missing data: analysis of a challenging data set using multiple imputation." International Journal of Research & Method in Education 39.1 (2016): 19-37.

Perez, Adriana, et al. "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia." Statistics in medicine 21.24 (2002): 3885-3896.

Pigott, Therese D. "A review of methods for missing data." Educational research and evaluation 7.4 (2001): 353-383.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, 85–95.

Royston, Patrick, John B. Carlin, and Ian R. White. "Multiple imputation of missing values: new features for mim." The Stata Journal 9.2 (2009): 252-264.

Rubin, Donald B. "Inference and missing data." Biometrika 63.3 (1976): 581-592.

Rubin, Donald B. "Multiple imputation after 18+ years." Journal of the American statistical Association 91.434 (1996): 473-489.

Rubin, Donald B. "Multiple imputation for survey nonresponse." (1987): 9780470316696.

SAS Institute Inc. PROC MI.SAS Procedures Guide,Version 92. SAS
Institute Inc, Cary, NC, 2008.

Schafer, Joseph L. Analysis of incomplete multivariate data. CRC press, 1997.

Schafer, Joseph L. "Multiple imputation: a primer." Statistical methods in medical research 8.1 (1999): 3-15.

Schafer, Joseph L., and John W. Graham. "Missing data: our view of the state of the art." Psychological methods 7.2 (2002): 147.

Shrive, Fiona M., et al. "Dealing with missing data in a multi-question depression scale: a comparison of imputation methods." BMC medical research methodology 6 (2006): 1-10.

Somasundaram, R. S., and R. Nedunchezhian. "Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values." International Journal of Computer Applications 21.10 (2011): 14-19.

Srebotnjak, Tanja, et al. "A global Water Quality Index and hot-deck imputation of missing data." Ecological indicators 17 (2012): 108-119.

StataCorp. Stata Statistical Software: Release 12. StataCorp LP, College Station, TX, 2011.

Statistical Analysis System Version 8.1. Cary, NC, SAS Instititute; 2000.

Su, Yu-Sung, et al. "Multiple imputation with diagnostics (mi) in R: Opening windows into the black box." Journal of Statistical Software 45 (2011): 1-31.

Suthar, Bhavisha, Hemant Patel, and Ankur Goswami. "A survey: classification of imputation methods in data mining." International Journal of Emerging Technology and Advanced Engineering 2.1 (2012): 309-12.

Tang, Lingqi, et al. "A comparison of imputation methods in a longitudinal randomized clinical trial." Statistics in medicine 24.14 (2005): 2111-2128.

Templ, Matthias, Alexander Kowarik, and Peter Filzmoser. "Iterative stepwise regression imputation using standard and robust methods." Computational Statistics & Data Analysis 55.10 (2011): 2793-2806.

Twisk, Jos, and Wieke de Vente. "Attrition in longitudinal studies: How to deal with missing data." Journal of clinical epidemiology 55.4 (2002): 329-337.

van Buuren, S. (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16, 219–242.

van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999) Multiple imputation of missing blood presure covariates in survival analysis. Statistics in Medicine, 18, 681–694.

Van Buuren, Stef, and Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." Journal of statistical software 45 (2011): 1-67.

von Hippel, Paul T. "Should a normal imputation model be modified to impute skewed variables?." Sociological Methods & Research 42.1 (2013): 105-138.

Wang, Guoli, Andrew V. Kossenkov, and Michael F. Ochs. "LS-NMF: a modified non-negative matrix factorization algorithm utilizing uncertainty estimates." BMC bioinformatics 7.1 (2006): 1-10.

Wayman, Jeffrey C. "Multiple imputation for missing data: What is it and how can I use it." Annual Meeting of the American Educational Research Association, Chicago, IL. Vol. 2. 2003.

Wilson, D. Randall, and Tony R. Martinez. "Reduction techniques for instance-based learning algorithms." Machine learning 38.3 (2000): 257-286.

Yuan, Yang C. "Multiple imputation for missing data: Concepts and new development." Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference. Vol. 267. No. 11. 2000.