Introduction to Data Science
2016

Team 16

Missing-data Imputation Notebook

1. Introduction and Motivation

This study material aims to explain the problem of missing data and they key methods to handle the issue. This problem arises in almost all serious statistical analyses due to unavailability of the data, censoring of the data, non-response rate or just human error. This issues will make it harder to reach the desired analysis, therefore the data need to be cleaned in order for it to be analised fully. This paper will go through the main missing gata types, propose some solutions to fill the data with relevant information, propose some non trivial applications of the techniques on real world examples.

2. Missing-data Types

In order to be able to handle the missing data, its very important to know why the data is actually missing. For missing mechanisms are outlined in this section. For further easiness of understanding we define the term “variable” as an input/column in the data set

2.1 Missingness completely at random

A variable is missing completely at random infer that the probability of missingness is equal across all the data. This can be illustrated for example as if each survey respondent decides to answer to a certain variable by throwing a dice, and if “1” is the result he will not answer. In the case of missingness completely at random, replacing the missing data with the mean of the registered values, or simply removing the data will not influence in any way the final result of the analysis.

2.2 Missingness at random

In this case missingness is not longer randomly distributed, instead the missingness rate depends on the other observed variables. For example in an earnings survey, people why are white have a greater chance of not completing the earnings variable that other people. The data can be modeled by including the observed predictors when generating the missing data, in this way the nonresponse bias can be avoided.

2.3 Missingness that depends on unobserved predictors

Missingness depends on the information that has not been recorded, therefore the missingness rate is not at random, the unobserved data predicts the missing values. It must be explicitly modeled in order to avoid bias in the inferences.

2.4 Missingness that depends on the missing value itself

The probability of missingness depends on the variable itself. This situation may also be called censoring, the data which is missing cannot be modeled according to other predictors because the relevant predictor for it is the variable itself.

3. Main Study

Imputation is one of the key methods that researchers use to fill in missing data in a dataset. This can be done by using various methods in order to impute the most probable answer for more accurate data analyses. We can divide these various methods for imputing data in 3 different categories which are **Case Deletion**, **Single Imputation** and **Multiple Imputation**.

3.1 Case Deletion

Deletion techniques are most the basic and traditional methods to handle with missing data and it is also the most common method in statistical software. These techniques simply involves discarding and excluding data that are missing. There are two deletion methods which are **Complete Case Analysis** and **Available Case Analysis**.

Complete Case Analysis
With complete case analysis, or listwise deletion, all cases in the dataset with one or more missing variable will be excluded from the analysis. The advantages of this method is that we will have a complete dataset excluding the missing values. On the other hand, we will have a smaller dataset with reduced power. This method is the default option in many statistical procedures in many statistical software packages and it is also the most frequently used method.

Available Case Analysis
In available case analysis, or pairwise deletion, missing values are excluded based on the analysis we want to do. If the case contributes to one analysis, we will analyse the available data, if not it will be excluded, but the case is not fully excluded from the dataset compared to **Complete Case Analysis**. With this method, the sample size will remain the same for some analyses and will be reduced for others. The disadvantage is that the assumption of the MCAR (Missing completely at Random) mechanism to produce unbiased estimates. Also, by using varying sample size, it can lead to problems in computing standard errors.

3.2 Single Imputation

With single imputation, we will replace the missing value with a value. WIth this method, the sample is retrieved. The imputed data are assumed to be the real values of the dataset when the data would have been complete. For single imputation, we have 4 methods to impute missing data and these methods are **Mean imputation**, **Last Observation Carried Forward**, **Regression imputation** and **Hot-deck imputation**.

Mean Imputation

The mean imputation is one of the easiest ways to replace missing data. The process implies replacing the missing values with the mean of the registered values. In figure one this is illustrated by representing the imputation points with pink

