Class 4: Missingness in machine learning

Andrew Parnell
andrew.parnell@mu.ie

PRESS RECORD https://andrewcparnell.github.io/mda_course

In this class . . .

- Missing data analysis on large data sets
- Introduction to `mlr`/`mlr3` and missingness
- Other packages that perform missing data analysis
- Further resources on missingness

Large data missing data analysis

- In *omics data we often have number of variables greater than the number of samples (small n large p)
- Absent missing data, the usual approaches involve dimension reduction, variable selection or regularisation approaches such as the lasso
- With (ignorable) MAR data we need to incorporate the approaches into the imputation model which can often require bespoke code
- With NMAR data, selection models possibly the best approach as the high-dimensional regression model can be left unchanged (though the classification model will get harder to fit)
- Other problems might occur if the target variables are high dimensional (not covered here)

A high dimensional selection regression model

We might write such a model as:

$$y_i = f(X_i) + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

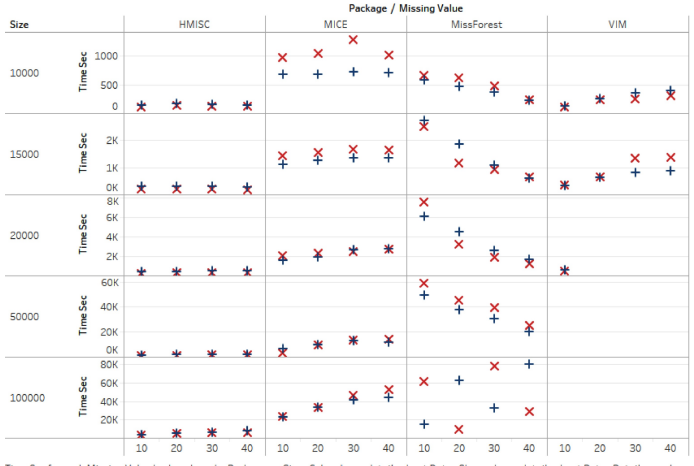
$$m_i \sim Bernoulli(p_i),\ \text{logit}(p_i) = \alpha + g(X_i) + \gamma y_i$$

- Now both f and g need to take account of the high-dimensionality of X_i
  - For the Fully Conditional Specification (FCS) versions of these models, an f_j will need to be proposed for every variable $y_i, X_{i1}, \ldots X_{ip}$

Long data FCS

- Yadav *et al*, Handling missing values: A study of popular imputation packages in R, Knowledge-Based Systems, 2018
- Compared R packages `VIM`, `mice`, `MissForest`, and `HMISC`
- Created 'fake' datasets by sub-sampling two classification data sets to contain 10k to 100k rows, and introduced 10-40% missingness. None of these
- They don't seem to have worried about whether this was MCAR, MAR or NMAR
- They evaluated their methods based on:
  - The time taken to do the imputation
  - The predictive performance of the classifier on the 'complete' data set
  - The variance of the imputed values (compared to the known true variances of the variable)

Results - time taken



Results - model performance