Durham University

Research Methods Centre

MATH43515: Multilevel Modelling

Lecture 9: Missing Data Methods

Module Convenor / Tutor:

Andy Golightly

Outline

1. Introduction
2. Missingness Mechanisms
3. Handling the missing data
   - Deletion
   - Imputation Methods
4. Additional resources

Introduction

Missing data handling is a continuously evolving and debated research topic.

Dealing with the loss of information is an open problem and subsequently the literature on it is vast and will keep expanding.

Because of the abundance of methods, it is impossible to comprehensively cover all of them in a single presentation.

This lecture: to showcase the most important aspects of missing data and some popular methods for handling them.

Missing value

Missing values/data defined as the data value that is not stored for a variable in the observation of interest. They are often encoded as NA, NaN, blank or any other place holders.

Missing values can occur in data for various reasons, such as survey non-responses or error in the data entry etc.

Missing data causes various problems:

- The absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false
- It can cause bias in the estimation of parameters
- Reduce the representativeness of the samples
- It may complicate the analysis of the study

…..Can lead to invalid conclusions

Types of missing data

Rubin (1996), first described and divided the types of missing data

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Not Missing At Random (MNAR)

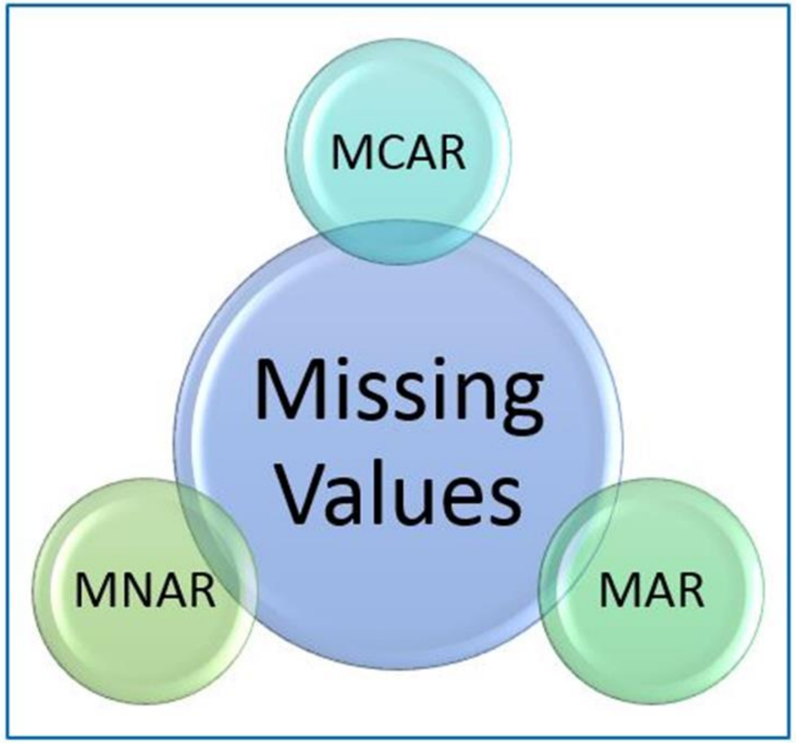


Figure 1 - Different Types of Missing Values in Datasets

Types of missing data

Handling of missing data largely depends on the likely reasons for the missingness observed. While there are many reasons why missing data may arise, missingness can be broadly classified into three categories:

Missing completely at random (MCAR): The missingness mechanism is not related to any element in our analysis.

- ➢ Complete case analysis (CCA) implies loss of efficiency but no bias.

Missing at random (MAR): The missingness mechanism is related to our dependent/independent variable(s) however it can be accounted for by the observed covariates.

- ➢ CCA implies loss of efficiency and bias (but can be corrected using MI).

Missing not at random (MNAR): The missingness mechanism is related to the dependent/independent variable(s) and this association cannot be accounted for by observed covariates.