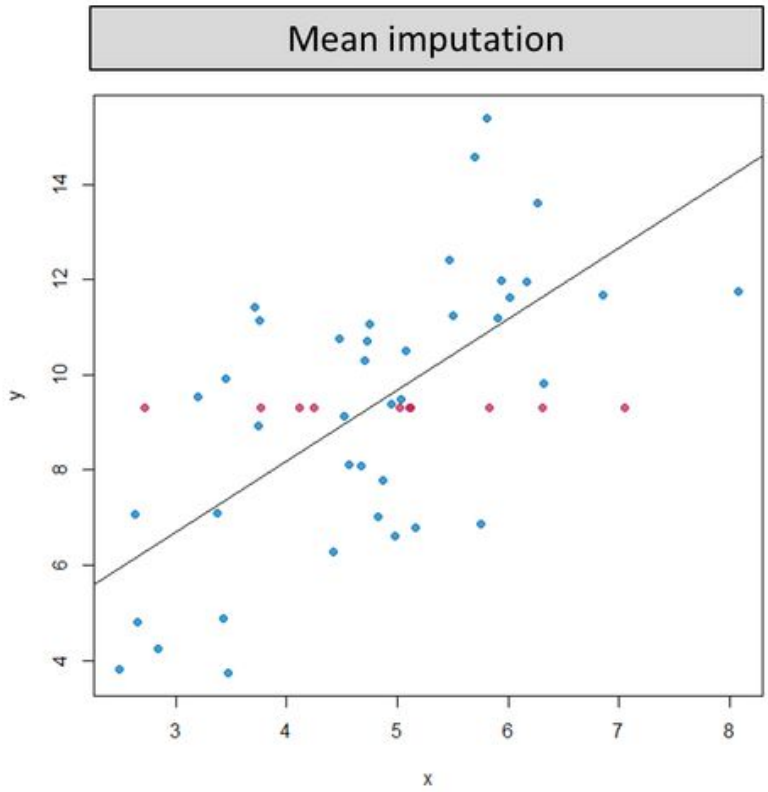


Figure 1. Mean Imputation Illustration

Downsides:

- May lead to distortion in the distribution of the variable
- Underestimates the standard deviation

Advantages:

- Very easy to implement
- Low computational resources needed
- May give good results if data is (almost) missing completely at random

Last Observation Carried Forward (LOCF)

The last Observation Carried Forward method replaces the missing data with the last available value, which is carried forward. In Figure 2, an example of applying the LOCF method is outlined, along with its possible downsides.

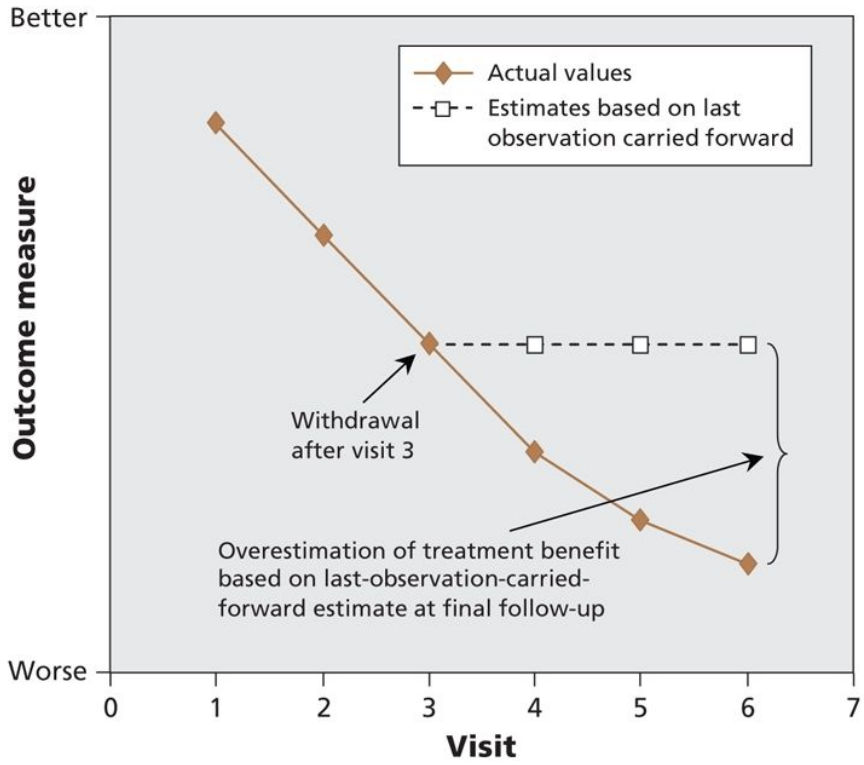


Figure 2. Example of LOCF method applied to clinical drug study

Downsides:

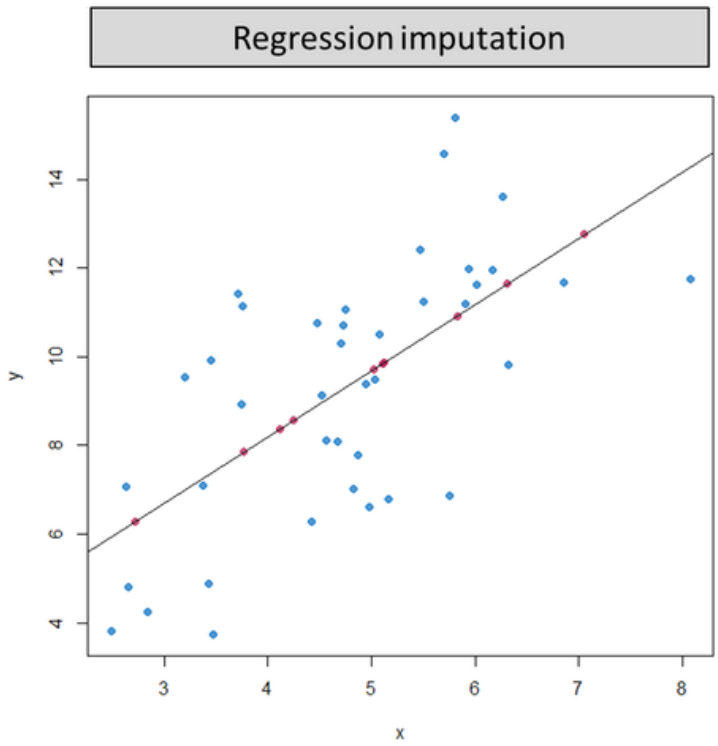
- Means and covariance structure are seriously distorted
- It does not take into consideration the trend of previous versions

Advantages:

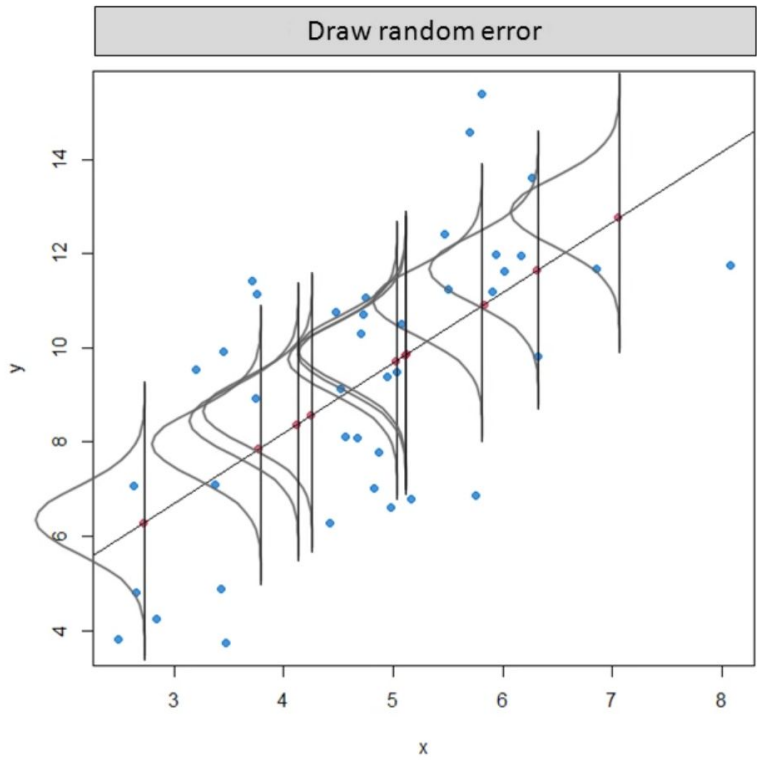
- No extra computational power needed
- Can perform good if data is missing completely at random

Regression Imputation

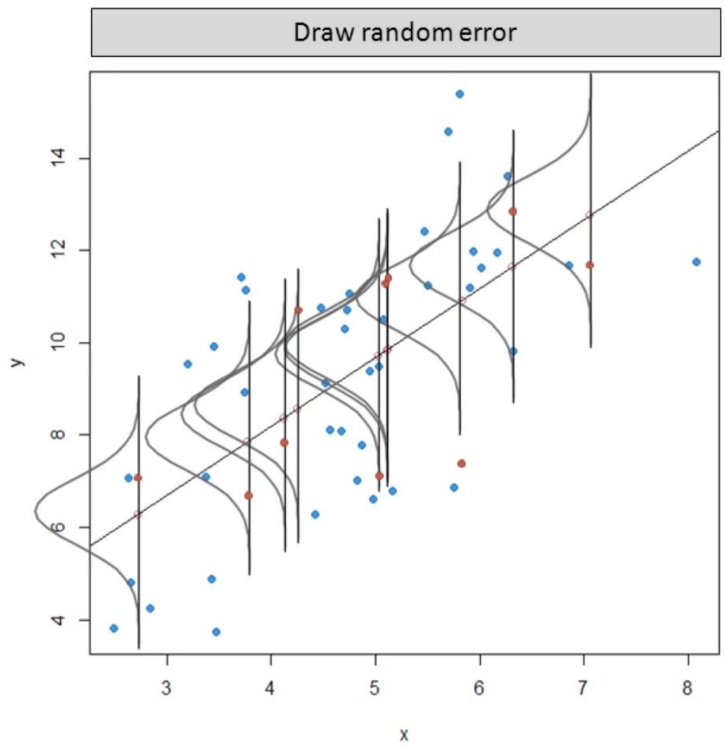
For regression imputation, we have two ways of imputing data based on a regression which are single regression imputation and stochastic regression imputation. In single regression imputation, the imputed data is predicted from a regression equation. The complete observations in the dataset are used to predict the missing observations. This regression method assumes that the imputed values fall directly on a regression line with a non-zero slope meaning the correlation between the predictors and the missing values is 1. With this method, the correlations will be overestimated. However, the variances and covariances are underestimated. Here below, you may see the imputed values (in pink) are predicted based on a regression line of the complete observations.



