University of Groningen

Introduction to Data Science

Missing values

Group 16:
Otte Tjepkema *(s3237184)*
José Rodrigues *(s4169328)*
Andrei Miculita *(s1234567)*
Robert Riesebos *(s3220672)*

September 23, 2019

Contents

1 Introduction

Missing values are a common occurrence when dealing with cleaning and analysing data. There are multiple types of missing data, which will be laid out in this report. These types, along with the quantity of missing data, determine how to deal with the absence of certain values. In some cases missing values can be ignored, but more often than not we need to use imputation methods to substitute missing values. These methods are discussed in this report and illustrated with the help of an example dataset.

2 Types of missing values

When we want to impute missing values inside a dataset it is important to know the nature of the missing data. Generally the missing data is classified into three categories: missing completely at random, missing at random and missing not at random [1]. These different categories will be elaborated upon in the following subsections.

2.1 Missing completely at random (MCAR)

The first category of missing data, missing completely at random (MCAR), includes data where the probability of values being missing is the same for all cases, or to say it in a different way: the causes of the missing data are unrelated to the data itself. This is the easiest case, since we can ignore many of the complexities that arise from missing data, such as biasing. An example of this would be that survey results get lost or that the batteries on your measuring device run out. However while this data type is the most convenient, it is often unrealistic for the data at hand, since most of the time there will be a relation between features in the dataset.

2.2 Missing at random (MAR)

The second category is missing at random (MAR), where the probability of missing depends on the other variables that are known. For example, if we know that for a weighing scale the probability of missing values is higher when the scale is placed on a soft surface and we know all the surfaces that were used, then the data is missing at random. Generally missing at random is used as an assumption when working with missing data.

2.3 Missing not at random (MNAR)

The final category is missing not at random (MNAR), where the probability of missing depends on data that is unknown to us. A common example of this occurs in public opinion research, where people with weaker opinions respond less often. This type of missing data is also the most difficult to handle. Common approaches for this are to perform more research to find the cause of the missing values, or to perform what if analyses to see how sensitive the results are under various scenarios.

3 Methods to deal with missing values

3.1 Discarding missing data

The simplest way to deal with missing data is to simply discard data objects that contain missing values. This method has its drawbacks however since even partially specified data objects contain some information, and with a higher proportion of objects that have missing values it becomes increasingly harder to make a reliable analysis of the data. If the data only has a few objects that have missing values, discarding them can be a perfectly fine and straightforward solution [2].

3.2 Single imputation methods

3.2.1 Mean/Median imputation

In mean imputation all the missing values of a feature are replaced with the mean/median value of the values that are available for that feature. The advantage of this method is that it is very simple to understand and it is computationally inexpensive. Also this method can easily be applied to datasets where missing values exist in multiple features. In this case the mean imputation is applied separately for each feature. Mean/Median imputation cannot be used when there is a strong correlation between features, since it does not take this into account. Besides this mean imputation will also underestimate the variance and almost any estimate other than the mean. Therefore mean imputation should only be used for exploratory data analysis or when a small number of variables are missing and should be avoided in general. Other techniques similar to this are mode or most frequent value imputation. An example showing mean, median and most frequent value imputation is shown in section 3.6.

3.2.2 Regression imputation

In single regression imputation the imputed value is predicted from a regression model. The first step in this process is to create a model using the complete feature vectors in the dataset. A prediction for incomplete cases can then be made by inputting the information that is available for that class and using the output of the model to replace the missing value. A simple case for regression is demonstrated in section 3.6.

If the data are MCAR, regression imputation will yield unbiased results of the mean and the regression coefficients. Even if the data are

MAR, regression imputation will result in correct regression coefficients if the variables that affect the missingness are included in the regression model. The disadvantages of this method are that regression imputation artificially strengthens the relations between variables in the dataset. Also the variability of the data is underestimated.