CCA implies loss of efficiency and bias (can be corrected in some cases but non-trivial)

Missing mechanism

Type of missing value	Description	Examples	Acceptable solutions
Missing completely at random	All observations have the same likelihood of being missing.	• Electronic time observations are missing, independent of what lane a swimmer is in. • A scale is equally likely to produce missing values when placed on a soft surface or a hard surface (Van Buren, 2018). • Geographical location data is equally likely to be missing for all locations.	Deletion, Imputation
Missing at random	Likelihood that a data point is missing is not related to the missing data but may be related to other observed data.	• A certain swimming lane is more likely to have missing electronic time observations. • A scale produces more missing values when placed on a soft surface than a hard surface (Van Buren, 2018). • Childhood health assessment data is more likely to be missing in lower median income counties.	Deletion, Imputation
Missing not at random	Likelihood of a missing observation is related to its values.	• When surveyed people with more income are less likely to report their income. • On a health survey illicit drug users are less likely to respond to a question about illicit drug use. • Individuals surveyed about their age are more likely to leave the age question blank when they are older.	Imputation

Missing value treatment - techniques

- Do nothing
- Deletion
    - List wise | Pairwise | Deleting the variable
- Imputation
    - Averaging techniques
        - Mean/Median/Mode or zero constant values
    - Predictive techniques
        - Linear regression
        - k- Nearest Neighbours (k-NN)
        - ….

Handling missing data – Deletion

The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully.

When data is MCAR and MAR deletion may be a suitable method for dealing with missing values. However, when data is MNAR, deletion of missing observations can lead to bias.

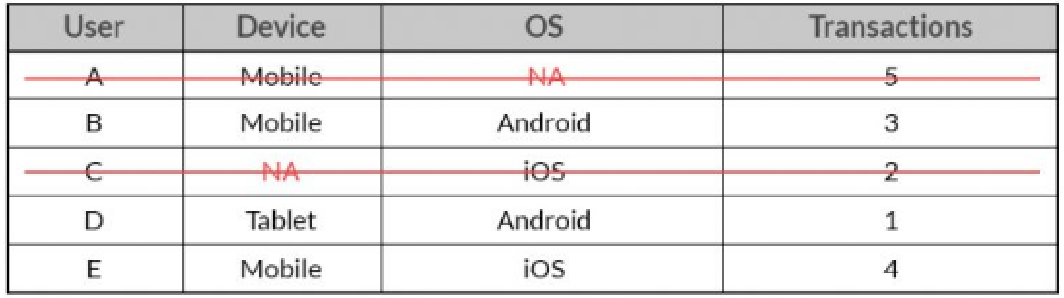
There are three methods of data deletion for missing values:

- Listwise deletion
- Pairwise deletion
- Variable deletion

Handling missing data – Deletion > Listwise/case deletion

Simply omit those cases with the missing data and analyse the remaining data. Here, rows containing missing variables are deleted.

User A and User C will be ignored for listwise deletion



User	Device	OS	Transactions
~~A~~	~~Mobile~~	~~NA~~	~~5~~
B	Mobile	Android	3
~~C~~	~~NA~~	~~iOS~~	~~2~~
D	Tablet	Android	1
E	Mobile	iOS	4

If there is a large enough sample, where power is not an issue, and the assumption of MCAR is satisfied, the listwise deletion may be a reasonable strategy.

However, when there is not a large sample, or the assumption of MCAR is not satisfied, the listwise deletion is not the optimal strategy.

Handling missing data – Deletion > Pairwise

- This method calculates means and (co)variances based on all observed data.
- For example, the mean of a variable X1 is based on all observed data on that variable.
- For the covariance between X1 and X2, all data are used for which both X1 and X2 have non-missing scores.
- Subsequently, the resulting summaries are used in the desired multivariate analysis, for example, the *lavaan* package in R allows this feature.

- The method is simple and, under MCAR, produces consistent estimates of quantities of interest.
- However, estimates are biased if the data are not MCAR.
- Further, the resulting covariance matrix might not be positive definite, which is a requirement for most multivariate analyses.

