Bayesian variable selection with missing data: An application to cardiology

Stefano Cabras[1], *María Eugenia Castellanos*[2], *Anabel Forte*[3], *Gonzalo García-Donato*[4], *Alicia Quirós*[5]

[1]stefano.cabras@uc3m.es, Department of Statistics, Universidad Carlos III de Madrid
[2]maria.castellanos@urjc.es, Department of Informatics and Statistics, Universidad Rey Juan Carlos
[3]anabel.forte@uv.es, Department of Statistics and OR, Universitat de Valencia
[4]gonzalo.garciadonato@uclm.es, Department of Economy and Finance, Universidad de Castilla-La Mancha
[5]alicia.quiros@unileon.es, Department of Mathematics, Universidad de León

Linear regression models are widely used in medical studies to identify the factors that influence a particular outcome or medical condition. Most commonly, data will have some missing values, and rarely is the missing mechanism completely at random. In medicine, the main methods of dealing with missing data is to eliminate cases with incomplete data or impute it using the mean of each variable. However, these may lead to inconsistency in the analyses and the potential alteration of the relationships between the response and the explanatory variables.

The main challenge faced by many researchers is variable selection with missing data, a topic that has received very little attention in the literature. Current proposals use multiple imputation to complete the database and Rubin's rules to combine the results, but their theoretical underpinnings are still unclear.

To investigate this, we propose an application of a new Bayesian variable selection method for linear regression models in the presence of missing data. We analysed the dataset of the PREDICT-MVI study, which aimed to test whether certain physiological indices calculated during the intervention for acute myocardial infarction could predict the extent of microvascular damage caused by the infarction. The dataset had a 63% of missing data, meaning that there were 22 complete observations, out of the 60 patients recruited.

The employed method reliably calculates Bayes factors (and thus the model posterior probabilities) for all possible models by performing a Monte Carlo approximation of the data densities in each model (known as marginals), which takes into account the variability due to data imputation. The benefits of our approach are: i) A Bayesian method with a probabilistic foundation; ii) The results of variable selection are directly interpretable as they are the probabilities that those are included into the true model; iii) All collected data are used and no subjects are discarded; iv) The method is readily available as it uses already implemented R packages.

Keywords: Model posterior probabilities; Objective priors; Myocardial Physiology indices.