Train In Data

MICE

MICE

Multivariate Imputation of Chained Equations

A series of models whereby each variable is modelled conditional upon the other variables in the data.

Each incomplete variable is imputed by a separate model.

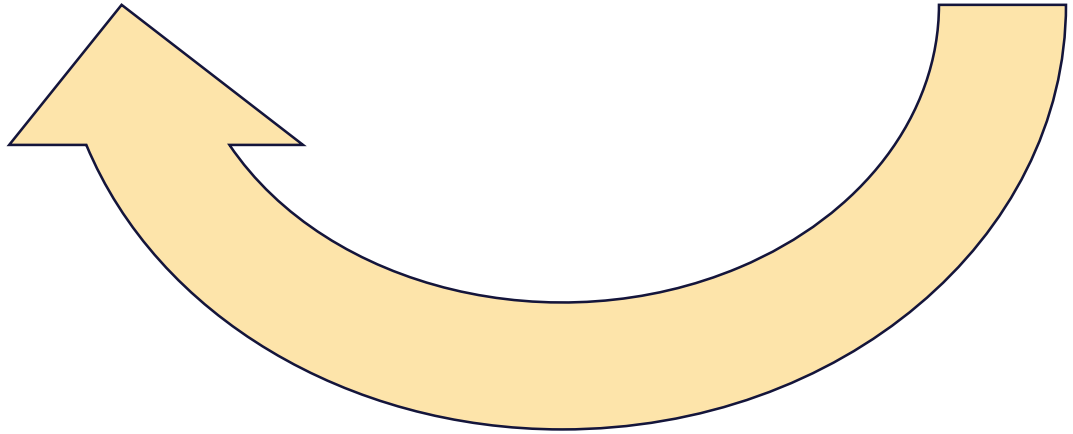
MICE: framework

Mean imputation to fill NA

Values in 1 variable are reverted to NA

The variable is modelled upon the other variables

NA are replaced by the model predictions



MICE: framework

After all variables with NA have been modelled based on the other variables, 1 round of imputation is completed.

The procedure repeats itself n times, usually 10 imputation cycles are enough to find stable parameters for the models.

MICE: why multiple rounds?

In the first round, we are modelling the variables based on the other ones, which themselves may contain NA.

- Thus, the predictions might be biased.

As we continue to regress one variable upon the others:

- We obtain better estimates for the NA,
- These estimates are used to regress the other variables,
- Thus returning more accurate predictions.

MICE: assumptions

- Data is MAR
- The NA in the variables can be modelled by the other variables in the dataset, and does not depend on external sources.

MICE

Considerations

MICE: variable relationship

Variables may have linear or non-linear relationships

- Find best model to predict the missing data, i.e., Linear Regression, Bayes, tree based algorithms, etc.
- Optimise the model parameters.

MICE: variable nature

Depending on the nature of our variables, we should use different models.

- Binary variables should be modelled with classification algorithms
- Continuous variables should be modelled with regression algorithms
- Discrete variables should be modelled with Poison

Not possible to automate with current tools. We would have to train each model manually.

MICE: Which variables should we use as predictors?

Authors suggest that using every available bit of available information yields multiple imputations that have minimal bias and maximal certainty.

➔ the number of predictors in should be as large as possible.

- Include all variables that will be used in the final model
- Add variables thought to be somehow related to the introduction of missing data.

MICE: more considerations

- "Circular" dependence can occur: same observations show NA on several variables ➔ the variables may be correlated

Train In Data

THANK YOU

www.trainindata.com