YIQIAN WU

Part1: Mice (discuss all methods)-----

In this project, we use several methods to perform imputation. The following are imputation methods we used in Mice library : Random Forest, PMM, Cart, Sample, Fast PMM, Norm.boot, 2lOnly.norm.

Multivariate Imputation by Chained Equations (MICE) library generates multiple imputations for incomplete multivariate data by Gibbs sampling. It uses the posterior distribution to build Markov Chain. After given vector variables with specific distribution, the Gibbs sampler uses the set of full conditionals of the posterior distribution of interest to sample indirectly from the full posterior distribution. Algorithm of Gibbs sampling is described as following.

1 set t = 0

2 generate an initial state x0 ∈ domain of p(x1,x2,x3,…xn)

3 repeat until t = M

4 for each dimension I = 1…n

5 draw xt from p(xi|x1,x2,…,xi-1,xi+1,…,xn)

6 set t = t + 1

In MICE library, each incomplete column must act as a target column, and has its own specific set of predictors which always are other covariates in the dataset. Then, MICE uses different models selected by users to perform imputation.

Random Forest imputation method calls randomForest() function for classification and regression and then use the predict result to fill missing values

Predictive Mean Matching (PMM) estimates a linear regression of missing value and produces a set of coefficients. After that, it will use Gibbs sampling to make a random draw from the posterior predictive distribution of coefficients and create a new set of coefficients. These new coefficients will be used in linear model to predict missing values.

Cart is an imputation method using classification and regression trees. Like random forest, it will call a tree function and then using this function and other covariates to create a model to predict missing values.

Sample method uses Gibbs sampling to perform a random sample from the observed values. The imputation of missing values is come from samples of the same covariate without apply any model.

Like PMM, FastPMM also use predictive mean matching to impute dataset. It is a wrap of C++ function applied in R to achieve a faster speed.

Norm.boot imputes univariate missing data using linear regression with boostrap. It draws a bootstrap from covariates and calculates regression weights and imputes with normal residuals.

2lOnly.norm imputes univariate missing data at level 2 using Bayesian linear regression analysis. Variables are level 1 are aggregated at level2 .

Part2: What data is needed to change a ‘other’ level in GOOD or BAD?

When the return output probability is within (0.3,0.7) it need to change a ‘other’ level.