Handling missing data – Deletion > Variable

If a particular variable is having more missing values (> 50%) than the rest of the variables in the dataset, then removing that variable could be a reasonable strategy unless it is a really important predictor that makes a lot of practical sense.

Of course, this scenario may suggest a problem with the design of the data collection mechanism in the first place…

Imputation

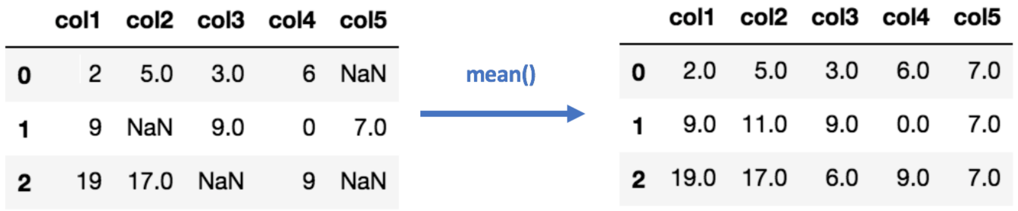
- The options described above are *ad-hoc* and not without issues.
- Instead of deletion, data scientists have multiple solutions to impute the value of missing data. Depending on why the data are missing, imputation methods can deliver reasonably reliable results.
- Imputing data replaces missing values with statistically determined values. Methods of imputation can vary from simply replacing missing values with the mean to sophisticated multiple imputation processes.
- Which method of imputation should be used depends on the characteristics of the data.
- Imputation is a simulation-based statistical technique for handling missing data. The goal is to replace missing values with imputed values so as to allow statistical inference with minimum loss of efficiency.

Imputation > Averaging techniques > mean/median/mode values

This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others.

It can only be used with numeric data.

All missing values are replaced with the variable mean, median or mode.



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Imputation > Averaging techniques > mean/median/mode values

Advantages:

- Easy and fast.
- Preserves the mean of the data

Disadvantages:

- Distorts the distribution of the data.
- Will give poor results on encoded categorical features (do NOT use it on categorical features).
- Reduces data variance.
- Not very accurate.
- Doesn’t account for the uncertainty in the imputations.
- Results in biased estimates.

***Zero or Constant imputation** — as the name suggests — it replaces the missing values with either zero or any constant value you specify

Imputing time series data | Carrying the last observation forward (LOCF)

Not all imputation methods are appropriate for time series data.

Time series data may contain time trends or seasonality, all of which should be addressed when imputing missing data.

Carrying the last observation forward (LOCF) or carrying the next observation backward (NOCB).

Use last observed data value as a replacement for missing data

Pros:

- Appropriate for time series data.
- Easy implementation.

Cons:

- Can results in biased estimates even when data is MCAR.
- Modelling techniques should address that data has been imputed by LOCF.
- May incorrectly suggest stability across time stretches if used to fill successions of missing data.

Imputation > Predictive > Linear regression

In regression imputation, the existing variables are used to make a prediction, and then the predicted value is substituted as if an actual obtained value.

Pros:

- Simple to implement.
- Uses all available information.
- Avoids drastically altering the shape of the distribution.

Cons:

- Underestimates variability.
- Falsely strengthens relationship between variables.

Imputation > Predictive > k-nearest neighbours algorithm (*k-NN*)

The k nearest neighbours is an algorithm that is used for simple classification or regression. The algorithm uses ‘feature similarity’ to predict the values of any new data points.

New point is assigned a value based on how closely it resembles nearby points. This can be very useful in making predictions about the missing values by finding the k closest neighbours to the observation with missing data and then imputing based on the non-missing values in the neighbourhood.

Usually uses inverse distance weighting for the regression task.

