CHL5212 Assignment 1: Imputation

The purpose of this assignment is to compare the results of imputation techniques for different data structures. For each of the three datasets, split the data into a training and test set using the code provided (the first 250 observations) and do the following:

1. Randomly set 15 and 40 percent of the training set predictors to missing using the code provided and impute using three methods (mean, MICE with 'pmm' over 20 iterations and random forest with 500 trees over 5 iterations).
2. Train a linear regression model with all variables as predictors using the original training set and the two imputed training sets. Also train a null model (intercept only).
3. Get a measure of MSE on the test set for each of the three fitted models and the null model. Calculate the residual error by dividing the MSEs by the null model MSE.

Repeat the above, except use non-random missingness. Generate six plots (random, non-random missingness for each dataset) showing percentage missing vs. residual error for each of the three imputation methods.

1. Compare the covariance structures of the three datasets. How would you expect each to affect the quality of the imputation?

2. Compare the distribution of variables in each dataset. How would you expect this to impact imputation performance?

3. Contrast the performance of MICE and random forest, in what instance would they produce differentiated results?

4. Compare the results with random and non-random missingness, what scenarios/data structures produce the greatest discrepancies between MICE and random forest? Comment on the limits of imputation for maintaining predictive validity under different data structures and missingness types.

5. Residual error is a very aggregate measure, how else can you compare imputation quality?

6. What can you do to improve some of the metrics?

7. Comment on the computation times for each method vs. their performance, when is it most justified?

Evaluation:
Plots/outputs (35%): Correctness and presentation (labels, captions, etc.)
Interpretation (35%): Provide clear and concise answers of one or two sentences.
Implementation/coding style (30%): Follow proper coding practices (modular code, commented, etc.)

Submit a reasonably-sized PDF by the due date.

Due: May 19, 2020 before class (12:00pm).