R Packages: VIM and VIMGUI

Presented by Geoffrey Hubona

VIM and VIMGUI

- The R package **VIM** (Templ et al., 2013, 2012) was developed to:
  1) explore and analyze the structure of missing values in data using graphical methods;
  2) to impute these missing values;
  3) to verify the imputation process using visualization tools; and
  4) to produce high-quality graphics for publications.
- **VIMGUI** was developed as a graphical user interface version of **VIM** to give access to users with limited R skills.

VIM and VIMGUI Imputations

- The R package **VIM** (Templ et al., 2013, 2012) has three imputation technique implemented:
  1) Hot-deck;
  2) K-Nearest Neighbor (**kNN**); and
  3) Iterative Robust Model-based Implementation (**IRMI**).
- **VIMGUI** also supports:
  4) Individual regression imputation where users can specify a (formulaic) model for regression imputation using ‘point and click.’

Hot-Deck imputation

- Uses popular sequential and random **hot-deck** algorithm with the option to use it within a ‘domain’
- **Hot-deck** is faster in computational speed that the others but may not produce the same quality imputations
- The most important **hot-deck** arguments are:
  - **`data:`** data set containing missing values;
  - **`variable:`** vector of variable names for which missing values should be imputed;
  - **`ord_var:`** vector of variable names for sorting; and
  - **`domain_var:`** vector of variable names for building domains and to impute within these.

kNN imputation

- Is based on ***Gower distance***, but distance variables can be binary, categorical, ordered, continuous, and semi-continuous
- Entire distance matrix is not calculated, so can be used with large data sets
- Most important **kNN** arguments are:
  - **`data, variable:`** see previous;
  - **`dist_var:`** vector of variable names used to calculate distance;
  - **`weights:`** vector of weights to be used for each distance variable;
  - **`numFun:`** function for aggregating k nearest neighbors if numerical, defaults to median;
  - **`catFun:`** function for aggregating if categorical;
  - **`addRandom:`** boolean variable if needed to add variable with only random numbers to avoid multiple selection of same donor.

Iterative Robust Model-Based Imputation (irmi)

- In each step of the inner loop iteration, one variable is used as a response variable and the remaining variables serve as regressors
  - It repeats until convergence.
- Data can be mix of binary, categorical, count, continuous and semi-continuous.
  - **irmi algorithm** selects correct regression method, based on data types, in an automatized manner.
- Most important **irmi** arguments are:
  - **`robust:`** boolean variable to enable robust regression;
  - **`step:`** boolean variable to enable stepwise selection of regressors;
  - **`mixed:`** column index of the semi-continuous variables;
  - **`count:`** count index of the count variables.

Individual Regression Imputation

- Through the **VIMGUI** a formula can be specified for defining a model that describes a single variable using a combination of explanatory variables for regression imputation.
    - **VIMGUI** guides user to formulate a certain regression model, or formula, using a ‘point and click’ mechanism.