Processing of the missing values, multiple imputation, MICE

Grigory Malinovsky

Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics

MIPT, Moscow, 2019

Paper presentation

- https://www.jstatsoft.org/article/view/v045i03/v45i03.pdf
- https://arxiv.org/pdf/1509.04992.pdf
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/

Motivation

In statistics, imputation is the process of replacing missing data with substituted values.

problems of missing data

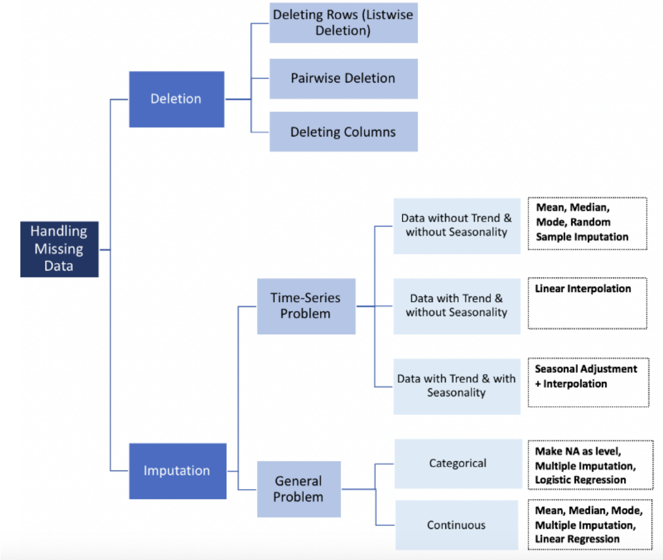
- introduce a substantial amount of bias
- make the handling and analysis of the data more arduous
- create reductions in efficiency

Types of missing data

Types

- Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value or missing value is dependent on some other variable's value

Overview



Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Deletion

- Listwise deletion (complete-case analysis) removes all data for an observation that has one or more missing values. It assumes that the missing data are MCAR.
- Pairwise deletion analyses all cases in which the variables of interest are present and thus maximizes all data available by an analysis basis. It assumes that the missing data are MCAR. If you delete pairwise then you'll end up with different numbers of observations contributing to different parts of your model, which can make interpretation difficult.
- Dropping variables: sometimes you can drop variables if the data is missing for more than 60% observations but only if that variable is insignificant.

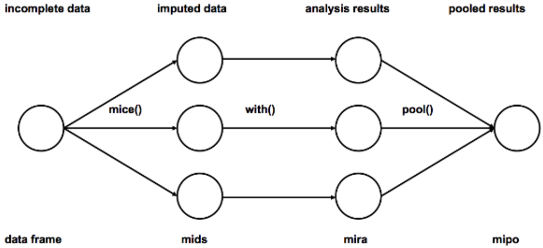
Hot and cold deck imputation

- A once-common method of imputation was hot-deck imputation where a missing value was imputed from a randomly selected similar record. The term "hot deck"dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot"because it was currently being processed.
- Cold-deck imputation, by contrast, selects donors from another dataset. Due to advances in computer power, more sophisticated methods of imputation have generally superseded the original random and sorted hot deck imputation techniques.

Single imputation

- Mean, Median and Mode
  Computing the overall mean, median or mode is a very basic imputation method, it is the only tested function that takes no advantage of the time series characteristics or relationship between the variables. It is very fast, but has clear disadvantages. One disadvantage is that mean imputation reduces variance in the dataset.
- Regression
  A regression model is estimated to predict observed values of a variable based on other variables, and that model is then used to impute values in cases where the value of that variable is missing. The regression model predicts the most likely value of missing data but does not supply uncertainty about that value.

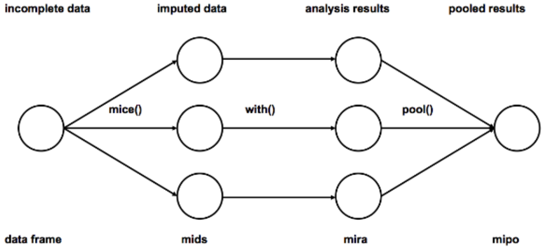
Multiple imputation



Source: http://www.stefvanbuuren.nl/publications/mice%20in%20r%20-%20draft.pdf

- Imputation – Similar to single imputation, missing values are imputed. However, the imputed values are drawn m times from a distribution rather than just once. At the end of this step, there should be m completed datasets.
- Analysis – Each of the m datasets is analyzed. At the end of this step there should be m analyses.
- Pooling – The m results are consolidated into one result by calculating the mean, variance, and confidence interval of the variable of concern.

Multiple imputation



Source: http://www.stefvanbuuren.nl/publications/mice%20in%20r%20-%20draft.pdf

- Imputation – Similar to single imputation, missing values are imputed. However, the imputed values are drawn m times from a distribution rather than just once. At the end of this step, there should be m completed datasets.
- Analysis – Each of the m datasets is analyzed. At the end of this step there should be m analyses.
- Pooling – The m results are consolidated into one result by calculating the mean, variance, and confidence interval of the variable of concern.

MICE

Algorithm

- Step 1: A simple imputation, such as imputing the mean, is performed for every missing value in the dataset. These mean imputations can be thought of as “place holders.”
- Step 2: The “place holder” mean imputations for one variable (“var”) are set back to missing.
- Step 3: The observed values from the variable “var” in Step 2 are regressed on the other variables in the imputation model, which may or may not consist of all of the variables in the dataset. In other words, “var” is the dependent variable in a regression model and all the other variables are independent variables in the regression model.
- Step 4: The missing values for “var” are then replaced with predictions (imputations) from the regression model. When “var” is subsequently used as an independent variable in the regression models for other variables, both the observed and these imputed values will be used.

This demands MAR assumption.

MICE

- Once the designated number of cycles has been completed, the entire imputation process is repeated to generate multiple imputed datasets.
- Initial research indicated that 5–10 imputed datasets was sufficient; however, recent research suggests that, depending upon the amount of missing information in the data, increasing that to as many as 40 imputed datasets can improve power.
- For example, creating a single imputed dataset that has hundreds of variables, thousands of cases, and missingness ranging from less than 5% to 80% could take hours to run and therefore, it may be impractical to create 40 imputed datasets.

Suggestions

- The second aspect of having an imputation model that is more general then the analysis model is including additional ("auxiliary") variables in the imputation process—variables that are not going to be used in the analysis but that can improve the imputations
- To further enhance the imputation model and the creation of valid imputations, bounds and restrictions are useful specifications to impose upon some variables, and these are very easy to specify in some MICE software packages.
- Large datasets naturally lead to the possibility of a very large number of variables to include in the imputation regression models, and it may not always be possible to include all of those variables identified for potential inclusion. Stepwise regression is one method of identifying variables for inclusion in the individual regression models.