3.3 LOCF and BOCF

LOCF and BOCF stand for last observation carried forward and baseline observation carried forward respectively. This is a technique that can be used in longitudinal data, which is data obtained from repeated observations of the same variables, for example blood tests of people who have regular checkups. In this method the value which was last observed is used to replace the missing value. An example of this method is given in the figure below. Here we see that if the data is erratic this method yields implausible results since the real values are likely to vary. This method can only be applied with confidence if the data that we are trying to fill in has a stable pattern, for example the address of a patient. The disadvantages of this method is that it can either lead to conservative or anticonservative results depending on actions that were undertaken between observations of the data. Also it has been shown that LOCF can result in biased estimates even for MCAR data

3.3.1 Hot deck imputation

Hot deck imputation replaces missing values in a data vector with those of a data vector that is most "similar" in its features. This similarity can be defined in different ways. When distance is used the technique is called nearest neighbor approach. This method is used when the auxiliary values are numerical in nature. Since we are taking the distance it is important that the auxiliary values that are used are normalized or properly weighted. It is also possible to take values from K nearest neighbors and average their values, this method is called K nearest neighbors (Knn).

Another approach is the random hot deck imputation. In this method imputation classes are formed based on matching auxiliary variables between data vectors. Then a random data vector is selected from these imputation classes, which is then used to fill in the missing values. Since this method can only effectively be used when all of the values are categorical in nature, this method is most often used in survey research.

This method preserves the the variable distribution, however the standard errors and variability in the data is underestimated [3].

3.4 Multiple imputation methods

Most disadvantages of single imputation methods can be ameliorated through a method called multiple imputation. Multiple imputation consists of three steps:

1. Imputation

   Generate multiple datasets, by replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute.

2. Analysis

   The datasets are then analysed using the standard procedures for complete data.

3. Pooling

   Valid statistical inferences are obtained by combining results from different imputed data sets.

The biggest advantage of multiple imputation is its ability to measure uncertainty surrounding parameter estimates. This is due to the fact that it does not use a single imputed value, but multiple possible values.

Its only downside is the fact that it may end up being more computationally expensive. The 3 steps are iterative, and require parsing the dataset multiple times to work. A simple case for regression is demonstrated in section 3.6.

3.5 Maximum likelihood method

Generally multiple imputation methods are preferred over single imputation for its superior statistical properties. However there also exists another methods which yields equally good results which is called Maximum likelihood[4].

The main idea of Maximum likelihood implementation is the fitting of a statistical model to a dataset. We will first explain this concept using the assumption that no data is missing from the dataset as this will be easiest to understand. The concept is then easily generalised to datasets with missing values.

The fitting of distributions to datasets is done through the maximum likelihood function

$$L = \prod_{i=1}^{n} f_i(y_{i1}, y_{i2}, ..., y_{ik}; \theta) \tag{1}$$

where n is the number of independent observations, and k is the number of variables $(y_{i1}, y_{i2}, ..., y_{ik})$, f_i is the probability density function and θ is a set of parameters of the f_i that are to be estimated. Using this formula we can find the best fitting distribution by maximising L by trying different values of θ.

The probability function that is most commonly used is the normal distribution, therefore we will use this to demonstrate further concepts. For other distributions such as an exponential or multinomial the reader is redirected to [5], which presents an excellent overview. The likelihood function for a normal distribution is equal to [6]

$$L = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-.5(y_i-\mu)^2}{\sigma^2}} \right) \tag{2}$$

where μ is the population mean and σ^2 is the variance. This formula can be interpreted as the probability of drawing a collection of k scores $(y_1, y_2, ..., y_n)$ from a normal distribution with mean μ and a variance σ^2. The part to the left of the exponent is normalisation parameter, which ensures that the area under the graph is always one, which is expected from a probability distribution. The part in the exponent is a distance term, which should be a small as possible for a good likelihood score.