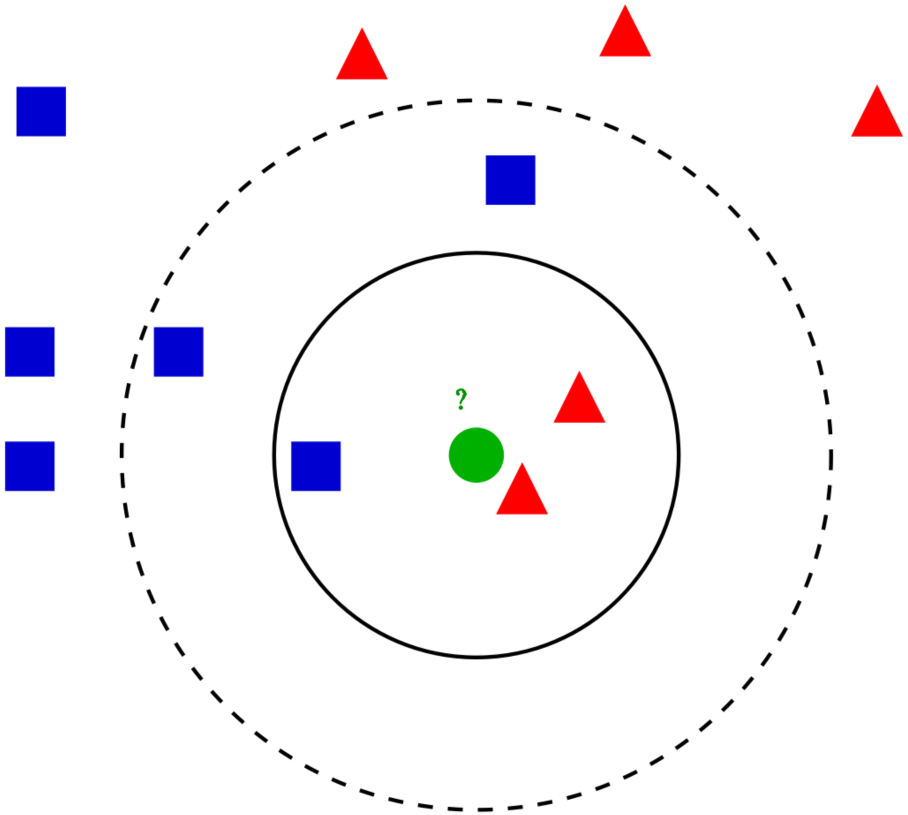
Imputation > Predictive > k-nearest neighbours algorithm (*k-NN*)

Example of k-*NN*.

The test sample (green dot) should be classified either to blue squares or to red triangles.

If k = 3 (solid line circle) it is assigned to the red triangles because there are 2 triangles and only 1 square inside the inner circle.

If k = 5 (dashed line circle) it is assigned to the blue squares (3 squares vs. 2 triangles inside the outer circle).



Source: Wikipedia

Imputation > Predictive > k-nearest neighbours algorithm (*k-NN*)

Advantages:

Can be much more accurate than the mean, median or most frequent imputation methods (It depends on the dataset).

Disadvantages:

- Computationally expensive. kNN works by storing the whole (training) dataset in memory.
- kNN is quite sensitive to outliers in the data.
- Choice of k? Could use cross validation (expensive). Trying to balance bias and variance.

Imputation > Predictive > Multiple Imputation by Chained Equation (MICE)