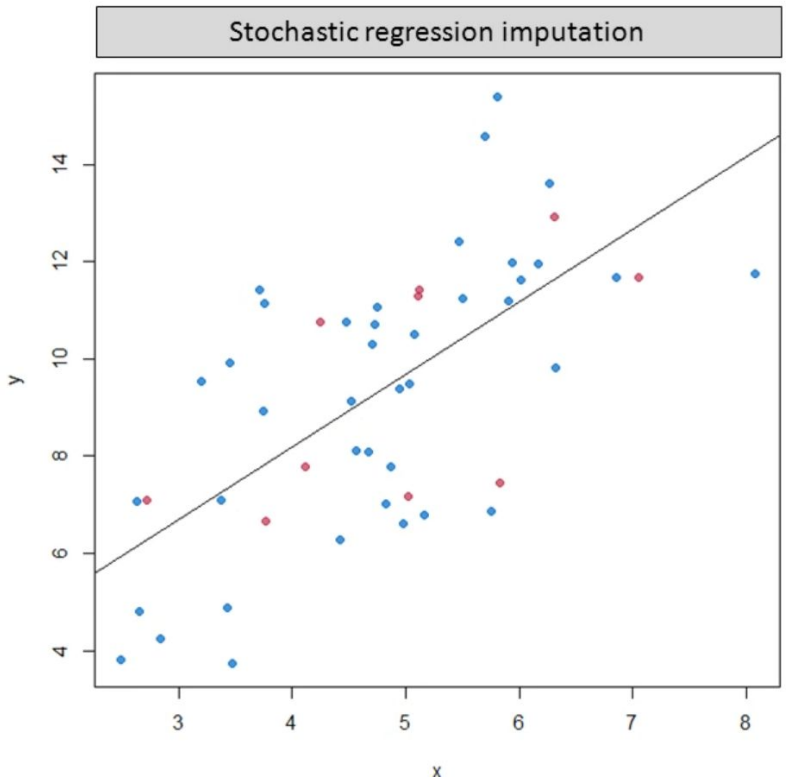
On the other hand, we also have stochastic regression imputation which aims to reduce the bias by an extra step with a residual term. The residual term is normally distributed with a mean of zero and a variance equal to the residual variance from the regression of the predictor on the outcome. The error from a normal distribution is added to the imputed value. By using this method, we can see more variability in the data and the estimates of each parameter are unbiased with MAR (Missing at Random) data. The disadvantage is that the standard error is underestimated, because the uncertainty about imputed values is not included, which can increase the risk of getting type I errors. Below, you may find the illustrations how the predicted values are computed.



From the regression line, we can draw a normal distribution on each imputed value.



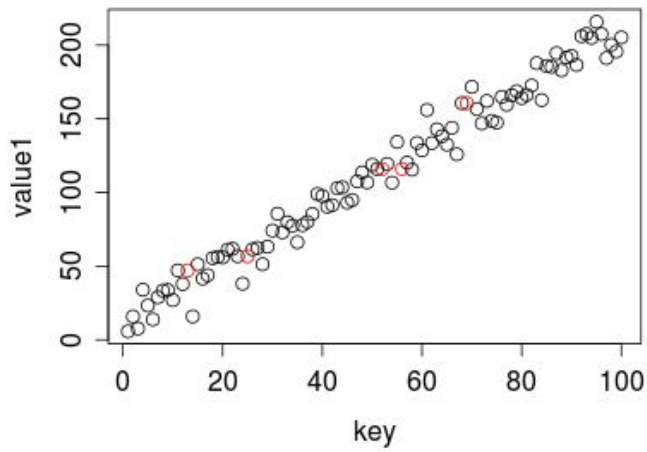
We can then randomly add the errors from the normal distribution