One can imagine that as the number of samples increases the likelihood function takes very small values. This is difficult to work with and introduces rounding errors, therefore is is common to take the natural logarithm of the log-likelihood function. It also has the additional benefit of converting the product in function (2) to a sum, which is easier to work with. The resulting log-likelihood function is as follows

$$\log L = \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-.5(y_i-\mu)^2}{\sigma^2}} \right) \tag{3}$$

Since the non-log likelihood returns a value that is always between zero and one, the log-likelihood will return a negative value. The value that is least negative corresponds to the best fit of μ and σ^2

The question that remains is how to find the optimal values of μ and σ^2 such that the likelihood is maximised. A simple way to do this is differenti-

ation. The first step is to take the derivative of equation (3) with respect to μ or σ^2 and equate it to zero. Then we can solve for either of the variables, which then gives us the optimal value. This is the easiest method but not always viable, sometimes the derivative of a likelihood function can be too complicated to solve analytically. In this case iterative methods are used to find the best solution.

Now that we have introduced the theory of maximum likelihood with univariate examples it is time to extend the theory to multivariate data. Hopefully it will make sense that the multivariate methods use matrices and vectors as this is the natural language for problems that involve multiple parameters. We will first write down the likelihood function for a multivariate normal distribution for a single entry.

$$L_i = \frac{1}{(2\pi)^{k/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-.5(\boldsymbol{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu})} \tag{4}$$

where $\boldsymbol{Y}_i$ is the score vector which contains the values of k variables, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. The key element in this formula is the Mahalanobis distance $\left(\boldsymbol{Y}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{Y}_i - \boldsymbol{\mu}\right)$, which replaces the 1D distance value $(y_i - \mu)$ in the univariate case. Again we want to find the optimal values for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to find the maximum likelihood fit. Again we can also define a log likelihood function which is easier to work with.

$$L = \sum_{i=1}^{n} \frac{1}{(2\pi)^{k/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-.5(\boldsymbol{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu})} \tag{5}$$

Also similar to the univariate case we try to find these optimum values by solving the first derivatives, the only difference is that the equations are now contained in matrix form.

Finally we list the formula that is used for log-likelihood with missing data.

$$L = \sum_{i=1}^{n} \frac{1}{(2\pi)^{k_i/2} \left|\boldsymbol{\Sigma}_i\right|^{1/2}} e^{-.5(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)} \tag{6}$$

we see that this equation is the same as equation (5), with the exception that now all of the parameter matrices now also have a subscripts. The interpretation of this is that the size and contents of the matrices can vary between individuals, such that we can exclude any missing parameters.

The technique of maximum likelihood has some advantages over multiple imputation[4]. We will list the most important ones here:

For a given dataset, multiple imputation will give different results each time your run it, while the results of maximimum likelihood stay the same. This makes it easier for researchers to compare their results.

The implementation of multiple imputation requires more decisions than maximum likelihood. For multiple imputation the number of used datasets needs to be chosen, the number of iterations between datasets and how to incorporate non-linearities, just to name a few. With maximum likelihood the only thing that needs to be chosen is the model of interest, after this the method handels all the decision making. This makes multiple implementation much cleaner.

Finally with multiple implementation there is a change of conflict between the imputation model and the analysis model, this is not possible in maximum likelihood, since their is only one model. Examples of conflicting imputation and analysis models would be if the analysis model contains variables that were not included in the imputation model. Or if the imputation model is strictly linear, but the analysis model contains non-linearities. When the user is careful these problems can be avoided, however since it is easy to not be careful, this is regarded as a major problem for multiple implementation.

3.6 Case Study

In order to demonstrate each method's advantages and disadvantages and illustrate them in a more user-friendly way, a real dataset is used. This dataset displays 178 samples of wine bottles and classifies them based on 12 different features (alcohol percentage, malic acid, colour intensity...). After a random deletion of 35 numerical values related to the alcohol percentage, some imputation methods were tested and compared in order differentiate the accuracy of each one.

