Missing Values in Data

university of groningen
faculty of science and engineering

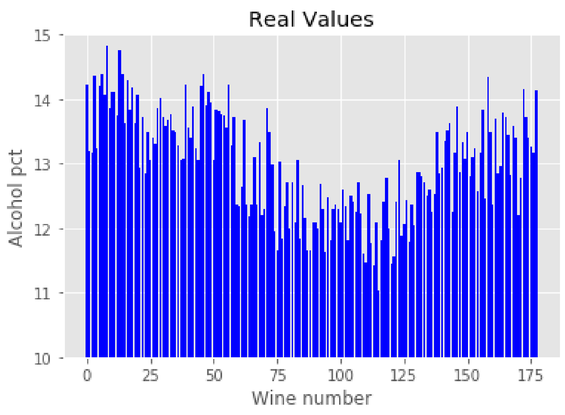
Otte Tjepkema, José Rodrigues, Andrei Miculita and Robert Riesebos
University of Groningen - Introduction to Data Science

Introduction and Problem Statement

Missing values are a common occurrence when dealing with cleaning and analysing data. There are multiple reasons why data goes missing, which will be laid out on this poster. These reasons 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. The advantages and disadvantages of these methods are discussed on this poster and illustrated with the help of an example dataset.

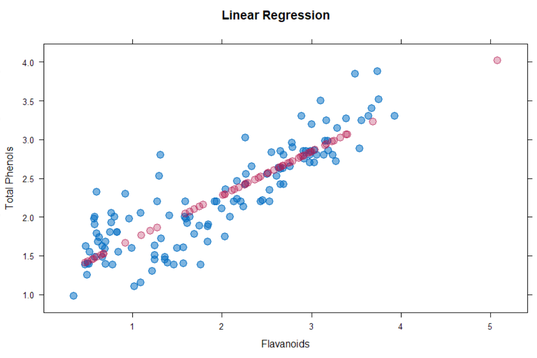
Dataset

- The dataset used is available in the scikit-learn python package [1].
- The data is the results of a chemical analysis of wines grown in the same region in Italy by three different cultivators. There are thirteen different measurements taken for different constituents found in a total of 178 types of wine.
- To reduce the scope only the first feature, the alcohol percentage, is plotted.
- The original distribution is included for reference below.



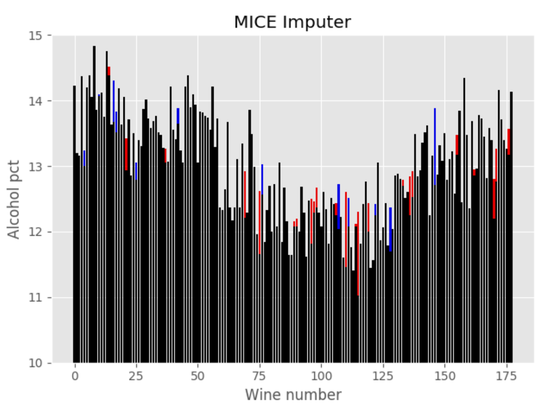
Regression Imputation

- The imputed values are predicted from a regression model created from the complete data vectors in the dataset.
- Advantages:
  - Takes into account correlation between features and preserves their relation
- Disadvantages:
  - Regression imputation artificially strengthens the relations between variables in the dataset.
  - The variability of the data is underestimated
  - The correct type of regression model needs to be chosen



Multiple Imputation [4]

- Generate multiple datasets, by replacing each missing value with a set of plausible values that represent the uncertainty about the right value to impute.
- The datasets are then analysed using the standard procedures for complete data. Valid statistical inferences are obtained by combining results from different imputed data sets.
- Advantages:
  - Never use a single imputed value
  - Can measure uncertainty surrounding parameter estimates.
- Disadvantages:
  - More computationally expensive, may need to parse entire dataset in order to perform multiple imputation.

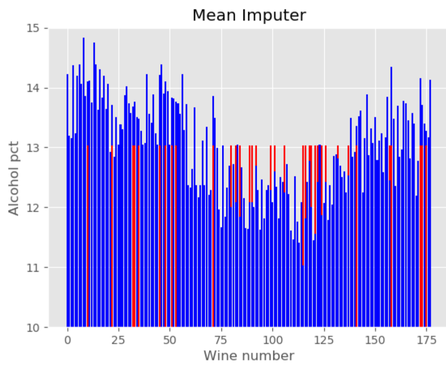


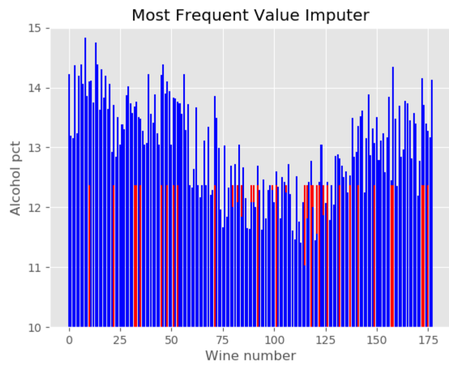
Types of missing data [2, 3]

- Missing completely at random (MCAR): The probability of missing is the same for all cases.
- Missing at random (MAR): The probability of missing depends on the other variables that we know.
- Not missing at random (NMAR): The probability of being missing varies for reasons that are unknown.

Mean imputation

- All the missing values of a feature are replaced with the mean value of the values that are available for that feature. Similar methods to this are median imputation or most frequent value imputation.



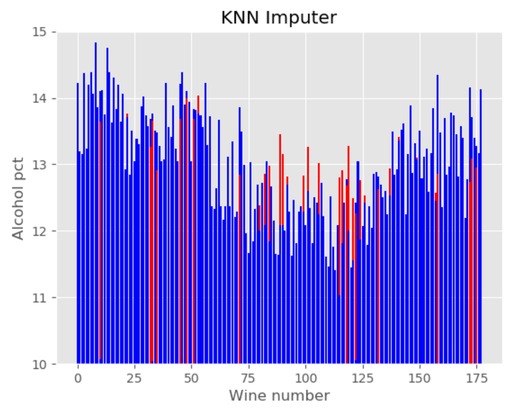


- Advantages:
  - Extremely easy to use and computationally inexpensive
  - No information about other features needs to be known
- Disadvantages:
  - Cannot be used when there exists a correlation between features
  - Underestimates the variance and disturbs the relations between variables

Hot and cold 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. K nearest neighbours (Knn) uses distance as a measure of similarity, but it is also possible to match categorical values (random hot deck imputation).
- Another method orders the dataset, usually with respect to a time variable, then if a value is missing it can be replaced with the entry before it.

- Advantages:
  - Variability of distribution is preserved
- Disadvantages:
  - Standard errors and variability are underestimated



Conclusion

- There are various imputation methods, each with its advantages and disadvantages. Multiple imputation seems to be the most advantageous when computational resources are not an issue. It is, however, arguable whether it is better for prediction and classification than single imputation when the uncertainty due to missing values does not have to be measured.

References and Acknowledgements

[1] Scikit-learn package. `https://scikit-learn.org/stable/`.

[2] S. Buuren. *Flexible imputation of missing data*. CRC Press, 2nd edition, 2018.

[3] Pannekoek J. Scholtus S. de Waal, T. *Handbook of statistical data editing and imputation*. Wiley, 2011.

[4] Donald B. Rubin. Multiple imputation for nonresponse in surveys. 1987.