Here, we can see a more varied dataset

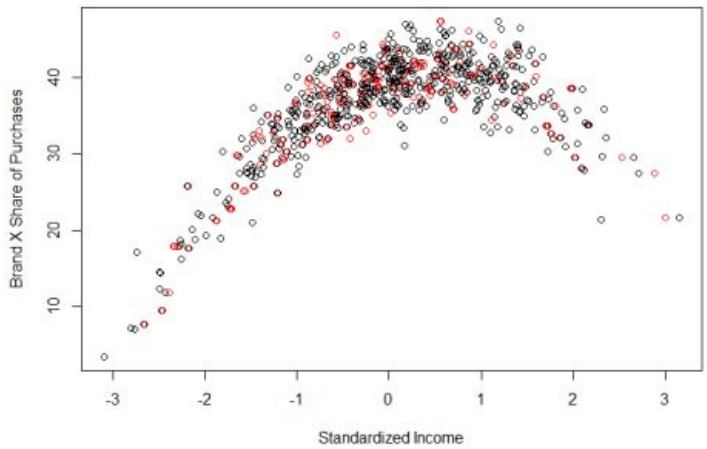
Hot-Deck Imputation

Hot-deck imputation is a technique where non-respondents are matched with respondents and the missing value is imputed from a value of the similar respondent. There are two hot-deck approaches which are the distance function approach and the pattern matching approach. The distance function approach, also called nearest neighbor approach, imputes the missing value with from the nearest neighbor with the smallest squared distance statistic to the case with the missing value.



The red points are the imputed values from their nearest neighbor

The matching pattern approach is more common, where the data is classified into separate homogenous groups. The missing value is then imputed randomly from another case in the same group in which the case with the missing value belongs to.



Here you can see how data looks like when using the matching pattern approach

Hot-deck imputation replaces the missing data by realistic scores that preserve the variable distribution. However, this can underestimate the standard errors and the variability. Hot-deck is most commonly used in survey research.

3.3 Multiple Imputation

- **What is multiple imputation ?**

  Multiple imputation is a statistical technique for analyzing incomplete data sets for which some entries are missing. The main difference between single and multiple imputation is that each missing value is replaced with a set of slightly different values. This set consists of predictions about the probable value but each prediction is derived from a different imputation method.

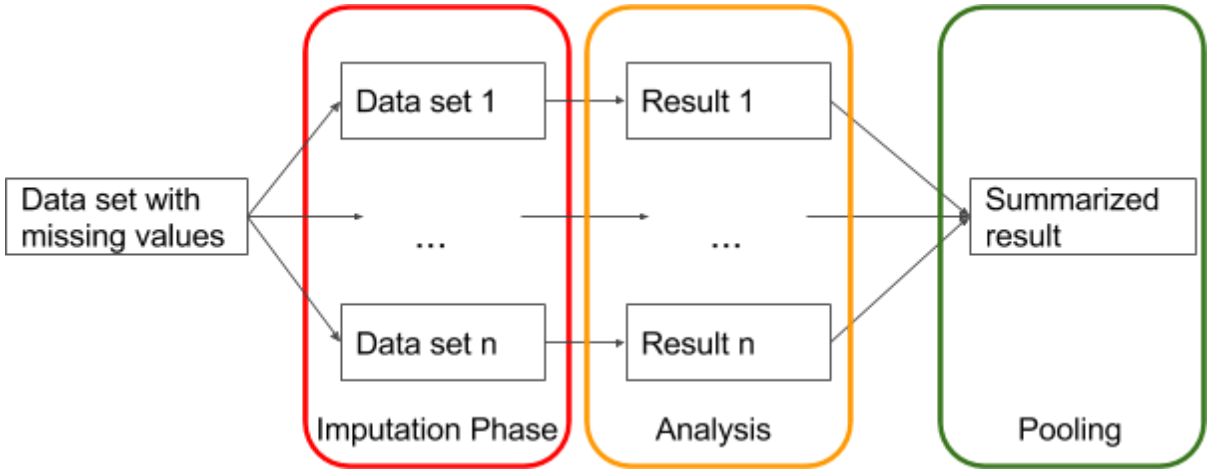
- **When do we use it ?**

  Multiple imputation gives great flexibility as it works well when the data is missing at random (MAR), missing completely at random or the data is missing not at random. Since MI achieves higher accuracy in comparison to single imputation it comes handy when we want to reduce the noise or achieve higher precision in analysis. Another advantage is that SI methods process the missing data with certainty (as we may know it) which may result in biased results and may lead to wrong conclusions while MI yields results which are closer to the real world (as if we knew them).