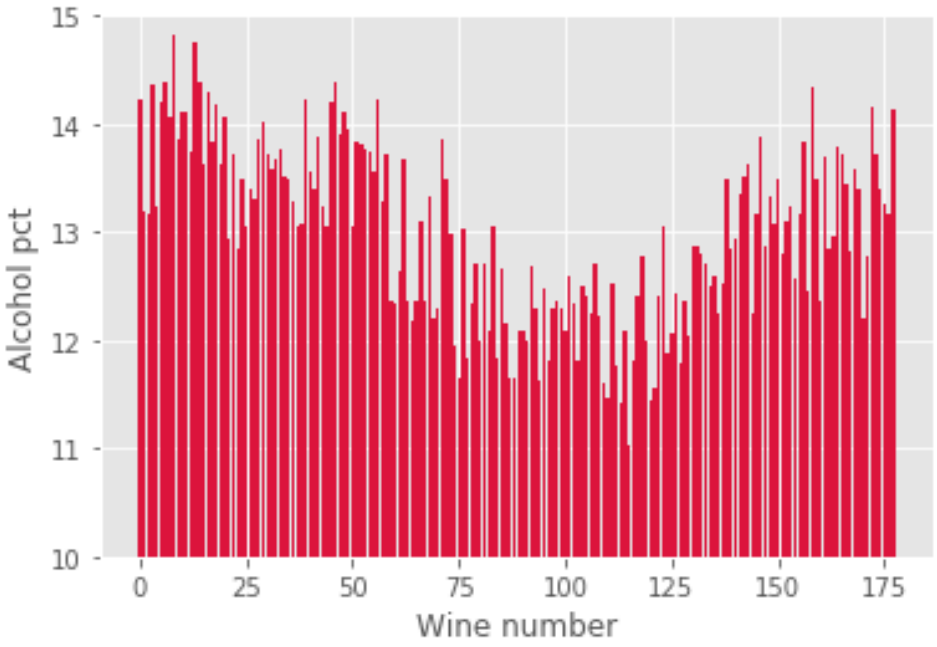

Figure 1: Real Data

In a primary approach, simple imputation methods (mean, median and most frequent values) were tested and the imputed datasets compared to the real data. Given the nature of this imputation methods, there's a clear underestimation of the variance in the resulting datasets (even though that's not necessarily true for the median and most frequent value imputation).

The three first examples of imputed datasets are illustrated bellow in bar charts.

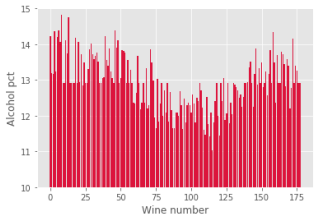

Figure 2: Mean

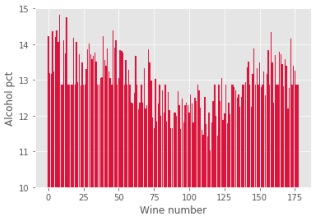

Figure 3: Median

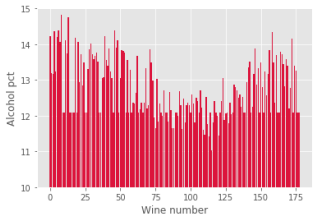

Figure 4: MFV

After the first tries with simple imputers, a different approach was made in order to get a more accurate model of the imputed data. Using K-Nearest Neighbours, new values based on the most similar samples where computed for the missing data. In order to get more input on various imputation methods, a 3rd degree polynomial fit was made, as well as an iterative multivariate imputation method that models each feature with missing values as a function of other features in a Round-Robin fashion.

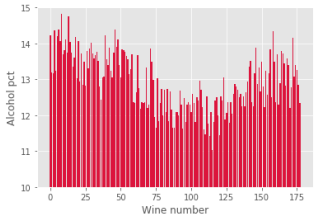

Figure 5: KNN

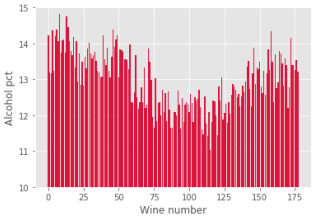

Figure 6: Multivariate

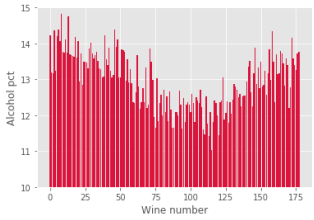

Figure 7: Polynomial

In order to compare all these different imputation methods, a cumulative relative error (CRE) was computed for each imputed dataset. This was done by computing the relative error of every data point of the imputed datasets when compared to the real values and then summing them all. To make it easier to visualise the accuracy of each one, a vector containing all the CRE's was normalised and then subtracted to a vector of the same shape (1,6) containing only ones. This way, a relative score was calculated for each tested imputation method.