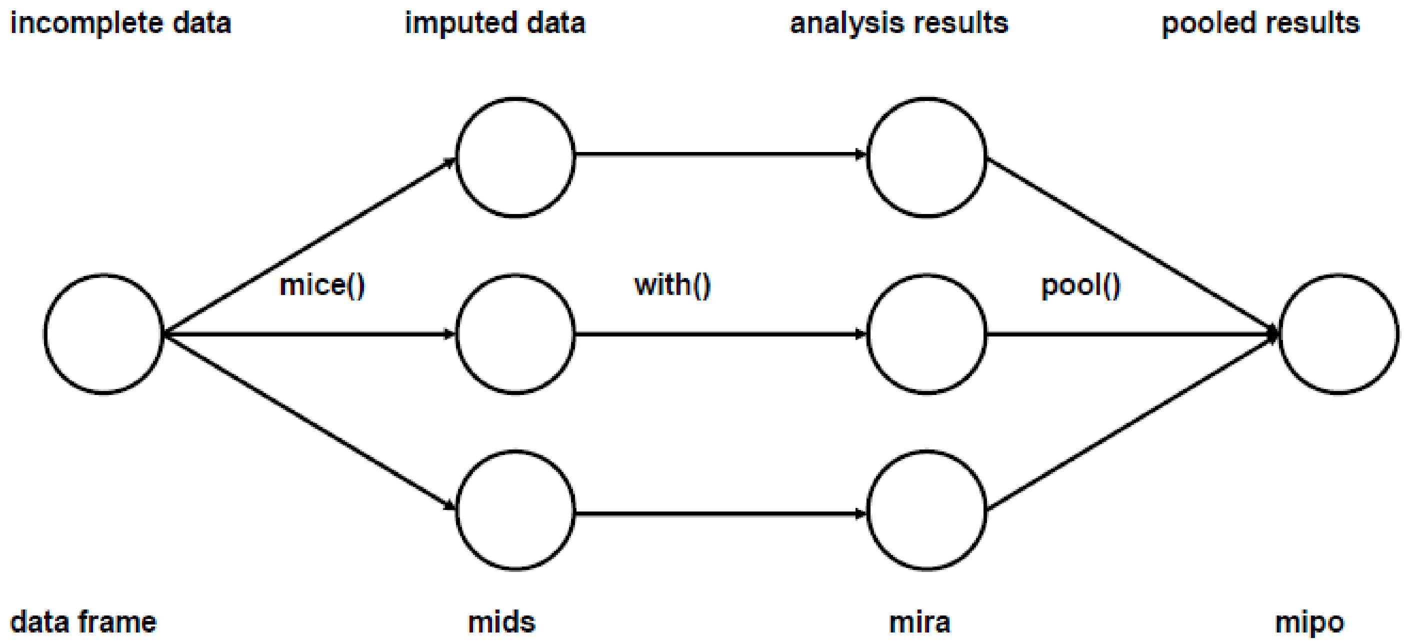


Figure 1: Main steps used in multiple imputation.

Imputation > Predictive > Multiple Imputation by Chained Equation (MICE)

- ✓ **Step 1:** A simple imputation, such as imputing the mean, is performed for every missing value in the dataset for all variables to be imputed (***V1…Vn***). These imputations are only temporary.

- ✓ **Step 2:** The temporary imputations for one variable, ***V1***, are set back to missing.

- ✓ **Step 3:** The observed values from the variable ***V1*** in Step 2 are regressed on the other variables in the imputation model.

- ✓ **Step 4:** The missing values for ***V1*** are then replaced with predictions (imputations(!)) from the regression model. Then ***V1*** is subsequently used as an independent variable in the regression for other variables with both the observed and imputed values used.

Imputation > Predictive > Multiple Imputation by Chained Equation (MICE)

- ✓ **Step 5:** Steps 2–4 are then repeated for each unimputed variable (***V2…Vn***) that has missing data. At the end of one iteration, all of the missing values have been replaced with predictions from regressions that reflect the relationships observed in the data.

- ✓ **Step 6:** Steps 2–5 are repeated for ***K*** number of iterations, with the imputations being updated at each iteration. At the end of ***K*** iterations an imputed dataset is generated.

- ✓ **Step 7:** Repeat step 6 ***M*** times and perform original model analysis for each of those ***M*** imputed datasets. Then use Rubin's rules (Rubin, 1987) to pool estimates from all ***M*** number of model results to obtain final estimates.

References and additional resources

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software.

https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Imputation diagnostics:

- Abayomi et al. (2008)
- He et al. (2010)
- Stuart et al. (2009).

Imputation for Multilevel models:

- Goldstein et al. (2009)
- Huque et al. (2018)

Imputation for Non-normal data:

- Liu (1995)
- He and Raghunathan (2006)
- Demirtas and Hedeker (2008)

Great all-in-one (free) book on Imputation:

- Stef Van Buuren (2018)

Thanks for your attention!

Durham University
Research Methods Centre