- **What are the steps?**

In 1987 Rubin developed a method for averaging the imputation results across multiple imputed data sets.

According to him each imputation method should follow the described steps :



1.The **imputation phase** consists of creating a number of copies of the data set (i.e. 5,10) which differ only in the imputed values of the missing data.

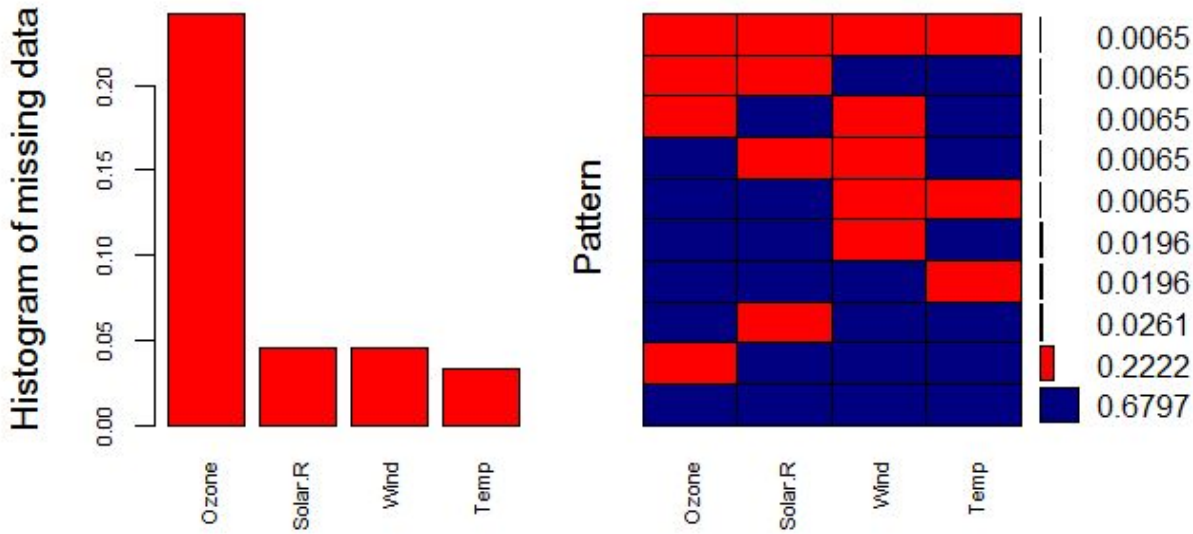
2.**Analysis phase** : each of the imputed datasets is analyzed independently as if the data was complete.

3.**Pooling :** The results from the analyzed sets are consolidated into a single one. For each missing value of the initial data now there are several predictions which reflect our uncertainty about the real value. A final result can be calculated by averaging the values from the imputed data sets.

One of the the most used methods which comply this specification is **M**ultiple **I**mputation by **C**hained **E**quations (MICE). This method consists of the following steps