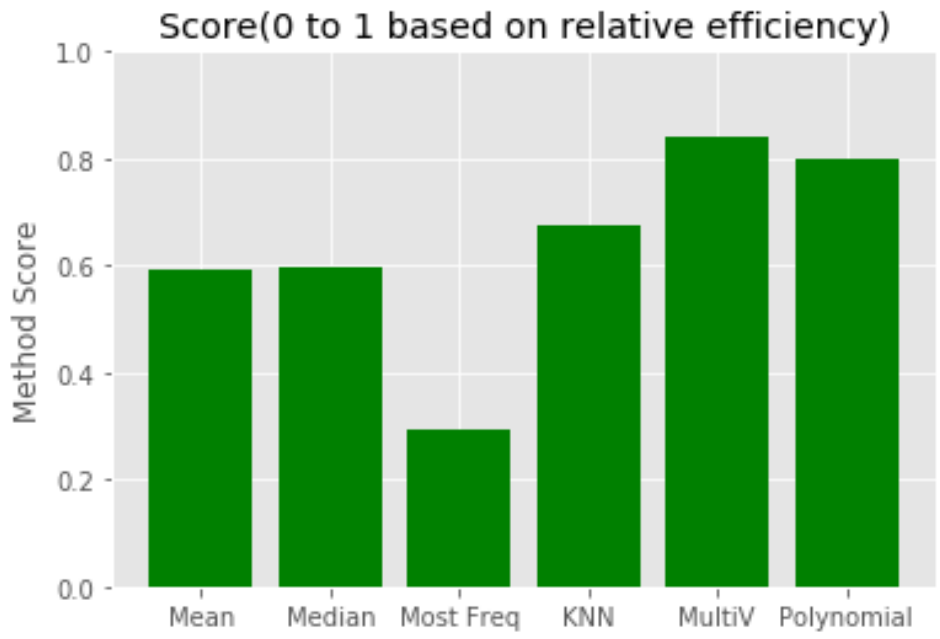

Figure 8: Relative Scores

The end result of the tests was predictable overall. The mean, median and especially the most frequent values imputers were not the most accurate and the KNN and multivariate were clearly better suited for this problem. However, the polynomial fit did surprisingly well when compared to a most complex and expensive (computation-wise) KNN imputer. Because of this, a new trial was made, using multiple polynomial fits (linear, 2nd degree and 3rd degree):

Figure 9: Linear

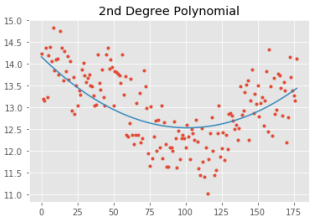

Figure 10: 2nd Degree

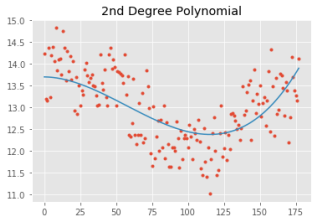

Figure 11: 3rd Degree

Just as in the previous imputation methods, relative scores were computed:

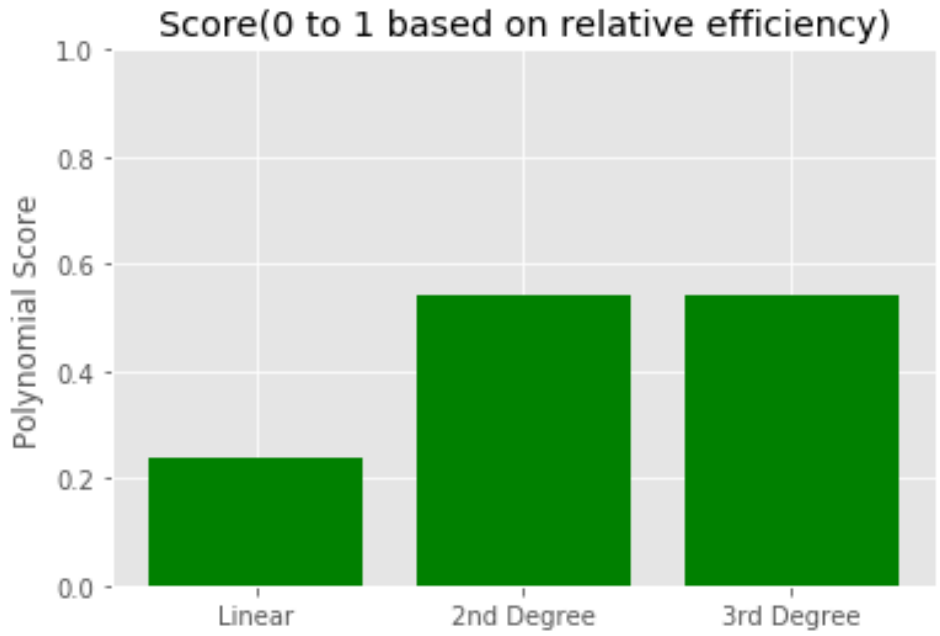

Figure 12: Relative Scores (Polynomials)

Even though there's a big difference between the accuracy of linear and 2nd Degree fits, if the scores of the 2nd Degree fit and the 3rd Degree fit are compared, we can conclude that there's no point in doing higher level polynomial fits. It is important to notice that these don't represent absolute scores, for instance it's pointless to compare this linear fit score to a previous mean imputer score.

In order to show the accuracy of multiple imputation when compared to single imputation, a MICE (Multiple Imputation by Chained Equations) was made and the resulting imputed data illustrated in the next bar chart.

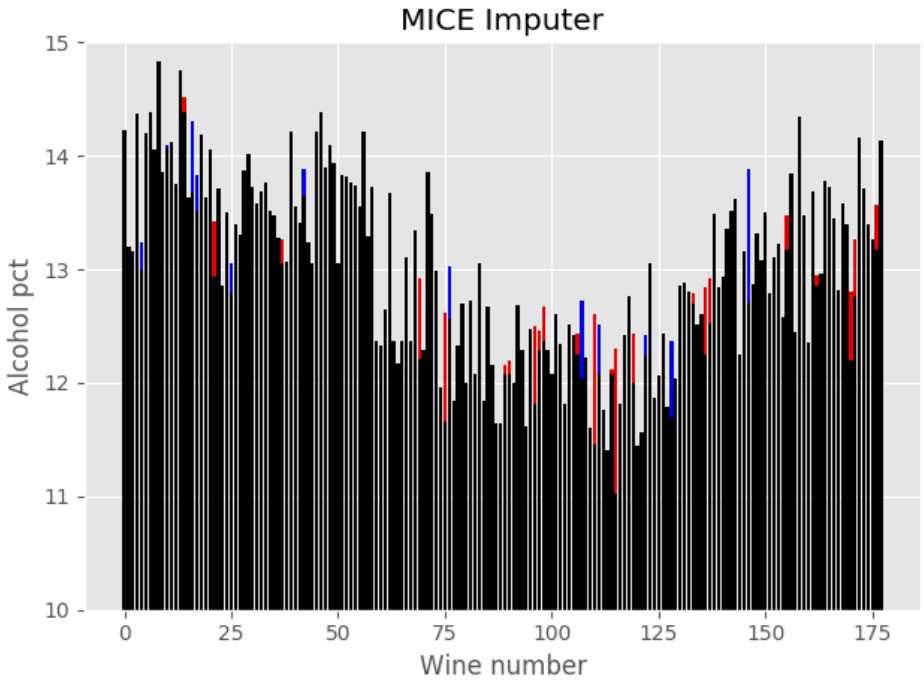

Figure 13: Imputed values for Multiple Imputation by Chained Equations. Overestimations are shown with red lines and underestimations are shown with blue lines

All in all, every imputation method has it's advantages and disadvantages (as mentioned above in this article), and their usage must be pondered based on the problem in question. For instance, it is remarkable that if a random deletion of values is repeated in every trial, the relative accuracy of each method may vary.