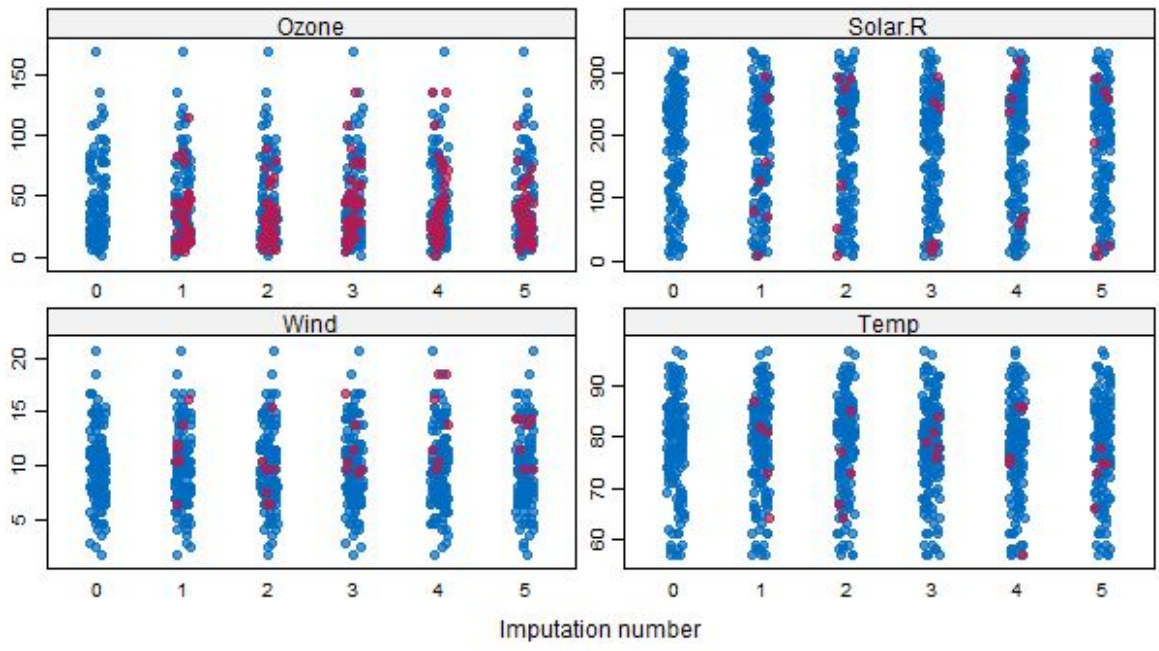
1. **Perform simple imputation** - Once the data set is replicated several (i.e. 5) times all missing values are replaced using a simple, single imputation method (i.e. mean imputation). As a result there are several “complete” sets.
2. **Set back to missing the imputed values** - In case more than one variable is missing then set back the values of one of the missing data variables to empty.
3. **Perform regression for the missing values** - Using the imputed values for the other missing variables apply regression in order to predict the possible values.
4. **Set the predicted values** - After the regression, replace the missing values with the result
5. **Repeat 2-4 for another variable** - After the execution of the process for the first variable apply this method for another variable.
6. **Repeat 2-4 until convergence** - Repeat the process until the predicted values for each variable converge

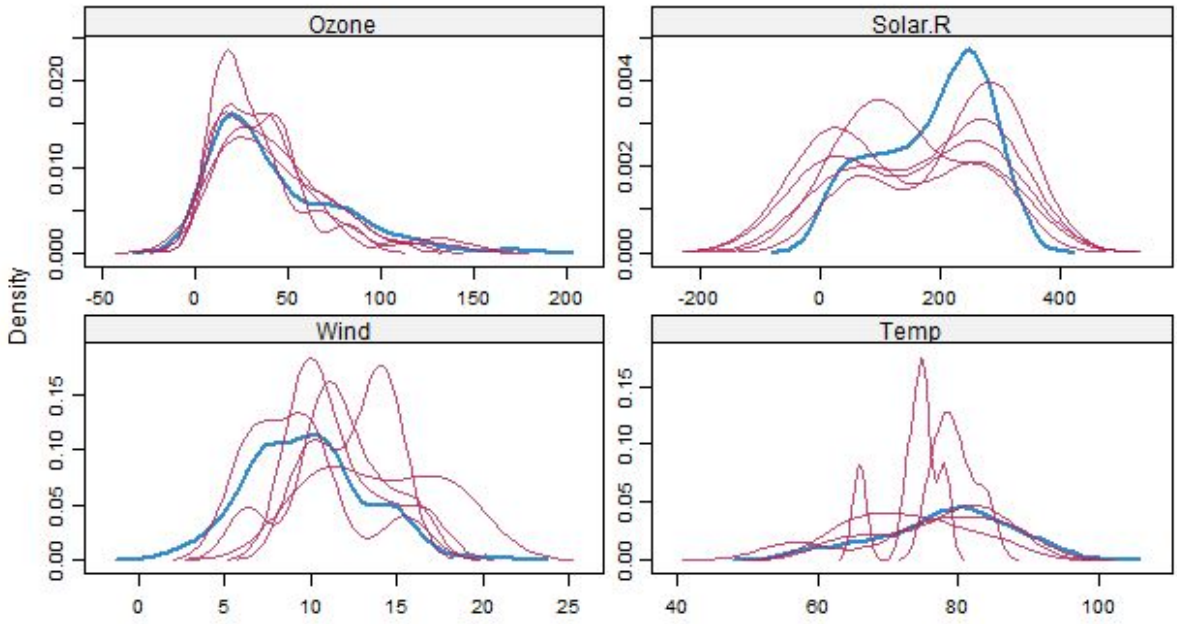
Both the Analysis and Pooling phases are also topic of interest but they are much more simpler than the Imputation Phase. That's why the following example observes mainly this phase. Let's say we are given a data set which contains information about the air quality of a certain city, which has missing values for some measurements (some of them are intentionally removed for experimental purposes). The following graph shows the distribution of the missing parameters, and the pattern of their missingness.



As seen from the graph ~68% of the data is "complete", ~1% has missing the Temp data etc. Having this set-up the process of imputing may begin. The following experiment uses MICE method.

The following graphs show the original data (in blue) compared to the final data set (in magenta).





As seen from the graphs while some of the methods give results close to the real data, others produce high error.
The following tables show the comparison between the original data set, and the final result with filled missing values.

Initial data set

	Ozone	**Solar.R**	**Wind**	**Temp**
Min	1.00	7.0	1.700	56.00
1st Qu.	18.00	115.8	7.400	72.00
Median	31.50	205.0	9.700	79.00
Mean	42.13	185.9	9.958	77.88
3rd Qu.	63.25	258.8	11.500	85.00
Max	168.00	334.0	20.700	97.00
NA's	37	7	0	0

Data set after pooling

	Ozone	**Solar.R**	**Wind**	**Temp**
Min	1.00	7.0	1.700	56.00
1st Qu.	18.00	**115.0**	7.400	72.00

Median	31.50	**203.0**	9.700	79.00
Mean	**40.76**	**185.8**	9.958	77.88
3rd Qu.	63.00	**259.0**	11.500	85.00
Max	168.00	334.0	20.700	97.00

As seen from the bolded values the result after applying the multiple imputation is really close to the initial data set which means that the error rate should be small.
For full explanation of the experiment see [1].

4. Non-trivial questions

4.1.Case deletion

"When is the Complete Case Analysis useful despite the reduction of the sample size?"

- Complete Case Analysis can be used when the data is Missing Completely at Random (MCAR), because the cases with missing data are no different than the complete cases - they are purely random subset of the data. When the incomplete cases that are dropped differ from the complete cases still in the sample, then the carefully selected random sample is no longer reflective of the entire population.
- Or when you have sufficient power anyway, even though you lost part of your data set. If the percentage of missing data is very small or you had an overly large sample to begin with, you may still have adequate power to detect meaningful effects.

4.2.Single Imputation

"When to use single imputation or multiple imputation?"

Single imputation involves less computation, and provides the dataset with a specific number in place of the missing data. In general, for single imputation, the missing data is replaced with another data available in the data set, or with a variation of all the data available, such as the mean value.
When only a little bit of data is missing, single imputation provides a useful enough tool. It fills in the data points well and the variance between the results of your analyses is unlikely to be altered by any significant margin. A drawback for using single imputation is that after imputing the data, it's treated as normal data, so adding more data to your initial set allows for misleading analysis.
Multiple imputations use simulation models that take from a set of possible responses, and impute in succession to try to come up with a variance/confidence interval that one can use to better understand the differences between imputed datasets, depending on the numbers that the simulation chooses to use for the missing data.

As a conclusion, if very little data is missing, single imputation is the simpler way to solve the problem without any serious errors and without the need to do the complex computations multiple imputation needs. But, if the data set is too large, you may need to account for the variability of the imputed data in order to find a range of possible responses that will generate a satisfactory result.

4.3.Multiple Imputation

"*How many imputed data sets do I need?"*

In his book Rubin provides an answer for this question. He proves that in most cases 3-5 imputed sets are enough for achieving excellent results. The approximate efficiency of an estimate is
$(1+\frac{\alpha}{m})^{-1}$
where α is the fraction of missing data and m is the number of imputed sets.

5. Conclusions

Although missing data can be a serious problem there are solutions. There are number of methods which range from easy to implement, which raise higher error for the imputed values, to hard to implement which give predictions close to the real world. If the missing values are not subject of the analysis then a method such as means imputation, last observation carried forward or any of the others which raise high error is sufficient. If the missing values are important for the analysis then a more sophisticated method should be used. In this case the complexity pays off in terms of accuracy. As already stated even when the missingness is 50% a more complex method manages to restore the data with accuracy ~90%.

6.References

[1] https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
[3] https://en.wikipedia.org/wiki/Imputation_(statistics)#Multiple_imputation
[4] http://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf
[5] http://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/
[6] http://statisticalhorizons.com/more-imputations
[7] http://sci2s.ugr.es/keel/pdf/specific/articulo/graham_olchowski_07.pdf
[8] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.
[9] http://surveymethods.com/blog/when-to-use-single-imputation-or-multiple-imputation/

7.Division of work

Student	Student Number	Topic
Isac Andrei	S3257053	Introduction, problem statement, single imputation
Win Leong Xuan	S3208435	Case Deletion, Regression imputation and Hot-Deck imputation
David Pavlov	S3187330	Multiple Imputation