4 Non-trivial questions

- **Q:** What to do if it turns out that the assumption that the missing data is of type MAR doesn't hold?
  A: When faced with this problem there are multiple ways to proceed. Firstly we can try to make the assumption hold by finding additional data that is strongly predictive of the missingness, and include these into the imputation model. If all possibilities for such data are exhausted and if the assumption is still suspect, we have no choice but to accept that the missing data is of type MNAR. To deal with this type of data we have to perform a concise simulation study as in [7] customized for the problem at hand with the goal of finding out how extreme the MNAR mechanism needs to be to influence the parameters of scientific interest. Finally, use a non-ignorable imputation model to correct the direction of imputations created under MAR. Vary the most critical parameters, and study their influence on the final inferences [1].

- **Q:**How to determine if an imputation method is adequate for a particular problem?
  A:There's still a scarcity of tools to check if a certain imputation method is appropriate for a specific problem. There are however some statistical tests that can be performed in order to diagnose eventual problems in the imputation models. For instance, the Kolmogorov-Smirnoff (KS) test is a great way of flagging differences between observations and imputed values. The goal of this test is to compare distributions of imputed and observed variables and evaluate if a certain imputed value needs to be investigated in a more detailed fashion (using p-values as guidelines).

  Another more formal method is the usage of analysis of variance (ANOVA) where the outcome variable is the variable being imputed and the factors are the response stratum. Given that the accuracy of the imputation method is based on how many times the ANOVA test is rejected for a number of datasets.

  Standard regression diagnosis is yet another way of trying to figure out the accuracy of an imputation method. This is based on fitting a regression model to the missing values dataset (before imputation) and then performing the regression diagnostics proving if the regression is a good fit or not.

- **Q:** When should you use single imputation versus multiple imputation?

  A: We know that single imputation can be used when only a small percentage of the data is missing. However we have not quantitatively defined what "small" means. A great drawback of multiple imputation its complex nature. In order to get good results familiarity with the analysis is needed, but also how to combine the results. This can be complicated for people without a rigorous background in statistics. Therefore when the result from multiple imputation is the same as single imputation we would prefer to use single imputation. Much research has been done comparing the two methods, one study has compared six different imputation techniques on missing data in the Zung Self-reported Depression scale, of which five were single imputation and one multiple imputation [8]. Both missing at random and missing not at random simulations were performed. It was found out that when 10% of values were missing all of the single imputation methods except for random single imputation produced near perfect results. When 20% of the values were missing mean imputation and single regression still produced very comparable results to multiple imputation. Only when 30% of the data was missing multiple imputation started to produce substantially better results, but the single imputation methods were still in substantial agreement. Although results may differ for different datasets this should be an indication that multiple imputation is not always the best solution. When picking a method of imputation single imputation should also be considered based on the accuracy, statistical expertise of the researcher, and easy of interpretability for the reader.

5 Conclusion

Missing data can be classified as missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). Missing data that's not missing at random is the most difficult to deal with, while dealing with MAR and MCAR type data is significantly easier. There exist various methods to deal with missing data. Missing data can either be discarded, or we can use imputation methods to substitute missing values. There exist two types of imputation methods, single imputation methods and multiple imputation methods. Methods that fall into the second category generally produce better results, while methods from the single imputation category are less computationally expensive and are sufficient for simple cases.

Bibliography

[1] Stef van Buuren. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press LLC, 2018.

[2] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Pearson Addison Wesley, 2006.

[3] Iris eekhout — missing data. https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/, 2019.

[4] Paul J. Allison. Handling missing data by maximum likelihood. 2012.

[5] Roderick J. A Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 3rd edition, 2019.

[6] Craig K Enders. *Applied missing data analysis*. Guilford Publications, 1st edition, 2010.

[7] Linda Collins, Joseph Schafer, and Chi-Ming Kam. A comparison of restrictive strategies in modern missing data procedures. *Psychological methods*, 6:330–51, 01 2002.

[8] Fiona M Shrive, Heather Stuart, Hude Quan, and William A Ghali. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. *BMC Medical Research Methodology*, 6(1), 2006.